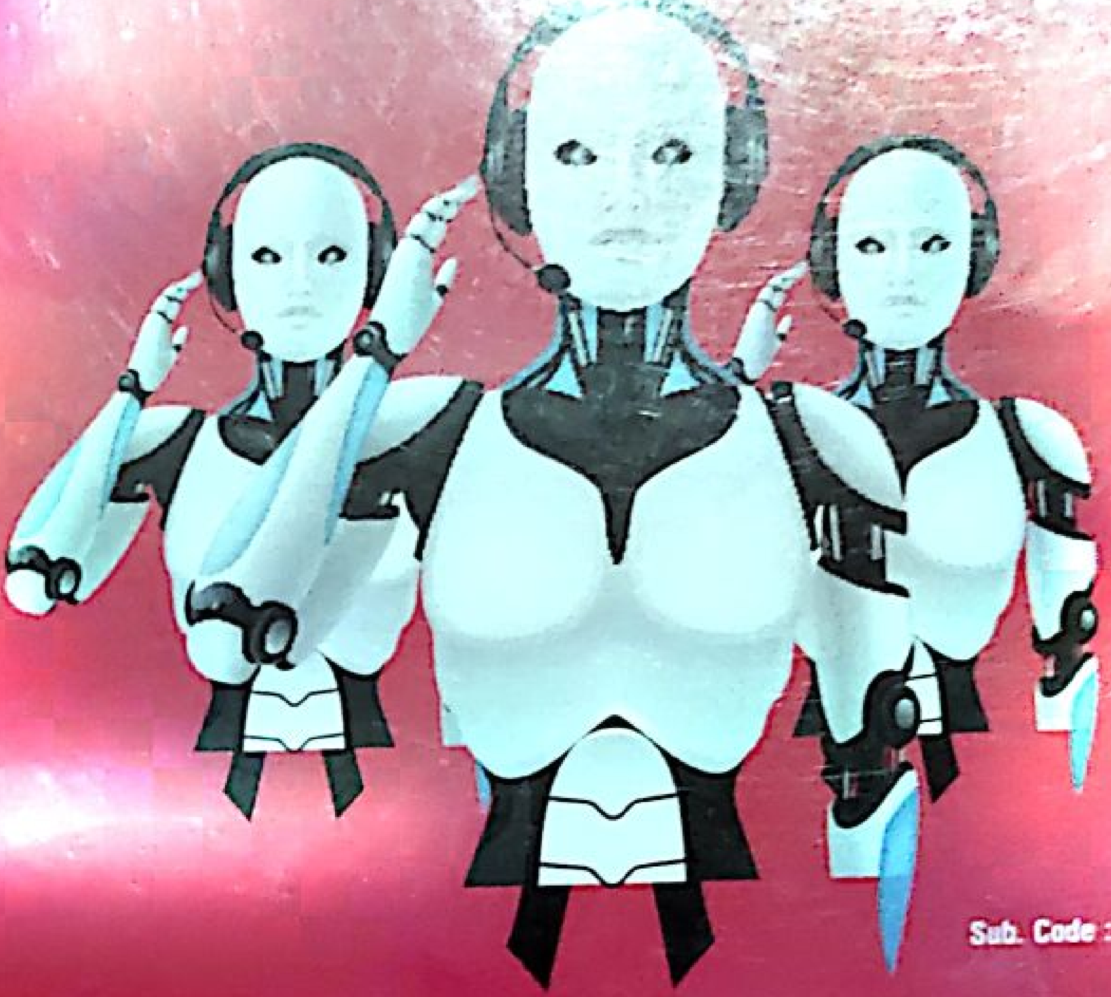


Sub. Code : 3170724

Machine Learning

AS PER NEW SYLLABUS - GTU - SEM - VII (CE/CSE/ICT) Professional Elective - VI

• Simplified & Conceptual Approach • Fill in the Blanks with Answers
• Multiple Choice Questions with Answers

first edition : july 2021

Price : ₹ 250/-
ISBN 978-93-90770-16-8

TECHNICAL PUBLICATIONS
SINCE 1993
An Up-Thrust for Knowledge

I. A. Dhotre

SUBJECT CODE : 3170724

As per New Syllabus of

GUJARAT TECHNOLOGICAL UNIVERSITY

Semester - VII (CE / CSE / ICT) Professional Elective - VI

MACHINE LEARNING

Iresh A. Dhotre

M.E. (Information Technology)

Ex-Faculty, Sinhgad College of Engineering,

Pune.

TECHNICAL PUBLICATIONS

SINCE 1993

An Up-Thrust for Knowledge

PREFACE

The importance of **Machine Learning** is well known in various engineering fields. Overwhelming response to my books on various subjects inspired me to write this book. The book is structured to cover the key aspects of the subject **Machine Learning.**

The book uses plain, lucid language to explain fundamentals of this subject. The book provides logical method of explaining various complicated concepts and stepwise methods to explain the important topics. Each chapter is well supported with necessary illustrations, practical examples and solved problems. All the chapters in the book are arranged in a proper sequence that permits each topic to build upon earlier studies. All care has been taken to make students comfortable in understanding the basic concepts of the subject.

The book not only covers the entire scope of the subject but explains the philosophy of the subject. This makes the understanding of this subject more clear and makes it more interesting. The book will be very useful not only to the students but also to the subject teachers. The students have to omit nothing and possibly have to cover nothing more.

I wish to express my profound thanks to all those who helped in making this book a reality. Much needed moral support and encouragement is provided on numerous occasions by my whole family. I wish to thank the **Publisher** and the entire team of **Technical Publications** who have taken immense pain to get this book in time with quality printing.

Any suggestion for the improvement of the book will be acknowledged and well appreciated.

Author
D. A. Dhotre

Dedicated to God

MACHINE LEARNING

Subject Code : 3170724

Semester - VII (CE / CSE / ICT) Professional Elective - VI

Published by :

TECHNICAL PUBLICATIONS
An Up-Thrust for Knowledge

Amit Residency, Office No.1, 412, Shaniwar Peth,
Pune - 411030, M.S. INDIA Ph.: +91-020-24495496/97
Email : sales@technicalpublications.org Website : www.technicalpublications.org

Printer :
Yogiraj Printers & Binders
Sr. No. 10/1A,
Ghule Industrial Estate, Nanded Village Road,
Tal. - Haveli, Dist. - Pune - 411041.

ISBN 978-93-90770-16-8

Course 18

9789390770168 [1]

PREFACE

The importance of **Machine Learning** is well known in various engineering fields. Overwhelming response to my books on various subjects inspired me to write this book. The book is structured to cover the key aspects of the subject **Machine Learning.**

The book uses plain, lucid language to explain fundamentals of this subject. The book provides logical method of explaining various complicated concepts and stepwise methods to explain the important topics. Each chapter is well supported with necessary illustrations, practical examples and solved problems. All the chapters in the book are arranged in a proper sequence that permits each topic to build upon earlier studies. All care has been taken to make students comfortable in understanding the basic concepts of the subject.

The book not only covers the entire scope of the subject but explains the philosophy of the subject. This makes the understanding of this subject more clear and makes it more interesting. The book will be very useful not only to the students but also to the subject teachers. The students have to omit nothing and possibly have to cover nothing more.

I wish to express my profound thanks to all those who helped in making this book a reality. Much needed moral support and encouragement is provided on numerous occasions by my whole family. I wish to thank the **Publisher** and the entire team of **Technical Publications** who have taken immense pain to get this book in time with quality printing.

Any suggestion for the improvement of the book will be acknowledged and well appreciated.

Author
D. A. Dhotre

Dedicated to God

SYLLABUS

Machine Learning - (3170724)

Credits	Examination Marks				Total Marks
C	Theory Marks		Practical Marks		
	ESE (E)	PA(M)	ESE (V)	PA (I)	
4	70	30	30	20	150

1. **Introduction to Machine Learning :**
   Overview of Human Learning and Machine Learning, Types of Machine Learning, Applications of Machine Learning , Tools and Technology for Machine Learning. **(Chapter - 1)**
2. **Preparing to Model :**
   Machine Learning activities, Types of data in Machine Learning, Structures of data, Data quality and remediation, Data Pre-Processing: Dimensionality reduction, Feature subset selection. **(Chapter - 2)**
3. **Modelling and Evaluation :**
   Selecting a Model: Predictive/Descriptive, Training a Model for supervised learning, model representation and interpretability, Evaluating performance of a model, Improving performance of a model. **(Chapter - 3)**
4. **Basics of Feature Engineering :**
   Feature and Feature Engineering, Feature transformation: Construction and extraction, Feature subset selection : Issues in high-dimensional data, key drivers, measure and overall process. **(Chapter - 4)**
5. **Overview of Probability :**
   Statistical tools in Machine Learning, Concepts of probability, Random variables, Discrete distributions, Continuous distributions, Multiple random variables, Central limit theorem, Sampling distributions, Hypothesis testing, Monte Carlo Approximation. **(Chapter - 5)**
6. **Bayesian Concept Learning :**
   Importance of Bayesian methods, Bayesian theorem, Bayes' theorem and concept learning, Bayesian Belief Network. **(Chapter - 6)**
7. **Supervised Learning : Classification and Regression :**
   Supervised Learning, Classification Model, Learning steps, Classification algorithms, Regression, Regression algorithms. **(Chapter - 7)**
8. **Unsupervised Learning :**
   Supervised vs. Unsupervised Learning, Applications, Clustering, Association rules **(Chapter - 8)**
9. **Neural Network :**
   Introduction to neural network, Biological and Artificial Neurons, Types of Activation functions, Implementation of ANN, Architecture, Leaning process, Backpropogation, Deep Learning. **(Chapter - 9)**

TABLE OF CONTENTS

Chapter - 3 Modelling and Evaluation (3 - 1) to (3 - 22)

Notes

1 Introduction to Machine Learning

Syllabus

Overview of Human Learning and Machine Learning, Types of Machine Learning, Applications of Machine Learning , Tools and Technology for Machine Learning .

Contents

1.1 Overview of Human Learning

- Learning is the process of acquiring new understanding, knowledge, behaviours, skills, values, attitudes and preferences. Learning process happens when you observe a phenomenon and recognize a pattern.
- Learning is a phenomenon and process which has manifestations of various aspects. Learning process includes gaining of new symbolic knowledge and development of cognitive skills through instruction and practice. It is also discovery of new facts and theories through observation and experiment.
- All human learning is observing something, identifying a pattern, building a theory (model) to explain this pattern and testing this theory to check if its fits in most or all observations.
- Fig. 1.1.1 shows human learning.

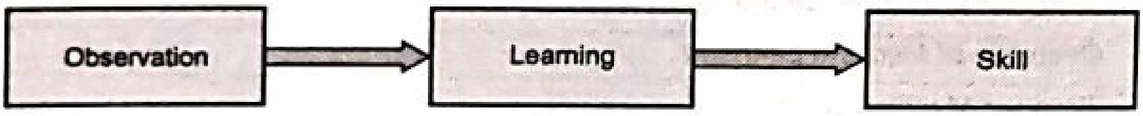


Fig. 1.1.1 Human learning

- Both human as well as machine learning generate knowledge, one residing in the brain the other residing in the machine.
- Human learning process varies from person to person. Once a learning process is set into the minds of people, it is difficult to change it.
- Fig. 1.1.2 shows relation between human and machine learning.

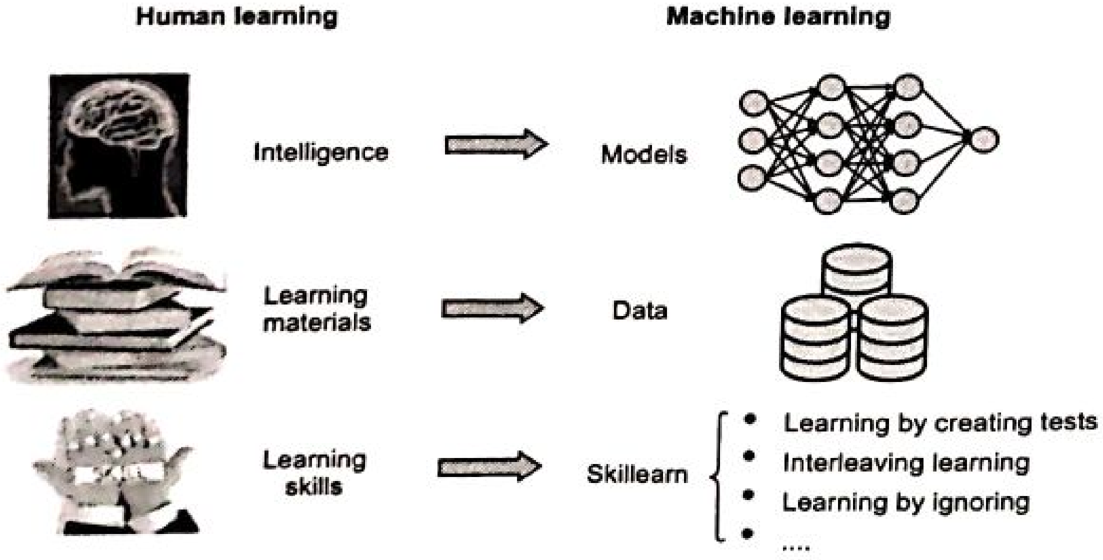


Fig. 1.1.2

Types of human learning

- Human learning take place in following way :

1. Self-learning : Human try many times after multiple attempts, some being unsuccessful.
2. Knowledge gained from expert : We build our own notion indirectly based on what we have learnt from the expert in the past.
3. Learning directly from expert : Either somebody who is an expert in the subject directly teaches us.

- Humans acquire knowledge through experience either directly or shared by others. Humans begin learning by memorizing. After few years, he realizes that mere capability to memorize is not intelligence.
- In humans, learning speed depends on individuals and in machines, learning speed depends on the algorithm selected and the volume of examples exposed to it.

1.1.1 Difference between Human and Machine Learning

Human learning	Machine learning
Humans acquire knowledge through experience either directly or shared by others.	Machines acquire knowledge through experience shared in the form of past data.
Model-free and model-based mechanisms can be found in human learning.	Knowledge based learning in machine learning.
Observation ➡ Learning ➡ Skill	Data ➡ Machine Learning ➡ Skill

1.2 Overview of Machine Learning

- Machine Learning (ML) is a sub-field of Artificial Intelligence (AI) which concerns with developing computational theories of learning and building learning machines.
- **Learning** is a phenomenon and process which has manifestations of various aspects. Learning process includes gaining of new symbolic knowledge and development of cognitive skills through instruction and practice. It is also discovery of new facts and theories through observation and experiment.
- **Machine Learning Definition** : A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

- Machine learning is programming computers to optimize a performance criterion using example data or past experience. Application of machine learning methods to large databases is called **data mining.**
- It is very hard to write programs that solve problems like recognizing a human face. We do not know what program to write because we don't know how our brain does it. Instead of writing a program by hand, it is possible to collect lots of examples that specify the correct output for a given input.
- A machine learning algorithm then takes these examples and produces a program that does the job. The program produced by the learning algorithm may look very different from a typical hand-written program. It may contain millions of numbers. If we do it right, the program works for new cases as well as the ones we trained it on.
- Main goal of machine learning is to devise learning algorithms that do the learning automatically without human intervention or assistance. The machine learning paradigm can be viewed as "programming by example." Another goal is to develop computational models of human learning process and perform computer simulations.
- The goal of machine learning is to build computer systems that can adapt and learn from their experience.
- Algorithm is used to solve a problem on computer. An algorithm is a sequence of instruction. It should carry out to transform the input to output. For example, for addition of four numbers is carried out by giving four number as input to the algorithm and output is sum of all four numbers. For the same task, there may be various algorithms. It is interested to find the most efficient one, requiring the least number of instructions or memory or both.
- For some tasks, however, we do not have an algorithm.

Why is Machine Learning Important ?

- Machine learning algorithms can figure out how to perform important tasks by generalizing from examples.
- Machine Learning provides business insight and intelligence. Decision makers are provided with greater insights into their organizations. This adaptive technology is being used by global enterprises to gain a competitive edge.
- Machine learning algorithms discover the relationships between the variables of a system (input, output and hidden) from direct samples of the system.
- **Following are some of the reasons :**
  1. Some tasks cannot be defined well, except by examples. For example : Recognizing people.

2. Relationships and correlations can be hidden within large amounts of data. To solve these problems, machine learning and data mining may be able to find these relationships.
3. Human designers often produce machines that do not work as well as desired in the environments in which they are used.
4. The amount of knowledge available about certain tasks might be too large for explicit encoding by humans.
5. Environments change time to time.
6. New knowledge about tasks is constantly being discovered by humans.

- Machine learning also helps us find solutions of many problems in computer vision, speech recognition and robotics. Machine learning uses the theory of statistics in building mathematical models, because the core task is making inference from a sample.

How Machines Learn ?

- Machine learning typically follows three phases :

1. **Training :** A training set of examples of correct behavior is analyzed and some representation of the newly learnt knowledge is stored. This is some form of rules.
2. **Validation :** The rules are checked and, if necessary, additional training is given. Sometimes additional test data are used, but instead, a human expert may validate the rules, or some other automatic knowledge - based component may be used. The role of the tester is often called the opponent.
3. **Application :** The rules are used in responding to some new situation.

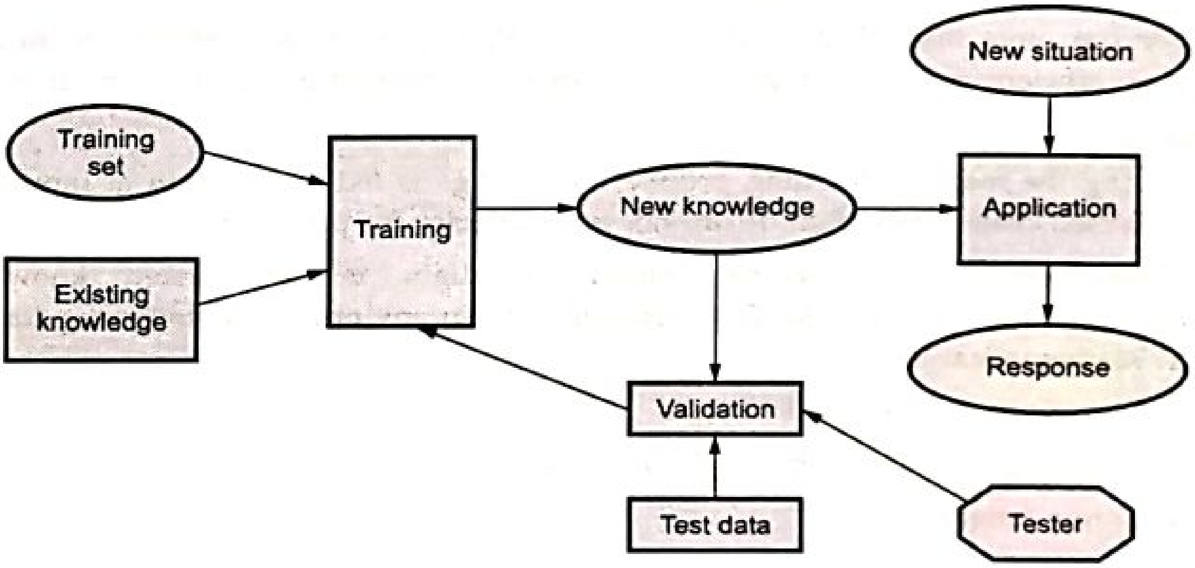


Fig. 1.2.1

1.2.1 How do Machine Learn?

- Machine learning process in divided into three parts : Data inputs, abstraction and generalization.
- Fig. 1.2.2 shows machine learning process.

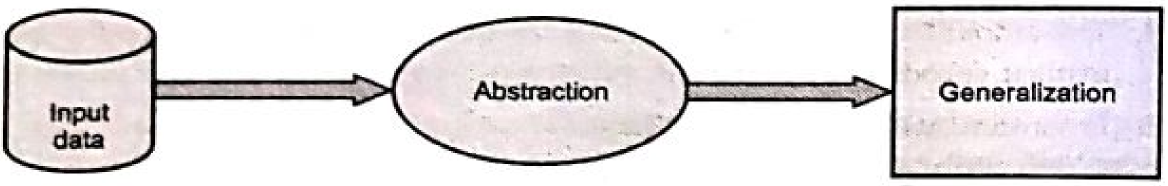


Fig. 1.2.2 Machine learning process

- **Data input :** Information is used for future decision making.
- **Abstraction :** Input data is represented in broader way through the underlying algorithm.
- **Generalization :** It forms framework for making decision.
- Machine learning is a form of Artificial Intelligence (AI) that teaches computers to think in a similar way to how humans do : Learning and improving upon past experiences. It works by exploring data and identifying patterns and involves minimal human intervention.
- Algorithm is used to solve a problem on computer. An algorithm is a sequence of instruction. It should carry out to transform the input to output. For example, for addition of four numbers is carried out by giving four number as input to the algorithm and output is sum of all four numbers.
- For the same task, there may be various algorithms. It is interested to find the most efficient one, requiring the least number of instructions or memory or both.

Abstraction

- During the machine learning process, knowledge is fed in the form of input data. Collected data is raw data. It can not used directly for processing.
- Model known in machine leaning paradigm is summarized knowledge representation of raw data. The model may be in any one of the following forms :
  1. Mathematical equations.
  2. Specific data structure like trees.
  3. Logical grouping of similar observations.
  4. Computational blocks.

- Choice of the model used to solve specific learning problem is the human task. Some of the parameters are as follows :
  a) Type of problem to be solved.
  b) Nature of the input data.
  c) Problem domain.

1.2.2 Well Posed Learning Problem

- **Definition : A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.**
- A (machine learning) problem is well-posed if a solution to it exists, if that solution is unique, and if that solution depends on the data / experience but it is not sensitive to (reasonably small) changes in the data / experience.
- Identify three features are as follows :
  1. Class of tasks
  2. Measure of performance to be improved
  3. Source of experience
- What are T, P, E ? How do we formulate a machine learning problem ?
- A Robot Driving Learning Problem
  1. **Task T :** Driving on public, 4-lane highway using vision sensors.
  2. **Performance measure P :** Average distance traveled before an error (as judged by human overseer).
  3. **Training experience E :** A sequence of images and steering commands recorded while observing a human driver.
- A Handwriting Recognition Learning Problem.
  1. **Task T :** Recognizing and classifying handwritten words within images.
  2. **Performance measure P :** Percent of words correctly classified.
  3. **Training experience E :** A database of handwritten words with given classifications.
- Text Categorization Problem.
  1. Task T : Assign a document to its content category.
  2. Performance measure P : Precision and Recall.
  3. Training experience E : Example pre-classified documents.

1.3 Types of Machine Learning

- Learning is constructing or modifying representation of what is being experienced. Learn means to get knowledge of by study, experience or being taught.
- Machine learning is a scientific discipline concerned with the design and development of the algorithm that allows computers to evolve behaviours based on empirical data, such as form sensors data or database.
- Machine learning is usually divided into three types : Supervised, unsupervised and reinforcement learning.
- Why do machine learning ?
  1. To understand and improve efficiency of human learning.
  2. Discover new things or structure that is unknown to humans.
  3. Fill in skeletal or incomplete specifications about a domain.

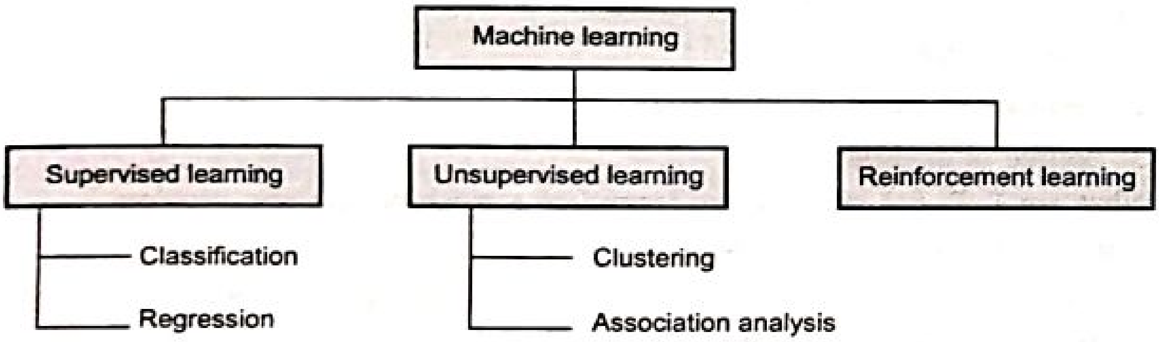


Fig. 1.3.1

1.3.1 Supervised Learning

- Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. The task of the supervised learner is to predict the output behavior of a system for any set of input values, after an initial training phase.
- **Supervised learning** in which the network is trained by providing it with input and matching output patterns. These input-output pairs are usually provided by an external teacher.
- Human learning is based on the past experiences. A computer does not have experiences.
- A computer system learns from data, which represent some "past experiences" of an application domain.

- To learn a target function that can be used to predict the values of a discrete class attribute, e.g., approve or not-approved, and high-risk or low risk. The task is commonly called : Supervised learning, Classification or inductive learning.
- Training data includes both the input and the desired results. For some examples the correct results (targets) are known and are given in input to the model during the learning process. The construction of a proper training, validation and test set is crucial. These methods are usually fast and accurate.
- Have to be able to generalize : Give the correct results when new data are given in input without knowing a priori the target.
- Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object and a desired output value.
- A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier or a regression function. Fig. 1.3.2. shows supervised learning process.

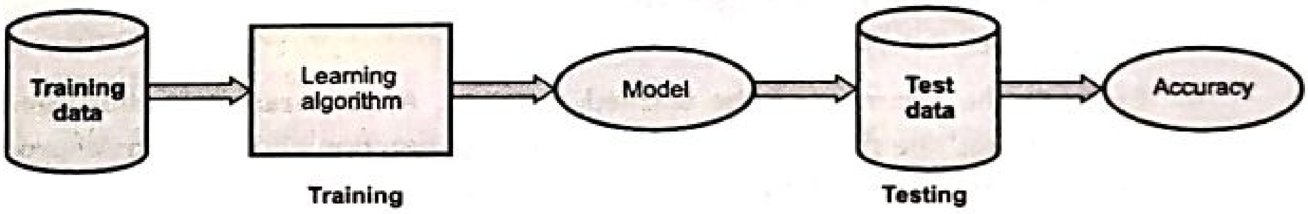


Fig. 1.3.2 Supervised learning process

- The learned model helps the system to perform task better as compared to no learning.
- Each input vector requires a corresponding target vector.

Training Pair = (Input Vector, Target Vector)

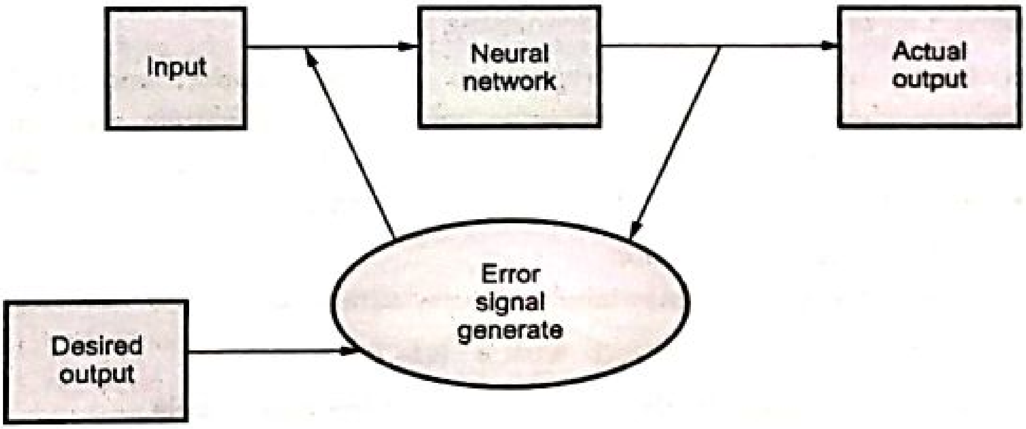


Fig. 1.3.3

- Supervised learning denotes a method in which some input vectors are collected and presented to the network. The output computed by the net-work is observed and the deviation from the expected answer is measured. The weights are corrected according to the magnitude of the error in the way defined by the learning algorithm.
- Supervised learning is further divided into methods which use reinforcement or error correction. The perceptron learning algorithm is an example of supervised learning with reinforcement.
- In order to solve a given problem of supervised learning, following steps are performed :
  1. Find out the type of training examples.
  2. Collect a training set.
  3. Determine the input feature representation of the learned function.
  4. Determine the structure of the learned function and corresponding learning algorithm.
  5. Complete the design and then run the learning algorithm on the collected training set.
  6. Evaluate the accuracy of the learned function. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.

1.3.1.1 Classification

- Classification predicts categorical labels (classes), prediction models continuous-valued functions. Classification is considered to be supervised learning.
- Classifies data based on the training set and the values in a classifying attribute and uses it in classifying new data. **Prediction** means models continuous-valued functions, i.e., predicts unknown or missing values.
- Preprocessing of the data in preparation for classification and prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes, and data transformation, such as generalizing the data to higher level concepts or normalizing data.
- Fig. 1.3.4 shows the classification.

Aim : To predict categorical class labels for new samples.

Input : Training set of samples, each with a class label.

Output : Classifier is based on the training set and the class labels.

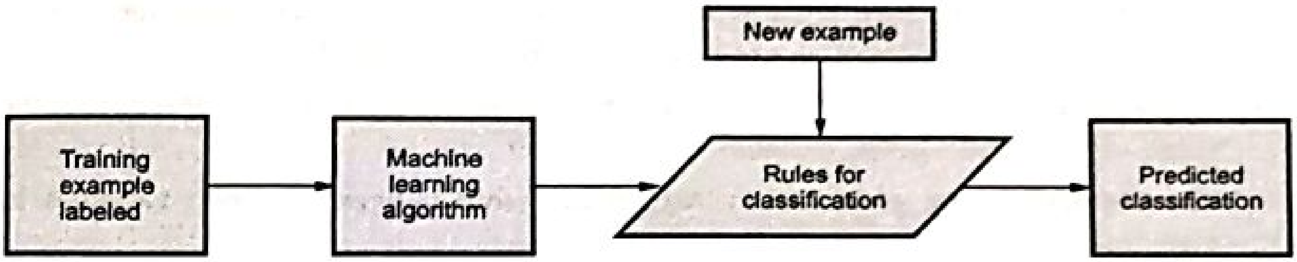


Fig. 1.3.4 Classification

- **Prediction** is similar to classification. It constructs a model and uses the model to predict unknown or missing value.
- Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data.
- Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process.
- Numeric prediction is the task of predicting continuous values for given input. For example, we may wish to predict the salary of college employee with 15 years of work experience, or the potential sales of a new product given its price.
- Some of the classification methods like back-propagation, support vector machines, and k-nearest-neighbor classifiers can be used for prediction.

1.3.1.2 Regression

- For an input x, if the output is continuous, this is called a regression problem. For example, based on historical information of demand for tooth paste in your supermarket, you are asked to predict the demand for the next month.
- Regression is concerned with the prediction of continuous quantities. Linear regression is the oldest and most widely used predictive model in the field of machine learning. The goal is to minimize the sum of the squared errors to fit a straight line to a set of data points.
- For regression tasks, the typical accuracy metrics are Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). These metrics measure the distance between the predicted numeric target and the actual numeric answer.

Regression Line

- **Least squares :** The least squares regression line is the line that makes the sum of squared residuals as small as possible. Linear means "straight line".
- **Regression line** is the line which gives the best estimate of one variable from the value of any other given variable.
- **The regression line** gives the average relationship between the two variables in mathematical form.
- For two variables X and Y, there are always two lines of regression.
- **Regression line of X on Y :** Gives the best estimate for the value of X for any specific given values of Y :

$$X = a + bY$$

where

a = X - intercept

b = Slope of the line

X = Dependent variable

Y = Independent variable

- **Regression line of Y on X :** Gives the best estimate for the value of Y for any specific given values of X :

$$Y = a + bx$$

where

a = Y - intercept

b = Slope of the line

Y = Dependent variable

x = Independent variable

- By using the least squares method (a procedure that minimizes the vertical deviations of plotted points surrounding a straight line) we are able to construct a best fitting straight line to the scatter diagram points and then formulate a regression equation in the form of :

$$\hat{y} = a + bX$$

$$\hat{y} = \bar{y} + b(x - \bar{x})$$

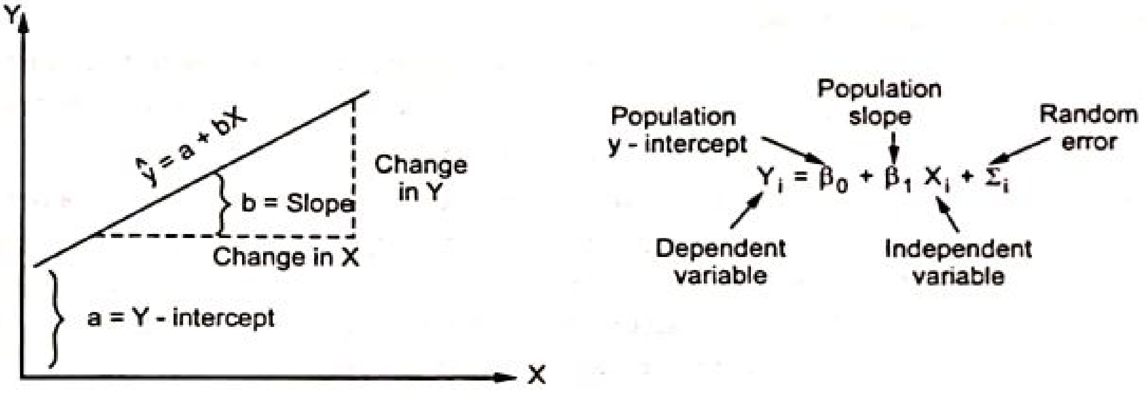


Fig. 1.3.5

- Regression analysis is the art and science of fitting straight lines to patterns of data. In a linear regression model, the variable of interest ("dependent" variable) is predicted from k other variables ("independent" variables) using a linear equation. If Y denotes the dependent variable, and $X_1, \ldots, X_k$, are the independent variables, then the assumption is that the value of Y at time t in the data sample is determined by the linear equation :

$$Y_1 = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + \varepsilon_t$$

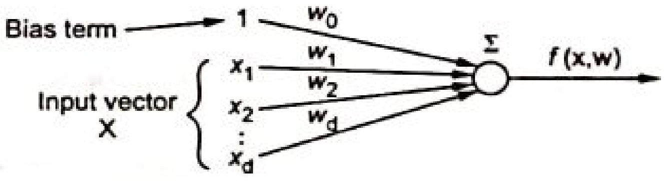


Fig. 1.3.6

where the betas are constants and the epsilons are independent and identically distributed normal random variables with mean zero.

- In a regression tree the idea is this : Since the target variable does not have classes, we fit a regression model to the target variable using each of the independent variables. Then for each independent variable, the data is split at several split points.

- At each split point, the "error" between the predicted value and the actual values is squared to get a "Sum of Squared Errors (SSE)". The split point errors across the variables are compared and the variable/point yielding the lowest SSE is chosen as the root node/split point. This process is recursively continued.

- Error function measures how much our predictions deviate from the desired answers.

$$\text{Mean-squared error } J_n = \frac{1}{n} \sum_{i=1\_n} (y_i - f(x_i))^2$$

- **Multiple linear regression** is an extension of linear regression, which allows a response variable, y, to be modeled as a linear function of two or more predictor variables.

Evaluating a Regression Model

- Assume we want to predict a car's price using some features such as dimensions, horsepower, engine specification, mileage etc. This is a typical regression problem, where the target variable (price) is a continuous numeric value.
- We can fit a simple linear regression model that, given the feature values of a certain car, can predict the price of that car. This regression model can be used to score the same dataset we trained on. Once we have the predicted prices for all of the cars, we can evaluate the performance of the model by looking at how much the predictions deviate from the actual prices on average.

Advantages :

a. Training a linear regression model is usually much faster than methods such as neural networks.

b. Linear regression models are simple and require minimum memory to implement.

c. By examining the magnitude and sign of the regression coefficients you can infer how predictor variables affect the target outcome.

Assessing Performance of Regression- Error Measures

- The **training error** is the mean error over the training sample. The test error is the expected prediction error over an independent test sample.
- Fig. 1.3.7 shows the relationship between training set and test set.

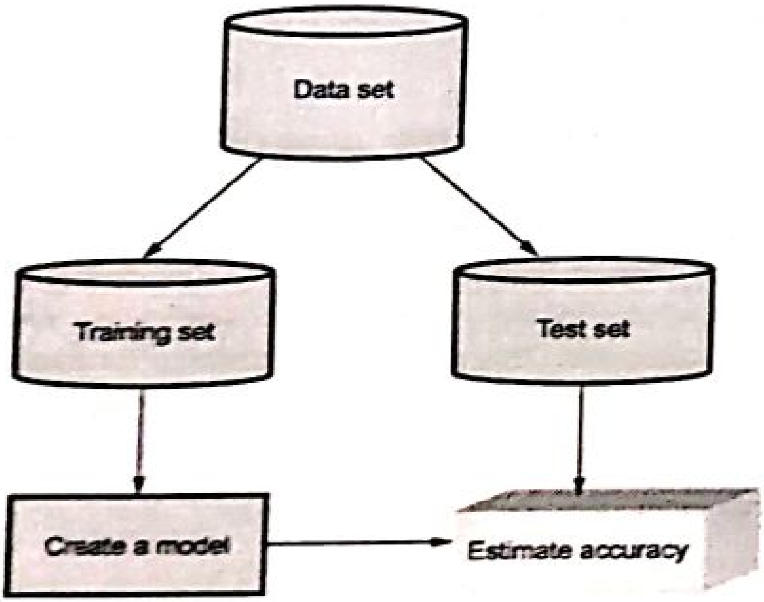


Fig. 1.3.7

- Unlike decision trees, regression trees and model trees are used for prediction. In regression trees, each leaf stores a continuous-valued prediction. In model trees, each leaf holds a regression model.

1.3.2 Un - Supervised Learning

- The model is not provided with the correct results during the training. It can be used to cluster the input data in classes on the basis of their statistical properties only. Cluster significance and labeling.
- The labeling can be carried out even if the labels are only available for a small number of objects representative of the desired classes. All similar inputs patterns are grouped together as clusters.
- If matching pattern is not found, a new cluster is formed. There is no error feedback.
- External teacher is not used and is based upon only local information. It is also referred to as **self-organization.**
- They are called unsupervised because they do not need a teacher or super-visor to label a set of training examples. Only the original data is required to start the analysis.
- In contrast to supervised learning, unsupervised or self-organized learning does not require an external teacher. During the training session, the neural network receives a number of different input patterns, discovers significant features in these patterns and learns how to classify input data into appropriate categories.
- Unsupervised learning algorithms aim to learn rapidly and can be used in real-time. Unsupervised learning is frequently employed for data clustering, feature extraction etc.
- Another mode of learning called recording learning by Zurada is typically employed for associative memory networks. An associative memory networks is designed by recording several idea patterns into the networks stable states.

1.3.2.1 Clustering

- Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. Clustering can be considered the most important unsupervised learning problem.
- A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Fig. 1.3.8 shows cluster.

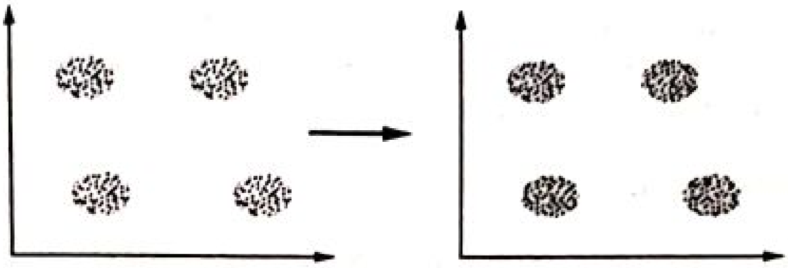
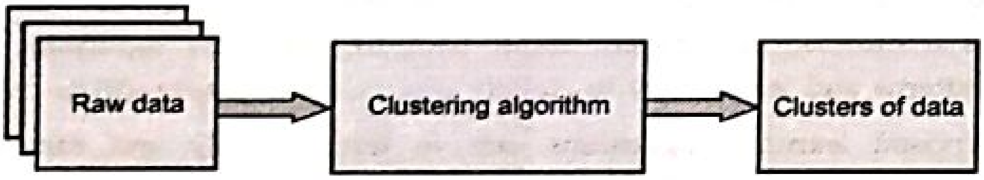
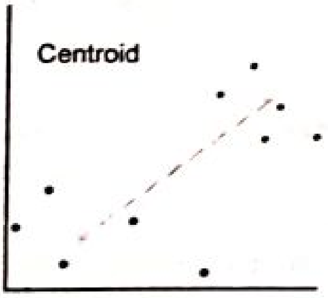


Fig. 1.3.8 Cluster

- In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance : two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called **distance-based clustering.**
- Clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.
- A clustering algorithm attempts to find natural groups of components or data based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets.
- To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.



- **Cluster centroid :** The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters. Each cluster has a well defined centroid.



- **Distance :** The distance between two points is taken as a common metric to as see the similarity among the components of a population. The commonly used distance measure is the Euclidean metric which defines the distance between two points $p = (p_1, p_2, ...)$ and $q = (q_1, q_2, ...)$ is given by :

$$d = \sum_{i=1}^{k} (p_i - q_i)^2$$

- The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering ? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
- Clustering analysis helps construct meaningful partitioning of a large set of objects. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, etc.
- Clustering algorithms may be classified as listed below :
  1. Exclusive clustering
  2. Overlapping clustering
  3. Hierarchical clustering
  4. Probabilistic clustering
- A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Examples of Clustering Applications

1. **Marketing :** Help marketers discover distinct groups in their customer bases and then use this knowledge to develop targeted marketing programs.
2. **Land use :** Identification of areas of similar land use in an earth observation database.
3. **Insurance :** Identifying groups of motor insurance policy holders with a high average claim cost.
4. **Urban planning :** Identifying groups of houses according to their house type, value, and geographical location.
5. **Seismology :** Observed earth quake epicenters should be clustered along continent faults.

1.3.3 Reinforcement Learning

- User will get immediate feedback in supervised learning and no feedback from unsupervised learning. But in the reinforced learning, you will get delayed scalar feedback.
- Reinforcement learning is learning what to do and how to map situations to actions. The learner is not told which actions to take. Fig. 1.3.9 shows concept of reinforced learning.
- Reinforced learning is deals with agents that must sense and act upon their environment. It combines classical Artificial Intelligence and machine learning techniques.

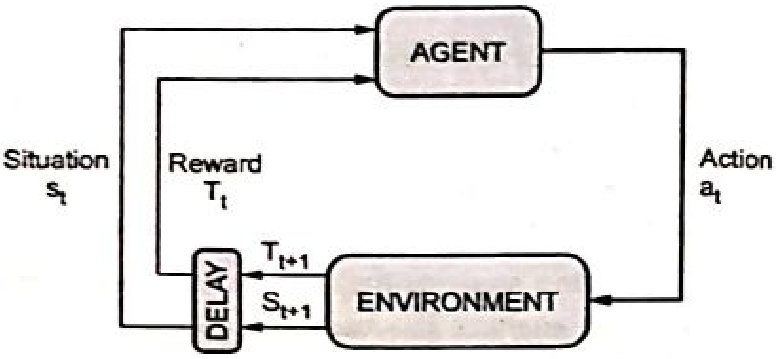


Fig. 1.3.9 Reinforced learning

- It allows machines and software agents to automatically determine the ideal behavior within a specific context, in order to maximize its performance. Simple reward feedback is required for the agent to learn its behavior; this is known as the reinforcement signal.
- Two most important distinguishing features of reinforcement learning is trial-and-error and delayed reward.
- With reinforcement learning algorithms an agent can improve its performance by using the feedback it gets from the environment. This environmental feedback is called the reward signal.
- Based on accumulated experience, the agent needs to learn which action to take in a given situation in order to obtain a desired long term goal. Essentially actions that lead to long term rewards need to reinforced. Reinforcement learning has connections with control theory, Markov decision processes and game theory.
  - **Example of Reinforcement Learning :** A mobile robot decides whether it should enter a new room in search of more trash to collect or start trying to find its way back to its battery recharging station. It makes its decision based on how quickly and easily it has been able to find the recharger in the past.

1.3.3.1 Elements of Reinforcement Learning

- Reinforcement learning elements are as follows :

  1. Policy 2. Reward Function

  3. Value Function 4. Model of the environment

- Fig. 1.3.10 shows

- **Policy :** Policy defines the learning agent behavior for given time period. It is a mapping from perceived states of the environment to actions to be taken when in those states.

- **Reward Function :** Reward function is used to define a goal in a reinforcement learning problem. It also maps each perceived state of the environment to a single number.

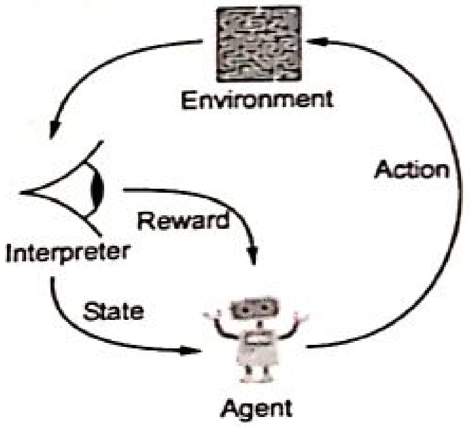


Fig. 1.3.10 : Elements of reinforcement learning

- **Value function :** Value functions specify what is good in the long run. The value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state.
- **Model of the environment :** Models are used for planning.
- Credit assignment problem : Reinforcement learning algorithms learn to generate an internal value for the intermediate states as to how good they are in leading to the goal.
- The learning decision maker is called the agent. The agent interacts with the environment that includes everything outside the agent.
- The agent has sensors to decide on its state in the environment and takes an action that modifies its state.
- The reinforcement learning problem model is an agent continuously interacting with an environment. The agent and the environment interact in a sequence of time steps. At each time step t, the agent receives the state of the environment and a scalar numerical reward for the previous action, and then the agent then selects an action.
- Reinforcement Learning is a technique for solving Markov Decision Problems.

- Reinforcement learning uses a formal framework defining the interaction between a learning agent and its environment in terms of states, actions, and rewards. This framework is intended to be a simple way of representing essential features of the artificial intelligence problem.

1.3.4 Difference between Supervised, Unsupervised and Reinforcement Learning

Supervised learning	Unsupervised learning	Reinforcement learning
Supervised learning requires that the target variable is well defined and that a sufficient number of its values are given.	For unsupervised learning typically either the target variable is unknown or has only been recorded for too small a number of cases.	Reinforcement learning is learning what to do and how to map situations to actions. The learner is not told which actions to take.
Supervised learning deals with two main tasks regression and classification.	Unsupervised Learning deals with clustering and associative rule mining problems.	Reinforcement learning deals with exploitation or exploration, Markov's decision processes, policy learning, deep learning and value learning.
The input data in supervised learning in labelled data.	Unsupervised learning uses unlabelled data.	The data is not predefined in reinforcement learning.
Learns by using labelled data.	Trained using unlabelled data without any guidance.	Works on interacting with the environment.
Maps the labeled inputs to the known outputs.	Understands patterns and discovers the output.	Follows the trial and error method.

1.4 Applications of Machine Learning

- Examples of successful applications of machine learning :
  1. Learning to recognize spoken words.
  2. Learning to drive an autonomous vehicle.
  3. Learning to classify new astronomical structures.
  4. Learning to play world-class backgammon.
  5. Spoken language understanding: within the context of a limited domain, determine the meaning of something uttered by a speaker to the extent that it can be classified into one of a fixed set of categories.

Face Recognition

- Face recognition task is effortlessly and every day we recognize our friends, relative and family members. We also recognition by looking at the photographs.

In photographs, they are in different pose, hair styles, background light, makeup and without makeup.

- We do it subconsciously and cannot explain how we do it. Because we can't explain how we do it, we can't write an algorithm.
- Face has some structure. It is not a random collection of pixel. It is symmetric structure. It contains predefined components like nose, mouth, eye, ears. Every person face is a pattern composed of a particular combination of the features. By analyzing sample face images of a person, a learning program captures the pattern specific to that person and uses it to recognize if a new real face or new image belongs to this specific person or not.
- Machine learning algorithm creates an optimized model of the concept being learned based on data or past experience.

Healthcare :

- With the advent of wearable sensors and devices that use data to access health of a patient in real time, ML is becoming a fast-growing trend in healthcare.
- Sensors in wearable provide real-time patient information, such as overall health condition, heartbeat, blood pressure and other vital parameters.
- Doctors and medical experts can use this information to analyse the health condition of an individual, draw a pattern from the patient history and predict the occurrence of any ailments in the future.
- The technology also empowers medical experts to analyze data to identify trends that facilitate better diagnoses and treatment.

Financial services :

- Companies in the financial sector are able to identify key insights in financial data as well as prevent any occurrences of financial fraud, with the help of machine learning technology.
- The technology is also used to identify opportunities for investments and trade.
- Usage of cyber surveillance helps in identifying those individuals or institutions which are prone to financial risk and take necessary actions in time to prevent fraud.

1.5 Tools and Technology for Machine Learning

1.5.1 Python

- Python is a high-level scripting language which can be used for a wide variety of text processing, system administration and internet-related tasks.

- Python is a true object-oriented language and is available on a wide variety of platforms.
- Python was developed in the early 1990's by Guido van Rossum, then at CWI in Amsterdam and currently at CNRI in Virginia. Python 3.0 was released in Year 2008.
- Python statements do not need to end with a special character. Python relies on modules, that is, self-contained programs which define a variety of functions and data types.
- A module is a file containing Python definitions and statements. The file name is the module name with the suffix .py appended. Within a module, the module's name (as a string) is available as the value of the global variable __name__.
- If a module is executed directly however, the value of the global variable __name__ will be "__main__".
- Modules can contain executable statements aside from definitions. These are executed only the first time the module name is encountered in an import statement as well as if the file is executed as a script.
- Integrated Development Environment (IDE) is the basic interpreter and editor environment that you can use along with Python. This typically includes an editor for creating and modifying programs, a translator for executing programs and a program debugger. A debugger provides a means of taking control of the execution of a program to aid in finding program errors.
- Python is most commonly translated by use of an interpreter. It provides the very useful ability to execute in interactive mode. The window that provides this interaction is referred to as the Python shell.
- Python support two basic modes : Normal mode and interactive mode.
- Normal mode : The normal mode is the mode where the scripted and finished . py files are run in the Python interpreter. This mode is also called as script mode.
- Interactive mode is a command line shell which gives immediate feedback for each statement, while running previously fed statements in active memory.
  - Start the Python interactive interpreter by typing python with no arguments at the command line.
  - To access the Python shell, open the terminal of your operating system and then type "python". Press the enter key and the python shell will appear.

```
C:\Windows\system32>python
Python 3.5.0(v.3.5.0:374f501f4567, Sep 13 2015, 2:27:37)[MSCv.1900 64 bit (AMD64)] on win32
Type "help", copyright,"credits" or "license" for more information.
>>>
```

- The >>> indicates that the Python shell is ready to execute and send your commands to the Python intrepreter. The result is immediately displayed on the Python shell as soon as the Python interpreter interpreters the command.
- For example, to print the text "Hello World", we can type the following :

```
>>> print("Hello World")
Hell World
>>>
```

- In script mode, a file must be created and saved before executing the code to get results. In interactive mode, the result is returned immediately after pressing the eneter key.
- In script mode, you are provided with a direct way of editing your code. This is not possible in interactive mode.
- A variable is a way of referrring to a memory location used by a computer program.
- A variable is a symbolic name for this physical location. This memory location contains values, like numbers, text or more complicated types.
- A variable is a name that refers to a value. The equal (=) operator is used to assign value to a variable.
- Python's data types include : Numbers, strings, lists, dictionaries, tuples and files.
- Python has no additional commands to declare a variable. As soon as the value is assigned to it, the variable is declared.
- Rules for varibles are as follows :
  a. Special characters are not allowed.
  b. Variables are case sensitive.
  c. Variable can only contain aplha-numeric characters and underscores.
  d. Variable name always start with character, not with number.

Features of Puython programming

1. Python is a high-level, interpreted, interactive and object-oriented scripting language.
2. It is simple and easy to learn.
3. It is portable.
4. Python is free and open source programming langauage.
5. Python can perform complex tasks using a few lines of code.

6. Python can run equally on different platforms such as Window, Linux, UNIX and Macintosh etc.
7. It provides a vast range of libraries for the various fields such as machine learing, web, developer and also for the scripting.

Advantages of Python

- Ease of programming.
- Minimizes the time to develop and maintain code.
- Modular and object-oriented.
- Large community of users.
- A large standard and user-constributed library.

Disadvantages of Python

- Interpreted and therefore slower than compiled languages.
- Decentralized with pacakges.

1.5.2 R Programming Language

- R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.
- R is often used for statistical computing and graphical presentation to analyse and visualize data.
- To use a function in a package, the package needs to be loaded in memory. Command for this is library( ), for example : library(affy).
- R is case sensitive, so take care when typing in the commands. Multiple commands can be written on the same line.
- Command can have many arguments. These are always giving inside the brackets. Numeric (1, 2, 3...) or logic (T/F) values and names of existing objects are given for the arguments without quotes, but string values, such as file names, are always put inside quotes.
- For example : mas5(dat3, normalize = T, analysis = "absolute").
- Vectors and matrices in R are two ways to work with a collection of objects.
- Lists provide a third method. Unlike a vector or a matrix a list can hold different kinds of objects. One entry in a list may be a number, while the next is a matrix, while a third is a character string.
- Statistical functions of R usually return the result in the form of lists. So we must know how to unpack a list using the $ symbol.

1.5.3 MATLAB

- MATLAB is a programming language developed by MathWorks. It started out as a matrix programming language where linear algebra programming was simple. It can be run both under interactive sessions and as a batch job.
- MATLAB is a high-performance language for technical computing. It integrates computation, visualization and programming environment.
- MATLAB is an interactive system whose basic data element is an array that does not require dimensioning.
- The name MATLAB stands for matrix laboratory. MATLAB was originally written to provide easy access to matrix software developed by the LINPACK and EISPACK projects, which together represent the state-of-the-art in software for matrix computation.
- The MATLAB system consists of five main parts :
  1. The MATLAB language. This is a high-level matrix/array language with control flow statements, functions, data structures, input/output and object-oriented programming features.
  2. The MATLAB working environment. This is the set of tools and facilities that you work with as the MATLAB user or programmer. It includes facilities for managing the variables in your workspace and importing and exporting data.
  3. It handle graphics. This is the MATLAB graphics system. It includes high-level commands for two-dimensional and three-dimensional data visualization, image processing, animation and presentation graphics.
  4. The MATLAB mathematical function library. This is a vast collection of computational algorithms ranging from elementary functions like sum, sine, cosine and complex arithmetic, to more sophisticated functions like matrix inverse, matrix eigenvalues, Bessel functions and fast Fourier transforms.
  5. The MATLAB Application Program Interface (API). This is a library that allows you to write C and Fortran programs that interact with MATLAB.

1.6 Fill in the Blanks

Q.1 Machine learning is a sub-field of _______ which concerns with developing computational theories of learning and building learning machines.

Q.2 _______ learning in which the network is trained by providing it with input and matching output patterns.

Q.3 Both human as well as machine learning generate knowledge, one residing in the _______ the other residing in the _______.

Q.4 Humans acquire ________ through experience either directly or shared by others.

Q.5 Supervised learning and unsupervised learning are the types of ________.

Q.6 Python is a true ________ language and is available on a wide variety of platforms.

Q.7 MATLAB is a programming language developed by ________.

Q.8 Vectors and matrices in R are two ways to work with a collection of ________.

Q.9 Machine learning algorithms discover the relationships between the variables of a system from direct ________ of the system.

Q.10 Human learning is based on the past ________.

Q.11 A ________ learning algorithm analyses the training data and produces an inferred function, which is called a classifier or a regression function.

Q.12 Supervised learning deals with two main tasks ________ and ________.

Q.13 Unsupervised learning uses ________ data.

Q.14 CART stands for ________.

Q.15 ________ can be viewed as the task of searching through a large space of hypotheses implicitly defined by the hypothesis representation.

Q.16 ________ learning is deals with agents that must sense and act upon their environment. It combines classical artificial intelligence and machine learning techniques.

Q.17 With reinforcement learning algorithms an agent can improve its performance by using the feedback it gets from the environment. This environmental feedback is called the ________.

Q.18 Supervised learning is also called ________ learning.

Q.19 Unsupervised learning is also called ________ learning.

Q.20 When we are trying to predict a categorical or nominal variable, the problem is known as a ________ problem.

Q.21 When we are trying to predict a real-valued variable, the problem falls under the category of ________.

1.7 Multiple Choice Questions

Q.1 A computer program is said to learn from ________ E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

a training
b experience
c testing
d algorithm

Q.2 Jarvis Patrick Clustering algorithm is a _______ clustering technique.

a grid based
b graph based
c density based
d all of these

Q.3 Which of the following is hierarchical clustering method :

a Agglomerative
b Divisive clustering
c PAM
d A and B

Q.4 The k-means algorithm is sensitive to _______ because an object with an extremely large value may substantially distort the distribution of data.

a outliers
b text data
c boasting
d cluster

Q.5 _______ hierarchical clustering method works by grouping data objects into a tree of clusters.

a PAM
b Density-based method
c Hierarchical
d Grid-Based method

Q.6 In DIANA, all of the objects are used to form _______ initial cluster.

a one
b two
c four
d eight

Q.7 If the clustering process is terminated when the distance between nearest clusters exceeds an arbitrary threshold, it is called a _______.

a dendrogram
b nearest-neighbor clustering algorithm
c minimal spanning tree algorithm
d single-linkage algorithm

Q.8 Which of the following is NOT type of clusters ?

a Well-separated clusters
b Prototype-based clusters
c Contiguity-based clusters
d DBSCAN clusters

Q.9 Shared nearest neighbors is a _______clustering.

a density-based b well-separated

c contiguity based d graph based

Q.10 Unsupervised learning deals with _______ and _______ mining problems.

a classification, regression b clustering, classification

c clustering, associative rule d label, unlabelled data

Q.11 _______ learning deals with two main tasks regression and classification.

a Reinforcement b Deep

c Un supervised d Supervised

Q.12 The individual tuples making up the training set are referred to as _______ and are selected from the database under analysis.

a learning tuples b training tuples

c samples d database

Q.13 Machine learning is inherently a multi disciplinary field.

a Inter disciplinary b Multi disciplinary

c Single d None

Q.14 _____ methods have been used to train computer-controlled vehicles to steer correctly when driving on a variety of road types.

a Machine learning b Data mining

c Neural networks d Robotics

Q.15 The individual tuples making up the training set are reffered to as ______ and are selected from the database under analysis.

a learing tupes b training tupes

c sampels d database

Q.16 Training perceptron is based on _______.

a supervised learning technique b unsupervised learning

c reinforced learning d stochastic learning

Q.17 List the elements of reinforcement learning.

a Policy b Reward function

c Value function d All of these

Answer Keys for Fill in the Blanks

Q.1	artificial intelligence	Q.2	Supervised	Q.3	brain, machine
Q.4	knowledge	Q.5	machine learning	Q.6	object-oriented
Q.7	MathWorks	Q.8	objects	Q.9	samples
Q.10	experiences	Q.11	supervised	Q.12	Regression,Classification
Q.13	unlabelled	Q.14	Classification and Regression Trees	Q.15	Concept learning
Q.16	Reinforcement	Q.17	reward signal	Q.18	predictive
Q.19	descriptive	Q.20	classification	Q.21	regression

Answer Keys for Multiple Choice Questions

Q.1	b	Q.2	b	Q.3	d
Q.4	a	Q.5	c	Q.6	a
Q.7	d	Q.8	d	Q.9	a
Q.10	c	Q.11	d	Q.12	b
Q.13	b	Q.14	a	Q.15	b
Q.16	a	Q.17	d		

2 Preparing to Model

Syllabus

Machine Learning activities, Types of data in Machine Learning, Structures of data, Data quality and remediation, Data Pre-Processing : Dimensionality reduction, Feature subset selection.

Contents

2.1 Machine Learning Activities

- Following are the typical preparation activities for model :

  a) Understand the types of input data

  b) Find protentional issue in the data

  c) Identify the nature and quality of data

  d) Find out the relationship between data

  e) Apply pre-processing
- Input data is divided into two parts : Training data and testing data.
- Machine learning is about learning some properties of a data set and applying them to new data. This is why a common practice in machine learning to evaluate an algorithm is to split the data at hand in two sets, one that we call a training set on which we learn data properties and one that we call a testing set, on which we test these properties.
- In training data, data is assigning the labels. In test data, data labels are unknown but not given. The training data consist of a set of training examples.
- The real aim of supervised learning is to do well on test data that is not known during learning. Choosing the values for the parameters that minimize the loss function on the training data is not necessarily the best policy.
- The training error is the mean error over the training sample. The test error is the expected prediction error over an independent test sample.
- Problem is that training error is not a good estimator for test error. Training error can be reduced by making the hypothesis more sensitive to training data, but this may lead to over fitting and poor generalization.
- **Training set :** A set of examples used for learning, where the target value is known.
- **Test set :** It is used only to assess the performances of a classifier. It is never used during the training process so that the error on the test set provides an unbiased estimate of the generalization error.
- Training data is the knowledge about the data source which we use to construct the classifier.
- Fig. 2.1.1 shows four step process of machine learning.

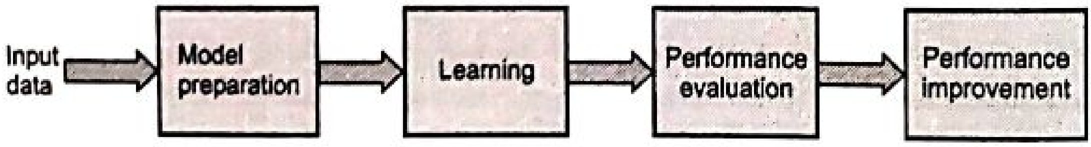


Fig. 2.1.1 Process of machine learning

Step 1	Model preparation	• Understand the types of input data • Find protentional issue in the data • Identify the nature and quality of data • Find out the relationship between data • Apply pre-processing
Step 2	Learning	• Data partitioning • Model selection • Cross-validation
Step 3	Performance evaluation	• Examine model performance • Visualize performance
Step 4	Performance improvement	• Tuning model • Ensembling • Bagging • Bosting

2.2 Types of Data in Machine Learning

- Data set is collection of related records or information. The information may be on some entity or some subject area.
- Collection of data objects and their attributes. Attributes captures the basic characteristics of an object.
- Each row of a data set is called a record. Each data set also has multiple attributes, each of which gives information on a specific characteristic.
- Following is an example of data set.

