TABLE OF CONTENTS

Chapter - 7 Deep Learning - More to Know (7 - 1) to (7 - 10)

1 Introduction to Machine Learning

Syllabus

Overview of Human Learning and Machine Learning, Types of 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.

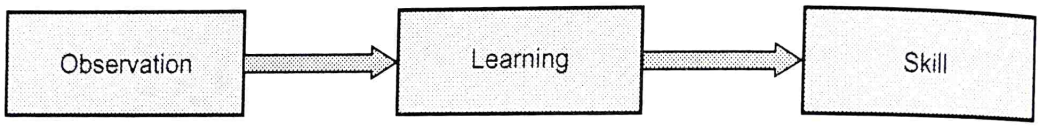


Fig. 1.1.1 Human learning

- 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.

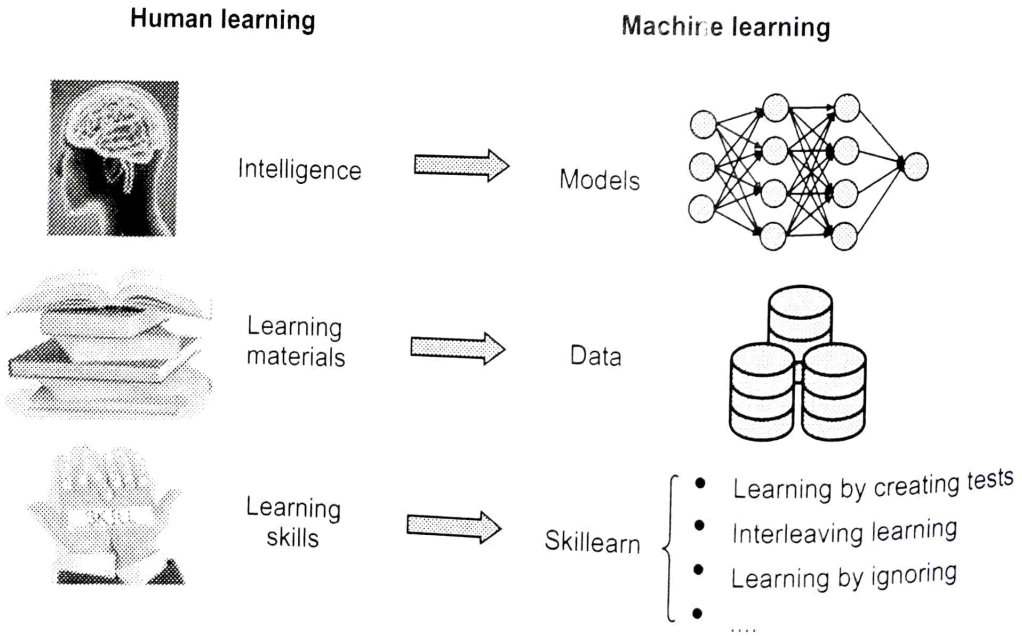


Fig. 1.1.2

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.

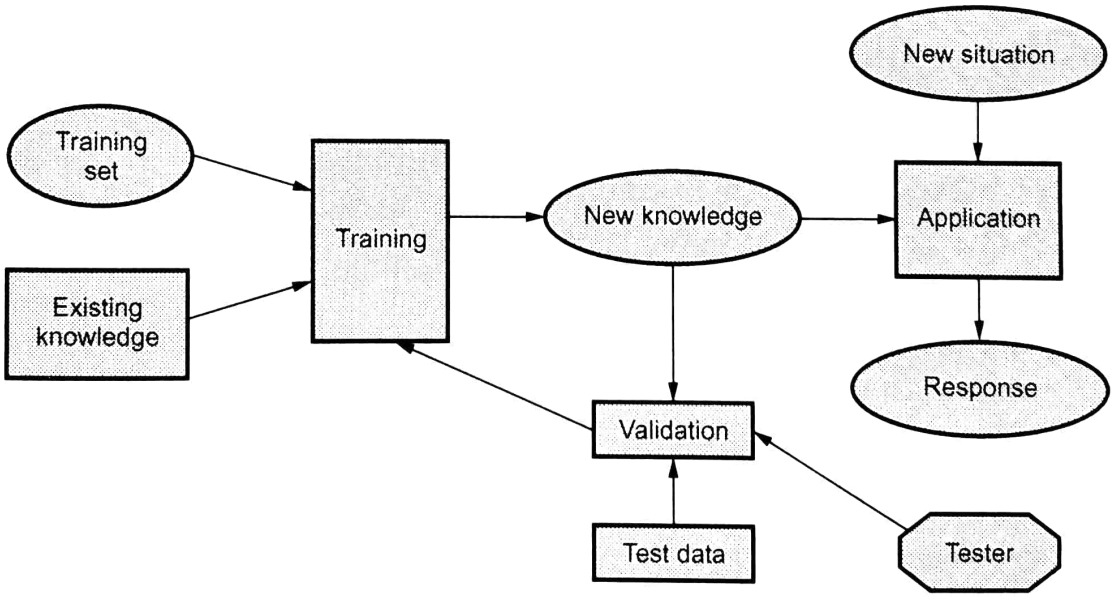


Fig. 1.2.1

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.

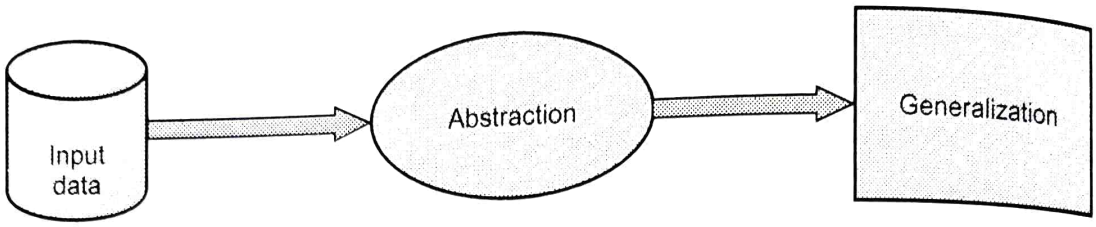


Fig. 1.2.2 Machine learning process

- **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.

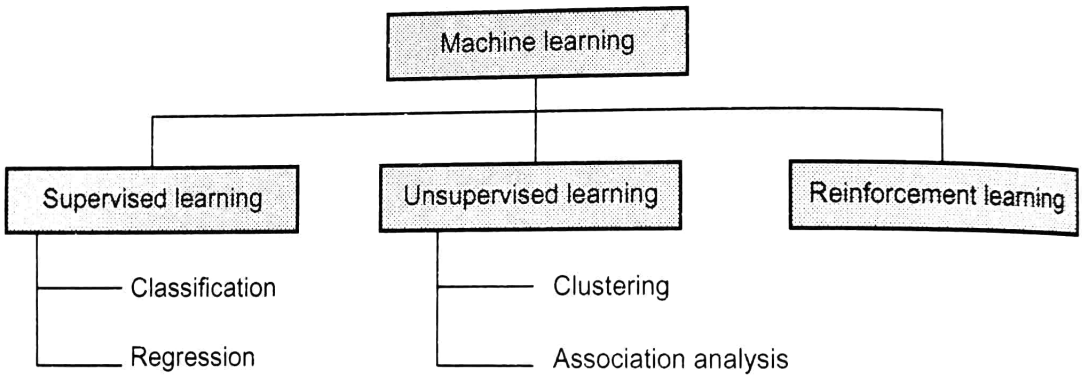


Fig. 1.3.1

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.

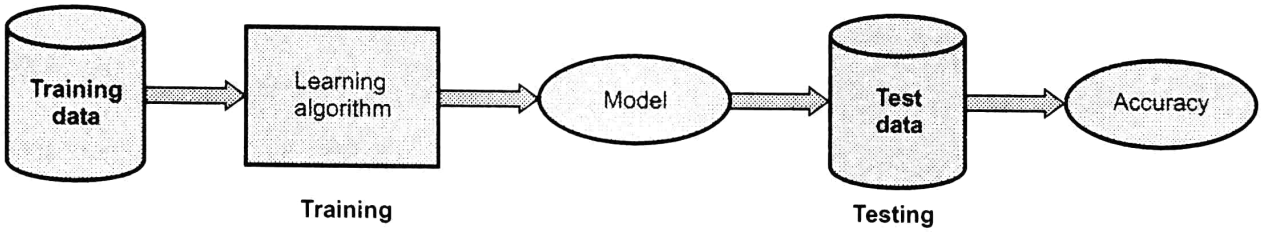


Fig. 1.3.2 Supervised learning process

- 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)

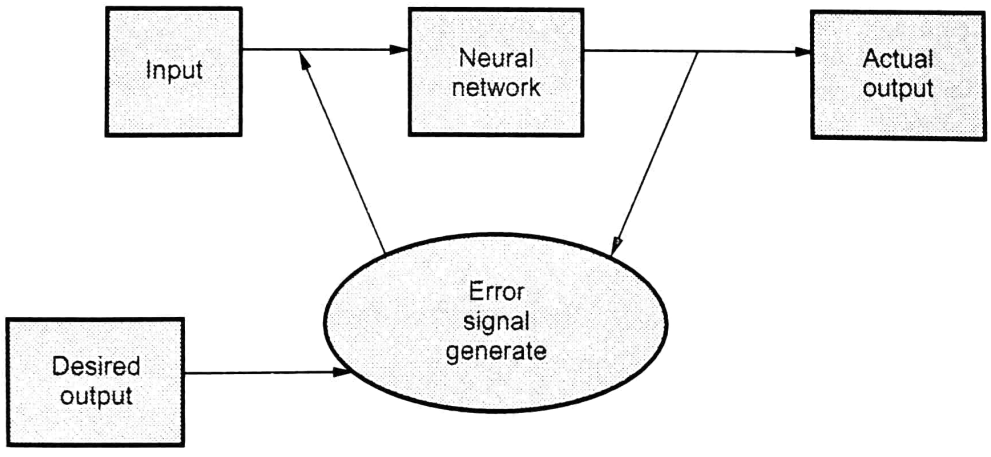


Fig. 1.3.3

- 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.

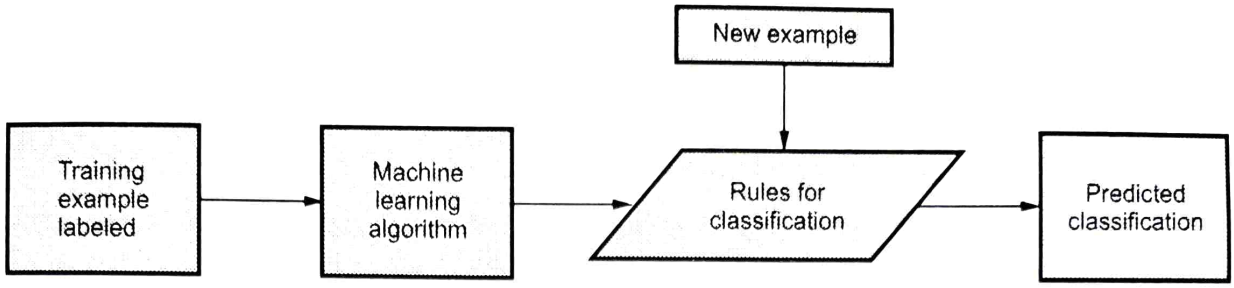


Fig. 1.3.4 Classification

- **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 :

$$X = a + b\,Y$$

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 :

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

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

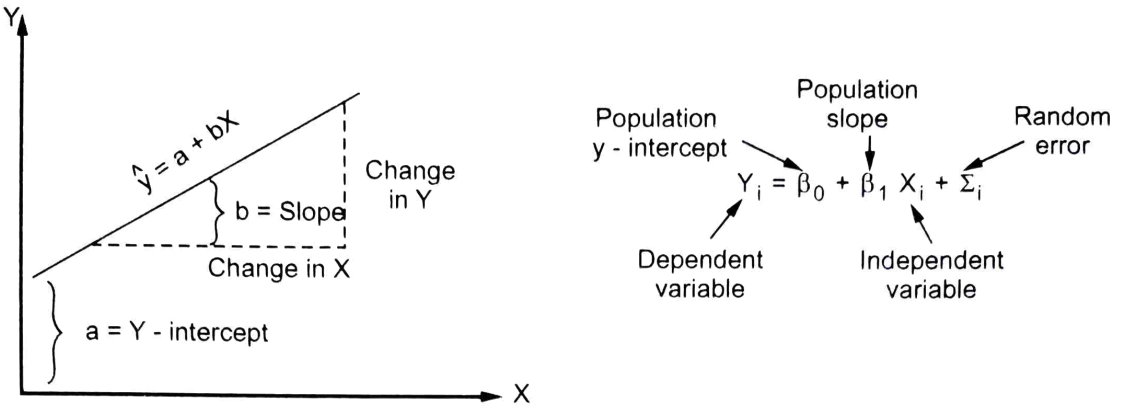


Fig. 1.3.5

- 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 $X_1, ..., X_k$, 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 :

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

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

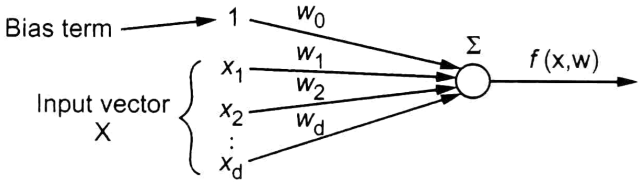


Fig. 1.3.6

- 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.

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

- **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.

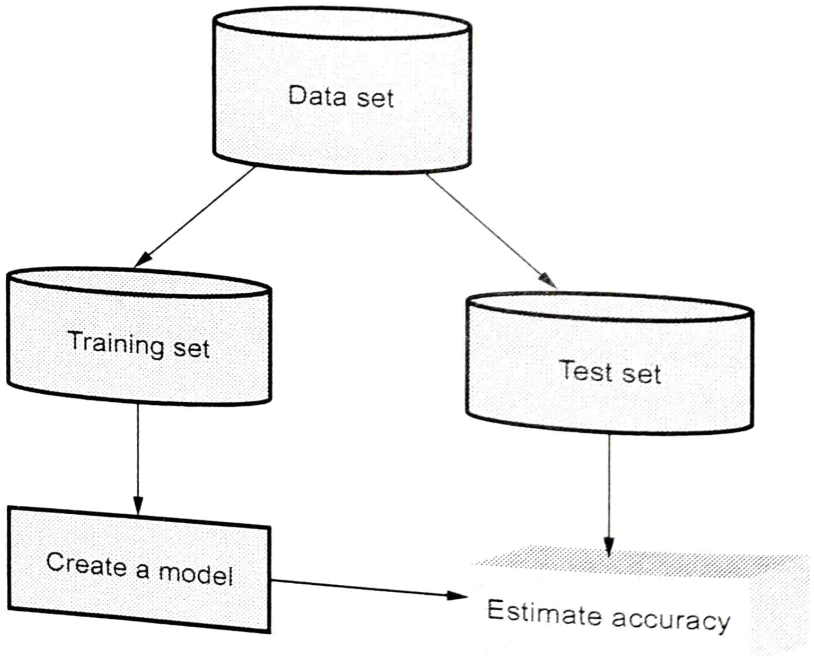


Fig. 1.3.7

- 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.

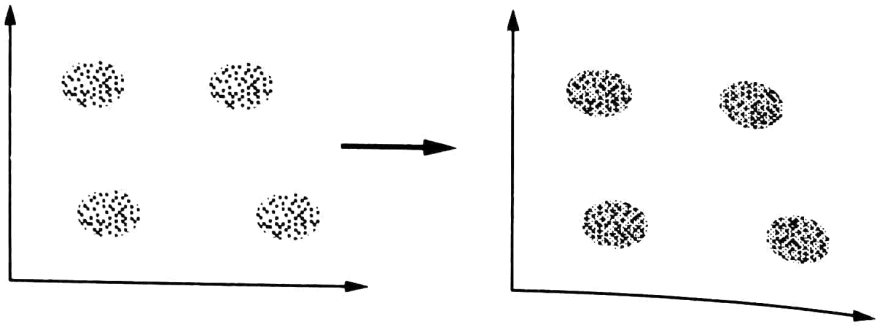
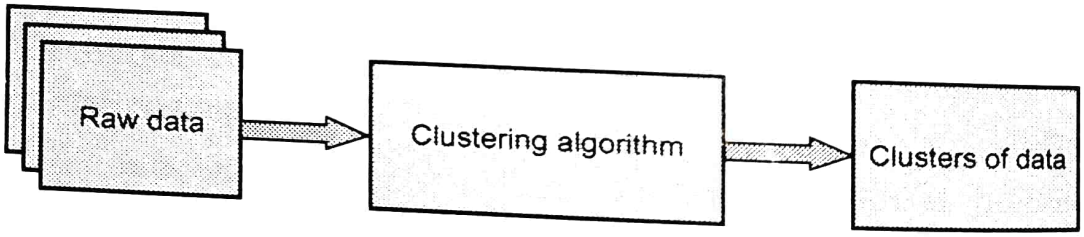
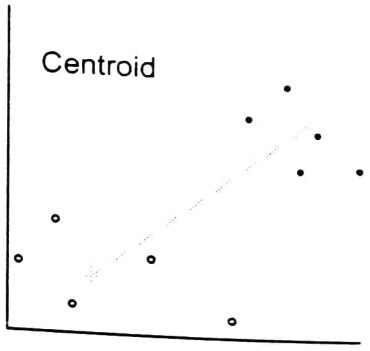


Fig. 1.3.8 Cluster

- 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.

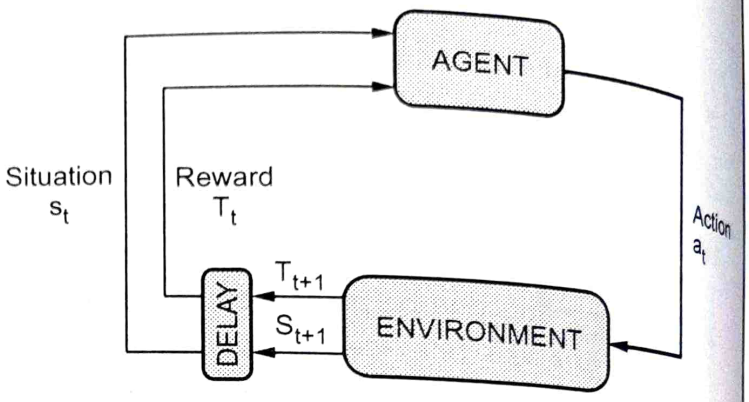


Fig. 1.3.9 Reinforced learning

- 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 elements of reinforcement learning.
- **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.

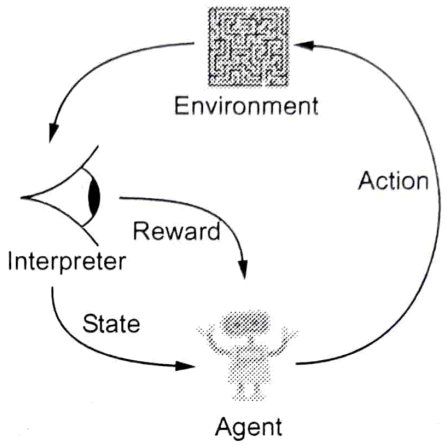


Fig. 1.3.10 : Elements of reinforcement learning

- **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 _____.

2 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 space and inductive bias, Evaluation and Cross Validation, Hypothesis testing, Monte Carlo Approximation.

Contents

2.1 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

2.2 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.

2.2.1 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.

2.2.2 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 :

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

2.2.3 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.

2.2.4 Definition of Probability

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

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

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.,

$$\text{Probability, } P(A) = \lim_{N \to \infty} \frac{N_A}{N} \quad \ldots (2.2.1)$$

$$= \frac{\text{Number of possible favourable outcomes}}{\text{Total number of outcomes}} \quad \ldots (2.2.2)$$

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

Combination of 'n' taken 'r' at a time, $n_{C_r} = \frac{n!}{(n-r)!r!}$ ... (2.2.3)

Permutations of 'n' taken 'r' at a time, $n_{P_r} = \frac{n!}{(n-r)!}$ ... (2.2.4)

Examples for Understanding

Example 2.2.1 *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.

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

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.

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

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

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

Examples with Solutions

Example 2.2.2 *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 2.2.3 *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 $N = {}^{8}C_{3} = \frac{8!}{(8-3)!3!} = \frac{8\times7\times6\times5!}{5!\times3\times2\times1} = 56$

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

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

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

Example 2.2.4 *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 ${}^{4}C_{1}$ ways. i.e.,

$$N(\text{red}) = {}^{4}C_{1} = 4$$

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

ii) *P* (not black)

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

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

$$N(\text{white}) = {}^{3}C_{1} = 3$$

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

iii) *P*(black or white)

Here $N(\text{black}) = {}^{5}C_{1} = 5$

and $N(\text{white}) = {}^{3}C_{1} = 3$

$\therefore$ P(black or white) = P(black) + P(white)

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

Example 2.2.5 *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.

$$\therefore \quad N = {}^{52}C_2 = \frac{52!}{(52-2)!2!} = \frac{52\times51\times50!}{50!\times2\times1} = 1326 \text{ ways}$$

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 $= \frac{13}{1326}\times\frac{12}{1326} = 88.723\times10^{-6}$

ii) Probability that both cards are hearts

$$N_A = {}^{13}C_2 = \frac{13!}{(13-2)!2!} = \frac{13\times12\times11!}{11!\times2\times1} = 78$$

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

Example 2.2.6 *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

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

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

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

ii) Probability of white or yellow ball

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

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

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

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.

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

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

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

2.2.5 Axioms (Properties) of Probability

The outcomes of the trial are said to be *mutually exclusive*, if the occurrence of one o 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 o 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. Ther probability of this event is unity. i.e.,

$$P(A) = P(S) = 1 \quad \text{... (2.2.5)}$$

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

$$0 \leq P(A) \leq 1 \quad \text{... (2.2.6)}$$

- 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

$$P(A + B) = P(A) + P(B) \quad \text{... (2.2.7)}$$

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

Property 1 : $P(\overline{A}) = 1 - P(A)$ ... (2.2.8)

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,

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

$\therefore$ $1 = P(A) + P(\overline{A})$, ... Since $P(S) = 1$ by equation (2.2.5)

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

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

then $$P(A_1) + P(A_2) + \ldots\ldots P(A_M) = 1 \quad \ldots (2.2.9)$$

Proof : The mutually exclusive events satisfy following relation,

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

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

$\therefore$ $P(A_1 + A_2 + \ldots\ldots A_M) = 1$, ... $P(S) = 1$ *for certain event.*

$P(A_1) + P(A_2) + \ldots\ldots + P(A_M) = 1$ ... By equation (2.2.7)

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,

$$P(A+B) = P(A) + P(B) - P(AB) \quad \ldots (2.2.10)$$

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,

$$P(AB) = \lim_{N \to \infty} \frac{N_{AB}}{N} \quad \ldots (2.2.11)$$

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

2.2.6 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.

$$P(B/A) = \frac{P(AB)}{P(A)} \quad \text{and} \quad P(A/B) = \frac{P(AB)}{P(B)} \quad \ldots (2.2.12)$$

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

The joint probability has commutative property i.e.,

$$P(AB) = P(BA) \quad \ldots (2.2.13)$$

2.2.7 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.

$$P(B/A) = \frac{P(AB)}{P(A)} \quad \text{... By equation 2.2.12} \quad \ldots (2.2.14)$$

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.,

$$P(B/A) = P(B) \quad \ldots (2.2.15)$$

With this result equation (2.2.14) becomes,

$$P(AB) = P(A)\,P(B) \quad \ldots (2.2.16)$$

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.

$$\therefore \quad P(A/B) = P(A) \quad \ldots (2.2.17)$$

With above result equation (2.2.12) can be written as, $P(AB) = P(A) \cdot P(B)$, which is same as equation (2.2.16)

Example 2.2.7 *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 2.2.8 *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 2.2.9 *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 2.2.10 *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

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

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

Hence $P(A/B) = 0$

Example 2.2.11 *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 ∪ 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 2.2.12 *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) $P(\overline{A}/B)$

$P(A \cup B) = P(A) + P(\overline{A} \cap B)$, since A and $\overline{A} \cap B$ 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) $P(\overline{A}/\overline{B})$

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

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

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

Example 2.2.13 *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 2.2.14 *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 2.2.15 *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. 2.2.1 shows the digital communication channel. Various probabilities are shown in the figure.

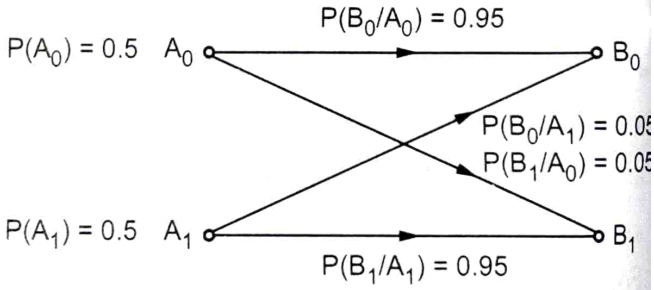


Fig. 2.2.1 Digital communication channel

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

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

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 2.2.16 *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,

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

2.2.8 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,

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

- 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 :

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

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

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

2.2.9 Bayes' Rule

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

$$P = (B_i / A) = \frac{P(B_i)\, P(A / B_i)}{\sum_{i=1}^{n} P(B_i)\, P(A / B_i)} \quad \ldots (2.2.18)$$

This relation is called Bayes' rule or Bayesian Policy.

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

$$P(A) = P(AB_1) + P(AB_2) + \ldots + P(AB_n) \quad \ldots (2.2.19)$$

$$= \sum_{i=1}^{n} P(A B_i) \quad \ldots (2.2.20)$$

$$\therefore \quad P(AB) = P(A)P(B/A) = P(B)P(A/B) \quad \text{... By equation 2.2.19} \quad \text{... (2.2.21)}$$

Now if B has multiple mutually exclusive events B_1, B_2, B_3 ..... B_n then equation (2.2.21) can be written as,

$$P(AB_i) = P(A)\ P(B_i/A) = P(B_i)\ P\ (A/B_i) \quad \text{... (2.2.22)}$$

i.e. $P(A)\ P(B_i/A) = P(B_i)\ P(A/B_i)$

$$\therefore \quad P(B_i/A) = \frac{P(B_i)\ P(A\ /\ B_i)}{P\ (A)} \quad \text{... (2.2.23)}$$

Let us substitute value of $P(AB_i)$ from equation (2.2.22) into equation (2.2.20). i.e.,

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

$$\therefore \quad P(A) = \sum_{i=1}^{n} P\ (B_i\ )\ P(A/\ B_i) \quad \text{... (2.2.24)}$$

Putting value of $P(A)$ from above equation in equation (2.2.23) gives,

$$P(B_i/A) = \frac{P(B_i)\ P(A\ /\ B_i)}{\sum_{i=1}^{n} P(B_i)\ P\ (A\ /\ B_i)} \quad \text{... (2.2.25)}$$

This is the complete proof of Bayes' Rule.

Example 2.2.17 *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.,

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

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

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

P(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 2.2.18 *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 2.2.19 *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

$\therefore \quad P(B_1) = \frac{5}{9}$ 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 2.2.20 *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.,

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

2.3 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 $\{e_1, e_2, e_3, ..., e_i ..., e_n\}$ 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.

2.3.1 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.

2.3.2 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.

2.3.3 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 $\{x_1 < X \leq x_2\}$ with $x_2 > x_1$. The set $\{x_1, x_2\}$ is nonempty and $\in$. Hence

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

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

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

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

or

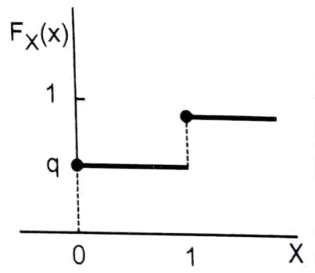
$$p[x_1 < X \leq x_2] = F_X(x_2) - F_X(x_1) \geq 0 \quad \text{for} \quad x_2 > x_1.$$

Some formula :

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

- 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,

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

$$= \lim_{e \to 0} p[x - \in < X \leq x]$$

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

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

Example :

Consider tossing a coin four times. The possible outcomes are contained in t 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 2.3.1 *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 2.3.2 *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	$\frac{1}{2}$

2.3.4 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 $A \in F$, $P(A) = \sum_{\omega \in A} p(\omega)$	For a set $A \in F$, $p(A) = \int_{A}^{-A} f_X(x)\,dx$
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 2.3.3 *Find the constant k such that the function*

$$f(x) = \begin{cases} kx^2 & 0 < x < 3 \\ 0 & otherwise \end{cases}$$ *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

$$k = \frac{1}{9}$$ 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 2.3.4 *If x is a continuous random variable with probability density function given by*

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

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

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

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 :

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

... 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 2.3.6 *A continous random variable X has the distribution function*

$F(X) = 0$ *if* $X \leq 1$

$= k(x-1)^4$ *if* $1 \leq X \leq 3$

$= 1$ *if* $x > 3$

find k and probability density function.

Solution : Probability density function = f(x)

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

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

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

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

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

$$\int_1 = 1$$

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

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

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

$$16k = 1$$

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

Example 2.3.7 *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 ≤ 2), P (0 < x < 2)

c) What is the smallest value of k for which

$F(x \leq k) > 1/2$ *?*

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

$$= k + 2k$$

$$= 3k$$

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

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

$$P(x \leq 2) = P(x=0) + P(x=1) + P(x=2)$$
$$= k + 2k + 3k$$
$$= 6k$$
$$= 6 \times 1/6$$
$$= 1$$

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

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

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

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

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

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 2.3.8 *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,

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

Mean $\mu = nP$

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

$$\mu = 3$$

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

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

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

$$= 0.5767$$

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

Example 2.3.9 *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 2.3.10 *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 2.3.11 *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 : ${}^nC_x\ p^x q^{n-x}$

The probability that none is '1' :

$$P(X = 0) = {}^nC_x\ p^x q^{n-x}$$

$$= {}^4C_0\ (1/9)^0\ (8/9)^{4-0}$$

$$= \frac{4!}{0!\,(4-0)!} \times \frac{1}{1} \times \frac{4096}{6561} = \frac{98304}{6561} = 14.98$$

Example 2.3.12 *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 : ${}^{n}C_{x}\, p^{x} q^{n-x}$

i) All five men

The probability of all the five men being alive is

$$P(X = 5) = {}^{n}C_{x}\, p^{x} q^{n-x}$$

$$= {}^{5}C_{5}\, (2/3)^{5}\, (1/3)^{5-5}$$

$$= \frac{5!}{5!\,(5-5)!} \times \frac{32}{243} \times \frac{1}{1} = \frac{32}{243} = 0.1316$$

ii) At least one man

$$P(X < 1) = 1 - P(x = 0)$$

$$= 1 - {}^{5}C_{0} (2/3)^{0}\, (1/3)^{5-0}$$

$$= 1 - \frac{5!}{0!\,(5-0)!} \times (1) \times \frac{1}{243} = 1 - \frac{1}{243} = \frac{242}{243}$$

iii) Almost three will be alive

The probability of almost three will be alive is :

$$P(x <= 3) = 1 - P(x > 3)$$

$$= 1 - [P(x = 4) + P(x = 5)]$$

$$= 1 - [{}^{5}C_{4}\, (2/3)^{4}\, (1/3)^{5-4} + {}^{5}C_{5}\, (2/3)^{5}\, (1/3)^{5-5}]$$

$$= 1-\left[\frac{5!}{4!\,(5-4)!}\times\frac{16}{81}\times\frac{1}{3}+\frac{32}{243}\right]$$

$$= 1-\left[\frac{120}{24}\times\frac{16}{81}\times\frac{1}{3}+\frac{32}{243}\right]$$

$$= 1-\left[\frac{80}{243}+\frac{32}{243}\right] = \frac{243-112}{243} = \frac{131}{243}$$

Examples on Discrete Distribution

Example 2.3.13 *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, $x = 0, 1, 2, 3, ..., 50.$

$$\binom{50}{x} = \frac{50!}{x!\,(50-x)!}$$

$$P(x \le 1) = P(x = 0) + P(x = 1)$$

$$= P(0) + P(1) = \binom{n}{x}P^x(1-P)^{n-x}$$

$$= \sum_{x=0}^{1}\binom{50}{x}(0.01)^x\,(0.99)^{50-x}$$

$$= \frac{50!}{0!\,(50-0)!}(0.99)^{50}\,(0.01)^0 + \frac{50!}{1!\,(50-1)!}(0.99)^{49}\,(0.01)^1$$

$$= \frac{50!}{50!}(0.605)\,(1) + \frac{50!}{49!}(0.611)\,(0.01)$$

$$= 0.605 + 0.30555$$

$$= 0.91055$$

Example 2.3.14 *Suppose that X has pdf* $f(x) = \frac{24}{x^4}$, $x > 2$. *Evaluate the variance of X.*

Solution : By definition

$$E(X) = \int_2^\infty x\, f(x)\, dx$$

$$= \int_2^\infty x \frac{24}{x^4}\, dx$$

$$= 24 \int_2^\infty x^{-3} dx$$

$$= 24 \left| -\frac{1}{2x^2} \right|_2^\infty$$

$$= 3$$

$$E(X^2) = \int_2^\infty x^2 f(x)\, dx$$

$$= \int_2^\infty x^2 \frac{24}{x^4}\, dx$$

$$= 24 \int_2^\infty x^{-2} dx$$

$$= 24 \left| -\frac{1}{x} \right|_2^\infty = 12$$

$$V(X) = E(X^2) - E(X)^2$$

$$= 12 - (3)^2$$

$$= 12 - 9$$

$$= 3$$

Example 2.3.15 *Consider the function :*

$$f(x) = \begin{cases} 2x, & 0 < x < 1 \\ 0, & elsewhere \end{cases}$$

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 2.3.16 *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 2.3.17 *Let X be a random variable with probability density function*

$$f(x) = \begin{cases} C(1-x^2) & -1 < x < 1 \\ 0 & otherwise \end{cases}$$

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 2.3.18 *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 2.3.19 *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

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

$\therefore$ $$\int_{-3}^{3} f(x)\,dx = 1$$

$$= \int_{-3}^{-1} \frac{1}{16}(3+x)^2\,dx + \int_{-1}^{1} \frac{1}{16}(2-6x)^2 + \int_1^3 \frac{1}{16}(3-x)^2\,dx = 1$$

$$= \int_{-3}^{-1} \frac{1}{16}(3+x)^2\,dx = \int_{-3}^{-1} \frac{1}{16}(9+6x+x^2)\,dx$$

$$= \frac{1}{16}\left[9x+6\frac{x^2}{2}+\frac{x^3}{3}\right]_{-3}^{-1}$$

$$= \frac{1}{16}\left[9(-1)+6\frac{(-1)^2}{2}+\frac{(-1)^3}{3}-\left(9(-3)+6\frac{(-3)^2}{2}+\frac{(-3)^3}{3}\right)\right]$$

$$= \frac{1}{16}\left[\left(-9+\frac{6}{2}-\frac{1}{3}\right)-\left(-27+\frac{54}{2}-\frac{27}{3}\right)\right]$$

$$= \frac{1}{16}\left[\left(-6-\frac{1}{3}\right)-(-27+27-9)\right]$$

$$= \frac{1}{16}\left[\left(\frac{-18-1}{3}\right)-(-9)\right] = \frac{1}{16}\left[\frac{-19}{3}+9\right] = -\frac{19}{16\times 3}+\frac{9}{16}$$

$$= \frac{9}{16}-\frac{19}{48} = \frac{(9\times 3)-19}{48} = \frac{27-19}{48} = \frac{8}{48}$$

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

Mean of $f(x) = \int_{-\infty}^{\infty} x\,f(x)\,dx$

$$= \int_{-3}^{-1} \frac{(3+x)^2}{16}\,x\,dx + \int_{-1}^{3} \frac{6-2x^2}{16}\,x\,dx \int_{-1}^{3} \frac{(3-x)^2}{16}\,x\,dx$$

$$= \frac{1}{16}\int_{-3}^{-1} (9x+6x^2+x^3)\,dx + 0 + \frac{1}{16}\int_{1}^{3} (9x-6x^2+x^3)\,dx$$

$$= \frac{1}{16}\left[\frac{9x^2}{2}+\frac{6x^3}{3}+\frac{x^4}{4}\right]_{-3}^{-1} + \frac{1}{16}\left[\frac{9x^2}{2}-\frac{6x^3}{3}+\frac{x^4}{4}\right]_{1}^{3}$$

$$= \frac{1}{16}\left[\left(\frac{9}{2}-2+\frac{1}{4}\right)-\left(\frac{81}{2}-54+\frac{81}{4}\right)+\frac{1}{16}\right]+\frac{1}{16}\left[\left(\frac{81}{2}-54+\frac{81}{4}\right)-\left(\frac{9}{2}-2+\frac{1}{4}\right)\right]$$

$$= \frac{1}{16}\left[\left(\frac{18-8+1}{4}\right)-\left(\frac{162-216+81}{4}\right)+\left(\frac{162-216+81}{4}\right)-\left(\frac{18-8+1}{4}\right)\right]$$

$$= \frac{1}{16}[0] = 0$$

2.4 Discrete Distributions

- A discrete distribution is a distribution of data in statistics that has discrete va Discrete values are countable, finite, non-negative integers, such as 1, 10, 15, e
- For discrete data key distributions are : Bernoulli, Binomial, Poisson, Multinomial

2.4.1 Binomial Distribution

- Binomial means 'two numbers'.
- The outcomes of health research are often measured by whether they have occ or not. For example, recovered from disease, admitted to hospital, died etc.
- The binomial distribution occurs in games of chance, quality inspection, opi polls, medicine and so on.
- It may be modelled by assuming that the number of events 'n' has a bino distribution with a fixed probability of event p. Binomial distribution distribution for a series of Bernoulli trials.
- Binomial distribution written as B (n, p) where n is the total number of events 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 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 $\binom{n}{r}$ and

$${}^nC_r = \frac{n(n-1)\ldots(n-r+1)}{r!}$$

$$= \frac{n!}{r!(n-r)!}$$

$\binom{n}{r}$ is called **binomial coefficient**.

2.4.1.1 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) :

$$V(X) = E(X^2) - [E(X)]^2$$

$$= \sum_{x=0}^{n} x^2 p(x) - \mu^2$$

$$= \sum_{x=0}^{n} {}^nC_x p^x q^{n-x} x^2 - \mu^2$$

$$= 1 \times 2\,{}^nC_2\, p^2 q^{n-2} + 3 \times 2\, p^3 q^{n-3} + \ldots + {}^nC_n\, n(n-1)\, np$$

$$+ \sum {}^nC_x\, x\, p^x q^{n-x} \mu^2$$

$$= n(n-1)\, p^2 \sum_{x=2}^{n} {}^{n-2}C_{x-2} p^{x-2} q^{n-x} + np - n^2 p^2$$

$$= n(n-1)\, p^2 (p+q)^{n-2} + np - n^2 p^2$$

$$= n(n-1)\, p^2 + np - n^2 p^2$$

$$= n^2 p^2 - np^2 + np - n^2 p^2$$

$$= np - np^2$$

$$= np(1-p)$$

$$\sigma^2 = npq$$

(q = 1 –

- The standard deviation (σ) of the binomial distribution is $\sqrt{npq}$.

An examples of the binomial distribution

Example 2.4.1 *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 rando variable.

Some values of X will be more likely to occur than others. Each value of X will ha a probability of occuring. Consider the probability of obtaining just one yellow ball, 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,

$$P(X = 1) = 5\times\left(\frac{2}{3}\times\left(\frac{1}{3}\right)^4\right)$$

$$= 5\times\frac{2}{3}\times\frac{1}{81}$$

$$= 0.04115$$

We calculate the probability P (X = 2) :

$$P(X = 2) = \text{Number of patterns} \times \text{Probability of pattern}$$

$$= {}^5C_2\times(2/3)^2\times(1/3)^3$$

$$= \frac{5!}{2!(5-2)!}\times\frac{4}{9}\times\frac{1}{27} = \frac{120}{12}\times\frac{4}{9}\times\frac{1}{27}$$

$$= 10\times\frac{4}{243} = 10\times 0.01646$$

$$= 0.1646$$

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 :

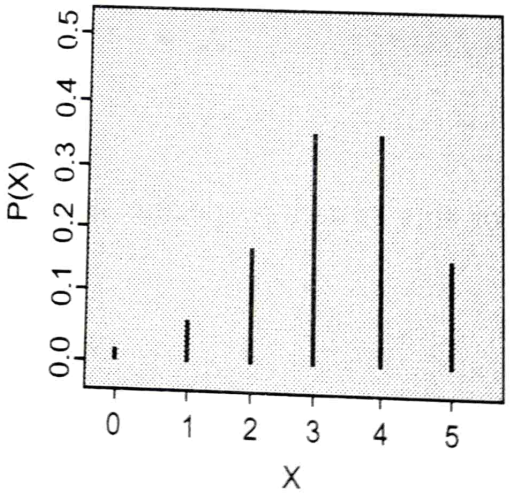


Fig. 2.4.1 A plot of the Binomial (5, 2/3) probabilities

Example 2.4.2 *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 :

$$\mu = xP(X=0) + xP(X=1) + xP(X=2) + xP(X=3) + xP(X=4) + xP(X=5)$$

$$= 0\times(0.004) + 1\times(0.0041) + 2\times(0.165) + 3\times(0.329) + 4\times(0.329) + 5\times(0.132)$$

$$= 0 + 0.0041 + 0.33 + 0.987 + 1.316 + 0.66$$

$$= 3.2971$$

2.4.1.2 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 2.4.3 *Find the variance and standard deviation for the following set of test marks T = {75, 80, 82, 87, 96}*

Solution :

$$\text{Mean} = \frac{75+80+82+87+96}{5}$$

$$= \frac{420}{5}$$

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 2.4.4 *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 2.4.5 *If X is a normal variate with mean 30 and standard deviation 5. Find the probability that,*

a) $26 \leq x \leq 40$ *b)* $x \geq 45$.

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 2.4.6 *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 ≥ 6), P (0 < X < 5) and P(0 ≤ X ≤ 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.

$P(X \leq 1) = P(X = 0) + P(X = 1)$

$= 0 + K$

$= K$

$= 0.1$

$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

$= 0 + K + 2K$

$= 3K$

$= 3 \times 0.1$

$= 0.3$

$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$$= 0 + K + 2K + 2K$$

$$= 5K$$

$$= 5 \times 0.1$$

$$= 0.5$$

$P(X \leq 4) = 0.8$ (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) = $P(X \leq x)$
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)

$$\mu = \sum_{i=0}^{7} p_i x_i$$

$$= 0(0) + 1(K) + 2(2K) + 3(2K) + 4(3K) + 5(K^2) + 6(2K^2) + 7(7K^2 + K)$$

$$= 0 + K + 4K + 6K + 12K + 5K^2 + 12K^2 + 49K^2 + 7K$$

$$= 30K + 66K^2$$

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 2.4.7 *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

$$= {}^{10}C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-n}$$

$$= {}^{10}C_x \left(\frac{1}{2}\right)^{10}$$ (Where x = 0, 1, 2, 3, ..., 10)

$$P(X = 3) = {}^{10}C_3 \left(\frac{1}{2}\right)^{10}$$

$$= \frac{10!}{3!\,(10-3)!} \times \frac{1}{1024}$$

$$= \frac{3628800}{6 \times 5040} \times \frac{1}{1024}$$

$$= \frac{120}{1024}$$

$$P(X = 3) = 0.11718$$

$$P(X = 4) = {}^{10}C_4 \left(\frac{1}{2}\right)^{10}$$

$$= \frac{10!}{4!\,(10-4)!} \times \frac{1}{1024}$$

$$= \frac{3628800}{24 \times 720} \times \frac{1}{1024}$$

$$= \frac{210}{1024}$$

$$P(X = 4) = 0.2050$$

$$P(X = 5) = {}^{10}C_5 \left(\frac{1}{2}\right)^{10}$$

$$= \frac{10!}{5!\,(10-5)!} \times \frac{1}{1024}$$

$$= \frac{3628800}{120 \times 120} \times \frac{1}{1024}$$

$$= \frac{252}{1024}$$

$$= 0.246$$

Example 2.4.8 *The mean of binomial distribution is 3 and variance is 9/4. Find*

i) The value of n

ii) $p(x \geq 7)$

iii) $p(1 \leq x \leq 6)$

olution : 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

$$= {}^{12}C_1\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^{12-1} + {}^{12}C_2\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^{12-2} + {}^{12}C_3\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^{12-3}$$

$$+ {}^{12}C_4\left(\frac{1}{4}\right)^4\left(\frac{3}{4}\right)^{12-4} + {}^{12}C_5\left(\frac{1}{4}\right)^5\left(\frac{3}{4}\right)^{12-5}$$

$$= 0.1267 + 0.2322 + 0.2581 + 0.1935 + 0.1032$$

$$= 0.9137$$

2.4.2 The Poisson Distribution

- Poisson distribution, named after its invertor simeon poisson who was a Frenc mathematician. He found that if we have a rare event (i.e. p is small) and w know the expected or mean (or μ) number of occurances, the probabilities of 0, 2 ... events are given by :

$$P(R) = \frac{e^{-\mu}\mu^R}{R!}$$

Poisson distribution : Is a distribution the number of rare events that occur in a un 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 2.4.9 *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

$$= \frac{e^{-\mu}\mu^{x}}{x!}$$

$$= \frac{e^{-2}\,2^{3}}{3!}$$

$$= 0.18$$

2.4.3 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.