Emp-ID	Name	Department	Age
TE1	Vilas Bagade	Account	39
TE2	Iresh Dhotre	EDP	40
TE3	Ganesh Patil	Sales	33
TE4	Rupali Tambe	Account	54
TE5	Rakshita Kale	Account	28
TE6	Mahesh Awati	EDP	34
TE7	Ranjeet Bhosale	HR	57

- For example, in the data set on Emp, there are four attributes namely Emp-ID, Name, Department and Age, each of which understandably is a specific characteristic about the employee entity.
- Attributes can also be termed as feature, variable, dimension or field. A row or record represents a point in the four-dimensional data space as each row has specific values for each of the four attributes or features.

2.2.1 Qualitative and Quantitative Data

- Data can broadly be divided into following two types :

1. Qualitative data
2. Quantitative data

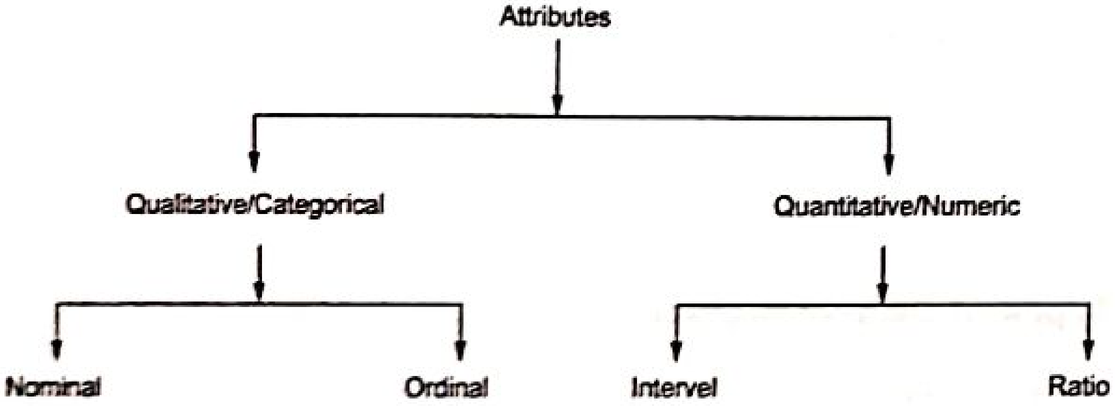


Fig. 2.2.1

Qualitative data :

- **Qualitative data** provides information about the quality of an object or information which cannot be measured. Qualitative data cannot be expressed as a number. Data that represent nominal scales such as gender, economic status, religious preference are usually considered to be qualitative data.
- **Qualitative data** is data concerned with descriptions, which can be observed but cannot be computed. Qualitative data is also called categorical data. Qualitative data can be further subdivided into two types as follows :

1. Nominal data
2. Ordinal data

Nominal data

- A nominal data is the 1st level of measurement scale in which the numbers serve as "tags" or "labels" to classify or identify the objects.

- A nominal data usually deals with the non-numeric variables or the numbers that do not have any value. While developing statistical models, nominal data are usually transformed before building the model.
- It is also known as categorical variables

Characteristics of nominal data :

1. A nominal data variable is classified into two or more categories. In this measurement mechanism, the answer should fall into either of the classes.
2. It is qualitative. The numbers are used here to identify the objects.
3. The numbers don't define the object characteristics. The only permissible aspect of numbers in the nominal scale is "counting."

- Example :
  1. Gender : Male, Female, Other.
  2. Hair color : Brown, Black, Blonde, Red, Other.

Ordinal data

- Ordinal data is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude.
- Ordinal represents the "order." Ordinal data is known as qualitative data or categorical data. It can be grouped, named and also ranked.
- Characteristics of the ordinal data :
  a) The ordinal data shows the relative ranking of the variables.
  b) It identifies and describes the magnitude of a variable.
  c) Along with the information provided by the nominal scale, ordinal scales give the rankings of those variables.
  d) The interval properties are not known.
  e) The surveyors can quickly analyze the degree of agreement concerning the identified order of variables.
- Examples :
  a) University ranking : 1st, 9th, 87th...
  b) Socioeconomic status : Poor, middle class, rich.
  c) Level of agreement : Yes, maybe, no.
  d) Time of day : Dawn, morning, noon, afternoon, evening, night.

Quantitative data

- **Quantitative data** is the one that focuses on numbers and mathematical calculations and can be calculated and computed.

- **Quantitative data** are anything that can be expressed as a number, or quantified. Examples of quantitative data are scores on achievement tests, number of hours of study, or weight of a subject. These data may be represented by ordinal, interval or ratio scales and lend themselves to most statistical manipulation.
- There are two types of quantitative data : Interval data and Ratio data

Interval data :

- Interval data corresponds to a variable in which the value is chosen from an interval set.
- It is defined as a quantitative measurement scale in which the difference between the two variables is meaningful. In other words, the variables are measured in an exact manner, not as in a relative way in which the presence of zero is arbitrary.
- Characteristics of interval data :
  a) The interval data is quantitative as it can quantify the difference between the values.
  b) It allows calculating the mean and median of the variables
  c) To understand the difference between the variables, you can subtract the values between the variables
  d) The interval scale is the preferred scale in statistics as it helps to assign any numerical values to arbitrary assessment such as feelings, calendar types, etc.
- Examples :
  1. Celsius temperature.
  2. Fahrenheit temperature.
  3. Time on a clock with hands.

Ratio data :

- Any variable for which the ratios can be computed and are meaningful is called ratio data.
- It is a type of variable measurement scale. It allows researchers to compare the differences or intervals. The ratio scale has a unique feature. It possesses the character of the origin or zero points.
- Characteristics of ratio data :
  a) Ratio scale has a feature of absolute zero.
  b) It doesn't have negative numbers, because of its zero - point feature.
  c) It affords unique opportunities for statistical analysis. The variables can be orderly added, subtracted, multiplied, divided. Mean, median, and mode can be calculated using the ratio scale.

d) Ratio data has unique and useful properties. One such feature is that it allows unit conversions like kilogram - calories, gram - calories, etc.

- Examples : Age, Weight, Height, Ruler measurements, Number of children

2.2.2 Difference between Qualitative and Quantitative Data

Qualitative data	Quantitative data
Qualitative data provides information about the quality of an object or information which cannot be measured	Quantitative data relates to information about the quantity of an object; hence it can be measured
Types : Nominal data and Ordinal data	Types : Interval data and Ratio data
Narratives often make use of adjectives and other descriptive words to refer to data on appearance, color, texture, and other qualities	Measure's quantities such as length, size, amount, price, and even duration.
They are descriptive rather than numerical in nature	Expressed in numerical form.
For example : • The team is well prepared. • The leaf feels waxy. • The river is peaceful.	For example : • The team has 7 players. • The leaf weighs 2 ounces. • The river is 25 miles long.

2.3 Structures of Data

- A data dictionary is a centralized repository of metadata. Metadata is data about data.
- A data dictionary is a repository of names, definitions, and attributes that provides contextual information about data. A data dictionary traditionally refers to a database dictionary, metadata repository or business glossary. It primarily focuses on the meaning or definition of all columns in a data table.
- In case the data dictionary is not available, we need to use standard library function of the machine learning tool that we are using and get the details.

2.3.1 Exploring Numerical Data

- There are two most effective mathematical plots to explore numerical data : Box plot and histogram

1) Understanding central tendency :

- Central tendency is a descriptive summary of a dataset through a single value that reflects the center of the data distribution

- To understand the nature of numeric variables, we can apply the measures of central tendency of data, i.e. mean and median

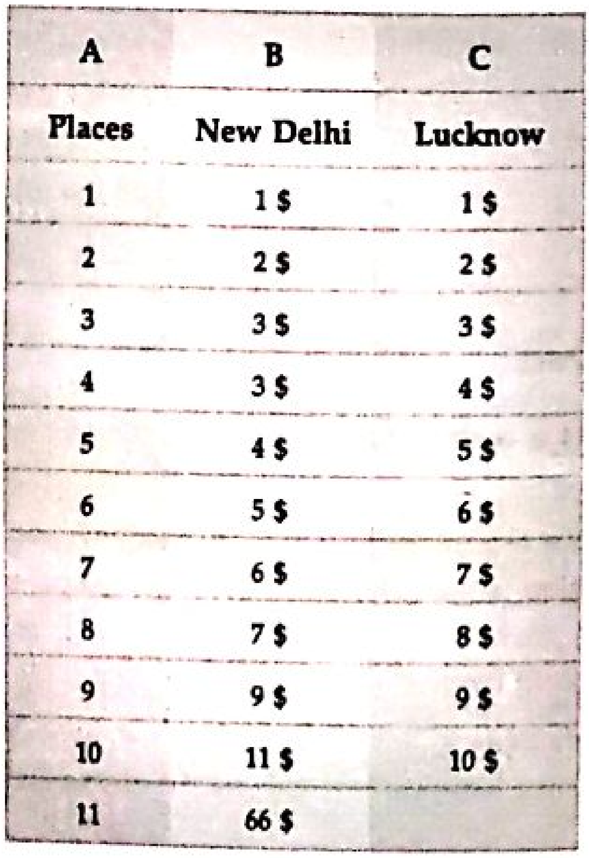
Let $x_1, x_2, x_3, \ldots x_n$ be the set 'n' values of the variate, then arithmetic mean or mean is given as,

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

Median :

Let the values of the variable are arranged in the ascending order of magnitude. Then median is the middle item, if number of values are odd and median will be mean of two middle terms if the number of values in even.

- Median is the mid-value that divide total frequency in two equal parts.
- Example : Below is the data set of pizza price is given cities. Find Mean and Median of both the cities .



A	B	C
Places	New Delhi	Lucknow
1	1 $	1 $
2	2 $	2 $
3	3 $	3 $
4	3 $	4 $
5	4 $	5 $
6	5 $	6 $
7	6 $	7 $
8	7 $	8 $
9	9 $	9 $
10	11 $	10 $
11	66 $	

Solution :

Mean of New Delhi pizza price $= \frac{1+2+3+3+4+5+6+7+9+11+66}{11} = 10.636$

Mean of New Lucknow pizza price $= \frac{1+2+3+4+5+6+7+8+9+10}{10} = 5.5$

Median of New Delhi pizza price $= \frac{N+1}{2}$ obs $= \frac{11+1}{2}$ obs $=$ 6^{th} obs

Here 6^{th} obs = 5

Median = 5

Median of Lucknow pizza price = (N/2) + ((N+1)/2) = (5/2) + (6/2) = 5.5

Median = 5.5

- The mean has one main disadvantage : It is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below :

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15 K	18 K	16 K	14 K	15 K	15 K	12 K	17 K	90 K	95 K

- The mean salary for these ten staff is $30.7 K. However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the $12 K to 18 K range.
- The mean is being skewed by the two large salaries. Therefore, in this situation, we would like to have a better measure of central tendency. As we will find out later, taking the median would be a better measure of central tendency in this situation..

2) Understanding data spread :

Definition : It is the scatteredness or spread of data about an average value.

- It gives an idea about how individual values difffer from the central value, i.e. whether they are closely packed around central value or widely scattered away from it.

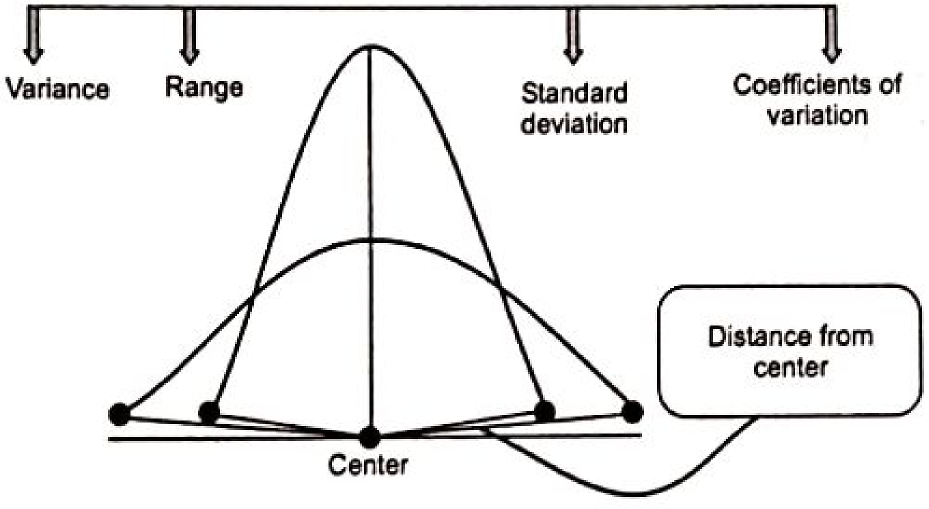


Fig. 2.3.1 Measures of dispersion

- The magnitude of the variation is called dispersion.
- Fig. 2.3.1 shows measures of dispersion.

Variance :

- The second central moment is called varidation. It is given as,

$$\sigma_x^2 = \text{Var}[X] = E[(X-m_x)]^2 = \int_{-\infty}^{\infty} (x-m_x)^2 f_X(x)\,dx$$

or

$$\sigma_x^2 = \overline{X^2} - m_x^2 = E[X^2] - m_x^2$$

- Variance can also be given as,

$$\sigma_x^2 = \frac{1}{N}\sum_i f_i (x_i - \bar{x})^2 \qquad \text{Here } N = \sum_i f_i$$

Let 'A' be assumed mean, 'h' be the magnitude of the class interval and let $d = \frac{x-A}{h}$

Then mean $\bar{x} = A + \frac{h\sum fd}{N}$ and $\sigma_x^2 = h^2\left[\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2\right]$

Standard deviation :

- It is the measure of spread over the values of 'X' relative to mean value. It is given as,

$$\sigma_x = \sqrt{\text{Variance}} = \sqrt{E(X^2) - m_x^2}$$

$$\text{S.D}\sigma_x = \sqrt{\frac{1}{N}\sum_i f_i (x_i - \bar{x})^2}$$

- Standard deviation of a data is measured as follows :

  Standard deviation (x) = $\sqrt{\text{Variance (x)}}$

- Larger value of variance or standard deviation indicates more dispersion in the data and vice versa.

Example 2.3.1 *Consider the data values of two attributes.*

Attribute 1 values : 44, 46, 48, 45, 47 Calculate variance

Solution :

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2$$

$$= \frac{44^2+46^2+48^2+45^2+47^2}{5} - \left(\frac{44+46+48+45+47}{5}\right)^2$$

$$= \frac{1936+2116+2304+2025+2209}{5} - \left(\frac{230}{5}\right)^2$$

$$= \frac{10590}{5} - (46)^2 = 25$$

Difference between standard deviation and variance :

Standard deviation	Variance
Standard deviation is a measure of dispersion of the values of a data set from their mean.	It is the statistical measure of how far the numbers are spread in a data set from their average.
It is a common term in statistical theory to calculate central tendency	Variance is primarily used for statistical probability distribution to measure volatility from the mean
It measures the absolute variability of the dispersion	It helps determine the size of the data spread.
It is calculated by taking the square root of the variance.	It is calculated by taking the average of the squared deviation of each value in the data set from the mean
The standard deviation is symbolized by the Greek letter sigma "σ" as in lower case sigma	The notation for the variance of a variable is "σ^2" sigma squared
$\sigma = \sqrt{\sum(x-M)^2/n}$	$\sigma^2 = \sum(x-M)^2/n$
where M = Mean, x = A values in a data set, and n = Number of values	where M = Mean, x = Each value in the data set, n = Number of values in the data set
Used in finance sector as a measure of market and security volatility.	Used in asset allocation

2.3.2 Plotting and Exploring Numerical Data

1. Box plots

- The box plot is a useful graphical display for describing the behaviour of the data in the middle as well as at the ends of the distributions. The box plot uses the median and the lower and upper quartiles. If the lower quartile is Q_1 and the upper quartile is Q_3, then the difference $(Q_3 - Q_1)$ is called the interquartile range or IQ.

- Box plot is also called whisker plot. It shows data using the middle value of the data and the quartiles, or 25 % divisions of the data.
- Box plot shows the five-number summary of a set of data : Minimum, lower quartile, median, upper quartile and maximum.

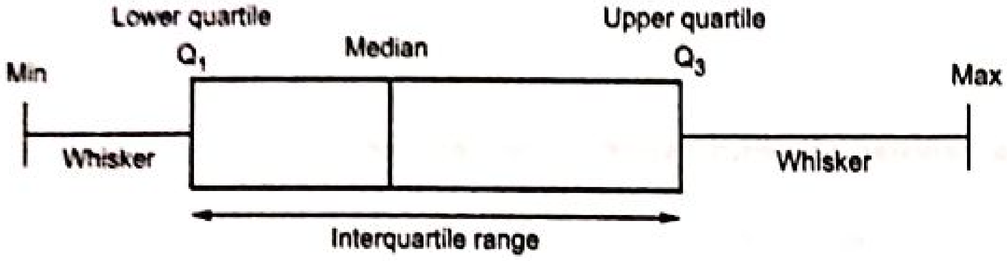


Fig. 2.3.2

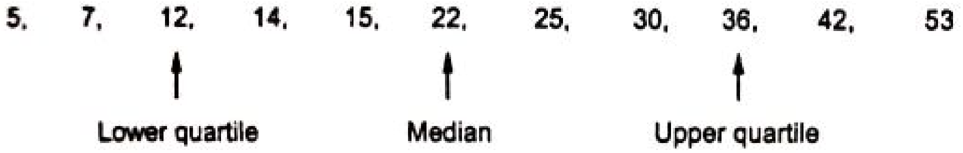
Example 2.3.2 *Construct a box plot for the following data :*

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

Solution :

Step 1 : Arrange the data in ascending order.

Step 2 : Find the median, lower, upper quartile

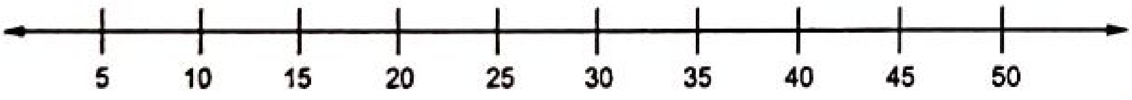


Median (middle value) = 22

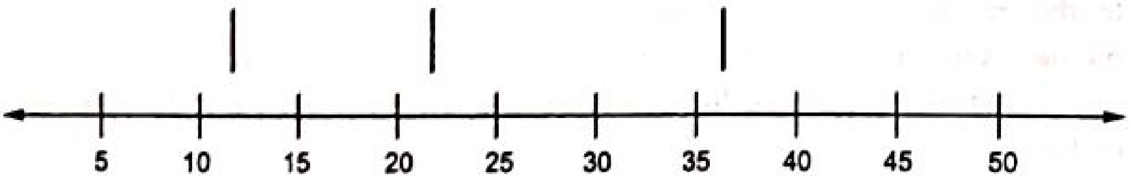
Lower quartile (middle value of the lower half) = 12

Upper quartile (middle value of the upper half) = 36

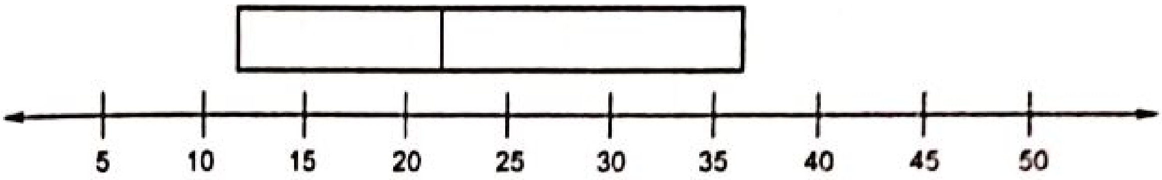
Step 3 : Draw a number line that will include the smallest and the largest data.



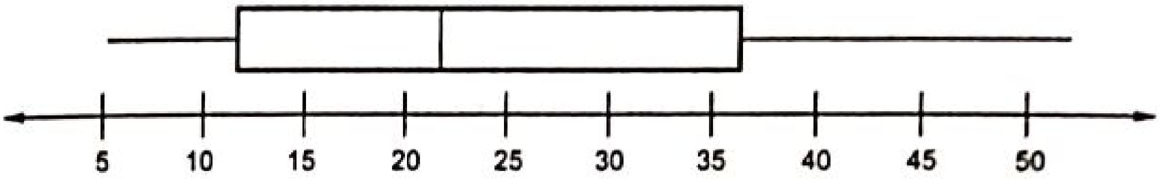
Step 4 : Draw three vertical lines at the lower quartile (12), median (22) and the upper quartile (36), just above the number line.



Step 5 : Join the lines for the lower quartile and the upper quartile to form a box.



Step 6 : Draw a line from the smallest value (5) to the left side of the box and draw a line from the right side of the box to the biggest value (53).



Histogram :

- In a histogram, the data are grouped into ranges (e.g. 10 - 19, 20 - 29) and then plotted as connected bars. Each bar represents a range of data.
- The width of each bar is proportional to the width of each category, and the height is proportional to the frequency or percentage of that category.
- Fig. 2.3.3 shows **distributions of a Histogram.**

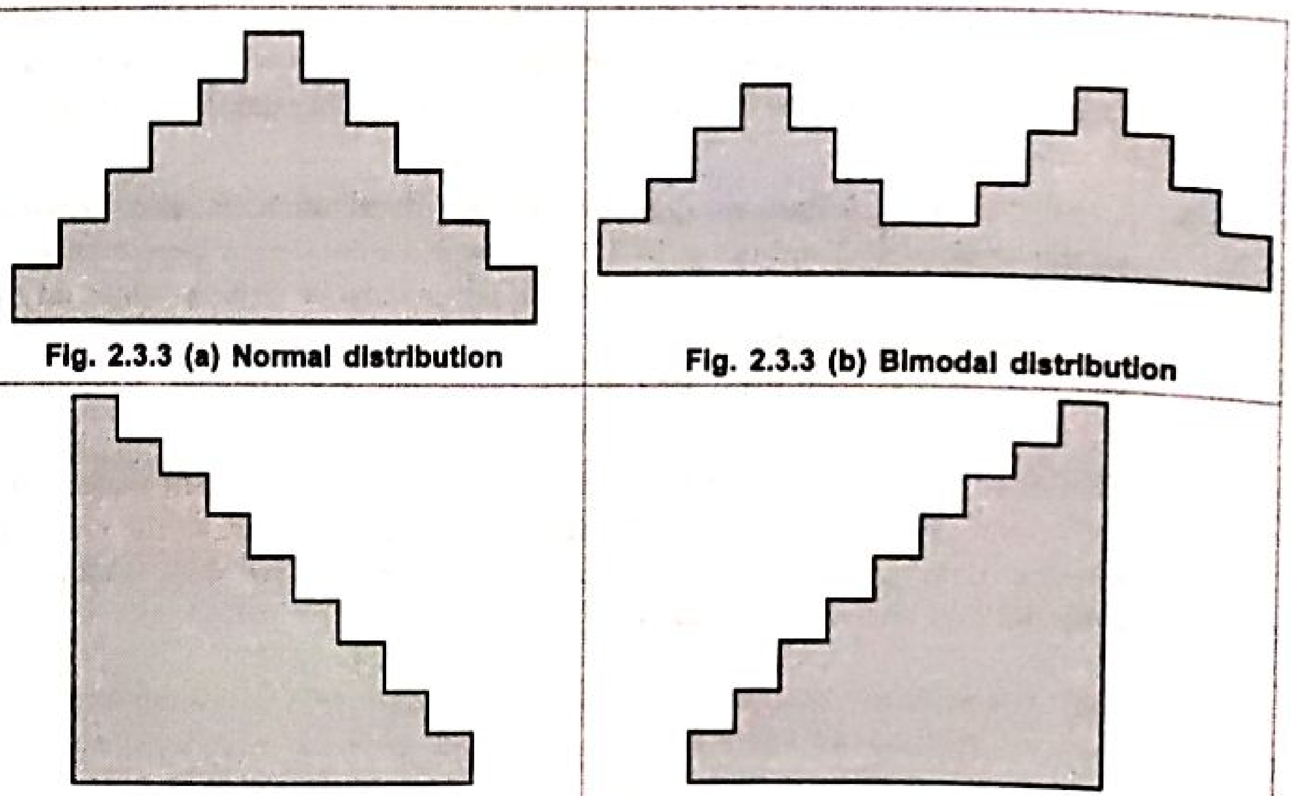


Fig. 2.3.3 (a) Normal distribution

Fig. 2.3.3 (b) Bimodal distribution

Fig. 2.3.3 (c) Right-skewed distribution

Fig. 2.3.3 (d) Left-skewed distribution

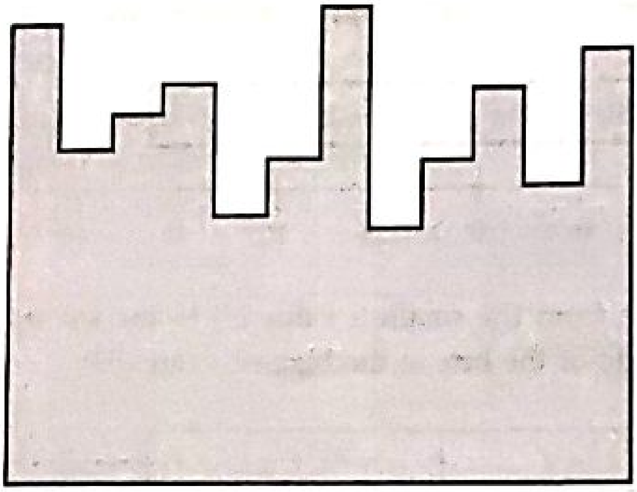


Fig. 2.3.3. (e) Random distribution

1. A normal distribution : In a normal distribution, points on one side of the average are as likely to occur as on the other side of the average.
2. A bimodal distribution : In a bimodal distribution, there are two peaks. In a bimodal distribution, the data should be separated and analyzed as separate normal distributions.
3. A right-skewed distribution : A right-skewed distribution is also called a positively skewed distribution. In a right-skewed distribution, a large number of data values occur on the left side with a fewer number of data values on the right side. A right-skewed distribution usually occurs when the data has a range boundary on the left-hand side of the histogram. For example, a boundary of 0.
4. A left-skewed distribution : A left-skewed distribution is also called a negatively skewed distribution. In a left-skewed distribution, a large number of data values occur on the right side with a fewer number of data values on the left side. A right-skewed distribution usually occurs when the data has a range boundary on the right-hand side of the histogram. For example, a boundary such as 100.
5. A random distribution : A random distribution lacks an apparent pattern and has several peaks. In a random distribution histogram, it can be the case that different data properties were combined. Therefore, the data should be separated and analyzed separately.

2.3.3 Exploring Relationship between Variables

Scatter plot :

- It displays collection of all the points for the set of data limited only for two values. It also called scatter plot, X-Y graph.
- While working with statistical data it is often observed that there are connections between sets of data. For example, the mass and height of persons are related, the taller the person the greater his/her mass.
- To find out whether or not two sets of data are connected scatter diagrams can be used. Fig. 2.3.4 shows scatter diagram.

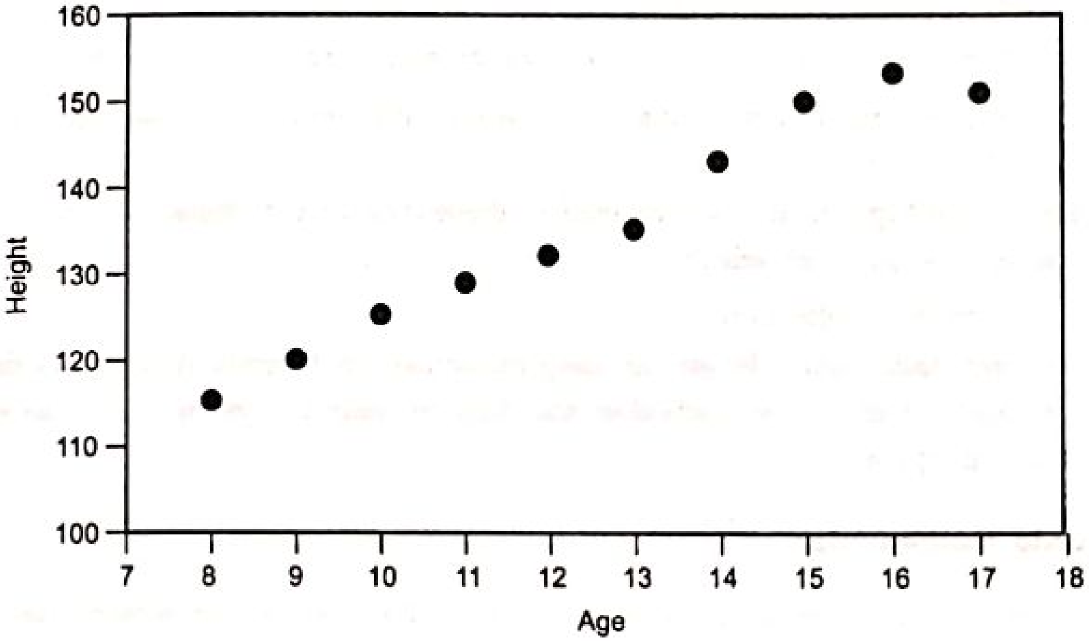


Fig. 2.3.4 Scatter diagram

- Scatter diagram shows the relationship between children's age and height. A scatter diagram is a tool for analyzing relationship between two variables. One variable is plotted on the horizontal axis and the other is plotted on the vertical axis.
- The pattern of their intersecting points can graphically show relationship patterns. Commonly a scatter diagram is used to prove or disprove cause-and-effect relationships.
- While scatter diagram shows relationships, it does not by itself prove that one variable causes other. In addition to showing possible cause and effect relationships, a scatter diagram can show that two variables are from a common cause that is unknown or that one variable can be used as a surrogate for the other.

Two - way cross - tabulations

- Two - way cross - tabulations is also called cross - tab or contingency table. It is used to understand the relationship of two categorical attributes in a concise way

2.4 Data Quality and Remediation

- Data remediation is the process of cleansing, organizing and migrating data so that it's properly protected and best serves its intended purpose

2.4.1 Data Quality

- A data which has the right quality helps to achieve better prediction accuracy, in case of supervised learning. Data quality problems are
  1. Certain data elements without a value or data with a missing value.
  2. Data elements having value surprisingly different from the other elements, which we term as outliers
- There are multiple factors which lead to these data quality issues.
  a) Incorrect sample set selection

  b) Errors in data collection
- Measuring data quality levels can help organizations identify data errors that need to be resolved and assess whether the data in their IT systems is fit to serve its intended purpose.

2.4.2 Data Remediation

- Outliers are data elements with an abnormally high value which may impact prediction accuracy, especially in regression models.
- An **outlier** is an observation that lies an abnormal distance from other values in a random sample from a population.
- First quartile (Q_1) : The first quartile is the value, where 25 % of the values are smaller than Q_1 and 75 % are larger.
- Third quartile (Q_3) : The third quartile is the value, where 75 % of the values are smaller than Q_3 and 25 % are larger.
- Outliers are data elements with an abnormally high value which may impact prediction accuracy, especially in regression models
- Outlier detection is the process of detecting and subsequently excluding outliers from a given set of data.
- Fig. 2.4.1 shows outliers detection. Here O_1 and O_2 seem outliers from the rest.

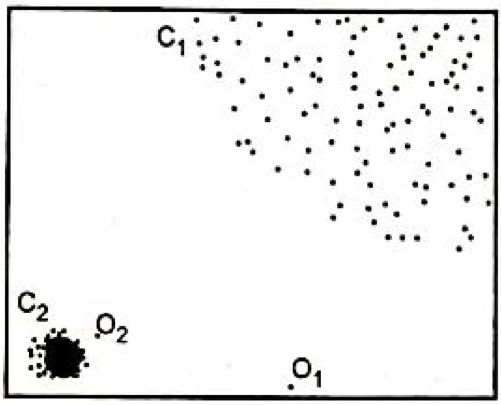


Fig. 2.4.1 Outliers detection

- An outlier may be defined as a piece of data or observation that deviates drastically from the given norm or average of the data set. An outlier may be caused simply by chance, but it may also indicate measurement error or that the given data set has a heavy - tailed distribution

Handling missing values

- In a data set, one or more data elements may have missing values in multiple records.
- These dirty data will affects on miming procedure and led to unreliable and poor output. Therefore it is important for some data cleaning routines.

How to handle noisy data in data mining ?

- Following methods are used for handling noisy data :
  1. **Ignore the tuple :** Usually done when the class label is missing. This method is not good unless the tuple contains several attributes with missing values.
  2. **Fill in the missing value manually :** It is time-consuming and not suitable for a large data set with many missing values.
  3. **Use a global constant to fill in the missing value :** Replace all missing attribute values by the same constant.
  4. Use the attribute mean to fill in the missing value : For example, suppose that the average salary of staff is Rs 65000/- . Use this value to replace the missing value for salary.
  5. Use the attribute mean for all samples belonging to the same class as the given tuple
  6. Use the most probable value to fill in the missing value

2.5 Data Pre-Processing

- Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Aim to reduce the data size, find the relation between data and normalized them. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing.

- Data which capture from various source is not pure. It contains some noise. It is called dirty data or incomplete data. In this data, there is lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. For example : occupation=" "
- Noisy data which contains errors or outliers. For example : Salary="-10"
- Inconsistent data which contains discrepancies in codes or names. For example : Age = "51" Birthday="03/08/1998"
- Incomplete, noisy, and inconsistent data are commonplace properties of large real - world databases and data warehouses. Incomplete data can occur for a number of reasons

2.5.1 Dimensionality Reduction

- Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.
- Most machine learning and data mining techniques may not be effective for high-dimensional data. Query accuracy and efficiency degrade rapidly as the dimension increases.
- The "dimensionality" simply refers to the number of features (i.e. input variables) in your dataset.
- When the number of features is very large relative to the number of observations in your dataset, certain algorithms struggle to train effective models. This is called the "Curse of Dimensionality," and it's especially relevant for clustering algorithms that rely on distance calculations.
- Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.
- It reduces the time and storage space required. Removal of multi-collinearity improves the interpretation of the parameters of the machine learning model.
- There are many methods to perform dimension reduction.
  1. Missing values : While exploring data, if we encounter missing values, what we do ? Our first step should be to identify the reason then impute missing values / drop variables using appropriate methods. But, what if we have too many missing values ? Should we impute missing values or drop the variables ?

2. Low variance : Let's think of a scenario where we have a constant variable in our data set.
3. Desicion trees : It can be used as a ultimate solution tackle multiple challenges like missing values, outliers and identifying significant variavbles.
4. Random forest : Similar to decision tree is random forest.
5. High coreelation : Dimensions exhitbiting higher correlation can lower down the performance of model. Moreover, it is not good to have multipule variables of similar information or variation also known as "Multicollinearity".

Advantages of dimensionality reduction

- It helps in data compression, and hence reduced storage space.
- It reduces computation time.
- It also helps remove redundant features, if any.

Disadvantages of dimensionality reduction

- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which is sometimes undesirable.
- PCA fails in cases where mean and covariance are not to define datasets.
- We may not know how many principal components to keep in practice, some thumb rules are applied.

2.5.1.1 Principal Component Analysis

- If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called lossless. If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called **lossy.**
- Lossy dimensionality reduction methods are Principal Components Analysis (PCA) and wavelet transforms.
- Principal Component Analysis (PCA) is to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, retains most of the sample's information and useful for the compression and classification of data.
- In PCA, it is assumed that the information is carried in the variance of the features, that is, the higher the variation in a feature, the more information that feature carries.
- Hence, PCA employs a linear transformation that is based on preserving the most variance in the data using the least number of dimensions.

- The most common approach for dimensionality reduction is known as Principal Component Analysis (PCA). PCA is a statistical technique to convert a set of correlated variables into a set of transformed, uncorrelated variables called principal components. The principal components are a linear combination of the original variables.
- A Discrete Wavelet Transform (DWT) is a transform that decomposes a given signal into a number of sets, where each set is a time series of coefficients describing the time evolution of the signal in the corresponding frequency band.
- Another commonly used technique which is used for dimensionality reduction is Singular Value Decomposition (SVD).

2.5.2 Feature Subset Selection

- A good feature representation is central to achieving high performance in any machine learning task.
- Consider an example of text categorization. Assume that we need to train a model for classifying a given document as spam and not spam. If we represent a document as a bag of words, the feature space consists of a vocabulary of all unique words present in all the documents in the training set.
- For a collection of 100,000 to 1,000,000 documents, we can easily expect hundreds of thousands of features. If we further extend this document model to include all possible bigrams and trigrams, we could easily get over a million features.
- A feature tree is a tree such that each internal node is labelled with a feature, and each edge emanating from an internal node is labelled with a literal. The set of literals at a node is called a split.
- Each leaf of the tree represents a logical expression, which is the conjunction of literals encountered on the path from the root of the tree to the leaf. The extension of that conjunction is called the instance space segment associated with the leaf.
- Two features are redundant if they are highly correlated, regardless of whether they are correlated with the task or not.

Feature construction and transformation

- Feature construction involves transforming a given set of input features to generate a new set of more powerful features which can then use for prediction.
- Feature construction methods may be applied to pursue two distinct goals : Reducing data dimensionality and improving prediction performance.
- Steps :
  1. Start with an initial feature space F_0

2. Transform F_0 to construct a new feature space F_N
3. Select a subset of features F_i from F_N
4. If some terminating criteria is achieved : Go back to step 3 otherwise set $F_T = F_i$
5. F_T is the newly constructed feature space

- The initial feature space F_0 consists of manually constructed features that often encode some basic domain knowledge.
- The task of constructing appropriate features is often highly application specific and labour intensive. Thus, building auto-mated feature construction methods that require minimal user effort is challenging. In particular we want methods that :

  1. Generate a set of features that help improve prediction accuracy.
  2. Are computationally efficient.
  3. Are generalizable to different classifiers.
  4. Allow for easy addition of domain knowledge.

- Genetic programming is an evolutionary algorithm - based technique that starts with a population of individuals, evaluates them based on some fitness function and constructs a new population by applying a set of mutation and crossover operators on high scoring individuals and eliminating the low scoring ones.
- In the feature construction paradigm, genetic programming is used to derive a new feature set from the original one.
- Individuals are often tree like representations of features, the fitness function is usually based on the prediction performance of the classifier trained on these features while the operators can be applications specific.
- The method essentially performs a search in the new feature space and helps generate a high performing subset of features. The newly generated features may often be more comprehensible and intuitive than the original feature set, which makes GP-related methods well-suited for such tasks.
- In decision trees, the model explicitly selects features that are highly correlated with the label. In particular, by limiting the depth of the decision tree, one can at least hope that the model will be able to throw away irrelevant features.
- In the case of K-nearest neighbours, the situation is perhaps more terrible. Since KNN weighs each feature just as much as another feature, the introduction of irrelevant features can completely mess up KNN prediction.
- Feature extraction is a process that extracts a set of new features from the original features through some functional mapping.

- Transformation studies ways of mapping original attributes to new features. Different mappings can be employed to extract features.
- In general, the mappings can be categorized into linear or nonlinear transformations. One could categorize transformations along two dimensions linear and labeled, linear and non labeled, nonlinear and labeled, nonlinear and non labeled.

Feature selection

- Feature selection is a process that chooses a subset of features from the original features so that the feature space is optimally reduced according to a certain criterion.
- Feature selection is a critical step in the feature construction process. In text categorization problems, some words simply do not appear very often.
- Perhaps the word "groovy" appears in exactly one training document, which is positive. Is it really worth keeping this word around as a feature ? It's a dangerous endeavour because it's hard to tell with just one training example if it is really correlated with the positive class, or is it just noise.
- You could hope that your learning algorithm is smart enough to figure it out. Or you could just remove it.
- There are three general classes of feature selection algorithms : Filter methods, wrapper methods and embedded methods.
- The role of feature selection is as follows:

  1. To reduce the dimensionality of feature space.
  2. To speed up a learning algorithm.
  3. To improve the predictive accuracy of a classification algorithm.
  4. To improve the comprehensibility of the learning results.
- Features selection algorithms are as follows :

  1. Instance based approaches : There is no explicit procedure for feature subset generation. Many small data samples are sampled from the data. Features are weighted according to their roles in differentiating instances of different classes for a data sample. Features with higher weights can be selected.
  2. Nondeterministic approaches : Genetic algorithms and simulated annealing are also used in feature selection.
  3. Exhaustive complete approaches : Branch and bound evaluates estimated accuracy and ABB checks an inconsistency measure that is monotonic. Both start with a full feature set until the pre-set bound cannot be maintained.

2.6 Fill in the Blanks

Q.1 _______ set is collection of related records or information.

Q.2 Each row of a data set is called a _______ .

Q.3 Qualitative data is also called _______ data.

Q.4 _______ data provides information about the quality of an object or information which cannot be measured.

Q.5 Dimensionality reduction helps in reducing irrelevance and _______ in features.

Q.6 Dimensionality reduction refers to the techniques of reducing the dimensionality of a data set by creating new attributes by combining the original _______ .

Q.7 Lossy dimensionality reduction methods are _______ and wavelet transforms.

Q.8 An _______ is an observation that lies an abnormal distance from other values in a random sample from a population.

Q.9 Exploration of numerical data can be best done using _______ and _______.

Q.10 Data can be broadly divided into _______ data and _______ data.

2.7 Multiple Choice Questions

Q.1 Data can be broadly divided into _______ .

a qualitative data
b quantitative data
c qualitative and Quantitative data
d ratio data

Q.2 Feature selection tries to eliminate features which are _______ .

a rich
b redundant
c irrelevant
d relevant

Q.3 Principal component analysis is used for _______ .

a dimensionality Enhancement
b LU decomposition
c QR decomposition
d dimensionality reduction

Q.4 Which of the following methods to perform dimension reduction ?

a Missing values
b Decision tree
c Random forest
d All of these

Answer Keys for Fill In the Blanks

Q.1	Data	Q.2	record
Q.3	categorical	Q.4	Qualitative
Q.5	redundancy	Q.6	attributes
Q.7	principal components analysis	Q.8	outlier
Q.9	box plots, histograms	Q.10	Qualitative, Quantitative

Answer Keys for Multiple Choice Questions

Q.1	c	Q.2	c
Q.3	d	Q.4	d

3 Modelling and Evaluation

Syllabus

Selecting a Model : Predictive/Descriptive, Training a Model for supervised learning, model representation and interpretability, Evaluating performance of a model, Improving performance of a model.

Contents

3.1 Selecting a Model

- Structured representation of raw input data to the meaningful pattern is called a model. The model might have different forms. It might be a mathematical equation, it might be a graph or tree structure, it might be a computational block, etc.
- Given easy-to-use machine learning libraries like scikit-learn and Keras, it is straightforward to fit many different machine learning models on a given predictive modeling dataset
- Model selection is the task of selecting a statistical model from a set of candidate models, given data.
- The decision regarding which model is to be selected for a specific data set is taken by the learning task, based on the problem to be solved and the type of data.
- The process of assigning a model, and fitting a specific model to a data set is called model training.
- Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.
- Model selection is a process that can be applied both across different types of models (e.g. logistic regression, SVM, KNN, etc.) and across models of the same type configured with different model hyperparameters.
- Fitting models is relatively straightforward, although selecting among them is the true challenge of applied machine learning.
- All models have some predictive error, given the statistical noise in the data, the incompleteness of the data sample, and the limitations of each different model type. Therefore, the notion of a perfect or best model is not useful. Instead, we must seek a model that is "good enough."
- The best approach to model selection requires "sufficient" data, which may be nearly infinite depending on the complexity of the problem.
- In this ideal situation, we would split the data into training, validation, and test sets, then fit candidate models on the training set, evaluate and select them on the validation set, and report the performance of the final model on the test set.

3.1.1 Predictive models

- Predictive modelling is also called predictive analytics. It is a mathematical process that seeks to predict future events or outcomes by analyzing patterns that are likely to forecast future results.

- If you are trying to predict a continuous target, then you will need a regression model. But if you are trying to predict a discrete target, then you will need a classification model.
- The predictive models have a clear focus on what they want to learn and how they want to learn.
- Predictive analysis provides answers of the future queries that move across using historical data as the chief principle for decisions
- It involves the supervised learning functions used for the prediction of the target value. The methods fall under this mining category are the classification, time-series analysis and regression.
- Data modeling is the necessity of the predictive analysis, which works by utilizing some variables to anticipate the unknown future data values for other variables.
- It provides organizations with actionable insights based on data. It provides an estimation regarding the likelihood of a future outcome.
- To do this, a variety of techniques are used, such as machine learning, data mining, modeling and game theory.
- Predictive modeling can, for example, help to identify any risks or opportunities in the future.
- Predictive analytics can be used in all departments, from predicting customer behaviour in sales and marketing, to forecasting demand for operations or determining risk profiles for finance.
- A very well-known application of predictive analytics is credit scoring used by financial services to determine the likelihood of customers making future credit payments on time. Determining such a risk profile requires a vast amount of data, including public and social data.
- Historical and transactional data are used to identify patterns, and statistical models and algorithms are used to capture relationships in various datasets.
- Predictive analytics has taken off in the big data era, and there are many tools available for organisations to predict future outcomes.
- The target feature is known as a class and the categories to which classes are divided into are called levels. The k-Nearest Neighbor, Naïve Bayes, and decision tree are the popular classification models.
- Predictive models may also be used to predict numerical values of the target feature based on the predictor features. Popular regression models are Linear Regression and Logistic Regression.

3.1.2 Descriptive models

- A descriptive model is used for tasks that would benefit from the insight gained from summarizing data in new and interesting ways. The process of training a descriptive model is called unsupervised learning.
- So, in unsupervised learning algorithm, we do not have any target or outcome variable to predict / estimate. It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: Apriori algorithm, K-means.
- Descriptive Analytics is the conventional form of business intelligence and data analysis, seeks to provide a depiction or "summary view" of facts and figures in an understandable format, to either inform or prepare data for further analysis.
- Two primary techniques are used for reporting past events : data aggregation and data mining.
- It presents past data in an easily digestible format for the benefit of a wide business audience.
- A set of techniques for reviewing and examining the data set to understand the data and analyze business performance.
- Descriptive analytics helps organisations to understand what happened in the past. It helps to understand the relationship between product and customers.
- The objective of this analysis is to understanding, what approach to take in the future. If we learn from past behaviour, it helps us to influence future outcomes.
- Company reports is an example of descriptive analytics which simply provides a historic review of company operations, stakeholders, customers and financials.
- It also helps to describe and present data in such format, which can be easily understood by a wide variety of business readers.
- The descriptive modeling task called pattern discovery is used to identify useful associations within data. Pattern discovery is often used for market basket analysis on retailers' transactional purchase data.
- Here, the goal is to identify items that are frequently purchased together, such that the learned information can be used to refine marketing tactics.
- For instance, if a retailer learns that swimming trunks are commonly purchased at the same time as sunglasses, the retailer might reposition the items more closely in the store or run a promotion to "up-sell" customers on associated items.

3.2 Training a Model for Supervised Learning

3.2.1 Holdout Method

- The data is split into two different datasets labelled as a training and a testing dataset. This can be a 60/40 or 70/30 or 80/20 split. This technique is called the **hold-out validation technique.**
- Suppose we have a database with house prices as the dependent variable and two independent variables showing the square footage of the house and the number of rooms.
- Now, imagine this dataset has 30 rows. The whole idea is that you build a model that can predict house prices accurately.
- To 'train' your model, or see how well it performs, we randomly subset 20 of those rows and fit the model.
- The second step is to predict the values of those 10 rows that we excluded and measure how well our predictions were.
- As a rule of thumb, experts suggest to randomly sample 80 % of the data into the training set and 20 % into the test set.
- Training set : Used to train the classifier.

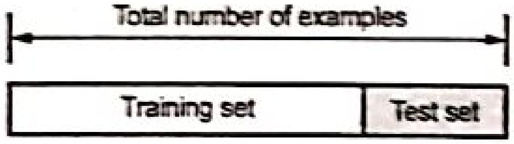


Fig. 3.2.1

- The holdout method has two, basic drawbacks :
  1. It requires extra dataset
  2. It is a single train-and-test experiment, the holdout estimate of error rate will be misleading if we happen to get an "unfortunate" split.

3.2.2 Cross-Validation

- Cross-validation is a technique for evaluating estimating performance by training several machine learning models on subsets of the available input data and evaluating them on the complementary subset of the data. Use cross-validation to detect overfitting, i.e., failing to generalize a pattern.
- In general, machine learning involves deriving models from data, with the aim of achieving some kind of desired behaviour, e.g., prediction or classification.
- Fig. 3.2.2 shows cross-validation.

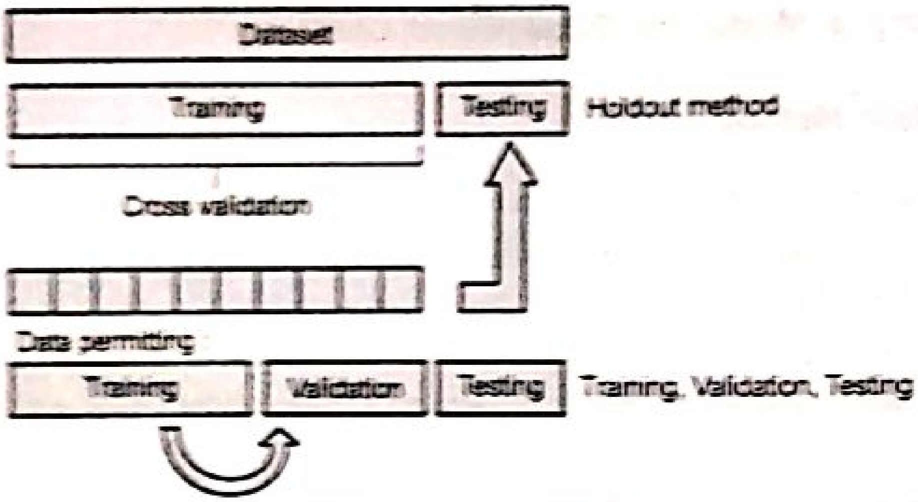


Fig. 3.2.2 : Cross validation

- But this generic task is broken down into a number of special cases. When training is done, the data that was removed can be used to test the performance of the learned model on "new" data. This is the basic idea for a whole class of model evaluation methods called **cross validation**.
- Types of cross validation methods are holdout, K-fold and Leave-one-out.
- The holdout method is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The function approximate fits a function using the training set only.
- The K-fold cross validation is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times.
- Each time, one of the k subsets is used as the test set and the other k – 1 subsets are put together to form a training set. Then the average error across all k trials is computed.
- Leave-one-out cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set.
- That means that N separate times, the function approximate is trained on all the data except for one point and a prediction is made for that point.
- Cross-validation ensures non-overlapping test sets.

K-fold cross-validation :

- In this technique, k – 1 folds are used for training and the remaining one is used for testing as shown in Fig. 3.2.3.
- The advantage is that entire data is used for training and testing. The error rate of the model is average of the error rate of each iteration.
- This technique can also be called a form the repeated hold-out method. The error rate could be improved by using stratification technique.

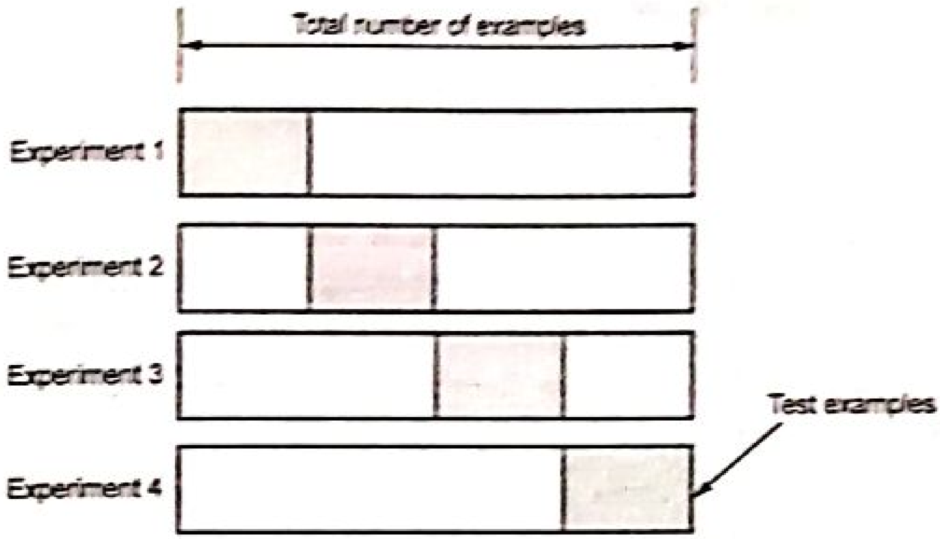


Fig. 3.2.3 K-fold cross validation

3.2.3 Bootstrap

- Ensemble classifiers such as bagging, boosting and model averaging are known to have improved accuracy and robustness over a single model. Although unsupervised models, such as clustering, do not directly generate label prediction for each individual, they provide useful constraints for the joint prediction of a set of related objects.
- For given a training set of size n, create m samples of size n by drawing n examples from the original data, with replacement. Each bootstrap sample will on average contain 63.2 % of the unique training examples, the rest are replicates. It combines the m resulting models using simple majority vote.
- In particular, on each round, the base learner is trained on what is often called a "bootstrap replicate" of the original training set. Suppose the training set consists of n examples.
- Then a bootstrap replicate is a new training set that also consists of n examples, and which is formed by repeatedly selecting uniformly at random and with replacement n examples from the original training set. This means that the same example may appear multiple times in the bootstrap replicate, or it may appear not at all.
- It also decreases error by decreasing the variance in the results due to unstable learners, algorithms (like decision trees) whose output can change dramatically when the training data is slightly changed.

3.2.4 Lazy vs. Eager Learner

- Eager learning : Given a set of training set, constructs a classification model before receiving new data to classify. For example, decision tree induction, Bayesian classification, rule-based classification etc.
- Lazy learning : Simply stores training data and waits until it is given a new instance. Lazy learners take less time in training but more time in predicting. For example, k-nearest-neighbor classifiers, case-based reasoning classifiers
- Instance-based methods are also known as lazy learning because they do not generalize until needed.
- The eager learner must create a global approximation. The lazy learner can create many local approximations.

3.3 Model Representation and Interpretability

- In addition to using models for prediction, the ability to interpret what a model has learned is receiving an increasing amount of attention.
- Interpretability has to do with how accurate a machine learning model can associate a cause to an effect.
- If a model can take the inputs, and routinely get the same outputs, the model is interpretable :
  1. If you overeat your magi at dinnertime and you always have troubles sleeping, the situation is interpretable.
  2. If all 2019 polls showed " ABC party" win and the "XYZ party" candidate took office, all those models showed low interpretability.
- Interpretability poses no issue in low-risk scenarios. If a model is recommending movies to watch, that can be a low-risk task
- Fitness of a target function approximated by a learning algorithm determines how correctly it is able to classify a set of data it has never seen.

3.3.1 Underfitting and Overfitting

- Training error can be reduced by making the hypothesis more sensitive to training data, but this may lead to overfitting and poor generalization.
- Overfitting occurs when a statistical model describes random error or noise instead of the underlying relationship. Overfitting is when a classifier fits the training data too tightly. Such a classifier works well on the training data but not on independent test data. It is a general problem that plagues all machine learning methods.

- Underfitting : If we put too few variables in the model, leaving out variables that could help explain the response, we are **underfitting.** Consequences :
  1. Fitted model is not good for prediction of new data - prediction is biased
  2. Regression coefficients are biased
  3. Estimate of error variance is too large
- Because of overfitting, low error on training data and high error on test data. Overfitting occurs when a model begins to memorize training data rather than learning to generalize from trend.
- The more difficult a criterion is to predict, the more noise exists in past information that need to be ignored. The problem is determining which part to ignore.
- Overfitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. We can determine whether a predictive model is underfitting or overfitting the training data by looking at the prediction error on the training data and the evaluation data. Fig. 3.3.1 shows underfiting and overfiting.

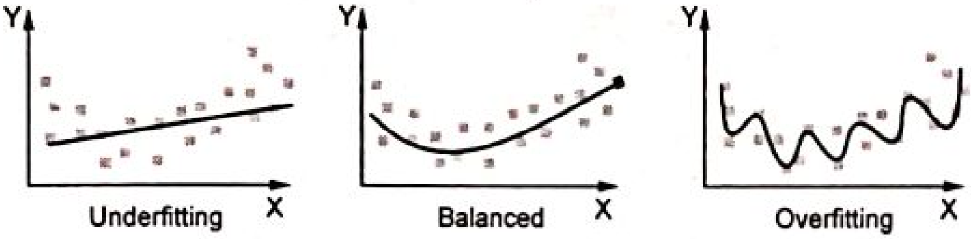


Fig. 3.3.1

- Reasons for overfitting
  1. Noisy data
  2. Training set is too small
  3. Large number of features
- In the machine learning the more complex model is said to show signs of overfitting, while the simpler model underfitting. Often several heuristic are developed in order to avoid overfitting, for example, when designing neural networks one may :
  1. Limit the number of hidden nodes
  2. Stop training early to avoid a perfect explanation of the training set, and
  3. Apply weight decay to limit the size of the weights, and thus of the function class implemented by the network

3.3.2 Bias- Variance

- In the experimental practice we observe an important phenomenon called the bias variance dilemma.
- In supervised learning, the class value assigned by the learning model built based on the training data may differ from the actual class value. This error in learning can be of two types, errors due to 'bias' and error due to 'variance'.
- Fig. 3.3.2 shows bias-variance trade off.

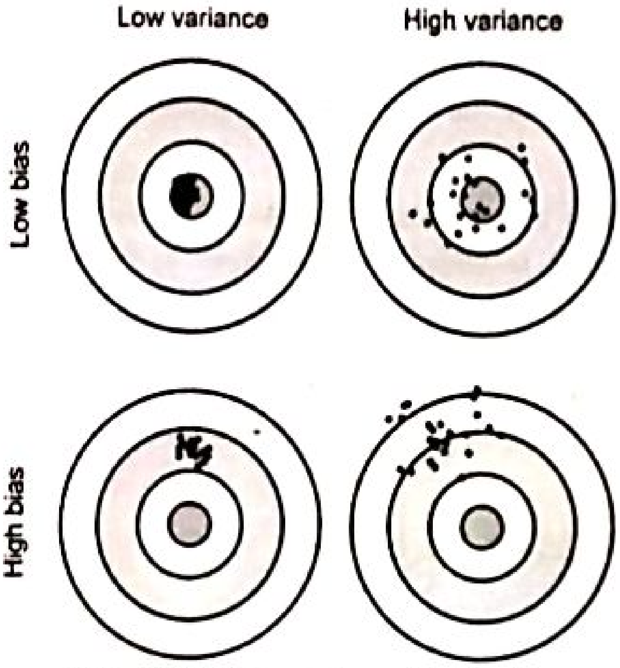


Fig. 3.3.2 Bias-variance trade off

- Give two classes of hypothesis (e.g. linear models and k-NNs) to fit to some training data set, we observe that the more flexible hypothesis class has a low bias term but a higher variance term. If we have parametric family of hypothesis, then we can increases the flexibility of the hypothesis but we still observe the increase of variance.
- The bias-variance-dilemma is the problem of simultaneously minimizing two sources of error that prevent supervised learning algorithm from generalizing beyond their training set :
  1. The bias is error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs.
  2. The variance is error from sensitivity to small fluctuations in the training set. High variance can cause overfitting : modeling the random noise in the training data, rather than the intended outputs.
- In order to reduce the model error, the designer can aim at reducing either the bias or the variance, as the noise components is irreducible.

- As the model increases in complexity, its bias is likely to diminish. However, as the number of training examples is kept fixed, the parametric identification of the model may strongly vary from one DN to another. This will increase the variance term.
- At one stage, the decrease in bias will be inferior to the increase in variance, warning that the model should not be too complex. Conversely, to decrease the variance term, the designer has to simplify its model so that it is less sensitive to a specific training set. This simplification will lead to a higher bias.

3.4 Evaluating Performance of a Model

3.4.1 Supervised Learning : Classification

- Classification is major task of supervised learning. The responsibility of the classification model is to assign class label to the target feature based on the value of the predictor features.
- When performing classification predictions, there's four types of outcomes that could occur. The evaluation measures in classification problems are defined from a matrix with the numbers of examples correctly and incorrectly classified for each class, named confusion matrix.
- Confusion matrix is also called a contingency table.
  1) True positives are when you predict an observation belongs to a class and it actually does belong to that class.
  2) True negatives are when you predict an observation does not belong to a class and it actually does not belong to that class.
  3) False positives occur when you predict an observation belongs to a class when in reality it does not.
  4) False negatives occur when you predict an observation does not belong to a class when in fact it does.
- Confusion matrix goes deeper than classification accuracy by showing the correct and incorrect (i.e. true or false) predictions on each class. In case of a binary classification task, a confusion matrix is a 2x2 matrix. If there are three different classes, it is a 3x3 matrix and so on.

Actual value	A	TP	FN
	B	FP	TN
		A	B
		Predicted value	

- For any classification model, model accuracy is given by total number of correct classifications (True Positive or True Negative) divided by total number of classifications done.

$$\text{Accuracy rate} = \frac{|\text{True negatives}|+|\text{True positives}|}{|\text{False negatives}|+|\text{True positives}|+|\text{True negatives}|+|\text{True positives}|}$$

- The complement of accuracy rate is the error rate, which evaluates a classifier by its percentage of incorrect predictions.

$$\text{Error rate} = \frac{|\text{False negatives}| + |\text{False positives}|}{|\text{False negatives}| + |\text{False positives}| + |\text{True negatives}| + |\text{True positives}|}$$

$$\text{Error rate} = 1 - (\text{Accuracy rate})$$

- The recall accuracy rate predicted as positive.
- The **specificity** is a statistical measure of how well a binary classification test correctly identifies the negatives cases.

$$\text{Recall (R)} = \frac{|\text{True negative}|}{|\text{True positivs}|+|\text{False negative}|}$$

$$\text{Specificity} = \frac{|\text{True positives}|}{|\text{False positives}|+|\text{True negatives}|}$$

- True Positive Rate (TPR) is also called sensitivity, hit rate and recall.