2.4.4 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

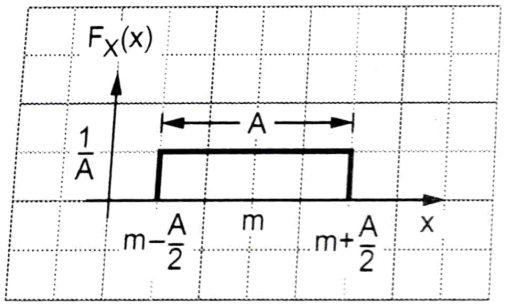
2.5 Continuous Distributions

- Continuous distributions are characterized by an infinite number of possible outcomes, together with the probability of observing a range of these outcomes.

2.5.1 Uniform Distribution

The PDF for a uniform distribution is given as,

$$\text{Uniform PDF: } f_X(x) = \begin{cases} 0 & \text{for } x < m - \frac{A}{2} \text{ and } x > m + \frac{A}{2} \\ \frac{1}{A} & \text{for } \left(m - \frac{A}{2}\right) \leq x \leq \left(m + \frac{A}{2}\right) \end{cases} \quad \ldots (2.5.1)$$



A = Peak to peak value of a random variable.

$\frac{1}{A}$ = Amplitude of all possible values of random variable

Fig. 2.5.1 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 2.5.1 *Show that the mean and variance of a random variable 'X' having a uniform distribution in the interval [a, b] are,*

$m_x = \frac{a+b}{2}$ *and* $\sigma_x^2 = \frac{(a-b)^2}{12}$

Solution : Fig. 2.5.2 shows the sketch of uniform distribution in the interval [a, b].

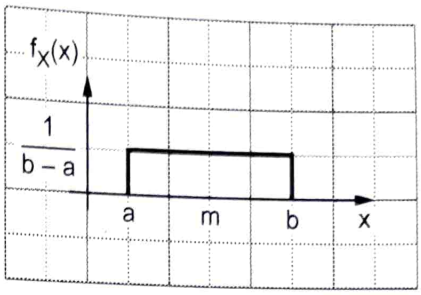


Fig. 2.5.2 Uniform distribution having interval [a, b]

To find mean value

The mean value of a continuous random variable is given by

$$m_x = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

$$= \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b$$

$$= \frac{1}{2(b-a)}[b^2 - a^2] = \frac{1}{2(b-a)}(b-a)\cdot(b+a)$$

$$= \frac{b+a}{2} \quad or \quad \frac{a+b}{2} \qquad \ldots(2.5.2)$$

To find variance

Variance is given by equation (2.2.8) 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 (2.5.2) in above equation we get,

$$\sigma_x^2 = \frac{1}{3(b-a)}\left[\left(b-\frac{a+b}{2}\right)^3 - \left(a-\frac{a+b}{2}\right)^3\right] = \frac{1}{3(b-a)}\left[\left(\frac{b-a}{2}\right)^3 - \left(\frac{a-b}{2}\right)^3\right]$$

$$= \frac{1}{-3(a-b)}\left[\left(-\frac{a-b}{2}\right)^3 - \left(\frac{a-b}{2}\right)^3\right] \quad \text{Here we have written } b-a = -(a-b)$$

$$= \frac{1}{-3(a-b)} \times \frac{-(a-b)^3}{4} = \frac{(a-b)^2}{12} \quad \ldots (2.5$$

Thus, for uniform distribution,

$$\text{Mean, } m_x = \frac{a+b}{2} = m \text{ and variance, } \sigma_x^2 = \frac{(a-b)^2}{12} \quad \ldots (2.5$$

2.5.2 Normal Distribution

Gaussian distribution is also called *Normal Distribution*. It is defined for continuo random variables. The *PDF* for a Gaussian random variable is given as,

$$\textbf{Gaussian PDF : } f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-m)^2/2\sigma^2} \quad \ldots (2.5$$

Here 'm' is mean and σ^2 is variance.

Fig. 2.5.3 shows the sketch of Gaussian pdf.

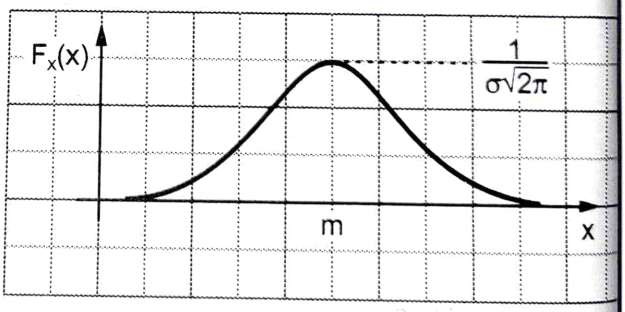


Fig. 2.5.3 Plot of Gaussian *PDF*

Properties of Gaussian *PDF*

Property 1 : The peak value occurs at $x = m$ (i.e. mean value). i.e.,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \quad \text{at} \quad x = m \quad \text{i.e. mean value} \quad \ldots (2.5$$

Property 2 : The plot of Gaussian *PDF* has even symmetry around mean value i.e.,

$$f_X(m-\sigma) = f_X(m+\sigma) \quad \ldots (2.5$$

Property 3 : The area under the *PDF* curve is 1/2 for all values of x below mean va and 1/2 for all values of x above mean value. i.e.,

$$P(X \leq m) = P(X > m) = \frac{1}{2} \qquad \ldots (2.5.8)$$

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 2.5.2 *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 (2.5.5) in above equation,

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2}\, dx \qquad \ldots (2.5.9)$$

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 (2.5.9) we get,

$$F_X(x) = \int_{\infty}^{\frac{m-x}{\sigma\sqrt{2}}} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-z^2} \cdot (-\sigma\sqrt{2}\, dz) = -\frac{1}{\sqrt{\pi}} \int_{\infty}^{\frac{m-x}{\sigma\sqrt{2}}} \cdot e^{-z^2} \cdot dz$$

$$= \frac{1}{\sqrt{\pi}} \int_{\frac{m-x}{\sigma\sqrt{2}}}^{\infty} \cdot e^{-z^2} \cdot dz \qquad \text{By interchanging the limits.}$$

$$= \frac{1}{2} \cdot \frac{2}{\sqrt{\pi}} \int_{\frac{m-x}{\sigma\sqrt{2}}}^{\infty} \cdot e^{-z^2} \cdot dz \qquad \text{By rearranging} \qquad \ldots (2.5.10)$$

The above integration is represented by error function. It is given as,

$$erfc(u) = \frac{2}{\sqrt{\pi}} \int_{u}^{\infty} e^{-z^2} \, dz \quad \ldots (2.5.11)$$

Gaussian CDF : $F_X(x) = \frac{1}{2} erfc\left(\frac{m - x}{\sigma\sqrt{2}}\right)$ ... (2.5.12)

Example 2.5.3 *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

$$\int_{-\infty}^{\infty} f_X(x) \, dx = 1 \quad \ldots (2.5.13$$

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 $f_X(x)$ ... (2.5.14

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 (2.5.14) becomes,

$$I = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-y^2/2} \cdot \sigma \, dy = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \, dy \quad \ldots (2.5.15$$

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.,

$$I^2 = \underbrace{\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dy \right)}_{\text{Variable is changed from y to x in this term. It will not change value of integration}} \times \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\, dy \right)$$

$$= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\, dx\, dy \qquad \ldots (2.5.16)$$

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 (2.6.12) becomes,

$$I^2 = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2}\, r\, dr\, d\phi = \frac{1}{2\pi} \int_0^{2\pi} d\phi \int_0^{\infty} e^{-r^2/2}\, r\, dr \qquad \ldots (2.5.17)$$

Put $\frac{r^2}{2} = t \qquad \Rightarrow \qquad r\, dr = dt$

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.

$$I = 1.$$

Area under Gaussian PDF is unity : $I = \int_{-\infty}^{\infty} f_X(x)\, dx = 1$ ... (2.5.18)

Example 2.5.4 *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,

$$m_x = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/2\sigma^2} dx$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-(x-m)^2/2\sigma^2}\, dx \quad \ldots (2.5.19)$$

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,

Mean value of Gaussian distribution $= m_x = m$ ... (2.5.20)

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 $\frac{x - m}{\sqrt{2}\,\sigma} = z \quad \Rightarrow \quad dx = \sigma\sqrt{2}\, dz$

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'.

$$= \frac{4\sigma^2}{\sqrt{\pi}} \int_{0}^{\infty} z^2 e^{-z^2} dz \qquad \ldots (2.5.21)$$

Here use the standard relation given in Appendix i.e;

$$\int_0^{\infty} x^{2n} e^{-ax^2} \, dx = \frac{1 \cdot 3 \cdot 5 \ldots (2n-1)}{2^{n+1} a^n} \sqrt{\frac{\pi}{a}}$$

In equation (2.5.21), $n = 1$ 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 : $\sigma_x^2 = \sigma^2$ ... (2.5.22)

Example 2.5.5 *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,

$$P(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f_X(x)\,dx \qquad \ldots (2.5.2$$

i) $P(X > 20)$: Above probability can be written as $P(20 < X \leq \infty)$. Putting values equation (2.6.19),

$$P(X > 20) = P(20 < X \leq \infty) = \int_{20}^{\infty} \frac{1}{20\sqrt{2\pi}} e^{-(x-10)^2/800}\,dx \qquad \ldots (2.5.2$$

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 (2.5.24),

$$P(X > 20) = \int_{0.5}^{\infty} \frac{1}{20\sqrt{2\pi}} e^{-z^2/2} \cdot 20\,dz = \frac{1}{\sqrt{2\pi}} \int_{0.5}^{\infty} e^{-z^2/2}\,dz$$

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 (2.5.24),

$$P(10 < X \leq 20) = P(X > 10) - P(X > 20) = P(10 < X \leq \infty) - 0.31$$

$$= \int_{10}^{\infty} \frac{1}{20\sqrt{2\pi}} e^{-(x-10)^2/800}\,dx - 0.31$$

Put $\frac{x-10}{20} = z \quad \therefore dx = 20\,dz$

When $x = 10$, $z = 0$ and when $x = \infty$, $z = \infty$

$$\therefore P(10 < X \leq 20) = \int_0^{\infty} \frac{1}{20\sqrt{2\pi}} e^{-z^2/2}\, 20\,dz - 0.31 = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-z^2/2}\, dz - 0.31$$

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,

$$P(10 < X \leq 20) = Q(0) - 0.31 = 0.5 - 0.31 \quad \text{since } Q(0) = 0.5 \text{ given}$$
$$= 0.19$$

iii) $P(0 < X \leq 20)$: Fig. 2.5.4 shows the plot of given Gaussian pdf. If has the mean value of $m = 10$. We have obtained $P(10 < X \leq 20)$. Due to symmetry of the pdf curve, $P(0 < X \leq 20)$ will be twice of $P(10 < X \leq 20)$. Thus,

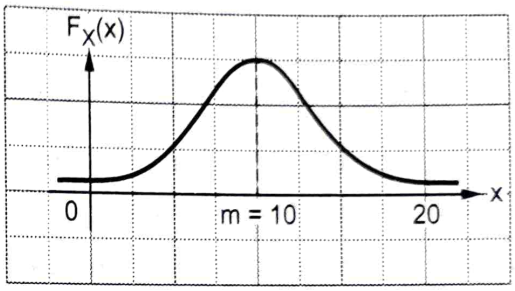


Fig. 2.5.4 Sketch of pdf

$$P(0 < X \leq 20) = 2 \times P(10 < X \leq 20)$$
$$= 2 \times 0.19 = 0.38$$

iv) $P(X > 0)$: From the pdf curve of Fig. 2.5.4,

$$P(X > 0) = P(0 < X \leq 20) + P(X > 20) = 0.38 + 0.31 = 0.69$$

v) $P(X < 20)$: From the pdf curve of Fig. 2.5.4,

$$P(X < 20) = 1 - P(X > 20) = 1 - 0.31 = 0.69$$

2.6 Multiple Random Variables

2.6.1 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. $f_{XY}(x, y) \geq 0$ 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.

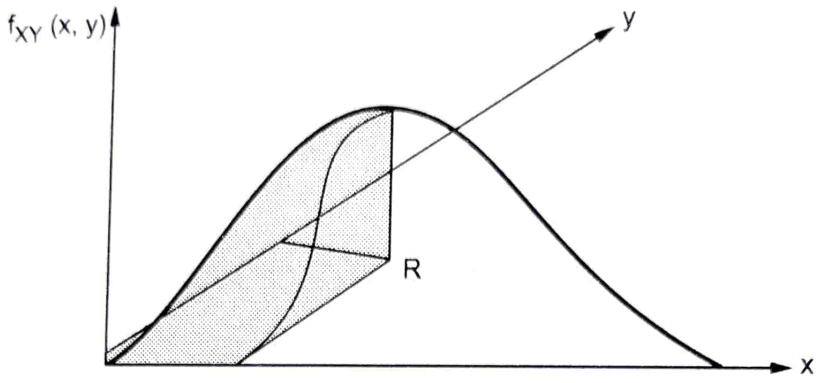


Fig. 2.6.1 Region R

- 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.

2.6.2 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) $f_{XY}(x, y) = P(X = x, Y = y)$

2.6.3 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 $\{(x, y) : a \leq x \leq b, c \leq y \leq d\}$,

$$P[(XY) \in A] = \int_a^b \int_c^d f(x\ y)\ dx\ dy$$

$P[(XY) \in A]$ = Volume under density surface above A.

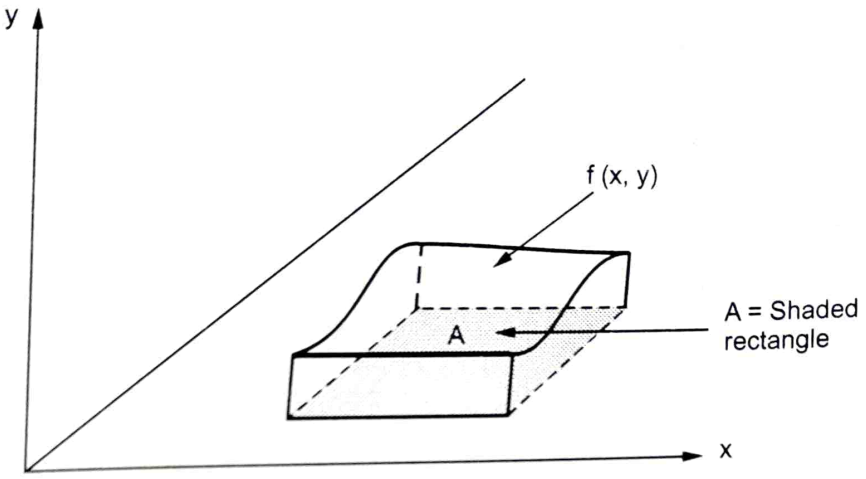


Fig. 2.6.2

Example 2.6.1 *X and Y are jointly continuous with joint pdf*

$$f(x,y) = \begin{cases} Cx^2 + \frac{xy}{3}, & 0 \leq x \leq 1, 0 \leq y \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

i) Find C

ii) Find marginal pdf of X and of Y.

iii) Find $P(X + Y \geq 1)$

Solution :

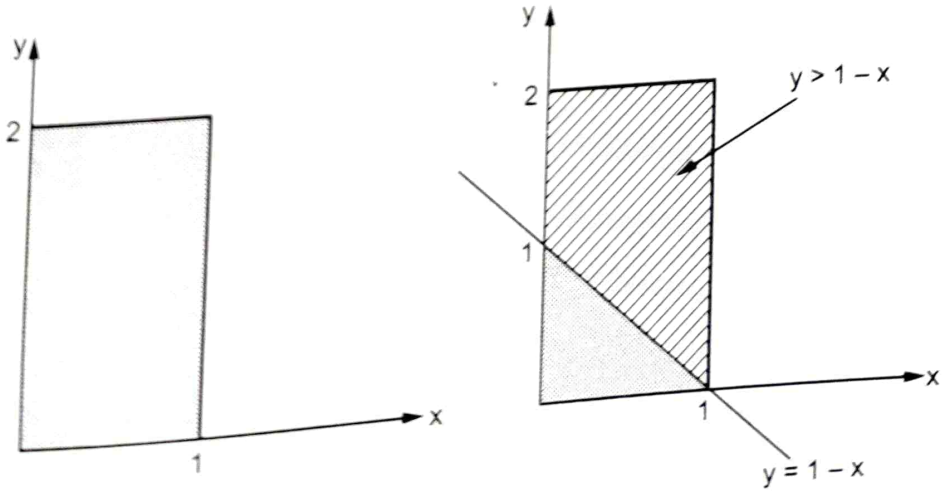


Fig. 2.6.3

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 2.6.2 *Compute the covariance between X and Y for following joint probability density function.*

$$f_{XY}(x,y) = \begin{cases} \frac{1}{3} y \ exp(-xy), & if\ x \in (0,\infty) \ and \ y \in (1,4) \\ 0, & otherwise \end{cases}$$

$$R_{XY} = (0,\infty) \times (1,4)$$

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 $$E(X) = \int_{-\infty}^{\infty} f_X(x)\, dx$$

$$= \int_0^{\infty} x \left[\int_1^4 \frac{1}{3} y \exp(-xy)\, dy \right] dx$$

$$= \frac{1}{3} \int_1^4 \left(\int_0^{\infty} xy \exp(-xy)\, dx \right) dy$$

$$= \frac{1}{3} \int_1^4 \left(\frac{1}{y} \int_0^{\infty} t \exp(-t)\, dt \right) dy \qquad [\because t = x$$

$$= \frac{1}{3} \int_1^4 \frac{1}{y} \left([-t \exp(-t)]_0^{\infty} + \int_0^{\infty} \exp(-t)\, dt \right) dy$$

$$= \frac{1}{3} \int_1^4 \frac{1}{y} (0 + [-\exp(-t)]_0^{\infty})\, dy$$

$$= \frac{1}{3} \int_1^4 \frac{1}{y} dy = \frac{1}{3} [\ln(y)]_1^4$$

$$E(X) = \frac{1}{3} \ln(4)$$

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

$$Cov[X, Y] = E[XY] - E[X]\,E[Y]$$

$$= 1 - \left(\frac{1}{3}\ln(4)\right)\left(\frac{5}{2}\right) = 1 - \frac{5}{6}\ln(4)$$

Example 2.6.3 *Let X and Y be two random variables such that*

$Var[X] = 4 \quad Cov[X, Y] = 2$

Calculate the variance for following

$Cov[3X, X + 3Y]$

Solution :

$$Cov[3X, X+3Y] = 3\,Cov[X, X+3Y]$$

$$= 3\,Cov[X, X] + 9\,Cov[X, Y]$$

$$= 3\,Var[X] + 9\,Cov[X, Y]$$

$$= 3(4) + 9(2) = 12 + 18 = 30$$

Example 2.6.4 *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 function.

Example 2.6.5 *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 2.6.6 *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 :

$$f_X(x) = \int_x^{\infty} e^{-y}dy = -e^{-y}\Big|_x^0 = e^{-x},\quad 0 < x < \infty$$

$$f_Y(y) = \int_0^{y} e^{-y}dx = ye^{-y},\quad 0 < y < \infty$$

Example 2.6.7 *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 :

$$P(X + Y < 1) = \int_0^1 \int_0^{1-x} (x+y)\, dy\, dx$$

$$= \int_0^1 xy + \frac{y^2}{2}\Big|_{y=0}^{y=1-x} dx$$

$$= \frac{1}{2}\int_0^1 (1-x^2)\, dx = \frac{1}{2}\left(x - \frac{x^3}{3}\Big|_0^1 \right)$$

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

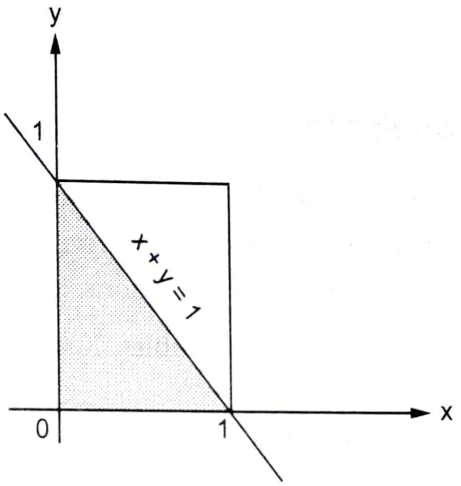


Fig. 2.6.4

Example 2.6.8 *Compute Var (X) when X is roll of a fair die outcome.*

Solution : For fair die

$$P\{X = i\} = \frac{1}{6}$$

where $i = 1, 2, 3, 4, 5, 6$

$$E(X^2) = \sum_{i=1}^{6} i^2 P\{X = i\}$$

$$= \frac{1}{6}[(1)^2 + (2)^2 + (3)^2 + (4)^2 + (5)^2 + (6)^2]$$

$$= \frac{[1 + 4 + 9 + 16 + 25 + 36]}{6}$$

$$E(X^2) = \frac{91}{6}$$

$$E(X) = 1\left(\frac{1}{6}\right)+2\left(\frac{1}{6}\right)+3\left(\frac{1}{6}\right)+4\left(\frac{1}{6}\right)+5\left(\frac{1}{6}\right)+6\left(\frac{1}{6}\right)$$

$$= \frac{1}{6}+\frac{2}{6}+\frac{3}{6}+\frac{4}{6}+\frac{5}{6}+\frac{6}{6} = \frac{21}{6}$$

$$E(X) = \frac{7}{2}$$

$$Var(X) = E(X^2) - (E[X])^2$$

$$= \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{364-294}{24} = \frac{70}{24}$$

$$Var(X) = \frac{35}{12}$$

2.6.4 Covariance and Correlation

- Covariance is a measure of association between two random variables. It positive if the deviations of the two variables from their respective means tend 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 defined as follows :

$$Cov[X, Y] = E[(X - E[X])\,(Y - E[Y])]$$

or

$$Cov[X, Y] = E[(X-\mu_x)(Y-\mu_y)]$$

- Covariance indicates how two variables are related. A positive covariance me the variables are positively related, while a negative covariance means variables are inversely related. The formula for calculating covariance of sam data is shown below.

$$Cov[X, Y] = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

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

$\bar{Y}$ = The mean of dependent variable y

$$\begin{aligned} Cov[X, Y] &= E(XY - \mu_x Y - \mu_y X + \mu_x \mu_y) \\ &= E(XY) - \mu_x E[Y] - \mu_y E[X] + \mu_x \mu_y \\ &= E(XY) - \mu_x \mu_y - \mu_y \mu_x + \mu_x \mu_y \\ &= E[XY] - E[X]\ E[Y] \end{aligned}$$

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.