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives}+\text{Number of false negative}}$$

- **Precision** measures how good our model is when the prediction is positive.

$$\text{Precision} = \frac{TP}{TP+FP}$$

- The focus of precision is positive predictions. It indicates how many positive predictions are true.
- F_1 score is the weighted average of precision and recall.

$$F_1\text{\_score} = 2\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

- F_1 score is a more useful measure than accuracy for problems with uneven class distribution because it takes into account both false positive and false negatives.
- Kappa value of a model indicates the adjusted the model accuracy

$$\text{Kappa} = \frac{\text{Total accuracy} - \text{Random accuracy}}{1 - \text{Random accuracy}}$$

- Total accuracy is simply the sum of true positive and true negatives, divided by the total number of items, that is :

$$\text{Total accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

- Random Accuracy is defined as the sum of the products of reference likelihood and result likelihood for each class. That is,

$$\text{Random accuracy} = \frac{\text{Actual False} * \text{Predicted False} + \text{Actual True} * \text{Predicted True}}{\text{Total} * \text{Total}}$$

- In terms of false positives etc., random accuracy can be written as :

$$\text{Random accuracy} = \frac{(TN+FP) * (TN+FN) + (FN+TP) * (FP+TP)}{\text{Total} * \text{Total}}$$

Example 3.4.1 *Consider the following three-class confusion matrix.*

	Predicted		
Actual	15	2	3
	7	15	8
	2	3	45

Calculate precision and recall per class. Also calculate weighted average precision and recall for the classifier.

Solution :

	Predicted			
Actual	15	2	3	20
	7	15	8	30
	2	3	45	50
	24	20	56	100

$$\text{Classifier accuracy} = \frac{15+15+45}{100} = \frac{75}{100} = 0.75$$

Calculate per-class precision and recall :

First class $= \frac{15}{24} = 0.63$ and $\frac{15}{20} = 0.75$

Second class $= \frac{15}{20} = 0.75$ and $\frac{15}{30} = 0.50$

Third class $= \frac{45}{56} = 0.8$ and $\frac{45}{50} = 0.9$

Example 3.4.2 *Calculate accuracy, precision and recall for the following :*

	Predicted +	Predicted –
Actual +	60	15
Actual –	10	15

Solution :

$$\text{Accuracy} = \frac{60+15}{60+10+15+15} = \frac{75}{100} = 0.75 = 75\ \%$$

$$\text{Precision} = \frac{60}{60+10} = 0.8571$$

$$\text{Recall} = \frac{60}{60+15} = 0.8$$

Example 3.4.3 *Calculate true negative rate (t_{nr}), accuracy and pos for the following*

	Predicted +	Predicted –
Actual +	50	25
Actual –	5	20

Solution :

$$\text{Accuracy} = \frac{50+25}{50+5+25+20} = \frac{75}{100} = 0.75 = 75\ \%$$

$$\text{Precision} = \frac{50}{50+5} = 0.9090$$

- True negative rate is also called as specificity.

$$\text{Specificity} = \frac{|\text{True negatives}|}{|\text{False positives}|+|\text{True negatives}|}$$

$$\text{True negative rate} = \frac{20}{5+20} = 0.8$$

ROC Curve :

- Receiver Operating Characteristics (ROC) graphs have long been used in signal detection theory to depict the trade-off between hit rates and false alarm rates over noisy channel. Recent years have seen an increase in the use of ROC graphs in the machine learning community.

- ROC curve summarizes the performance of the model at different threshold values by combining confusion matrices at all threshold values. ROC curves are typically used in binary classification to study the output of a classifier.
- An ROC plot plots true positive rate on the Y-axis against false positive rate on the X-axis; a single contingency table corresponds to a single point in an ROC plot.
- The performance of a ranker can be assessed by drawing a piecewise linear curve in an ROC plot, known as an ROC curve. The curve starts in (0, 0), finishes in (1, 1), and is monotonically non-decreasing in both axes.
- A useful technique for organizing classifiers and visualizing their performance. Especially useful for domains with skewed class distribution and unequal classification error costs.
- It allows to create ROC curve and a complete sensitivity/specificity report. The ROC curve is a fundamental tool for diagnostic test evaluation.
- In a ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate for different cut-off points of a parameter.
- Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve is a measure of how well a parameter can distinguish between two diagnostic groups.
- Each point on an ROC curve connecting two segments corresponds to the true and false positive rates achieved on the same test set by the classifier obtained from the ranker by splitting the ranking between those two segments.
- An ROC curve is convex if the slopes are monotonically non-increasing when moving along the curve from (0, 0) to (1, 1). A concavity in an ROC curve, i.e., two or more adjacent segments with increasing slopes, indicates a locally worse than random ranking.
- **True Positive Rate (TPR)** is a synonym for recall and is therefore defined as follows:

$$\text{True Positive Rate TPR} = \frac{TP}{TP+FN}$$

- **False Positive Rate (FPR)** is defined as follows :

$$\text{False Positive Rate FPR} = \frac{FP}{FP+TN}$$

3.4.2 Supervised Learning : Regression

- A regression model which ensures that the difference between predicted and actual values is low can be considered as a good model.
- For example, a regression model could be used to predict the values of a data warehouse based on web-marketing, number of data entries, size and other factors.
- A regression task begins with a data set in which the target values are known. Regression analysis is a good choice when all of the predictor variables are continuous valued as well.
- Fig. 3.4.1 shows linear regression model.

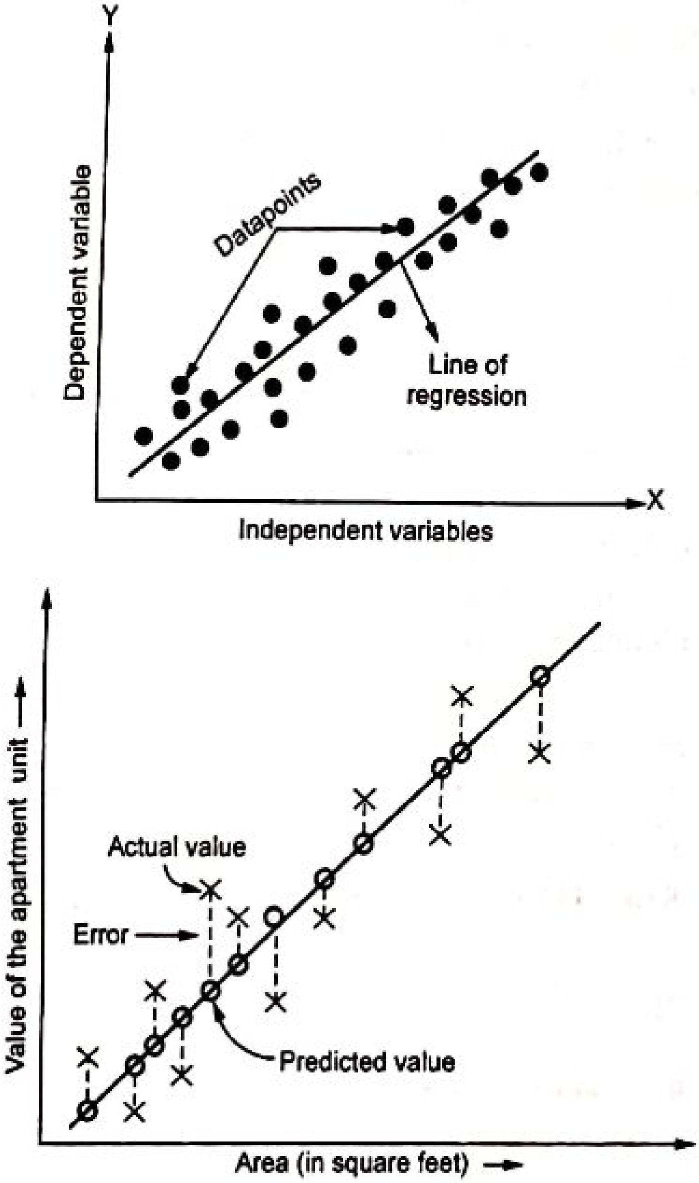


Fig. 3.4.1 Linear regression model

- If 'area' is the predictor variable (say x) and 'value' is the target variable (say y), the linear regression model can be represented in the form : $y = \alpha + \beta x$.

- In this equation
  1. y is the output variable. It is also called the target variable in machine learning, or the dependent variable in statistical modeling. It represents the continuous value that we are trying to predict.
  2. x is the input variable. In machine learning, x is referred to as the feature, while in statistics, it is called the independent variable. It represents the information given to us at any given time.
  3. β is the regression coefficient or scale factor.
- It assumes that there exists a linear relationship between a dependent variable and independent variable(s). The value of the dependent variable of a linear regression model is a continuous value i.e. real numbers.
- Linear regression is a statistical tool that determines how well a straight line fits a set of paired data. The straight line that best fits that data is called the **least squares** regression line.
- The distance between the actual value and predicted values is called **residual.**
- If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function.
- **R-squared** is a good measure to evaluate the model fitness. It is also known as the coefficient of determination. R-squared is the fraction by which the variance of the errors is less than the variance of the dependent variable.
- It is called R-squared because in a simple regression model it is just the square of the correlation between the dependent and independent variables, which is commonly denoted by "r".
- In a multiple regression model R-squared is determined by pairwise correlations among all the variables, including correlations of the independent variables with each other as well as with the dependent variable.

3.4.3 Unsupervised Learning : Clustering

- Clustering groups data points based on their similarities. Each group is called a cluster and contains data points with high similarity and low similarity with data points in other clusters.
- The objective of clustering is to segregate groups with similar traits and bundle them together into different clusters.
- Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point

in one cluster is to points in the neighboring clusters. This measure has a range of [– 1, 1].

- Silhouette coefficients near + 1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.
- Many clustering algorithms use distance measures to determine the similarity or dissimilarity between any pair of data points. A valid distance measure should be symmetric and obtains its minimum value (usually zero) in case of identical data points. By computing the distance or (dis) similarity between each pair of observations, a dissimilarity or distance matrix is obtained.

3.5 Improving Performance of a Model

- When we build random forest classifier we can tune the number of trees to build, the number of variables to choose for splitting etc.
- Similarly, when we build deep learning algorithm we can specify how many layers we would need, how many neurons we want in each layer, which activation function we want. Tuning parameter enhances model performance if we use the right type of parameters in an algorithm.
- One effective way to improve model performance is by tuning model parameter. Model parameter tuning is the process of adjusting the model fitting options.

3.6 Fill in the Blanks

Q.1 Structured representation of raw input data to the meaningful pattern is called a ______.

Q.2 The process of assigning a model, and fitting a specific model to a data set is called model ______.

Q.3 In bias-variance, when the value of 'k' is decreased, the model becomes simpler to fit and ______ increases.

Q.4 In bias-variance , When the value of 'k' is increased, the variance ______.

Q.5 Both underfitting and overfitting result in poor classification quality which is reflected by low classification ______ .

Q.6 Overfitting refers to a situation where the model has been designed in such a way that it emulates the ______ data too closely.

Q.7 A typical case of underfitting may occur when trying to represent a ______ data with a linear model as demonstrated by both cases of underfitting

Q.8 Lazy learning are also called ______ learning.

Q.9 The distance between the actual value and predicted values is called ________.

Q.10 Errors due to bias arise from simplifying assumptions made by the model whereas errors due to variance occur from over-aligning the model with the ________ data sets.

Q.11 One of the most popular algorithms for lazy learning is________.

Q.12 ________ is another measure of model performance which combines the precision and recall

Q.13 For any classification model, model ________ is the primary indicator of the goodness of the model

Q.14 In k-fold cross-validation technique, the data set is divided into k- completely separate random partitions called ________

Q.15 R-squared is a good measure to evaluate the model fitness. It is also known as the coefficient of ________.

Q.16 A high value of ________ is more desirable than a high value of accuracy.

Q.17 ________ is also another good measure to indicate a good balance of a model being excessively conservative or excessively aggressive.

3.7 Multiple Choice Questions

Q.1 A matrix containing correct and incorrect predictions in the form of TPs, FPs, FNs and TNs is known as _______.

a sparse matrix b confusion matrix

c zero matrix d all of these

Q.2 In ________ analysis, based on the purchase pattern available in the transactional data, the possibility of purchasing one product based on the purchase of another product is determined.

a market basket b cluster

c regression d decision tree

Q.3 Some of the popular classification models include _______.

a k-Nearest Neighbor b Naïve Bayes

c Decision Tree d All of these

Q.4 Some of the algorithms which adopt eager learning approach include _______.

a Decision Tree b Support Vector Machine

c Neural Network d All of these

Q.5 ______ occur when you predict an observation belongs to a class when in reality it does not.

a False negatives b False positives

c true positives d true negatives

Q.6 ______ occur when you predict an observation does not belong to a class when in fact it does.

a False negatives b False positives

c true positives d true negatives

Q.7 Overfitting can be avoided by ______.

a using re-sampling techniques like k-fold cross validation

b hold back of a validation data set

c remove the nodes which have little or no predictive power for the given machine learning problem.

d All of these

Q.8 Underfitting can be avoided by ______.

a using more training data

b using re-sampling techniques

c remove the nodes

d increasing features by effective feature selection

Answer Keys for Fill in the Blanks

1.	model	2.	training
3.	bias	4.	increases
5.	accuracy	6.	training
7.	non-linear	8.	non-parametric
9.	residual	10.	training
11.	k-nearest neighbor	12.	F-measure
13.	accuracy	14.	folds
15.	determination	16.	sensitivity
17.	Specificity		

Answer Keys for Multiple Choice Questions

1.	b	2.	a
3.	d	4.	d
5.	b	6.	a
7.	d	8.	a

Notes

4 Basics of Feature Engineering

Syllabus

Feature and Feature Engineering, Feature transformation : Construction and extraction, Feature subset selection : Issues in high-dimensional data, key drivers, measure and overall process.

Contents

4.1 Feature and Feature Engineering

- In machine learning, features are individual independent variables that act like input in your system. Feature is an attribute of a data set and used in a machine learning process. Selection of the subset of features which are meaningful for machine learning is a sub-area of feature engineering.
- The features in a data set are also called its dimensions. So a data set having 'n' features is called an n-dimensional data set.
- A good feature representation is central to achieving high performance in any machine learning task.
- Consider an example of text categorization. Assume that we need to train a model for classifying a given document as spam and not spam. If we represent a document as a bag of words, the feature space consists of a vocabulary of all unique words present in all the documents in the training set.
- For a collection of 100,000 to 1,000,000 documents, we can easily expect hundreds of thousands of features. If we further extend this document model to include all possible bigrams and trigrams, we could easily get over a million features.
- A feature tree is a tree such that each internal node is labelled with a feature, and each edge emanating from an internal node is labelled with a literal. The set of literals at a node is called a split. Each leaf of the tree represents a logical expression, which is the conjunction of literals encountered on the path from the root of the tree to the leaf. The extension of that conjunction is called the instance space segment associated with the leaf.
- Two features are redundant if they are highly correlated, regardless of whether they are correlated with the task or not.
- Feature engineering is the process of creating features (also called "attributes") that don't already exist in the dataset. This means that if your dataset already contains enough "useful" features, you don't necessarily need to engineer additional features.
- Feature engineering refers to the process of translating a data set into features such that these features are able to represent the data set more effectively and result in a better learning performance.
- If feature engineering is performed properly, it helps to improve the power of prediction of machine learning algorithms by creating the features using the raw data that facilitate the machine learning process.
- Elements of feature engineering is **feature transformation and feature subset selection.**

4.2 Feature Transformation

- **Feature transformation** transforms the data, structured or unstructured, into a new set of features which can represent the underlying problem which machine learning is trying to solve.
- There are two distinct goals of feature transformation :
  1. Achieving best reconstruction of the original features in the data set
  2. Achieving highest efficiency in the learning task
- There are two variants of feature transformation :
  1. Feature construction
  2. Feature extraction

4.2.1 Feature Construction

- Feature construction involves transforming a given set of input features to generate a new set of more powerful features which can then used for prediction.
- Feature construction methods may be applied to pursue two distinct goals : Reducing data dimensionality and improving prediction performance.
- **Steps :**
  1. Start with an initial feature space F_0 .
  2. Transform F_0 to construct a new feature space F_N .
  3. Select a subset of features F_i from F_N .
  4. If some terminating criteria is achieved : Go back to step 3 otherwise set $F_T = F_i$.
  5. F_T is the newly constructed feature space.
- Feature construction process discovers missing information about the relationships between features and augments the feature space by creating additional features.
- Hence, if there are 'n' features or dimensions in a data set, after feature construction 'm' more features or dimensions may get added. So at the end, the data set will become 'n + m' dimensional.
- The task of constructing appropriate features is often highly application specific and labour intensive. Thus building auto-mated feature construction methods that require minimal user effort is challenging. In particular we want methods that :
  1. Generate a set of features that help improve prediction accuracy.
  2. Are computationally efficient.
  3. Are generalizable to different classifiers.
  4. Allow for easy addition of domain knowledge.

- Genetic programming is an evolutionary algorithm-based technique that starts with a population of individuals, evaluates them based on some fitness function and constructs a new population by applying a set of mutation and crossover operators on high scoring individuals and eliminating the low scoring ones.
- In the feature construction paradigm, genetic programming is used to derive a new feature set from the original one. Individuals are often tree like representations of features, the fitness function is usually based on the prediction performance of the classifier trained on these features while the operators can be applications specific.
- The method essentially performs a search in the new feature space and helps generate a high performing subset of features. The newly generated features may often be more comprehensible and intuitive than the original feature set, which makes GP-related methods well-suited for such tasks.
- In decision trees, the model explicitly selects features that are highly correlated with the label. In particular, by limiting the depth of the decision tree, one can at least hope that the model will be able to throw away irrelevant features.

4.2.2 Feature Extraction

- Feature extraction is a process that extracts a set of new features from the original features through some functional mapping. Feature extraction method creates a new feature set.
- Feature extraction increases the accuracy of learned models by extracting features from the input data. This phase of the general framework reduces the dimensionality of data by removing the redundant data.
- A characteristic of these large data sets is a large number of variables that require a lot of computing resources to process.
- Feature extraction is the name for methods that select and/or combine variables into features, effectively reducing the amount of data that must be processed, while still accurately and completely describing the original data set.
- The process of feature extraction is useful when you need to reduce the number of resources needed for processing without losing important or relevant information.
- Feature extraction can also reduce the amount of redundant data for a given analysis. Also, the reduction of the data and the machine's efforts in building variable combinations (features) facilitate the speed of learning and generalization steps in the machine learning process.

Principal Component Analysis :

- Principal Component Analysis (PCA) is to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, retains most of the sample's information and useful for the compression and classification of data.
- In PCA, it is assumed that the information is carried in the variance of the features, that is, the higher the variation in a feature, the more information that feature carries.
- Hence, PCA employs a linear transformation that is based on preserving the most variance in the data using the least number of dimensions.
- The objective of PCA is to make the transformation in such a way that,
  1. The new features are distinct, i.e., covariance between the new features, i.e. the principal components is 0.
  2. The principal components are generated in order of the variability in the data that it captures. Hence, the first principal component should capture the maximum variability, the second principal component should capture the next highest variability etc.
  3. The sum of variance of the new features or the principal components should be equal to the sum of variance of the original features.

4.3 Feature Subset Selection

- Feature selection is critical pre-processing activity in any machine learning project. The subset of features is expected to give better results than the full set.
- Reasons to use feature selection are :
  a) It enables the machine learning algorithm to train faster.
  b) It reduces the complexity of a model and makes it easier to interpret.
  c) It improves the accuracy of a model if the right subset is chosen.
  d) It reduces overfitting.

4.3.1 Issues in High - Dimensional Data

- Feature selection approach solves the dimensionality problem by removing irrelevant and redundant features.
- High - dimensional' refers to the high number of variables or attributes or features present in certain data sets.
- A feature selection algorithm can be measured from both the efficiency and effectiveness points. The efficiency composes of the time required to find a subset of features, whereas effectiveness is related to the accuracy of the subset of

features finally selected. The objective of feature selections are increasing the prediction performance of the predictors, providing correct and fast result.

- In high dimensional data like genes data or DNA data contains large number of attributes which affect the computational cost and decrease the learning accuracy. The increase of data size in terms of number of instances and number of features becomes a great challenge for the feature selection algorithms.
- The objective of feature selection is as follows :
  1. Having faster and more cost-effective learning model
  2. Improving the efficiency of the learning model
  3. Having a better understanding of the underlying model that generated the data

4.3.2 Key Drivers

- Key drivers of feature selection is **feature relevance and redundancy.**

1. Feature relevance

- Each of the predictor variables, is expected to contribute information to decide the value of the class label. In case a variable is not contributing any information, it is said to be **irrelevant**.
- In case the information contribution for prediction is very little, the variable is said to be **weakly relevant**. Remaining variables, which make a significant contribution to the prediction task are said to be strongly **relevant variables.**
- Certain variables do not contribute any useful information for deciding the similarity of dissimilarity of data instances. These variables are marked as irrelevant variables in the context of the unsupervised machine learning task.

2. Redundancy

- All features having potential redundancy are candidates for rejection in the final feature subset.

4.3.3 Measures of Feature Relevance and Redundancy

- The mutual information measures the amount of information contained in a variable or a group of variables, in order to predict the dependent one.
- **Measures of Feature redundancy** : There are multiple measures of similarity of information.

Contribution :

1. Correlation-based measures
2. Distance-based measures, and
3. Other coefficient-based measure

- Correlation is a measure of linear dependency between two random variables.
- Correlation coefficients provide a numerical measurement of the association between two variables. They can be used to determine the similarly between two objects when they are merged into a cluster.
- Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, age and blood pressure.
- **Pearson's correlation** coefficient is a measure related to the strength and direction of a linear relationship. We calculate this metric for the vectors x and y in the following way :

$$\text{CORR}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The Pearson's correlation can take a range of values from –1 to +1.

Distance - based similarity measure :

- The most common distance measure is the Euclidean distance, which, between two features F_1 and F_2 are calculated as :

$$d(F_1, F_2) = \sqrt{\sum_{i=1}^{n}(F_{1_i} - F_{2_i})^2}$$

where F_1 and F_2 are features of an n-dimensional data set.

- A more generalized form of the Euclidean distance is the Minkowski distance, measured as

$$d(F_1, F_2) = \sqrt{\sum_{i=1}^{n}(F_{1_i} - F_{2_i})^r}$$

4.3.4 Overall Feature Selection Process

- Typical feature selection process consists of four steps :
  1. Generation of possible subsets
  2. Subset evaluation
  3. Stop searching based on some stopping criterion
  4. Validation of the result

- Fig. 4.3.1 shows feature selection process.

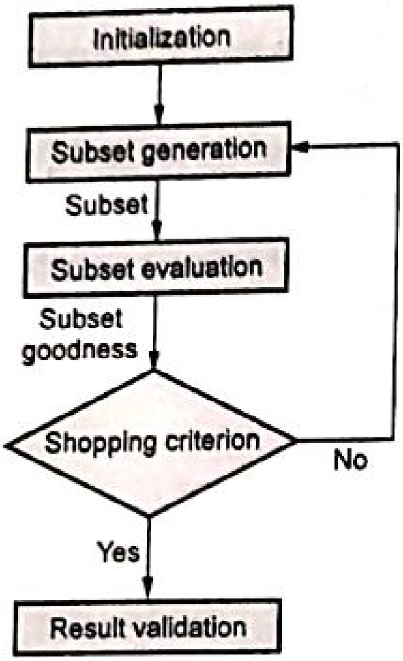


Fig. 4.3.1 Feature selection process

4.3.5 Feature Selection Approaches

- There are three types of approach for feature selection :
  1. Filter approach
  2. Wrapper approach
  3. Embedded approach

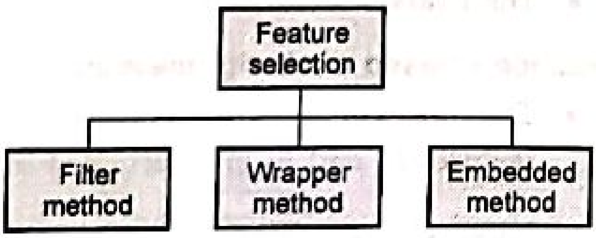


Fig. 4.3.2

1. Filter method :

- Filter methods are generally used as a preprocessing step. The selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable.
- Fig. 4.3.3 shows filter method.
- The filter feature selection methods make use of statistical techniques to predict the relationship between each

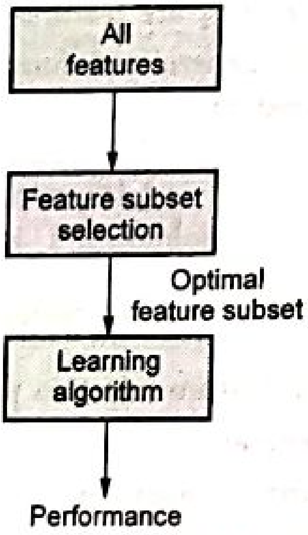


Fig. 4.3.3 Filter method

independent input variable and the output (target) variable. Which assigns scores for each feature.

- Correlation based Feature Selection (CFS) is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function.
- Example : coorelation, chi-square test, ANOVA, information gain etc.

2. Wrapper Methods :

- In Wrapper Methods, the Learner is considered a black-box. Interface of the black-box is used to score subsets of variables according to the predictive power of the learner when using the subsets.
- Fig. 4.3.4 shows wrapper method.
- The feature selection algorithm searches for a good feature subset using the induction algorithm itself as a part of the evaluation function.
- Wrapper methods are recursive feature elimination, sequential feature selection algorithms and genetic algorithms.

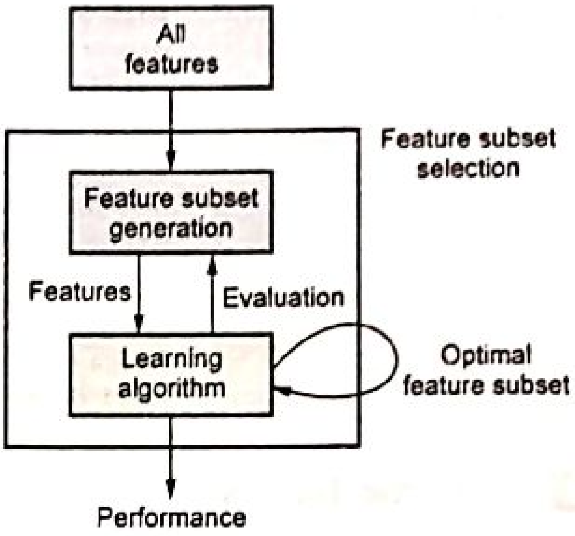


Fig. 4.3.4 Wrapper method

3. Embedded Methods :

- Embedded methods, are quite similar to wrapper methods since they are also used to optimize the objective function or performance of a learning algorithm or model.
- It's implemented by algorithms that have their own feature selection methods in them.
- A learning algorithm takes advantage of its own variable selection process and performs feature selection and classification/regression at the same time.
- The most Common embedded technique are the tree algorithm's like Random Forest, LASSO with the L1 penalty and Ridge with the L2 penalty for constructing a linear model.
- Tree algorithms select a feature in each recursive step of the tree growth process and divide the sample set into smaller subsets. The more child nodes in a subset are in the same class, the more informative the features are.

4.3.6 Difference between Filter, Wrapper and Embedded Method

Filter methods	Wrapper methods	Embedded methods
Generic set of methods which do not incorporate a specific machine learning algorithm.	Evaluates on a specific machine learning algorithm to find optimal features.	Embeds (fix) features during model building process. Feature selection is done by observing each iteration of model training phase.
Much faster compared to Wrapper methods in terms of time complexity.	High computation time for a dataset with many features.	Sits between Filter methods and Wrapper methods in terms of time complexity.
Less prone to over-fitting.	High chances of over-fitting because it involves training of machine learning models with different combination of features.	Generally used to reduce over-fitting by penalizing the coefficients of a model being too large.
Examples - Correlation, Chi-square test, ANOVA, Information gain, etc.	Examples - Forward Selection, Backward elimination, Stepwise Selection, etc.	Examples - LASSO, Elastic Net, Ridge Regression, etc.

4.4 Fill in the Blanks

Q.1 A feature is an ______ of a data set that is used in a machine learning process.

Q.2 The features in a data set are also called its ________ .

Q.3 Features are extracted from ________ .

Q.4 PCA works based on a process called _______ decomposition of a covariance matrix of a data set.

Q.5 Correlation is a measure of _______ dependency between two random variables.

Q.6 In the wrapper approach, identification of best feature subset is done using the induction algorithm as a _______ .

4.5 Multiple Choice Questions

Q.1 PCA stands for _______.

a Principal Component Analysis b Principal Component Approach

c Principal Correlation Analysis d Process Component Analysis

Q.2 Which of the following are the feature selection methods ?

a Filter approach b Wrapper approach

c Embedded approach d All of these

Q.3 PCA is a technique for ______.

a Feature extraction b Feature construction

c Feature selection d None of these

Q.4 Cosine similarity is most popularly used in ______.

a image classification b text classification

c feature selection d all of these

Q.5 Coorelation, chi-square test, ANOVA are the example of ______.

a embedded method b wrapper method

c filter method d wrapper and filter method

Q.6 Some popular feature extraction algorithms used in machine learning :

a Principal Component Analysis b Singular Value Decomposition

c Linear Discriminant Analysis d All of these

Answer Keys for Fill in the Blanks

Q.1	attribute	Q.2	dimensions	Q.3	raw data
Q.4	eigenvalue	Q.5	linear	Q.6	black box

Answer Keys for Multiple Choice Questions

Q.1	a	Q.2	d	Q.3	a
Q.4	d	Q.5	c	Q.6	d

Notes

5 Overview of Probability

Syllabus

Statistical tools in Machine Learning, Concepts of probability, Random variables, Discrete distributions, Continuous distributions, Multiple random variables, Central limit theorem, Sampling distributions, Hypothesis testing, Monte Carlo Approximation.

Contents

5.1 Statistical Tools in Machine Learning

- In machine learning, we train the system by using a limited data set called 'training data' and based on the confidence level of the training data we expect the machine learning algorithm to depict the behaviour of the larger set of actual data.
- Probability theory provides a mathematical foundation for quantifying uncertainty of the knowledge.
- ML is focused on making predictions as accurate as possible, while traditional statistical models are aimed at inferring relationships between variables.
- We make observations using the sensors in the world. Based on the observations, we intend to make decisions. Given the same observations, the decision should be the same. However, the world changes, observations change, our sensors change, the output should not change.
- We build models for predictions; can we trust them? Are they certain? Many applications of machine learning depend on good estimation of the uncertainty:

  a) Forecasting

  b) Decision making

  c) Learning from limited, noisy, and missing data

  d) Learning complex personalised models

  e) Data compression

  f) Automating scientific modelling, discovery, and experiment design

5.2 Concepts of Probability

- A signal is called random if its occurrence can not be predicted. Such signal can not be represented by any mathematical equation.
- The random signals are represented collectively by a random variable. The random variable takes its value from the specified set of values. But which particular value will be taken at particular time is not known.
- The random variables are analyzed statistically with the help of probability, probability density functions and statistical averages such as mean, variance etc.

5.2.1 Experiment

Definition : It is the process which is conducted to get some results.

- An experiment is also called trial. For example, throw of a coin is an experiment or trial.

- The trial or an experiment has outcomes. For example throwing a coin has two outcomes head *(H)* or tail *(T)*.
- Outcomes of an experiment are call equally likely if all of them have equal chance of occurring. For example, head and tail are equally likely.

5.2.2 Sample Space (S)

> ***Definition : A set of all possible outcomes of an experiment is called sample space of that experiment.***

- **Examples :** If a coin is thrown, outcomes are head *(H)* and tail *(T)*. Hence sample space will be,

$$S = \{H, T\}$$

If three coins are tossed simultaneously, then each experiment will have an outcome which will be combination of *H* or *T*. The sample will be as follows :

$$S = \{H_1 H_2 H_3, \quad H_1 H_2 T_3, \quad H_1 T_2 H_3, \quad T_1 H_2 H_3, \quad H_1 T_2 T_3, \quad T_1 H_2 T_3, \quad T_1 T_2 H_3, \quad T_1 T_2 T_3 \}$$

5.2.3 Event

- **Definition :** The expected subset of the sample space or happening is called an event.
- **Example :** Consider an experiment of throwing a dice.

  Then sample space will be,

$$S = \{1, 2, 3, 4, 5, 6\}$$

An event *'A'* for setting number greater than 4 will be,

$$A = \{4, 5, 6\}$$

- **Elementary event :** Event contains only one outcome.
- **Null event :** Event not possible.
- **Contain event :** Event contains all outcomes of sample space.
- **Independent event :** If happening of 'A' has nothing to do with happening of 'B', then A and B are independent.
- **Dependant event :** If outcome of one event is affected by other, then they are called dependant events.

5.2.4 Definition of Probability

Relative Frequency : For event 'A' relative frequency is defined as,

$$\text{Relative frequency} = \frac{\textit{Number of times an event occurs } (N_A)}{\textit{Total number of trials } (N)} = \frac{N_A}{N}$$

As number of trials approach infinity, relative frequency is called probability.

Probability of event 'A' is defined as the ratio of number of possible favourable outcomes to total number of outcomes. i.e.,

$$\text{Probability, } P(A) = \lim_{N \to \infty} \frac{N_A}{N} \quad \text{... (5.2.1)}$$

$$= \frac{\textit{Number of possible favourable outcomes}}{\textit{Total number of outcomes}} \quad \text{... (5.2.2)}$$

Example : Probability of getting head in tossing a coin is,

$$P(A) = \frac{1(Head)}{2(Head+Tail)} = 0.5$$

Here favourable outcome is only one, i.e. head and total number of outcomes are two, i.e. head and tail.

Permutations and Combinations

$$\text{Combination of '}n\text{' taken '}r\text{' at a time, } n_{C_r} = \frac{n!}{(n-r)!r!} \quad \text{... (5.2.3)}$$

$$\text{Permutations of '}n\text{' taken '}r\text{' at a time, } n_{P_r} = \frac{n!}{(n-r)!} \quad \text{... (5.2.4)}$$

Examples for Understanding

Example 5.2.1 *If 3 of 20 tubes are defective and 4 of them are randomly chosen for inspection. What is the probability that only one of the defective tubes will be included ?*

Solution : Four tubes can be selected out of 20 in 20_{C_4} ways.

$$\text{Possible ways} = 20_{C_4}$$

$$\text{We know that } n_{C_r} = \frac{n!}{(n-r)!r!}$$

$$N = \frac{20!}{(20-4)!\,4!} = \frac{20!}{16!4!} = \frac{20 \times 19 \times 18 \times 17 \times 16!}{16! \times 4 \times 3 \times 2 \times 1} = 4845$$

Now there are three defective tubes. Now since only one defective tube should be included in set of four, this tube can be chosen in 3_{C_1} ways.

Thus in the set of four defective tubes one tube should be defective and three tubes should be non defective. That is, those three tubes can be selected in 17_{C_3} ways.

$$\therefore \quad N_A = 3_{C_1} \times 17_{C_3} = \frac{3!}{(3-1)!1!} \times \frac{17!}{(17-3)!3!}$$

$$= \frac{3\times 2!}{2!\times 1} \times \frac{17\times 16\times 15\times 14!}{14!\times 3\times 2\times 1} = 2040$$

$$\therefore \quad P(A) = \frac{\textit{Number of favourable ways}\,(NA)}{\textit{Total possible ways}\,(N)} = \frac{2040}{4845} = 0.42$$

Examples with Solutions

Example 5.2.2 *From a well shuffled pack of cards three cards are drawn at random. Find the probability that they form a King, Queen, Jack combination.*

Solution : Three cards can be drawn in 52_{C_3} ways. i.e.

$$N = 52_{C_3} = \frac{52!}{(52-3)!3!} = \frac{52\times 51\times 50\times 49!}{49!\times 3\times 2\times 1} = 22100$$

There are 4 Kings, 4 Queens and 4 Jacks in total. Hence a King, Queen and Jack can be chosen each in 4_{C_1}ways.

$$\therefore \quad N_A = 4_{C_1}\times 4_{C_1}\times 4_{C_1} = 64$$

$$\therefore \text{ Probability} = \frac{N_A}{N} = \frac{64}{22100} = 2.89 \times 10^{-3}$$

Example 5.2.3 *A room contains three sockets for bulbs. From the collection of 8 bulbs out of which 4 are defective, 3 bulbs are selected at random and put in the sockets. Find the probability that the room is lit.*

Solution : Three bulbs can be selected out of eight bulbs in 8_{C_3} ways. The room will lit if one, two or three bulbs are non defective. Therefore it is better to calculate the probability that room will not lit. That is all three lamps are defective. These three defective bulbs can be selected out of total four defective bulbs by using 4_{C_3} ways.

$$\text{i.e.} \quad N_A = 4_{C_3} = \frac{4!}{(4-3)!3!} = \frac{4\times 3!}{1!\times 3!} = 4$$

and $$N = {}^8C_3 = \frac{8!}{(8-3)!3!} = \frac{8 \times 7 \times 6 \times 5!}{5! \times 3 \times 2 \times 1} = 56$$

$\therefore$ Probability that room will not lit (dark) will be $P(\textit{room dark}) = \frac{4}{56} = 1/14.$

$\therefore$ P(room lits) + P(room dark) = 1

$$\therefore P(\text{room lits}) = 1 - P(\text{room dark}) = 1 - \frac{1}{14} = 0.928$$

Example 5.2.4 *A box contains 3 white, 4 red and 5 black balls. A ball is drawn at random. Find the probability that it is :*

i) Red ii) Not black iii) Black or white

Solution : There are total 12 balls. Hence one ball can be drawn from 12 balls in, ${}^{12}C_1$ ways. i.e.,

$$N = {}^{12}C_1 = 12$$

i) P(red)

Out of 4 red balls one ball can be drawn in ${}^4C_1$ ways. i.e.,

$$N(\text{red}) = {}^4C_1 = 4$$

$$\therefore \quad P(\text{red}) = \frac{N(\text{red})}{N} = \frac{4}{12} = \frac{1}{3}$$

ii) P (not black)

Then probability that ball will not be black is same as probability that it will be white or red. Hence,

$$N(\text{red}) = 4$$

$$N(\text{white}) = {}^3C_1 = 3$$

$$\therefore P(\text{not black}) = P(\text{red}) + P(\text{white}) = \frac{N(\text{red})}{N} + \frac{N(\text{white})}{N} = \frac{4}{12} + \frac{3}{12} = \frac{7}{12}$$

iii) P(black or white)

Here $N(\text{black}) = {}^5C_1 = 5$

and $N(\text{white}) = {}^3C_1 = 3$

$$\therefore \quad P(\text{black or white}) = P(\text{black}) + P(\text{white})$$

$$= \frac{N(\text{black})}{N} + \frac{N(\text{white})}{N} = \frac{5}{12} + \frac{3}{12} = \frac{8}{12} = \frac{2}{3}$$

Example 5.2.5 *Two cards are drawn from a 52 card deck successively without replacing the first :*

i) Given the first one is heart, what is the probability that second is also a heart ?

ii) What is the probability that both cards will be hearts ?

Solution : Two cards can be drawn in 52_{C_2} ways.

$$\therefore \quad N = {}^{52}C_2 = \frac{52!}{(52-2)!2!} = \frac{52\times 51\times 50!}{50!\times 2\times 1} = 1326 \text{ ways}$$

i) Probability that second is also heart

Probability that first is heart $= \frac{{}^{13}C_1}{N} = \frac{13}{1326}$

Now 12 heart cards are remaining in the pack.

Probability that second is heart $= \frac{{}^{12}C_1}{N} = \frac{12}{1326}$

Probability that second is also heart $= \frac{13}{1326}\times\frac{12}{1326} = 88.723\times 10^{-6}$

ii) Probability that both cards are hearts

$$N_A = {}^{13}C_2 = \frac{13!}{(13-2)!2!} = \frac{13\times 12\times 11!}{11!\times 2\times 1} = 78$$

Probability that both cards are hearts $= \frac{N_A}{N} = \frac{78}{1326} = 0.059$

Example 5.2.6 *A box contains five white balls, 6 blue balls and three yellow balls. A ball is drawn at random. Find probability that :*

i) ball is not yellow *ii) ball is either white or yellow*

In the second random experiment if two balls are drawn in succession, then what is the probability that the second ball is blue if the first ball is white.

Solution : i) Probability that ball is not yellow

There are total 5 + 6 + 3 = 14 balls. One ball can be drawn in total $N = {}^{14}C_1 = 14$ ways. The ball is not yellow means it can be white or blue. Hence

$$N\,(\text{white}) = {}^{5}C_1 = 5 \text{ ways}$$

$$N\,(\text{blue}) = {}^{6}C_1 = 6 \text{ ways}$$

P (Not yellow) = P (white) + P (blue) $= \frac{5}{14}+\frac{6}{14} = \frac{11}{14}$

ii) Probability of white or yellow ball

$$N\text{ (white)} = {}^{5}C_1 = 5 \text{ ways}$$

$$N\text{ (yellow)} = {}^{3}C_1 = 3 \text{ ways}$$

$$P\text{ (white + yellow)} = P\text{ (white)} + P\text{ (yellow)} = \frac{5}{14}+\frac{3}{14} = \frac{8}{14} = \frac{4}{7}$$

iii) Probability that second is blue if first ball is white

First ball is drawn in $N_1 = {}^{14}C_1 = 14$ ways.

After first ball is draw, 13 balls are left. Hence second ball is drawn in $N_2 = {}^{13}C_1 = 13$ ways.

$$N\text{ (white)} = {}^{5}C_1 = 5 \text{ ways}$$

$$N\text{ (blue)} = {}^{6}C_1 = 6 \text{ ways}$$

$$P\text{ (1}^{st}\text{ white and 2}^{nd}\text{ blue)} = \frac{5}{14}\times\frac{6}{13} = \frac{30}{182} = 0.165.$$

5.2.5 Axioms (Properties) of Probability

The outcomes of the trial are said to be *mutually exclusive,* if the occurrence of one of them precludes the occurrence of all other outcomes. For example in tossing a coin, events Head and Tail are mutually exclusive. In throw of a die the occurrence of number '4' will automatically exclude the occurrence of numbers 1, 2, 3, 5 and 6.

- If an event contains all the outcomes then it is called certain event. Then probability of this event is unity. i.e.,

$$P(A) = P(S) = 1 \quad \text{... (5.2.5)}$$

- We also know that probability of any event is always less than or equal to '1' and non negative. i.e.,

$$0 \leq P(A) \leq 1 \quad \text{... (5.2.6)}$$

- Just now we have defined mutually exclusive events. The occurrence of such events precludes over each other. Then if $A + B$ is the union of two mutually exclusive events, then

$$P(A + B) = P(A) + P(B) \quad \text{... (5.2.7)}$$

which states that probability of union of mutually exclusive events is equal to sum of their independent probabilities.

Property 1 : $\boxed{P(\overline{A}) = 1 - P(A)}$ ... (5.2.8)

Here $\overline{A}$ denotes the complement of event A.

Proof : Let the sample space be the union of two mutually exclusive events A and $\overline{A}$.

i.e. $$S = A + \overline{A}$$

By taking probability of both sides,

$$P(S) = P(A) + P(\overline{A})$$

$\therefore$ $$1 = P(A) + P(\overline{A}), \quad \text{... Since } P(S) = 1 \text{ by equation (5.2.5)}$$

or $$P(\overline{A}) = 1 - P(A)$$

Property 2 : If $A_1, A_2, \ldots A_M$ are mutually exclusive events,

then $$\boxed{P(A_1) + P(A_2) + \ldots\ldots P(A_M) = 1} \quad \text{... (5.2.9)}$$

Proof : The mutually exclusive events satisfy following relation,

$$A_1 + A_2 + \ldots\ldots + A_M = S$$

$\therefore$ $$P(A_1 + A_2 + \ldots\ldots A_M) = P(S)$$

$\therefore$ $$P(A_1 + A_2 + \ldots\ldots A_M) = 1, \quad \text{... } P(S) = 1 \textit{ for certain event.}$$

$$P(A_1) + P(A_2) + \ldots\ldots + P(A_M) = 1 \quad \text{... By equation (5.2.7)}$$

If all events $A_1, A_2, \ldots\ldots A_M$ have same possibility of occurrence (equally likely),

then $$P(A_1) = P(A_2) = P(A_3) = \ldots\ldots = P(A_M) = \frac{1}{M}$$

Property 3 : If events A and B are not mutually exclusive events, then the probability of the union of A or B is given as,

$$\boxed{P(A+B) = P(A) + P(B) - P(AB)} \quad \text{... (5.2.10)}$$

Here $P(AB)$ is called the probability of events A and B both occuring simultaneously. Such event is called *joint event* of A and B, and the probability $P(AB)$ is called *joint probability* it is defined as,

$$P(AB) = \lim_{N \to \infty} \frac{N_{AB}}{N} \quad \text{... (5.2.11)}$$

If events A and B are mutually exclusive, then the joint probability $P(AB) = 0$.

5.2.6 Conditional Probability

Definition : Probability of B given that A has occurred is represented by $P(B/A)$. Alternately $P(A/B)$ represents probability of A given that B has occurred. $P(B/A)$ and $P(A/B)$ are called conditional probabilities.

$$\boxed{P(B/A) = \frac{P(AB)}{P(A)}} \quad \text{and} \quad \boxed{P(A/B) = \frac{P(AB)}{P(B)}} \qquad \text{... (5.2.12)}$$

Here $P(AB)$ is the joint probability of A and B.

The joint probability has commutative property i.e.,

$$P(AB) = P(BA) \qquad \text{... (5.2.13)}$$

5.2.7 Independent Events Probability

> ***Definition :** If A and B are the two events possible from an experiment, and possibility of occurrence of B simply does not depend on occurrence of event A then these events are called statistically independent events.*

$$P(B/A) = \frac{P(AB)}{P(A)} \qquad \text{... By equation 5.2.12} \qquad \text{... (5.2.14)}$$

This gives probability of B given that event A has occurred. If the occurrence of B does not depend on event A, probability of event B is same as conditional probability $P(B/A)$. i.e.,

$$P(B/A) = P(B) \qquad \text{... (5.2.15)}$$

With this result equation (5.2.14) becomes,

$$P(AB) = P(A)\, P(B) \qquad \text{... (5.2.16)}$$

Similarly since events A and B are statistically independent, probability of event A is same as conditional probability of A given that event B has occurred.

$$\therefore \quad P(A/B) = P(A) \qquad \text{... (5.2.17)}$$

With above result equation (5.2.12) can be written as, $P(AB) = P(A) \cdot P(B)$, which is same as equation (5.2.16)

Example 5.2.7 *Find out the number of permutations of fair letters A,B,C and D taken two at a time.*

Solution : Here n = 4 and k = 2.

Hence

$$^{n}P_{k} = \frac{n!}{(n-k)!}$$

$$\therefore \quad ^{4}P_{2} = \frac{4!}{(4-2)!} = \frac{4!}{2!} = \frac{4\times3\times2!}{2!} = 12$$

Example 5.2.8 *Consider an experiment of drawing two cards at random from a bag containing four cards marked with the integers 1 through 4. Find the sample space of the experiment if the first card is replaced before the second is drawn.*

Solution : The sample space will contain 16 ordered pairs

(i, j) $1 \le i \le 4$ and $1 \le j \le 4$. i.e.

$$S = \begin{Bmatrix} 1,1 & 1,2 & 1,3 & 1,4 \\ 2,1 & 2,2 & 2,3 & 2,4 \\ 3,1 & 3,2 & 3,3 & 3,4 \\ 4,1 & 4,2 & 4,3 & 4,4 \end{Bmatrix}$$

Example 5.2.9 *In a competitive examination 30 candidates are to be selected In all 600 candidates appear in a written test and 100 will be called for interview. What is the probability that a person will be called for the intrview ? Determine the probability of a person getting selected, if he has been called for interview*

Solution : Let event A be the person called for an interview and event B be the person selected.

Hence $$P(A) = \frac{\text{Called for interview}}{\text{Total candidates}} = \frac{100}{600} = \frac{1}{6}$$

and $$P(B/A) = \frac{\text{Selected candidates}}{\text{Called for interview}} = \frac{30}{100} = \frac{3}{10}$$

Example 5.2.10 *If $P(A) = \frac{1}{3}$, $P(B) = \frac{3}{4}$ and $P(A \cup B) = \frac{11}{12}$. Then find P(A/B).*

Solution : Here

$$P(A \cup B) = P(A) + P(B) \text{ form given data}$$

i.e. $$= \frac{1}{3} + \frac{3}{4} = \frac{11}{12} \text{ which is given.}$$

Hence A and B area mutually exclusive events.

For such events $P(AB) = P(A \cap B) = 0$ and

$$P(AB) = P(B) \cdot P(A/B)$$

$$\therefore \quad 0 = \frac{3}{4} \cdot P(A/B)$$

Hence $P(A/B) = 0$

Example 5.2.11 *If A and B are two independent events, where* $P(A) = \frac{1}{4}$, $P(B) = \frac{2}{3}$, *find* $P(A \cup B)$.

Solution : Hence $P(A + B)$ or $P(A \cup B)$ is given as,

$$P(A \cup B) = P(A) + P(B) - P(AB)$$

For independent events $P(AB) = P(A) \cdot P(B)$, i.e.;

$$P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$$

$$= \frac{1}{4} + \frac{2}{3} - \frac{1}{4} \times \frac{2}{3} = \frac{3}{4}$$

Example 5.2.12 *If A and B are two events such that* $P(A) = 0.3$, $P(B) = 0.4$, $P(A \cap B) = 0.2$ *find :*

i) $P(A \cup B)$ *ii)* $P(\overline{A}/B)$ *iii)* $P(\overline{A}/\overline{B})$ *iv)* $P(\overline{A} \cup \overline{B})$.

Solution : Here note that $P(A+B) = P(A \cup B)$ and $P(AB) = P(A \cap B)$

i) $P(A \cup B)$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.3 + 0.4 - 0.2 = 0.5$$

ii) $P(\overline{A}/B)$

$$P(A \cup B) = P(A) + P(\overline{A} \cap B), \text{ since } A \text{ and } \overline{A} \cap B \text{ are disjoint events}$$

$$\therefore \quad P(\overline{A} \cap B) = P(A \cup B) - P(A) = 0.5 - 0.3 = 0.2$$

$$\therefore \quad P(\overline{A}/B) = \frac{P(\overline{A} \cap B)}{P(B)} \quad \text{since } P(A|B) = \frac{P(AB)}{P(B)} = \frac{0.2}{0.4} = 0.5$$

iii) $P(\overline{A}/\overline{B})$

$$P(\overline{A} \cup \overline{B}) = P(\overline{B}) + P(\overline{A} \cap B) \quad \text{since } \overline{B} \text{ and } \overline{A} \cap B \text{ are disjoint events}$$

$$= 1 - P(B) + P(\overline{A} \cap B), \text{ since } P(\overline{B}) = 1 - P(B) = 1 - 0.4 + 0.2 = 0.8$$

$$P(\overline{A} \cup \overline{B}) = P(\overline{A}) + P(\overline{B}) - P(\overline{A} \cap \overline{B})$$

$$\therefore \quad P(\overline{A} \cup \overline{B}) = 1 - P(A) + 1 - P(B) - P(\overline{A} \cap \overline{B})$$

$$\therefore \quad 0.8 = 1 - 0.3 + 1 - 0.4 - P(\overline{A} \cap \overline{B})$$

$$\therefore \quad P(\overline{A} \cap \overline{B}) = 0.5$$

$$\therefore \quad P(\overline{A}/\overline{B}) = \frac{P(\overline{A} \cap \overline{B})}{P(\overline{B})} \quad \text{since } P(A/B) = \frac{P(AB)}{P(B)}$$

$$= \frac{0.5}{1 - 0.4}, \text{ since } P(\overline{B}) = 1 - P(B) = 1 - 0.4 = 0.8333$$

iv) $P(\overline{A} \cup \overline{B})$

$$P(\overline{A} \cup \overline{B}) = 0.8 \text{ [as obtained in part (iii)]}$$

Example 5.2.13 *In the experiment of rolling six face dice find the probability of occurrence of 4 if it is known that even face has appeared.*

Solution : Let event 'A' denote occurrance of even face.

Let event 'B' denote occurrance of 4.

Then AB denote occurrance of 4 with even face.

Since AB can occurs only once out of six possible out comes,

$$P(AB) = \frac{1}{6}$$

Total possible out comes are 3 i.e. 2, 4, 6 out of 6 outcomes. Hence,

$$P(A) = \frac{3}{6} = \frac{1}{2}$$

The probability of occurrance of 4, known that even face has appeared is $P(B/A)$. It is given as,

$$P(B/A) = \frac{P(AB)}{P(A)} = \frac{1/6}{1/2} = \frac{1}{3}$$

Example 5.2.14 *Each letter of the word ATTRACT is written on a separate card. The cards are then thoroughly shuffled and four of them are drawn in succession. What is the probability of getting result as TACT ?*

Solution : Alphabet of ATTRACT will be written on 7 separate cards.

- Out of 7 cards the card drawn should be T.
- Out of remaining 6 cards the card drawn should be A.
- Out of remaining 5 cards the card drawn should be C.
- Out of remaining 4 cards the card drawn should be T.

Following table illustrates calculation of probabilities of above events :

	N	N(alphabet)	P(alphabet)
T	Initially there are all 7 cards. One card can be drawn from 7 cards in ${}^7C_1 = 7$ ways. $\therefore N = 7$	First alphabet should be 'T'. There are three cards written with alphabet 'T'. Hence one card can be drawn from 3 'T' cards in ${}^3C_1 = 3$ ways. Hence $N(T) = 3$.	$P(T) = \frac{N(T)}{N} = \frac{3}{7}$
A	One card is already drawn. Hence there are '6' cards remaining. Now one card can be drawn from 6 cards in ${}^6C_1 = 6$ ways $\therefore N = 6$	In remaining '6' cards there are 'two' cards written alphabet 'A'. Hence one card can be drawn from 2 'A' cards in ${}^2C_1 = 2$ ways. $\therefore N(A) = 2$	$P(A) = \frac{N(A)}{N} = \frac{2}{6} = \frac{1}{3}$

C	Two cards are already drawn. Hence there are '5' cards. One card can be drawn out of 5 cards in ${}^5C_1 = 5$ ways $\therefore N = 5$	In remaining '5' cards there is only one 'C' card. Hence one 'C' card can be drawn in ${}^1C_1 = 1$ ways. $\therefore N(C) = 1$	$P(C) = \frac{N(C)}{N} = \frac{1}{5}$
T	Three cards are already drawn. There are only four cards left. One card can be drawn out of 4 cards in ${}^4C_1 = 4$ ways. $\therefore N = 4$	There were total '3' T cards. One 'T' card is already drawn. Hence only '2' *T* cards are left. One card can be drawn from there '2' *T* cards in ${}^2C_1 = 2$ ways. $\therefore N(T) = 2$	$P(T) = \frac{N(T)}{N} = \frac{2}{4} = \frac{1}{2}$

Since the cards are drawn in succession,

$$P(TACT) = P(T) \times P(A) \times P(C) \times P(T) = \frac{3}{7} \times \frac{1}{3} \times \frac{1}{5} \times \frac{1}{2} = \frac{1}{70}$$

Example 5.2.15 *In a digital communication channel the probability of sending '0' or '1' is 0.5. If the probability of error due to noise in channel is 0.05, find the probability of sending '0' when the received bit is '1'.*

Solution : Fig. 5.2.1 shows the digital communication channel. Various probabilities are shown in the figure.

The probability of error means $P(B_1/A_0)$ or $P(B_0 / A_1)$. It is 0.05 and shown in the figure. Hence $P(B_0/A_0) = P(B_1/A_1) = 1 - 0.05 = 0.95$.

Probability of sending '0' when received bit is '1' means $P(A_0/B_1)$. This probability is to be evaluated. For the communication channel

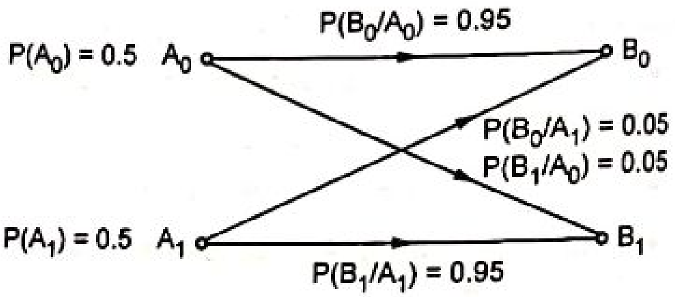


Fig. 5.2.1 Digital communication channel

$$P(B_1) = P(B_1/A_1)P(A_1) + P(B_1/A_0)P(A_0) = 0.95 \times 0.5 + 0.05 \times 0.5 = 0.5$$

By standard relations,

$$P(A_0B_1) = P(A_0/B_1)P(B_1)$$

also $$P(A_0B_1) = P(B_1/A_0)P(A_0)$$

From above two equations,

$$P(A_0/B_1)P(B_1) = P(B_1/A_0)P(A_0)$$

$$\therefore \quad P(A_0/B_1) = \frac{P(B_1/A_0)P(A_0)}{P(B_1)} = \frac{0.05 \times 0.5}{0.5} = 0.05$$

Thus $P(A_0/B_1) = 0.05$ i.e. probability of sending '0' when received bit is 1.

Example 5.2.16 *A certain computer becomes inoperative, if two components A and B both fail. The probability that A fails is 0.01 and the probability that B fails is 0.005. However the probability that B fails increases by a factor of 4, if A has failed. Calculate the probability that the computer becomes inoperable. Also find the probability that A will fail if B has failed. Comment on the result of conditional probability.*

Solution : The given data is,

$$P(A) = 0.01$$

$$P(B) = 0.005$$

The probability that B fails if A has failed is $P(B/A)$. It is given as,

$$P(B/A) = P(B) \times 4 = 0.005 \times 4 = 0.02$$

i) Probability that computer becomes inoperable :

Computer is inoperative if A and B both fail simultaneously. Hence this probability will be represented by joint probability $P(AB)$. From equation (5.2.12) we have,

$$P(B/A) = \frac{P(AB)}{P(A)}$$

$$\therefore \quad P(AB) = P(B/A)\, P(A) = 0.02 \times 0.01 = 0.0002$$

ii) Probability that A fails if B has failed :

This probability is $P(A/B)$. From equation 5.2.12 we have,

$$P(A/B) = \frac{P(AB)}{P(B)} = \frac{0.0002}{0.005} = 0.04$$

5.2.8 Joint Probability

- A joint probability is a probability that measures the likelihood that two or more events will happen concurrently.
- If there are two independent events A and B, the probability that A and B will occur is found by multiplying the two probabilities. Thus for two events A and B, the special rule of multiplication shown symbolically is :

  P(A and B) = P(A) P(B).
- The general rule of multiplication is used to find the joint probability that two events will occur. Symbolically, the general rule of multiplication is,

$$P(\text{A and B}) = P(A)\,P(B\,|\,A).$$

- The probability $P(A \cap B)$ is called the joint probability for two events A and B which intersect in the sample space. Venn diagram will readily shows that

$$P(A \cap B) = P(A) + P(B) - P\,(A \cup B)$$

Equivalently :

$$P(A \cap B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$$

- The probability of the union of two events never exceeds the sum of the event probabilities.
- A tree diagram is very useful for portraying conditional and joint probabilities. A tree diagram portrays outcomes that are mutually exclusive.
- Based on joint distribution on two events p(A, B),we can define the marginal distribution as follows :

$$p(A) = \sum_{b} p(A, B) = \sum_{b} p(A|B = b)\, p(B = b)$$

Summing up the all probable states of B gives the total probability formulae, which is also called sum rule or the rule of total probability.

- p(B) can be defined as

$$p(B) = \sum_{a} p(A, B) = \sum_{a} p(B|A = a)\, p(A = a)$$

5.2.9 Bayes' Rule

Let $B_1, B_2, B_3, \ldots\ldots B_n$ be mutually exclusive events and event A occurs only when any one of $B_1, B_2, B_3, \ldots\ldots B_n$ occurs. Then,

$$\boxed{P = (B_i\,/\,A) = \frac{P(B_i)\,P(A\,/\,B_i)}{\sum\limits_{i=1}^{n} P(B_i)\,P(A\,/\,B_i)}} \qquad \ldots (5.2.18)$$

This relation is called Bayes' rule or Bayesian Policy.

Proof : We know from statement of Bayes' rule that $B_1, B_2, B_3 \ldots\ldots B_n$ are mutually exclusive events and event A occurs only when any one of $B_1, B_2, B_3 \ldots\ldots B_n$ occurs. That is event A occurs jointly with any one of $B_1, B_2, B_3 \ldots\ldots B_n$. In other words '$A$' occurs certainly whenever AB_1 or AB_2 or AB_3 or .... AB_n occurs. Therefore we can define probability of event A in terms of joint events $AB_1, AB_2, AB_3, \ldots\ldots AB_n$. i.e.

$$P(A) = P\,(AB_1) + P\,(AB_2) + \ldots\ldots + P\,(AB_n) \qquad \ldots (5.2.19)$$

$$= \sum_{i=1}^{n} P\,(A\,B_i) \qquad \ldots (5.2.20)$$

$$\therefore \quad P(AB) = P(A)P(B/A) = P(B)P(A/B) \quad \text{... By equation 5.2.19} \quad \text{... (5.2.21)}$$

Now if B has multiple mutually exclusive events $B_1, B_2, B_3 \; ..... \; B_n$ then equation (5.2.21) can be written as,

$$P(AB_i) = P(A)\,P(B_i/A) = P(B_i)\,P\,(A/B_i) \quad \text{... (5.2.22)}$$

$$\text{i.e. } P(A)\,P(B_i/A) = P(B_i)\,P(A/B_i)$$

$$\therefore \quad P(B_i/A) = \frac{P(B_i)\;P(A\,/\,B_i)}{P\,(A)} \quad \text{... (5.2.23)}$$

Let us substitute value of $P(\text{AB}_i)$ from equation (5.2.22) into equation (5.2.20). i.e.,

$$P(A) = \sum_{i=1}^{n} P(A\,B_i) = \sum_{i=1}^{n} P(B_i)\,P(A/B_i)$$

$$\therefore \quad P(A) = \sum_{i=1}^{n} P(B_i)\,P(A/B_i) \quad \text{... (5.2.24)}$$

Putting value of $P(A)$ from above equation in equation (5.2.23) gives,

$$P(B_i/A) = \frac{P(B_i)\;P(A\,/\,B_i)}{\sum_{i=1}^{n} P(B_i)\,P\,(A\,/\,B_i)} \quad \text{... (5.2.25)}$$

This is the complete proof of Bayes' Rule.

Example 5.2.17 *Consider that there are three identical bags A,B and C. The bag A contains 2 gold coins, bag B contains 2 silver coins and bag C contains 1 silver and 1 gold coin. What is the probabilty that if the coin is gold, it is taken from bag 'A'.*

Solution : Let, B_1, B_2 and B_3 be the events that bags A, B, and C are selected respectively.

And, Let A be the event that gold coin is selected.

There are three bags and probability of selecting any one bag is same for all the three i.e.,

$$P(B_1) = P(B_2) = P(B_3) = \frac{1}{3}$$

$$\text{P(selecting gold coin from bag } A\text{)} = P\left(\frac{A}{B_1}\right) = \frac{2}{2} = 1$$

$$\text{P(selecting gold coin from bag } B\text{)} = P\left(\frac{A}{B_2}\right) = \frac{0}{2} = 0$$

$$P(\text{selecting gold coin from bag } C) = P\left(\frac{A}{B_3}\right) = \frac{1}{2}$$

Now Probability that the coin is gold, it is taken from bag A will be $P(B_1/A)$. It is obtained using Bayers' theorem i.e.,

$$P(B_1/A) = \frac{P(B_1)P(A/B_1)}{P(B_1)P(A/B_1)+P(B_2)P(A/B_2)+P(B_3)P(A/B_3)}$$

$$= \frac{\frac{1}{3}\times 1}{\frac{1}{3}\times 1+\frac{1}{3}\times 0+\frac{1}{3}\times\frac{1}{2}} = \frac{2}{3}$$

Example 5.2.18 *When the machine is set correctly, it produces 25 % defectives ; otherwise it produces 60 % defectives. From the past knowledge and experience, the manufacturer knows that the chances that the machine is set correctly or wrongly are 50 : 50. The machine was set and before commencement of production, one piece was inspected and found to be defective. What is the probability of machine set up being correct ?*

Solution : Let, A indicates that piece is defective

B_1 indicates that set up was correct

B_2 indicates that set up was wrong

$P(B_1)$ = Probability that set up was correct = 0.5

$P(B_2)$ = Probability that set up was wrong = 0.5

$P(A/B_1)$ = Probability that sample is defective given that set up was correct = 0.25

$P(A/B_2)$ = Probability that sample is defective given that set up was wrong = 0.60

Using Bayers' theorem,

$P(B_1/A)$ = Probabiity of set up being correct given that sample was defective

$$= \frac{P(B_1)P(A/B_1)}{P(B_1)P(A/B_1)+P(B_2)P(A/B_2)} = \frac{0.5\times 0.25}{0.5\times 0.25+0.5\times 0.6} = 0.294$$

Example 5.2.19 *Suppose box A contains 4 red and 5 blue chips and box B contains 6 red and 3 blue chips. A chip is chosen at random from box A and placed in box B. Finally, a chip is chosen at random from box B. What is the probability a blue chip was transferred from box A to box B given that the chip chosen from box B is red ?*

Solution : Let us define the following events :

A = Chip chosen from box B is red

B_1 = Blue chip is transferred from box A to box B

B_2 = Red chip is transferred from box A to box B

We have to find $P(B_1 / A)$ i.e. blue chip is transferred from box A to box B , given that chip chosen from box B is red.

Here,

A = {4 Red 5 Blue} Total 9 chips

B = {6 Red 3 Blue} Total 9 chips

$\therefore$ $P(B_1) = \frac{5}{9}$ and $P(B_2) = \frac{4}{9}$

$P(A / B_1)$ = Probability of selecting red chip form box 'B' given that blue chip was transferred from box A to box B.

$= \frac{6}{10}$ (After transferring there will be 10 chips in box B)

$P(A / B_2)$ = Probability of selecting red chip from box 'B' given that red chip was transferred from box A to box B.

$= \frac{7}{10}$ (After transferred red chip there will be 7 red chips and total 10 chips in box B)

Using Bayers' theorem,

$$P(B_1/A) = \frac{P(B_1)\,P(A/B_1)}{P(B_1)\,P(A/B_1)+P(B_2)\,P(A/B_2)} = \frac{\frac{5}{9}\times\frac{6}{10}}{\frac{5}{9}\times\frac{6}{10}+\frac{4}{9}\times\frac{7}{10}} = \frac{15}{29}$$

Example 5.2.20 *Suppose that a laboratory test to detect a certain disease has the following statistics.*

Let A = event that the tested person has the disease

B = event that the test result is positive

It is known that P(B / A) = 0.99 and P(B / A^c) = 0.005

and 0.1% of the population actually has the disease. What is the probability that a person has the disease given that the test result is positive ?

Soluiton : Let A^c = event that the tested person does not have a disease

$P(A) = 0.001$. Hence, $P(A^c) = 1 - P(A) = 1 - 0.001 = 0.999$

We have to find $P(A/B)$ i.e. Person has the disease given that the test result is positive i.e.,

$$P(A/B) = \frac{P(A)P(B/A)}{P(A)P(B/A)+P(A^c)P\left(B/A^c\right)} = \frac{0.001\times 0.99}{0.001\times 0.99+0.999\times 0.005} = 0.1654$$

5.3 Random Variables

- The distribution function F(x) or the density f(x) completely characterizes the behavior of a random variable X. The concept of a random variable will enable us to replace the original probability space with one in which events are set of numbers.
- Whenever you run and experiment, flip a coin, roll a die, pick a card, you assign a number to represent the value to the outcome that you get. This assign is called **a random variable.**
- A random variable is a variable X that assigns a real number [x], for each and every outcome of a random experiment. If S is the sample space containing all the 'n' outcomes $\{e_1, e_2, e_3, \ldots, e_i \ldots, e_n\}$ of random experiment, and X is a random variable defined as a function X(e) on S, then for every outcome e_i (where i = 1, 2, 3, ..., n) that is in S the random variable $X(e_i)$ will assign a real value x_i.
- Advantages of random variables is that user can define certain probability functions that make it both convenient and easy to compute the probabilities of various events.

5.3.1 Discrete Random Varaible

- The random variable is called a **discrete random variable** if it is defined over a sample space having a finite or a countable infinite number of sample points. In this case, random variable takes on discrete values and it is possible to enumerate all the values it may assume.
- A discrete random variable can only have a specific (or finite) number of numerical values.
- We can have **infinite discrete random variables** if we think about things that we know have an estimated number. Think about the number of stars in the universe. We know that there are not a specific number that we have a way to count so this is an example of an infinite discrete random variable.
- Another example would be with investments with share market. If you were to invest ₹ 1 lakh at the start of year, you could only estimate the amount you would have at the end of year.

5.3.2 Continuous Random Variable

- In the case sample space having an uncountable infinite number of sample points, the associated random variable is called a **continuous random variable,** with its values distributed over one or more continuous intervals on the real line. We make this distinction because they require different probability assignment considerations.
- A continous random variable is one having continuous range of values. It cannot be produced from a discrete sample space because of our requirement that all random variables be single valued functions of all sample space points.
- Both types of random variables are important in science and engineering.
- **Maxed random** variable is one for which some of its values are discrete and some are continuous.

5.3.3 Probability Distributions

- The behavior of a random variable is characterized by its probability distribution, that is, by the way probabilities are distributed over the values it assumes. A probability mass function are two ways to characterize this distribution for a discrete random variable.
- They are equivalent in the sense that the knowledge of either one completely specifies the random variable. The corresponding functions for a continuous random variable are the probability distribution function, defined in the same way as it the case of discrete random variable and the probability density function.
- If X is random variable, then the function F(x) is defined by

$$F(x) = P\{X \leq x\}$$

is called the **probability distributed function (PDF)** of X. All probabilities concerning X can be stated in terms of F. The argument 'x' is any real number ranging from ∞ to ∞.

- The probability distribution function is also called **Cumulative Distribution Function (CDF).**

Properties

1. $F_X(-\infty) = 0$
2. $F_X(\infty) = 1$
3. $0 \leq F_X(x) \leq 1$
4. $F_X(x_1) \leq F_X(x_2)$ if $x_1 < x_2$

Proof of (4)

Consider the event $\{x_1 < X \le x_2\}$ with $x_2 > x_1$. The set $\{x_1, x_2\}$ is nonempty and ∈.
Hence

$$0 \le p[x_1 < X \le x_2] \le 1$$

But $\{X \le x_2\} = \{X \le x_1\} \cup \{x_1 < X \le x_2\}$

and $\{X \le x_1\}\{x_1 < X \le x_2\} = \phi$

Hence $F_X(x_2) = F_X(x_1) + p[x_1 < X \le x_2]$

or

$$p[x_1 < X \le x_2] = F_X(x_2) - F_X(x_1) \ge 0 \quad \text{for } x_2 > x_1.$$

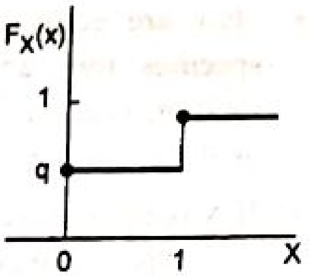
Some formula :

1. $p[a \le X \le b] = F_X(b) - F_X(a) + p[X = a]$
2. $p[a < x < b] = F_X(b) - p(X = b) - F_X(a)$
3. $p[a \le X < b] = F_X(b) - p[X = a] - F_X(a) + p[x = a]$

- Distribution functions of discrete random variables grows only by jumps, whereas the distribution functions of continuous random variables are continuous functions and hence have no jumps.
- If $F_X(x)$ is a continuous function of x, then

$$F_X(x) = F_X(x^-)$$

- However, if $F_X(x)$ is discontinuous at the point x then,



$$F_X(x) - F_X(x^-) = p[x^- < X \le x]$$

$$= \lim_{\epsilon \to 0} p[x - \epsilon < X \le x]$$

$$\triangleq p[X = x]$$

Typically $p[X = x]$ is a discontinuous function of x; it is zero whenever $F_X(x)$ is continuous and nonzero only at discontinuities in $F_X(x)$.

Example :

Consider tossing a coin four times. The possible outcomes are contained in the following table and the value of f in equation.

Tossing a coin four times

Elements of sample space	Probability	Value of random variable X (x)
HHHH	1/16	4
HHHT	1/16	3
HHTH	1/16	3
HTHH	1/16	3
THHH	1/16	3
HHTT	1/16	2
HTHT	1/16	2
HTTH	1/16	2
THHT	1/16	2
THTH	1/16	2
TTHH	1/16	2
HTTT	1/16	1
THTT	1/16	1
TTHT	1/16	1
TTTH	1/16	1
TTTT	1/16	0

Example 5.3.1 *Probability of a function of the number of Heads from tossing a coin four times. Determine the cumulative distribution function.*

Solution :

$$F(0) = f(0) = \frac{1}{16}$$

$$F(1) = f(0) + f(1)$$

$$= \frac{1}{16} + \frac{4}{16} = \frac{5}{16}$$

$$F(2) = f(0) + f(1) + f(2)$$

$$= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} = \frac{11}{16}$$

$$F(3) = f(0) + f(1) + f(2) + f(3)$$

$$= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16}$$

$$= \frac{1+4+6+4}{16} = \frac{15}{16}$$

$$F(4) = f(0) + f(1) + f(2) + f(3) + f(4)$$

$$= \frac{1}{16}+\frac{4}{16}+\frac{6}{16}+\frac{4}{16}+\frac{1}{16}$$

$$= \frac{1+4+6+4+1}{16} = \frac{16}{16} = 1$$

Example 5.3.2 *What is the probability distribution for the toss of one fair coin ?*

Solution :

$$P\text{ (Heads)} = \frac{1}{2}$$

$$P\text{ (Tails)} = \frac{1}{2}$$

Let heads denote the coin landing head side up.

Let tails denote the coin landing tail side up.

The possible outcomes are for the coin to land head side up or tail side up.

Using the alternative notation.

$$P(X = \text{Heads}) = \frac{1}{2}$$

$$P(X = \text{Tails}) = \frac{1}{2}$$

X	P (X)
Heads	$\frac{1}{2}$
Tails	$\frac{1}{2}$

5.3.4 Difference between Discrete and Continuous Random Variable

Sr. No.	Discrete	Continuous
1.	It uses countable set	It uses set of interval on R.
2.	F is set of all subset of Ω.	F is made from sub-intervals of Ω with set operations.

3.	For a set A ∈ F, $P(A) = \sum_{\omega \in A} p(\omega)$	For a set A ∈ F, $p(A) = \int_{A}^{-A} f_X(x)\,dx$
4.	Distribution function (Cdf) : $F_X(x) = \sum_{\omega \leq x} p\omega$	Distrinbution function (Cdf) : $F_X(x) = \int_{-\infty}^{x} f_X(x)\,dt$

Example 5.3.3 *Find the constant k such that the function*

$$f(x) = \begin{cases} kx^2 & 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$ *it is a density function, then find P (1 < X < 2).*

Solution :

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{3} kx^2 dx$$

$$= \left|\frac{kx^3}{3}\right|_0^3 = \frac{k(3)^3 - k(0)}{3}$$

$$= \frac{27k}{3} = 9k$$

This must be equal to 1, so we have

$$k = \frac{1}{9} \quad \text{and density function}$$

$$f(x) = \begin{cases} \frac{1}{9}x^2 & 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

$$P(1 < X < 2) = \int_{1}^{2} \frac{1}{9} x^2 dx$$

$$= \left.\frac{x^3}{27}\right|_1^2$$

$$= \frac{(2)^3 - (1)^3}{27} = \frac{8-1}{27}$$

$$= \frac{7}{27}$$

Example 5.3.4 *If x is a continuous random variable with probability density function given by*

$$f(x) = \begin{cases} k_x & when \quad 0< x< 2 \\ 2k & when \quad 2< x< 4 \\ k(b-x) & when \quad 4< x< 6 \\ 0 & otherwise \end{cases}$$

Find the value of k and also find the cumulative distribution function F(x).

Solution : By defination of probability density function :

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

$$\int_0^2 kx\,dx + \int_2^4 2k\,dx + \int_4^6 k(6-x)\,dx = 1$$

$$\left|\frac{kx^2}{2}\right|_0^2 + |2kx|_2^4 + \left|k(6x - \frac{x^2}{2})\right|_4^6 = 1$$

$$\frac{k((2)^2-(0)^2)}{2} + 2k(4-2) + k\left(\left(36-\frac{36}{2}\right)-\left(24-\frac{16}{2}\right)\right) = 1$$

$$\frac{4k}{2} + 4k + 2k = 1$$

$$2k + 4k + 2k = 1$$

$$k = \frac{1}{8}$$

So that

$$F(x) = p(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

For 0 < x < 2

$$f(x) = \int_0^x \frac{x}{8}dx = \left|\frac{x^2}{16}\right|_0^x = \frac{x^2}{16}$$

For 2 < x < 4

$$F(x) = \int_0^2 \frac{x}{8}dx + \int_2^x \frac{2}{8}dx$$

$$= \left|\frac{x^2}{16}\right|_0^2 + \left|\frac{2x}{8}\right|_2^x = \frac{4}{16} + \frac{2x-4}{8}$$

$$= \frac{1}{4} + \frac{x}{4} - \frac{1}{2} = \frac{1+x-2}{4}$$

$$= \frac{x-1}{4}$$

For 4 < x< 6

$$F(x) = \int_0^2 \frac{x}{8} dx + \int_2^4 \frac{2}{8} dx + \int_4^x \frac{1}{8}(6-x)dx$$

$$= \left|\frac{x^2}{16}\right|_0^2 + \left|\frac{2x}{8}\right|_2^4 + \left|\frac{1}{8}(6x - \frac{x^2}{2})\right|_4^x$$

$$= \frac{4-0}{16} + \frac{8-4}{8} + \frac{1}{8}\left[\left(6x - \frac{x^2}{2}\right) - \left(24 - \frac{16}{2}\right)\right]$$

$$= \frac{1}{4} + \frac{1}{2} + \frac{1}{8}\left(\left(6x - \frac{x^2}{2}\right) - 16\right)$$

$$= \frac{1}{4} + \frac{1}{2} + \frac{6x}{8} - \frac{x^2}{16} - 2$$

$$= \frac{4+8+12x-x^2-32}{16}$$

$$= \frac{12x-x^2-20}{16}$$

Therefore :

$$F(x) = \begin{cases} 0 & x< 0 \\ \frac{x^2}{16} & 0< x< 2 \\ \frac{x-1}{4} & 2< x< 4 \\ \frac{12x-x^2-20}{16} & 4< x< 6 \\ 1 & x \geq 6 \end{cases}$$

Example 5.3.5 *Probability density function of random variable x is $\frac{1}{2}\sin x$ in $0 \le x \le \pi$ otherwise 0. Find the mean, mode and median for the distribution and also find the probability between 0 and $\frac{\pi}{2}$.*

Solution : Given data

$$f(x) = \frac{1}{2}\sin x \quad D \le x \le \pi$$

$$= 0 \qquad \text{elsewhere}$$

Mean :

$$\text{Mean} = \int_{-\infty}^{\infty} xf(x)dx$$

$$= \int_{-\infty}^{\infty} xf(x)dx + \int_{0}^{\pi} xf(x)dx + \int_{\pi}^{\infty} xf(x)dx$$

$$= 0 + \frac{1}{2}\int_{0}^{\pi} x\sin x\, dx + 0$$

$$= \frac{1}{2}[-x\cos x + \sin x]_0^{\pi}$$

$$\text{Mean} = \frac{\pi}{2}$$

Mode :

f(x) is maximum for mode.

$\therefore$ f(x) = 0 and f′(x) is negative value.

$$f'(x) = \frac{1}{2}\cos x = 0 \quad \text{when } x = \frac{\pi}{2}$$

$$f''(x) = -\frac{1}{2}\sin x, \quad \text{when } x = \frac{\pi}{2}$$

So mode = $\frac{\pi}{2}$

Median :

$$\text{Medium} = \int_{0}^{m} f(x)dx$$

$$= \int_{m}^{\pi} f(x)dx$$

$$= \frac{1}{2}$$

$$\frac{1}{2}\int_{0}^{m} \sin x dx = \frac{1}{2}$$

$$= \left[\frac{1}{2} - \cos x\right]_{0}^{m}$$

$$= \frac{1}{2}(1 - \cos m)$$

$$= \frac{1}{2} - \frac{1}{2}\cos m$$

$$m = \frac{\pi}{2} \qquad (\cos m = 0)$$

Example 5.3.6 *A continous random variable X has the distribution function*

$$F(X) = 0 \text{ if } X \leq 1$$

$$= k(x-1)^4 \text{ if } 1 \leq X \leq 3$$

$$= 1 \quad \text{if} \quad x > 3$$

find k and probability density function.

Solution : Probability density function = f(x)

$$X = \frac{d}{dx}[F(x)]$$

i.e. $$f_x(x) = \frac{d}{dx} F_x(x)$$

$$\int_{-\infty}^{\infty} f_x(x)\, dx = 1$$

$$\int_{-\infty}^{\infty} f_x(x)\, dx + \int_{1}^{3} f_x(x)\, dx + \int_{3}^{\infty} f_x(x)\, dx = 1$$

$$0 + 4k\int_{1}^{3} (x-1)^3\, dx + 0 = 1$$

$$4k\left[\frac{(x-1)^4}{4}\right]_1^3 = 1$$

$$4k\left[\frac{(3-1)^4}{4} - \frac{(1-1)^4}{4}\right] = 1$$

$$4k\left[\frac{(2)^4}{4} - 0\right] = 1$$

$$4k\left[\frac{16}{4}\right] = 1$$

$$16k = 1$$

$$k = \frac{1}{16}$$

Example 5.3.7 *The random variable X has a probability function of the following form :*

$$f(x) = \begin{cases} k & \text{if} \quad x = 0 \\ 2k & \text{if} \quad x = 1 \\ 3k & \text{if} \quad x = 2 \end{cases}$$

otherwise where k is some number

a) Determine the value of k.

b) Find P $(x < 2)$, P $(x \le 2)$, P $(0 < x < 2)$

c) What is the smallest value of k for which
$F(x \le k) > 1/2$?

d) Determine the distribution of X.

Solution : a) Value of k

$$\sum_{i=0}^{2} P(x) = 1$$

$$\therefore \quad k + 2k + 3k = 1$$

$$6k = 1$$

$$k = \frac{1}{6}$$

b) $$P(x < 2) = P(x = 0) + P(x = 1)$$

$$= k + 2k$$

$$= 3k$$

$$= 3 \times \frac{1}{6} \qquad \left(\therefore k = \frac{1}{6}\right)$$

$$P(x < 2) = \frac{1}{2}$$

$$\begin{aligned} P(x \leq 2) &= P(x=0) + P(x=1) + P(x=2) \\ &= k + 2k + 3k \\ &= 6k \\ &= 6 \times 1/6 \\ &= 1 \end{aligned}$$

$$\begin{aligned} P(0 < x < 2) &= P(x=0) + P(x=1) \\ &= k + 2k \\ &= 3k \\ &= 3 \times \frac{1}{6} \end{aligned}$$

$$P(0 < x < 2) = \frac{1}{2}$$

c) Smallest value of k for which $F(x \leq k) > \frac{1}{2}$

$$\begin{aligned} P(x \leq 1) &= P(X = 0) + P(X = 1) \\ &= k + 2k \\ &= 3k \\ &= 3 \times \frac{1}{6} \\ &= \frac{1}{2} \end{aligned}$$

$$\begin{aligned} P(x \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= k + 2k + 3k \\ &= 6k \\ &= 6 \times \frac{1}{6} \\ &= 1 \end{aligned}$$

The smallest value of k for which $F(x \leq k) > \frac{1}{2}$ is $k = 2$.

d) Distribution of X.

X	F(X) = P (X≤x)
0	0
1	1/2
2	1

Example 5.3.8 *If two cards are drawn from a pack of 52 cards which are diamends. Using Poission distribution find the probability of getting two diamonds at least three times in 51 consecutive trials of two cards drawing each time.*

Solution : P = Probability of getting two diamends from a pack of 52.

n = 51

we can write,

$$\text{Randomly} = \frac{{}^{13}C_2}{{}^{52}C_2} = \frac{\dfrac{13\,!}{2!\,(13-2)\,!}}{\dfrac{52\,!}{2\,(52-2)\,!}} = \frac{3}{51}$$

$$\text{Mean} \qquad \mu = nP$$

$$= 51 \times \frac{3}{51}$$

$$\mu = 3$$

$$P\,(x \geq 3) = 1 - P\,(X = 0) - P\,(X = 1) - P\,(X = 2)$$

$$= 1 - e^{-3} - e^{-3} - e^{-3} \times \frac{9}{2}$$

$$= 1 - e^{-3}\left(\frac{17}{2}\right)$$

$$= 0.5767$$

Probability of getting two diamends at least three times is 0.5767.

Example 5.3.9 *If X is a poisson variant such that P (X = 0) = P (X = 1) find P (X = 0) and using recurrence formula find the probability at x = 1, 2, 3, 4 and 5.*

Solution : Given data

$$P\,(X = 0) = P\,(X = 1)$$

$$\therefore \quad \frac{e^{-\lambda}\lambda^0}{0!} = \frac{e^{-\lambda}\lambda^1}{1!} \Rightarrow \lambda = 1$$

$$P(X = 0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-1} = 0.3678$$

Poission distribution using recurrence formula

$$P(r+1) = \frac{\lambda}{r+1} P(r)$$

$$= \frac{1}{r+1} P(r)$$

$$P(1) = P(0+1) = \frac{1}{0+1} P(0)$$

$$= \frac{1}{1}(0.3678)$$

$$= 0.3678$$

$$P(2) = P(1+1) = \frac{1}{1+1} P(1)$$

$$= \frac{1}{2}(0.3678)$$

$$= 0.1839$$

$$P(3) = P(2+1) = \frac{1}{2+1} P(2)$$

$$= \frac{1}{3}(0.1839)$$

$$= 0.0613$$

$$P(4) = P(3+1) = \frac{1}{3+1} P(3)$$

$$= \frac{1}{4}(0.0613)$$

$$= 0.015325$$

$$P(5) = P(4+1) = \frac{1}{4+1} P(4)$$

$$= \frac{1}{5}(0.015325)$$

$$= 0.003065$$

Probability at x

x =	1	2	3	4	5
Probability	0.3678	0.1839	0.0613	0.015325	0.003065

Example 5.3.10 *Births in a hospital occur randomly at an average rate of 1.8 births per hour. What is probability of observing 4 births in a given hour at the hospital ?*

Solution : Let X be the number of births in a given hour.

Mean rate $\lambda = 1.8$

Probability of observing 4 births per hour

$$P(X=4) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$= \frac{e^{-1.8}(1.8)^{4}}{4!}$$

$$= \frac{e^{-1.8}\,(10.4976)}{24}$$

$$= \frac{0.16259 \times 10.4976}{24}$$

$$P(X=4) = 0.072297$$

Example 5.3.11 *A box contains 9 cards numbered 1 to 9. If four cards are drawn with replacement. What is the probability that none is 1 ?*

Solution : Given data : n = 9

The probability of getting one on the card = P = 1/9.

$$q = 1 - p$$

$$= 1 - \frac{1}{9} = \frac{9-1}{9} = \frac{8}{9}$$

By using Binomial distribution formula : ${}^{n}C_{x}\, p^{x} q^{n-x}$

The probability that none is '1' :

$$P(X=0) = {}^{n}C_{x}\, p^{x} q^{n-x}$$

$$= {}^{4}C_{0}\ (1/9)^{0}\ (8/9)^{4-0}$$

$$= \frac{4!}{0!\,(4-0)!} \times \frac{1}{1} \times \frac{4096}{6561} = \frac{98304}{6561} = 14.98$$

Example 5.3.12 *An insurance agent accepts policies of 5 men all identical age and good in health. The probability that a man of this age will be alive 30 years is 2/3. Find the probability that in 30 years :*

i) All five men ii) At least one man iii) Almost three will be alive.

Solution : Given data

The probability that a man of identical age and good in health will be alive 30 years : $p = \frac{2}{3}$.

$$n = 5$$

$$q = 1 - p$$

$$= 1 - \frac{2}{3} = \frac{3-2}{3} = \frac{1}{3}$$

By using Binomial distribution formula : $^nC_x\ p^x q^{n-x}$

i) All five men

The probability of all the five men being alive is

$$P(X = 5) = {}^nC_x\ p^x q^{n-x}$$

$$= {}^5C_5\ (2/3)^5\ (1/3)^{5-5}$$

$$= \frac{5!}{5!\,(5-5)!} \times \frac{32}{243} \times \frac{1}{1} = \frac{32}{243} = 0.1316$$

ii) At least one man

$$P(X < 1) = 1 - P(x = 0)$$

$$= 1 - {}^5C_0 (2/3)^0\ (1/3)^{5-0}$$

$$= 1 - \frac{5!}{0!\,(5-0)!} \times (1) \times \frac{1}{243} = 1 - \frac{1}{243} = \frac{242}{243}$$

iii) Almost three will be alive

The probability of almost three will be alive is :

$$P(x <= 3) = 1 - P(x > 3)$$

$$= 1 - [P(x = 4) + P(x = 5)]$$

$$= 1 - [{}^5C_4\ (2/3)^4\ (1/3)^{5-4} + {}^5C_5\ (2/3)^5\ (1/3)^{5-5}]$$

$$= 1-\left[\frac{5\,!}{4!\,(5-4)!}\times\frac{16}{81}\times\frac{1}{3}+\frac{32}{243}\right]$$

$$= 1-\left[\frac{120}{24}\times\frac{16}{81}\times\frac{1}{3}+\frac{32}{243}\right]$$

$$= 1-\left[\frac{80}{243}+\frac{32}{243}\right]=\frac{243-112}{243}=\frac{131}{243}$$

Examples on Discrete Distribution

Example 5.3.13 *A manufacturing process produces thousands of capacitor per day. Every hour, supervisor selects a random sample of 50 capacitor and classifies each capacitor in the sample as conforming or non confirming. Find the probability of finding one or fewer nonconforming parts of capacitor.*

Solution : Let x be the random variable representing the number of nonconforming parts in the sample.

$$P(x) = \binom{50}{x}(0.01)^{x}\,(0.99)^{50-x}$$

where, $x = 0, 1, 2, 3, \ldots., 50.$

$$\binom{50}{x} = \frac{50\,!}{x!\,(50-x)\,!}$$

$$\begin{aligned}
P(x \leq 1) &= P(x=0)+P(x=1)\\
&= P(0)+P(1) = \binom{n}{x}P^{x}(1-P)^{n-x}\\
&= \sum_{x=0}^{1}\binom{50}{x}(0.01)^{x}\,(0.99)^{50-x}\\
&= \frac{50\,!}{0!\,(50-0)\,!}(0.99)^{50}\,(0.01)^{0}+\frac{50\,!}{1!\,(50-1)\,!}(0.99)^{49}\,(0.01)^{1}\\
&= \frac{50\,!}{50\,!}(0.605)\,(1)+\frac{50\,!}{49\,!}(0.611)\,(0.01)\\
&= 0.605+0.30555\\
&= 0.91055
\end{aligned}$$

Example 5.3.14 *Suppose that X has pdf $f(x) = \frac{24}{x^4}$, $x > 2$. Evaluate the variance of X.*

Solution : By definition

$$
\begin{aligned}
E(X) &= \int_2^\infty x\, f(x)\, dx \\
&= \int_2^\infty x \frac{24}{x^4}\, dx \\
&= 24 \int_2^\infty x^{-3} dx \\
&= 24 \left| -\frac{1}{2x^2} \right|_2^\infty \\
&= 3 \\
E(X^2) &= \int_2^\infty x^2 f(x)\, dx \\
&= \int_2^\infty x^2 \frac{24}{x^4}\, dx \\
&= 24 \int_2^\infty x^{-2} dx \\
&= 24 \left| -\frac{1}{x} \right|_2^\infty = 12 \\
V(X) &= E(X^2) - E(X)^2 \\
&= 12 - (3)^2 \\
&= 12 - 9 \\
&= 3
\end{aligned}
$$

Example 5.3.15 *Consider the function :*

$$f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 0, & \text{elsewhere} \end{cases}$$

Since, $0 < x < 1$, $f(x) \geq 0$ for all x.

Solution :

$$\int_0^1 f(x)\,dx = \int_0^1 2x\,dx$$

$$= 2\left|\frac{x^2}{2}\right|_0^1$$

$$= 2\left|\frac{1}{2}-0\right|$$

$$= 1$$

$$P(0.5< X\leq 1) = \int_{0.5}^{1} 2x\,dx$$

$$= 2\left|\frac{x^2}{2}\right|_{0.5}^{1}$$

$$= 2\left|\frac{1}{2}-\frac{(0.5)^2}{2}\right| = 2\left|\frac{1}{2}-\frac{0.25}{2}\right|$$

$$= 2|\,0.5-0.125|$$

$$= 0.75$$

Example 5.3.16 *The probability distribution of daily demand for a product is*

d	1	2	3	4	5
p (d)	0.1	0.1	0.3	0.3	0.2

Evaluate E(D)

Solution : By definition

$$E(D) = \sum_{i=1}^{n} d_i\ p(d_i)$$

$$= \sum_{1}^{5} d_i\ p(d_i)$$

$$= 1\ (0.1) + 2\ (0.1) + 3\ (0.3) + 4\ (0.3) + 5\ (0.2)$$

$$= 0.1 + 0.2 + 0.9 + 1.2 + 1.0$$

$$= 3.4$$

Example 5.3.17 *Let X be a random variable with probability density function*

$$f(x) = \begin{cases} C(1-x^2) & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the value of C.

Solution :

f(x) to be a probability distribution $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$1 = \int_{-1}^{1} C(1-x^2)\,dx$$

$$= \left[Cx - \frac{Cx^3}{3}\right]_{-1}^{1}$$

$$= \left(C - \frac{C}{3}\right) - \left(-C + \frac{C}{3}\right)$$

$$= \left(\frac{3C - C}{3}\right) - \left(\frac{-3C + C}{3}\right)$$

$$= \left(\frac{2C}{3}\right) - \left(\frac{-2C}{3}\right)$$

$$1 = \frac{4C}{3}$$

$$C = \frac{3}{4}$$

Example 5.3.18 *A random variable x has a p.d.f. f(x) where $f(x) = e^{-x}$, $0 \le x \le \infty$. Find the probability that a) $0 \le x \le 2$, b) $x > 1$ c) $x < 0.5$.*

Solution :

$$\int_{0}^{\infty} f(x)\,dx = \int_{0}^{\infty} e^{-x} dx$$

$$= (-e^{x})_0^{\infty} = -(e^{-\infty} - e^{0})$$

$$= 1$$

Hence $f(x) = e^{-x}$ is a suitable function for a pdf.

a) $$P(0 \le x \le 2) = \int_{0}^{\infty} e^{-x} dx$$

$$= \int_{0}^{2} e^{-x} dx = (-e^{-x})_0^2 = (1 - e^{-2})$$

$$= 0.865$$

b) $$P(x > 1) = \int_{1}^{\infty} e^{-x} dx$$

$$= (-e^{-x})_1^{\infty} = (-e^{-1} - 1)$$

$$= 0.368$$

c) $$P(x < 0.5) = \int_{0}^{0.5} e^{-x} dx$$

$$= (-e^{-x})_0^{0.5} = 1 - e^{-0.5}$$

$$= 0.393$$

Example 5.3.19 *The probability density function of the continuous variable x is given by :*

$$f(x) = \begin{cases} \frac{1}{16}(3+x)^2; & -3 \le x \le -1 \\ \frac{1}{16}(2-6x)^2; & -1 \le x \le 1 \\ \frac{1}{16}(3-x)^2, & 1 \le x \le 3 \end{cases}$$

Show that the area under the curve above x-axis is unity. Also find the mean of the distribution.

Solution : As per definition, we have

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

$$\therefore \quad \int_{-3}^{3} f(x)\, dx = 1$$

$$= \int_{-3}^{-1} \frac{1}{16}(3+x)^2\, dx + \int_{-1}^{1} \frac{1}{16}(2-6x)^2 + \int_{1}^{3} \frac{1}{16}(3-x)^2 dx = 1$$

$$= \int_{-3}^{-1} \frac{1}{16}(3+x)^2\,dx = \int_{-3}^{-1} \frac{1}{16}(9+6x+x^2)\,dx$$

$$= \frac{1}{16}\left[9x+6\frac{x^2}{2}+\frac{x^3}{3}\right]_{-3}^{-1}$$

$$= \frac{1}{16}\left[9(-1)+6\frac{(-1)^2}{2}+\frac{(-1)^3}{3}-\left(9(-3)+6\frac{(-3)^2}{2}+\frac{(-3)^3}{3}\right)\right]$$

$$= \frac{1}{16}\left[\left(-9+\frac{6}{2}-\frac{1}{3}\right)-\left(-27+\frac{54}{2}-\frac{27}{3}\right)\right]$$

$$= \frac{1}{16}\left[\left(-6-\frac{1}{3}\right)-(-27+27-9)\right]$$

$$= \frac{1}{16}\left[\left(\frac{-18-1}{3}\right)-(-9)\right] = \frac{1}{16}\left[\frac{-19}{3}+9\right] = -\frac{19}{16\times 3}+\frac{9}{16}$$

$$= \frac{9}{16}-\frac{19}{48} = \frac{(9\times 3)-19}{48} = \frac{27-19}{48} = \frac{8}{48}$$

$$= \frac{1}{6}$$

$$\text{Mean of } f(x) = \int_{-\infty}^{\infty} x\,f(x)\,dx$$

$$= \int_{-3}^{-1} \frac{(3+x)^2}{16}\,xdx + \int_{-1}^{3} \frac{6-2x^2}{16}\,x\,dx \int_{-1}^{3} \frac{(3-x)^2}{16}\,xdx$$

$$= \frac{1}{16}\int_{-3}^{-1}(9x+6x^2+x^3)\,dx + 0 + \frac{1}{16}\int_{1}^{3}(9x-6x^2+x^3)\,dx$$

$$= \frac{1}{16}\left[\frac{9x^2}{2}+\frac{6x^3}{3}+\frac{x^4}{4}\right]_{-3}^{-1} + \frac{1}{16}\left[\frac{9x^2}{2}-\frac{6x^3}{3}+\frac{x^4}{4}\right]_{1}^{3}$$

$$= \frac{1}{16}\left[\left(\frac{9}{2}-2+\frac{1}{4}\right)-\left(\frac{81}{2}-54+\frac{81}{4}\right)+\frac{1}{16}\right]+\frac{1}{16}\left[\left(\frac{81}{2}-54+\frac{81}{4}\right)-\left(\frac{9}{2}-2+\frac{1}{4}\right)\right]$$

$$= \frac{1}{16}\left[\left(\frac{18-8+1}{4}\right)-\left(\frac{162-216+81}{4}\right)+\left(\frac{162-216+81}{4}\right)-\left(\frac{18-8+1}{4}\right)\right]$$

$$= \frac{1}{16}[0] = 0$$

5.4 Discrete Distributions

- A discrete distribution is a distribution of data in statistics that has discrete values. Discrete values are countable, finite, non-negative integers, such as 1, 10, 15, etc.
- For discrete data key distributions are : Bernoulli, Binomial, Poisson, and Multinomial

5.4.1 Binomial Distribution

- Binomial means 'two numbers'.
- The outcomes of health research are often measured by whether they have occured or not. For example, recovered from disease, admitted to hospital, died etc.
- The binomial distribution occurs in games of chance, quality inspection, opinion polls, medicine and so on.
- It may be modelled by assuming that the number of events 'n' has a binomial distribution with a fixed probability of event p. Binomial distribution is distribution for a series of Bernoulli trials.
- Binomial distribution written as B (n, p) where n is the total number of events and p = probability of an event.
- Properties of binomial distribution :
  1. Experiment consist of n identical trials.
  2. Each trial has only two outcomes.
  3. The probability of one outcome is p and the other is q = 1 – p.
  4. The trials are independent.
  5. We are interested in x, the number of success observed during the n trials.
- Trials satisfying the above properties are called **Bernoulli trials.**
- The probability function X,

$$f(x) = \binom{n}{x} p^x q^{n-x}$$

and f (x) = 0 otherwise. The distribution of X with probability function is called the binomial distribution or Bernoulli distribution.

- The mean μ (mu) of the binomial distribution is

$$\mu = np$$

- The variance is,

$$\sigma^2 = npq$$

- The mean and variance of binomial distribution with parameters (n, p) are given as,

$$\text{Mean} = \mu = E(X) = \sum_{i=1}^{n} E(X_i) = np$$

$$\text{Variance } \sigma^2 = V(X) = \sum_{i=1}^{n} V(X_i) = np(1-p) = npq$$

- A combination of n different objects taken r at a time is a selection of r out of n objects with attention not given to order of arrangments. It is denoted by ${}^nC_r$ or $C(n, r)$ or $\begin{pmatrix} n \\ r \end{pmatrix}$ and

$$ {}^nC_r = \frac{n(n-1)\ldots(n-r+1)}{r!} = \frac{n!}{r!\,(n-r)!}$$

$\begin{pmatrix} n \\ r \end{pmatrix}$ is called **binomial coefficient.**

5.4.1.1 Mean and Variance of the Binomial Distribution

$$\text{Mean } (\mu) = \sum_{i=0}^{n} x\, {}^nC_r p^x q^{n-x}$$

$$= nC_1 pq^{n-1} + 2nC_2 p^2 q^{n-2} + \ldots n\, nC_n\, p^n$$

$$= {}^nC_1 pq^{n-1} + \frac{2n(n-1)}{1\times 2} p^2 q^{n-2} + \ldots + \frac{n(n-1)}{1\times 2\times 3\times\ldots n} p^n$$

$$= np\left[q^{n-1} + (n-1)pq^{n-2} + \frac{(n-1)(n-2)}{2!} p^2 q^{n-3} + \ldots + p^{n-1}\right]$$

$$= np\,(q+p)^{n-1}$$

Using binomial theorem (p + q = 1).

Therefore $\mu = np\,(1)$

$$\mu = np$$

Variance (σ^2) V (X) :

$$
\begin{aligned}
V(X) &= E(X^2) - [E(X)]^2 \\
&= \sum_{x=0}^{n} x^2 p(x) - \mu^2 \\
&= \sum_{x=0}^{n} {}^nC_x p^x q^{n-x} x^2 - \mu^2 \\
&= 1 \times 2\, {}^nC_2\, p^2 q^{n-2} + 3 \times 2\, p^3 q^{n-3} + \ldots + {}^nC_n n\,(n-1)\, np \\
&\quad + \sum\ {}^nC_x x\, p^x q^{n-x} \mu^2 \\
&= n\,(n-1)\, p^2 \sum_{x=2}^{n} n-2C_{x-2} p^{x-2} q^{n-x} + np - n^2p^2 \\
&= n\,(n-1)\, p^2\, (p+q)^{n-2} + np - n^2p^2 \\
&= n\,(n-1)\, p^2 + np - n^2p^2 \\
&= n^2p^2 - np^2 + np - n^2p^2 \\
&= np - np^2 \\
&= np\,(1-p) \\
\sigma^2 &= npq \qquad\qquad (q = 1 - p)
\end{aligned}
$$

- The standard deviation (σ) of the binomial distribution is $\sqrt{npq}$.

An examples of the binomial distribution

Example 5.4.1 *Suppose a box contains a very large number of balls. Black ball are 2/3 and rests of the balls are red. We draw 5 balls from box from the box. How many black balls do we get ?*

Solution : Let,

X = Number of black balls in 5 draws.

So X can take on any of the values 0, 1, 2, 3, 4 and 5 and X is a discrete random variable.

Some values of X will be more likely to occur than others. Each value of X will have a probability of occuring. Consider the probability of obtaining just one yellow ball, i.e. X = 1.

One possible way of obtaining one yellow ball is if we observe the pattern BRRRR. The probability of obtaining this patterns is,

$$P\ (BRRRR) = \frac{2}{3}\times\frac{1}{3}\times\frac{1}{3}\times\frac{1}{3}\times\frac{1}{3}$$

There are 32 possible patterns of black and red balls we might observe, 5 of the patterns contain just one black ball.

BBBBB RBBBB BRBBB BBRBB BBBRB BBBBR RRBBB RBRBB
RBBRB RBBBR BRBRB BRBRB BRBBR BBRRB BBRBR BBBRR
RRRBB RRBRB RRBBR RBRRB RBRBR RBBRR BRRRB BRRBR
BRBRR BBRRR BRRRR RBRRR RRBRR RRRBR RRRRB RRRRR

The other 5 possible combinations all have the same probability so the probability of obtaining one head in 5 coin tosses is,

$$\begin{aligned} P\ (X = 1) &= 5\times\left(\frac{2}{3}\times\left(\frac{1}{3}\right)^4\right) \\ &= 5\times\frac{2}{3}\times\frac{1}{81} \\ &= 0.04115 \end{aligned}$$

We calculate the probability P (X = 2) :

$$\begin{aligned} P\ (X = 2) &= \text{Number of patterns} \times \text{Probability of pattern} \\ &= {}^5C_2 \times (2/3)^2 \times (1/3)^3 \\ &= \frac{5!}{2!(5-2)!}\times\frac{4}{9}\times\frac{1}{27} = \frac{120}{12}\times\frac{4}{9}\times\frac{1}{27} \\ &= 10\times\frac{4}{243} = 10\times 0.01646 \\ &= 0.1646 \end{aligned}$$

To write down a formula for this situation specific situation in which we toss a coin 5 times.

$$P\ (X = x) = {}^5C_x\times\left(\frac{2}{3}\right)^x\times\left(\frac{1}{3}\right)^{(5-x)}$$

Using this above formula, we can tabulate the probabilities of each possible value of X.

$$P(X = 0) = {}^5C_0 \times \left(\frac{2}{3}\right)^0 \times \left(\frac{1}{3}\right)^5 = 0.0041$$

$$P(X = 1) = {}^5C_1 \times \left(\frac{2}{3}\right)^1 \times \left(\frac{1}{3}\right)^4 = 0.0412$$

$$P(X = 2) = {}^5C_2 \times \left(\frac{2}{3}\right)^2 \times \left(\frac{1}{3}\right)^3 = 0.1646$$

$$P(X = 3) = {}^5C_3 \times \left(\frac{2}{3}\right)^3 \times \left(\frac{1}{3}\right)^2 = 0.3292$$

$$P(X = 4) = {}^5C_4 \times \left(\frac{2}{3}\right)^4 \times \left(\frac{1}{3}\right)^1 = 0.3292$$

$$P(X = 5) = {}^5C_5 \times \left(\frac{2}{3}\right)^5 \times \left(\frac{1}{3}\right)^0 = 0.1317$$

The distribution functions of X :

X	F (X) = P (X < x)
0	0.0041
1	0.0412
2	0.1646
3	0.3292
4	0.3292
5	0.1317

We plot the graph using distribution function value :

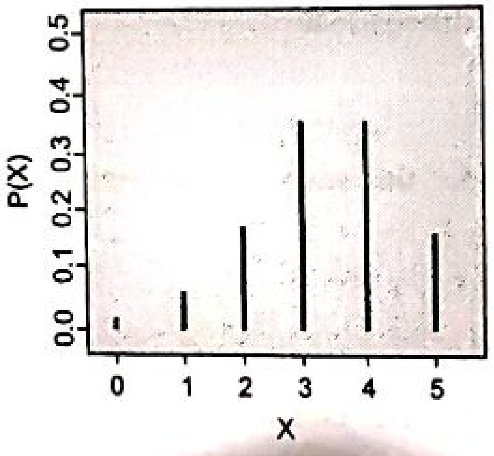


Fig. 5.4.1 A plot of the Binomial (5, 2/3) probabilities

Example 5.4.2 *Consider the example of the Binomial distribution*

X	0	1	2	3	4	5
P (X = x)	0.004	0.041	0.165	0.329	0.329	0.132

Calculate the mean value of distribution.

Solution :

$$\begin{aligned}\mu &= xP(X=0)+xP(X=1)+xP(X=2)+xP(X=3)+xP(X=4)+xP(X=5)\\ &= 0\times(0.004)+1\times(0.0041)+2\times(0.165)+3\times(0.329)+4\times(0.329)+5\times(0.132)\\ &= 0+0.0041+0.33+0.987+1.316+0.66\\ &= 3.2971\end{aligned}$$

5.4.1.2 Mean and Variance of Distribution

- The mean μ and variance σ^2 of a random variable X and of its distribution are the theoretical counterparts of the mean $\bar{x}$ and variance s^2 of a freqquency distribution.
- The mean μ (mu) is defined by :

$$\mu = \sum_j x_j f(x_j) \qquad \text{for discrete distribution}$$

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \qquad \text{for continuous distribution}$$

and the variance σ^2 (Sigma square) by :

$$\sigma^2 = \sum_j (x_j-\mu)^2 f(x_j) \qquad \text{for discrete distribution}$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx \qquad \text{for continuous distribution}$$

- The mean (μ) is also denoted by E (X) and is called the expectation of X because it gives the average value of X to be expected in many trials.
- Let us compute the variance of a normal distribution. If X has an $N(\mu,\sigma^2)$ distribution, then :

$$\text{Var}(X) = E[(X-E[X])^2]$$

$$= \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$= \sigma^2 \int_{-\infty}^{\infty} Z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2} dx$$

Here we substituted $Z = (x-\mu)/\sigma$.

Using integration,

$$\int_{-\infty}^{\infty} Z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}Z^2} dz = 1$$

Example 5.4.3 *Find the variance and standard deviation for the following set of test marks $T = \{75, 80, 82, 87, 96\}$*

Solution :

$$\text{Mean} = \frac{75+80+82+87+96}{5}$$

$$= \frac{420}{5}$$

$$\text{Mean} = 84$$

$$\text{Variance} = \frac{[(x_1-\mu)^2 + (x_2-\mu)^2 + \ldots + (x_n-\mu)^2]}{n}$$

$$\sigma^2 = \frac{[(75-84)^2 + (80-84)^2 + (82-84)^2 + (87-84)^2 + (96-84)^2}{5}$$

$$= \frac{(-9)^2 + (-4)^2 + (-2)^2 + (3)^2 + (12)^2}{5}$$

$$= \frac{81+16+4+9+144}{5} = \frac{254}{5}$$

$$\sigma^2 = 50.8$$

Standard Deviation (σ)

$$\sigma = \sqrt{\sigma^2} = \sqrt{50.8} = 7.1274$$

Examples on mean and median

Example 5.4.4 *In order to control costs, a company collects data on the weekly number of meals claimed on expense accounts. The numbers for five weeks are 15, 14, 2, 27 and 13.*

Solution :

$$\text{The mean} = \bar{x} = \frac{15+14+2+27+13}{5}$$

$$= \frac{71}{5} = 14.2$$

Median : Ordering the data from smallest to largest, we get

2, 13, 14, 15, 27

⇑

the medium is the third largest value i.e. 14.

Example 5.4.5 *If X is a normal variate with mean 30 and standard deviation 5. Find the probability that,*

a) $26 \le x \le 40$ *b)* $x \ge 45$.

Solution : Given data :

Mean $\mu = 30$

Standard deviation $\sigma = 5$.

i) $x_1 = 26$ and $x_2 = 40$

$$Z = \frac{x-\mu}{\sigma}$$

$$Z_1 = \frac{x_1-\mu}{\sigma}$$

$$= \frac{26-30}{5}$$

$$= \frac{4}{5}$$

$$= -0.8$$

$$Z_2 = \frac{x_2-\mu}{\sigma}$$

$$= \frac{40-30}{5}$$

$$= \frac{10}{5}$$

$$= 2$$

$$P(26 \leq x \leq 40) = P(-0.8 \leq z \leq 2)$$

ii) $x \geq 45$

$$Z = \frac{x-\mu}{\sigma}$$

$$= \frac{45-30}{5}$$

$$= \frac{15}{5}$$

$$= 3$$

Example 5.4.6 *A random variable X has the following probability function :*

x	0	1	2	3	4	5	6	7
P (x)	0	K	2K	2K	3K	K^2	$2K^2$	$7K^2+K$

Determine :

i) K ii) Evaluate P (X < 6), $P(X \geq 6)$, P (0 < X < 5) and $P(0 \leq X \leq 4)$.

iii) If $P(X \leq K) > \frac{1}{2}$ find the minimum value of K.

iv) Determine the distribution function of X.

v) Mean vi) Variance.

Solution : i) K

$$\sum_{x=0}^{7} P(x) = 1$$

$$K + 2K + 2K + 3K + K^2 + 2K^2 + 7K^2 + K = 1$$

$$10K^2 + 9K - 1 = 0$$

$$(K+1)(10K-1) = 0$$

$$K + 1 = 0 \quad \text{and} \quad 10K - 1 = 0$$

$$K = -1 \quad \text{and} \quad K = \frac{1}{10}$$

We discard K = – 1 value. Therefore $K = \frac{1}{10} = 0.1$.

ii) $P(X < 6) = P(X=0) + P(X=1) + P(X=2) + \ldots + P(X = 5)$

$$= 0 + K + 2K + 2K + 3K + K^2$$

Put $K = 0.1$

$= 0+0.1+2(0.1)+2(0.1)+3(0.1)+(0.1)^2$

$= 0.1 + 0.2 + 0.2 + 0.3 + 0.01$

$P(X < 6) = 0.81$

$P(X \geq 6) = 1 - P(X < 6)$

$= 1 - 0.81$

$P(X \geq 6) = 0.19$

$P(0 < X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$

$= K + 2K + 2K + 3K$

$= 8K$

$= 8 \times 0.1$ $(K = 0.1)$

$P(0 \leq X < 5) = 0.8$.

$P(0 \leq X \leq 4) = P(X=0)+P(X=1)+P(X=2)+P(X=3)+P(X=4)$

$= 0 + K + 2K + 2K + 3K$

$= 8K$

$= 8 \times 0.1$

$P(0 \leq X \leq 4) = 0.8$

iii) If $P(X \leq K) > \frac{1}{2}$, minimum value of K.

$P(X \leq 1) = P(X = 0) + P(X = 1)$

$= 0 + K$

$= K$

$= 0.1$

$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

$= 0 + K + 2K$

$= 3K$

$= 3 \times 0.1$

$= 0.3$

$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$$= 0 + K + 2K + 2K$$

$$= 5K$$

$$= 5 \times 0.1$$

$$= 0.5$$

$$P(X \leq 4) = 0.8 \text{ (We already calculated)}$$

But the condition is $P(X \leq K) > \frac{1}{2}$.

So K = 4 is suitable for this minimum value of K = 4.

iv) Distribution function of X.

X	F (X) = P (X≤x)
0	0
1	0.1
2	0.3
3	0.5
4	0.8
5	0.81
6	0.83
7	$9K + 10K^2 = 1$

v) Mean (μ)

$$\mu = \sum_{i=0}^{7} p_i x_i$$

$$= 0(0) + 1(K) + 2(2K) + 3(2K) + 4(3K) + 5(K^2) + 6(2K^2) + 7(7K^2 + K)$$

$$= 0 + K + 4K + 6K + 12K + 5K^2 + 12K^2 + 49K^2 + 7K$$

$$= 30K + 66K^2$$

$$\text{Substitute } K = 1/10$$

$$= 30 \times \frac{1}{10} + 66 \times \left(\frac{1}{10}\right)^2$$

$$= \frac{30}{10} + \frac{66}{100}$$

$$= 3 + 0.66$$

$$\mu = 3.66$$

vi) Variance (σ^2)

$$\sigma^2 = \sum_{i=0}^{7} p_i x_i^2 - \mu^2$$

$$= K + 8K + 18K + 48K + 25K^2 + 72K^2 + 343K^2 + 49K - (3.66)^2$$

$$= 440K^2 + 124K - 13.3956$$

$$= 440(0.1)^2 + 124(0.1) - 13.3956$$

$$= 4.4 + 12.4 - 13.3956$$

$$\sigma^2 = 3.4044$$

Example 5.4.7 *Find the probability of getting an even number 3 or 4 or 5 times in throwing 10 dice using binomial distribution.*

Solution :

P = Probability of getting even number in throw of a die.

$$P = \frac{3}{6} = \frac{1}{2}$$

$$q = 1 - p$$

$$= 1 - \frac{1}{2}$$

$$= \frac{1}{2}$$

n = 10 (Given data)

x = Probability of getting even number

$$P(X = x) = {}^{10}C_x p^x q^{n-x}$$

Substituting value of p, q, n, we get

$$= {}^{10}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-n}$$

$$= {}^{10}C_x \left(\frac{1}{2}\right)^{10} \qquad \text{(Where } x = 0, 1, 2, 3, ..., 10\text{)}$$

$$P(X = 3) = {}^{10}C_3 \left(\frac{1}{2}\right)^{10}$$

$$= \frac{10\,!}{3!\,(10-3)\,!} \times \frac{1}{1024}$$

$$= \frac{3628800}{6 \times 5040} \times \frac{1}{1024}$$

$$= \frac{120}{1024}$$

$$P\,(X = 3) = 0.11718$$

$$P\,(X = 4) = {}^{10}C_4 \left(\frac{1}{2}\right)^{10}$$

$$= \frac{10!}{4!\,(10-4)!} \times \frac{1}{1024}$$

$$= \frac{3628800}{24 \times 720} \times \frac{1}{1024}$$

$$= \frac{210}{1024}$$

$$P\,(X = 4) = 0.2050$$

$$P\,(X = 5) = {}^{10}C_5 \left(\frac{1}{2}\right)^{10}$$

$$= \frac{10!}{5!\,(10-5)!} \times \frac{1}{1024}$$

$$= \frac{3628800}{120 \times 120} \times \frac{1}{1024}$$

$$= \frac{252}{1024}$$

$$= 0.246$$

Example 5.4.8 *The mean of binomial distribution is 3 and variance is 9/4. Find*

i) The value of n

ii) $p\,(x \geq 7)$

iii) $p\,(1 \leq x \leq 6)$

Solution : Given data :

$$\mu = 3 \quad \sigma^2 = \frac{9}{4} = npq$$

i) Value of n

$$npq = \frac{9}{4}$$

$$3q = \frac{9}{4}$$

$$q = \frac{9}{4} \times \frac{1}{3} = \frac{3}{4}$$

$$p = 1 - q$$

$$= 1 - \frac{3}{4}$$

$$p = \frac{1}{4}$$

$$np = 3$$

$$n \times \frac{1}{4} = 3$$

$$n = 12$$

ii) $P(x \geq 7) = P(x=7) + P(x=8) + P(x=9) + P(x=10) + P(x=11) + P(x=12)$

Using binomial distribution

$$= {}^{n}C_{x} p^{x} q^{n-x}$$

$$P(x \geq 7) = {}^{12}C_7 \left(\frac{1}{4}\right)^7 \left(\frac{3}{4}\right)^{12-7} + {}^{12}C_8 \left(\frac{1}{4}\right)^8 \left(\frac{3}{4}\right)^{12-8}$$

$$+ {}^{12}C_9 \left(\frac{1}{4}\right)^9 \left(\frac{3}{4}\right)^{12-9} + {}^{12}C_{10} \left(\frac{1}{4}\right)^{10} \left(\frac{3}{4}\right)^{12-10}$$

$$+ {}^{12}C_{11} \left(\frac{1}{4}\right)^{11} \left(\frac{3}{4}\right)^{12-11} + {}^{12}C_{12} \left(\frac{1}{4}\right)^{12} \left(\frac{3}{4}\right)^{12-12}$$

$$= \frac{1}{(4)^{12}} [792 (3)^5 + 495 (3)^4 + 220 (3)^3 + 66 (3) + 12 (3) + 1]$$

$$= \frac{1}{(4)^{12}} [192456 + 40095 + 5940 + 594 + 36 + 1]$$

$$= \frac{239122}{16777216}$$

$$P(x \geq 7) = 0.0142$$

iii) $P(1 \leq x < 6) = P(x=1) + P(x=2) + P(x=3) + P(x=4) + P(x=5)$

Using binomial distribution

$$= {}^{12}C_1\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^{12-1} + {}^{12}C_2\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^{12-2} + {}^{12}C_3\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^{12-3}$$

$$+ {}^{12}C_4\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^{12-4} + {}^{12}C_5\left(\frac{1}{4}\right)^5\left(\frac{3}{4}\right)^{12-5}$$

$$= 0.1267 + 0.2322 + 0.2581 + 0.1935 + 0.1032$$

$$= 0.9137$$

5.4.2 The Poisson Distribution

- Poisson distribution, named after its invertor simeon poisson who was a French mathematician. He found that if we have a rare event (i.e. p is small) and we know the expected or mean (or μ) number of occurances, the probabilities of 0, 1, 2 ... events are given by :

$$\boxed{P(R) = \frac{e^{-\mu}\mu^R}{R!}}$$

Poisson distribution : Is a distribution the number of rare events that occur in a unit of time, distance, space and so on.

Examples :

1. Number of insurance claims in a unit of time.
2. Number of accidents in a ten-mile highway.
3. Number of airplane crash in triangle area.

- When there is a large number of trials, but a small probability of success, binomial calculate becomes impractical. Example : Number of deaths from horse kicks in the army in different years. The mean number of successes from n trials is $\mu = np$.
  - If we substitute μ/n for p, and let n tend to infinity, the binomial distribution becomes the Poisson distribution :

$$P(x) = \frac{e^{-\mu}\mu^x}{x!}$$

- Poisson distribution is applied where random events in space or time are expected to occur. Deviation from poisson distribution may indicate some degree of non-randomness in the events under study.
- Example : 64 deaths in 20 years from thousands of soldiers.

- If a mean or average probability of an event happening per unit time/per page/per mile cycled etc., is given and you ar asked to calculate a probability of n events happening in a given time/number of pages/number of miles cycled, then the **Poisson distribution** is used.
- If on the other hand, an exact probability of an event happenng is given, or implied, in the question, and you are asked to calculate the probability of this event happening k times out n, then the **Binomial distribution** must be used.

Example 5.4.9 *In oil exploration, the probability of an oil strike in the north sea is 1 in 500 drillings. What is the probability of having exactly 3 oil producing wells in 1000 explorations ?*

Solution : Given data :

$$n = 1000, \quad p = \frac{1}{500}$$

$$\mu = np = 1000 \times \frac{1}{500} = 2$$

The desired probability

$$= \frac{e^{-\mu}\mu^{x}}{x!}$$

$$= \frac{e^{-2}\,2^{3}}{3!}$$

$$= 0.18$$

5.4.3 Bernoulli Distribution

- The most basic of all discrete random variables is the Bernoulli. X is said to have a Bernoulli distribution if X = 1 occurs with probability Π and X = 0 occurs with probability 1 – Π.

$$f(x) = \begin{cases} \Pi & x-1 \\ 1-\Pi & x=0 \\ 0 & \text{otherwise} \end{cases}$$

- Suppose an experiment has only two possible outcomes, "success" and "failure," and let Π be the probability of a success. If we let X denote the number of successes (either zero or one), then X will be Bernoulli.

5.4.4 Multinomial Distribution

- The multinomial distribution is a generalization of the binomial distribution to k categories instead of just binary (success/fail).
- For n independent trials each of which leads to a success for exactly one of k categories, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.
- The multinomial distribution can be used to compute the probabilities in situations in which there are more than two possible outcomes.
- Example : Rolling a die N times

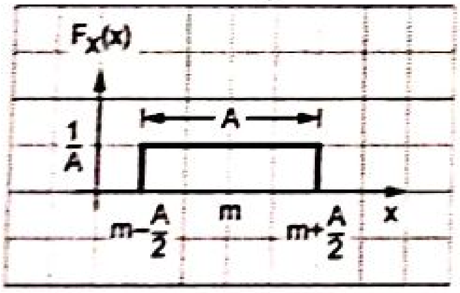
5.5 Continuous Distributions

- Continuous distributions are characterized by an infinite number of possible outcomes, together with the probability of observing a range of these outcomes.

5.5.1 Uniform Distribution

The PDF for a uniform distribution is given as,

$$\textbf{Uniform PDF:}\ f_X(x) = \begin{cases} 0 & \text{for } x < m - \frac{A}{2} \text{ and} \\ & x > m + \frac{A}{2} \\ \frac{1}{A} & \text{for } \left(m - \frac{A}{2}\right) \leq x \leq \left(m + \frac{A}{2}\right) \end{cases} \quad \dots (5.5.1)$$



A = Peak to peak value of a random variable.

$\frac{1}{A}$ = Amplitude of all possible values of random variable

Fig. 5.5.1 PDF of uniformly distributed random variable. The peak to peak value is 'A' and amplitude is uniform (i.e. A)

The value of PDF, $f_X(x)$ is same for all possible values of a random variable. Therefore this distribution is called *Uniform Distribution*.

Example 5.5.1 *Show that the mean and variance of a random variable 'X' having a uniform distribution in the interval [a, b] are,*

$$m_x = \frac{a+b}{2} \quad \text{and} \quad \sigma_x^2 = \frac{(a-b)^2}{12}$$

Solution : Fig. 5.5.2 shows the sketch of uniform distribution in the interval [a, b].