2.7 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.

2.8 Sampling Distributions

2.8.1 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.

2.8.2 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

2.8.3 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.

$$C_n^N = \frac{N!}{n!(N-n)!} = \frac{4!}{2!(4-2)!} = 6$$

2.8.4 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 ($n < 30$).

2.8.5 Mean, Medium and Mode

1. Mean : The mean of a data set is the average of all the data values. The sampl mean $\bar{x}$ is the point estimator of the population mean μ.

$$\textbf{Sample mean } \bar{x} = \frac{\text{Sum of the values of the n observations}}{\text{Number of observations in the sample}} = \frac{\sum x_i}{n}$$

$$\textbf{Population mean } \mu = \frac{\text{Sum of the values of the N observations}}{\text{Number of observations in the population}} = \frac{\sum x_i}{N}$$

2. Median

- The median of a data set is the value in the middle when the data items ar arranged in ascending order. Whenever a data set has extreme values, the media 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 : $S^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$

Population variance : $\sigma^2 = \dfrac{\sum (x_i - \mu)^2}{N}$

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 :

$$\text{Population standard deviation} = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$

$$\text{Sample standard deviation} = S = \sqrt{S^2} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

2.8.6 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	
n = 5	Total = 798		2.56
	(Step 1) Mean m = 159.6		**(Step 4)** Sum of squared deviations $\sum (m-i)^2 = 253.2$

Step 9 : Divide by number of measurements – 1 :

$$\frac{\sum (m-i)^2}{n-1} = \frac{253.2}{5-1} = \mathbf{63.3}$$

Step 10 : Standard deviation $= \dfrac{\text{Square root of} \sum (m-i)^2}{n-1} = \dfrac{\sqrt{63.3}}{4} = 1.9890$

Step 11 : Standard error $= \dfrac{\text{Standard deviation}}{\sqrt{n}} = \dfrac{1.9890}{\sqrt{4}} = 0.9945$

Step 12 : m ± 1SE = 159.6 ± 0.9945

= 159.6 + 0.9945 or 159.6 – 0.9945

= 160.5945 or 158.6055

Example 2.8.1 *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

$$\text{Sample mean} = \frac{188+214+204}{3} = \frac{606}{3} = 202$$

$$\text{Sample variance} = \frac{(188-202)^2+(214-202)^2+(204-202)^2}{3-1} = \frac{196+144+4}{2} = \frac{344}{2} = 172$$

Standard deviation = $\sqrt{172} = 13.11$

Example 2.8.2 *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 $\bar{x} = \dfrac{\sum x_i}{n}$

$$\bar{x} = \frac{x_1+x_2+x_3+x_4+x_5+x_6}{6} = \frac{27+15+20+32+18+26}{6} = \frac{138}{6} = 23$$

Estimate of variance $(S^2) = \dfrac{\sum (x_i - \bar{x})^2}{n-1}$

$$S^2 = \frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+(x_4-\bar{x})^2+(x_5-\bar{x})^2+(x_6-\bar{x})^2}{6-1}$$

$$= \frac{(27-23)^2+(15-23)^2+(20-23)^2+(32-23)^2+(18-23)^2+(26-23)^2}{5}$$

$$= \frac{16+64+9+81+25+9}{5} = \frac{204}{5} = 40.8$$

$S^2 = 40.8$

$S = 6.387$

$$t = \frac{x-\mu}{S/\sqrt{n}} = \frac{23-20}{6.387/\sqrt{6}} = \frac{3}{6.387/2.449}$$

$t = 1.15$

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.

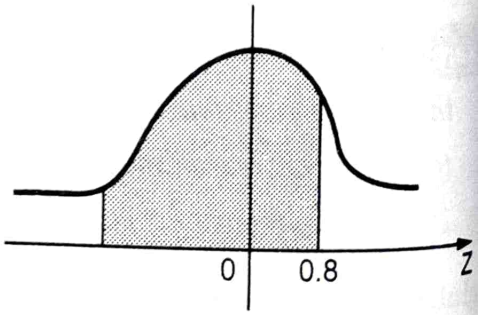
Example 2.8.3 *A normal population has a mean of 0.1 and standard deviation of 2.1. Find the probability that the mean of simple, sample of 900 members will be negative.*

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 2.8.4 *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 2.8.5 *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 2.8.6 *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 :*

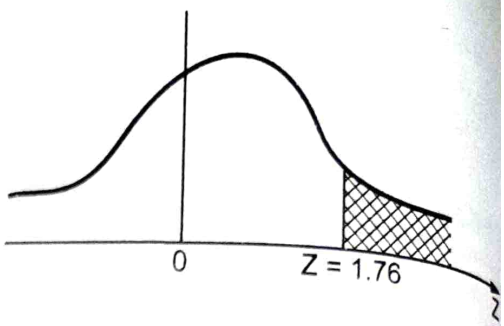
i) Exceed 52.9 ii) Fall between 50.5 and 52.3 iii) Be less than 50.6.

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



$$P(\bar{x} \text{ exceed } 52.9) = P(\bar{x} > 52.9)$$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{52.9 - 51.4}{0.85} = 1.76$$

$$P(\bar{x} > 52.9) = P(Z > 1.76)$$

$$= 0.5 - P(0 < Z < 1.76)$$

$$= 0.5 - 0.4608$$

$$= 0.03982$$

ii) Fall between 50.5 and 52.3

$$P(50.5 < \bar{x} < 52.3) = P(\bar{x}_1 < \bar{x} < \bar{x}_2)$$

$$\bar{x}_1 = 50.5 \quad \text{and} \quad \bar{x}_2 = 52.3$$

$$Z_1 = \frac{\bar{x}_1 - \mu}{\sigma_{\bar{x}}} = \frac{50.5 - 51.4}{0.85} = \frac{-0.9}{0.85} = -1.06$$

$$Z_2 = \frac{\bar{x}_2 - \mu}{\sigma_{\bar{x}}} = \frac{52.3 - 51.4}{0.85} = \frac{0.9}{0.85} = 1.06$$

$$P(50.5 < \bar{x} < 52.3 = P(-1.06 < Z < 1.06)$$

$$= P(-1.06 < Z < 0) + P(0 < Z < 1.06)$$

$$= 0.3554 + 0.3554 = 0.7108$$

iii) BC less than 50.6

$$P(\bar{x} < 50.6)$$

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{50.6 - 51 - 4}{0.85} = -0.94$$

$$P(\bar{x} < 50.6) = P(Z < -0.94)$$

$$= 0.5 - P(0.94 < Z < 0)$$

$$= 0.5 - 0.3264 = 0.1736$$

2.8.7 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 $\bar{X}$ 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 2.8.7 *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 $\sigma^2 = 225$

$$\sigma = \sqrt{225} = 15$$

$$\text{Mean } \mu = 75$$

$$\bar{x}_1 = 72, \quad \bar{x}_2 = 77$$

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$Z_1 = \frac{\bar{x}_1 - m}{\sigma / \sqrt{n}} = \frac{72-75}{15/\sqrt{144}} = \frac{-3}{15/12} = -2.4$$

$$Z_2 = \frac{\bar{x}_2 - \mu}{\sigma / \sqrt{n}} = \frac{77-75}{15/\sqrt{144}} = \frac{2}{15/12} = 1.6$$

$$P(72 < \bar{x} < 77) = P(\bar{x}_1 < \bar{x} < \bar{x}_2)$$

$$= P(-2.4 < Z < 0) + P(0 < Z < 1.06)$$

$$= P(0 < Z < 2.4) + P(0 < Z < 1.6)$$

$$= 0.4918 + 0.4452 = 0.9370$$

Example 2.8.8 *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 $\mu = 76$

Variance $\sigma^2 = 256$

$\sigma = 16$

$n = 100, \quad \bar{x}_1 = 75, \quad \bar{x}_2 = 78$

$$Z_1 = \frac{\bar{x}_1 - \mu}{\sigma / \sqrt{n}} = \frac{75-76}{16/\sqrt{100}} = \frac{-1}{16/10} = -0.625$$

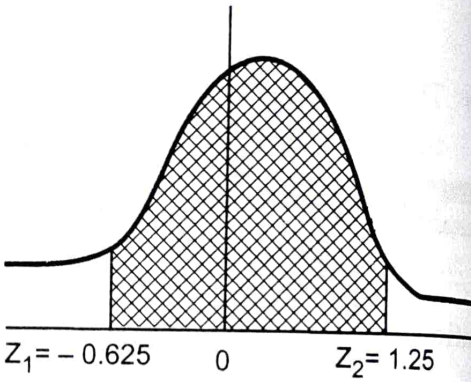
$$Z_2 = \frac{\bar{x}_2 - \mu}{\sigma / \sqrt{n}} = \frac{78-76}{16/\sqrt{100}} = \frac{2}{16/10} = 1.25$$

$$P(75 < \bar{x} < 78) = P(-0.625 < Z < 1.25)$$

$$= P(-0.625 < Z < 0) + P(0 < Z < 1.25)$$

$$= P(0 < Z < 0.625) + P(0 < Z < 1.25)$$

$$= 0.2324 + 0.3944 = 0.6268$$



Example 2.8.9 *When a sample is taken from an infinite population, what happen to th 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

$n_1 = 800$ 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 2.8.10 *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)

$N^n = (4)^2 = 16$ (sample size = 2)

N = Population size

n = 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

$$\sigma^2_{\bar{x}} = \frac{1}{16}\begin{bmatrix} (2-3.5)^2+(2.5-3.5)^2+(3-3.5)^2+(3.5-3.5)^2 \\ +(2.5-3.5)^2+(3-3.5)^2+(3.5-3.5)^2+(4-3.5)^2 \\ +(3-3.5)^2+(3.5-3.5)^2+(4-3.5)^2+(4.5-3.5)^2 \\ +(3.5-3.5)^2+(4-3.5)^2+(4.5-3.5)^2+(5-3.5)^2 \end{bmatrix}$$

$$= \frac{\begin{bmatrix} 2.25+1+0.25+0+1+0.25+0+0.25+0.25 \\ +0.25+1+0+0.25+1+2.25 \end{bmatrix}}{16}$$

$$= \frac{10}{16} = 0.625 = \sqrt{0.625} = 0.79$$

e) s.d. of SD mean

$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} = \frac{(1.118)^2}{2} = 0.6249$$

Example 2.8.11 *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 $\mu_{\bar{X}} = 14$

So, considering $\mu_{\bar{X}} = \mu$

iv) Standard deviation of the sampling distribution of means

$$\sigma^2_{\bar{x}} = \frac{1}{15}[(6-14)^2 + (8-14)^2 + (10-14)^2 + (12-14)^2 + (14-14)^2 + (10-14)^2 + (12-14)^2 + (14-14)^2 + (16-14)^2 + (14-14)^2 + (16-14)^2 + (18-14)^2 + (18-14)^2 + (20-14)^2 + (22-14)^2]$$

$$= \frac{1}{15}[64+36+16+4+0+16+4+0+4+0+4+16+16+36+64]$$

$$= \frac{280}{15} = 18.66$$

Standard deviation of sampling distribution of means is

$$\sigma_{\bar{x}} = \sqrt{18.66} = 4.319$$

Example 2.8.12 *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 ($\mu_{\bar{x}} = \mu$)

$$\sigma^2_{\bar{x}} = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$\sigma^2_{\bar{x}} = \frac{1}{15}\begin{bmatrix} (7.5-14)^2 + (9.5-14)^2 + (11.5-14)^2 + (9-14)^2 \\ + (14.5-14)^2 + (12-14)^2 + (14-14)^2 + (11.5-14)^2 \\ + (17-14)^2 + (16-14)^2 + (13.5-14)^2 + (19-14)^2 \\ + (15.5-14)^2 + (21-14)^2 + (18.5-14)^2 \end{bmatrix}$$

$$= \frac{1}{15}\begin{bmatrix} 42.25+20.25+6.25+25+0.25+4+0+6.25 \\ + 9+4+0.25+25+2.25+49+20.25 \end{bmatrix} = \frac{214}{15}$$

$$\sigma^2_{\bar{x}} = 14.2666$$

$$\sigma_{\bar{x}} = 3.777$$

2.9 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.

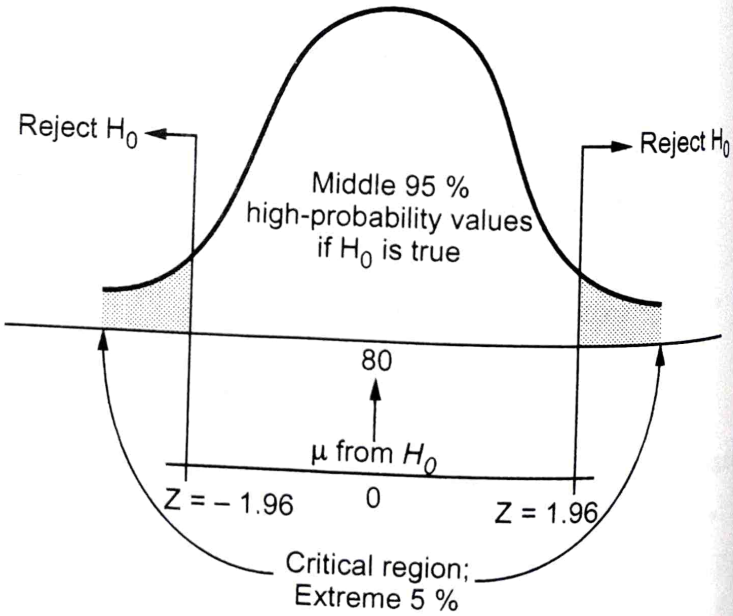


Fig. 2.9.1

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.

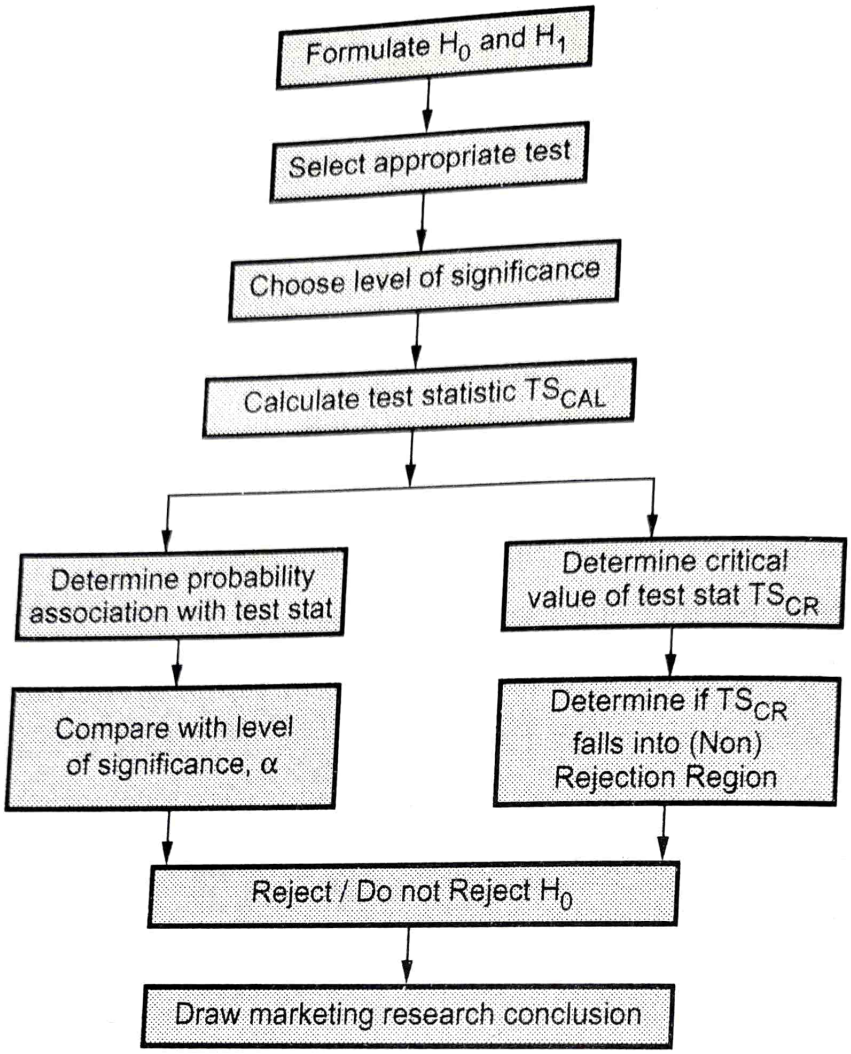


Fig. 2.9.2

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. 2.9.3.
2) For normal population, acceptance and rejection regions are shown in Fig. 2.9.4.

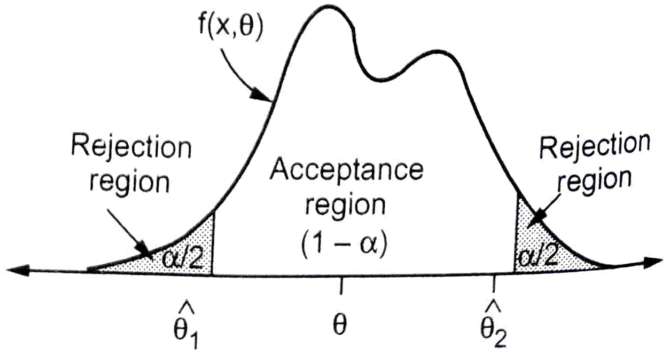


Fig. 2.9.3 Arbitrary population

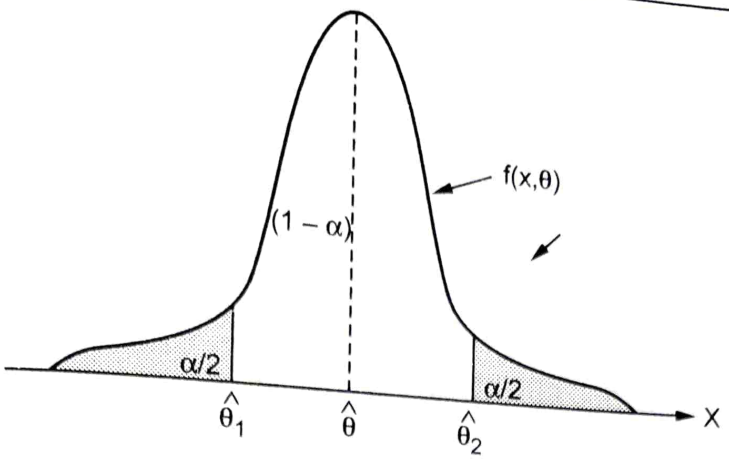


Fig. 2.9.4 Normal population

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.

2.9.1 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 : $>, <, \neq$
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.

2.10 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

2.11 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.

3 Bayesian Concept Learning

Syllabus

Importance of Bayesian methods, Bayesian theorem, Bayes' theorem and concept learning, Bayesian Belief Network

Contents

3.1 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.

3.2 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 $B_1, B_2, B_3 \ldots B_n$ partition the outcomes of an experiment and that A is another event. For any number, k, with $1 \leq k \leq n$, 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 3.2.1 *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

D = {bolt is defective},

A = {bolt is from machine A},

B = {bolt is from machine B},

C = {bolt is from machine C}.

Given data : $P(A) = 0.25$, $P(B) = 0.35$, $P(C) = 0.4$.

$P(D \mid A) = 0.05$, $P(D \mid B) = 0.04$, $P(D \mid C) = 0.02$.

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 3.2.2 *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 :

$$P(F/T) = \frac{P(T/F)\,P(F)}{P(T/F)\,P(F) + P(T/M)\,P(M)} = \frac{\frac{1}{100} \times \frac{3}{5}}{\frac{1}{100} \times \frac{3}{5} + \frac{4}{100} \times \frac{2}{5}} = \frac{\frac{3}{500}}{\frac{3}{500} + \frac{8}{500}}$$

$$P(F/T) = \frac{3}{11}$$

3.2.1 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

3.2.2 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 $X_1 = X_1, x_2 = x_2, \cdots, X_n = x_n$,

  $$lik(\theta) = f(x_1, x_2, \cdots, x_n|\theta)$$

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 $lik(\theta)$: 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

$$P(X = x) = \theta^x (1-\theta)^{1-x}$$

3.3 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.

3.3.1 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

3.3.2 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.

3.3.3 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

3.4 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. 3.4.1 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.

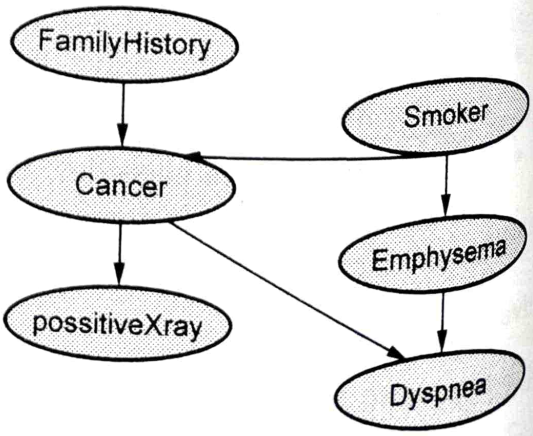


Fig. 3.4.1 simple belief network

- 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 :

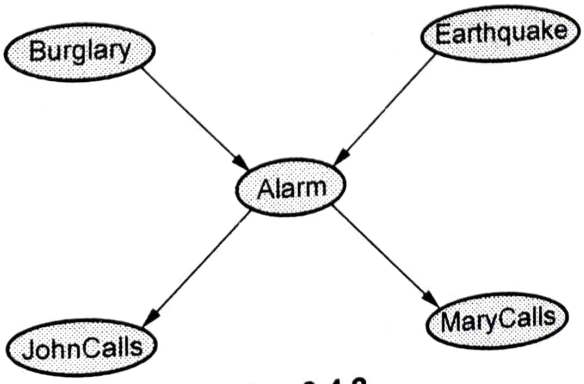


Fig. 3.4.2

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.

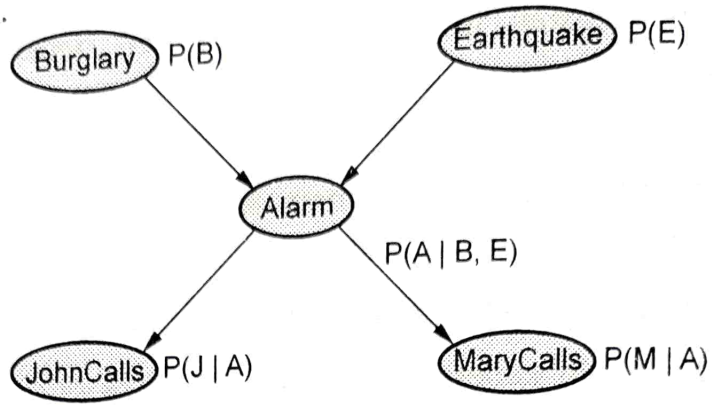


Fig. 3.4.3

Local conditional distributions : Relate variables and their parents.