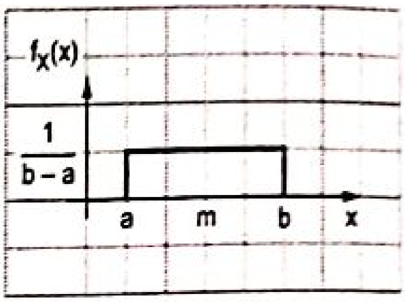


Fig. 5.5.2 Uniform distribution having interval [a, b]

To find mean value

The mean value of a continuous random variable is given by

$$m_x = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

$$= \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b$$

$$= \frac{1}{2(b-a)}[b^2-a^2] = \frac{1}{2(b-a)}(b-a)\cdot(b+a)$$

$$= \frac{b+a}{2} \text{ or } \frac{a+b}{2} \qquad \text{...(5.5.2)}$$

To find variance

Variance is given by equation (5.2.8) as,

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x-m_x)^2 f_X(x)\, dx$$

$$= \int_a^b (x-m_x)^2 \cdot \frac{1}{b-a}\, dx \quad \text{since } f_X(x) = \frac{1}{b-a}$$

Let $x - m_x = y$ then we have $dx = dy$.

And the limits will be,

when $x = a$, $y = a - m_x$ and when $x = b$, $y = b - m_x$

$$\therefore \quad \sigma_x^2 = \int_{a-m_x}^{b-m_x} y^2 \cdot \frac{1}{b-a}\, dy = \frac{1}{b-a}\left[\frac{y^3}{3}\right]_{a-m_x}^{b-m_x}$$

$$= \frac{1}{3(b-a)}[(b-m_x)^3-(a-m_x)^3]$$

Putting the value of $m_x = \frac{a+b}{2}$ from equation (6.3.12) in above equation we get,

$$\sigma_x^2 = \frac{1}{3(b-a)}\left[\left(b-\frac{a+b}{2}\right)^3-\left(a-\frac{a+b}{2}\right)^3\right] = \frac{1}{3(b-a)}\left[\left(\frac{b-a}{2}\right)^3-\left(\frac{a-b}{2}\right)^3\right]$$

$$= \frac{1}{-3(a-b)}\left[\left(-\frac{a-b}{2}\right)^3-\left(\frac{a-b}{2}\right)^3\right] \quad \text{Here we have written } b-a=-(a-b)$$

$$= \frac{1}{-3(a-b)} \times \frac{-(a-b)^3}{4} = \frac{(a-b)^2}{12} \quad \text{... (5.5.3)}$$

Thus, for uniform distribution,

$$\text{Mean, } m_x = \frac{a+b}{2} = m \text{ and variance, } \sigma_x^2 = \frac{(a-b)^2}{12} \quad \text{... (5.5.4)}$$

5.5.2 Normal Distribution

Gaussian distribution is also called *Normal Distribution*. It is defined for continuous random variables. The *PDF* for a Gaussian random variable is given as,

$$\textbf{Gaussian PDF : } f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \quad \text{... (5.5.5)}$$

Here 'm' is mean and σ^2 is variance.

Fig. 5.5.3 shows the sketch of Gaussian pdf.

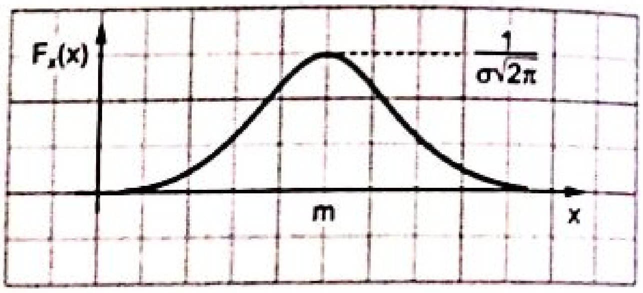


Fig. 5.5.3 Plot of Gaussian *PDF*

Properties of Gaussian *PDF*

Property 1 : The peak value occurs at $x = m$ (i.e. mean value). i.e.,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \quad \text{at} \quad x = m \quad \text{i.e. mean value} \quad \text{... (5.5.6)}$$

Property 2 : The plot of Gaussian *PDF* has even symmetry around mean value i.e.,

$$f_X(m-\sigma) = f_X(m+\sigma) \quad \text{... (5.5.7)}$$

Property 3 : The area under the *PDF* curve is $1/2$ for all values of x below mean value and $1/2$ for all values of x above mean value. i.e.,

$$P(X \leq m) = P(X > m) = \frac{1}{2} \quad \text{... (5.5.8)}$$

Property 4 : As $\sigma \to 0$ the Gaussian function approaches to δ (i.e. impulse) function located at $x = m$. This is because the area under the *PDF* curve is always unity. And the area of impulse function is also unity.

Significance : The Gaussian distribution is used for continuous random variables. The random motion of the thermally agitated electrons produces thermal noise. This thermal noise has Gaussian distribution. The random errors in the experimental measurements cause the measured values to have Gaussian distribution about the true value.

Example 5.5.2 *Find out the CDF of the Gaussian random variable.*

Solution : $F_X(x) = \int_{-\infty}^{x} f_X(x)\, dx$

Putting the value of $f_X(x)$ from equation (5.5.5) in above equation,

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}\, dx \quad \text{... (5.5.9)}$$

Put $\frac{m-x}{\sigma\sqrt{2}} = z$ in the above equations

$\therefore \quad -\frac{dx}{\sigma\sqrt{2}} = dz \quad \Rightarrow \quad dx = -\sigma\sqrt{2}\, dz$

These limits will be,

as $x \to -\infty$, $z \to +\infty$. and as $x \to x$, $z \to \frac{m-x}{\sigma\sqrt{2}}$

Putting these values in equation (5.6.5) we get,

$$F_X(x) = \int_{\infty}^{\frac{m-x}{\sigma\sqrt{2}}} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-z^2} \cdot (-\sigma\sqrt{2}\, dz) = -\frac{1}{\sqrt{\pi}} \int_{\infty}^{\frac{m-x}{\sigma\sqrt{2}}} \cdot e^{-z^2} \cdot dz$$

$$= \frac{1}{\sqrt{\pi}} \int_{\frac{m-x}{\sigma\sqrt{2}}}^{\infty} \cdot e^{-z^2} \cdot dz \quad \text{By interchanging the limits.}$$

$$= \frac{1}{2} \cdot \frac{2}{\sqrt{\pi}} \int_{\frac{m-x}{\sigma\sqrt{2}}}^{\infty} \cdot e^{-z^2} .\, dz \quad \text{By rearranging} \quad \text{... (5.5.10)}$$

The above integration is represented by error function. It is given as,

$$erfc\,(u) = \frac{2}{\sqrt{\pi}} \int_{u}^{\infty} e^{-z^2}\, dz \qquad \text{... (5.5.11)}$$

$$\textbf{Gaussian CDF : } F_X(x) = \frac{1}{2}\, erfc\left(\frac{m-x}{\sigma\sqrt{2}}\right) \qquad \text{... (5.5.12)}$$

Example 5.5.3 *A Gaussian distributed random variable has PDF given as follows :*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-m)^2/2\sigma^2}$$

Prove that the area under the Gaussian PDF curve defined by above equation is equal to 1.

Solution : We have to prove that

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1 \qquad \text{... (5.5.13)}$$

Let us represent the above integral by 'I' i.e.,

$$I = \int_{-\infty}^{\infty} f_X(x)\, dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-m)^2/2\sigma^2}\, dx$$

Putting value of $f_X(x)$ ... (5.5.14)

Put $\frac{x-m}{\sigma} = y$ in the above relation.

we have, $dx = \sigma\, dy$

And limits will be,

as $x \to -\infty$, $y \to -\infty$ and as $x \to +\infty$, $y \to +\infty$

With these values equation (5.5.14) becomes,

$$I = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-y^2/2} \cdot \sigma\, dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\, dy \qquad \text{...(5.5.15)}$$

Let us make the square of the above integration i.e.,

$$I \cdot I = \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\, dy\right) \times \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\, dy\right)$$

In the above equation, if we change the variable from y to some other variable say x, it will not change value of integration i.e.,

$$I^2 = \underbrace{\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dy\right)}_{\substack{\text{Variable is changed from} \\ \text{y to x in this term. It will} \\ \text{not change value of integration}}} \times \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\, dy\right)$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\, dx\, dy \qquad \text{... (5.5.16)}$$

Now let us change the variables to polar co-ordinates. i.e.,

$$x^2 + y^2 = r^2 \quad \text{and} \quad \phi = \tan^{-1}\left(\frac{y}{x}\right)$$

And $dx\, dy = r\ dr\ d\phi$

And limits are : r varies from 0 to ∞

and ϕ varies from 0 to 2π

With this conversion equation (5.6.12) becomes,

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2}\ r\ dr\ d\phi = \frac{1}{2\pi} \int_0^{2\pi} d\phi \int_0^{\infty} e^{-r^2/2}\, r\ dr \qquad \text{... (5.5.17)}$$

Put $\dfrac{r^2}{2} = t \qquad \Rightarrow \qquad r\ dr = dt$

And limits will be : as $r \to 0,\ t \to 0$

And limits are : as $r \to \infty,\ t \to \infty$

With this substitutions above equation will be,

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} d\phi \int_0^{\infty} e^{-t}\ dt = \frac{1}{2\pi}\left\{ [\phi]_0^{2\pi} \left[\frac{e^{-t}}{-1}\right]_0^{\infty} \right\}$$

$$= \frac{1}{2\pi}\left\{[2\pi - 0] \cdot [e^{-\infty} + e^0]\right\} = \frac{1}{2\pi} \times 2\pi \times 1 = 1$$

Thus $I^2 = 1$, therefore $\sqrt{I^2} = \sqrt{1}$ By taking root on both sides.

$$I = 1.$$

$$\boxed{\textbf{Area under Gaussian PDF is unity : } I = \int_{-\infty}^{\infty} f_X(x)\ dx = 1} \qquad \text{... (5.5.18)}$$

Example 5.5.4 *For the Gaussian distribution, where PDF is given as,*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}$$

Prove that i) mean $(m_x) = m$ and ii) variance $(\sigma_x^2) = \sigma^2$

Solution : i) To find mean value

The mean value of a continuous random variable is given as,

$$m_x = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-(x-m)^2/2\sigma^2}\, dx \qquad \text{... (5.5.19)}$$

Put $\frac{x-m}{\sigma} = y \quad \therefore x = \sigma y + m \quad \Rightarrow \quad dx = \sigma\, dy$

And limits will be $(-\infty, \infty)$ for y. With these substitutions above equation will be,

$$m_x = \frac{1}{2\pi} \int_{-\infty}^{\infty} (\sigma y + m)\; e^{-y^2/2} dy$$

$$= \underbrace{\frac{\sigma}{2\pi} \int_{-\infty}^{\infty} y\; e^{-y^2/2} dy}_{\text{This term will be `zero' since integer and is odd function and integration is evaluated over symmetrical limits}} + \frac{m}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y\; e^{-y^2/2} dy$$

$$= 0 + m \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} y\; e^{-y^2/2} dy}_{\text{This is the integration of Gaussian PDF. Its value is equal to '1' as obtained by equation 6.4.28.}} = m$$

That is,

$$\boxed{\textbf{Mean value of Gaussian distribution} = m_x = m} \qquad \text{... (5.5.20)}$$

II) To find variance

Variance is given as,

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - m_x)^2\; f_X(x)\; dx = \int_{-\infty}^{\infty} (x - m_x)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} dx$$

Since $m_x = m$, the above equation becomes,

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - m_x)^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} \, dx$$

Put $\dfrac{x - m}{\sqrt{2}\,\sigma} = z \qquad \Rightarrow \qquad dx = \sigma\sqrt{2}\, dz$

And integration limits will be $-\infty$ to ∞. Then above equation becomes,

$$\sigma_x^2 = \int_{-\infty}^{\infty} 2\sigma^2 z^2 \cdot \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-z^2} \cdot \sigma\sqrt{2}\, dz$$

$$= \int_{-\infty}^{\infty} \frac{2\sigma^2}{\sqrt{\pi}} \cdot z^2 e^{-z^2} \cdot dz = 2\int_{0}^{\infty} \frac{2\sigma^2}{\sqrt{\pi}} \cdot z^2 e^{-z^2} \cdot dz$$

Since z is even function and integration limits are symmetric around '0'.

$$= \frac{4\sigma^2}{\sqrt{\pi}} \int_0^{\infty} z^2 e^{-z^2} dz \qquad \text{... (5.5.21)}$$

Here use the standard relation given in Appendix i.e;

$$\int_0^{\infty} x^{2n} e^{-ax^2} \, dx = \frac{1 \cdot 3 \cdot 5 \ldots\ldots (2n-1)}{2^{n+1} a^n} \sqrt{\frac{\pi}{a}}$$

In equation (5.6.17), $n = 1$ and $a = 1$ then the result will be,

$$\sigma_x^2 = \frac{4\sigma^2}{\sqrt{\pi}} \cdot \frac{1}{2^2 \cdot 1} \cdot \sqrt{\frac{\pi}{1}} = \sigma^2$$

Thus,

Variance of Gaussian random variable : $\sigma_x^2 = \sigma^2$ ... (5.5.22)

Example 5.5.5 *A random noise voltage X is known to be Gaussian with mean $m_x = 10$ and variance $\sigma^2 = 400$. Find the probability that it :*

i) Exceeds 20 V ii) Falls between 10 V and 20 V

iii) Falls between 0 V and 20 V iv) Exceeds 0 V

v) Falls below 20 V.

You can use the Q-function defined as :

$$Q(z) = \frac{1}{\sqrt{2\pi}} \int_z^{\infty} e^{-\lambda^2/2} \, d\lambda$$

and Q(1) = 0.158 ; Q(0.5) = 0.31 ; Q(0) = 0.5.

Solution : Here $m_x = 10$ and $\sigma^2 = 400$

Gaussian pdf is given as,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}$$

Putting values of $m = m_x = 10$ and $\sigma = 20$,

$$f_X(x) = \frac{1}{20\sqrt{2\pi}} e^{-(x-10)^2/800}$$

Probability that value of 'X' lies between x_1 and x_2 is given as,

$$P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f_X(x)\, dx \qquad \text{... (5.5.23)}$$

i) $P(X > 20)$ **:** Above probability can be written as $P(20 < X \leq \infty)$. Putting values in equation (5.6.19),

$$P(X > 20) = P(20 < X \leq \infty) = \int_{20}^{\infty} \frac{1}{20\sqrt{2\pi}} e^{-(x-10)^2/800}\, dx \qquad \text{... (5.5.24)}$$

Put $\frac{x-10}{20} = z \quad \therefore \quad dx = 20\, dz$

and when $x = 20$, $z = 0.5$ Similarly when $x = \infty$, $z = \infty$

Putting these values in equation (5.6.20),

$$P(X > 20) = \int_{0.5}^{\infty} \frac{1}{20\sqrt{2\pi}} e^{-z^2/2} \cdot 20\, dz = \frac{1}{\sqrt{2\pi}} \int_{0.5}^{\infty} e^{-z^2/2}\, dz$$

We know that $Q(u) = \frac{1}{2\pi} \int_{u}^{\infty} e^{-z^2/2}\, dz$. Then above equation becomes,

$$P(X > 20) = Q(0.5) = 0.31 \text{ given.}$$

Thus $P(X > 20) = 0.31$

ii) $P(10 < X \leq 20)$ **:** Putting values in equation (5.619),

$$P(10 < X \leq 20) = P(X > 10) - P(X > 20) = P(10 < X \leq \infty) - 0.31$$

$$= \int_{10}^{\infty} \frac{1}{20\sqrt{2\pi}} e^{-(x-10)^2/800}\, dx - 0.31$$

Put $\frac{x-10}{20} = z \quad \therefore \ dx = 20\,dz$

When $x = 10$, $z = 0$ and when $x = \infty$, $z = \infty$

$$\therefore \ P(10 < X \le 20) = \int_0^{\infty} \frac{1}{20\sqrt{2\pi}} e^{-z^2/2}\, 20\,dz - 0.31 = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-z^2/2}\, dz - 0.31$$

Here the integration term is $Q(0)$, since $Q(u) = \frac{1}{\sqrt{2\pi}} \int_u^{\infty} e^{-z^2/2}\, dz$. Hence above equation will be,

$$P(10 < X \le 20) = Q(0) - 0.31 = 0.5 - 0.31 \quad \text{since } Q(0) = 0.5 \text{ given}$$

$$= 0.19$$

iii) ***P* (0 < *X* ≤ 20) :** Fig. 5.5.4 shows the plot of given Gaussian pdf. If has the mean value of $m = 10$. We have obtained $P(10 < X \le 20)$. Due to symmetry of the pdf curve, $P(0 < X \le 20)$ will be twice of $P(10 < X \le 20)$. Thus,

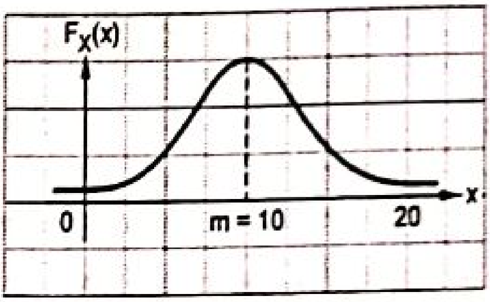


Fig. 5.5.4 Sketch of pdf

$$P(0 < X \le 20) = 2 \times P(10 < X \le 20)$$

$$= 2 \times 0.19 = 0.38$$

iv) ***P* (*X* > 0) :** From the pdf curve of Fig. 5.5.4,

$$P(X > 0) = P(0 < X \le 20) + P(X > 20) = 0.38 + 0.31 = 0.69$$

v) ***P* (*X* < 20) :** From the pdf curve of Fig. 5.5.4,

$$P(X < 20) = 1 - P(X > 20) = 1 - 0.31 = 0.69$$

5.6 Multiple Random Variables

5.6.1 Joint Distribution Function

- A joint probability density function for the continuous random variables X and Y, denoted as $f_{XY}(x, y)$, satisfies the following properties :

1. $f_{XY}(x, y) \ge 0$ for all x, y.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx\, dy = 1$
3. for any range R of two-dimensional space.

$$P([X, Y] \in R) = \int_R \int f_{XY}(x, y)\, dxdy$$

- The probability that (X, Y) assumes a value in the region R equals the volume of the shaded region.

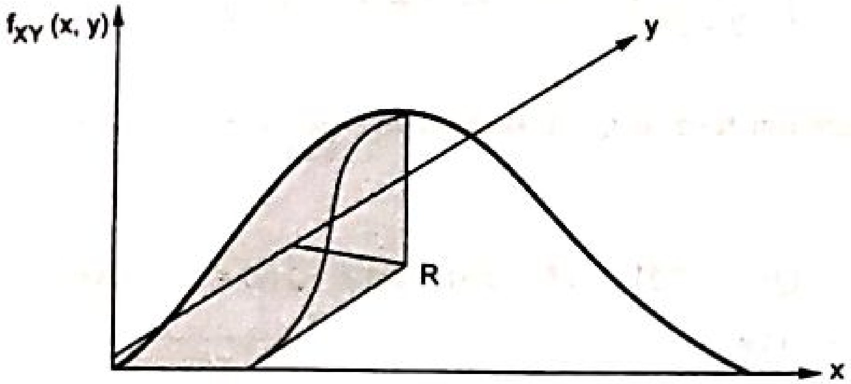


Fig. 5.6.1 Region R

- In general, if X and Y are two random variables, the probability distribution that defines their simultaneous behaviour is called a **joint probability distribution.**
- If X and Y are discrete, this distribution can be described with a **joint probability mass function.**
- If X and Y are continuous, this distribution for X and Y, we can obtain the individual probability distribution for X or for Y (and these are called the Marginal probability distributions).
- The individual probability distribution of a random variable is referred to as its marginal probability distribution.

5.6.2 Joint Probability Mass Function

The joint probability mass function of the discrete random variables X and Y, denoted as $f_{XY}(x, y)$, satisfies.

1) $f_{XY}(x, y) \geq 0$

2) $\sum_x \sum_y f_{XY}(x, y) = 1$

3) $f_{XY}(x, y) = P(X = x, Y = y)$

5.6.3 Joint Probability Density Function

- Let X and Y be continuous random variables. Then f(x, y) is a joint probability density function for X and Y if for any two-dimentional set A.

$$P[(XY) \in A] = \int_A \int f(x\, y)\, dx\, dy$$

- If A is two dimentional rectangle $\{(x, y) : a \le x \le b, c \le y \le d\}$,

$$P[(XY) \in A] = \int_a^b \int_c^d f(x\ y)\, dx\, dy$$

$P[(XY) \in A]$ = Volume under density surface above A.

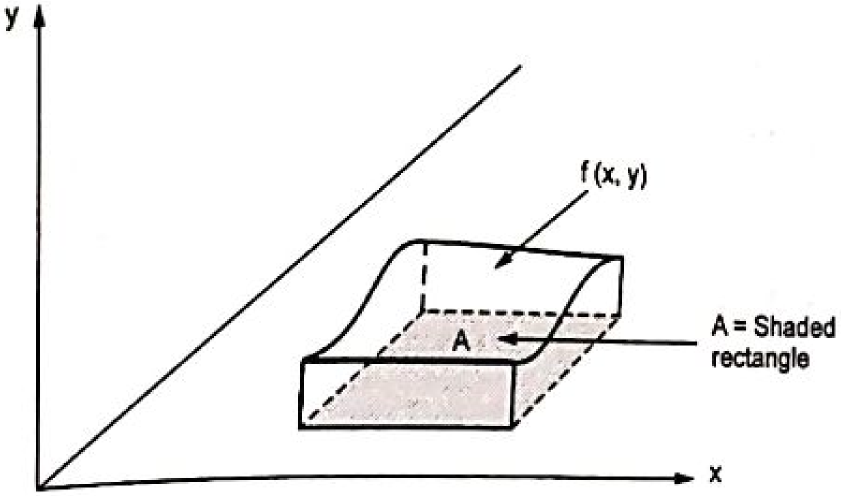


Fig. 5.6.2

Example 5.6.1 *X and Y are jointly continuous with joint pdf*

$$f(x,y) = \begin{cases} Cx^2 + \dfrac{xy}{3}, & 0 \le x \le 1, 0 \le y \le 2 \\ 0, & \textit{otherwise} \end{cases}$$

i) Find C

ii) Find marginal pdf of X and of Y.

iii) Find $P(X+Y \ge 1)$

Solution :

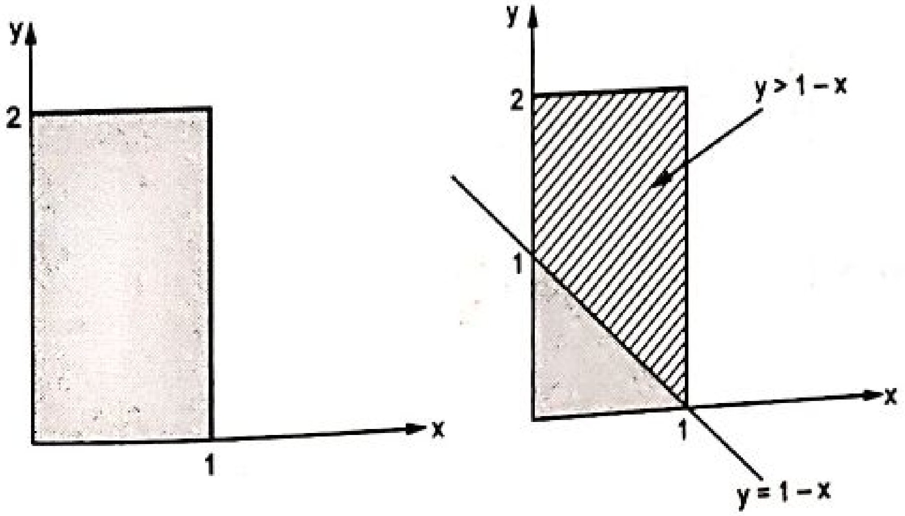


Fig. 5.6.3

i)
$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy$$

$$= \int_{0}^{1}\int_{0}^{2}\left(Cx^2 + \frac{xy}{3}\right)dx\,dy$$

$$1 = \frac{2C}{3} + \frac{1}{3}$$

$$1 = \frac{2C+1}{3}$$

$$2C + 1 = 3$$

$$2C = 3 - 1$$

$$C = \frac{2}{2}$$

$$\boxed{C = 1}$$

ii) Marginal pdf

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy$$

$$= \begin{cases} \int_{0}^{2}\left(x^2 + \frac{xy}{3}\right)dy = 2x^2 + \frac{2x}{3} \\ 0 \end{cases}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx$$

$$= \begin{cases} \int_{0}^{1}\left(x^2 + \frac{xy}{3}\right)dx = \frac{1}{3} + \frac{y}{6} \\ 0 \end{cases}$$

iii) $P(X + Y \geq 1) = \int_{0}^{1}\int_{1-x}^{2}\left(x^2 + \frac{xy}{3}\right)dy\,dx = \frac{65}{72}$

Example 5.6.2 *Compute the covariance between X and Y for following joint probability density function.*

$$f_{XY}(x, y) = \begin{cases} \frac{1}{3} y \exp(-xy), & \text{if } x \in (0, \infty) \text{ and } y \in (1, 4) \\ 0, & \text{otherwise} \end{cases}$$

$$R_{XY} = (0, \infty) \times (1, 4)$$

Solution : Given data :

$$R_Y = (1, 4)$$

Marginal probability density function of Y is :

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx$$

$$= \int_{0}^{\infty} \frac{1}{3} y \exp(-xy)\, dx = \frac{1}{3} [-\exp(-xy)]_0^{\infty}$$

$$= \frac{1}{3} [0 - (-1)] = \frac{1}{3}$$

then, $$f_Y(y) = \begin{cases} \frac{1}{3}, & \text{if } y \in (1, 4) \\ 0, & \text{otherwise} \end{cases}$$

Expected value of Y :

$$E[Y] = \int_{-\infty}^{\infty} y\, f_Y(y)\, dy = \int_{1}^{4} y \frac{1}{3} dy$$

$$= \left[\frac{1}{6} y^2\right]_1^4 = \frac{1}{6} [(4)^2 - (1)^2] = \frac{1}{6} [16 - 1] = \frac{15}{6}$$

$$E(Y) = \frac{5}{2}$$

Support of X is :

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy = \int_{1}^{4} \frac{1}{3} y \exp(-xy)\, dy$$

$$f_X(x) = \begin{cases} \int_{1}^{4} \frac{1}{3} y \exp(-xy) & \text{if } x \in (0, \infty) \\ 0, & \text{otherwise} \end{cases}$$

then $$E(X) = \int_{-\infty}^{\infty} f_X(x)\, dx$$

$$= \int_{0}^{\infty} x \left[\int_{1}^{4} \frac{1}{3} y \exp(-xy)\, dy \right] dx$$

$$= \frac{1}{3} \int_{1}^{4} \left(\int_{0}^{\infty} xy \exp(-xy)\, dx \right) dy$$

$$= \frac{1}{3} \int_{1}^{4} \left(\frac{1}{y} \int_{0}^{\infty} t \exp(-t)\, dt \right) dy \qquad [\because t = xy]$$

$$= \frac{1}{3} \int_{1}^{4} \frac{1}{y} \left([-t \exp(-t)]_0^{\infty} + \int_{0}^{\infty} \exp(-t)\, dt \right) dy$$

$$= \frac{1}{3} \int_{1}^{4} \frac{1}{y} \left(0 + [-\exp(-t)]_0^{\infty} \right) dy$$

$$= \frac{1}{3} \int_{1}^{4} \frac{1}{y}\, dy = \frac{1}{3} [\ln(y)]_1^4$$

$$E(X) = \frac{1}{3} \ln(4)$$

Expected value of E[XY] :

$$E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f_{XY}(xy)\, dy\, dx$$

$$= \int_{0}^{\infty} \left(\int_{1}^{4} xy \frac{1}{3} y \exp(-xy)\, dy \right) dx$$

$$= \frac{1}{3} \int_{1}^{4} y \left(\int_{0}^{\infty} xy \exp(-xy)\, dx \right) dy$$

$$= \frac{1}{3} \int_{1}^{4} y \left(\frac{1}{y} \int_{0}^{\infty} t \exp(-t)\, dt \right) dy$$

$$= \frac{1}{3} \int_{1}^{4} \left([-t \exp(-t)]_0^{\infty} + \int_{0}^{\infty} \exp(-t)\, dt \right) dy$$

$$= \frac{1}{3} \int_{1}^{4} (0 + [-\exp(-t)]_0^{\infty}\, dy = \frac{1}{3} \int_{1}^{4} dy = \frac{1}{3} [4-1] = \frac{3}{3}$$

$$E[XY] = 1$$

Covariance between X and Y is

$$\text{Cov}[X, Y] = E[XY] - E[X]\,E[Y]$$

$$= 1-\left(\frac{1}{3}\ln(4)\right)\left(\frac{5}{2}\right) = 1-\frac{5}{6}\ln(4)$$

Example 5.6.3 *Let X and Y be two random variables such that*

Var[X] = 4 *Cov*[X, Y] = 2

Calculate the variance for following

Cov [3X, X + 3Y]

Solution :

$$\begin{aligned}\text{Cov}[3X, X+3Y] &= 3\,\text{Cov}[X, X+3Y]\\ &= 3\,\text{Cov}[X, X] + 9\,\text{Cov}[X, Y]\\ &= 3\,\text{Var}[X] + 9\,\text{Cov}[X, Y]\\ &= 3(4) + 9(2) = 12 + 18 = 30\end{aligned}$$

Example 5.6.4 *Is the following function is a joint density function ?*

$$f(x,y) = \begin{cases} x+y, & \text{if } 0 \le x \le 1 \text{ and } 0 \le y \le 1 \\ 0, & \text{otherwise}\end{cases}$$

Solution :

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy = 1$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)\,dx\,dy$$

$$= \int_{0}^{1}\int_{0}^{1} (x+y)\,dx\,dy = \int_{0}^{1}\left(\frac{x^2}{2}+xy\Big|_0^1\right)dy$$

$$= \int_{0}^{1}\left(\left(\frac{1}{2}+y\right)-(0+0)\right)dy = \int_{0}^{1}\left(\frac{1}{2}+y\right)dy$$

$$= \frac{y}{2}+\frac{y^2}{2}\Big|_0^1$$

$$= \left(\frac{1}{2}+\frac{1}{2}\right)-\left(\frac{0}{2}+\frac{0}{2}\right) = \frac{1}{2}+\frac{1}{2} = 1$$

So, it is a joint density funtion.

Example 5.6.5 *The diameter of a metal cylinder is a random variable X with a probability density funtion given by,*

$f_X(x) = C[1-4(x-50)^2]$, $49.5 \le x \le 50.5$

Compute the value of C.

Solution :

$$1 = \int_{49.5}^{50.5} C\,[1-4\,(x-50)^2\, dx$$

$$= C\left[\int_{49.5}^{50.5} dx - 4\int_{49.5}^{50.5} (x-50)^2\, dx\right] = C\left[1-4\int_{-0.5}^{0.5} x^2 dx\right] = C\left[1-\frac{4}{3}x^3\Big|_{-0.5}^{0.5}\right]$$

$$1 = \frac{2}{3}C$$

$$C = \frac{3}{2}$$

Example 5.6.6 *Let Y have a continuous probability distribution with p.d.f. $f_Y(y) = ye^{-y}$, $0 < y < \infty$ then let X have a conditional distribution that is uniform on the interval from 0 to Y. Find the marginal density functions for X and Y.*

Solution :

$$f_X(x) = \int_{x}^{\infty} e^{-y} dy = -e^{-y}\Big|_{x}^{0} = e^{-x},\ 0 < x < \infty$$

$$f_Y(y) = \int_{0}^{y} e^{-y} dx = ye^{-y},\ 0 < y < \infty$$

Example 5.6.7 *The joint density function of X and Y is*

$$f(x,y) = \begin{cases} x+y, & 0<x<1, 0<y<1 \\ 0, & \text{otherwise} \end{cases}$$

find $P(X+Y<1)$

Solution :

$$P(X+Y<1) = \int_0^1 \int_0^{1-x} (x+y)\,dy\,dx$$

$$= \int_0^1 xy + \frac{y^2}{2}\Big|_{y=0}^{y=1-x} dx$$

$$= \frac{1}{2}\int_0^1 (1-x^2)\,dx = \frac{1}{2}\left(x - \frac{x^3}{3}\Big|_0^1\right)$$

$$= \frac{1}{2}\left(1-\frac{1}{3}\right) = \frac{1}{2}\left(\frac{2}{3}\right) = \frac{1}{3}$$

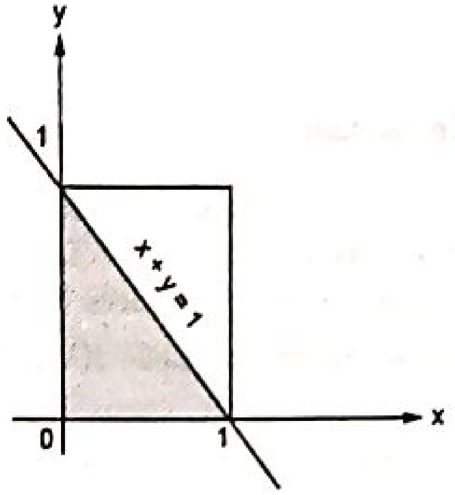


Fig. 5.6.4

Example 5.6.8 *Compute Var (X) when X is roll of a fair die outcome.*

Solution : For fair die

$$P\{X=i\} = \frac{1}{6}$$

where $i = 1, 2, 3, 4, 5, 6$

$$E(X^2) = \sum_{i=1}^{6} i^2 P\{X=i\}$$

$$= \frac{1}{6}[(1)^2+(2)^2+(3)^2+(4)^2+(5)^2+(6)^2]$$

$$= \frac{[1+4+9+16+25+36]}{6}$$

$$E(X^2) = \frac{91}{6}$$

$$E(X) = 1\left(\frac{1}{6}\right)+2\left(\frac{1}{6}\right)+3\left(\frac{1}{6}\right)+4\left(\frac{1}{6}\right)+5\left(\frac{1}{6}\right)+6\left(\frac{1}{6}\right)$$

$$= \frac{1}{6}+\frac{2}{6}+\frac{3}{6}+\frac{4}{6}+\frac{5}{6}+\frac{6}{6} = \frac{21}{6}$$

$$E(X) = \frac{7}{2}$$

$$\text{Var}(X) = E(X^2)-(E[X])^2$$

$$= \frac{91}{6}-\left(\frac{7}{2}\right)^2 = \frac{91}{6}-\frac{49}{4} = \frac{364-294}{24} = \frac{70}{24}$$

$$\text{Var}(X) = \frac{35}{12}$$

5.6.4 Covariance and Correlation

- Covariance is a measure of association between two random variables. It is positive if the deviations of the two variables from their respective means tend to have the same sign and negative if the deviations tend to have opposite signs.
- The covariance between two random variables X and Y, denoted by Cov[X, Y] is defined as follows :

$$\text{Cov}[X, Y] = E[(X - E[X])\ (Y - E[Y])]$$

or

$$\text{Cov}[X, Y] = E[(X-\mu_x)(Y-\mu_y)]$$

- Covariance indicates how two variables are related. A positive covariance means the variables are positively related, while a negative covariance means the variables are inversely related. The formula for calculating covariance of sample data is shown below.

$$\text{Cov}[X, Y] = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1}$$

Where x = The independent variable

y = The dependent variable

n = Number of data points in the sample

$\overline{X}$ = The mean of the independent variable x

$$\bar{Y} = \text{The mean of dependent variable y}$$

$$\begin{aligned}\text{Cov}[X, Y] &= E(XY - \mu_x Y - \mu_y X + \mu_x \mu_y) \\ &= E(XY) - \mu_x E[Y] - \mu_y E[X] + \mu_x \mu_y \\ &= E(XY) - \mu_x \mu_y - \mu_y \mu_x + \mu_x \mu_y \\ &= E[XY] - E[X]\, E[Y]\end{aligned}$$

Correlation :

- When one measurement is made on each observation, uni-variate analysis is applied. If more than one measurement is made on each observation, multivariate analysis is applied. Here we focus on bivariate analysis, where exactly two measurements are made on each observation.
- The two measurements will be called X and Y. Since X and Y are obtained for each observation, the data for one observation is the pair (X, Y).
- Some examples :
  1. Height (X) and weight (Y) are measured for each individual in a sample.
  2. Stock market valuation (X) and quarterly corporate earnings (Y) are recorded for each company in a sample.
- A **positive correlation** is where the two variables react in the same way, increasing or decreasing together. Temperature in Celsius and Fahrenheit has a positive correlation.
- The term "correlation" refers to a measure of the strength of association between two variables.
- **Covariance** is the extent to which a change in one variable corresponds systematically to a change in another. Correlation can be thought of as a standardized covariance.
- The correlation coefficient r is a function of the data, so it really should be called the sample correlation coefficient. The (sample) correlation coefficient r estimates the population correlation coefficient ρ.
- If either the X_i or the Y_i values are constant (i.e. all have the same value), then one of the sample standard deviations is zero, and therefore the correlation coefficient is not defined.

5.7 Central Limit Theorem

- The sampling distribution of the sample mean, $\bar{x}$ is approximated by a normal distribution when the sample is a simple random sample and the sample size, n, is large.

- In this case, the mean of the sampling distribution is the population mean, μ, and the standard deviation of the sampling distribution is the population standard deviation, σ, divided by the square root of the sample size. The latter is referred to as the **standard error** of the mean.
- A sample size of 100 or more elements is generally considered sufficient to permit using the CLT. If the population from which the sample is drawn is symmetrically distributed, $n > 30$ may be sufficient to use the CLT.
- The central limit theorem states that the mean of the sampling distribution of the mean will be the unknown population mean. The standard deviation of the sampling distribution of the mean is called the standard error. In fact, it is just another standard deviation, we just call it the standard error so we know we're talking about the standard deviation of the sample means instead of the standard deviation of the raw data. The standard deviation of data is the average distance values are from the mean.

5.8 Sampling Distributions

5.8.1 Population

- A population is any entire collection of people, animals, plants or things from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.
- Population is a collection of objects. It may be finite or infinite according to the number of objects in the population.
- A population can be defined as including all people or items with the characteristic one wishes to understand. Because there is very rarely enough time or money to gather information from everyone or everything in a population, the goal becomes finding a representative sample (or subset) of that population.
- In order to make any generalizations about a population, a sample, that is meant to be representative of the population, is often studied. For each population there are many possible samples. A sample statistic gives information about a corresponding population parameter. For example, the sample mean for a set of data would give information about the overall population mean.
- It is important that the investigator carefully and completely defines the population before collecting the sample, including a description of the members to be included.
- **Example :** The population for a study of infant health might be all children born in the UK in the 1980's. The sample might be all babies born on 7th May in any of the years.

- When such measures like the mean, median, mode, variance and standard deviation of a population distribution are computed, they are referred to as parameters. A parameter can be simply defined as a summary characteristic of a population distribution.

5.8.2 Sample

- A sample is a group of units selected from a larger group (the population). By studying the sample it is hoped to draw valid conclusions about the larger group.
- A sample is a subset of a population. Sample is a smaller group, the part of the population of interest that we actually examine in order to gather the information.
- A sample is "a smaller collection of units from a population used to determine truths about that population".
- A sample is generally selected for study because the population is too large to study in its entirety. The sample should be representative of the general population. This is often best achieved by random sampling. Also, before collecting the sample, it is important that the researcher carefully and completely defines the population, including a description of the members to be included.
- **Example :** The population for a study of infant health might be all children born in the UK in the 1980's. The sample might be all babies born on 7th May in any of the years.
- **Symbols for population and sample descriptive measures**

Parameter	Population	Sample
Mean	M	X
Variance	σ^2	var
Standard deviation	σ	sd

5.8.3 Types of Sampling

- Two general approaches to sampling are used.
- **Probability (Random) Samples**

  a. Simple random sample

  b. Systematic random sample

  c. Stratified random sample

  d. Multistage sample

  e. Multiphase sample

  f. Cluster sample

- **Non-Probability Samples**
  1. Convenience sample
  2. Purposive sample
  3. Quota
- With *probability sampling*, all elements (e.g., persons, households) in the population have some opportunity of being included in the sample and the mathematical probability that any one of them will be selected can be calculated.
- With *nonprobability sampling*, in contrast, population elements are selected on the basis of their availability or because of the researcher's personal judgment that they are representative. The consequence is that an unknown portion of the population is excluded. One of the most common types of non-probability sample is called a *convenience* sample.
- Any sampling method where some elements of population have *no* chance of selection, or where the probability of selection can't be accurately determined. It involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is non-random, non-probability sampling not allows the estimation of sampling errors.

1.Random sampling :

- Applicable when population is small, homogeneous and readily available.
- All subsets of the frame are given an equal probability. Each element of the frame thus has an equal probability of selection.
- It provides for greatest number of possible samples. This is done by assigning a number to each unit in the sampling frame.
- A table of random number or lottery system is used to determine which units are to be selected.
- Estimates are easy to calculate.
- Simple random sampling is always an EPS design, but not all EPS designs are simple random sampling.

Disadvantages

- If sampling frame large, this method impracticable.
- Minority subgroups of interest in population may not be present in sample in sufficient numbers for study.

2. Stratified sampling

- Where population embraces a number of distinct categories, the frame can be organized into separate "strata." Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected.
- Every unit in a stratum has same chance of being selected.
- Using same sampling fraction for all strata ensures proportionate representation in the sample.
- Adequate representation of minority subgroups of interest can be ensured by stratification and varying sampling fraction between strata as required.
- Finally, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata.

Drawbacks to using stratified sampling.

- First, sampling frame of entire population has to be prepared separately for each stratum.
- Second, when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design and potentially reducing the utility of the strata.
- Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods.

Some terms used in sampling

1. **Sampled population** - Population from which sample drawn.
2. **Frame** - List of elements that sample selected from. E.g. telephone book, city business directory. May be able to construct a frame.
3. **Parameter** - Characteristics of a population. E.g. Total (annual GDP or exports), proportion p of population that votes Liberal in federal election. Also μ or σ of a probability distribution is termed parameters.
4. **Statistic** - Numerical characteristics of a sample. E.g. monthly unemployment rate, pre-election polls.
5. **Sampling distribution** of a statistic is the probability distribution of the statistic.

Selecting a sample

1. ***N*** is the symbol given for the size of the population or the number of elements in the population.
2. ***n*** is the symbol given for the size of the sample or the number of elements in the sample.

3. **Simple random sample** is a sample of size n selected in a manner that each possible sample of size n has the same probability of being selected.

4. In the case of a random sample of size n = 1, each element has the same chance of being selected.

- **The sampling process comprises several stages :**
  1. Defining the population of concern.
  2. Specifying a sampling frame, a set of items or events possible to measure.
  3. Specifying a sampling method for selecting items or events from the frame.
  4. Determining the sample size.
  5. Implementing the sampling plan.
  6. Sampling and data collecting.
  7. Reviewing the sampling process.

Selecting a simple random sample

- **Sample with replacement** - After any element randomly selected, replace it and randomly select another element. But this could lead to the same element being selected more than once.
- More common to **sample without replacement.** Make sure that on each stage, each element remaining in the population has the same probability of being selected.
- Use a random number table or a computer generated random selection process. Or use a coin, die or bingo ball popper, etc.
- **Simple random sample of size 2 from a population of 4 elements - without replacement**
  1. Population elements are A, B, C, D then N = 4 and n = 2.
  2. The first element selected could be any one of the 4 elements and this leaves 3, so there are 4 × 3 = 12 possible samples, each equally likely : AB, AC, AD, BA, BC, BD, CA, CB, CD, DA, DB, DC.

$$P_n^N = \frac{N!}{(N-n)!} = \frac{4!}{(4-2)!} = 12$$

  3. If the order of selection does not matter (i.e. we are interested only in what elements are selected), then this reduces to 6 combinations. If {AB} is AB or BA, etc., then the equally likely random samples are {AB}, {AC}, {AD}, {BC}, {BD}, {CD}. This is the number of combinations.

$$C_n^N = \frac{N!}{n!(N-n)!} = \frac{4!}{2!(4-2)!} = 6$$

5.8.4 Sampling Distribution of the Mean

- A theoretical probability distribution of sample means that would be obtained by drawing from the population all possible samples of the same size.
- The standard deviation of the sampling distribution is called the standard error.
- The **sampling error** is the difference between the point estimate (value of the estimator) and the value of the parameter. This is the error caused by sampling only a subset of elements of a population, rather than all elements in a population. A researcher hopes to minimize the sampling error, but all samples have some such error associated with them.
- The sample is a sampling distribution of the sample means. When all of the possible sample means are computed, then the following properties are true :
  1. The mean of the sample means will be the mean of the population
  2. The variance of the sample means will be the variance of the population divided by the sample size.
  3. The standard deviation of the sample means (known as the standard error of the mean) will be smaller than the population mean and will be equal to the standard deviation of the population divided by the square root of the sample size.
  4. If the population has a normal distribution, then the sample means will have a normal distribution.
  5. If the population is not normally distributed, but the sample size is sufficiently large, then the sample means will have an approximately normal distribution. Some books define sufficiently large as at least 30 and others as at least 31.

The formula for a Z-score when working with the sample means is :

$$Z = \frac{\overline{X}-\mu}{\sigma / \sqrt{n}}$$

Finite Population Correction Factor

- If the sample size is more than 5 % of the population size and the sampling is done without replacement, then a correction needs to be made to the standard error of the means.
- In the following, N is the population size and n is the sample size. The adjustment is to multiply the standard error by the square root of the quotient of the difference between the population and sample sizes and one less than the population size.

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$$

Random sample from a normally distributed population

	Normally distributed population	Sampling distribution of $\bar{x}$ when sample is random
Number of elements	N	n
Mean	μ	μ
Standard deviation	σ	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Classification of Samples

- Samples are classified as two types : Large sample and Small sample.
  1. Large sample : The sample is said to be large if the size of sample ($n \geq 30$).
  2. Small sample : The sample is said to be large if the size of sample ($n < 30$).

5.8.5 Mean, Medium and Mode

1. Mean : The mean of a data set is the average of all the data values. The sample mean $\bar{x}$ is the point estimator of the population mean μ.

$$\text{Sample mean } \bar{x} = \frac{\text{Sum of the values of the n observations}}{\text{Number of observations in the sample}} = \frac{\sum X_i}{n}$$

$$\text{Population mean } \mu = \frac{\text{Sum of the values of the N observations}}{\text{Number of observations in the population}} = \frac{\sum x_i}{N}$$

2. Median

- The median of a data set is the value in the middle when the data items are arranged in ascending order. Whenever a data set has extreme values, the median is the preferred measure of central location.
- The median is the measure of location most often reported for annual income and property value data. A few extremely large incomes or property values can inflate the mean.
- For an **odd number** of observations :

7 observations = 26, 18, 27, 12, 14, 29, 19

Numbers in ascending order = 12, 14, 18, 19, 26, 27, 29

The median is the middle value.

Median = 19

- For an **even number** of observations :

8 observations = 26, 18, 29, 12, 14, 27, 30, 19

Numbers in ascending order = 12, 14, 18, 19, 26, 27, 29, 30

The median is the average of the middle two values.

Median = (19 + 26)/2 = 22.5

3. **Mode :** The mode of a data set is the value that occurs with greatest frequency. The greatest frequency can occur at two or more different values. If the data have exactly two modes, the data are bimodal. If the data have more than two modes, the data are multimodal.

4. Range

- The range of a data set is the difference between the largest and smallest data values.
- It is the simplest measure of variability. It is very sensitive to the smallest and largest data values.

Range = Largest value – Smallest value

5. Variance

- The variance is a measure of variability that utilizes all the data. It is based on the difference between the value of each observation (x_i) and the mean ($\bar{x}$ for a sample, μ for a population).
- The variance is the average of the squared differences between each data value and the mean.
- The variance is computed as follows :

Sample variance : $S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

Population variance : $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$

6. Standard deviation

- The standard deviation of a data set is the positive square root of the variance. It is measured in the same units as the data, making it more easily interpreted than the variance.
- The standard deviation is computed as follows :

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

$$\text{Sample standard deviation} = S = \sqrt{S^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

5.8.6 Standard Error

- The standard deviation of the sampling distribution is of a statistic. Standard error is a statistical term that measures the accuracy with which a sample represents a population. In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error.
- The term "standard error" is used to refer to the standard deviation of various sample statistics such as the mean or median. For example, the "standard error of the mean" refers to the standard deviation of the distribution of sample means taken from a population.

Standard Error Calculation Procedure :

Step 1 : Calculate the mean (Total of all samples divided by the number of samples).

Step 2 : Calculate each measurement's deviation from the mean (i.e. Mean minus the individual measurement).

Step 3 : Square each deviation from mean. Squared negatives become positive.

Step 4 : Sum the squared deviations.

Step 5 : Divide that sum from step 4 by one less than the sample size (n – 1)

Step 6 : Take the square root of the number in step 5. That gives you the "Standard Deviation (S.D.)."

Step 7 : Divide the standard deviation by the square root of the sample size (n). That gives you the "standard error".

Step 8 : Subtract the standard error from the mean and record that number. Then add the standard error to the mean and record that number. You have plotted mean ± 1 standard error , the distance from 1 standard error below the mean to 1 standard error above the mean.

Let us consider the following table :

Name	Height to nearest	(Step 2) Deviations (m – i)	(Step 3) Squared deviations $(m - i)^2$
Rupali	150	9.6	92.16
Rakshita	170	– 10.4	108.16
Sangeeta	165	– 5.4	29.16
Rutuja	155	4.6	21.16

Rushi	158	1.6	2.56
n = 5	Total = 798 (Step 1) Mean m = 159.6		(Step 4) Sum of squared deviations $\sum (m-i)^2 = 253.2$

Step 5 : Divide by number of measurements – 1 :

$$\frac{\sum (m-i)^2}{n-1} = \frac{253.2}{5-1} = 63.3$$

Step 6 : Standard deviation $= \frac{\text{Square root of} \sum (m-i)^2}{n-1} = \frac{\sqrt{63.3}}{4} = 1.9890$

Step 7 : Standard error $= \frac{\text{Standard deviation}}{\sqrt{n}} = \frac{1.9890}{\sqrt{4}} = 0.9945$

Step 8 : m ± 1SE $= 159.6 \pm 0.9945$

$= 159.6 + 0.9945$ or $159.6 - 0.9945$

$= 160.5945$ or 158.6055

Example 5.8.1 *A bowler claims that she has a 215 average. In her latest performance, she scores 188, 214 and 204. Assume that her bowling scores are normally distributed. Calculate the sample mean, variance and standard deviation*

Solution : The sample mean, variance, and standard deviation

$$\text{Sample mean} = \frac{188+214+204}{3} = \frac{606}{3} = 202$$

$$\text{Sample variance} = \frac{(188-202)^2+(214-202)^2+(204-202)^2}{3-1} = \frac{196+144+4}{2} = \frac{344}{2} = 172$$

$$\text{Standard deviation} = \sqrt{172} = 13.11$$

Example 5.8.2 *The following are the times between six calls for an ambulance in a city and the patient's arrival at the hospital : 27, 15, 20, 32, 18 and 26 minutes. Use these figures to judge the reasonableness of the ambulance services claim that it takes on the average 20 minutes between the call for an ambulance and patient's arrival at the hospital.*

Solution : Given data : n = 6, Average minutes to reach the hospital (μ) = 20

$x_1 = 27,\ x_2 = 15,\ x_3 = 20,\ x_4 = 32,\ x_5 = 18,\ x_6 = 26$

Then, Arithmetic mean $\bar{x} = \frac{\sum x_i}{n}$

$$\bar{x} = \frac{x_1+x_2+x_3+x_4+x_5+x_6}{6} = \frac{27+15+20+32+18+26}{6} = \frac{138}{6} = 23$$

$$\text{Estimate of variance } (S^2) = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$S^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + (x_3-\bar{x})^2 + (x_4-\bar{x})^2 + (x_5-\bar{x})^2 + (x_6-\bar{x})^2}{6-1}$$

$$= \frac{(27-23)^2 + (15-23)^2 + (20-23)^2 + (32-23)^2 + (18-23)^2 + (26-23)^2}{5}$$

$$= \frac{16+64+9+81+25+9}{5} = \frac{204}{5} = 40.8$$

$$S^2 = 40.8$$

$$S = 6.387$$

$$t = \frac{x-\mu}{S/\sqrt{n}} = \frac{23-20}{6.387/\sqrt{6}} = \frac{3}{6.387/2.449}$$

$$t = 1.15$$

Now, $t_{n-1,\alpha} \Rightarrow t_{6-1,\alpha} \Rightarrow t_{5,\alpha} = 2.015$

(for $\alpha = 0.05$)

For $\alpha = 0.05 \quad t_5 = 2.015$

$\alpha = 0.1 \quad t_5 = 1.476$

So $t = 1.15 < 1.476$

So claim is rejected.

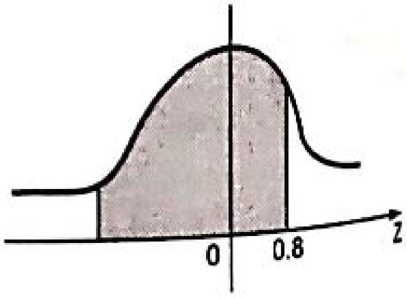
Example 5.8.3 *A normal population has a mean of 0.1 and standard deviation of 2.1. Find the probability that the mean of simple, sample of 900 members will be negative.*

Solution : Given data :

Mean of population $\mu = 0.1$

Standard deviation of the population $\sigma = 2.1$

Sample size $n = 900$



$$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{x}-0.1}{2.1/\sqrt{900}} = \frac{\bar{x}-0.1}{0.07}$$

$$Z = \frac{\bar{x}}{0.07} - 1.428$$

$\bar{x}$ is negative if $Z < -1.428$

$$P(\bar{x} < 0) = P(Z < -1.428)$$

$$= P(Z < 1.428)$$

$$= \int_0^{\infty} \phi(Z)\,dZ - \int_0^{1.428} \phi(Z)\,dZ$$

$$= 0.5 - 0.4236 = 0.0764$$

Example 5.8.4 *The mean height of the students in a college is 155 cms and standard deviation is 15. What is the probability that the mean height of 38 students is less than 157 cms ?*

Solution : Given data : Mean height of the student $\mu = 155$ cms,

Standard deviation $\sigma = 15$

Sample size $n = 36$, Mean of sample $\bar{x} = 157$ cms

Then

$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{157-155}{15/\sqrt{36}} = \frac{2}{15/6} = \frac{2}{2.5} = 0.8$$

$$P(\bar{x} \leq 157) = P(Z < 0.8)$$

$$P(Z < 0.8) = 0.5 + P(0 \leq Z \leq 0.8) = 0.5 + 0.2881 = 0.7881$$

Probability of height = 0.7881

Example 5.8.5 *A sample of size 400 is taken from a population whose standard deviation is 16. Find the standard error.*

Solution : Given data : Standard deviation of population $\sigma = 16$,

Size of the sample $n = 400$, Standard error = ?

$$\text{Standard error} = \frac{\sigma}{\sqrt{n}} = \frac{16}{\sqrt{400}} = \frac{16}{20}$$

$$\text{S.E.} = 0.8$$

Example 5.8.6 *A random sample of size 64 is taken from a normal population with $\mu = 51.4$ and $\sigma = 68$. What is the probability that the mean of the sample will :*

i) Exceed 52.9 ii) Fall between 50.5 and 52.3 iii) Be less than 50.6.

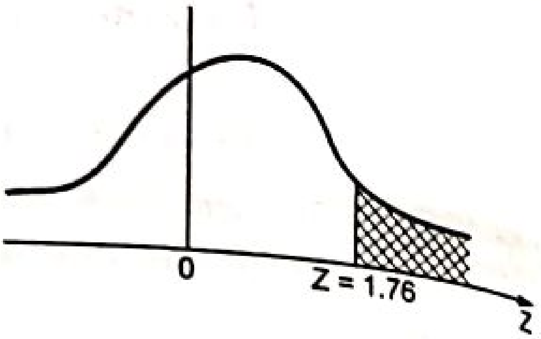
Solution : Given data : Size of the sample $n = 64$, Mean of the population $\mu = 51.4$,

Standard deviation $\sigma = 6.8$

$$\text{Standard error } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{6.8}{\sqrt{64}} = 0.85$$

i) Exceed 52.9

$P(\bar{x} \text{ exceed } 52.9) = P(\bar{x} > 52.9)$



$$Z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{52.9-51.4}{0.85} = 1.76$$

$$\begin{aligned} P(\bar{x} > 52.9) &= P(Z > 1.76) \\ &= 0.5 - P(0 < Z < 1.76) \\ &= 0.5 - 0.4608 \\ &= 0.03982 \end{aligned}$$

ii) Fall between 50.5 and 52.3

$P(50.5 < \bar{x} < 52.3) = P(\bar{x}_1 < \bar{x} < \bar{x}_2)$

$$\bar{x}_1 = 50.5 \quad \text{and} \quad \bar{x}_2 = 52.3$$

$$Z_1 = \frac{\bar{x}_1-\mu}{\sigma_{\bar{x}}} = \frac{50.5-51.4}{0.85} = \frac{-0.9}{0.85} = -1.06$$

$$Z_2 = \frac{\bar{x}_2-\mu}{\sigma_{\bar{x}}} = \frac{52.3-51.4}{0.85} = \frac{0.9}{0.85} = 1.06$$

$$\begin{aligned} P(50.5 < \bar{x} < 52.3 &= P(-1.06 < Z < 1.06) \\ &= P(-1.06 < Z < 0) + P(0 < Z < 1.06) \\ &= 0.3554 + 0.3554 = 0.7108 \end{aligned}$$

iii) BC less than 50.6

$P(\bar{x} < 50.6)$

$$Z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{50.6-51-4}{0.85} = -0.94$$

$$\begin{aligned} P(\bar{x} < 50.6) &= P(Z < -0.94) \\ &= 0.5 - P(0.94 < Z < 0) \\ &= 0.5 - 0.3264 = 0.1736 \end{aligned}$$

5.8.7 Sampling Distribution of the Mean (σ-unknown)

A population consisting of all real numbers is an example of an infinite population.

1. Arithmetic mean :

If $x_1 + x_2 + x_3 + \ldots + x_n$ are the values in a sample then the arithmetic mean is given by $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$

2. Variance :

$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

- Sampling distribution of $\overline{X}$ is normally distributed even for small samples of size $n < 30$ provided sampling is from normal population.
- When σ is unknown, it can be substituted by S.
- t-distribution with the parameter $v = n-1$ is given by

$$t = \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n}}} \quad \text{where} \quad v = \text{Degree of freedom}$$

- The standard normal distribution provide a good approximation to the t-distribution for samples of size 30 or more.

Example 5.8.7 *A random sample of size 144 is taken from an infinite population having the mean 75 and variance 225. What is probability that $\bar{x}$ will be between 72 and 77 ?*

Solution : Given data : Size of sample n = 144, Variance $\sigma^2 = 225$

$$\sigma = \sqrt{225} = 15$$

$$\text{Mean } \mu = 75$$

$$\bar{x}_1 = 72, \quad \bar{x}_2 = 77$$

$$Z = \frac{\bar{x}-\mu}{\sigma / \sqrt{n}}$$

$$Z_1 = \frac{\bar{x}_1 - m}{\sigma / \sqrt{n}} = \frac{72-75}{15 / \sqrt{144}} = \frac{-3}{15/12} = -2.4$$

$$Z_2 = \frac{\bar{x}_2 - \mu}{\sigma / \sqrt{n}} = \frac{77-75}{15 / \sqrt{144}} = \frac{2}{15/12} = 1.6$$

$$\begin{aligned} P(72 < \bar{x} < 77) &= P(\bar{x}_1 < \bar{x} < \bar{x}_2) \\ &= P(-2.4 < Z < 0) + P(0 < Z < 1.06) \\ &= P(0 < Z < 2.4) + P(0 < Z < 1.6) \\ &= 0.4918 + 0.4452 = 0.9370 \end{aligned}$$

Example 5.8.9 *A random sample of size 100 is taken from an infinite population having the mean $\mu = 76$ and the variance $\sigma^2 = 256$. What is the probability that $\bar{x}$ will be between 75 and 78 ?*

Solution : Given data : Mean $\mu = 76$

$$\text{Variance } \sigma^2 = 256$$

$$\sigma = 16$$

$$n = 100, \quad \bar{x}_1 = 75, \quad \bar{x}_2 = 78$$

$$Z_1 = \frac{\bar{x}_1 - \mu}{\sigma/\sqrt{n}} = \frac{75-76}{16/\sqrt{100}} = \frac{-1}{16/10} = -0.625$$

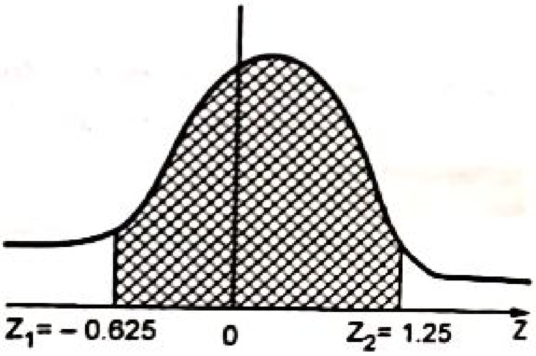
$$Z_2 = \frac{\bar{x}_2 - \mu}{\sigma/\sqrt{n}} = \frac{78-76}{16/\sqrt{100}} = \frac{2}{16/10} = 1.25$$

$$P(75 < \bar{x} < 78) = P(-0.625 < Z < 1.25)$$

$$= P(-0.625 < Z < 0) + P(0 < Z < 1.25)$$

$$= P(0 < Z < 0.625) + P(0 < Z < 1.25)$$

$$= 0.2324 + 0.3944 = 0.6268$$



Example 5.8.10 *When a sample is taken from an infinite population, what happen to the standard error of the mean if the same size is decreased from 800 to 200 ?*

Solution : Mean for standard error $= \frac{\sigma}{\sqrt{n}}$

Sample size = n

$$n_1 = 800 \text{ and } n_2 = 200$$

$$\text{Standard error } (SE_1) = \frac{\sigma}{\sqrt{n_1}} = \frac{\sigma}{\sqrt{800}} = \frac{\sigma}{\sqrt{400 \times 2}} = \frac{\sigma}{20\sqrt{2}}$$

$$\text{Standard error } (SE_2) = \frac{\sigma}{\sqrt{n_2}} = \frac{\sigma}{\sqrt{200}} = \frac{\sigma}{\sqrt{100 \times 2}} = \frac{\sigma}{10\sqrt{2}}$$

$$SE_2 = \frac{\sigma}{10\sqrt{2}} = 2\ [SE_1] = 2\left[\frac{\sigma}{20\sqrt{2}}\right]$$

If a sample size is reduced then standard error of mean will be multiplied by 2.

Example 5.8.11 *A population consists of four numbers 2, 3, 4, 5. Consider all possible distinct samples of size two with replacement find :*

a) The population mean b) The population standard deviation (s.d)

c) The sampling distribution of means d) The mean of the S.D of means

e) s.d. of S.D of means. Verify (c) and (e) directly from (a) and (b) by use of suitable formulae.

Solution : a) Population mean (μ)

$$\mu = \frac{2+3+4+5}{4} = \frac{14}{4} = 3.5$$

b) The population standard deviation

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{(2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2}{4}$$

$$= \frac{2.25+0.15+0.25+2.25}{4} = \frac{5}{4}$$

$$\sigma^2 = 1.25$$

$$\sigma = 1.118$$

c) The sampling distribution of means (Sampling with replacement)

$$N^n = (4)^2 = 16 \text{ (sample size = 2)}$$

$$N = \text{Population size}$$

$$n = \text{Sample size listing}$$

Sampling distribution is :

$$\begin{Bmatrix} (2,2),(2,3),(2,4),(2,5) \\ (3,2),(3,3),(3,4),(3,5) \\ (4,2),(4,3),(4,4),(4,5) \\ (5,2),(5,3),(5,4),(5,5) \end{Bmatrix}$$

Sample value	Total of sample values	Distribution means
2, 2	4	2
2, 3	5	2.5
2, 4	6	3
2, 5	7	3.5

3, 2	5	2.5
3, 3	6	3
3, 4	7	3.5
3, 5	8	4
4, 2	6	3
4, 3	7	3.5
4, 4	8	4
4, 5	9	4.5
5, 2	7	3.5
5, 3	8	4
5, 4	9	4.5
5, 5	10	5

$$\mu_{\bar{x}} = \frac{\text{Sum of all sample means}}{16}$$

$$= \frac{2+2.5+3+3.5+2.5+3+3.5+4+3+3.5+4+4.5+3.5+4+4.5+5}{16}$$

$$= \frac{56}{16} = 3.5$$

Considering $\mu_{\bar{x}} = \mu$

d) The mean of the S.D. of means

$$\sigma_{\bar{x}}^2 = \frac{1}{16}\begin{bmatrix}(2-3.5)^2+(2.5-3.5)^2+(3-3.5)^2+(3.5-3.5)^2 \\ +(2.5-3.5)^2+(3-3.5)^2+(3.5-3.5)^2+(4-3.5)^2 \\ +(3-3.5)^2+(3.5-3.5)^2+(4-3.5)^2+(4.5-3.5)^2 \\ +(3.5-3.5)^2+(4-3.5)^2+(4.5-3.5)^2+(5-3.5)^2\end{bmatrix}$$

$$= \frac{\begin{bmatrix}2.25+1+0.25+0+1+0.25+0+0.25+0.25 \\ +0.25+1+0+0.25+1+2.25\end{bmatrix}}{16}$$

$$= \frac{10}{16} = 0.625 = \sqrt{0.625} = 0.79$$

e) s.d. of SD mean

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{(1.118)^2}{2} = 0.6249$$

Example 5.8.12 *A population consists of six numbers 4, 8, 12, 16, 20, 24 consider all samples of size two which can be drawn without replacement from this population. Find*

i) The population mean ii) The population standard deviation

iii) The mean of the sampling distribution of means

iv) The standard deviation of the sampling distribution of means verify (iii) and (iv) from (i) and (ii) by use of suitable formulae.

Solution : i) Population mean (μ)

$$\mu = \frac{\sum x}{n} = \frac{4+8+12+16+20+24}{6} = \frac{84}{6} = 14$$

ii) Population standard deviation (σ^2)

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

Here $\bar{x} = 14$, $n = 6$

$$= \frac{1}{6}[(4-14)^2+(8-14)^2+(12-14)^2+(16-14)^2+(20-14)^2+(24-14)^2]$$

$$= \frac{1}{6}[(-10)^2+(-6)^2+(-2)^2+(2)^2+(6)^2+(10)^2]$$

$$= \frac{100+36+4+4+36+100}{6} = \frac{280}{6}$$

$$\sigma^2 = 46.66$$

iii) Mean of the sampling distribution of means

Number of samples = $^6C_2$

$$= \frac{6!}{2!(6-2)!} = \frac{720}{2! \times 4!} = \frac{720}{48} = 15$$

Sample number	Sample values	Total of sample values	Sample mean
1	4, 8	12	6
2	4, 12	16	8
3	4, 16	20	10
4	4, 20	24	12
5	4, 24	28	14

6	8, 12	20	10
7	8, 16	24	12
8	8, 20	28	14
9	8, 24	32	16
10	12, 16	28	14
11	12, 20	32	16
12	12, 24	36	18
13	16, 20	36	18
14	16, 24	40	20
15	20, 24	44	22
	Total		210

Mean of sample means $= \frac{210}{15} = 14$

The mean of sampling distribution of mean is $\mu_x = 14$

So, considering $\mu_x = \mu$

iv) Standard deviation of the sampling distribution of means

$$\sigma_{\bar{x}}^2 = \frac{1}{15}[(6-14)^2 + (8-14)^2 + (10-14)^2 + (12-14)^2 + (14-14)^2 + (10-14)^2$$

$$+ (12-14)^2 + (14-14)^2 + (16-14)^2 + (14-14)^2 + (16-14)^2 + (18-14)^2$$

$$+ (18-14)^2 + (20-14)^2 + (22-14)^2]$$

$$= \frac{1}{15}[64+36+16+4+0+16+4+0+4+0+4+16+16+36+64]$$

$$= \frac{280}{15} = 18.66$$

Standard deviation of sampling distribution of means is

$$\sigma_{\bar{x}} = \sqrt{18.66} = 4.319$$

Example 5.8.13 *A population consists of 5, 10, 14, 18, 13, 24. Consider all possible samples of size two which can be drawn without replacement from the population. Find*

a) The mean of the population b) The standard deviation of the population

c) The mean of the sampling distribution of means

d) The standard deviation of sampling distribution of means.

Solution : a) Mean of the population (μ)

$$\mu = \frac{5+10+14+18+13+24}{6} = \frac{84}{6} = 14$$

b) Standard deviation of the population

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$= \frac{[(5-14)^2 + (10-14)^2 + (14-14)^2 + (18-14)^2 + (13-14)^2 + (24-14)^2}{6}$$

$$= \frac{81+16+0+16+1+100}{6} = \frac{214}{6} = 35.6666$$

$$\sigma = 5.9721$$

c) The mean of the sampling distribution of means

Number of samples = ${}^6C_2$

$$= \frac{6!}{2!(6-2)!} = \frac{720}{48} = 15$$

Sample number	Sample values	Sample value total	Sample mean
1	5, 10	5 + 10 = 15	$\frac{15}{2} = 7.5$
2	5, 14	5 + 14 = 19	$\frac{19}{2} = 9.5$
3	5, 18	5 + 18 = 23	$\frac{23}{2} = 11.5$
4	5, 13	5 + 13 = 18	$\frac{18}{2} = 9$
5	5, 24	5 + 24 = 29	$\frac{29}{2} = 14.5$
6	10, 14	10 + 14 = 24	$\frac{24}{2} = 12$
7	10, 18	10 + 18 = 28	$\frac{28}{2} = 14$
8	10, 13	10 + 13 = 23	$\frac{23}{2} = 11.5$
9	10, 24	10 + 24 = 34	$\frac{34}{2} = 17$

10	14, 18	14 + 18 = 32	$\frac{32}{2} = 16$
11	14, 13	14 + 13 = 27	$\frac{27}{2} = 13.5$
12	14, 24	14 + 24 = 38	$\frac{38}{2} = 19$
13	18, 13	18 + 13 = 31	$\frac{31}{2} = 15.5$
14	18, 24	18 + 24 = 42	$\frac{42}{2} = 21$
15	13, 24	13 + 24 = 37	$\frac{37}{2} = 18.5$

$$= \frac{7.5+9.5+11.5+9+14.5+12+14+11.5+17+16+13.5+19+15.5+21+18.5}{15}$$

$$= \frac{210}{15} = 14$$

d) The standard deviation of sampling distribution of means $(\mu_{\bar{x}} = \mu)$

$$\sigma_{\bar{x}}^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$\sigma_{\bar{x}}^2 = \frac{1}{15}\begin{bmatrix} (7.5-14)^2 + (9.5-14)^2 + (11.5-14)^2 + (9-14)^2 \\ + (14.5-14)^2 + (12-14)^2 + (14-14)^2 + (11.5-14)^2 \\ + (17-14)^2 + (16-14)^2 + (13.5-14)^2 + (19-14)^2 \\ + (15.5-14)^2 + (21-14)^2 + (18.5-14)^2 \end{bmatrix}$$

$$= \frac{1}{15}\begin{bmatrix} 42.25+20.25+6.25+25+0.25+4+0+6.25 \\ + 9+4+0.25+25+2.25+49+20.25 \end{bmatrix} = \frac{214}{15}$$

$$\sigma_{\bar{x}}^2 = 14.2666$$

$$\sigma_{\bar{x}} = 3.777$$

5.9 Hypothesis Testing

General definition of a hypothesis : **"A hypothesis is a statement of a relationship between two or more variables"**. A statistical hypothesis is simply a particular kind of hypothesis.

- A hypothesis is a statement or claim regarding a characteristic of one or more populations. Hypothesis testing is a procedure, based on sample evidence and probability, used to test claims regarding a characteristic of one or more populations.

- The null hypothesis, denoted H_0 (read "H-naught"), is a statement to be tested. The null hypothesis is assumed true until evidence indicates otherwise. The alternative hypothesis, denoted H_1 (read "H-one"), is a claim to be tested. We are trying to find evidence for the alternative hypothesis.
- A statistical hypothesis is either
  1. A statement about the value of a population parameter (e.g., mean, median, mode, variance, standard deviation, proportion, total) or
  2. A statement about the kind of probability distribution that a certain variable obeys.
- Examples of statistical hypothesis :
  a. The mean age of all college students is 20.4 years. **(simple hypothesis)**
  b. The proportion of college students two are men is 60 %. **(simple hypothesis)**
  c. The proportion of books in the college library whose heights exceed 30 cm is less than or equal to 0.13. **(Composite hypothesis)**
- A statistical hypothesis that specifies a single value for a population parameter is called a simple hypothesis; every statistical hypothesis that is not simple is called composite.

Hypothesis Testing

- A statistical hypothesis test is a procedure for deciding between two possible statements about a population. The phrase significance test means the same thing as the phrase "hypothesis test."
- A hypothesis test is a statistical method that uses sample data to evaluate a hypothesis about a population. The general goal of a hypothesis test is to rule out chance as a plausible explanation for the results from a research study.
- The goal in hypothesis testing is to analyze a sample in an attempt to distinguish between population characteristics that are likely to occur and population characteristics that are unlikely to occur.

Basic assumption of hypothesis testing

- If the treatment has any effect, it is simply to add or subtract a constant amount to each individual's score.
- Remember that adding or subtracting constant changes the mean, but not the shape of the distribution for the population and/or the standard deviation.

- The population after treatment has the same shape and standard deviation as the population prior to treatment.
- If the individuals in the sample are noticeably different from the individuals in the original population, we have evidence that the treatment has an effect.

The purpose of the hypothesis test is to decide between two explanations :

1. The difference between the sample and the population can be explained by sampling error.
2. The difference between the sample and the population is too large to be explained by sampling error.

Steps in hypothesis testing

1. Specify the null hypothesis.
2. Specify the alternative hypothesis
3. Set the significance level (?)
4. Calculate the test statistic and corresponding P-value.
5. Display the conclusion.

Step 1 : Formulate the hypothesis

- A null hypothesis is a statement of the status quo, one of no difference or no effect. If the null hypothesis is not rejected, no changes will be made.
- An alternative hypothesis is one in which some difference or effect is expected.
- The null hypothesis refers to a specified value of the population parameter, not a sample statistic.

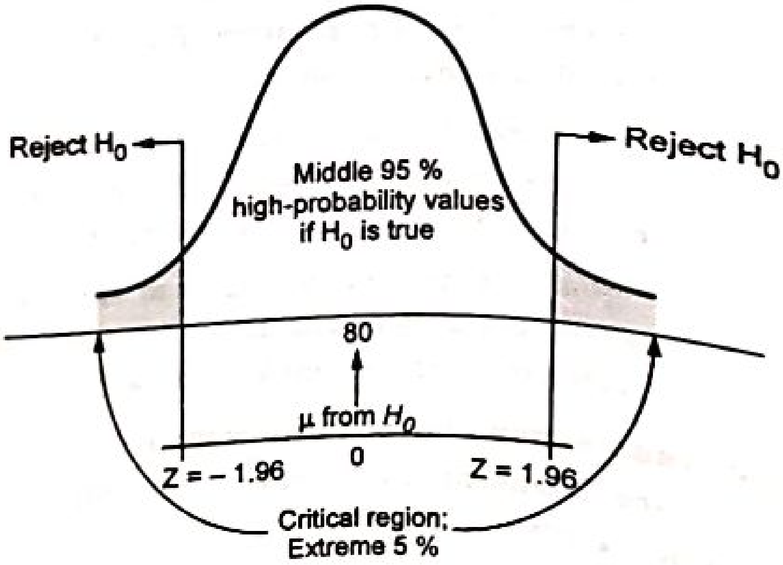

Reject H_0
Reject H_0
Middle 95 % high-probability values if H_0 is true
80
μ from H_0
0
Z = – 1.96
Z = 1.96
Critical region; Extreme 5 %

Fig. 5.9.1

Step 2 : Select an appropriate test

- The **test statistic** measures how close the sample has come to the null hypothesis.
- The test statistic often follows a well-known distribution (e.g., normal, t, or chi-square).
- Calculate Z statistic.

Step 3 : Choose level of significance

Type I Error

- Occurs if the null hypothesis is rejected when it is in fact true.
- The probability of type I error (α) is also called the **level of significance.**

Type II Error

- Occurs if the null hypothesis is not rejected when it is in fact false.
- The probability of type II error is denoted by β.
- Unlike α, which is specified by the researcher, the magnitude of β depends on the actual value of the population parameter (proportion).
- **It is necessary to balance the two types of errors.**
- The power of a test is the probability $(1 - \beta)$ of rejecting the null hypothesis when it is false and should be rejected. Although β is unknown, it is related to α.

Step 4 : Collect data and calculate test statistic

- The required data are collected and the value of the test statistic computed. The test statistic z can be calculated as follows :

$$Z_{cal} = \frac{\hat{P} - \pi}{\sigma_P}$$

Step 5 : Determine probability value/critical value

- Using standard normal tables.
- Note, in determining the critical value of the test statistic, the area to the right of the critical value is either α or $\alpha/2$. It is α for a one-tail test and $\alpha/2$ for a two-tail test.
- If the prob associated with the calculated value of the test statistic (Z_{cal}) is less than the level of significance (α), the null hypothesis is rejected.
- Alternatively, if the calculated value of the test statistic is greater than the critical value of the test statistic (z_α), the null hypothesis is rejected.

1. **Two-tailed alternative :** If the alternative states that a population parameter is different from a specific value. The corresponding test is called a two-tailed test.

2. **Right-tailed alternative :** If the alternative states that a population parameter is greater than a specific value. The corresponding test is called a right-tailed test.

3. **Left-tailed alternative :** If the alternative states that a population parameter is less than a specific value. The corresponding test is called a left-tailed test.

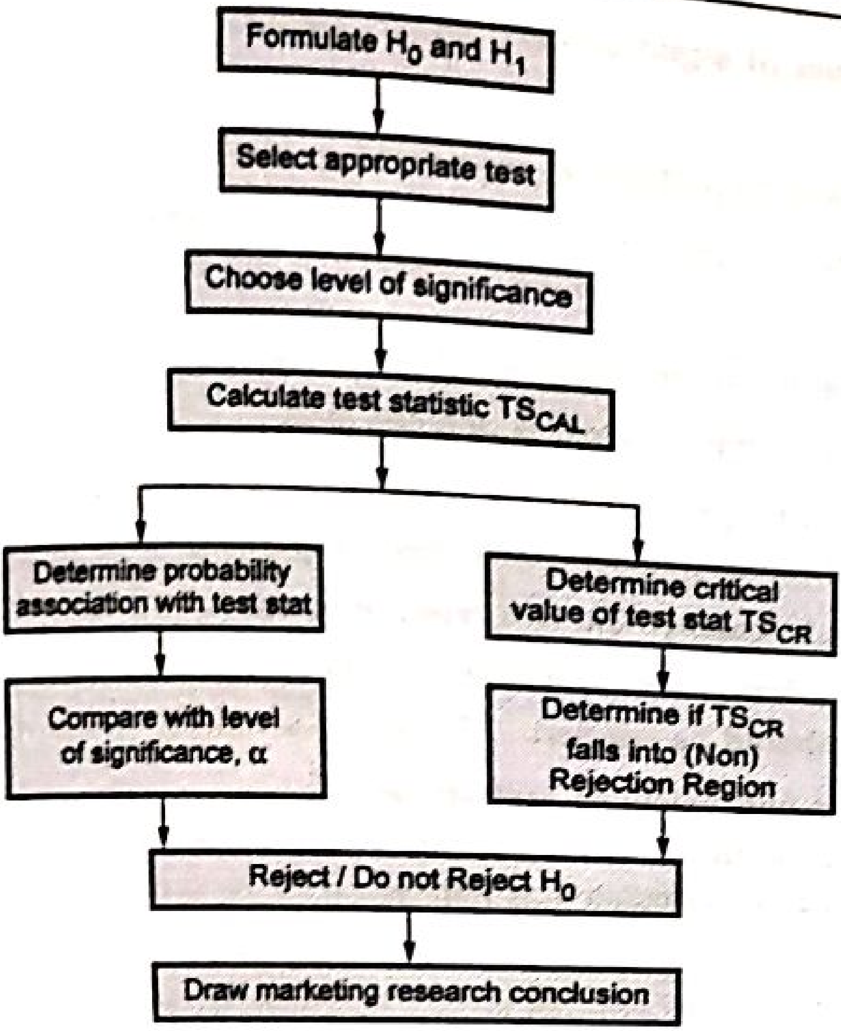

Formulate H_0 and H_1
Select appropriate test
Choose level of significance
Calculate test statistic TS_{CAL}
Determine probability association with test stat
Determine critical value of test stat TS_{CR}
Compare with level of significance, α
Determine if TS_{CR} falls into (Non) Rejection Region
Reject / Do not Reject H_0
Draw marketing research conclusion

Fig. 5.9.2

Decide the rejection region of the test

- Based on the test statistic and a given confidence level, we can determine the rejection region, the acceptance region, and the critical value of the test.
- Rejection region is the region in which we can reject the null-hypothesis when the test statistics falls in this region. Acceptance region is simply the complement of the rejection region.
- Critical value is the value on the boundary of the rejection region and acceptance region.

1) For arbitrary population, acceptance and rejection regions are shown in Fig. 5.9.3.

2) For normal population, acceptance and rejection regions are shown in Fig. 5.9.4.

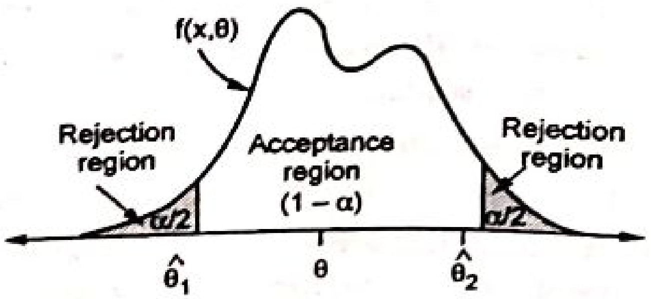


Fig. 5.9.3 Arbitrary population

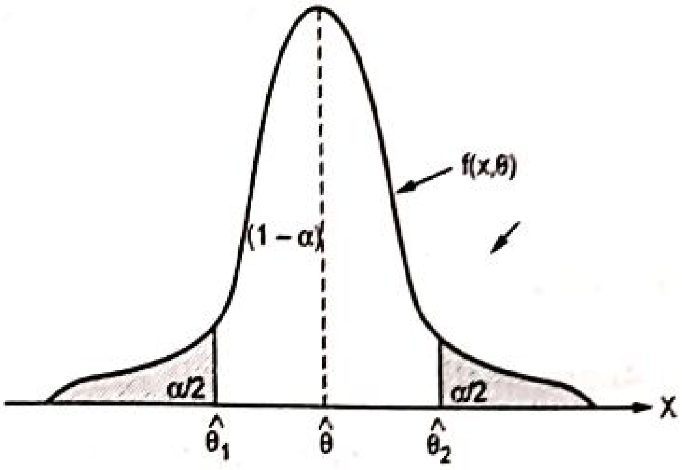


Fig. 5.9.4 Normal population

P-value and hypotheses testing

- As an alternative approach to the rejection/acceptance-region approach, we can calculate a probability related to the test statistic, called P-value, and base our decision of rejection/acceptance on the magnitude of the P-value.
- P-value is the probability to observe a value of the test statistic as extreme as the one observed, if the null hypothesis is true. So a small P-value indicates that the null hypothesis is not true and hence should be rejected.

In a hypothesis testing problem :

a) The null hypothesis will not be rejected unless the data are not unusual (given that the hypothesis is true).

b) The null hypothesis will not be rejected the P-value indicates the data are very unusual (given that the hypothesis is true).

c) The null hypothesis will not be rejected only if the probability of observing the data provides convincing vidence that it is true.

d) The null hypothesis is also called the research hypothesis ; the alternative hypothesis often represents the status quo.

e) The null hypothesis is the hypothesis that we would like to prove ; the alternative hypothesis is also called the research hypothesis.

5.9.1 Difference between Null and Alternative Hypothesis

Sr. No.	Null hypothesis	Alternative hypothesis
1.	Represented by H_0.	Represented by H_1.
2.	Statement about the value of a population parameter.	Statement about the value of a population parameter that must be true if the null hypothesis is false.

3.	Always stated as an equality.	Stated in on of three forms : >, <, ≠
4.	This is the hypothesis or claim that is initially assumed to be true.	This is the hypothesis or claim which we initially assume to be false but which we may decide to accept if there is sufficient evidence.
5.	Independent variable had no effect on the dependent variable.	Independent variable did have an effect on the dependent variable.

5.10 Monte Carlo Approximation

- Monte Carlo method is used for drawing a sample at random from the empirical distribution.
- Using the Monte Carlo technique, we can approximate the expected value of any function of a random variable by simply drawing samples from the population of the random variable, and then computing the arithmetic mean of the function applied to the samples.
- These methods are used in cases where analytical or numerical solutions don't exist or are too difficult to implement
- Monte-Carlo methods generally follow the following steps :
  1. Determine the statistical properties of possible inputs
  2. Generate many sets of possible inputs which follows the above properties
  3. Perform a deterministic calculation with these sets
  4. Analyze statistically the results
- Monte Carlo integration uses random sampling of a function to numerically compute an estimate of its integral. Suppose that we want to integrate the one-dimensional function f (x) from a to b :

$$F = \int_a^b f(x)\,dx$$

- We can approximate this integral by averaging samples of the function f at uniform random points within the interval

5.11 Fill in the Blanks

Q.1 The probability of the joint event A and B is defined as the ______ rule

Q.2 A set of all possible outcomes of an experiment is called ______ space of that experiment.

Q.3 The outcomes of the trial are said to be ______ , if the occurrence of one of them precludes of all other outcomes.

Q.4 A function which takes on any value from the sample space and its range is some set of numbers is called a ______ of an experiment.

Q.5 A ______ probability is a probability that measures the likelihood that two or more events will happen concurrently.

Q.6 If n independent Bernoulli trials are performed and X represents the number of success in those n trials, then X is called a ______ random variable

Q.7 ______ is a finite set of objects being investigated

Q.8 Random sample refers to a sample of objects drawn from a ______ in a way that every member of the population has the same chance of being chosen.

Q.9 Sampling distribution refers to the probability distribution of a ______ variable defined in a space of random samples.

Q.10 The arithmetic mean of a discrete random variable of the probability distribution is called as ______.

Q.11 The intersection of two events A and B, denoted by the symbol ______ , is the event containing all elements that are common to A and B.

Q.12 Random variables whose values can be written as a finite or infinite sequence are called ________.

Q.13 A ______ variable takes numerical value which is determined by the result of the random experiment.

Q.14 Cross-correlation is a measure of similarity of two series as a function of the displacement of one relative to the other. This is also known as a ______ .

Q.15 When the changes in one variable are associated or followed by the changes in the other is called ______ .

Q.16 Laplace distribution, which is also known as the ______ exponential distribution

Q.17 The ______ distribution is the probability distribution of the number of failures we get by repeating a Bernoulli experiment until we obtain the first success.

Q.18 Covariance can be between 0 and ______ .

Q.19 In a binomial distribution, when the number of trials n is large and the probability of success p is small, the distribution approaches the ______ distribution

Q.20 If μ is not an integer, the mode of Poisson distribution is integral part of μ. In this case the distribution is called ________ .

Q.21 In binomial distribution, formula of calculating mean is ______ .

Q.22 When sample sizes are small, as is often the case in practice, the ______ Theorem does not apply

Q.23 A ______ is a statement of a relationship between two or more variable.

5.12 Multiple Choice Questions

Q.1 Two items are chosen at random from a 12 items of which 4 are defective. A be the event that 'both items chosen are defective'. What is P(A)?

a 1/11 b 14/33

c 10/11 d 1/33

Q.2 Two events are not independent if

a Events are not mutually exclusive

b Events are mutually exclusive

c Outcome of one trial does not depend on the outcome of the other trial

d None of these

Q.3 A random variable is said to be discrete if its range set is

a Finite b Countably infinite

c Either (a) or (b) d Neither (a) nor (b)

Q.4 A pair of fair dice is tossed. What is the probability that the maximum of the two numbers is greater than 4?

a 4/36 b 20/36

c 2/36 d 6/36

Q.5 If X is the random variable representing the number of tails obtained when a coin is tossed four times, the maximum value taken by X is

a 0 b 3

c 4 d 16

Q.6 Which among the following is a sample space obtained while tossing a coin thrice?

a {(H,T),(T,H),(T,T),(H,H)}

b {(H,H,H),(H,T,T),(T,T,T)}

c {(H,H),(T,T)}

d {(H,H,H),(H,H,T), (H,T,T),(T,H,T),(H,T,H),(T,T,H),(T,H,H),(T,T,T)}

Q.7 If all values of a sample are same, then its variance is

a 1

b 0

c 2

d Cannot be determined

Q.8 If A={1} and B={2,3} in S={1,2,3,5,6},which is the event representing the occurrence of exactly one of events A, B ?

a {1,2,3}

b { }

c {2,3}

d S

Q.9 Baye's theorem is applicable to the events that are :

a Independent

b Conditional

c Disjoint

d All of these

Q.10 If P(A ∪ B) = 5/6 and P(A ∩ B) = 1/3 and $P(A^c) = \frac{1}{2}$ then P(A/B) is _____

a 1

b 0.25

c 0.75

d 0.50

Q.11 if S is a set of exhaustive events, then : _________

a 0 < P(S) < 1

b P(S) = 1

c P(S) = 0

d P(S) ≥ 0

Q.12 If the increase or decrease in one variable corresponds to an increase or decrease in the other, the correlation is said to be ______ correlation.

a Auto

b Cross

c Positive

d Negative

Q.13 Examples of few important continuous random distributions are _______

a Uniform distribution

b Gaussian distribution

c Laplace distribution

d All of these

Q.14 A family of parametric distribution in which mean = variance is

a Binomial distribution

b Gamma distribution

c Normal distribution

d Poisson distribution

Q.15 Let X is a binomial variate with parameters n and p. If n=1, the distribution of X reduces to ———

a Poisson distribution
b Binomial distribution
c Bernoulli distribution
d Uniform distribution

Q.16 Binomial distribution with parameters n and p is said to be symmetric if

a $q < p$
b $q = p$
c $q > p$
d $q \neq p$

Q.17 Which of the following real life situations follow Poisson distribution

a The number of printing mistakes per page of a book
b The number of defects per item produced
c The number of persons arriving in a queue
d All the above.

Q.18 The skewness of a binomial distribution will be zero if

a $P < 1 / 2$
b $p = 1 / 2$
c $p > 1 / 2$
d $P = q$

Q.19 The null and alternate hypothesis statements are important parts of the analytical methods collectively known as ________ inferential statistics.

a Inferential
b descriptive
c fixed
d variable

Q.20 Null hypothesis is represented by ________

a H_3
b H_2
c H_1
d H_0

Q.21 The probability of type I error is also called the ___________

a significance
b level of significance
c test statistic
d none of these

Q.22 Tolerance limits cannot be directly calculated using the ________ distribution table.

a binomial
b probability
c normal
d all of these

Q.23 The _____ interval provides a good estimate of the location of a future observation, which is quite different from the estimation of the sample mean value.

a prediction

b confidence

c alternative

d all of these

Answer Keys for Fill in the Blanks

Q.1	product	Q.2	sample	Q.3	mutually exclusive	Q.4	random variable
Q.5	joint	Q.6	binomial	Q.7	Population	Q.8	Population
Q.9	random	Q.10	expectation of random variable	Q.11	A ? B	Q.12	discrete random variables
Q.13	random	Q.14	sliding dot product	Q.15	correlation	Q.16	double-sided
Q.17	geometric	Q.18	infinity	Q.19	Poisson	Q.20	unimodal
Q.21	$\mu = np$	Q.22	Central Limit	Q.23	hypothesis		

Answer Keys for Multiple Choice Questions

Q.1	a	Q.2	b	Q.3	c	Q.4	b
Q.5	c	Q.6	d	Q.7	b	Q.8	a
Q.9	c	Q.10	d	Q.11	b	Q.12	c
Q.13	d	Q.14	d	Q.15	c	Q.16	b
Q.17	d	Q.18	c	Q.19	a	Q.20	d
Q.21	b	Q.22	c	Q.23	a		

Notes

6 Bayesian Concept Learning

Syllabus

Importance of Bayesian methods, Bayesian theorem, Bayes' theorem and concept learning, Bayesian Belief Network

Contents

6.1 Importance of Bayesian Methods

- Bayesian methods allow us to estimate model parameters, to construct model forecasts and to conduct model comparisons. Bayesian learning algorithms can calculate explicit probabilities for hypotheses.
- Bayesian classifiers use a simple idea that the training data are utilized to calculate an observed probability of each class based on feature values.
- When Bayesian classifier is used for unclassified data, it uses the observed probabilities to predict the most likely class for the new features.
- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.
- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis. In Bayesian learning, prior knowledge is provided by asserting a prior probability for each candidate hypothesis, and a probability distribution over observed data for each possible hypothesis.
- Bayesian methods can accommodate hypotheses that make probabilistic predictions. New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.
- Even in cases where Bayesian methods prove computationally intractable, they can provide a standard of optimal decision making against which other practical methods can be measured.
- Uses of Bayesian classifiers are as follows :

  1. Used in text-based classification for finding spam or junk mail filtering.
  2. Medical diagnosis.
  3. Network security such as detecting illegal intrusion.

6.2 Bayes Theorem

- Bayes' theorem is a method to revise the probability of an event given additional information. Bayes's theorem calculates a conditional probability called a posterior or revised probability.
- Bayes' theorem is a result in probability theory that relates conditional probabilities. If A and B denote two events, P(A | B) denotes the conditional probability of A occurring, given that B occurs. The two conditional probabilities P(A | B) and P(B | A) are in general different.
- Bayes theorem gives a relation between P(A | B) and P(B | A). An important application of Bayes' theorem is that it gives a rule how to update or revise the strengths of evidence-based beliefs in light of new evidence a posteriori.

- A **prior probability** is an initial probability value originally obtained before any additional information is obtained.
- A **posterior probability** is a probability value that has been revised by using additional information that is later obtained.
- Suppose that $B_1, B_2, B_3 \dots B_n$ partition the outcomes of an experiment and that A is another event. For any number, k, with $1 \le k \le n$, we have the formula :

$$P(B_k/A) = \frac{P(A/B_k)\cdot P(B_k)}{\sum_{i=1}^{n} P(A/B_i)\cdot P(B_i)}$$

Example 6.2.1 *A mechanical factory production line is manufacturing bolts using three machines, A, B and C. The total output, machine A is responsible for 25 %, machine B for 35 % and machine C for the rest. The machines that 5 % of the output from machine A is defective, 4 % from machine B and 2 % from machine C. A bolt is chosen at random from the production line and found to be defective. What is the probability that it came from*

i. machine A ii. machine B iii. machine C ?

Solution : Let

$$\begin{aligned} D &= \{\text{bolt is defective}\}, \\ A &= \{\text{bolt is from machine A}\}, \\ B &= \{\text{bolt is from machine B}\}, \\ C &= \{\text{bolt is from machine C}\}. \end{aligned}$$

Given data : $P(A) = 0.25$, $P(B) = 0.35$, $P(C) = 0.4$.

$P(D|A) = 0.05$, $P(D|B) = 0.04$, $P(D|C) = 0.02$.

From the Bayes' Theorem :

$$P(A/D) = \frac{P(D/A)\times P(A)}{P(D/A)\times P(A)+P(D/B)\times P(B)+P(D/C)\times P(C)}$$

$$= \frac{0.05\times 0.25}{0.05\times 0.25 + 0.04\times 0.35 + 0.02\times 0.4}$$

$$= \frac{0.0125}{0.0125 + 0.014 + 0.008}$$

$$\mathbf{P(A/D) = 0.3621}$$

Similarly :

$$P(B/D) = \frac{P(D/B)\times P(B)}{P(D/A)\times P(A)+P(D/B)\times P(B)+P(D/C)\times P(C)}$$

$$= \frac{0.04 \times 0.35}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4}$$

$$= \frac{0.014}{0.0125 + 0.014 + 0.008} = \frac{0.014}{0.0345}$$

$$P(B/D) = 0.4057$$

$$P(C/D) = \frac{P(D/C) \times P(C)}{P(D/A) \times P(A) + P(D/B) \times P(B) + P(D/C) \times P(C)}$$

$$= \frac{0.02 \times 0.4}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4}$$

$$= \frac{0.008}{0.0125 + 0.014 + 0.008} = \frac{0.008}{0.0345}$$

$$P(C/D) = 0.2318$$

Example 6.2.2 *At a certain university, 4 % of men are over 6 feet tall and 1 % of women are over 6 feet tall. The total student population is divided in the ratio 3 : 2 in favour of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman ?*

Solution : Let us assume following :

M = {Student is Male},

F = {Student is Female},

T = {Student is over 6 feet tall}.

Given data :

$$P(M) = 2/5,$$

$$P(F) = 3/5,$$

$$P(T|M) = 4/100$$

$$P(T|F) = 1/100.$$

We require to find $P(F|T)$?

Using Bayes' Theorem we have :

$$P(F/T) = \frac{P(T/F)\,P(F)}{P(T/F)\,P(F) + P(T/M)\,P(M)} = \frac{\frac{1}{100} \times \frac{3}{5}}{\frac{1}{100} \times \frac{3}{5} + \frac{4}{100} \times \frac{2}{5}} = \frac{\frac{3}{500}}{\frac{3}{500} + \frac{8}{500}}$$

$$P(F/T) = \frac{3}{11}$$

6.2.1 Prior and Posterior Probability

- In Bayesian learning, the best hypothesis means the most probable hypothesis, given the data D plus any initial knowledge about the prior probabilities of the various hypothesis in H.
- Bayes' theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probabilities of observing various data given the hypothesis and the observed data itself.
- Bayes' theorem is a method to revise the probability of an event given additional information.
- Bayes' theorem calculates a conditional probability called a posterior or revised probability.
- Bayes' theorem is a result in probability theory that relates conditional probabilities. If A and B denote two events, $P(A|B)$ denotes the conditional probability of A occurring, given that B occurs. The two conditional probabilities $P(A|B)$ and $P(B|A)$ are in general different.
- This theorem gives a relation between $P(A|B)$ and $P(B|A)$. An important application of Bayes' theorem is that it gives a rule how to update or revise the strengths of evidence-based beliefs in light of new evidence a porteriori.
- A prior probability is an initial probability value originally obtained before any additional information is obtained.
- The prior knowledge or belief about the probabilities of various hypotheses in H is called Prior in context of Bayes' theorem.
- The probability that a particular hypothesis holds for a data set based on the Prior is called the posterior probability or simply Posterior.
- A posterior probability is a probability value that has been revised by using additional information that is later obtained.
- If A and B are two random variables

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

- In the context of classifier hypothesis h and training data I.

$$p(h/I) = \frac{P(I/h)P(h)}{P(I)}$$

Where (h) = Prior probability of hypothesis h

(I) = Prior probability of training data I

$(h|I)$ = Probability of h given I

$P(I|h)$ = Probability of I given h

6.2.2 Maximum - Likelihood Estimation

- Maximum - Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model. When applied to a data set and given a statistical model, maximum - likelihood estimation provides estimates for the model's parameters. $X_1, X_2, X_3, \cdots, X_n$ have joint density denoted $f_\theta(x_1, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n|\theta)$. Given observed values $X_1 = X_1, x_2 = x_2, \cdots, X_n = x_n$,

$$\text{lik}(\theta) = f(x_1, x_2, \cdots, x_n|\theta)$$

Considered as a function of θ.

- If the distribution is discrete, f will be the frequency distribution function.
- The maximum likelihood estimate of θ is that value of that maximises lik(θ) : It is the value that makes the observed data the most probable.

Examples of maximizing likelihood :

- A random variable with this distribution is a formalization of a coin toss. The value of the random variable is 1 with probability θ and 0 with probability $1 - \theta$. Let X be a Bernoulli random variable and let x be an outcome of X, then we have

$$P(X = x) = \begin{bmatrix} \theta & \text{if} & x=1 \\ 1-\theta & \text{if} & x=0 \end{bmatrix}$$

- Usually, we use the notation P(.) for a probability mass and the notation P(.) for a probability density. For mathematical convenience write P(X) as

$$P(X = x) = \theta^x(1-\theta)^{1-x}$$

6.3 Bayes' Theorem and Concept Learning

A consistent learner is one that returns some hypothesis h from the hypothesis class H that is consistent with a random sequence of m examples. A consistent learner is a MAP learner, if all hypothesis are a-priori equally likely.

6.3.1 Consistent Learners

- The group of learners who commit zero error over the training data and output the hypothesis are called consistent learners.
- If the training data is noise free and deterministic and if there is uniform prior probability distribution over H, then every consistent learner outputs the MAP hypothesis

6.3.2 Bayes Optimal Classifier

- Bayes' classifier is a classifier that minimizes the error in a probabilistic manner. If it is Bayes' optimal, then the errors are weighed using the join probability distribution between the input and the output sets.
- The Bayes error is then the error of the Bayes classifier.

6.3.3 Naïve Bayes Classifier

- Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between the features.
- It is highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.
- A Naive Bayes classifier is a program which predicts a class value given a set of attributes.
- For each known class value,
  1. Calculate probabilities for each attribute, conditional on the class value.
  2. Use the product rule to obtain a joint conditional probability for the attributes.
  3. Use Bayes rule to derive conditional probabilities for the class variable.
- Once this has been done for all class values, output the class with the highest probability.
- Naive bayes simplifies the calculation of probabilities by assuming that the probability of each attribute belonging to a given class value is independent of all other attributes. This is a strong assumption but results in a fast and effective method.
- The probability of a class value given a value of an attribute is called the conditional probability. By multiplying the conditional probabilities together for each attribute for a given class value, we have a probability of a data instance belonging to that class.
- A key benefit of the naive Bayes classifier is that it requires only a little bit of training information to gauge the parameters essential for the classification.
- In the Naïve Bayes classifier, independent variables are always assumed, and only the changes of the factors/variables for each class should be determined and not the whole covariance matrix.

Advantages :

1. Simple to implement
2. Calculation is fast and produce effective result.

3. Suitable for noisy and missing data
4. Works well for small number of data.

Disadvantages :

1. Not suitable for large database
2. Estimated probabilities have relatively lower reliability

6.4 Bayesian Belief Network

- Bayesian Belief Networks (BBN) are also known as belief networks, Bayesian networks, and probabilistic networks. BBN is a special type of diagram (called a directed graph) together with an associated set of probability tables.
- The graph consists of nodes and arcs. The nodes represent variables, which can be discrete or continuous. The arcs represent causal relationships between variables.
- A belief network is defined by two components : Directed acyclic graph and a set of conditional probability tables.
- BBNs enable us to model and reason about uncertainty. BBNs accommodate both subjective probabilities and probabilities based on objective data. The most important use of BBNs is in revising probabilities in the light of actual observations of events.
- Each node in the directed acyclic graph represents a random variable. The variables may be discrete or continuous - valued. They may correspond to actual attributes given in the data or to "hidden variables" believed to form a relationship.
- Each arc represents a probabilistic dependence. If an arc is drawn from a node Y to a node Z, then Y is a parent or immediate predecessor of Z, and Z is a descendant of Y. Each variable is conditionally independent of its non - descendants in the graph, given its parents.
- Fig. 6.4.1 shows simple belief network.
- The diagram consists of nodes and arcs. The nodes represent the discrete or continuous variables for which we are interested to calculate the conditional probabilities. The arc represents the causal relationship of the variables.
- Belief network has one Conditional Probability Table (CPT) for each variable.

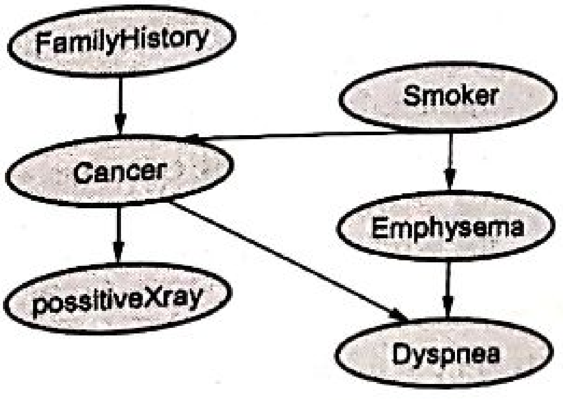


Fig. 6.4.1 simple belief network

- Probability theory is the body of knowledge that enables us to reason formally about uncertain events.
- The probability P of an uncertain event A, written P(A) is defined by the frequency of that event based on previous observations. This is called frequency based probability.
- In general, a person belief in a statement a will depend on some body of knowledge K. We write this as P(a|K). The expression P(a|K) thus represents a belief measure.
- Sometimes, for simplicity, when K remains constant we just write P(a), but you must be aware that this is a simplification.
- The notion of degree of belief P(A|K) is an uncertain event A is conditional on a body of knowledge K. In general, we write P(A|B) to rep represent a belief in A under the assumption that B is known.
- Bayesian belief network describes the joint probability distribution of a set of attributes in their joint space.
- Bayesian networks are used for modelling beliefs in domains like computational biology and bioinformatics such as protein structure and gene regulatory networks, medicines, forensics, document classification, information retrieval, image processing, decision support systems, sports betting and gaming.
- **Example :** Alarm system example.
- Assume your house has an alarm system against burglary. You live in the seismically active area and the alarm system can get occasionally set off by an earthquake.
- You have two neighbors, Mary and John, who do not know each other. If they hear the alarm they call you, but this is not guaranteed.
- We want to represent the probability distribution of events : Burglary, Earthquake, Alarm, Mary calls and John calls.

Causal relations :

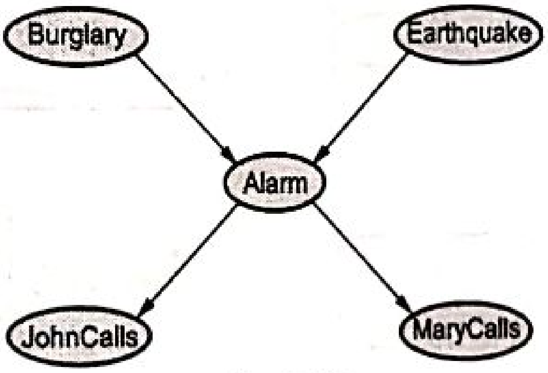


Fig. 6.4.2

Directed acyclic graph :

- Nodes = Random variables
  Burglary, Earthquake, Alarm, Mary calls and John calls
- Links = Direct (causal) dependencies between variables.
  The chance of Alarm is influenced by Earthquake, The chance of John calling is affected by the Alarm.

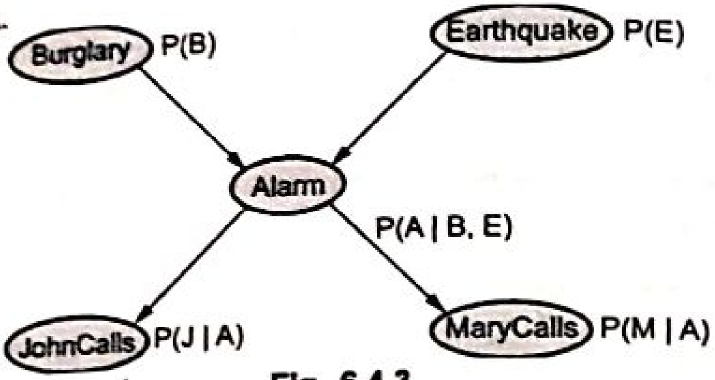


Fig. 6.4.3

Local conditional distributions : Relate variables and their parents.

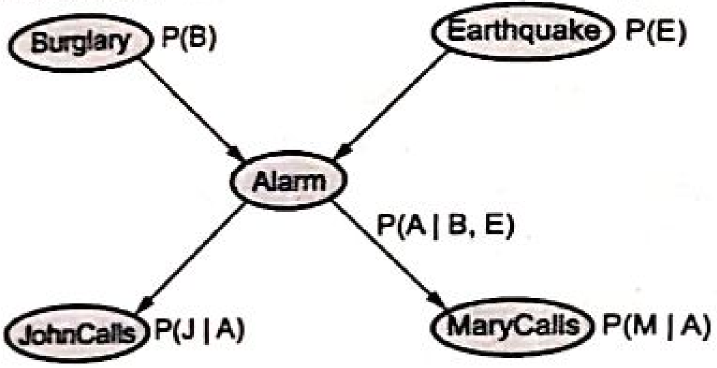
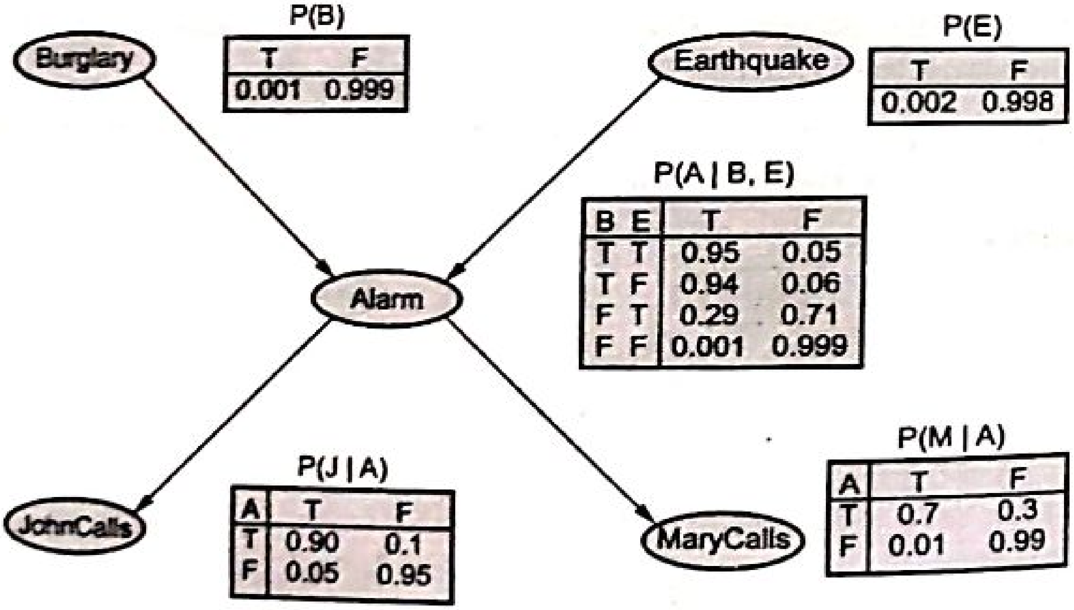


Fig. 6.4.4

Bayesian belief network



P(B)

T	F
0.001	0.999

P(E)

T	F
0.002	0.998

P(A | B, E)

B	E	T	F
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	0.999

P(J | A)

A	T	F
T	0.90	0.1
F	0.05	0.95

P(M | A)

A	T	F
T	0.7	0.3
F	0.01	0.99

Fig. 6.4.5

6.5 Fill in the Blanks

Q.1 The group of learners who commit zero error over the training data and output the hypothesis are called ________ learners.

Q.2 The Bayes rule, also known as ________ theorem, can be derived by combining the definition of conditional probability with the product and sum rules.

Q.3 The probability of the joint event A and B is defined as the ________ rule.

Q.4 If n independent Bernoulli trials are performed and X represents the number of success in those n trials, then X is called a ________ random variable.

Q.5 Population is a ________ set of objects being investigated.

Q.6 ________ refers to a sample of objects drawn from a population in a way that every member of the population has the same chance of being chosen.

Q.7 Sampling distribution refers to the ________ of a random variable defined in a space of random samples.

Q.8 ________ theorem calculates a conditional probability called a posterior or revised probability.

Q.9 A ________ estimate of a parameter consists of an interval of numbers along with a probability that the interval contains the unknown parameter.

Q.10 Bayes' theorem provides a way to calculate the probability of a hypothesis based on its ________ the probabilities of observing various data given the hypothesis and the observed data itself.

Q.11 PAC-learnability is largely determined by the number of training examples required by the ________.

Q.12 A learner is ________ if it outputs hypotheses that perfectly fit the training data, whenever possible.

Answer Keys for Fill in the Blanks

Q.1	consistent	Q.2	Bayes	Q.3	product
Q.4	binomial	Q.5	finite	Q.6	Random sample
Q.7	probability distribution	Q.8	Bayes'	Q.9	confidence interval
Q.10	prior probability	Q.11	learner	Q.12	consistent

□□□

Notes

7 Supervised Learning : Classification and Regression

Syllabus

Supervised Learning, Classification Model, Learning steps, Classification algorithms, Regression, Regression algorithms.

Contents

7.1 Supervised Learning Example

- Supervised Learning is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process.
- Supervised learning helps organizations solve for a variety of real - world problems at scale, such as classifying spam in a separate folder from your inbox.
- Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.
- Supervised learning can be separated into two types of problems when data mining, classification and regression :
- Classification uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined. Common classification algorithms are linear classifiers, Support Vector Machines (SVM), decision trees, k-nearest neighbour, and random forest, which are described in more detail below.
- Regression is used to understand the relationship between dependent and independent variables. It is commonly used to make projections, such as for sales revenue for a given business. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.
- Examples of supervised learning are as follows :

  a) Prediction of results of a game based on the past analysis of results

  b) Predicting whether a tumour is malignant or benign on the basis of the analysis of data

  c) Price prediction in domains such as real estate, stocks, etc.

7.2 Classification Model

- Classification is used to find out in which group each data instance is related within a given dataset. It is used for classifying data into different classes according to some constraints.
- Data classification is a two - step process : Learning and classification.
- During first step the model is created by applying classification algorithm on training data set then in second step the extracted model is tested against a predefined test data set of **measure** the model trained performance and accuracy.

- So classification is the process to assign class label from data set whose class label is unknown.
- Classification is the task of choosing the correct class label for a given input. In basic classification tasks, each input is considered in isolation from all other inputs, and the set of labels is defined in advance. Some examples of classification tasks are :
  a) Deciding whether an email is spam or not.
  b) Deciding what the topic of a news article is, from a fixed list of topic areas such as "sports," "technology," and "politics."
  c) Deciding whether a given occurrence of the world bank is used to refer to a river bank, a financial institution, the act of tilting to the side, or the act of depositing something in a financial institution.
- The basic classification task has a number of interesting variants. For example, in multi - class classification, each instance may be assigned multiple labels; in open - class classification, the set of labels is not defined in advance; and in sequence classification, a list of inputs are jointly classified.
- A classifier is called **supervised** if it is built based on training corpora containing the correct label for each input.
- Example :
  1. **Image and object - recognition :** Supervised learning algorithms can be used to locate, isolate, and categorize objects out of videos or images, making them useful when applied to various computer vision techniques and imagery analysis.
  2. **Predictive analytics :** A widespread use case for supervised learning models is in creating predictive analytics systems to provide deep insights into various business data points.
  3. **Customer sentiment analysis :** Using supervised machine learning algorithms, organizations can extract and classify important pieces of information from large volumes of data - including context, emotion and intent - with very little human intervention.
  4. **Spam detection :** Spam detection is another example of a supervised learning model.

7.3 Learning Steps

- Fig. 7.3.1 shows classification steps.

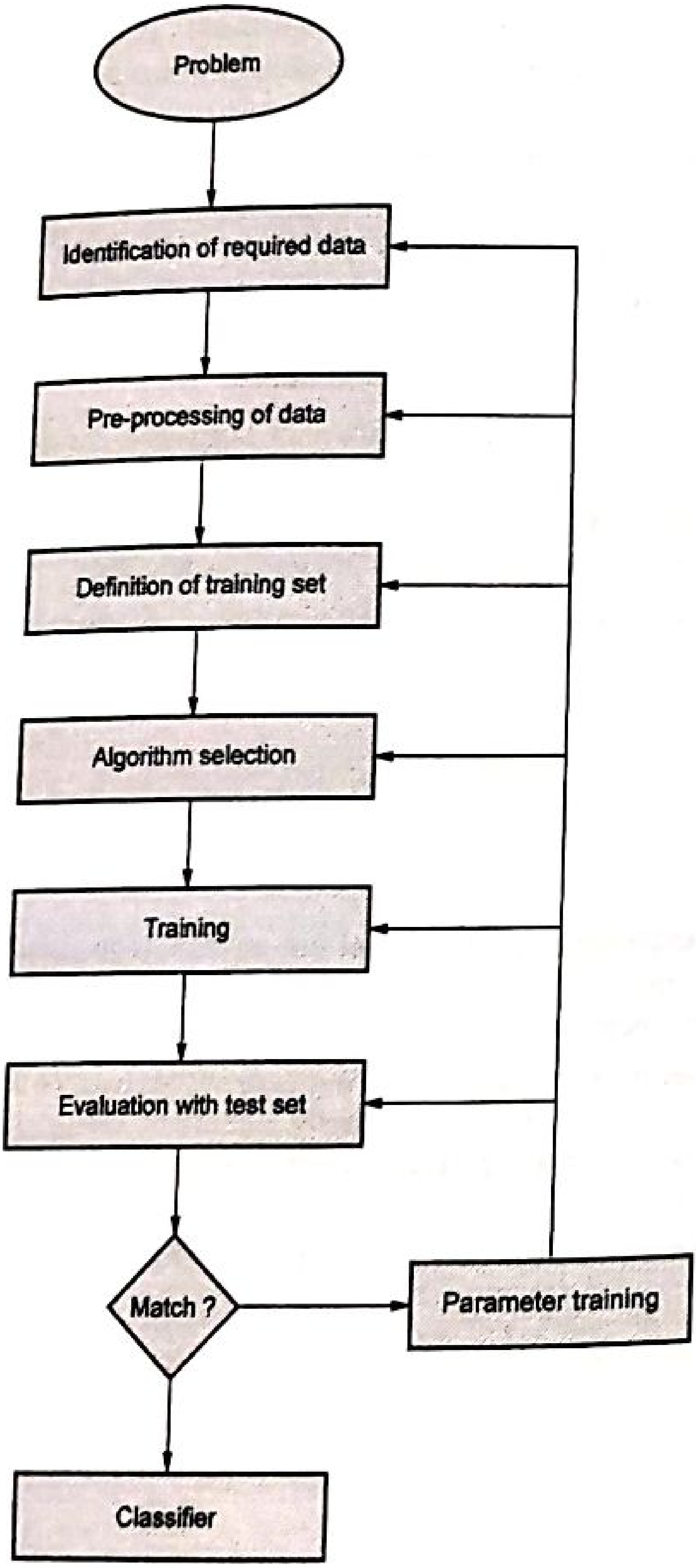


Fig. 7.3.1

1. **Problem identification :** First step of supervised learning is problem identification. Problem statement must be well defined. It contains goals and benefits.
2. **Identification of required data :** The required data set that precisely represents the identified problem needs to be identified/evaluated.
3. **Data pre-processing :** This is related to the cleaning/transforming the data set. This step ensures that all the unnecessary/irrelevant data elements are removed.
4. **Definition of training data set :** Before starting the analysis, the user should decide what kind of data set is to be used as a training set.
5. **Algorithm selection :** This involves determining the structure of the learning function and the corresponding learning algorithm.
6. **Training :** The learning algorithm identified is run on the gathered training set for further fine tuning.
7. **Evaluation with the test data set :** Training data is run on the algorithm, and its performance is measured here

7.4 Classification Algorithms

7.4.1 k-Nearest Neighbour (kNN)

- The k-nearest neighbour (kNN) is a classical classification method and requires no training effort, critically depends on the quality of the distance measures among examples.
- The kNN classifier uses Mahalanobis distance function. A sample is classified according to the majority vote of its nearest k training samples in the feature space. Distance of a sample to its neighbors is defined using a distance function.
- For all points x, y and z distance function F(., .), must satisfy the following conditions :

1.	Non-negativity	$F(x, y) \geq 0$
2.	Reflexivity	$F(x, y) = 0$ if and only if $x = y$
3.	Symmetry	$F(x, y) = F(y, x)$
4.	Triangle inequality	$F(x, y) + F(y, x) \geq F(x, y)$

- Mahalanobis distance is also called **quadratic distance.**
- Mahalanobis distance is a distance measure between two points in the space defined by two or more correlated variables. Mahalanobis distance takes the correlations within a data set between the variable into considerations.

- If there are two non-correlated variables, the Mahalanobis distance between the points of the variable in a 2D scatter plot is same as Euclidean distance.
- The Mahalanobis distance is the distance between an observations and the center for each group in m - dimensional space defined by m variables and their covariance. Thus, a small value of Mahalanobis distance increases the chance of an observation to be closer to the group's center and the more likely it is to be assigned to that group.
- Mahalanobis distance between two samples (x, y) of a random variable is defined as

$$d_{\text{Mahalanobis}}(x, y) = \sqrt{(x-y)^T \sum{}^{-1}(x-y)}$$

- The Mahalanobis metric is defined in independence of the data matrix.
- No pre - precessing of labeled data samples is needed before using kNN algorithm. A dominated class label in k - nearest neighbors of a data point is assigned as class label to that data point. A tie occurs when neighborhood has same amount of labels from multiple classes.
- To break the tie, the distances of neighbors can be summed up in each class that is tied and vector f is assigned to the class with minimal distance. Or, the class can be chosen with the nearest neighbor. Clearly, tie is still possible here, in which case an arbitrary assignment is taken.
- There distance functions that can be used in kNN classifier are :

1.	L_p norm	$L_p(x, y) = \left(\sum_{i=1}^{d} \lvert x_i - y_i \rvert^{P}\right)^{1/p}$
2.	L_2 norm (Euclidean distance)	$L_2(x, y) = \left(\sum_{i=1}^{d} \lvert x_i - y_i \rvert^{2}\right)^{1/2}$
3.	L_1 norm (Manhattan distance)	$L_1(x, y) = \sum_{i=1}^{d} \lvert x_i - y_i \rvert$

- Mahalanobis distance that takes into account the correlation S of the dataset :

$$L_m(x, y) = \sqrt{(x-y)S^{-1}(x-y)}$$

Advantages of kNN :

1. Simple to implement.
2. Good classification if the number of samples is large enough.
3. High performance accuracy.

Disadvantages :

1. Choosing k may be tricky.
2. Test stage is computationally expensive.
3. No training stage.

7.4.2 Decision Tree

- Decision tree learning is a method for approximating discrete - valued target functions, in which the learned function is represented by a decision tree.
- A decision tree is a tree where each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value).
- A decision tree or a classification tree is a tree in which each internal node is labeled with an input feature. The arcs coming from a node labeled with a feature are labeled with each of the possible value of the feature.
- A decision tree has two kinds of nodes,
  1. Each leaf node has a class lable, determined by majority vote of training examples reaching that leaf.
  2. Each internal node is a question on features. It branches out according to the answers.
- Decision tree learning is a method for approximating discrete - valued target functions. The learned function is represented by a decision tree.
- A decision tree is a tree where
  a. Each non - leaf node has associated with it an attribute (feature)
  b. Each leaf node has associated with it a classification (+ or –)
  c. Each arc has associated with it one of the possible values of the attribute at the node from which the arc is directed.
- Internal node denotes a test on an attribute. Branch represents an outcome of the test. Leaf nodes represent class labels or class distribution.
- A decision tree is a flow - chart - like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test and tree leaves represent classes or class distribution. Decision trees can easily be converted to classification rules.
- There are several steps involved in the building of decision tree.
  1. **Splitting :** The process of partitioning the data set into subsets. Splits are formed on a particular variable and in a particular location. For each split, two determinations are made : The predictor variable used for the split, called the

splitting variable and the set of values for the predictor variable, called the split point.

2. **Pruning :** The shortening of branches of the tree. Pruning is the process of reducing the size of the tree by turning some branch nodes into leaf nodes and removing the leaf nodes under the original branch.
3. **Tree selection :** The process of finding the smallest tree that fits the data. Usually this is the tree that yields the lowest cross - validated error.

- The Fig. 7.4.1 shows an example of a decision tree to determine what kind of contact lens a person way wear.

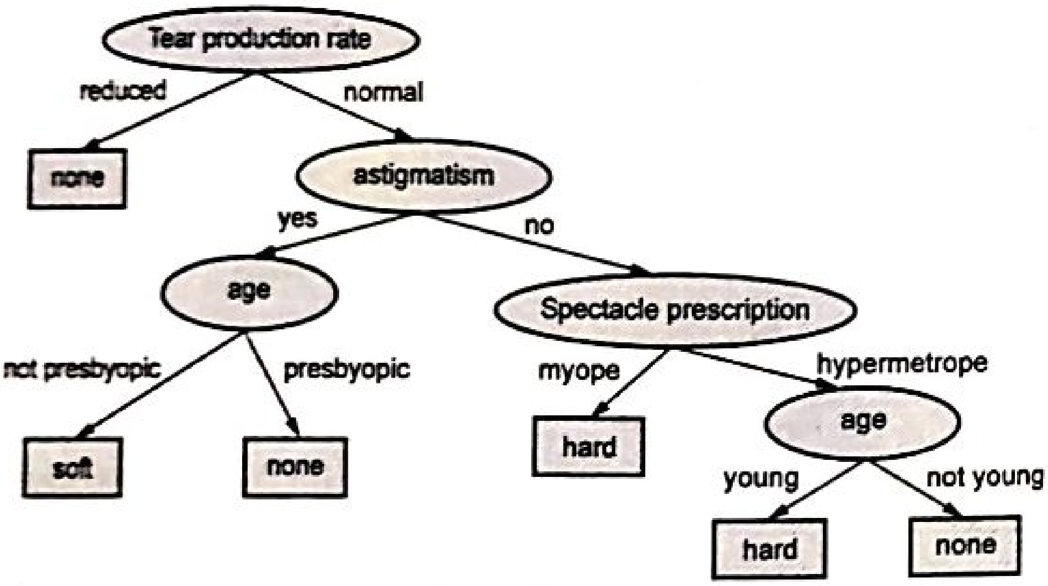


Fig. 7.4.1

- Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.
- Gini index, entropy and towing rule are some of the frequency used impurity measures.
- Gini Index for a given node t :

$$\text{GINI(t)} = \sum p(j \mid t)(1-p(j \mid t)) - \sum p(j \mid t)^2$$

Maximum number of classes when records are equally distributed among all classes is called maximal impurity.