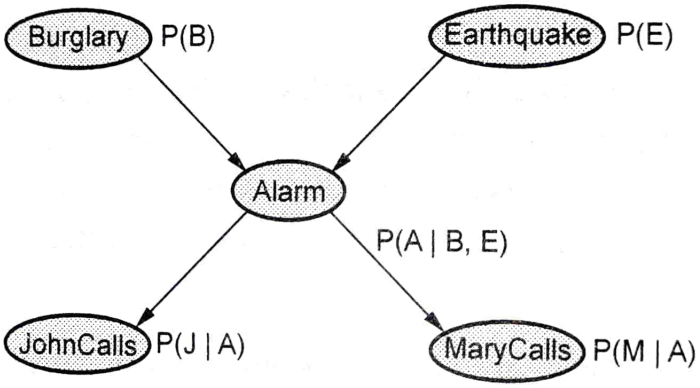


Fig. 3.4.4

Bayesian belief network

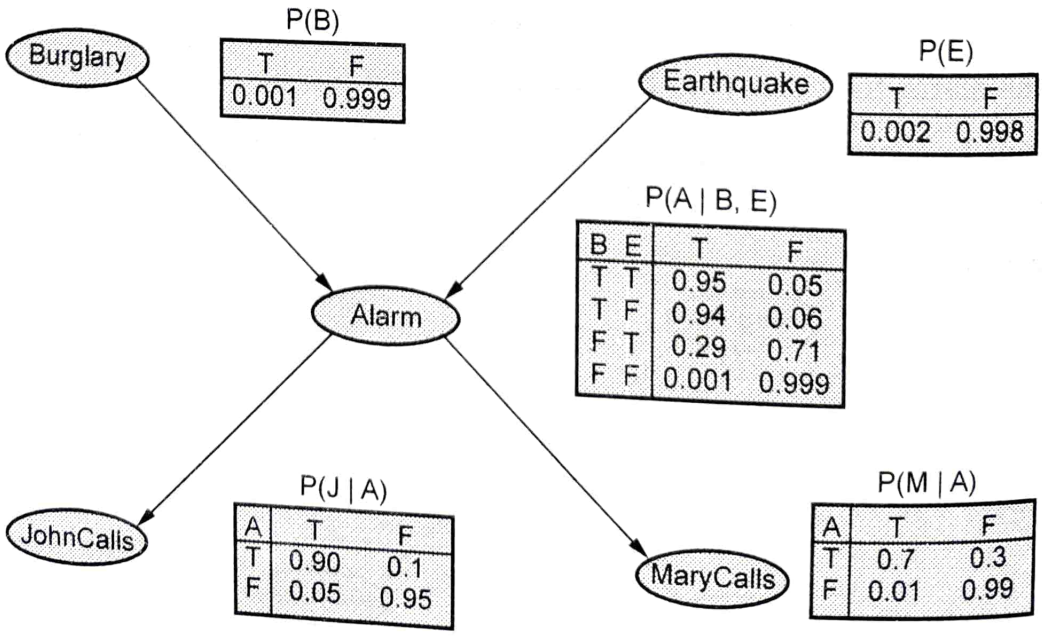


Fig. 3.4.5

3.5 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

❑❑❑

4 Classification and Regression

Syllabus

Supervised Learning vs Unsupervised Learning, Supervised Learning, Classification Model, Learning steps, Classification algorithms, Clustering, Association rules, Linear Regression, Multivariate Regression, Logistic Regression

Contents

4.1 Supervised Learning vs 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.

4.2 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.

4.3 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.

4.4 Learning Steps

- Fig. 4.4.1 (See Fig. 4.4.1 on next page) shows classification steps.
  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

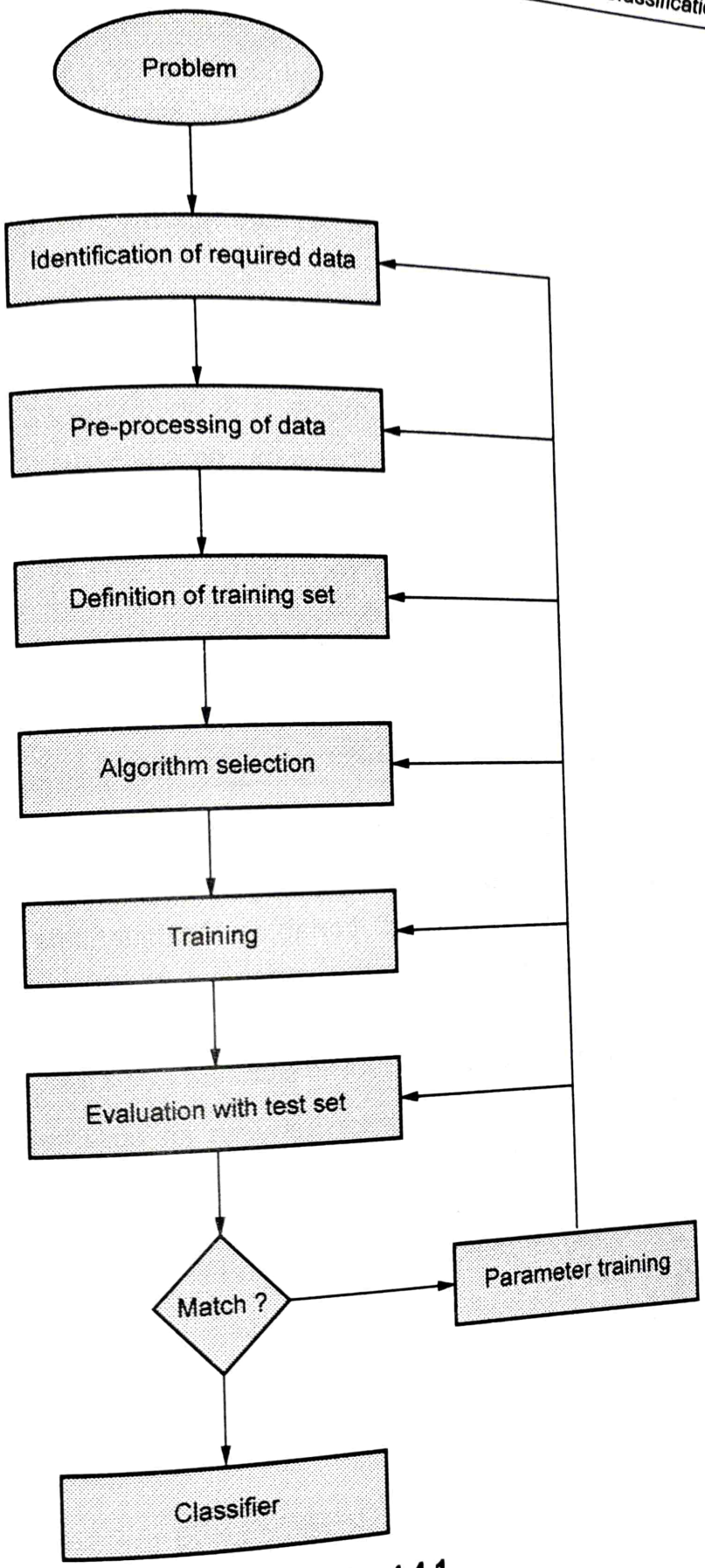

Problem
Identification of required data
Pre-processing of data
Definition of training set
Algorithm selection
Training
Evaluation with test set
Match ?
Parameter training
Classifier

Fig. 4.4.1

4.5 Classification Algorithms

4.5.1 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

$$d_{Mahalanobis}(x, y) = \sqrt{(x-y)^T \sum^{-1} (x-y)}$$

- 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.

4.5.2 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. 4.5.1 shows an example of a decision tree to determine what kind of contact lens a person way wear.
- 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.

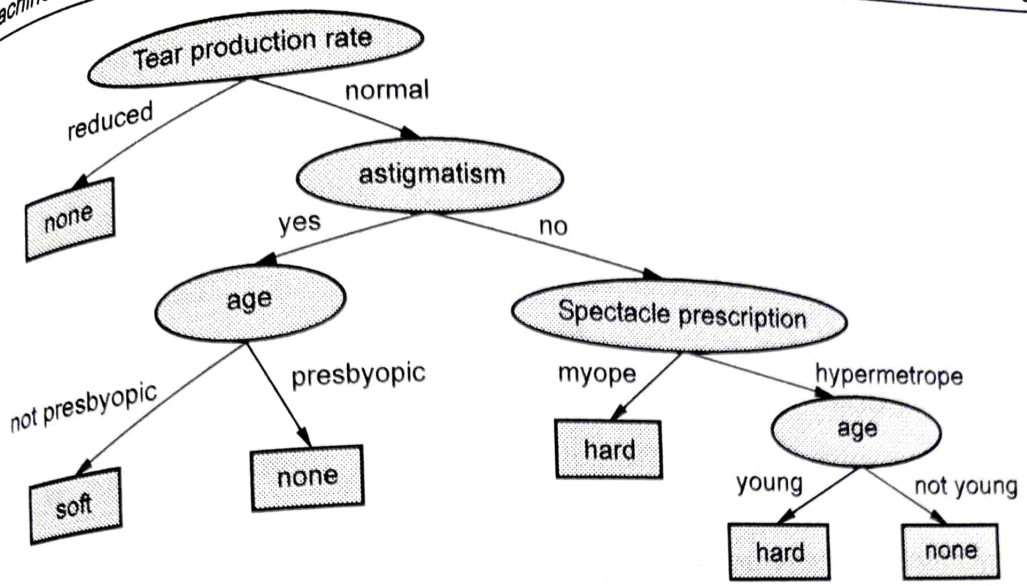


Fig. 4.5.1

- Gini Index for a given node t :

$$GINI(t) = \sum p(j \mid t)(1 - p(j \mid t)) - \sum p(j \mid t)^2$$

Maximum number of classes when records are equally distributed among all classes s 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 :

$$GIN_{Isplit} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i) = \sum_j p(j \mid t)^2$$

where n_i = Number of records at child i and n = Number of records at node P.

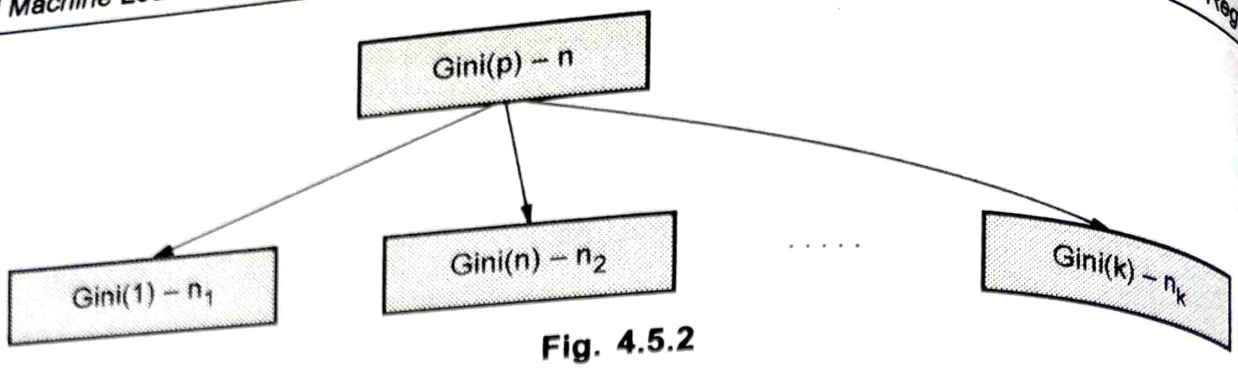


Fig. 4.5.2

4.5.2.1 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.

4.5.2.2 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 :

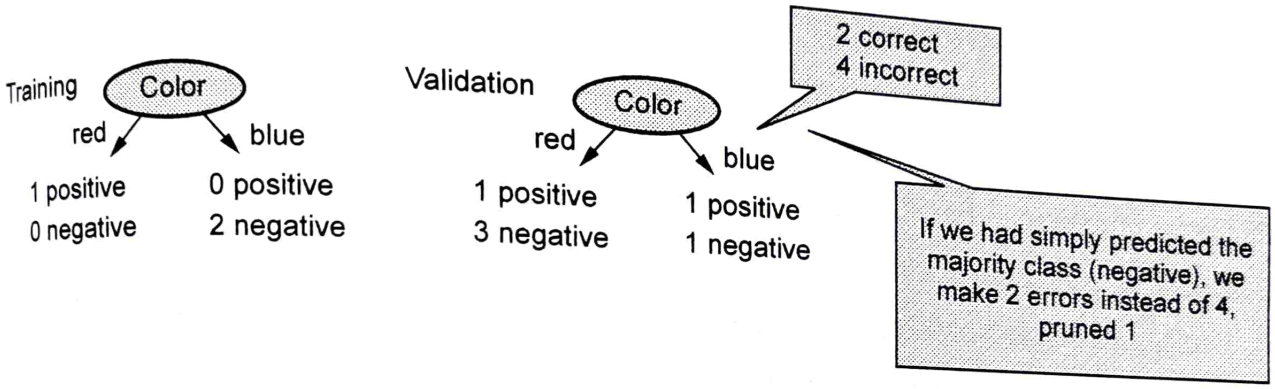


Fig. 4.5.3

4.5.2.3 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 no N
13. End of for loop
14. return N

4.5.2.4 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 4.5.1 *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 4.5.2 *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 :

$$\text{Entropy(S)} = -p_{\text{cinema}}\log_2(p_{\text{cinema}}) - p_{\text{tennis}}\log_2(p_{\text{tennis}}) - p_{\text{shopping}}\log_2(p_{\text{shopping}}) - p_{\text{stay\_in}}\log_2(p_{\text{stay\_in}})$$

$$= -(6/10)\times\log_2(6/10) - (2/10)\times\log_2(2/10) - (1/10)\times\log_2(1/10) - (1/10)\times\log_2(1/10)$$

$$= -(6/10)\times -0.737 - (2/10)\times -2.322 - (1/10)\times -3.322 - (1/10)\times -3.322$$

$$= 04422 + 0.4644 + 0.3322 + 0.3322 = 1.571$$

and we need to determine the best of :

$$\text{Gain(S, weather)} = 1.571 - (\text{IS}_{\text{sun}}\ \text{I}/10)\times\text{Entropy}(S_{\text{sun}}) - (\text{IS}_{\text{wind}}\ \text{I}/10)\times\text{Entropy}(S_{\text{wind}}) - (\text{IS}_{\text{rain}}\ \text{I}/10)\times\text{Entropy}(S_{\text{rain}})$$

$$= 1.571 - (0.3)\times\text{Entropy}(S_{\text{sun}}) - (0.4)\times\text{Entropy}(S_{\text{wind}}) - (0.3)\times\text{Entropy}(S_{\text{rain}})$$

$$= 1.571 - (0.3)\times(0.918) - (0.4)\times(0.81125) - (0.3)\times(0.918) = 0.70$$

$$\text{Gain(S, parents)} = 1.571 - (\text{IS}_{\text{yes}}\ \text{I}/10) \times \text{Entropy}(\text{S}_{\text{yes}}) - (\text{IS}_{\text{no}}\ \text{I}/10) \times \text{Entropy}(\text{S}_{\text{no}})$$
$$= 1.571 - (0.5) \times 0 - (0.5) \times 1.922 = 1.571 - 0.961 = 0.61$$

$$\text{Gain(S, money)} = 1.571 - (\text{IS}_{\text{rich}}\ \text{I}/10) \times \text{Entropy}(\text{S}_{\text{rich}}) - (\text{IS}_{\text{poor}}\ \text{I}/10) \times \text{Entropy}(\text{S}_{\text{poor}})$$
$$= 1.571 - (0.7) \times (1.842) - (0.3) \times 0 = 1.571 - 1.2894 = 0.2816$$

- 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 :

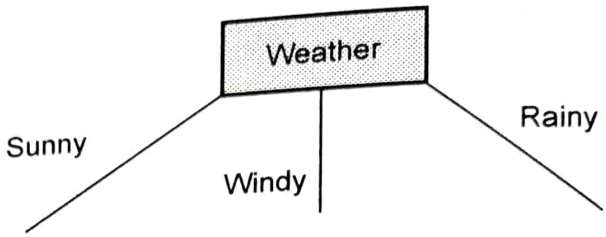


Fig. 4.5.4

- 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 :

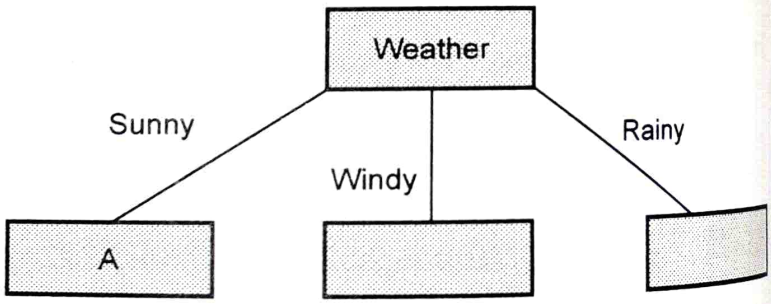


Fig. 4.5.5

- 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 :

$$Gain(S_{sunny}, parents) = 0.918 - (IS_{yes}\ I/ISI) \times Entropy(S_{yes}) - (IS_{no}\ I/ISI) \times Entropy(S_{no})$$
$$= 0.918 - (1/3) \times 0 - (2/3) \times 0 = 0.918$$

$$Gain(S_{sunny}, money) = 0.918 - (IS_{rich}\ I/ISI) \times Entropy(S_{rich}) - (IS_{poor}\ I/ISI) \times Entropy(S_{poor})$$
$$= 0.918 - (3/3) \times 0.918 - (0/3) \times 0 = 0.918 - 0.918 = 0$$

4.5.3 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. 4.5.6 shows empirical risk.

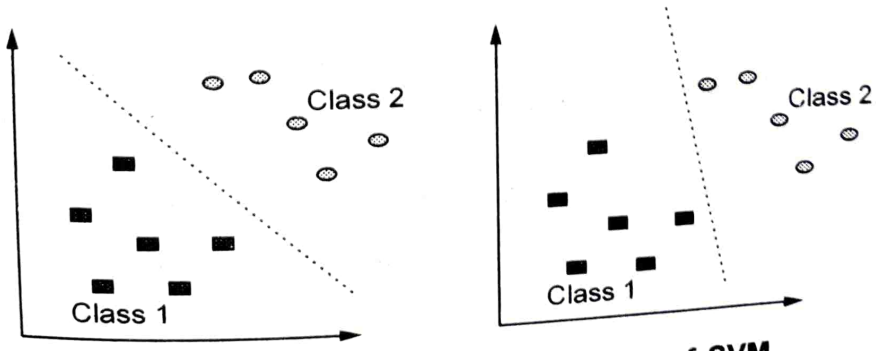


Fig. 4.5.6 Bab decision boundary of SVM

- 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.**

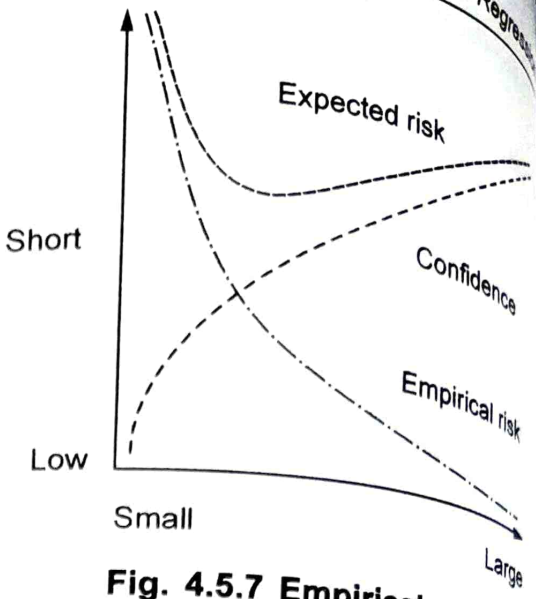


Fig. 4.5.7 Empirical risk

Applications :

- SVM has been used successfully in many real - world problems
  1. Text (and hypertext) categorization.
  2. Image classification.
  3. Bioinformatics (Protein classification, Cancer classification).
  4. Hand - written character recognition.
  5. Determination of SPAM email.

Limitations :

1. It is sensitive to noise.
2. The biggest limitation of SVM lies in the choice of the kernel.
3. Another limitation is speed and size.
4. The optimal design for multiclass SVM classifiers is also a research area.

4.6 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 :**

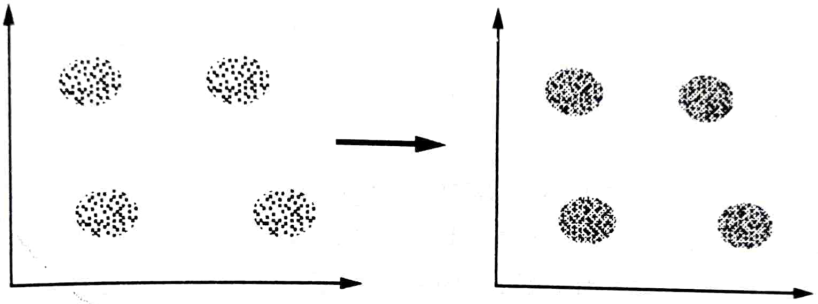


Fig. 4.6.1

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. 4.6.1 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.**

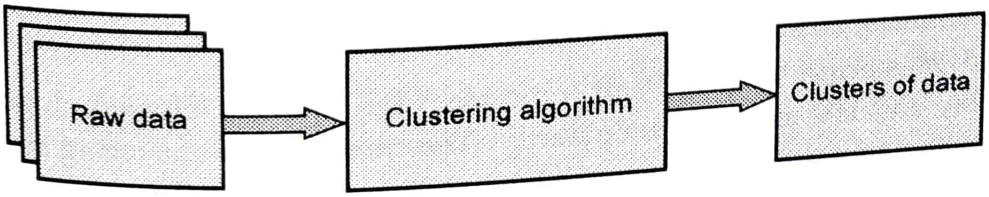


Fig. 4.6.2

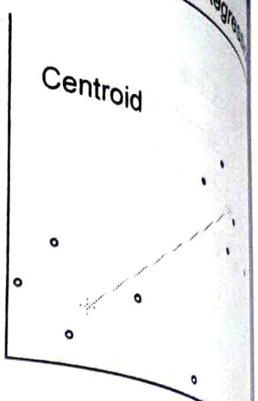


Fig. 4.6.3

- 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 basically a statistical description of the cluster centroids with the number components in each cluster.
- **Cluster centroid** : The centroid of a cluster is a point whose parameter values the mean of the parameter values of all the points in the cluster. Each cluster a well defined centroid.
- **Distance :** The distance between two points is taken as a common metric to as the similarity among the components of population. The commonly used dista measure is the euclidean metric which defines the distance between two po $p = (p_1, p_2, \ldots)$ and $q = (q_1, q_2, \ldots)$ 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 unlabele data. But how to decide what constitutes a good clustering ? It can be shown th 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 suc 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 patter 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-clas similarlity and low inter-class similarity. The quality of a clustering result depend on both the similarity measure used by the method and its implementation. Th

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.