- Minimum of 0 when all records belong to one class = Complete purity.
- Entropy at a given node by :

$$\text{Entropy(t)} = \sum_{j} p(j / t) \log p(j \mid t)$$

- Maximum $(\log n_c)$ when records are equally distributed among all classes (maximal impurity).
- Minimum (0.0) when all records belongs to one class (Maximal purity).
- Entropy is the only function that satisfies all of the following three properties :
  1. When node is pure, measure should be zero.
  2. When impurity is maximal (i.e. all classes equally likely), measure should be maximal.
  3. Measure should obey multistage property.
- When a node p is split into k partitions (children), then quality of the split is computed as a weighted sum :

$$\text{GIN}_{\text{Isplit}} = \sum_{i=1}^{k} \frac{n_i}{n} \text{GINI}(i) = \sum_j p(j \mid t)^2$$

where n_i = Number of records at child i and n = Number of records at node P.

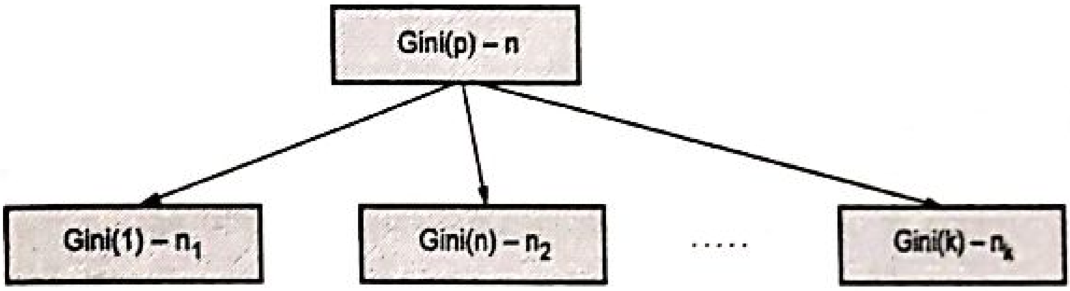


Fig. 7.4.2

7.4.2.1 Information Gain

- Entropy measures the impurity of a collection. Information gain is defined in terms of entropy.
- Information gain tells us how important a given attribute of the feature vectors is,
- Information gain of attribute A is the reduction in entropy caused by partitioning the set of examples S.

$$\text{Gain}(S, A) \equiv \text{Entropy}(S) - \sum_{v \in \text{values}} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

where values (A) is the set of all possible values for attributes A and S_v is the subset of S for which attribute A has value v.

Pruning by information gain :

- The simplest technique is to prune out portions of the tree that result in the least information gain.

- This procedure does not require any additional data and only bases the pruning on the information that is already computed when the tree is being built from training data.
- The process of information gain based pruning required us to identify "twigs", nodes whose children are all leaves.
- "Pruning" a twig removes all of the leaves which are the children of the twig and makes the twig a leaf.
- The algorithm for pruning is as follows :
  1. Catalog all twigs in the tree.
  2. Count the total number of leaves in the tree.
  3. While the number of leaves in the tree exceeds the desired number :
     a) Find the twig with the least information gain
     b) Remove all child nodes of the twig
     c) Relabel twig as a leaf
     d) Update the leaf count.

7.4.2.2 Tree Pruning

- If the classifier fits the training instances too closely, it may fit noisy instances and that reduces its usefulness. This phenomenon is called coverfitting.
- Pruning simplifies a classifier by merging disjuncts that are adjacent in instance space. This can improve the classifier's performance by eliminating error - prone components.
- Pruning of the decision tree is done by replacing a whole sub - tree by a leaf node. The replacement takes place if a decision rule establishes that the expected error rate in the sub - tree is greater than in the single leaf.
- For example :

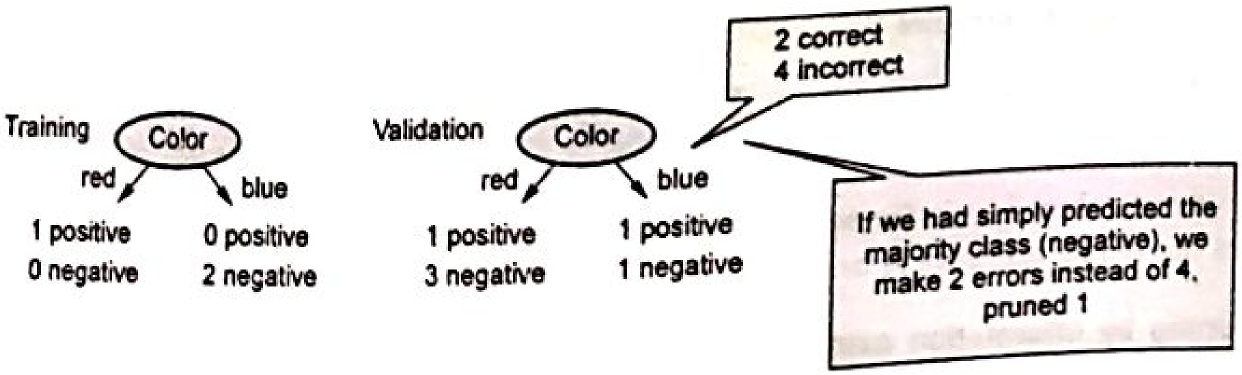


Fig. 7.4.3

7.4.2.3 Decision Tree Algorithm

- To generate decision tree from the training tuples of data partition D.

Input :

1. Data partition (D)
2. Attribute list
3. Attribute selection method.

Algorithm :

1. Create a node (N)
2. If tuples in D are all of the same class then
3. Return node (N) as a leaf node labeled with the class C.
4. If attributes list is empty then return N as a leaf node labeled with the majority class in D
5. Apply attribute selection method (D, attribute list) to find the "best" splitting criterion
6. Label node N with splitting criterion
7. If splitting attribute is discrete - valued and multiway splits allowed
8. Then attribute list -> attribute list -> splitting attributes
9. For each outcome j, select splitting criteria
10. Let D_j be the set of data tuples in D satisfying outcome j
11. If D_j is empty then attach a leaf labeled with the majority class in D to node N
12. Else attach the node returned by generate decision tree (D_j, attribute list) to node N
13. End of for loop
14. return N

7.4.2.4 Decision Tree Advantages and Disadvantages

Advantages :

1. Rules are simple and easy to understand.
2. Decision trees can handle both nominal and numerical attributes.
3. Decision trees are capable of handling datasets that may have errors.
4. Decision trees are capable of handling datasets that may have missing values.
5. Decision trees are considered to be a nonparametric method.
6. Decision trees are self-explantory.

Disadvantages :

1. Most of the algorithms require that the target attribute will have only discrete values.
2. Some problem are difficult to solve like XOR.
3. Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
4. Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.

Example 7.4.1 *If S is a collection of 14 examples with 9 YES and 5 NO examples then calculate entropy.*

Solution :

$$\text{Entropy(S)} = \Sigma - p(I) \log_2 p(I)$$

Where p(I) is the proportion of S belonging to class I.

Σ is over c.

$$\text{Entropy(S)} = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right)$$

$$= -0.940$$

Example 7.4.2 *Consider the following table :*

Weekend (Example)	Wheather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay in
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Calculate Entropy and Gain.

Solution :

$$\text{Entropy(S)} = -p_{cinema} \log_2(p_{cinema}) - p_{tennis} \log_2(p_{tennis})$$
$$-p_{shopping} \log_2(p_{shopping}) - p_{stay\_in} \log_2(p_{stay\_in})$$

$$=. -(6/10) * \log_2(6/10) - (2/10) * \log_2(2/10) - (1/10) * \log_2(1/10) - (1/10) * \log_2(1/10)$$

$$= -(6/10) * -0.737 - (2/10) * -2.322 - (1/10) * -3.322 - (1/10) * -3.322$$

$$= 04422 + 0.4644 + 0.3322 + 0.3322 = 1.571$$

and we need to determine the best of :

$$\text{Gain(S, weather)} = 1.571 - (IS_{sun}\ I/10) * \text{Entropy}(S_{sun}) - (IS_{wind}\ I/10) * \text{Entropy}(S_{wind})$$
$$-(IS_{rain}\ I/10) * \text{Entropy}(S_{rain})$$

$$= 1.571 - (0.3) * \text{Entropy}(S_{sun}) - (0.4) * \text{Entropy}(S_{wind}) - (0.3) * \text{Entropy}(S_{rain})$$

$$= 1.571 - (0.3) * (0.918) - (0.4) * (0.81125) - (0.3) * (0.918) = 0.70$$

$$\text{Gain(S, parents)} = 1.571 - (IS_{yes}\ I/10) * \text{Entropy}(S_{yes}) - (IS_{no}\ I/10) * \text{Entropy}(S_{no})$$
$$= 1.571 - (0.5) * 0 - (0.5) * 1.922 = 1.571 - 0.961 = 0.61$$

$$\text{Gain(S, money)} = 1.571 - (IS_{rich}\ I/10) * \text{Entropy}(S_{rich}) - (IS_{poor}\ I/10) * \text{Entropy}(S_{poor})$$
$$= 1.571 - (0.7) * (1.842) - (0.3) * 0 = 1.571 - 1.2894 = 0.2816$$

- This means that the first node in the decision tree will be the weather attribute. From the weather node, we draw a branch for the values that weather can take : Sunny, windy and rainy :

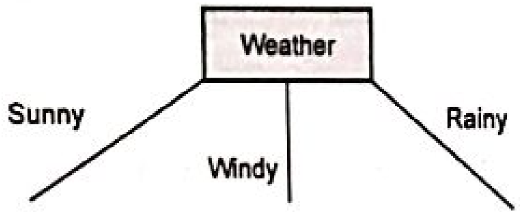


Fig. 7.4.4

- Now we look at the first branch. S_{sunny} = {W1, W2, W10}. This is not empty, so we do not put a default categorization leaf node here.
- The categorisations of W1, W2 and W10 are Cinema, Tennis and Tennis respectively. As these are not all the same, we cannot put a categorisation leaf node here. Hence we put an attribute node here, which we will leave blank for the time being.
- Looking at the second branch, S_{windy} = {W3, W7, W8, W9}. Again, this is not empty and they do not all belong to the same class, so we put an attribute node

here, left blank for now. The same situation happens with the third branch, hence our amended tree looks like this :

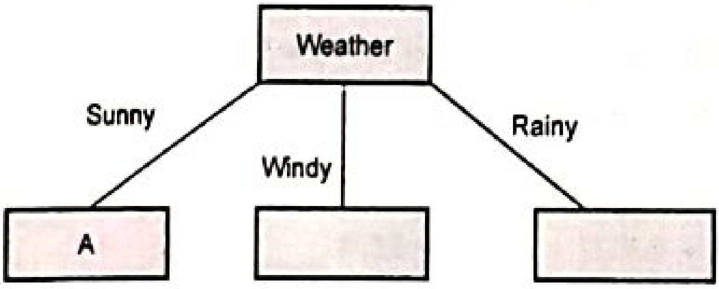


Fig. 7.4.5

- In effect, we are interested only in this part of the table :

Weekend (Example)	Wheather	Parents	Money	Decision (Category)
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W10	Sunny	No	Rich	Tennis

Hence we can calculate :

$$\text{Gain}(S_{sunny,}\text{parents}) = 0.918 - (|S_{yes}|/|S|) * \text{Entropy}(S_{yes}) - (|S_{no}|/|S|) * \text{Entropy}(S_{no})$$

$$= 0.918 - (1/3) * 0 - (2/3) * 0 = 0.918$$

$$\text{Gain}(S_{sunny,}\text{money}) = 0.918 - (|S_{rich}|/|S|) * \text{Entropy}(S_{rich}) - (|S_{poor}|/|S|) * \text{Entropy}(S_{poor})$$

$$= 0.918 - (3/3) * 0.918 - (0/3) * 0 = 0.918 - 0.918 = 0$$

7.4.3 SVM

- Support Vector Machines (SVMs) are a set of supervised learning methods which learn from the dataset and used for classification. SVM is a classifier derived from statistical learning theory by Vapnik and Chervonenkis.
- An SVM is a kind of large - margin classifier : It is a vector space based machine learning method where the goal is to find a decision boundary between two classes that is maximally far from any point in the training data.
- Given a set of training examples, each marked as belonging to one of two classes, an SVM algorithm builds a model that predicts whether a new example falls into one class or the other. Simply speaking, we can think of an SVM model as representing the examples as points in space, mapped so that each of the examples of the separate classes are divided by a gap that is as wide as possible.

- New examples are then mapped into the same space and classified to belong to the class based on which side of the gap they fall on.

Example of Bad Decision Boundaries :

- SVM are primarily two - class classifiers with the distinct characteristic that they aim to find the optimal hyperplane such that the expected generalization error is minimized. Instead of directly minimizing the empirical risk calculated from the training data, SVMs perform structural risk minimization to achieve good generalization.
- Fig. 7.4.6 shows empirical risk.

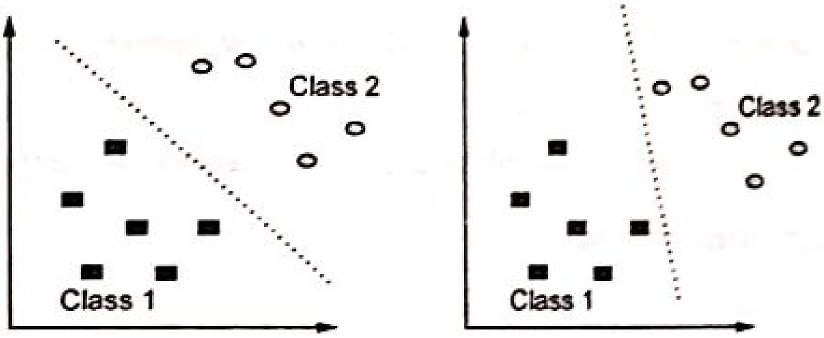


Fig. 7.4.6 Bab decision boundary of SVM

- The empirical risk is the average loss of an estimator for a finite set of data drawn from P. The idea of risk minimization is not only measure the performance of an estimator by its risk, but to actually search for the estimator that minimizes risk over distribution P. Because we don't know distribution P we instead minimize empirical risk over a training dataset drawn from P. This general learning technique is called **empirical rick minimization.**

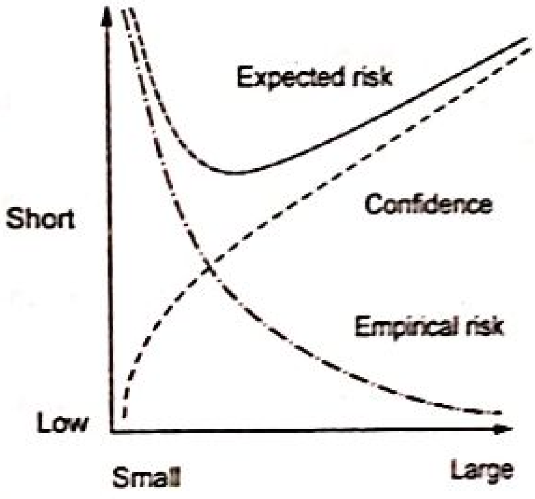


Fig. 7.4.7 Empirical risk

- Error function measures how much our predictions deviate from the desired answers.

  $$\text{Mean-squared error } J_n = \frac{1}{n} \sum_{i=1 \ldots n} (y_i - f(x_i))^2$$

Advantages :

a. Training a linear regression model is usually much faster than methods such as neural networks.

b. Linear regression models are simple and require minimum memory to implement.

c. By examining the magnitude and sign of the regression coefficients you can infer how predictor variables affect the target outcome.

7.5.2 Multiple Linear Regression

- **Multiple linear regression** is an extension of linear regression, which allows a response variable, y, to be modeled as a linear function of two or more predictor variables.
- In a multiple regression model, two or more independent variables, i.e. predictors are involved in the model. The simple linear regression model and the multiple regression model assume that the dependent variable is continuous.

Difference between simple and multiple regression :

Sr. No.	Simple regression	Multiple regression
1.	One dependent variable Y predicted from one independent variable X.	One dependent variable Y predicted from a set of independent variables $(X_1, X_2, \ldots, X_k)$
2.	One regression coefficient.	One regression coefficient for each independent variable.
3.	r^2 : Proportion of variation in dependent variable Y predictable from X.	R^2 : Proportion of variation in dependent variable Y predictable by set of independent variables (X's).

7.5.3 Logistic Regression

- Logistic regression is a form of regression analysis in which the outcome variable is binary or dichotomous. A statistical method used to model dichotomous or binary outcomes using predictor variables.
- **Logistic component :** Instead of modeling the outcome, Y, directly, the method models the log odds (Y) using the logistic function.

- **Regression component :** Methods used to quantify association between an outcome and predictor variables. It could be used to build predictive models as a function of predictors.
- **In simple** logistic regression, logistic regression with 1 predictor variable.

$$\ln\left(\frac{P(Y)}{1-P(Y)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- With logistic regression, the response variable is an indicator of some characteristic, that is, a 0/1 variable. Logistic - regression is used to determine whether other measurements are related to the presence of some characteristic, for example, whether certain blood measures are predictive of having a disease.
- Fig. 7.5.2 shows Sigmoid curve for logistic regression.

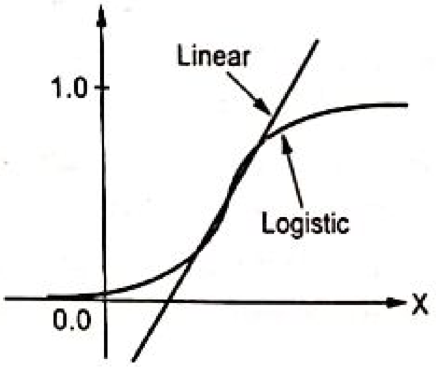


Fig. 7.5.2

- If analysis of covariance can be said to be test adjusted for other variables, then logistic regression can be thought of as a chi-square test for homogeneity of proportions adjusted for other variables. While the response variable in a logistic regression is a 0/1 variable, the logistic regression equation, which is a linear equation, does not predict the 0/1 variable itself.

7.5.4 Lasso and Ride Regression

- Ridge regression and the Lasso are two forms of regularized regression. These methods are seeking to improve the consequences of multicollinearity.
  1. When variables are highly correlated, a large coefficient in one variable may be alleviated by a large coefficient in another variable, which is negatively correlated to the former.
  2. Regularization imposes an upper threshold on the values taken by the coefficients, thereby producing a more parsimonious solution and a set of coefficients with smaller variance.
- Ridge estimation produces a biased estimator of the true parameter β.

$$E[\hat{\beta}^{ridge}|X] = (X^TX+\lambda I)^{-1}X\beta$$

$$= (X^TX+\lambda I)^{-1}(X^TX+\lambda I-\lambda I)\beta$$

$$= [I-\lambda(X^TX+\lambda I)^{-1}]\beta$$

$$= \beta-\lambda(X^TX+\lambda I)^{-1}\beta$$

- Ridge regression shrinks the regression coefficients by imposing a penalty on their size. The ridge coefficients minimize a penalized residual sum of squares.
- Ridge regression protects against the potentially high variance of gradients estimated in the short directions.

Lasso :

- One significant problem of ridge regression is that the penalty term will never force any of the coefficients to be exactly zero. Thus, the final model will include all p predictors, which creates a challenge in model interpretation. A more modern machine learning alternative is the lasso.
- The lasso works in a similar way to ridge regression, except it uses a different penalty term that shrinks some of the coefficients exactly to zero.
- **Lasso :** Lasso is a regularized regression machine learning technique that avoids over-fitting of training data and is useful for feature selection.

7.6 Fill in the Blanks

Q.1 The ________ of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data.

Q.2 Decision tree induction is the learning of decision trees from ________ training tuples.

Q.3 A ________ is a flowchart - like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test and each leaf node holds a class label.

Q.4 ________ uses information gain as its attribute selection measure.

Q.5 If we were to use the training set to estimate the error rate of a model, this quantity is known as the ________ error.

Q.6 CART stands for ________.

Q.7 A ________ set of class - labeled tuples is used to estimate cost complexity.

Q.8 A ________ network has one Conditional Probability Table (CPT) for each variable.

Q.9 True positives refer to the positive tuples that were correctly labeled by the ________.

Q.10 A belief network is defined by two components : a ______ and a set of ______.

Q.11 ______ classifiers use distance - based comparisons that intrinsically assign equal weight to each attribute.

Q.12 The Naive Bayesian classifier is based on ______ theorem with the independence assumptions between predictors.

7.7 Multiple Choice Questions

Q.1 The individual tuples making up the training set are referred to as ______ and are selected from the database under analysis.

a learning tuples

b training tuples

c samples

d database

Q.2 A ______ is a flowchart - like tree structure, where each internal node denotes a test on an attribute.

a decision tree

b binary tree

c cluster

d none of these

Q.3 ID3 stands for ______.

a induction decision tree

b iterative database

c iterative Dichotomiser

d iterative decision tree

Q.4 ID3 uses ______ as its attribute selection measure.

a decision tree

b Gini index

c information gain

d attributes

Q.5 Attribute selection measures based on the ______ principle have the least bias toward multi - valued attribute.

a maximum description length

b minimum description length

c minimum distance length

d maximum distance length

Q.6 In theory, Bayesian classifiers have the ________ error rate in comparison to all other classifiers.

a equal

b maximum

c minimum

d zero

Answer Keys for Fill in the Blanks

Q.1	accuracy	Q.2	class - labeled
Q.3	decision tree	Q.4	ID3
Q.5	resubstitution	Q.6	classification and regression trees
Q.7	pruning	Q.8	belief
Q.9	classifier	Q.10	directed acyclic graph, conditional probability tables
Q.11	Nearest-neighbor	Q.12	Bayes'

Answer Keys for Multiple Choice Questions

Q.1	b	Q.2	a
Q.3	c	Q.4	c
Q.5	b	Q.6	c

8 Unsupervised Learning

Syllabus

Supervised vs. Unsupervised Learning, Applications, Clustering, Association rules

Contents

8.1 Difference between Supervised and Unsupervised Learning

Sr. No.	Supervised Learning	Unsupervised Learning
1.	Desired output is given.	Desired output is not given.
2.	It is not possible to learn larger and more complex models than with supervised learning.	It is possible to learn larger and more complex models with unsupervised learning.
3.	Use training data to infer model.	No training data is used.
4.	Every input pattern that is used to train the network is associated with an output pattern.	The target output is not presented to the network.
5.	Trying to predict a function from labeled data.	Try to detect interesting relations in data.
6.	Supervised learning requires that the target variable is well defined and that a sufficient number of its values are given.	For unsupervised learning typically either the target variable is unknown or has only been recorded for too small a number of cases.
7.	Example : Optical character recognition	Example : Find a face in an image.
8.	We can test our model	We can not test our model
9.	Supervised learning is also called classification.	Unsupervised learning is also called clustering.

8.2 Applications of Unsupervised Learning

- The main applications of unsupervised learning include clustering, visualization, dimensionality reduction, finding association rules, and anomaly detection.
- Clustering allows you to automatically split the dataset into groups according to similarity. Often, however, cluster analysis overestimates the similarity between groups and doesn't treat data points as individuals. For this reason, cluster analysis is a poor choice for applications like customer segmentation and targeting.
- Anomaly detection can automatically discover unusual data points in your dataset. This is useful in pinpointing fraudulent transactions, discovering faulty pieces of hardware, or identifying an outlier caused by a human error during data entry.
- Association mining identifies sets of items that frequently occur together in your dataset. Retailers often use it for basket analysis, because it allows analysts to discover goods often purchased at the same time and develop more effective marketing and merchandising strategies.

- Latent variable models are commonly used for data preprocessing, such as reducing the number of features in a dataset (dimensionality reduction) or decomposing the dataset into multiple components.
- Segmentation of target consumer populations by an advertisement consulting agency on the basis of few dimensions such as demography, financial data, purchasing habits, etc. so that the advertisers can reach their target consumers efficiently
- Anomaly or fraud detection in the banking sector by identifying the pattern of loan defaulters.
- Image processing and image segmentation such as face recognition, expression identification, etc.
- Grouping of important characteristics in genes to identify important influencers in new areas of genetics.
- Utilization by data scientists to reduce the dimensionalities in sample data to simplify modelling.
- Document clustering and identifying potential labelling options.

8.3 Clustering

- Given a set of objects, place them in groups such that the objects in a group are similar (or related) to one another and different from (or unrelated to) the objects in other groups.
- Cluster analysis can be a powerful data-mining tool for any organization that needs to identity discrete groups of customers, sales transactions, or other types of behaviors and things. For example, insurance providers use cluster analysis to detect fraudulent claims and banks used it for credit scoring.
- Cluster analysis uses mathematical models to discover groups of similar customers based on the smallest variations among customers within each group.
- Cluster is a group of objects that belong to the same class. In another words the similar object are grouped in one cluster and dissimilar are grouped in other cluster.
- Clustering is a process of partitioning a set of data in a set of meaningful subclasses. Every data in the sub class shares a common trait. It helps a user understand the natural grouping or structure in a data set.
- Various types of clustering methods are partitioning methods, Hierarchical clustering, Fuzzy clustering, Density based clustering and Model based clustering.
- Cluster anlysis is process of grouping a set of data objects into clusters.

- **Desirable properties of a clustering algorithm are as follows :**

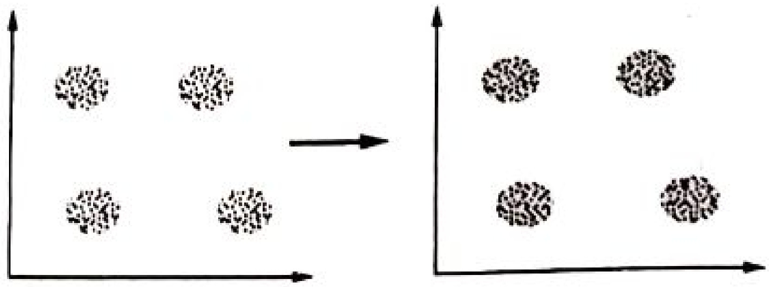


Fig. 8.3.1

1. Scalability (in terms of both time and space)
2. Ability to deal with different data types
3. Minimal requirements for domain knowledge to determine input parameters.
4. Interpretability and usability.

- Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. Clustering can be considered the most important unspervised learning problem.
- A cluster is therefore a collection of objects which are "similar" between them and are dissimilar" to the objects belonging to other clusters. Fig. 8.3.1 shows cluster.
- In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance : two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called **distance-based clustering.**

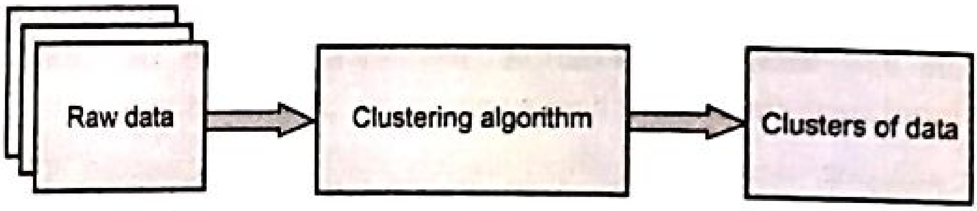


Fig. 8.3.2

- Clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.
- A clustering algorithm attempts to find natural groups components or data based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets.
- To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroids. The output from a clustering algorithm is

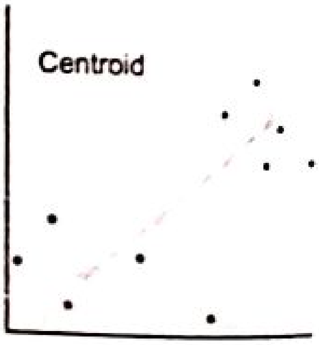


Fig. 8.3.3

basically a statistical description of the cluster centroids with the number of components in each cluster.

- **Cluster centroid** : The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the cluster. Each cluster has a well defined centroid.
- **Distance :** The distance between two points is taken as a common metric to as see the similarity among the components of population. The commonly used distance measure is the euclidean metric which defines the distance between two points $p = (p_1, p_2, ...)$ and $q = (q_1, q_2, ...)$ is given by.

$$d = \sum_{i=1}^{k} (p_i - q_i)^2$$

- The goal of clustering is to determine the intrinsic grouping in a set of unlableled data. But how to decide what constitutes a good clustering ? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply criterion, in such a way that the result of the clustering will suit their needs.
- Clustering analysis helps construct meaninful partitioning of a large set of objects Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing etc.
- Clustering algorithms may be classified as listed below :
  1. Exclusive clustering
  2. Overlapping clustering
  3. Hierarchical clustering
  4. Probabilisitic clustering
- A good clustering method will produce high quality clusters high intra-class similarlity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. The quality of a clustering method is also measured by it's ability to discover some or all of the hidden patterns.
- Clustering techniques types : The major clustering techniques are,
  a) Partitioning methods
  b) Hierarchical methods
  c) Density - based methods.

8.3.1 Partitioning Methods

- Partitioning clustering are clustering methods used to classify observations, within a data set, into multiple groups based on their similarity. The algorithms require the analyst to specify the number of clusters to be generated.
- Commonly used partitioning methods are k-means and k-medoids.
- In the k-means algorithm, the centroid of the prototype is identified for clustering, which is normally the mean of a group of points. Similarly, the k-medoid algorithm identifies the medoid which is the most representative point for a group of points.

8.3.1.1 K - mean Clustering

- K-Means clustering is heuristic method. Here each cluster is represented by the center of the cluster. "K" stands for number of clusters, it is typically a user input to the algorithm; some criteria can be used to automatically estimate K.
- This method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart.
- Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated everytime a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.
- Given K, the K-means algorithm consists of four steps :
  1. Select initial centroids at random.
  2. Assign each object to the cluster with the nearest centroid.
  3. Compute each centroid as the mean of the objects assigned to it.
  4. Repeat previous 2 steps until no change.
- The $x_1, \dots, x_N$ are data points or vectors of observations. Each observation (vector x_i) will be assigned to one and only one cluster. The C(i) denotes cluster number for the i^{th} observation. K-means minimizes within-cluster point scatter :

$$W(C) = \frac{1}{2} \sum_{K=1}^{K} \sum_{C(i)=K} \sum_{C(j)=K} \| x_i - x_j \|^2$$

$$= \sum_{K=1}^{K} N_k \sum_{C(i)=K}^{K} \| x_i - m_K \|^2$$

where

m_K is the mean vector of the K^{th} cluster.

N_K is the number of observations in K^{th} cluster.

K-Means Algorithm Properties

1. There are always K clusters.
2. There is always at least one item in each cluster.
3. The clusters are non-hierarchical and they do not overlap.
4. Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

The K-Means Algorithm Process

1. The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
2. For each data point.
   a. Calculate the distance from the data point to each cluster.
   b. If the data point is closest to its own cluster, leave it where it is.
   c. If the data point is not closest to its own cluster, move it into the closest cluster.
3. Repeat the above step until a complete pass through all the data points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
4. The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.

- K-means algorithm is iterative in nature. It converges, however only a local minimum is obtained. It works only for numerical data. This method easy to implement.
- **Advantages of K-Means Algorithm :**
  1. Efficient in computation
  2. Easy to implement
- **Weaknesses**
  1. Applicable only when *mean* is defined.
  2. Need to specify K, the *number* of clusters, in advance.
  3. Trouble with noisy data and *outliers*.
  4. Not suitable to discover clusters with *non-convex shapes*.

8.3.1.2 k-Medoids

- The K-medoids algorithm is a clustering algorithm related to the K-means algorithm and the medoidshift algorithm. K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into K clusters known a priori. A useful tool for determining K is the silhouette.

- The most common realisation of K-medoid clustering is the Partitioning Around Medoids (PAM) algorithm. PAM uses a greedy search which may not find the optimum solution, but it is faster than exhaustive search.
- Instead of taking the mean value of the object in a cluster as a reference point, medoids can be used, which is the most centrally located object in a cluster.
- A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal i.e. it is a most centrally located point in the given data set.

  1. The algorithm begins with arbitrary selection of the K objects as medoid points out of n data points (n > K)
  2. After selection of the k medoid points, associate each data object in the given data set to most similar medoid. The similarity here is defined using distance measure that can be euclidean distance, manhattan distance or minkowski distance
  3. Randomly select nonmedoid object O'
  4. Compute total cost , S of swapping initial medoid object to O'
  5. If S < 0, then swap initial medoid with the new one (if S < 0 then there will be new set of medoids)
  6. Repeat steps 2 to 5 until there is no change in the medoid.

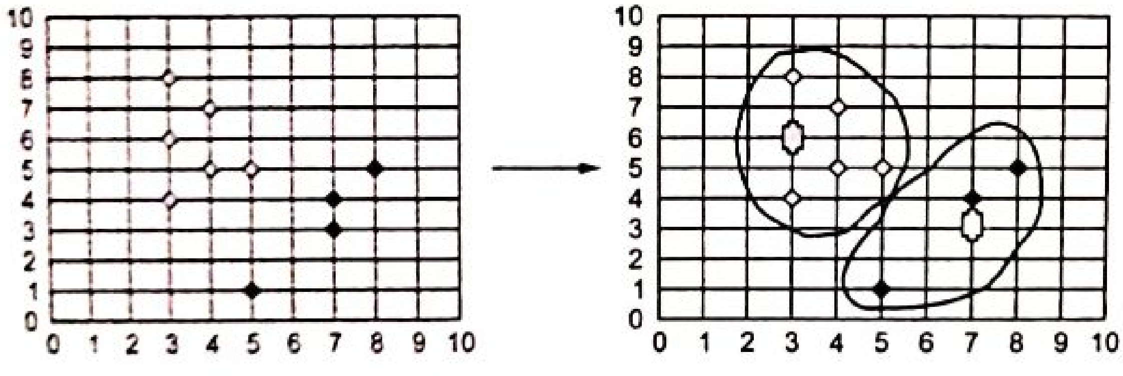


Fig. 8.3.4

8.3.2 Hierarchical Methods

- This method use distance matrix as clustering criteria. This method does not require the number of clusters K as an input, but needs a termination condition. Hierarchical clustering is a widely used data analysis tool.

- The idea is to build a binary tree of the data that successively merges similar groups of points. Visualizing this tree provides a useful summary of the data.
- Hierarchical clustering arranges items in a hierarchy with a treelike structure based on the distance or similarity between them. The graphical representation of the resulting hierarchy is a tree-structured graph called a **dendrogram.**
- Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion.

Agglomerative hierarchical clustering

- This bottom-up strategy starts by placing each object in its own cluster and then merges these atomic clusters into larger and larger clusters, until all of the objects are in a single cluster or until certain termination conditions are satisfied. Most hierarchical clustering methods belong to this category.
- Initially, AGNES places each objects into a cluster of its own. The clusters are then merged step-by-step according to some criterion. For example, cluster C_1 and C_2 may be merged if an object in C_1 and object in C_2 form the minimum Euclidean distance between any two objects from different clusters.

In the agglomerative hierarchical approach, we start by defining each data point to be a cluster and combine existing clusters at each step. Here are four different methods for doing this :

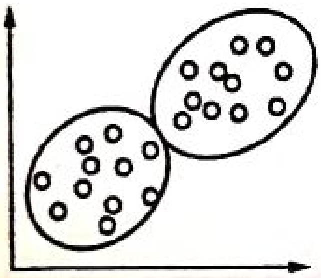
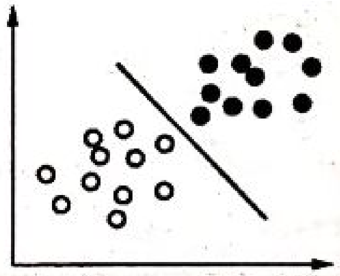
1. **Single linkage :** Smallest pairwise distance between elements from each cluster
2. **Complete linkage :** Largest distance between elements from each cluster
3. **Average linkage :** The average distance between elements from each cluster
4. **Centroid linkage :** Distance between cluster means

Divisive Hierarchical Clustering

This top-down strategy does the reverse of agglomerative hierarchical clustering by starting with all objects in one cluster. It subdivides the clusters into smaller and smaller pieces, until each object form a cluster on its own or until it satisfies certain termination conditions, such as a desired number of cluster or the diameter of each cluster is within a certain threshold.

Agglomerative	Divisive Hierarchical Clustering
Initially each item in its own cluster.	Initially all items in one cluster.
Iteratively clusters are merged together.	Large clusters are successively divided.
Bottom up.	Top down.

8.3.2.1 Difference between Clustering vs Classification

Clustering	Classification
This function maps the data into one of several clusters which is the grouping of data items based on the similarities between them.	This model function classifies the data into one of several predefined categorical classes.
Involved in unsupervised learning.	Involved in supervised learning .
Training sample is not provided.	Training sample is provided.
The number of cluster is not known before clustering. These are identified after the completion of clustering.	The number of classes is known before classification as there is predefined output based on input data.
Data is not labeled.	Labeled data points.
Asks how can I group this set of items?	Asks what class does this item belong to?
Unknown number of classes.	Known number of classes.
Used to understand data.	Used to classify future observations.
	

8.4 Association Rules

- Association rule presents a methodology that is useful for identifying interesting relationships hidden in large data sets. It is also known as association analysis.
- Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories.
- Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a transactional database, relational database or other information repository.

- An example of an association rule would be "If a customer buys a dozen eggs, he is 80 % likely to also purchase milk."
- Association rule mining can be viewed as a two-step process :
  1. Find all frequent item sets : By definition, each of these item sets will occur at least as frequently as a predetermined minimum support count, min sup.
  2. Generate strong association rules from the frequent item sets : By definition, these rules must satisfy minimum support and minimum confidence.
- An association rule is commonly understood to be an expression of the form : $X \Rightarrow Y$ where X and Y are sets of items such that $X \cap Y = \phi$.
- The association rule $X \Rightarrow Y$ means that transactions containing items from set X tend to contain items from set Y.
- Association rules show attribute value conditions that occur frequently together in a given data set. A typical example of association rule mining is Market Basket Analysis.
- Data is collected using bar-code scanners in supermarkets. Such market basket databases consist of a large number of transaction records.
- Each record lists all items bought by a customer on a single purchase transaction. Managers would be interested to know if certain groups of items are consistently purchased together.
- They could use this data for adjusting store layouts, for cross-selling, for promotions, for catalog design, and to identify customer segments based on buying patterns.
- Association rules provide information of this type in the form of if-then statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules are probabilistic in nature.
- In addition to the antecedent (if) and the consequent (then), an association rule has two numbers that express the degree of uncertainty about the rule.
- In association analysis, the antecedent and consequent are sets of items (called itemsets) that are disjoint (do not have any items in common).
- The first number is called the support for the rule. The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule.
- The other number is known as the confidence of the rule. Confidence is the ratio of the number of transactions that include all items in the consequent, as well as the antecedent (the support) to the number of transactions that include all items in the antecedent.

- **Market basket analysis** is an example of frequent itemset mining. The purpose of market basket analysis is to determine what products customers purchase together. It takes its name from the idea of customers throwing all their purchases into a shopping cart (a "market basket") during grocery shopping.
- Market Basket Analysis is a technique which identifies the strength of association between pairs of products purchased together and identify patterns of co-occurrence. A co-occurrence is when two or more things take place together.
- Market basket analysis takes data at transaction level, which lists all items bought by a customer in a single purchase.
- The technique determines relationships of what products were purchased with which other product(s). These relationships are then used to build profiles containing If-Then rules of the items purchased.
- The rules could be written as : **If {A} Then {B}**
- The If part of the rule (the {A} above) is known as the antecedent and the THEN part of the rule is known as the consequent (the {B} above).
- The antecedent is the condition and the consequent is the result. The association rule has three measures that express the degree of confidence in the rule, Support, Confidence, and Lift.
- For example, you are in a supermarket to buy milk. Based on the analysis, are you more likely to buy apples or cheese in the same transaction than somebody who did not buy milk ?
- In the following table, there are nine baskets containing varying combinations of milk, cheese, apples, and bananas.

Basket	Product 1	Product 2	Product 3
1	Milk	Cheese	
2	Milk	Apples	Cheese
3	Apples	Banana	
4	Milk	Cheese	
5	Apples	Banana	
6	Milk	Cheese	Banana
7	Milk	Cheese	
8	Cheese	Banana	
9	Cheese	Milk	

- **Support :** Support is the number of transactions that include items in the {A} and {B} parts of the rule as a percentage of the total number of transactions. It is a measure of how frequently the collection of items occur together as a percentage of all transaction.

$$\text{Support} = \frac{A+B}{\text{Total}}$$

- **Confidence :** Confidence of the rule is the ratio of the number of transactions that include all items in {B} as well as the number of transactions that include all items in {A} to the number of transactions that include all items in {A}.

$$\text{Confidence} = \frac{A+B}{A}$$

- **Lift or Lift ratio :** It is the ratio of confidence to expected confidence. Expected confidence is the confidence divided by the frequency of B. The Lift tells us how much better a rule is at predicting the result than just assuming the result in the first place. Greater lift values indicate stronger associations.

$$\text{Lift Ratio} = \frac{(A+B)}{(B/\text{Total})} = \frac{\text{Cofidence}}{(B/\text{Total})}$$

- **Leverage :** Leverage measures the difference in the probability of X and Y appearing together compared to statistical independence.

  Leverage $(X \rightarrow Y)$ = Support $(X \wedge Y)$ – Support$(X)^*$ Support(Y)
- Leverage = 0 if X and Y are statistically independent
- Leverage > 0 indicates degree of usefulness of rule
- Conviction : The conviction of a rule is defined as:

$$\text{conv } (X \rightarrow Y) = \frac{1-\text{supp}(Y)}{1-\text{conf}(X \rightarrow Y)}$$

The conviction of the rule **X=>Y** can be interpreted as the ratio of the expected frequency that **X** occurs without **Y** if **X** and **Y** were independent divided by the observed frequency of incorrect predictions

8.4.1 Frequent Itemsets and Closed Itemsets

1. A set of items is referred to as an itemset. An itemset is a unordered set of distinct items. An itemset that contains k items is a k-itemset.
2. The set {computer, antivirus software} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known as the frequency, support count, or count of the itemset.

3. Frequent itemsets that cannot be extended with any item without making them infrequent are called maximal frequent itemsets. Exact support counts of the subsets cannot be directly derived from support of the maximal frequent itemset.

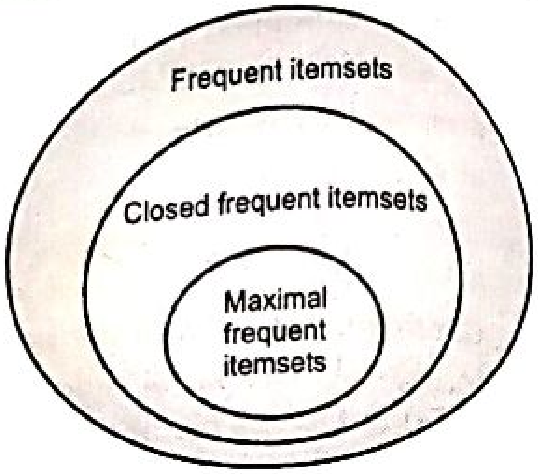


Fig. 8.4.1

Closed itemsets :

- An alternative approach is to try to retain some of the support information in the compacted representation.
- A closed itemset is an itemset whose all immediate supersets have different support count.
- A closed frequent itemset is a closed itemset that satisfies the minimum support threshold.
- Maximal frequent itemsets are closed by definition.
- An itemset X is closed in a data set S if there exists no proper super-itemset Y such that Y has the same support count as X in S. An itemset X is a closed frequent itemset in set S if X is both closed and frequent in S.
- An itemset is closed if none of its immediate supersets has the same support as the itemset.
- Closed itemset example 1 :

TID	Items
1	{A, B}
2	{B, C, D}
3	{A, B, C, D}
4	{A, B, D}
5	{A, B, C, D}

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A, B}	4
{A, C}	2
{A, D}	3
{B, C}	3
{B, D}	4
{C, D}	3

Itemset	Support
{A, B, C}	2
{A, B, D}	3
{A, C, D}	2
{B, C, D}	3
{A, B, C, D}	2

- Closed itemset are : {B}, {A, B}, {B, D}, {A, B, D}, {B, C, D}, {A, B, C, D}
- ·**Closed itemset example 2 :**

TID	Items
100	a, c, d, e, f
200	a, b, e
300	c, e, f
400	a, c, d, f
500	c, e, f

- Total Frequent itemsets : 20

{a}, {c}, {d}, {e}, {f}, {a, c} {a, d}, {a, e}, {a, f}, {c, d}, {c, e}, {c, f}, {d, f}, {e, f}, {a, c, d}, {a, c, f}, {a, d, f}, {c, d, f}, {c, e, f} {a, c, d, f}

Closed frequent itemsets :

{a, c, d, f}, {c, e, f}, {a, e}, {c, f}, {a}, {e}

8.4.2 The Apriori Algorithm

- The Apriori algorithmis an influential algorithm for mining frequent itemsets for boolean association rules.
- Innovative way to find association rules on large scale, allowing implication outcomes that consist of more than one item. It based on minimum support threshold.
- Let $I = \{i_1, i_2, ... i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, ..., t_n\}$ be a set of transactions called the database. Each transaction in D has a unique transaction ID and contains a subset of the items in I.
- A rule is defined as an implication of the form $X \rightarrow Y$ where $X, Y \subseteq I$ and $X > Y = \phi$. The sets of items X and Y are called antecedent (left-hand-side or LHS) and consequent (right-hand-side or RHS) of the rule respectively.
- The Apriori algorithm is used for mining frequent itemsets and devising association rules from a transactional database. The parameters "support" and "confidence" are used.
- Support refers to items' frequency of occurrence; confidence is a conditional probability.
- To improve the efficiency of level-wise generation of frequent itemsets, an important property is used called Apriori property which helps by reducing the search space.

- Major components of Apriori algorithm are **Support, Confidence and Lift.**
- The following are the main steps of the Apriori algorithm:
  1. Calculate the support of item sets (of size k = 1) in the transactional database. This is called generating the candidate set.
  2. Prune the candidate set by eliminating items with a support less than the given threshold.
  3. Join the frequent itemsets to form sets of size k + 1, and repeat the above sets until no more itemsets can be formed. This will happen when the set(s) formed have a support less than the given support.

Pseudocode of Apriori algorithm

```
$L_1$= {frequent items};
for (k= 2; $L_{k-1}$ !=ϕ; k++) do begin
    $C_k$ = candidates generated from $L_{k-1}$ (that is: cartesian product $L_{k-1} \times L_{k-1}$ and
            eliminating any k-1 size itemset that is not frequent);
    for each transaction t in database do
            increment the count of all candidates in
            $C_k$ that are contained in t
    $L_k$ = candidates in $C_k$ with min_sup
    end
return $C_k$ $L_k$;
```

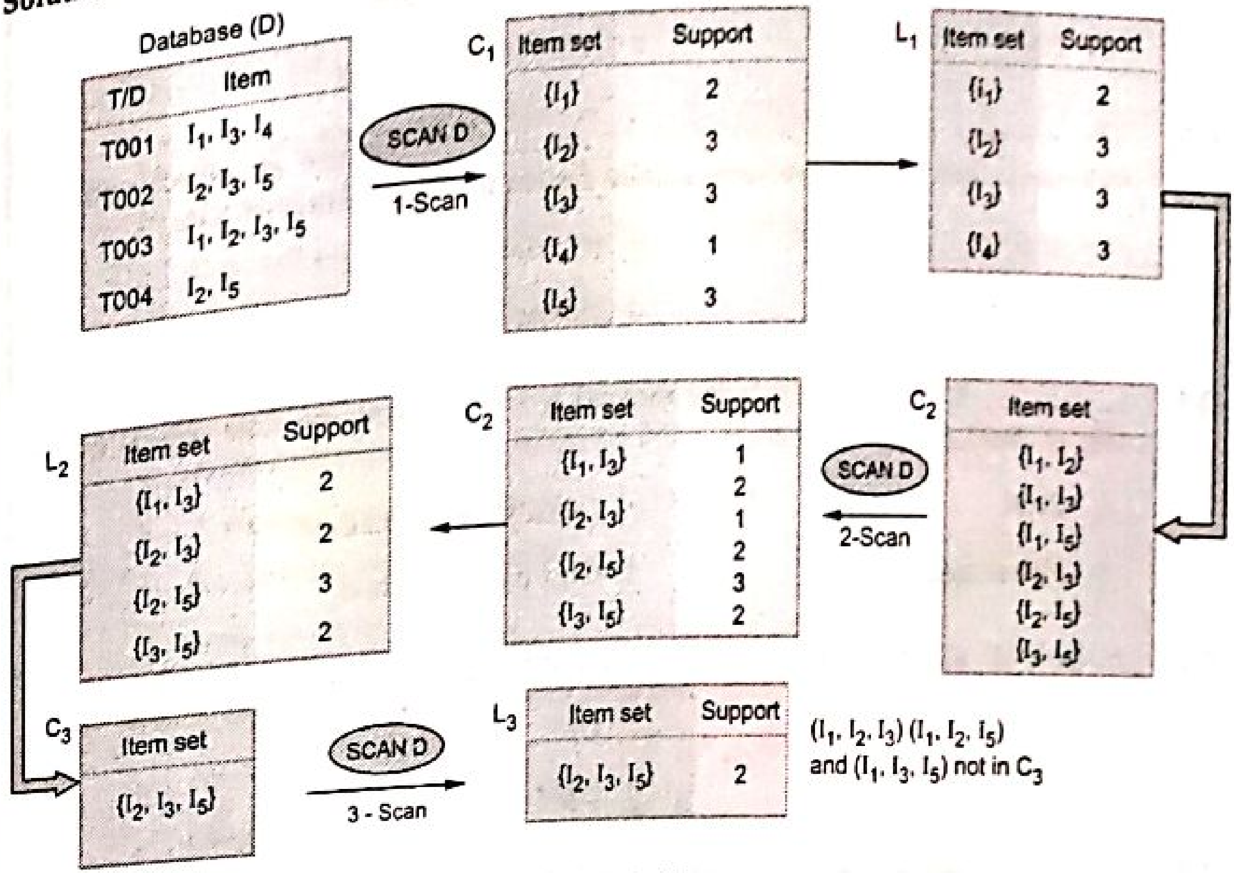
limitations of Apriori algorithm

1. Needs several iterations of the data.
2. Uses a uniform minimum support threshold.
3. Difficulties to find rarely occuring events.
4. Some competing alternative approaches focus on partition and sampling

Example 8.4.1 *Generate frequent itemsets and generate association rules based on it using aprori algorithm. Minimum support is 50 % and minimum confidence is 70 %.*

TID	Items
100	1, 3, 4
200	2, 3, 5
300	1, 2, 3, 5
400	2, 5

Solution : Apriori algorithm :



Database (D)

T/D	Item
T001	I_1, I_3, I_4
T002	I_2, I_3, I_5
T003	I_1, I_2, I_3, I_5
T004	I_2, I_5

C_1

Item set	Support
$\{I_1\}$	2
$\{I_2\}$	3
$\{I_3\}$	3
$\{I_4\}$	1
$\{I_5\}$	3

L_1

Item set	Support
$\{I_1\}$	2
$\{I_2\}$	3
$\{I_3\}$	3
$\{I_4\}$	3

C_2

Item set
$\{I_1, I_2\}$
$\{I_1, I_3\}$
$\{I_1, I_5\}$
$\{I_2, I_3\}$
$\{I_2, I_5\}$
$\{I_3, I_5\}$

C_2

Item set	Support
$\{I_1, I_2\}$	1
	2
$\{I_2, I_3\}$	1
	2
$\{I_2, I_5\}$	3
$\{I_3, I_5\}$	2

L_2

Item set	Support
$\{I_1, I_3\}$	2
$\{I_2, I_3\}$	2
$\{I_2, I_5\}$	3
$\{I_3, I_5\}$	2

C_3

Item set
$\{I_2, I_3, I_5\}$

L_3

Item set	Support
$\{I_2, I_3, I_5\}$	2

(I_1, I_2, I_3) (I_1, I_2, I_5) and (I_1, I_3, I_5) not in C_3

Fig. 8.4.2

8.5 Fill in the Blanks

Q.1 ________ clustering is a bottom-up technique which starts with individual objects as clusters and then iteratively merges them to form larger clusters.

Q.2 k-means and k-medoids are the most popular   techniques.

Q.3 Market basket analysis is an example of ________ mining.

Q.4 Major components of ________ algorithm are Support, Confidence and Lift.

Q.5 DBSCAN is one of the ________ clustering approaches that provide a solution to identify clusters of arbitrary shapes.

Q.6 The result of the ________ analysis is expressed as a set of association rules that specify patterns of relationships among items.

8.6 Multiple Choice Questions

Q.1 PAM stands for ________

a Partitioning Around Medoids b Partitioning Around Method

c Perception Around Medoids d All of these

Q.2 Which of the following is hierarchical clustering method ?

a Agglomerative b Divisive clustering

c PAM d A and B

Q.3 The k-means algorithm is sensitive to ________ because an object with an extremely large value may substantially distort the distribution of data.

a outliers b text data

c boasting d cluster

Q.4 ________ hierarchical clustering method works by grouping data objects into a tree of clusters.

a PAM b Density-based method

c Hierarchical d Grid-based method

Q.5 In DIANA, all of the objects are used to form ________ initial cluster.

a one b two

c four d eight

Q.6 If the clustering process is terminated when the distance between nearest clusters exceeds an arbitrary threshold, it is called a ________.

a dendrogram

b nearest-neighbor clustering algorithm

c minimal spanning tree algorithm

d single-linkage algorithm

Q.7 What are closed itemsets ?

a An itemset for which at least one proper super-itemset has same support

b An itemsetwhose no proper super-itemset has same support

c An itemset for which at least super-itemset has same confidence

d An itemsetwhose no proper super-itemset has same confidence

Q.8 What are maximal frequent itemsets ?

a A frequent itemsetwhose no super-itemset is frequent

b A frequent itemset whose super-itemset is also frequent c A non-frequent itemset whose super-itemset is frequent

d None of the above

Q.9 What does FP growth algorithm do ?

a It mines all frequent patterns through pruning rules with lesser support

b It mines all frequent patterns through pruning rules with higher support

c It mines all frequent patterns by constructing a FP tree

d All of the above

Q.10 What is the relation between candidate and frequent itemsets ?

a A candidate itemset is always a frequent itemset

b A frequent itemset must be a candidate itemset

c No relation between the two

d Both are same

Q.11 Which of these is not a frequent pattern mining algorithm ?

a Apriori b FP growth

c Decision trees d Eclat

Q.12 What will happen if support is reduced ?

a Number of frequent itemsets remains same

b Some itemsets will add to the current set of frequent itemsets

c Some itemsets will become infrequent while others will become frequent

d Can not say

Q.13 What are maximal frequent itemsets ?

a A frequent itemsetwhose no super-itemset is frequent

b A frequent itemset whose super-itemset is also frequent

c A non-frequent itemset whose super-itemset is frequent

d None of the above

Answer Keys for Fill in the Blanks

Q.1	Agglomerative	Q.2	partitioning	Q.3	frequent itemset
Q.4	Apriori	Q.5	density-based	Q.6	market basket

Answer Keys for Multiple Choice Questions

Q.1	a	Q.2	d	Q.3	a
Q.4	c	Q.5	a	Q.6	d
Q.7	b	Q.8	a	Q.9	c
Q.10	b	Q.11	c	Q.12	b
Q.13	a				

9 Neural Network

Syllabus

Introduction to neural network, Biological and Artificial Neurons, Types of Activation functions, Implementation of ANN, Architecture, Leaning process, Backpropogation, Deep Learning.

Contents

9.1 Introduction to Neural Network

- Neural networks consists of many numbers of simple elements (neurons) connected between them in system. Whole system is able to solve of complex tasks and to learn for it like a natural brain.
- Neural network is a system composed of many simple processing elements operating in parallel whose function is determined by network structure, connection strengths, and the processing performed at computing elements or nodes.
- For user, NN is black box with input vector (source data) and output vector (result).
  - A Neural network is usually structured into an input layer of neurons, one or more hidden layers one output layer.
  - Neurons belonging to adjacent layers are usually fully connected and the various types and architectures are indentified both by the different topologies adopted for the connections as well as by the choice of the activation function.
  - The values of the functions associated with the connections are called "weights".
  - The whole game of using NNs is in fact that, in order for the network to yield appropriate outputs for given inputs, the weight must be set to suitable values. The way this is obtained allows a further distinction among modes of operations.
  - A neural network is a processing device, either an algorithm or actual hardware, whose design was motivated by the design and functioning of human brains and components thereof.
  - Most neural networks have some sort of "training" rule whereby the weights of connections are adjusted on the basis of presented patterns.
  - In other words, neural networks "learn" from example, just like children learn to recognize dogs from examples of dogs, and exhibit some structural capability for generalization.
  - Neural networks normally have great potential for paralleism, since the computations of the components are independent of each other.
  - Neural networks are a different paradigm for computing :
    1. Von Neumann machines are based on the processing/memory abstraction of human information processing.
    2. Neural networks are based on the paralllel architecture of animal brains.
- Neural networks are a form of multiprocessor computer system, with

  a. Simple processing elements

b. A high degree of interconnections

c. Simple scalar messages

d. Adaptive interaction between elements

- The advantages of neural networks are due to its adaptive and generalization ability.
  a) Neural networks are adaptive methods that can learn without any prior assumption of the underlying data.
  b) Neural network, namely the feed forward multilayer perception and radial basis function network have been proven to be universal functional approximations.
  c) Neural networks are non-linear model with good generalization ability.
- **Useful properties and capabilities of neural network.**
  1. **Nonlinearity :** An artificial neuron can be linear or nonlinear. A neural network, made up of an interconnection of nonlinear neurons, is itself nonlinear.
  2. **Adaptivity :** Neural networks have a built-in capability to adapt their synaptic weights to changes in the surrounding environment.
  3. **Contextual information :** Knowledge is represented by the very structured and activation state of a neural network.
  4. **Evidential response :** In the context of pattern classification, a neural network can be designed to provide information not only about which particular pattern to select, but also about the confidence in the decision made.
  5. **Uniformity of analysis and design :** Neural networks enjoy universality as information processors.
  6. **VLSI implement-ability :** The massively parallel nature of a neural network makes it potentially fast for the computation of certain tasks.

9.1.1 Advantages of Neural Network

The advantages of neural networks are due to its adaptive and generalization ability.

a) Neural networks are adaptive methods that can learn without any prior assumption of the underlying data.

b) Neural network, namely the feed forward multilayer perception and radial basis function network have been proven to be universal functional approximations.

c) Neural networks are non-linear model with good generalization ability.

9.1.2 Application of Neural Network

Neural network applications can be grouped in following categories :

1. **Clustering :** A clustering algorithm explores the similarity between patterns and places similar patterns in a cluster. Best known applications include data compression and data mining.
2. **Classification/Pattern recognition :** The task of pattern recognition is to assign an input pattern (like handwritten symbol) to one of many classes. This category includes algorithmic implementations such as associative memory.
3. **Function approximation :** The tasks of function approximation is to find an estimate of the unknown function f() subject to noise. Various engineering and scientific disciplines require function approximation.
4. **Prediction/Dynamical systems :** The task is to forecast some future values of a time-sequenced data. Prediciton has a significant impact on decision support systems. Prediction differs from function approximation by considering time factor.

9.1.3 Difference between Digital Computer and Neural Networks

Sr. No.	Digital Computers	Neural Networks
1.	Deductive reasoning : We apply known rules input data to produce output.	Inductuve reasoning : Given input and output data (training examples), we construct the rules.
2.	Computation is centralized, synchronous and serial.	Coputation is collectivem asynchronous and parallel.
3.	Memory is packetted, literally stored and location addressable.	Memory is distributed, internalized and content addressable.
4.	Not fault toerant. One transistor goes and it no longer works.	Fault tolerant, redundancy and sharing of responsibilities.
5.	Fast. Measured in millionths of a second.	Slow. Measured in thousands of a second.
6.	Exact.	Inexact.
7.	Static connectivity.	Dynamic connectivity.
8.	Applicable if well defined rules with precise input data.	Applicable if rules are unknown or complicated or if data is noisy or partial.

9.2 Introduction of Artificial Neural Network

- Artificial Neural Network (ANN) is a computational system inspired by the structure, processing method, learning ability of a biological brain. An artificial neural network is composed of many artificial neurons that are linked together according to specific network architecture. The objective of the neural network is to transform the inputs into meaningful outputs.
- ANNs do not execute programmed instructions; they respond in parallel to the pattern of inputs presented to it. There are also no separate memory addresses for storing data. Instead, information is contained in the overall activation 'state' of the network. 'Knowledge' is thus represented by the network itself, which is quite literally more than the sum of its individual components.
- Fig 9.2.1 shows artificial neural network.

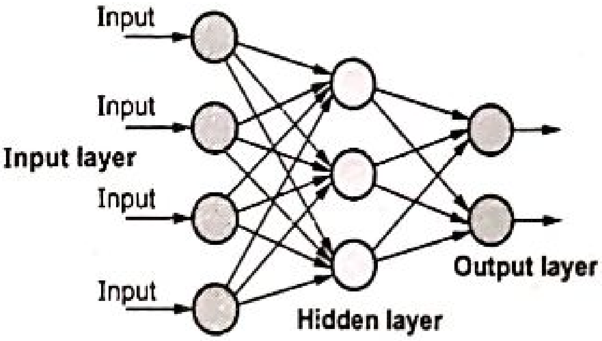


Fig. 9.2.1 Artificial neural network

- Elements of ANN are processing units, topology and learning algorithm.
- Tasks to be solved by artificial neural networks :
  1. Controlling the movements of a robot based on self-perception and other information;
  2. Deciding the category of potential food items in an artificial world;
  3. Recognizing a visual object;
  4. Predicting where a moving object goes, when a robot wants to catch it.
- Characteristics of artificial neural networks
  1. Large number of very simple processing neuron-like processing elements.
  2. Large number of weighted connections between the elements.
  3. Distributed representation of knowledge over the connections.
  4. Knowledge is acquired by network through a learning process.

Review Questions

1. *Write short note on : advantages of artificial neural network.*
2. *Write short note on : artificial neutral network application in communications.*
3. *List out the strength and weaknesses of artifical neural network.*
4. *List and explain performance issues of EBP.*
5. *List out the strength and weaknesses of EBP.*

9.3 Biological Neurons

- Artificial neural systems are inspired by biological neural systems. The elementary building block of biological neural systems is the neuron.
- The brain is a collection of about 10 billion interconnected neurons. Each neuron is a cell [right] that uses biochemical reactions to receive, process and transmit information. Fig. 9.3.1 shows biological neural systems.

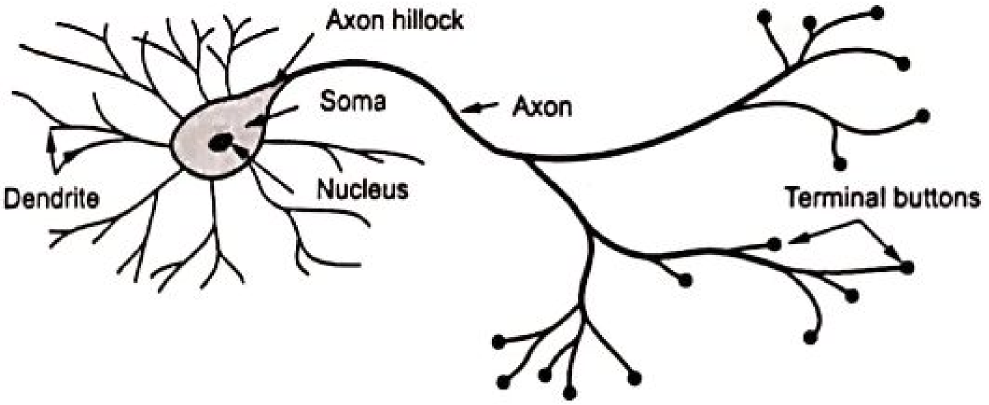


Fig. 9.3.1 Schematic of biological neuron

- The single cell neuron consists of the cell body or soma, the dendrites and the axon. The dendrites receive signals from the axons of other neurons. The small space between the axon of one neuron and the dendrite of another is the synapse. The afferent dendrites conduct impulses toward the soma. The efferent axon conducts impulses away from the soma.

Basic Components of Biological Neurons

1. The majority of ***neurons*** encode their activations or outputs as a series of brief electrical pulses (i.e. spikes or action potentials).
2. The neuron's ***cell body (soma)*** processes the incoming activations and converts them into output activations.

3. The neuron's ***nucleus*** contains the genetic material in the form of DNA. This exists in most types of cells, not just neurons.

4. ***Dendrites*** are fibres which emanate from the cell body and provide the receptive zones that receive activation from other neurons.

5. ***Axons*** are fibres acting as transmission lines that send activation to other neurons.

6. The junctions that allow signal transmission between the axons and dendrites are called ***synapses***. The process of transmission is by diffusion of chemicals called ***neurotransmitters*** across the synaptic cleft.

- Comparison between Biological NN and Artificial NN

Biological NN	Artificial NN
soma	unit
Axon, dendrite	dendrite
synapse	weight
potential	weighted sum
threshold	bias weight
signal	activation

9.4 Types of Activation Functions

- Activation functions also known as transfer function is used to map input nodes to output nodes in certain fashion.
- The activation function is the most important factor in a neural network which decided whether or not a neuron will be activated or not and transferred to the next layer.
- Activation functions help in normalizing the output between 0 to 1 or -1 to 1. It helps in the process of backpropagation due to their differentiable property. During backpropagation, loss function gets updated, and activation function helps the gradient descent curves to achieve their local minima.
- Activation function basically decides in any neural network that given input or receiving information is relevant or it is irrelevant.
- These activation function makes the multilayer network to have greater representational power than single layer network only when non-linearity is introduced.

- The input to the activation function is sum which is defined by the following equation.

$$\text{sum} = I_1 W_1 + I_2 W_2 + \ldots + I_n W_n = \sum_{j=1}^{n} I_j W_j + b$$

- **Activation Function : Logistic Function**

$$f(\text{sum}) = \frac{1}{(1 + e^{-s * \text{sum}})}$$

$$= (1 + s^{-s \times \text{sum}})^{-1}$$

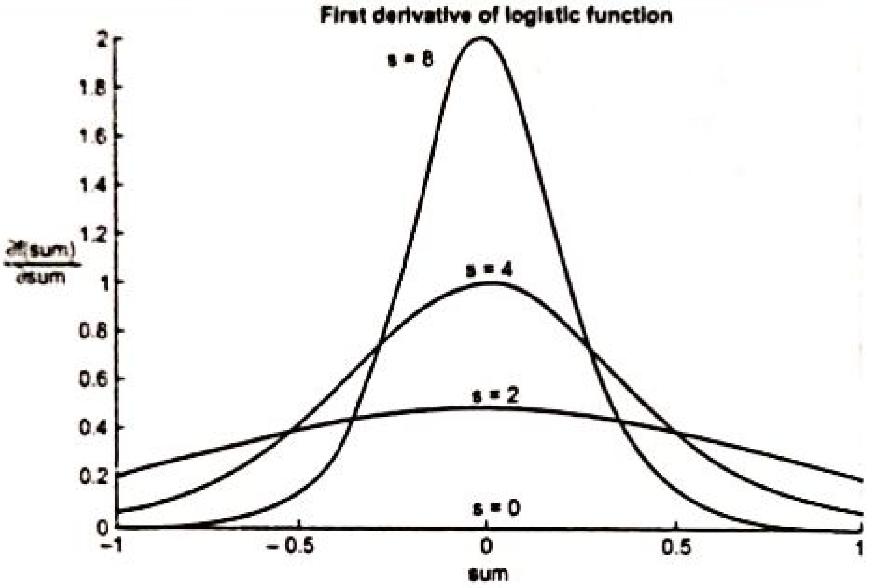


Fig. 9.4.1

- Logistic function monotonically increases from a lower limit (0 or - 1) to an upper limit (+1) as sum increases. In which values vary between 0 and 1, with a value of 0.5 when I is zero.

- **Activation Function : Arc Tangent**

$$f(\text{sum}) = \frac{2}{\pi} \tan^{-1} (s \times \text{sum})$$

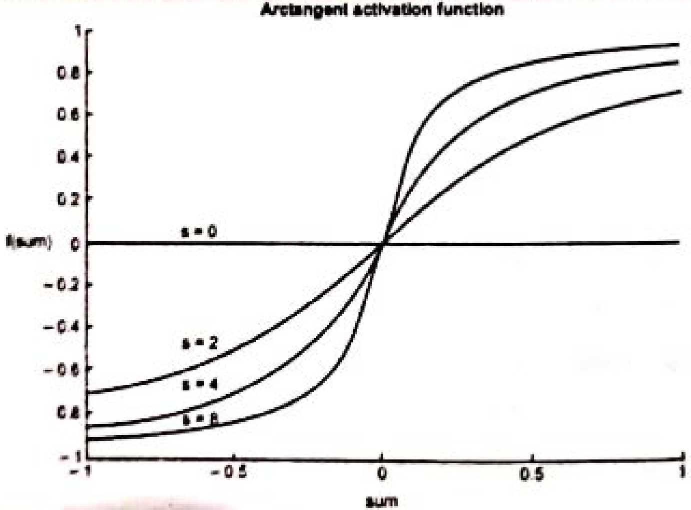


Fig. 9.4.2

- **Activation Function : Hyperbolic Tangent**

$$f(\text{sum}) = \tanh\ (s^{*}\ I)$$

$$= \frac{e^{s \times \text{sum}} - e^{-s \times \text{sum}}}{e^{s \times \text{sum}} + e^{-s \times \text{sum}}}$$

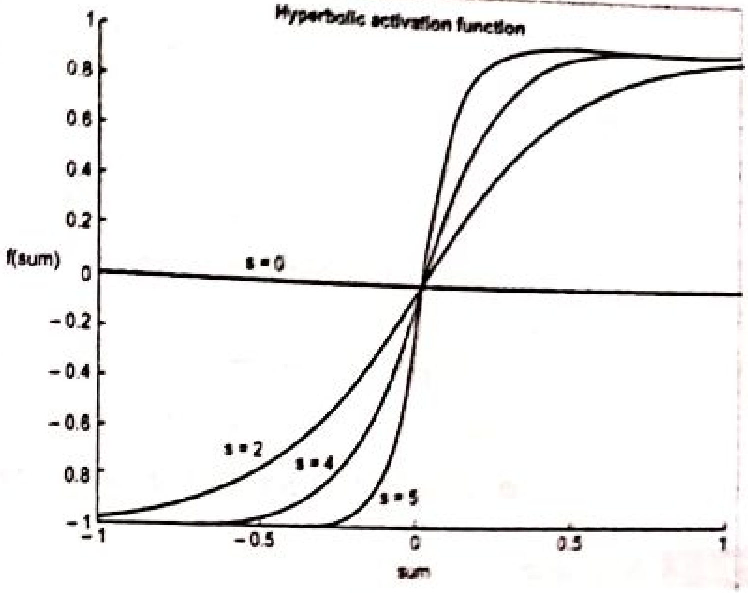


Fig. 9.4.3

9.4.1 Identity or Linear Activation Function

- A linear activation is a mathematical equation used for obtaining output vectors with specific properties.
- It is a simple straight line activation function where our function is directly proportional to the weighted sum of neurons or input.
- Linear activation functions are better in giving a wide range of activations and a line of a positive slope may increase the firing rate as the input rate increases.
- Fig. 9.4.4 shows identity function.

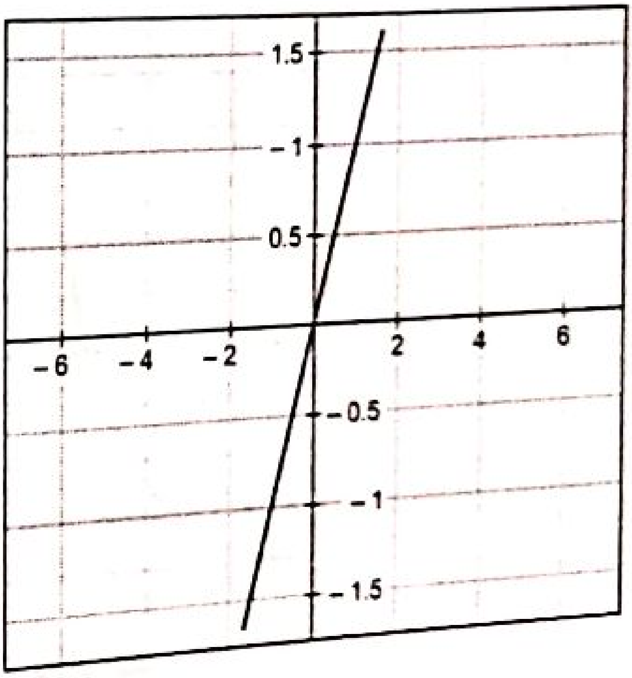


Fig. 9.4.4

- The equation for linear activation function is :

$$f(x) = a.x$$

When a = 1 then f(x) = x and this is a special case known as identity.

- **Properties :**
  1. Range is – infinity to + infinity
  2. Provides a convex error surface so optimisation can be achieved faster.
  3. df(x)/dx = a which is constant. So cannot be optimised with gradient descent.
- **Limitations :**
  1. Since the derivative is constant, the gradient has no relation with input.
  2. Back propagation is constant as the change is delta x.
  3. Activation function does not work in neural networks in practice.

9.4.2 Sigmoid

- A sigmoid function produces a curve with an "S" shape. The example sigmoid function shown on the left is a special case of the logistic function, which models the growth of some set.

$$\text{sig}(t) = \frac{1}{1+e^{-t}}$$

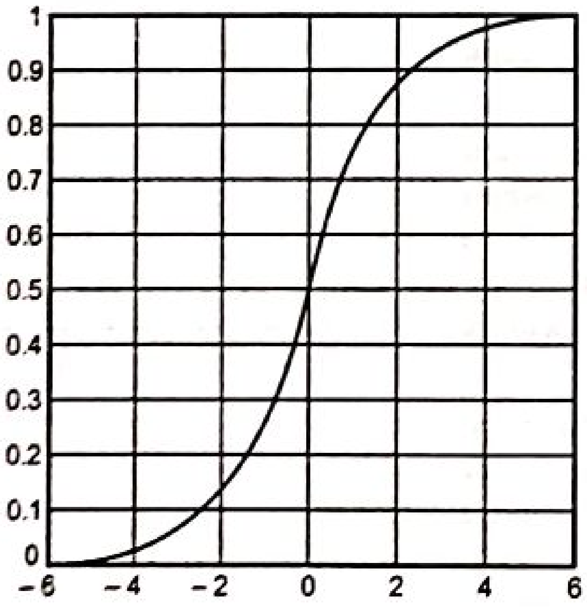


Fig. 9.4.5

- In general, a sigmoid function is real-valued and differentiable, having a non-negative or non-positive first derivative, one local minimum, and one local maximum.

- The logistic sigmoid function is related to the hyperbolic tangent as follows :

$$1 - 2\,\mathrm{sig}(x) = 1-2\frac{1}{1+e^{-x}} = -\tanh\frac{x}{2}$$

- Sigmoid functions are often used in artificial neural networks to introduce nonlinearity in the model.
- A neural network element computes a linear combination of its input signals, and applies a sigmoid function to the result.
- A reason for its popularity in neural networks is because the sigmoid function satisfies a property between the derivative and itself such that it is computationally easy to perform.

$$\frac{d}{dt}\mathrm{sig}(t) = \mathrm{sig}(t)(1-\mathrm{sig}(t))$$

- Derivatives of the sigmoid function are usually employed in learning algorithms.

9.4.3 ReLU Neuron

- Tanh is also like logistic sigmoid but better. The range of the tanh function is from (– 1 to 1). Tanh is also sigmoidal (s - shaped).
- Fig. 9.4.6 shows tanh v/s logistic sigmoid.

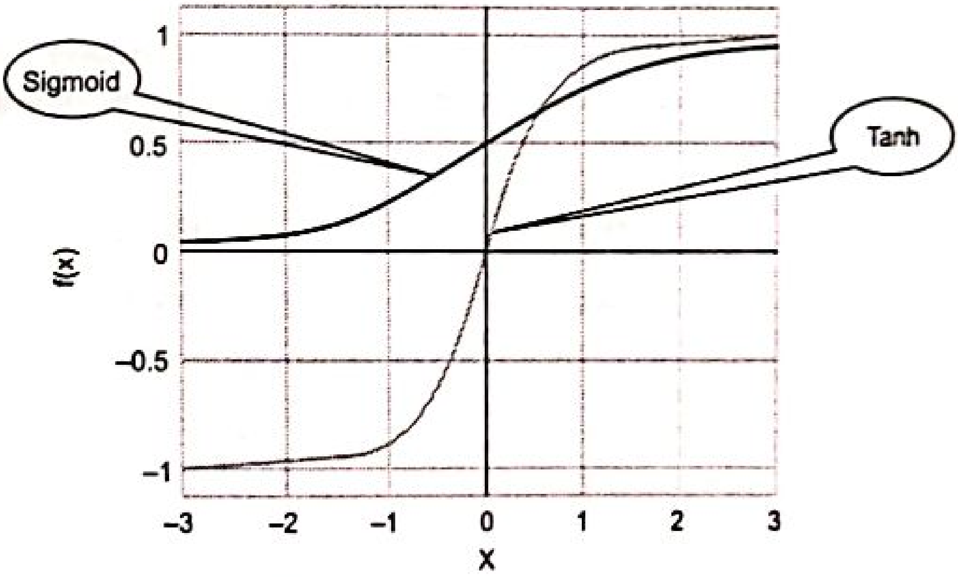


Fig. 9.4.6 Tanh v/s logistic sigmoid

- Tanh neuron is simply a scaled sigmoid neuron.
- Problems resolved by Tanh
  1. The output is not zero centered.
  2. Small gradient of sigmoid function.

- ReLU (Rectified Linear Unit) is the most used activation function in the world right now. Since, it is used in almost all the convolution neural networks or deep learning.
- Fig. 9.4.7 shows ReLU v/s Logistic Sigmoid.

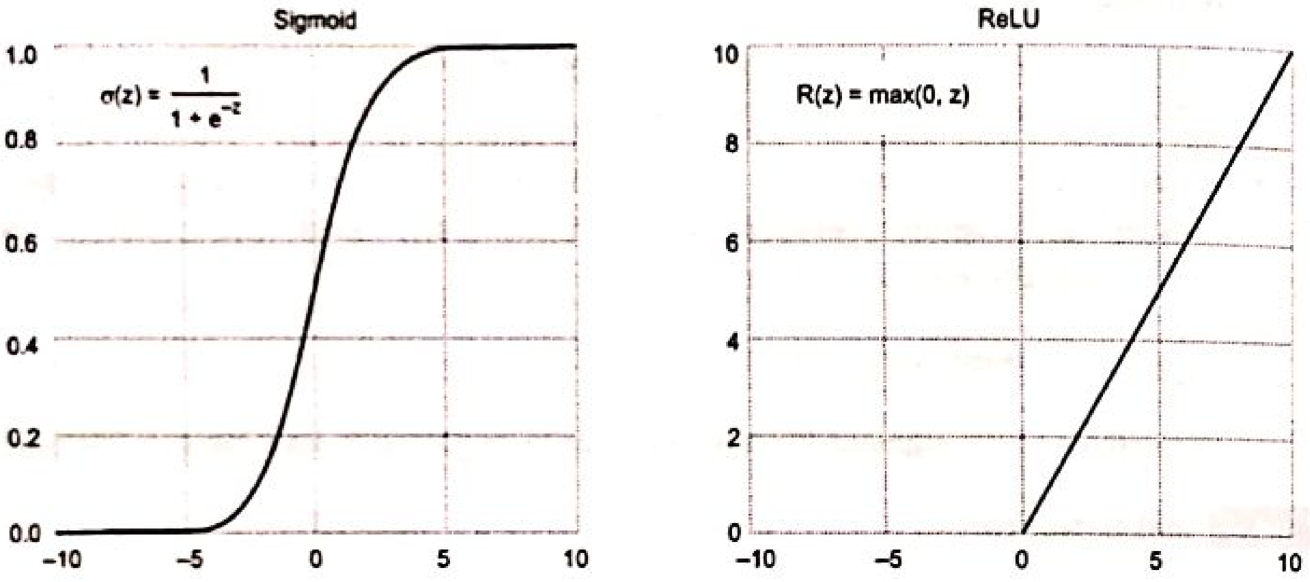


Fig. 9.4.7 ReLU v/s logistic sigmoid

- As you can see, the ReLU is half rectified (from bottom). f(z) is zero when z is less than zero and f(z) is equal to z when z is above or equal to zero.
- Compared to tanh/sigmoid neurons that involve expensive operations (exponentials, etc.), the ReLU can be implemented by simply thresholding a matrix of activations at zero.

Function	Advantages	Disadvantages
Sigmoid	1. Output in range (0,1)	1. Saturated neurons 2. Not zero centered 3. Small gradient 4. Vanishing gradient
Tanh	1. Zero centered 2. Output in range (− 1, 1)	1. Saturated Neurons
ReLU	1. Computational efficiency 2. Accelerated convergence	1. Dead Neurons 2. Not zero centered

9.5 Implementation of ANN

9.5.1 McCulloch Pitts Neuron

- The first mathematical model of a biological neuron was presented by McCulloch and Pitts. This model is known as McCulloch Pitt model. It is basic building block of neural network.
- Directed weight graph is used for connecting neurons.
- McCulloch and Pitts describe a neuron as a logical threshold element with two possible states. Such a threshold element has "N" input channels and one output channel. An input channel is either active (input 1) or silent (input 0).
- The activity states of all input channels thus encode the input information as a binary sequence of N bits. The state of the threshold element is then given by linear summation of all a different input signals x_i and comparison of the sum with a threshold value s.
- The system of neurons is static and acts synchronously. A processor (system) with multiple inputs and a single output.
- Effective input : Weighted sum of all inputs.
- Bias or threshold : If the effective input is larger than the bias, the neuron outputs a one, otherwise, it outputs a zero.
- Fig. 9.5.1 shows McCulloch Pitt model.

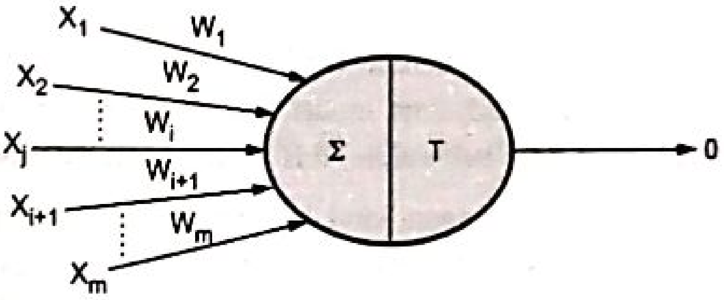


Fig. 9.5.1

- This model can be described in a mathematical formalism as follows :

$$0 = \theta(a)$$

Where

$$a = \sum W_j X_j - T$$

And

$\theta(x)$ is a function such that $\theta(x) = 1$ if $x > 0$, otherwise $\theta(x) = 1$.

- The parameters used to scale the inputs are called the weights. The effective input is the weighted sum of the inputs. The parameter to measure the switching level is the threshold or bias. Neuron fires (output of one) when its net input excitation exceeds a certain value called 'threshold.' Threshold is the minimum value of the sum of the weighted active inputs needed for the postsynaptic neuron to fire.
- The function for producing the final output is called the activation function, which is the step function in the McCulloch-Pitts model.

$$O = f\left(\sum_{j=1}^{N} W_j X_j - T\right)$$

$$f(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- Their "neurons" operated under the following assumptions :
  1. They are binary devices (0,1) .
  2. Each neuron has a fixed threshold (theta).
  3. The neuron receives inputs from excitatory synapses, all having identical weights.
  4. Inhibitory inputs have an absolute veto power over any excitatory inputs.
  5. At each time step the neurons are simultaneously (synchronously) updated by summing the weighted excitatory inputs and setting the output to 1 iff the sum is greater than or equal to the threshold AND if the neuron receives no inhibitory input.
- In general, there are many different kinds of activation functions. The step function used in the McCulloch-Pitts model is simply one of them. Because the activation function takes only two values, this model is called discrete neuron.
- To make the neuron learnable, some kind of continuous function is often used as the activation function. This kind of neurons is called continuous neurons. Typical functions used in an artificial neuron are sigmoid functions, radial basis function, sinusoidal functions, etc.

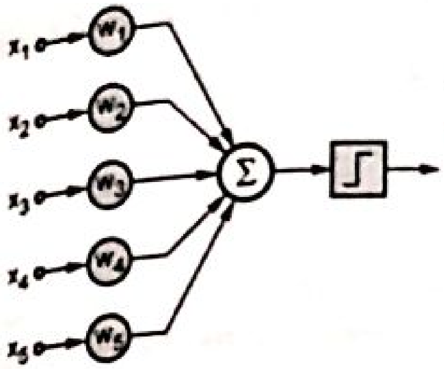
Problems with McCulloch-Pitts neurons

1. Weights and thresholds are analytically determined. We cannot learn.
2. It is very difficult to minimize size of a network.

9.5.2 Rosenblatt's Perceptron

- Rosenblatt's perceptron is built around a nonlinear neuron, namely, the McCulloch-Pitts model of a neuron.

- Rosenblatt perceptron is a binary single neuron model. The inputs integration is implemented through the addition of the weighted inputs that have fixed weights obtained during the training stage. If the result of this addition is larger than a given threshold θ the neuron fires. When the neuron fires its output is set to 1, otherwise it's set to 0.
- The equation can be re-written as follows including what it's known as the bias term :



$$h(x) = \begin{cases} 1 \text{ if } & w_1 \cdot x_1 + w_2 \cdot x_2 + ... + w_d \cdot x_d \geq \theta \\ 0 \text{ if } & w_1 \cdot x_1 + w_2 \cdot x_2 + ... + w_d \cdot x_d < \theta \end{cases}$$

The equation can be re-written as follows including what its known as the bias term : $x_0 = 1,\ w_0 = \theta$.

$$h(x) = \begin{cases} 1 \text{ if } & w_0 \cdot x_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + ... + w_d \cdot x_d \geq 0 \\ 0 \text{ if } & w_0 \cdot x_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + ... + w_d \cdot x_d < 0 \end{cases}$$

$$h(x) = \begin{cases} 1 \text{ if } & w^t \cdot x \geq 0 \\ 0 \text{ if } & w^t \cdot x < 0 \end{cases}$$

9.5.3 ADALINE Network Model

- ADALINE (Adaptive Linear Neuron) is an early single-layer artificial neural network.
- An important generalized of the perceptrons training algorithm was presented by Widrow and Hoff as the least mean square learning procedure also known as the delta rule.
- The learning rule was applied to the "**adaptive linear element**" also named Adaline.
- The perceptron learning rule uses the output of the thersold function for learning. The delta rule uses the net output without further mapping into output values –1 or + 1.
- Fig. 9.5.2 shows adaline.
- If the input conductances are denoted by w_i where m i = 0, 1, 2, ..., n and input and output signals by x_i and y respectively, then the output of the central block is defined to be :

$$y = \sum_{i=1}^{n} w_i x_i + \theta$$

Where, $\theta = w_0$

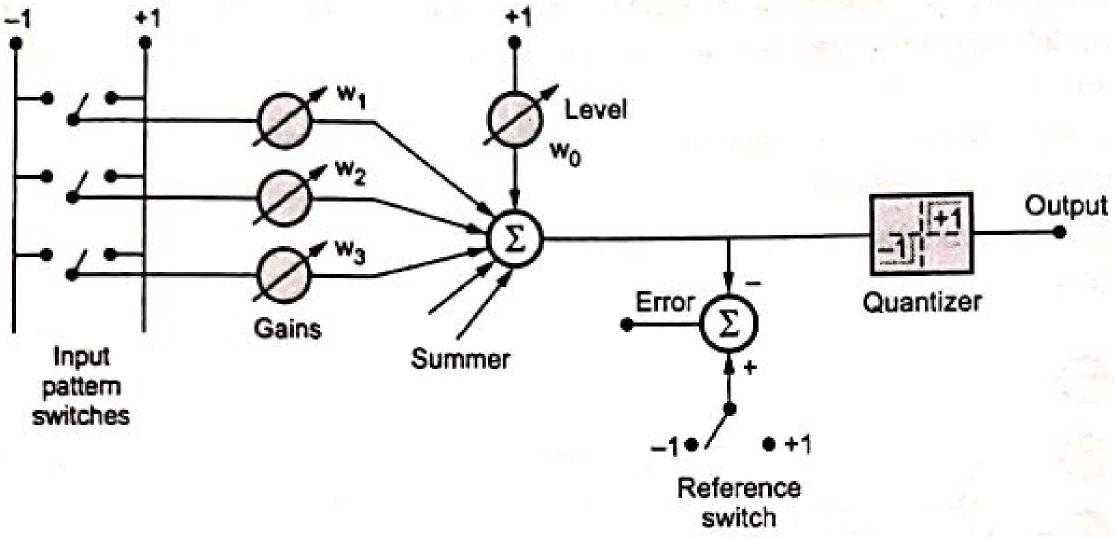


Fig. 9.5.2 Adaline

- In a simple physical implementation, this device consists of a set of controllable resistors connected to a circuit which can sum up currents caused by the input voltage signals. Usually the central block, the summer is also followed by a quantizer which outputs + 1 of – 1, depending on the polarity of the sum.
- The problem is to determine the coefficients w_i, where i = 0, 1 ..., n , in such way that the input output response is correct for a large number of arbitrarily chosen signal sets.
- If an exact mapping is not possible the average error must be minimized, for instance, in the sense of least squares.
- An adaptive operation means that there exists a mechanism by which the w_i can be adjusted, usually iteratively to attain the correct values.
- For the Adaline, Widrow introduced the delta rule to adjust the weights.
- For the pth input-output pattern, the earor measure of a single-output Adaline can be expressed as,

$$E_p = (t_p - o_p)^2$$

Where

$$t_p = \text{Target output}$$

$$o_p = \text{Actual output of the Adaline}$$

- The derivation of E_p with respect to each weight w_i is

$$\frac{\partial E_p}{\partial w_i} = -2(t_p - o_p)\, x_i$$

- To decrease E_p by gradient descent, the update formula for w_i on the p^{th} input-output pattern is

$$\Delta_p w_i = \eta(t_p - o_p)x_i$$

- The delta rule tries to minimize squared errors, it is also referred to as the least mean square learning procedure or Widrow - Hoff learning rule.

9.6 Architecture of Neural Network

9.6.1 Single Layer Feed Forward Network

- The architecture of the neural network refers to the arrangement of the connection between neurons, processing element, number of layers, and the flow of signal in the neural network.
- There are mainly two category of neural network architecture :

  a. Feed-forward

  b. Feedback (recurrent) neural networks.

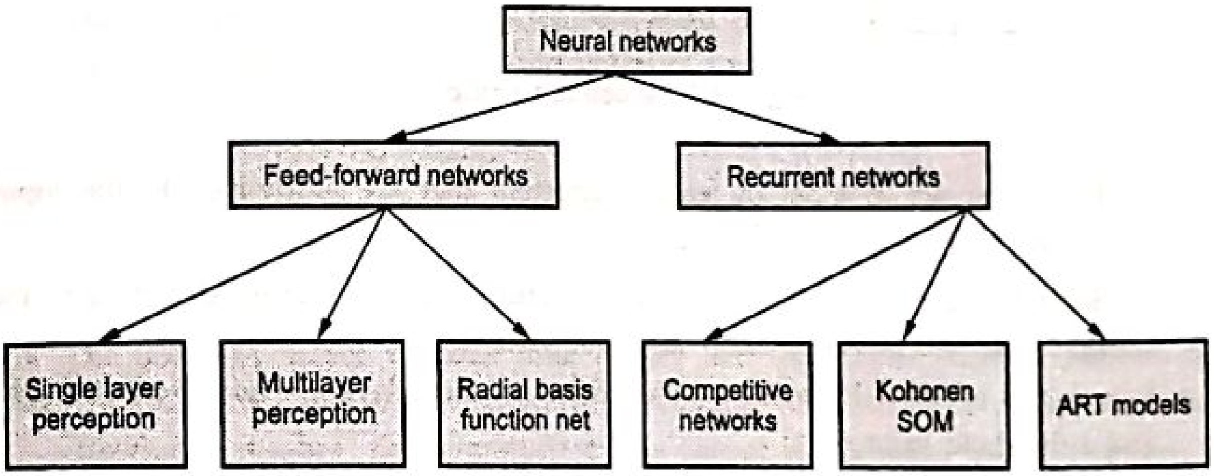


Fig. 9.6.1

1. Architecture and Learning Rule

- In late 1950s, Frank Rosenblatt introduced a network composed of the units that were enhanced version of McCulloch-Pitts Threshold Logic Unit (TLU) model.
- Rosenblatt's model of neuron, a perceptron, was the result of merger between two concepts from the 1940s, McCulloch-Pitts model of an artificial neuron and Hebbian learning rule of adjusting weights.
- In addition to the variable weight values, the perceptron model added an extra input that represents bias. Thus, the modified equation is now as follows :

$$\text{Sum} = \sum_{i=1}^{N} I_i \; W_i + b,$$

where **b** represents the bias value.

- Fig. 9.6.2 shows a typical perception setup for pattern recognition applications, in which visual patterns are represented as matrices of elements between 0 and 1.

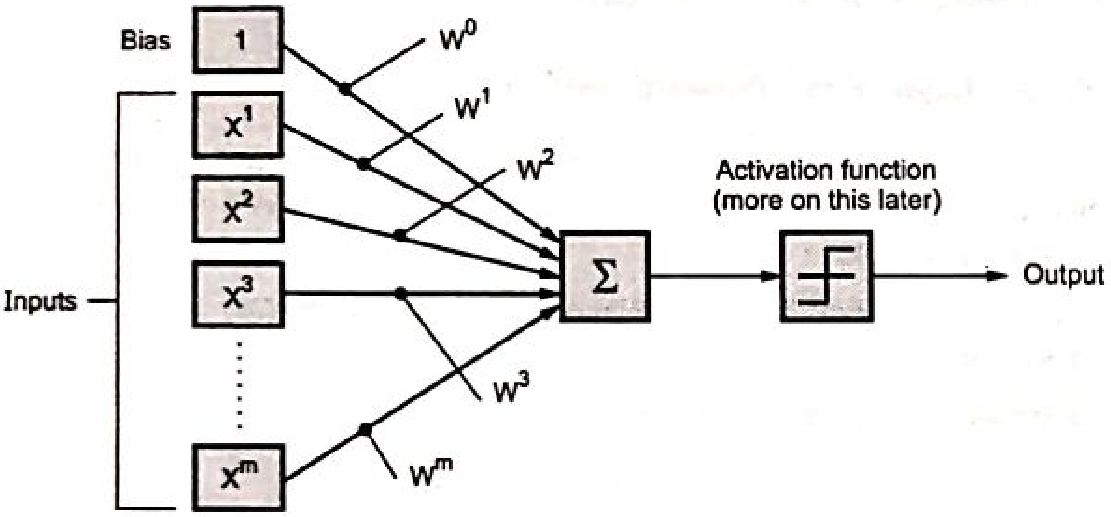


Fig. 9.6.2 Perception setup

1. First layer act as a set of feature detectors that are hardwired to the input signals to detect specific features.
2. Second layer i.e. output layer takes the outputs of the feature detectors in the first layer and classifies the given input pattern.

- Learning is initiated by making adjustments to the relevant connection strengths and a threshold value θ.
- Here we consider only two class problem. Here output layer usually has only a single node. For an n-class problem (n > 3), the output layer usually has n-nodes, each corresponding to a class and the output node with the largest value indicates which class the input vector belongs to.
- In the first stage, the linear combination of inputs is calculated. Each value of input array is associated with its weight value, which is normally between 0 and 1. Also, the summation function often takes an extra input value Theta with weight value of 1 to represent threshold or bias of a neuron.
  a. The term x_i is referred to as **active or excitatory** if its value is 1.
  b. If the value is 0 then it is **inactive.**
  c. If the value is –1 then it is **inhibitory.**
- The output unit is a linear threshold element with a threshold value θ :

$$0 = f\left(\sum_{i=1}^{n} w_i\ x_i - \theta\right)$$

$$= f\left(\sum_{i=1}^{n} w_i\ x_i + w_0\right), w_0 \equiv -\theta$$

$$= f\left(\sum_{i=1}^{n} w_i\ x_i\right), x_0 \equiv 1$$

where w_i is a modifiable weight associated with an incoming signal x_i.

- Fig. 9.6.3 shows the bias term w_0.

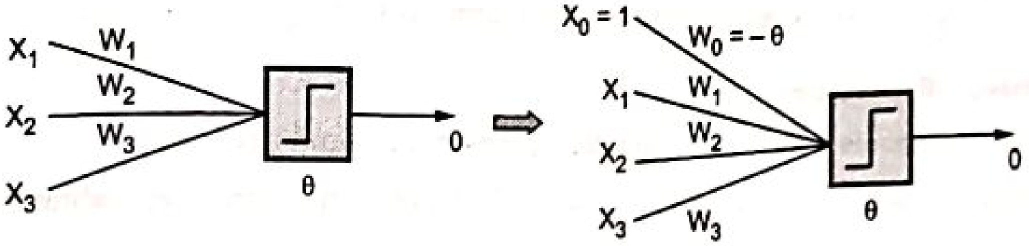


Fig. 9.6.3 Bias term w_0

- The function y = f(x) describes relationship, an input-output mapping from x to y.
- The equation (9.6.1), the f(.) is the **activation function** of the perceptron and it is typically either a **signum function** sgn(x) or **step function** step(x) :

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{otherwise} \end{cases}$$

$$\text{step}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases} \qquad \text{... (9.6.1)}$$

- The sum-of-product value is then passed into the second stage to perform the activation function which generates the output from the neuron. The activation function "squashes" the amplitude of the output in the range of [0, 1] or [-1, 1] alternately. The behavior of the activation function will describe the characteristics of an artificial neuron model.
- The basic learning algorithm for a single layer perceptron repeats the following steps until the weights converge :
  1. Select an input vector x from the training data set.
  2. If the perceptron gives an incorrect response, modify all connection weights w_i according to

$$\Delta\ w_i = \eta\ t_i\ X_i$$

Where t_i is a target output and η is a learning state.

Perceptron Convergence Theorem

Theorem : If there is a set of weights that correctly classify the (linearly separable) training patterns, then the learning algorithm will find one such weight set, w* in a finite number of iterations.

Assumptions :

1. At least one such set of weights, w*, exists, and
2. There are a finite number of training patterns.
3. The threshold function is uni-polar (output is 0 or 1).

2. Exclusive OR problem

- XOR problem is a pattern recognition problem in neural network.
- Neural networks can be used to classify boolean functions depending on their desired outputs.
- For a two input binary XOR problem , the desired output is given in the form of truth table.

	X	Y	Class
Desired I/O pair 1	0	0	0
Desired I/O pair 2	0	1	1
Desired I/O pair 3	1	0	1
Desired I/O pair 4	1	1	0

- The XOR problem is not **linearly separable**. We cannot use a single layer perceptron to construct a straight line to partition the two dimensional input space into two regions, each containing only data points of the same class.
- Let us consider following four equations :

$$0 \times w_1 + 0 \times w_2 + w_0 \le 0 \Leftrightarrow w_0 \le 0,$$

$$0 \times w_1 + 1 \times w_2 + w_0 \; 0 \Leftrightarrow w_0 > -w_2$$

$$1 \times w_1 + 0 \times w_2 + w_0 > 0 \Leftrightarrow w_0 > -w_1$$

$$1 \times w_1 + 1 \times w_2 + w_0 \le 0 \Leftrightarrow w_0 \le -w_1 - w_2$$

9.6.2 Multi-Layer Feed Forward Network

- A multilayer feed-forward neural network is a network consisting of multiple layers of units, all of which are adaptive. The network is not allowed to have cycles from later layers back to earlier layers, hence the name "feed-forward". Let

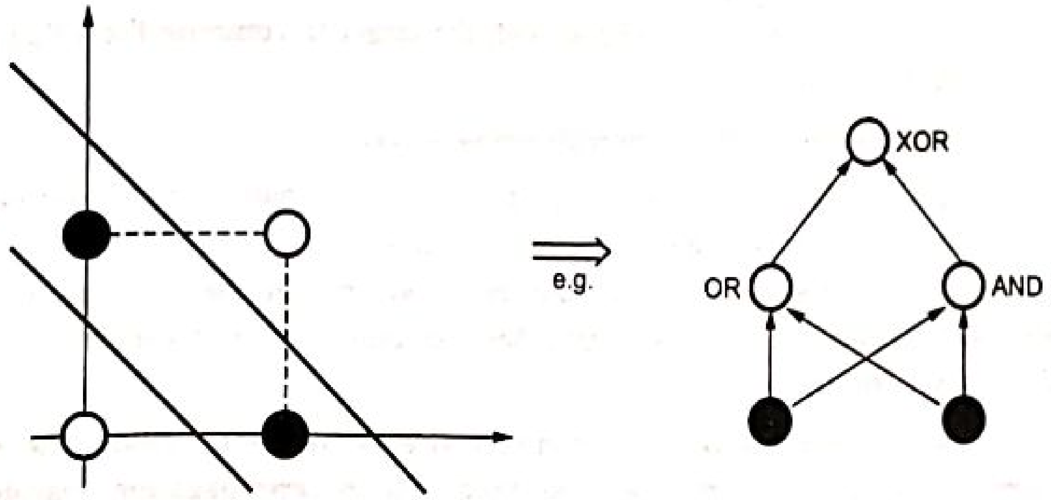
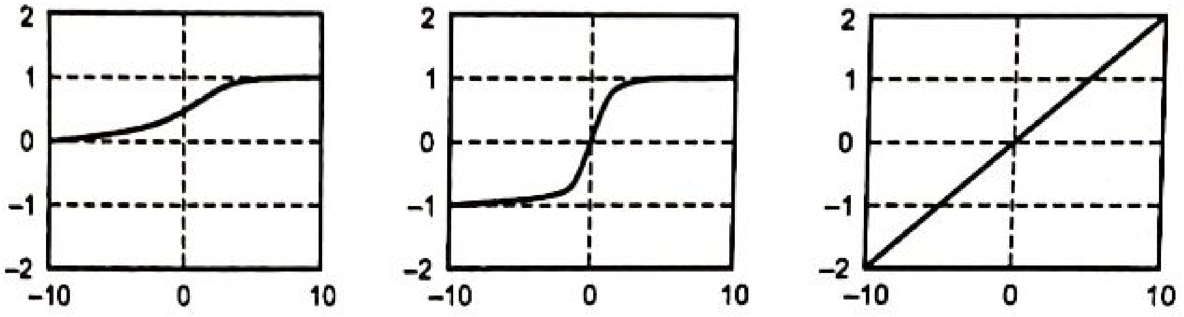


Fig. 9.6.4

us consider a network with a single complete hidden layer. i.e., the network consists of some input nodes, some output nodes, and a set of hidden nodes. Every hidden node takes inputs from each of the input nodes, and feeds into each of the output nodes.

- In multi-layer feed forward neural networks, the sigmoid activation function, defined by $g(x) = \frac{1}{1+e^{-x}}$ is normally used.
- A Multi-Layer Perceptron (MLP) has the same structure of a single layer perceptron with one or more hidden layers. An MLP is a network of simple neurons called perceptrons.
- A typical multilayer perceptron network consists of a set of source nodes forming the input layer, one or more hidden layers of computation nodes, and an output layer of nodes.
- It is not possible to find weights which enable single layer perceptrons to deal with non-linearly separable problems like XOR :
- Multi-layer perceptrons are able to cope with non-linearly separable problems.
- Each neuron in one layer has direct connections to all the neurons of the subsequent layer. MLP can implement nonlinear discriminants (for classification) and nonlinear regression functions (for regression).
- Historically, the problem was that there were no known learning algorithms for training MLPs. Fortunately; it is now known to be quite straightforward. The procedure for finding a gradient vector in the network structure is generally referred to as **backpropagation**. Because the gradient vector is calculated in the direction opposite to the flow of the output of each node.

- Procedure of backpropagation :
  1. The output values are compared with the target to compute the value of some predefined error function.
  2. The error is then fedback through the network.
  3. Using this information, the algorithm adjusts the weights of each connection in order to reduce the value of the error function.
- Continue this process until the connection weights in the network have been adjusted so that the network output has converged, to an acceptable level, with the desired output.
- If we use the gradient vector in a simple steepest descent method, the resulting learning paradigm is often referred to as the **backpropagation** learning rule. Backpropagation works by approximating the non-linear relationship between the input and the output by adjusting the weight values internally.
- Generally, the backpropagation network has two stages, training and testing. During the training phase, the network is "shown" sample inputs and the correct classifications. For example, the input might be an encoded picture of a face, and the output could be represented by a code that corresponds to the name of the person.
- Fig. 9.6.5 shows three most commonly used activation functions in backpropagation MLPs.



Logistic function

Hyperbolic tangent function

Identify function

Fig. 9.6.5 Activation function

Logistic function :

$$f(x) = \frac{1}{1+e^{-x}}$$

Hyperbolic tangent function :

$$f(x) = \tanh(x/2) = \frac{1-e^{-x}}{1+e^{-x}}$$

Identity function :

$$f(x) = x$$

- Both the hyperbolic tangent function and logistic function approximate the signum and step function respectively. Sometimes these two function are referred to as **squashing functions** since the inputs to these functions are squashed to the range [0, 1] or [– 1, 1].
- These functions are also called **sigmoidal functions** because their S-shaped curves exhibits smoothness and asymptotic properties.
- A learning process is organized through a learning algorithm, which is a process of updating the weights in such a way that a machine learning tool implements a given input/output mapping with no errors or with some minimal acceptable error.
- Any learning algorithm is based on a certain learning rule, which determines how the weights shall be updated if the error occurs.

Backpropagation Learning Rule

- The **net input** of a node is defined as the weighted sum of the incoming signals plus a bias term. Fig. 9.6.6 shows the backpropagation MLP for node j. The net input and output of node j is as follows :

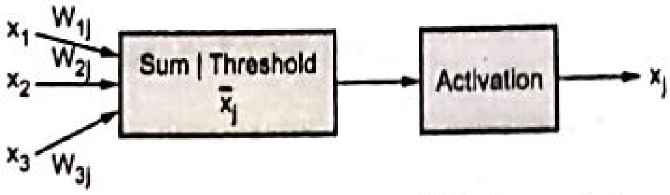


Fig. 9.6.6 Backpropagation MLP for node j

$$\bar{X}_j = \sum_i + W_{ij} + W_j$$

$$x_j = f(\bar{X}_j) = \frac{1}{1+\exp(-\bar{X}_j)}$$

Where

x_i is the output of node i located in any one of the pervious layers,

W_{ij} is the weight associated with the link connecting nodes i and j,

W_j is the bias of node j.

- Internal parameters associated with each node j is the weight W_{ij}. So changing the weights of the node will change the behaviour of the whole backpropagation MLP.
- Fig. 9.6.7 shows two layer backpropagation MLP.
- The above backpropagation MLP will refer to as a 3-4-3 network, corresponding to the number of nodes in each layer.

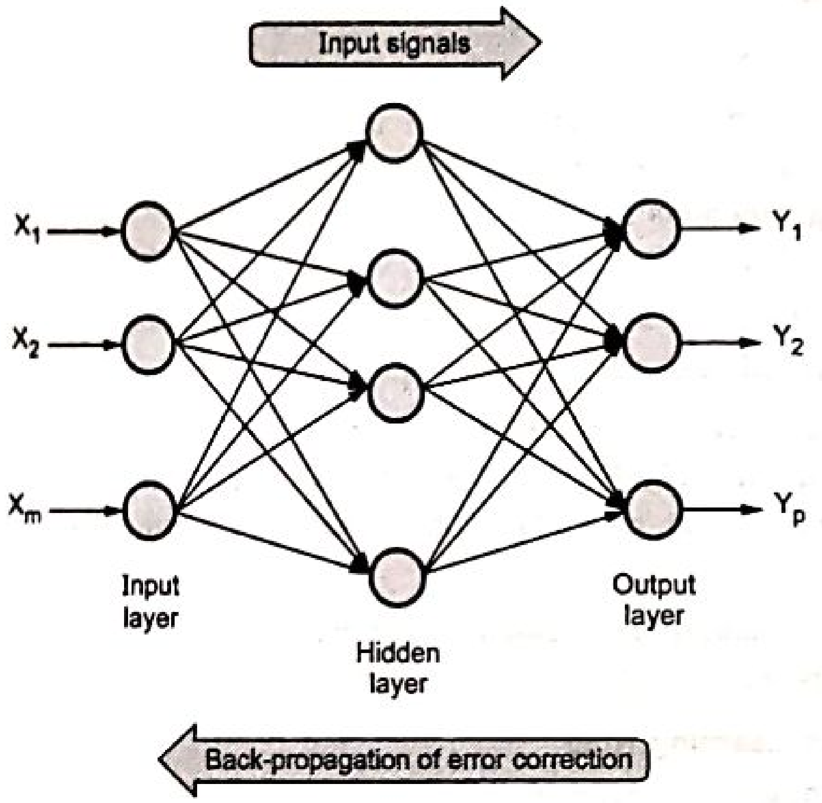


Fig. 9.6.7 Two layer backpropagation MLP

- The backward error propagation also known as the backpropagation (BP) or the Generalized Delta Rule (GDR). A squared error measure for the p^{th} input-output pair is defined as

$$E_p = \sum_k (d_k - x_k)^2$$

Where d_k is the desired output for node k and x_k is the actual output for node k when the input part of the p^{th} data pair is presented.

- To find the gradient vector, an error term $\bar{\epsilon}_i$ for node i is defined as :

$$\bar{\epsilon}_i = \frac{\partial + E_p}{\partial \overline{X}_i}$$

- The partial derivative can be rewritten as product of two terms using chain rule for partial differentiation :

$$\frac{\partial E(t)}{\partial w_{ij}(t)} = \frac{\partial E(t)}{\partial a_i(t)} \bullet \frac{\partial a_i(t)}{\partial w_{ij}(t)}$$

- Features of the delta rule are as follows :
  1. Simplicity

2. Distributed learning : Learning is not reliant on central control of the network.
3. Online learning : Weights are updated after presentation of each pattern.

Rules for Feedforward Multilayer Perceptron

- The training algorithm is called Error Back Propagation (EBP) training algorithm. If a submitted pattern provides an output far from desired value, the weights and thresholds are adjusted so that the current mean square classification error is reduced.
- The training is repeated for all patterns until the training set provide an acceptable overall error. Usually the mapping error is computed over the full training set.
- Error back propagation algorithm is working in two stages :
    1. The trained network operates feed-forward to obtain output of the network
    2. The weight adjustment propagate backward from output layer through hidden layer toward input layer.

9.6.3 Recurrent Neural Network

- A recurrent neural network is a type of neural network that contains loops, allowing information to be stored within the network.
- A RNN is particularly useful when a sequence of data is being processed to make a classification decision or regression estimate but it can also be used on non-sequential data. Recurrent neural networks are typically used to solve tasks related to time series data.
- Applications of recurrent neural networks include natural language processing, speech recognition, machine translation, character-level language modeling, image classification, image captioning, stock prediction, and financial engineering.
- Fig. 9.6.8 shows architecture of recurrent neural network.

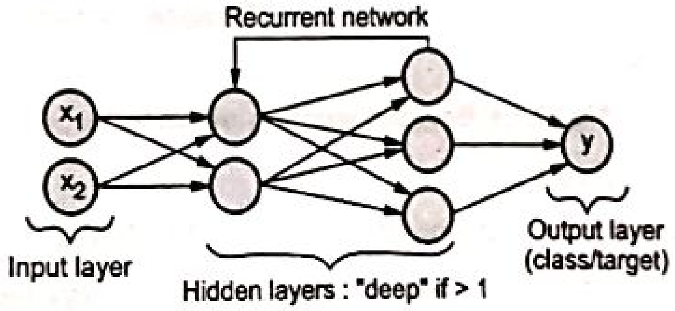


Fig. 9.6.8

- Recurrent Neural Networks can be thought of as a series of networks linked together. They often have a chain-like architecture, making them applicable for tasks such as speech recognition, language translation, etc.

- An RNN can be designed to operate across sequences of vectors in the input, output, or both. For example, a sequenced input may take a sentence as an input and output a positive or negative sentiment value. Alternatively, a sequenced output may take an image as an input, and produce a sentence as an output.

9.7 Backpropagation

- Backpropagation is a training method used for a multi-layer neural network. It is also called the generalized delta rule. It is a gradient descent method which minimizes the total squared error of the output computed by the net.
- The backpropagation algorithm looks for the minimum value of the error function in weight space using a technique called the delta rule or gradient descent. The weights that minimize the error function is then considered to be a solution to the learning problem.
- Backpropagation is a systematic method for training multiple layer ANN. It is a generalization of Widrow-Hoff error correction rule. 80 % of ANN applications uses backpropagation.
- Fig. 9.7.1 shows backpropagation network.

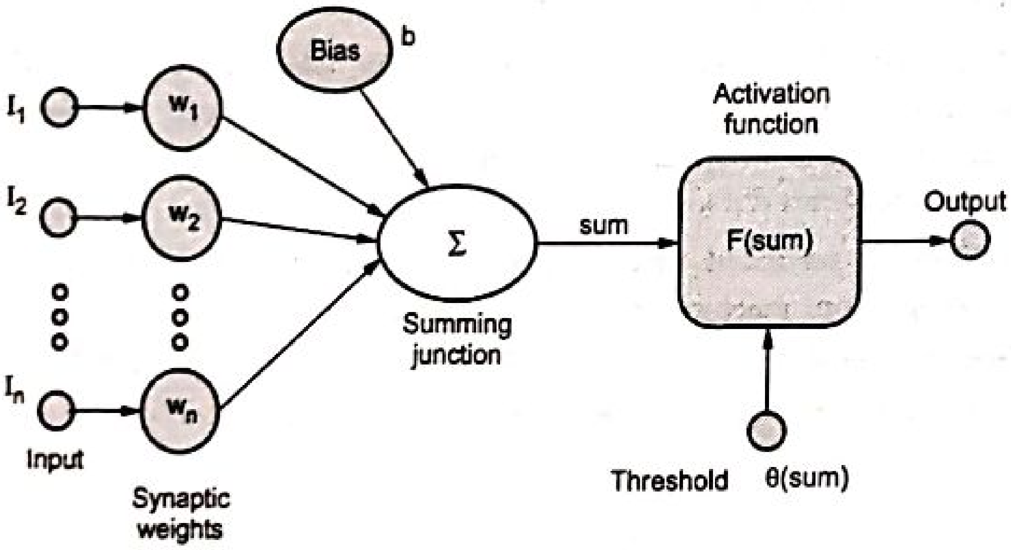


Fig. 9.7.1 Backpropagation network

- Consider a simple neuron :
  a. Neuron has a summing junction and activation function.
  b. Any non linear function which differentiable everywhere and increases everywhere with sum can be used as activation function.
  c. Examples : Logistic function, Arc tangent function, Hyperbolic tangent activation function.

- These activation function makes the multilayer network to have greater representational power than single layer network only when non-linearity is introduced.
- **Need of hidden layers :**
  1. A network with only two layers (input and output) can only represent the input with whatever representation already exists in the input data.
  2. If the data is discontinuous or non-linearly separable, the innate representation is inconsistent, and the mapping cannot be learned using two layers (Input and Output).
  3. Therefore, hidden layer(s) are used between input and output layers
- **Weights** connects unit (neuron) in one layer only to those in the next higher layer. The output of the unit is scaled by the value of the connecting weight, and it is fed forward to provide a portion of the activation for the units in the next higher layer.
- Backpropagation can be applied to an artificial neural network with any number of hidden layers. The training objective is to adjust the weights so that the application of a set of inputs produces the desired outputs.
- **Training procedure :** The network is usually trained with a large number of input-output pairs.
  1. Generate weights randomly to small random values (both positive and negative) to ensure that the network is not saturated by large values of weights.
  2. Choose a training pair from the training set.
  3. Apply the input vector to network input.
  4. Calculate the network output.
  5. Calculate the error, the difference between the network output and the desired output.
  6. Adjust the weights of the network in a way that minimizes this error.
  7. Repeat steps 2 - 6 for each pair of input-output in the training set until the error for the entire system is acceptably low.

Forward pass and backward pass :

- Backpropagation neural network training involves two passes.
  1. In the forward pass, the input signals moves forward from the network input to the output.

2. In the backward pass, the calculated error signals propagate backward through the network, where they are used to adjust the weights.

3. In the forward pass, the calculation of the output is carried out, layer by layer, in the forward direction. The output of one layer is the input to the next layer.

- In the reverse pass,

  a. The weights of the output neuron layer are adjusted first since the target value of each output neuron is available to guide the adjustment of the associated weights, using the delta rule.

  b. Next, we adjust the weights of the middle layers. As the middle layer neurons have no target values, it makes the problem complex.

- **Selection of number of hidden units :** The number of hidden units depends on the number of input units.

  1. Never choose h to be more than twice the number of input units.

  2. You can load p patterns of I elements into $\log_2$ p hidden units.

  3. Ensure that we must have at least 1/e times as many training examples.

  4. Feature extraction requires fewer hidden units than inputs.

  5. Learning many examples of disjointed inputs requires more hidden units than inputs.

  6. The number of hidden units required for a classification task increases with the number of classes in the task. Large networks require longer training times.

Factors influencing Backpropagation training

- The training time can be reduced by using :

  1. **Bias :** Networks with biases can represent relationships between inputs and outputs more easily than networks without biases. Adding a bias to each neuron is usually desirable to offset the origin of the activation function. The weight of the bias is trainable similar to weight except that the input is always +1.

  2. **Momentum :** The use of momentum enhances the stability of the training process. Momentum is used to keep the training process going in the same general direction analogous to the way that momentum of a moving object behaves. In backpropagation with momentum, the weight change is a combination of the current gradient and the previous gradient.

9.7.1 Advantages and Disadvantages

Advantages of backpropagation :

1. It is simple, fast and easy to program.

2. Only numbers of the input are tuned and not any other parameter.
3. No need to have prior knowledge about the network.
4. It is flexible.
5. A standard approach and works efficiently.
6. It does not require the user to learn special functions.

Disadvantages of backpropagation :

1. Backpropagation possibly be sensitive to noisy data and irregularity.
2. The performance of this is highly reliant on the input data.
3. Needs excessive time for training.
4. The need for a matrix-based method for backpropagation instead of mini-batch.

9.8 Deep Learning

- Deep Learning is a new area of machine learning research, which has been introduced with the objective of moving machine learning closer to one of its original goals. Deep Learning is about learning multiple levels of representation and abstraction that help to make sense of data such as images, sound, and text.
- 'Deep Learning' means using a neural network with several layers of nodes between input and output. It is generally better than other methods on image, speech and certain other types of data because the series of layers between input and output do feature identification and processing in a series of stages, just as our brains seem to.
- Deep Learning emphasizes the network architecture of today's most successful machine learning approaches. These methods are based on "deep" multi-layer neural networks with many hidden layers.

9.9 Fill in the Blanks

Q.1 Backpropagation is a training method used for a ________ neural network.

Q.2 ADALINE is an early ________ artificial neural network.

Q.3 Activation functions also known as ________ function is used to map input nodes to output nodes in certain fashion.

Q.4 ADALINE stands for ________ .

Q.5 Rosenblatt perceptron is a ________ single neuron model.

9.10 Multiple Choice Questions

Q.1 Backpropagation is a supervised learning algorithm, for training ______ perceptrons.

a single layer b multilayer

c any form of d none

Q.2 What is backpropagation ?

a It is another name given to the curvy function in the perceptron.

b It is the transmission of error back through the network to adjust the inputs.

c It is the transmission of error back through the network to allow weights to be adjusted so that the network can learn.

d None of the above.

Q.3 What are the general tasks that are performed with backpropagation algorithm ?

a pattern mapping b function approximation

c prediction d all of the above

Q.4 Adaline which stands for ________.

a Adaptive Linear Neuron b Address Linear Neuron

c Adaptive Linear Network d Adaptive Neural Neuron

Q.5 What is an activation value ?

a weighted sum of inputs b threshold value

c main input to neuron d none of the mentioned

Q.6 Why can't we design a perfect neural network ?

a full operation is still not known of biological neurons

b number of neuron is itself not precisely known

c number of interconnection is very large & is very complex

d all of these

Q.7 What was the main point of difference between the Adaline & perceptron model ?

a Weights are compared with output

b Sensory units result is compared with output

c Analog activation value is compared with output

d All of the mentioned

Q.8 The backpropagation algorithm is used to find a local minimum of the ______ .

a neural network b activation function

c error function d none of these

Q.9 Application of Neural Network includes

a Pattern Recognition b Classification

c Clustering d All of these

Q.10 Neural Networks are complex __________ with many parameters.

a Linear Functions b Nonlinear Functions

c Discrete Functions d Exponential Functions

Q.11 A possible neuron specification to solve the AND problem requires a minimum of ________.

a Single neuron b Two neurons

c Three neurons d Four neurons

Q.12 ADALINE is an early ________ layer artificial neural network.

a zero b single

c multi d All of these

Q.13 Neuron can send ________ signal at a time.

a multiple b one

c None d Any number of

Q.14 Internal state of neuron is called ________ , is the function of the inputs the neurons receives.

a weight b bias

c activation or activity level of neuron d All of these

Q.15 In artificial neural network interconnected processing elements are called ______.

a weights b nodes or neurons

c soma d axons

Q.16 _______ are fibres acting as transmission lines that send activation to other neurons.

a Soma b Nucleus

c Dendrites d Axons

Q.17 What are dendrites ?

a fibers of nerves b nuclear projections

c another name for nucleus d Soma

Q.18 In which ANN, loops are allowed ?

a FeedForward ANN b FeedBack ANN

c Both A and B d None of these

Q.19 The output at each node is called _______.

a axons b neurons

c weight d node value

Q.20 Training perceptron is based on _______.

a supervised learning technique b unsupervised learning

c reinforced learning d stochastic learning

Q.21 A perceptron takes a vector of real-valued inputs, calculates a linear combination of these inputs, then outputs.

a 1 or -1 b 0 or 1

c -1 or 0 d None

Q.22 Neural networks are complex _______ with many parameters.

a linear functions b nonlinear functions

c discrete functions d exponential functions

Q.23 What is perceptron in neural network ?

a It is an auto-associative neural network.

b It is a double layer auto-associative neural network.

c It is a single layer feed-forward neural network with pre-processing.

d It is a neural network that contains feedback.

Answer Keys for Fill in the Blanks

Q.1	multi-layer	Q.2	single-layer
Q.3	transfer	Q.4	Adaptive Linear Neuron
Q.5	binary		

Answer Keys for Multiple Choice Questions

Q.1	b	Q.2	c
Q.3	d	Q.4	a
Q.5	a	Q.6	d
Q.7	c	Q.8	c
Q.9	d	Q.10	a
Q.11	a	Q.12	b
Q.13	b	Q.14	c
Q.15	b	Q.16	d
Q.17	a	Q.18	b
Q.19	d	Q.20	a
Q.21	a	Q.22	a
Q.23	c		

Distributors

- **Patel Book Agency**
  1st Floor, Mahavir Smruti Complex,
  Gandhi Road, Ahmedabad-1,
  Ph: (079)25324741.

- **Atul Agencies**
  Fernandis Bridge,
  Gandhi Road, Ahmedabad - 1,
  Ph: (079)22160475.

- **Mahajan Book Depot**
  Gandhi Road, Ahmedabad - 1.
  Ph: (079)25356031.

Or Contact

- **Ajay Pachegaonkar (Technical Publications)**
  Mkt. Executive (Gujarat)
  Mob No.:09687118452,
  email: ajay@technicalpublications.org

ONLINE SELLER

Launching soon...
bookstation

Scan this QR Code & visit us at
www.technicalpublications.org
amazon Flipkart

ISBN 978-93-90770-16-8

9 789390 770168

Published by :

TECHNICAL PUBLICATIONS
SINCE 1993
An Up-Thrust for Knowledge

Find us

Website : www.technicalpublications.org
https://www.facebook.com/technicalpublications

Amit Residency, Office No.1, 412, Shaniwar Peth, Pune - 411030, M.S. INDIA.
Ph.: +91-020-24495496/97, | Email : sales@technicalpublications.org