4.6.1 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.

4.6.1.1 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 $x_1, \ldots, x_N$ 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*.

4.6.1.2 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'

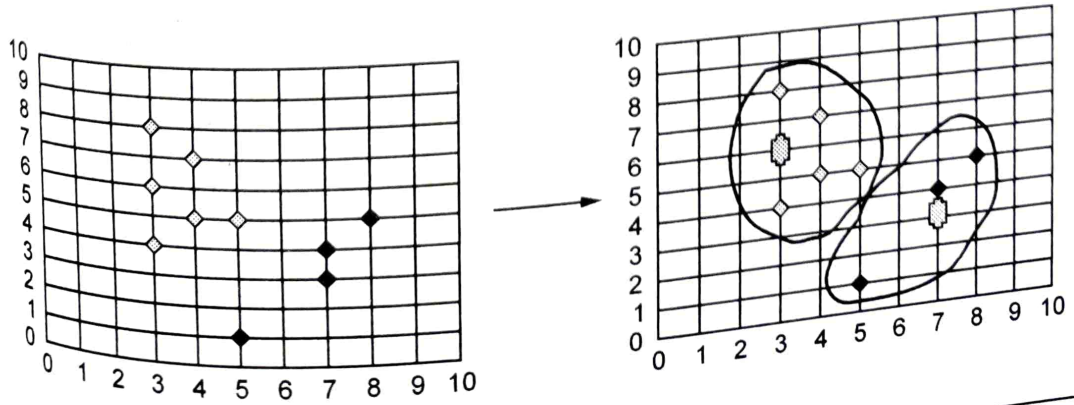


Fig. 4.6.4

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.

4.6.2 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

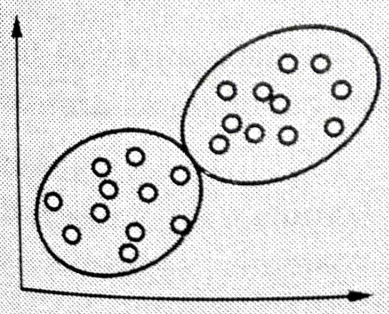
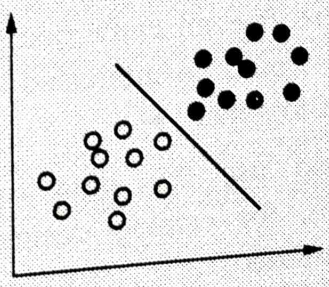
Divisive Hierarchical Clustering

- 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.

6.2.1 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.
	

4.7 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

4.7.1 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.

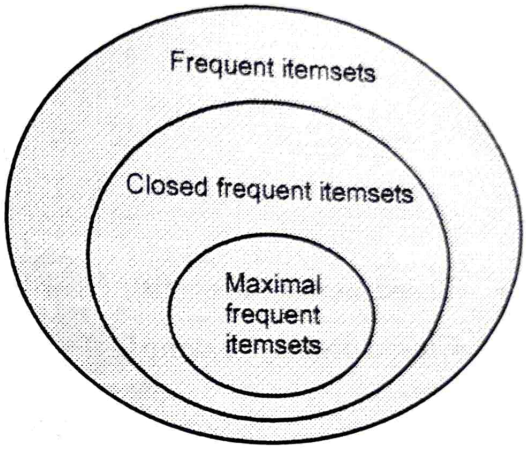


Fig. 4.7.1

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}

4.7.2 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 $I = \{i_1, i_2, \ldots i_n\}$ be a set of n binary attributes called items. Let $D = \{t_1, t_2, \ldots, t_n\}$ 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 4.7.1 *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 :

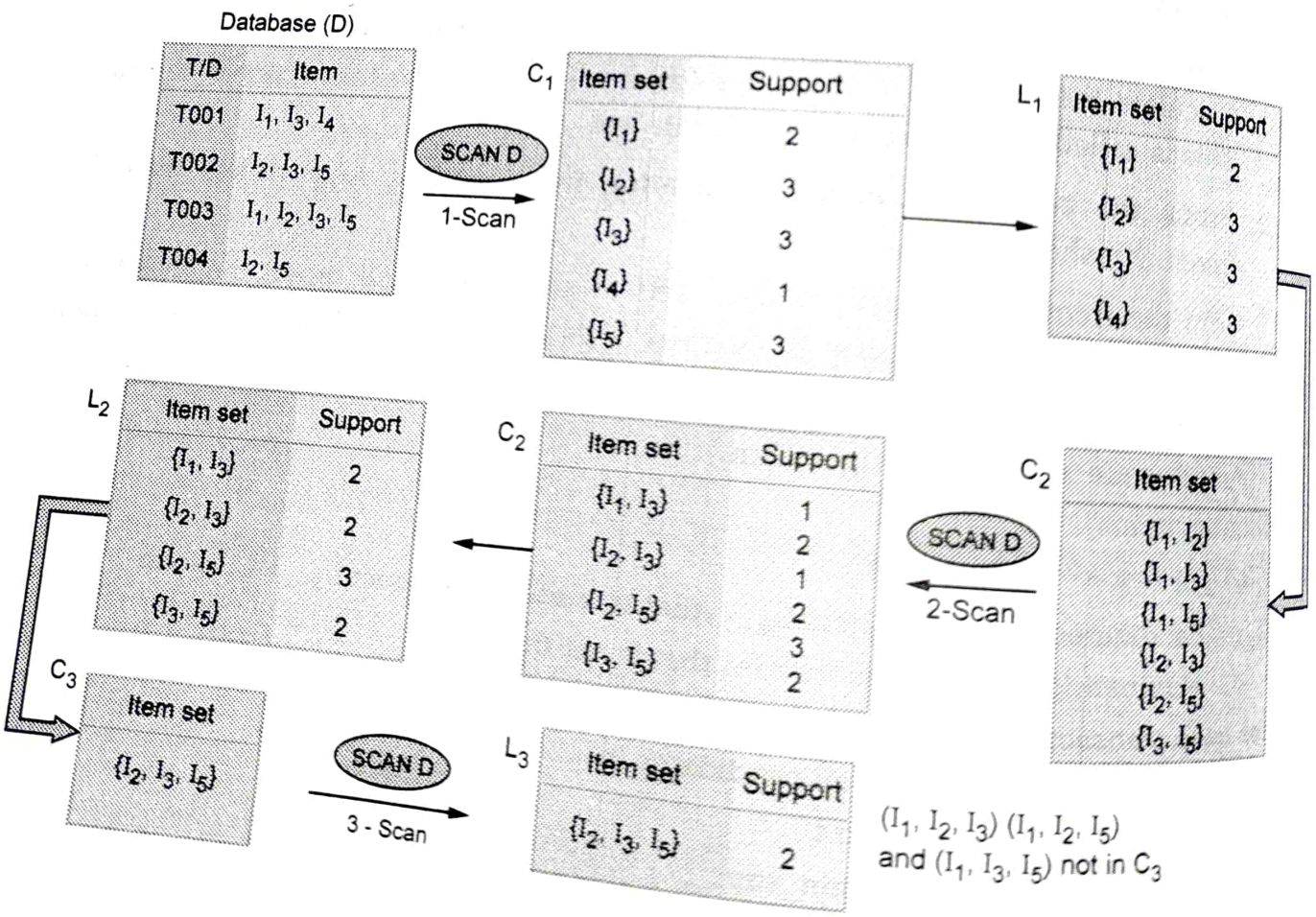


Fig. 4.7.2

4.8 Linear Regression

- The most common regression algorithms are,
  a) Simple linear regression
  b) Multiple linear regression
  c) Polynomial regression
  d) Multivariate adaptive regression splines
  e) Logistic regression
  f) Maximum likelihood estimation (Least squares)

4.8.1 Simple Linear Regression

- Regression model which involves only one predictor. Linear regression is a statistical method that allows us to summarize and study relationships between two continuous variables :
  1. One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.
  2. The other variable, denoted y, is regarded as the response, outcome, or dependent variable.
- Regression models predict a continuous variable, such as the sales made on a day or predict temperature of a city.
- Let's imagine that you fit a line with the training points you have. Imagine you want to add another data point, but to fit it, you need to change your existing model.
- This will happen with each data point that we add to the model; hence, linear regression isn't good for classification models.
- Regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.
- **Regression line of X on Y** gives the best estimate for the value of X for any specific given values of Y :

$$X = a + b\,Y$$

Where

a = X - intercept

b = Slope of the line

X = Dependent variable

Y = Independent variable

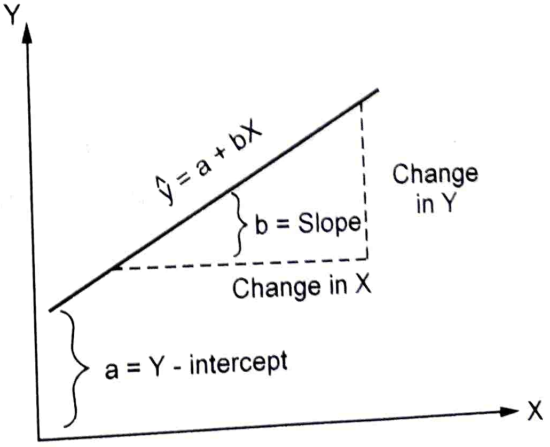


Fig. 4.8.1

- 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 :

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

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

- 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.

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

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.

4.8.2 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).

4.8.3 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. 4.8.2 shows Sigmoid curve for logistic regression.
- 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

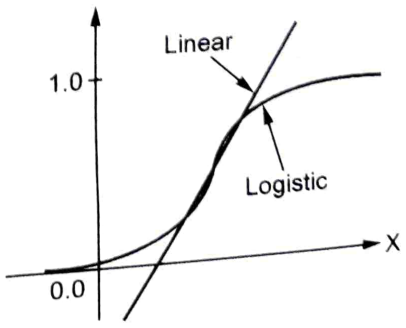


Fig. 4.8.2

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.

4.8.4 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.

$$E[\hat{\beta}^{ridge}|X] = (X^T X+\lambda I)^{-1} X\beta$$

$$= (X^T X+\lambda I)^{-1}(X^T X+\lambda I-\lambda I)\beta$$

$$= [I-\lambda(X^T X+\lambda I)^{-1}]\beta$$

$$= \beta-\lambda(X^T X+\lambda I)^{-1}\beta$$

- 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.

5 Neural Networks

Syllabus

Introduction, Early Models, Perceptron Learning, Backpropagation, Initialization, Training & Validation, Parameter Estimation - MLE, MAP, Bayesian Estimation.

Contents

5.1 Introduction

- 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.

5.1.1 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.

5.1.2 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.

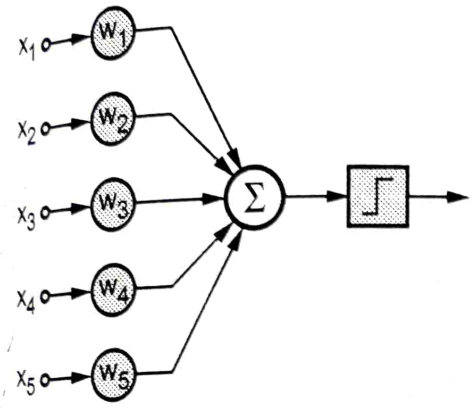
5.1.3 Difference between Digital Computer and Neural Networks

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.

5.2 Perceptron Learning

- 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 :



$$h(x) = \begin{cases} 1 \text{ if } & w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_d \cdot x_d \geq \theta \\ 0 \text{ if } & w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_d \cdot x_d < \theta \end{cases}$$

The equation can be re-written as follows including what its known as the bias term : $x_0 = 1$, $w_0 = \theta$.

$$h(x) = \begin{cases} 1 \text{ if } & w_0 \cdot x_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_d \cdot x_d \geq 0 \\ 0 \text{ if } & w_0 \cdot x_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_d \cdot x_d < 0 \end{cases}$$

$$h(x) = \begin{cases} 1 \text{ if } & w^t \cdot x \geq 0 \\ 0 \text{ if } & w^t \cdot x < 0 \end{cases}$$

5.2.1 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. 5.2.1 shows biological neural systems.

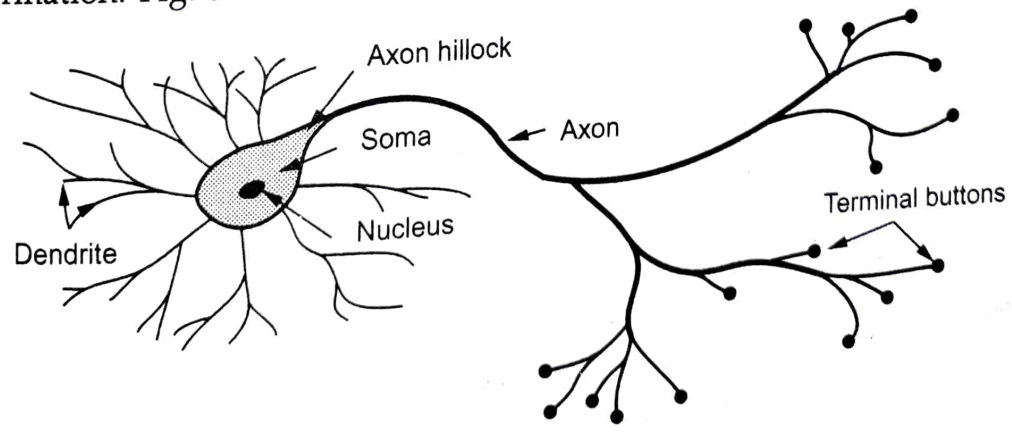


Fig. 5.2.1 Schematic of biological neuron

- 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

5.2.2 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. 5.2.2 shows adaline.

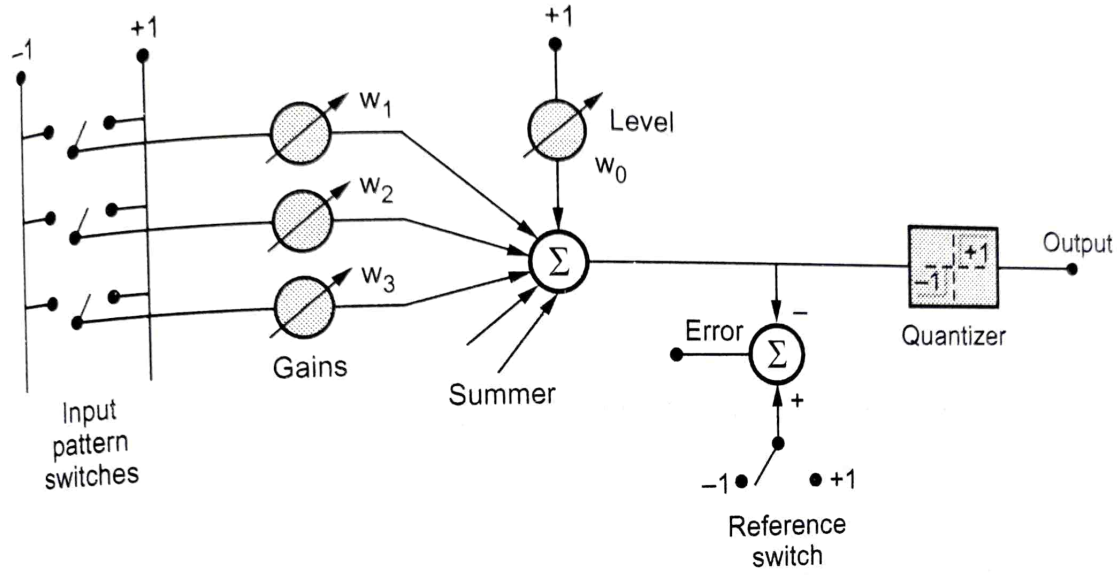


Fig. 5.2.2 Adaline

- 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, $\theta \equiv w_0$

- 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 p^{th} 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 = Target output

o_p = 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.

5.2.3 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. 5.2.3 shows McCulloch Pitt model.

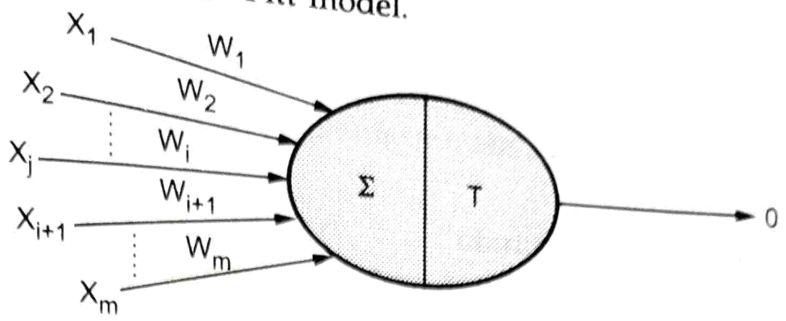


Fig. 5.2.3

- 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.

$$0 = f\left(\sum_{j=1}^{N} W_j X_j - T\right)$$

$$f(u) = \begin{cases} 1 & \text{if } u \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

- 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.

5.3 Architecture of Neural Network

5.3.1 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.

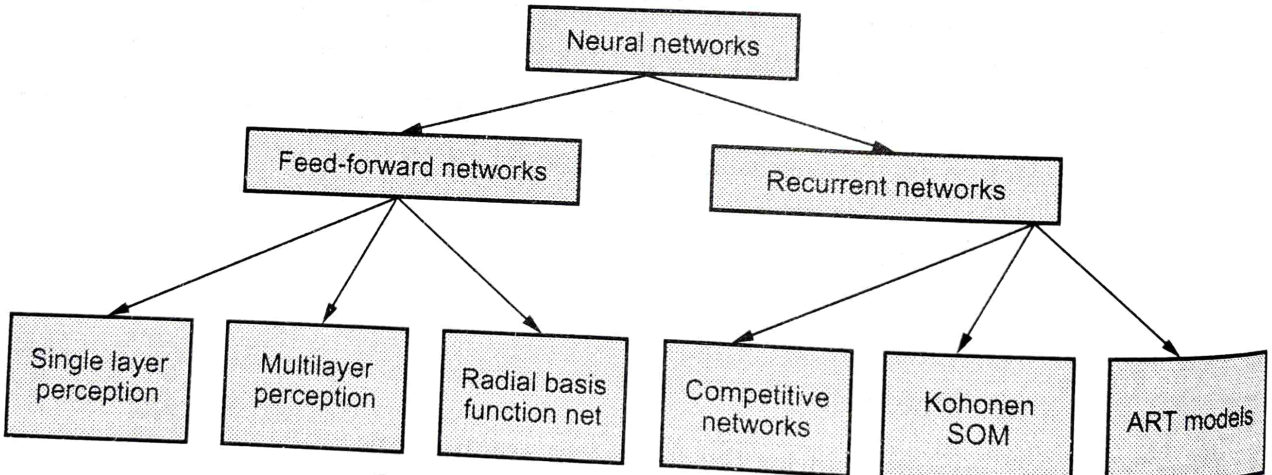


Fig. 5.3.1

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. 5.3.2 shows a typical perception setup for pattern recognition applications, in which visual patterns are represented as matrices of elements between 0 and 1.

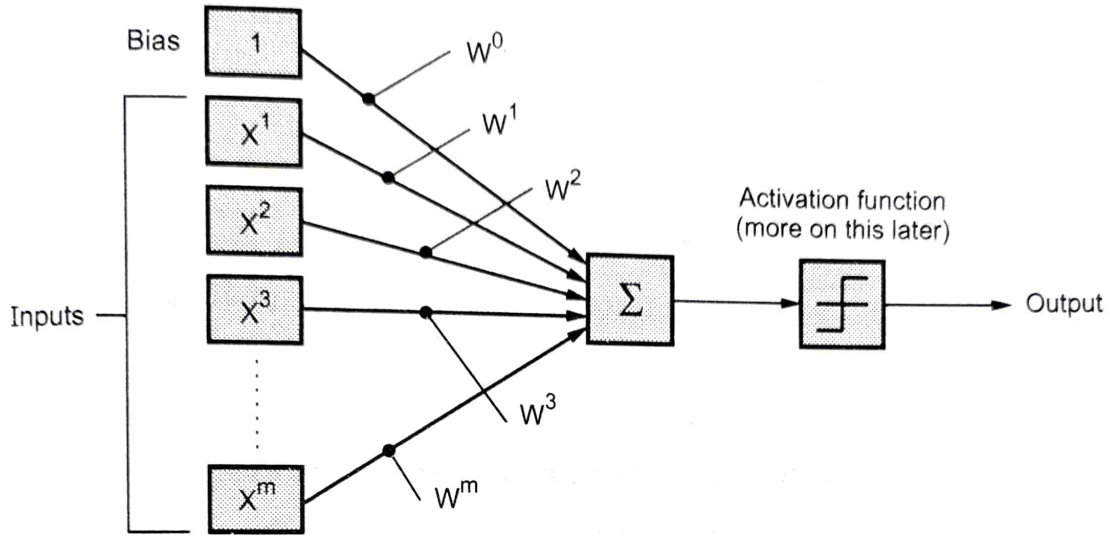


Fig. 5.3.2 Perception setup

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. 5.3.3 shows the bias term W_0.

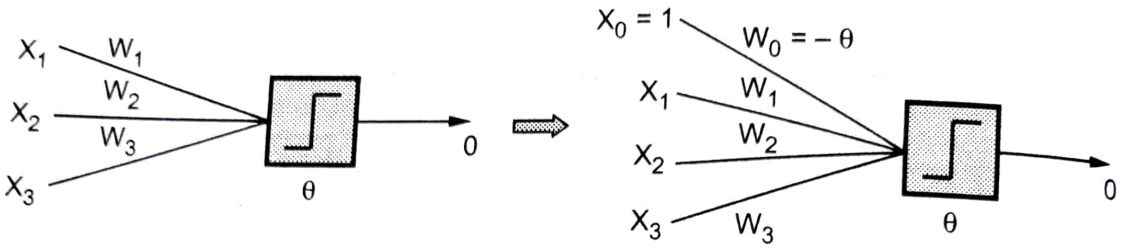


Fig. 5.3.3 Bias term W_0

- The function y = f(x) describes relationship, an input-output mapping from x to y.
- The equation (5.3.1), the f(.) is the **activation function** of the perceptron and it is typically either a **signum function** sgn(x) or **step function** step(x) :

$$\text{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{otherwise} \end{cases}$$

$$\text{step}(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0 & \text{otherwise} \end{cases} \quad \ldots (5.3.1)$$

- 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 :

$0 \times w_1 + 0 \times w_2 + w_0 \leq 0 \Leftrightarrow w_0 \leq 0,$

$0 \times w_1 + 1 \times w_2 + w_0\ 0 \Leftrightarrow w_0 > - w_2$

$1 \times w_1 + 0 \times w_2 + w_0 > 0 \Leftrightarrow w_0 > - w_1$

$1 \times w_1 + 1 \times w_2 + w_0 \leq 0 \Leftrightarrow w_0 \leq - w_1 - w_2$

5.3.2 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

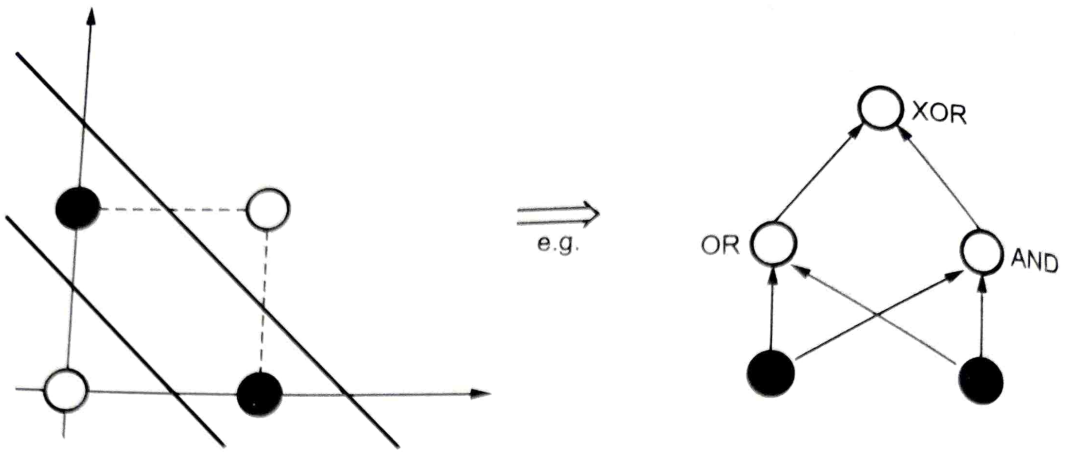


Fig. 5.3.4

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. 5.3.5 shows three most commonly used activation functions in backpropagation MLPs.

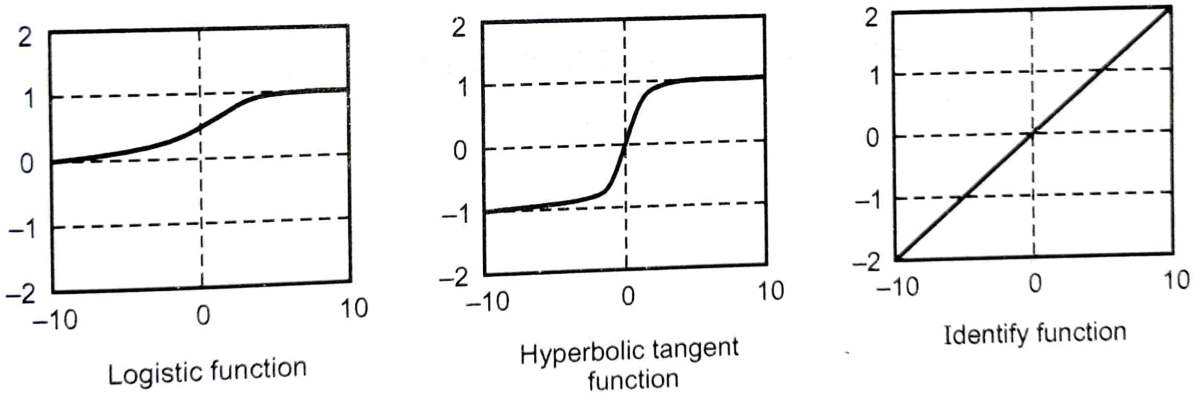


Fig. 5.3.5 Activation function

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. 5.3.6 shows the backpropagation MLP for node j. The net input and output of node j is as follows :

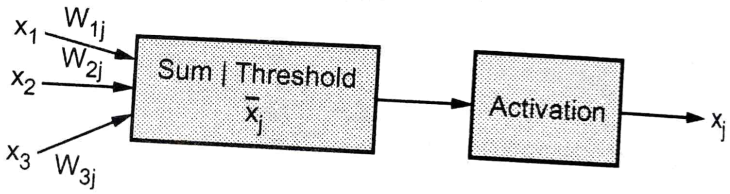


Fig. 5.3.6 Backpropagation MLP for node j

$$\bar{X}_j = \sum_i + W_{ij} + W_j$$

$$x_j = f(\bar{X}_j) = \frac{1}{1 + \exp(-\bar{X}_j)}$$

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. 5.3.7 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.

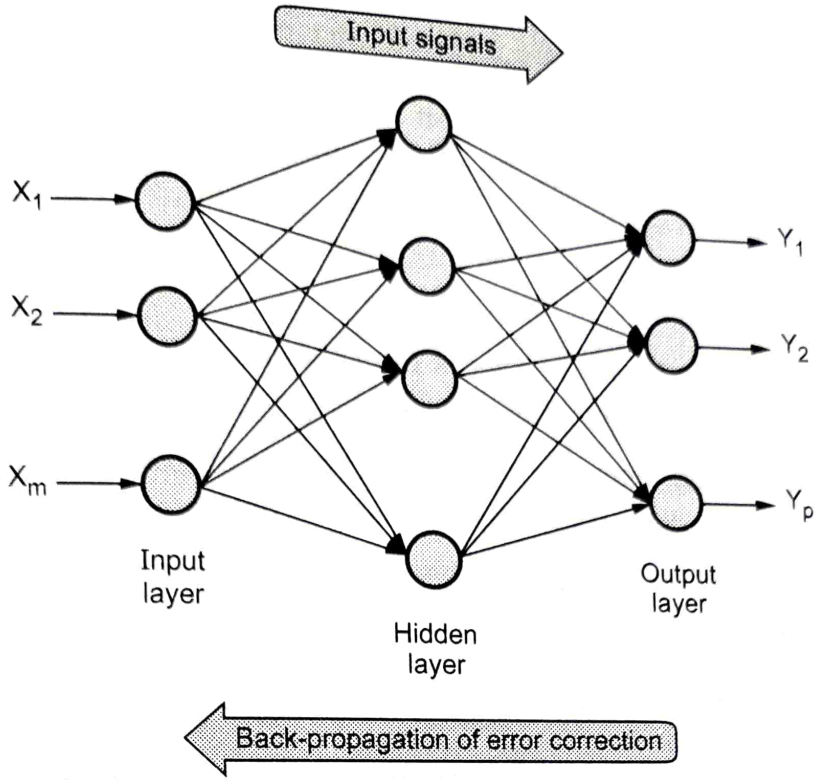


Fig. 5.3.7 Two layer backpropagation MLP

- 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.

5.3.3 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. 5.3.8 shows architecture of recurrent neural network.

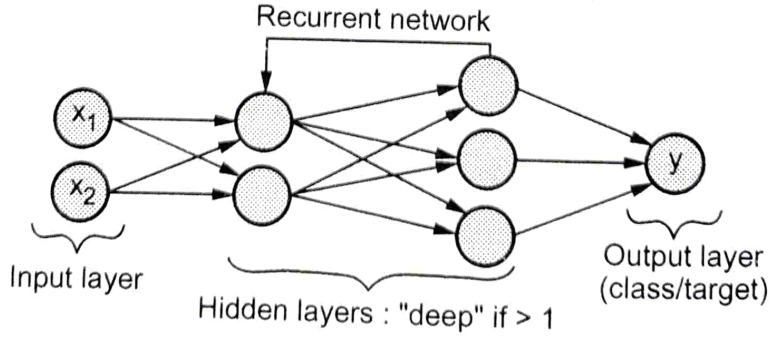


Fig. 5.3.8

- 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.

5.4 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. 5.4.1 shows backpropagation network.

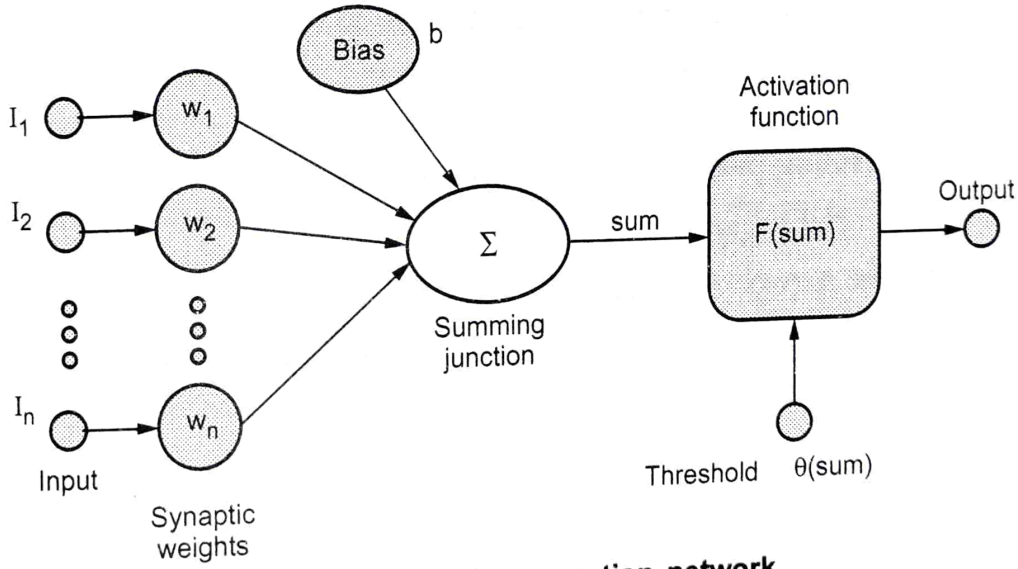


Fig. 5.4.1 Backpropagation network

- 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.

5.4.1 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.

5.5 Parameter Estimation

- Parameter estimation refers to the process of using sample data to estimate the parameters of the selected distribution. Several parameter estimation methods are available.
- MLE, MAP and Bayesian inference are methods to deduce properties of a probability distribution behind observed data.

5.5.1 MAP

- Maximum A Posteriori (MAP) estimation tries to find the estimate of the parameter θ by maximizing the posterior distribution. MAP selects a single most likely hypothesis given the data.
- The posterior distribution, $f_{x|y}(x|y)$ contains all the knowledge about the unknown quantity X. Therefore, we can use the posterior distribution to find point or interval estimates of X. One way to obtain a point estimate is to choose the value of x that maximizes the posterior PDF or PMF. This is called the Maximum a Posteriori (MAP) estimation.
- A MAP estimator finds the peak, or mode, of a posterior density.
- Fig. 5.5.1 shows the maximum a posteriori estimate of X given Y = y.

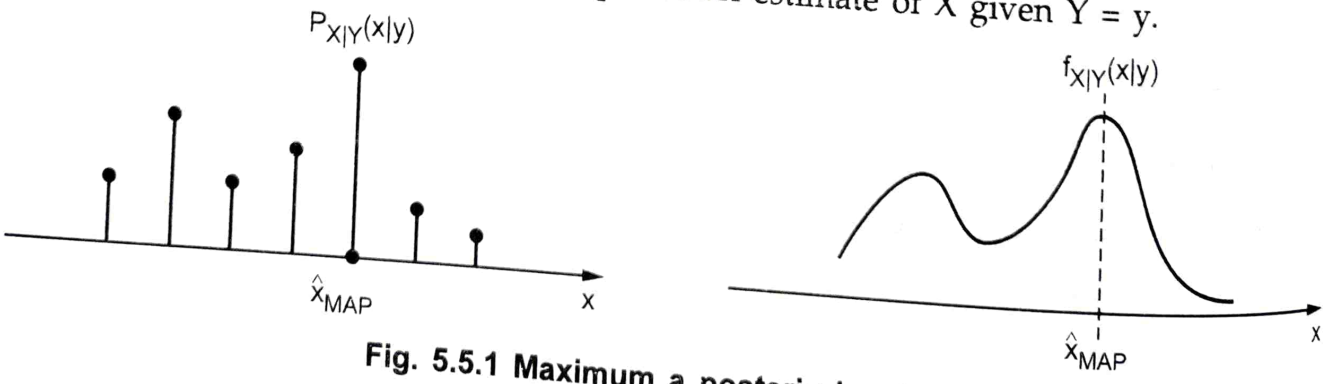


Fig. 5.5.1 Maximum a posteriori estimate

- The MAP estimate of the random variable X, given that we have observed Y = y, is given by the value of x that maximizes. Here
  a) $f_{X|Y}(x|y) \rightarrow$ If X is a continuous random variable,
  b) $P_{X|Y}(x|y) \rightarrow$ If X is a discrete random variable.

- To find the MAP estimate, we need to find the value of x that maximizes

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)\, f_X(X)}{f_y(Y)}$$

5.5.2 Bayesian Parameter Estimation

- A simple Bayesian model has three components.
  1. Observable data are generated as random variables y in some model from a model family with parameters θ.
  2. Prior to observing a particular set of data, however, we have beliefs/expectations about the possible model parameters θ; we call these beliefs I.
  3. These beliefs affect y only through the mediation of the model parameters, that is, y and I are conditionally independent given θ.
- In the Bayesian framework, both parameter estimation and density estimation simply involve the application of Bayes' rule.
- For example, parameter estimation means calculating the probability distribution over θ given observed data y and our prior beliefs I.
- We can use Bayes rule to write this distribution as follows :

$$P(\theta|y, I) = \frac{P(y|\theta, I)P(\theta|I)}{P(y|I)}$$

$$= \frac{\overbrace{P(y|\theta)}^{\text{Likelihood for } \theta} \quad \overbrace{P(\theta|I)}^{\text{Prior over } \theta}}{\underbrace{P(y|I)}_{\text{Likelihood marginalized over } \theta}} \text{ (because } y \perp I|\theta)$$

5.6 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.

6 Foundations of Neural Networks and Deep Learning, Techniques to Improve Neural Networks

Syllabus

Regularization and optimizations, hyperparameter tuning and deep learning frameworks (Tensorflow and Keras.), Convolutional Neural Networks, its applications, Recurrent Neural Networks and its applications.

Contents

6.1 A Quick Review on Neural Networks

- Neural networks, also known as artificial neural networks, can be viewed as a connection of processing elements that are capable to provide the dynamic outputs to the inputs provided to the system.
- These processing elements, preferred as "nodes" in this context, are organized in a layered structure that are interconnected with the nodes in other layers. These layers are generally "hidden layers", which does not take the input directly, but process the input that is provided to the external input layer.
- The input provided is then passed on to these hidden layers, where the actual processing is done, and based on that, patterns are generated, based on which, the output for the system is obtained.
- There might be more than one hidden layer, creating a network, which is typically a neural network. Each layer has is assigned a "weight", which defines the pattern it extracts. This weight is not pre-fixed, instead, it is dependent on the input provided.
- The weight of these layers is generally decided on the basis of a "learning rule", which is dependent on the input. For example, if we provide an image of a locomotive as an input, the features to be extracted can be the number of wheels, size, etc. So, on the basis of the input, the weight of the hidden layer is decided, such as the 1^{st} hidden layer will have more weightage as it counts the number of wheels in the image, the 2^{nd} hidden layer will have less weightage, which will calculate the size of the locomotive in the image and so on.
- These weighted layers process the input, extract some feature from the input and pass it on to the next hidden layer, where its preceding feature can be extracted. Similarly, the process goes on till the output is generated at the final layer.
- A typical neural network is depicted in the following figure :

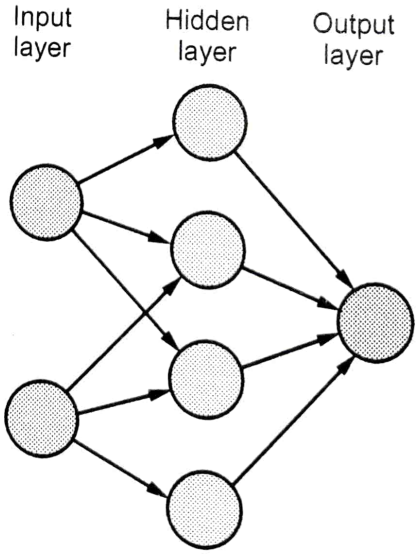


Fig. 6.1.1

6.1.1 Advantages of Neural Networks

- The artificial neural networks take the input and learn the features and patterns all by itself. The output produced by these neural networks are not dependent on the scope of the input provided. The network can learn and provide the output for the data not provided in the initial phases. This gives an advantage of real time processing of data to the network.
- As the input data is stored in the network, instead of database, one can store the pieces of an information across the whole network. Even if any single piece of information is distorted, there is no change in the information.
- Better fault tolerance capability of the network ensures the retain of the data even of a single node of the network becomes unfunctional.
- The network easily finds the fault if there is a node failure, that too without affecting the function of the whole network.
- One can perform multiple functions at a given time using the artificial neural network. Parallel processing of the tasks never affects the system performance.

6.1.2 Disadvantages of Neural Network

- The neural network is dependent on the hardware configuration of the system enabling the parallel processing. Hence, the knowledge which equipment to be used is highly required.
- Whenever you input some data to the network, it simply gives you the probable output, hiding the actual processing from the users. This creates trust issues for the normal users to rely on the neural networks.
- The network is built purely on the trial - and - error method, as there is no fixed provision to be followed to build a neural network.
- Artificial neural networks work with numerical data. Hence, any input provided should be translated first to its equivalent numerical value.
- The duration of the processing of input in the network can not be determined. It varies with the type of inputs provided.

Although the neural networks for implementing the machine learning algorithms provides accurate results to a larger extent, sometimes it may happen that these networks fail to provide the desired output, or the users are not satisfied with the outputs produced. These affect the performance of the neural network.

Hence, in this chapter, we are going to know about some of the models that may help improve the performance of the neural network. The models we are going to study in this chapter are Regularization, Optimization and Hyperparameter tuning.

So, let's get started with the chapter !

6.2 Regularization in Neural Networks

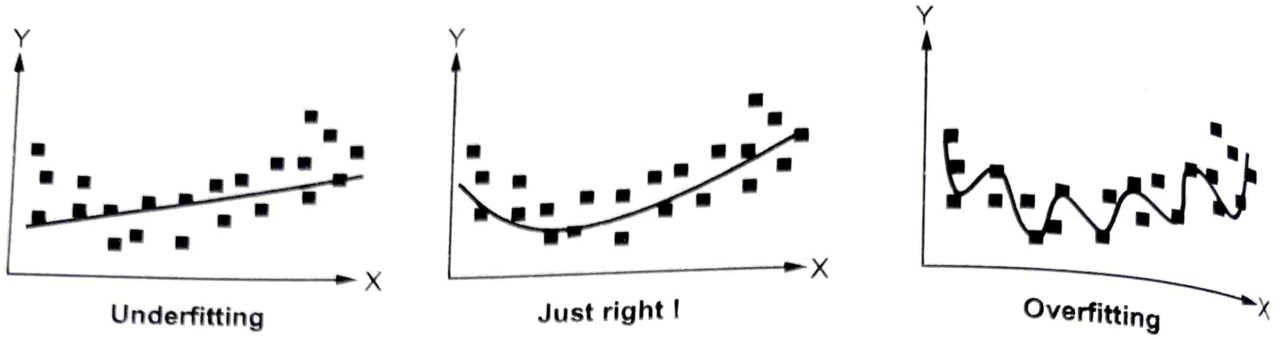


Fig. 6.2.1

- Just have a look at the above figure, and we can immediately predict that once we try to cover every minutest feature of the input data, there can be irregularities in the extracted features, which can introduce noise in the output. This is referred to as "Overfitting".
- This may also happen with the lesser number of features extracted as some of the important details might be missed out. This will leave an effect on the accuracy of the outputs produced. This is referred to as "Underfitting".
- This also shows that the complexity for processing the input elements increases with overfitting. Also, neural networks being a complex interconnection of nodes, the issue of overfitting may arise frequently.
- To eliminate this, regularization is used, in which we have to make the slightest modification in the design of the neural network, and we can get better outcomes.

6.2.1 Regularization in Machine Learning

- One of the most important factors that affect the machine learning model is overfitting.
- The machine learning model may perform poorly if it tries to capture even the noise present in the dataset applied for training the system, which ultimately results in overfitting. In this context, noise doesn't mean the ambiguous or false data, but those inputs which do not acquire the required features to execute the machine learning model.
- Analyzing these data inputs may surely make the model flexible, but the risk of overfitting will also increase accordingly.
- One of the ways to avoid this is to cross validate the training dataset, and decide accordingly the parameters to include that can increase the efficiency and performance of the model.

- Let this be the simple relation for linear regression :

$$Y \approx b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Where Y = Learned relation

β = Co-efficient estimators for different variables and/or predictors (X)

- Now, we shall introduce a loss function, that implements the fitting procedure, which is referred to as "Residual Sum of Squares" or RSS.
- The co-efficient in the function is chosen in such a way that it can minimize the loss function easily.

Hence,

$$RSS = \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2$$

- Above equation will help in adjusting the co-efficient function depending on the training dataset.
- In case noise is present in the training dataset, then the adjusted co-efficient won't be generalized when the future datasets will be introduced. Hence, at this point, regularization comes into picture and makes this adjusted co-efficient shrink towards zero.
- One of the methods to implement this is the ridge regression, also known as L2 regularization. Lets have a quick overview on this.

6.2.2 Ridge Regression (L2 Regularization)

- Ridge regression, also known as L2 regularization, is a technique of regularization to avoid the overfitting in training data set, which introduces a small bias in the training model, through which one can get long term predictions for that input.
- In this method, a penalty term is added to the cost function. This amount of bias altered to the cost function in the model is also known as ridge regression penalty.
- Hence, the equation for the cost function, after introducing the ridge regression penalty is as follows :

$$\sum_{i=1}^{m}(y_i - y'_i)^2 = \sum_{i=1}^{m}(Y_i - \sum_{j=1}^{n}\beta_j \times X_{ij})^2 + \lambda \sum_{j=0}^{n}\beta_j^2$$

Here, λ is multiplied by the square of the weight set for the individual feature of the input data. This term is ridge regression penalty.

- It regularizes the co-efficient set for the model and hence the ridge regression term deduces the values of the coefficient, which ultimately helps in deducing the complexity of the machine learning model.

- From the above equation, we can observe that if the value of λ tends to zero, the last term on the right - hand side will tend to zero, thus making the above equation a representation of a simple linear regression model.
- Hence, lower the value of λ, the model will tend to linear regression.
- This model is important to execute the neural networks for machine learning, as there would be risks of failure for generalized linear regression models, if there are dependencies found between its variables. Hence, ridge regression is used here.

6.2.3 Lasso Regression (L1 Regularization)

- One more technique to reduce the overfitting, and thus the complexity of the model is the lasso regression.
- Lasso regression stands for Least Absolute and Selection Operator and is also sometimes known as L1 regularization.
- The equation for the lasso regression is almost same as that of the ridge regression, except for a change that the value of the penalty term is taken as the absolute weights.
- The advantage of taking the absolute values is that its slope can shrink to 0, as compared to the ridge regression, where the slope will shrink it near to 0.
- The following equation gives the cost function defined in the Lasso regression :

$$\sum_{i=1}^{m}(y_i - y_i')^2 = \sum_{i=1}^{m}(Y_i - \sum_{j=0}^{n}\beta_j \times X_{ij})^2 + \lambda \sum_{j=0}^{n}|\beta_j|^2$$

- Due to the acceptance of absolute values for the cost function, some of the features of the input dataset can be ignored completely while evaluating the machine learning model, and hence the feature selection and overfitting can be reduced to much extent.
- On the other hand, ridge regression does not ignore any feature in the model and includes it all for model evaluation. The complexity of the model can be reduced using the shrinking of co-efficient in the ridge regression model.

6.3 Optimization in Machine Learning

- Optimization issue in machine learning algorithms is about finding the correct set of inputs for a function that gives the maximum effectiveness in the function evaluation.
- Be it fitting in the logical regression models in machine learning or training the neural networks with varying datasets, the problem of optimization arises in all phases of a machine learning model.

- Out of hundreds of optimization algorithms available, it becomes difficult to choose a single algorithm that can give the best performance while applied in our machine learning model. One suggestable approach to execute these optimization algorithms and make the maximum use of it is to group these algorithms and execute them on the machine learning.
- Most of the times, the "continuous function optimization" problem arises in majority of the machine learning algorithms, in which most of the input given are the real numbers, and so are the outputs. These kinds of problems that take only discrete values as input are generally referred to as "Combinatorial Optimization Problems".
- The optimization algorithms state that if more information about the target function can be made available, it becomes easier to optimize that function and that information can also be utilized effectively for further data processing.
- The major point that comes across the execution of an optimization algorithm is to decide whether we can differentiate an objective function at a given point or not.
- That is, can we calculate the first derivative of a function for a given solution or not ?
- Based on this point, the optimization algorithms are further classified into two categories : One that differentiates the function and other that does not.
- Hence, in this section, we shall discuss the "differentiable" and "non differentiable" objective functions that can be used to group multiple optimization algorithms, as discussed above.

6.3.1 Differentiable Objective Function

- If we can calculate the derivative of a function at any given point while input is given to the system such function can be referred to as "Differential Function".
- One can define the derivative of a function as the rate at which the function changes its value at a given point of time. This is often referred to as a slope too.
- One can apply the optimization techniques on these derivative functions using simple calculus.
- Optimization techniques can sound easier if the derivatives of these "continuous functions", as mentioned above, can be calculated. Some of the algorithms that uses these gradient values of the derivatives are as follows :

1. **Bracketing algorithms :**
   - This technique is helpful when there are problems having only one input variable and the optima exist in the pre - defined specific criteria or range.

- These algorithms can easily navigate this range that is known already and locate the optima. The only drawback is that the algorithm assumes that there is only one optima present in the model.
- The advantage of using this algorithm is that it can be utilized in a model even if sometimes there is no derivative information about the variables available.
- Some of the examples that use bracketing algorithms are Fibonacci search, Golden Section search, Bisection method, etc.

2. **Local descent algorithms :**
   - These algorithms work for the models in which there are multiple input variables with one global optima.
   - The algorithm is widely used in the line search problem.
   - This problem includes the definition of the direction to move during a search space and then it performs the bracketing type search in a line in the direction chosen.
   - The algorithm executes until no other iteration of finding the improved directions is possible.
   - These iterations make the algorithm costly as it continues its execution till an effective direction is obtained.
3. **First order algorithms :**
   - These algorithms use the first order derivatives (gradient) to decide the direction to move in the search space.
   - This algorithm works by first calculating the first derivative of the function, and then following it in the opposite direction, for example going downhill to minimum value for minimization problems, with the help of step size, also known as "learning rate".
   - This step size, or learning rate, is a hyperparameter in the algorithm, that decides the distance to cover, or how far to cover in a search space, which is opposite to the normally used local descent algorithms, which do not have this hyperparameter and performs a full line search in every directions specified.
   - These algorithms are also known as "Gradient Descent" algorithms and following the introduction to some of the minor extensions, these are also known as Momentum, Adagrad, RMSProp, Adam, etc.
   - These gradient descent algorithms are also helpful in training the artificial neural networks and implementing deep learning models in it, by providing

the template for Stochastic Gradient Descent, useful for artificial neural networks.

- Here, the gradient will be based on assumption, instead of direct calculation, using the prediction techniques on the trained data.

4. **Second order algorithms :**
   - These algorithms use the second order derivative of the input variables for choosing the direction of movement in the search space.
   - The algorithms work appropriately only for the objective functions in which the Hessian matrix needs to be calculated.
   - Some of the examples in which the second order algorithms are used.
     - Newton's method
     - Secant method
   - These algorithms are also known as Quasi Newton Methods.

6.3.2 Non - Differentiable Objective Function

- Although, optimization algorithms working on the derivatives of the objective functions are efficient and fast, there are certain objective functions whose derivatives cannot be calculated, the reason being the complexity of the function.
- Some of the reasons for the complexities in the function include.
  - Lack of analysis of the function
  - Multiple optima required.
  - Evaluation of stochastic functions
  - Objective functions are discontinuous.
- The optimization algorithms that do not make the compulsion for the first or second order derivatives for their objective functions are called as Black - box optimization algorithms. Some of these algorithms are :
  - Direct algorithms
  - Stochastic algorithms
  - Population algorithms

Let us have a brief of each of these :

1. **Direct algorithms :**
   - These algorithms are used when the calculation of the derivatives of the objective function is not possible.
   - The algorithms work with an assumption that the objective function contains single optima.

- These methods are also referred to as "pattern search" algorithms, since they analyze the search space using the geometrical shapes and patterns.
- The gradient information required to run the algorithm is calculated directly from the objective function by computing the difference between the scores obtained from the points in the search space.
- This information estimated are then helpful in choosing a direction to commute in the search space and cover the region of the present optima.
- Some of the examples that use these direct algorithms are Cyclic Coordinate search, Powell's method, Hooke-Jeeves method, etc.

2. **Stochastic algorithms :**
   - For the variables whose derivatives cannot be calculated, stochastic optimization algorithms use the randomness for those objective functions to commute in the search space.
   - Hence, due to the randomness involved, the stochastic algorithms involve many data sampling for the objective function.
3. **Population algorithms :**
   - These algorithms maintain a pool of solutions for a given input, often known as a population of candidate solutions, which are used to explore, sample the optima.
   - These algorithms are mostly used in the problems that are more challenging and also involves the evaluation of functions containing considerable noise in it, in addition to the presence of multiple global optima. The solutions of such algorithms are difficult to be found by other methods.
   - Genetic algorithms, differential evolution, particle swarm optimization, etc. are the examples of population algorithms.

6.4 Hyperparameter Tuning

- While designing a machine learning model, one always has multiple choices for the architectural design for the model. This creates a confusion on which design to choose for the model based on its optimality. And due to this, there are always trials for defining a perfect machine learning model.
- The parameters that are used to define these machine learning models are known as the hyperparameters and the rigorous search for these parameters to build an optimized model is known as hyperparameter tuning.
- The following questions can be answered by the parameters to build a perfect machine learning model :

- What should be the degree of polynomial features that can be used in my linear model ?
- What should be the maximum depth of my decision tree ?
- How many samples will be required at the leaf node of the decision tree ?
- How many trees should be included in the random forest ?
- How many neurons, layers should be provided in the neural network?
- What should be the learning rate for gradient descent ?

Hyperparameters are not model parameters, which can be directly trained from data. Model parameters usually specify the way to transform the input into the required output, whereas hyperparameters define the actual structure of the model that gives the required data.

The process of defining a hyperparameter can be :

1. Defining the model.
2. Defining the range of values for the hyperparameters.
3. Defining a method for sampling of these hyperparameter values.
4. Defining a criterion to evaluate the performance of a model.
5. Defining a method to cross validate a problem.

Let us discuss the methods used for hyperparameter tuning.

6.4.1 Hyperparameter Tuning Methods

Hyperparameter tuning methods depend on how the model architectures' candidates are sampled in the search space for the given possible hyperparameter values. This is well known as "searching" the hyperparameter space for optimized values.

Let us have a look at the hyperparameter tuning methods :

1. **Grid search :**
   - This is the simplest and the most basic hyperparameter tuning method.
   - Simply build the model using all the possible combinations of the hyperparameter values available, then evaluate each of these models and then select the model giving the best and the most relevant output.
   - Each of the defined model can be fit with a training data and validation data can be provided as an input to evaluate the model.
   - The disadvantage is that this method uses an exhaustive sampling of data in the hyperparameter space, making the method inefficient.

2. **Random search :**
   - Instead of providing a long range of values as a hyperparameter as done in grid search method, in random search, statistical distribution defined for each parameter can be used and sampling on the values can be done randomly on these distributions.
   - We can also fix the number of iterations in the search space to find the perfect model.
   - For every possible iteration, the hyperparameter values can be set by the sampling process.

Analyzing both the methods, the grid search method generally misses the optimal model while spending time in exploration of not so important hyperparameters, whereas the random search method focuses on the search of the optimal value for an important hyperparameter.

6.5 Deep Learning Frameworks

A deep learning framework can be defined as an interface or library or a tool using which one can build deep learning models with ease and in less time, even if we do not have the information about the depth of the algorithms. There are pre - built and customized components present in the framework, which enables the quick and easy build of deep learning models.

The main advantage of these deep learning frameworks is that we do not have to write longer lines of code, instead we can simply build the model using existing libraries. Some of the features of an efficient deep learning frameworks are :

- Optimization in performance of a model
- Easy to use, understand and implement,
- Better community reach
- Parallel processing
- Computes the gradients automatically.

There are several deep learning frameworks available - for example - TensorFlow, Keras, PyTorch, Caffe, DeepLearning4j, Microsoft cognitive toolkit etc.

In this section, we are going to see the details of TensorFlow and Keras. Let us have a brief of these :

6.5.1 TensorFlow

- TensorFlow is one of the most popular frameworks used to build deep learning models. The framework is developed by Google Brain Team.
- Languages like C++, R and Python are supported by the framework to create the models as well as the libraries. This framework can be accessed from both - desktop and mobile.
- The translator used by Google is the best example of TensorFlow. In this, the model is created by adding the functionalities of text classification, natural language processing, speech or handwriting recognition, image recognition, etc.
- The framework has its own visualization toolkit, named TensorBoard which helps in powerful data visualization of the network along with its performance.
- One more tool added in TensorFlow, TensorFlow Serving, can be used for quick and easy deployment of the newly developed algorithms without introducing any change in the existing API or architecture.
- TensorFlow framework comes along with a detailed documentation for the users to adapt it quickly and easily, making it the most preferred deep learning framework to model deep learning algorithms.
- Some of the characteristics of TensorFlow is :
  - Multiple GPU supported
  - One can visualize graphs and queues easily using TensorBoard.
  - Powerful documentation and larger support from community

6.5.2 Keras

- If you are comfortable in programming with Python, then learning Keras will not prove hard to you. This will be the most recommended framework to create deep learning models for ones having a sound of Python.
- Keras is built purely on Python and can run on the top of TensorFlow. Due to its complexity and use of low - level libraries, TensorFlow can be comparatively harder to adapt for the new users as compared to Keras. Users those who are beginners in deep learning, and find its models difficult to understand in TensorFlow generally prefer Keras as it solves all complex models in no time.
- Keras has been developed keeping in mind the complexities in the deep learning models, and hence it can run quickly to get the results in minimum time. Convolutional as well as Recurrent Neural networks are supported in Keras. The framework can run easily on CPU and GPU.

- The models in Keras can be classified into 2 categories :

1. Sequential model :

The layers in the deep learning model are defined in a sequential manner. Hence the implementation of the layers in this model will also be done sequentially.

2. Keras functional API :

Deep learning models that has multiple outputs, or has shared layers, i.e. more complex models can be implemented in Keras functional API.

6.6 Convolutional Neural Networks

- In today's era, where sharing of images has just become the part of our routines, the people are now using the images for their communication too. Hence, the machine recognizing the image to extract a pattern off it must be trained accordingly for the correct interpretation of data. To accomplish this complex task, convolutional neural networks are developed.
- To understand the Concept of Convolutional Neural Networks (CNNs), let us take an example of the images our brain can interpret.
- As soon as we see an image, our brain starts categorizing it based on the color, shape and sometimes also the message that image is conveying. Similar thing can be done through machines even after a rigorous training. But the difficulty is there is a huge difference in what humans interpret and what machine does. For a machine, the image is merely an array of pixels. There is a unique pattern included in each object present in the image and the computer tries to find out these patterns to get the information about the image.
- Machines can be trained giving tons of images to increase its ability to recognize the objects included in a given input image.
- Most of the digital companies have opted for CNNs for image recognition, some of these include Google, Amazon, Instagram, Pinterest, Facebook, etc.
- Hence, we define a convolutional neural network as : "A neural network consisting of multiple convolutional layers which are used mainly for image processing, classification, segmentation, and other correlated data."

6.6.1 Architecture of Convolutional Neural Network

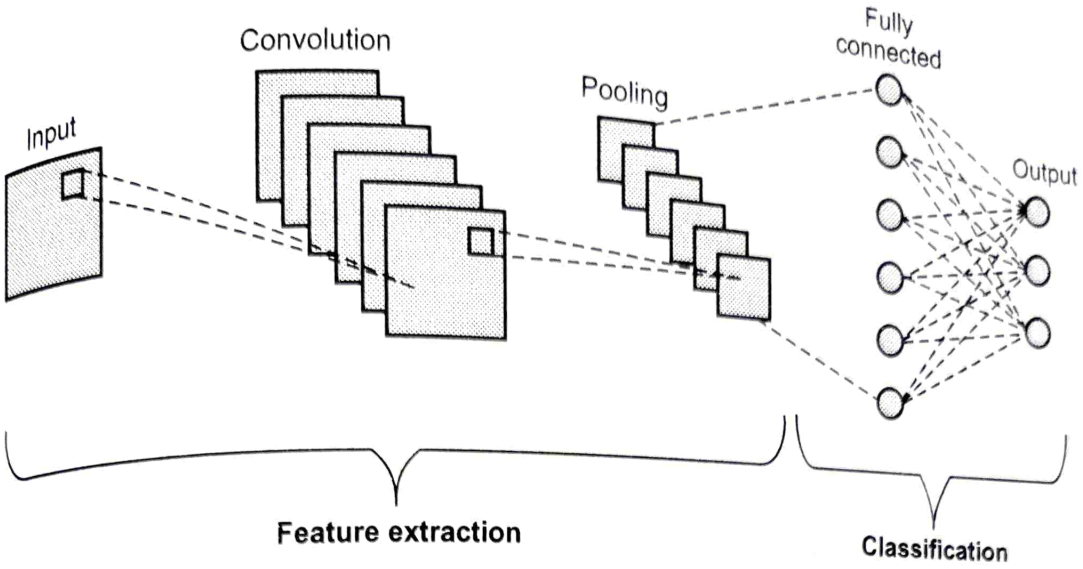


Fig. 6.6.1

- A convolutional neural network, as discussed above, has the following layers that are useful for various deep learning algorithms. Let us see the working of these layers taking an example of the image having dimension of $12 \times 12 \times 4$. These are :

1. **Input layer :** This layer will accept the image of width 12, height 12 and depth 4.
2. **Convolution layer :** It computes the volume of the image by getting the dot product between the image filters possible and the image patch. For example, there are 10 filters possible, then the volume will be computed as $12 \times 12 \times 10$.
3. **Activation function layer :** This layer applies activation function to each element in the output of the convolutional layer. Some of the well accepted activation functions are ReLu, Sigmoid, Tanh, Leaky ReLu, etc. These functions won't change the volume obtained at the convolutional layer and hence it will remain equal to $12 \times 12 \times 10$.
4. **Pool layer :** This function mainly reduces the volume of the intermediate outputs, which enables fast computation of the network model, thus preventing it from overfitting.

6.6.2 Applications of CNN

- CNN is mostly used for image classification, for example to determine the satellite images containing mountains and valleys, or recognition of handwriting, etc. image segmentation, signal processing, etc. are the areas where CNN are used.

Let us see some real life applications used in daily life. These are :

1. Face recognition system
2. Analysis of documents
3. Climate analysis
4. Handwriting recognition
5. Historic items detection
6. Advertisements
7. Recognition of grey areas and so on...

6.7 Recurrent Neural Networks

- The type of neural networks where the output from one layer is fed as input to the next layer are known as Recurrent Neural Networks (RNN).
- RNNs can be used in the cases where, for example, next word to place in a sentence can be predicted. Here, the information regarding the previous words is needed to predict the next word in the sentence. RNN can solve this issue with the help of hidden layers, which is one of the most important and unique features about the RNN. These hidden layers remember the previous information to be used in the next layers.

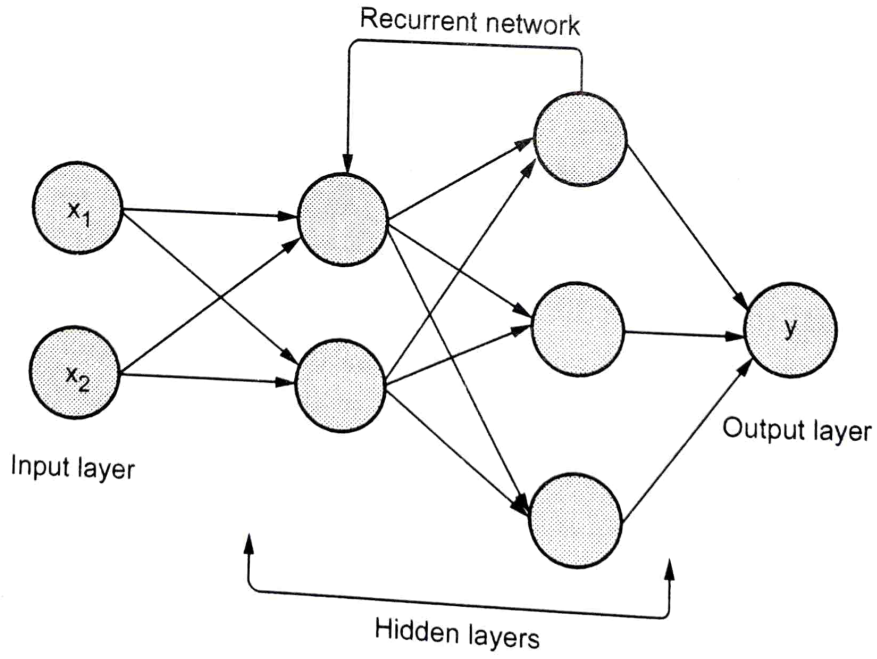


Fig. 6.7.1

6.7.1 Architecture of Recurrent Neural Network

Here are some of the advantages of using recurrent neural networks :

1. RNN can remember the previous information and can use it for upcoming computations. This is useful in time series prediction. This is generally referred to as Long Short Term Memory (LSTM).
2. These neural networks can also be implemented in the convolutional neural networks in case you want to expand the pixel neighborhood.

In spite of these major advantages, RNN have the following disadvantages :

1. Problems related to gradient vanishing may arise.
2. Training the recurrent neural network is not easy task.
3. It cannot process longer sequences of data.

Some of the applications of RNN include :

- Problems related to predictions
- Language and text generation
- Translations
- Recognizing speech
- Generating the descriptions from an image
- Tagging videos
- Summarizing text
- Detection of face, OCR, images, etc.
- Composing music notes, etc.

❑❑❑

7 Deep Learning - More to Know

Syllabus

Generative Adversarial Networks, Deep Reinforcement Learning, Adversarial Attacks

Contents

7.1 Generative Adversarial Networks

- Generative Adversarial Networks, abbreviated as GANs, are the method or approach to generate the models like Convolutional Neural Networks, using deep learning techniques.
- One type of modelling technique, that is Generative Modelling, is considered to be unsupervised learning technique in terms of Machine Learning tasks, which might involve discovering the patterns automatically, and learning this input patterns in such a way that the model will be capable to produce the output based on the input pattern, which may or may not be present in the available dataset.
- On the other hand, Generative Adversarial Networks (GANs) train the generative model by arranging the pattern in the form of a supervised learning model that might again consist of two sub models, namely - The generator model and the discriminator model.
- The generator model can be trained by us to get new samples, whereas in discriminator model can be helpful in classifying these input samples as real or fake.
- GANs are actively changing fields, generating models for realistic examples among a range of problems in the space. One such task carried out by a GAN is the translation from an image to image. These translations can be, for example, translation of photos of day to that of night, photos of summer to those of winter, and so on.
- These photo translations of objects, scenes, etc. seem so realistic that nobody can predict whether they are real or fake.

7.1.1 What are Generative Adversarial Networks ?

- Generative adversarial networks are the generative models that are based on the deep learning models.
- In other words, GANs are a type of architecture that trains a generative model, mostly used by deep learning architectures.
- GAN model architecture contains two sub - models :
  - Generator : Generates possible samples from the domain
  - Discriminator : Classifies these samples as real or fake

7.1.1.1 Generator Model

- A fixed length vector is given as an input to the generator model, which in turn generates samples from that domain.

- This fixed length vector is taken randomly from the Gaussian distribution, which is used as a base for this generator model.
- Once the training of these input vectors are done, the points in the multi dimensional vector space will be mapped with those points in the problem domain.
- This will form a compressed representation of the input data distribution.
- The vector space here can be referred to as latent space, or a space comprising of latent variables.
- These latent variables may or may not be visible, but are really important for a domain.
- In terms of GAN, the generator model gives a meaning to these latent variables chosen in the vector space, so that it can sample more points in the latent space, based on which more examples can be generated as an output of the generator model.
- Once the input vectors are trained, the generator model can be kept to further generate new samples.
- The process of generator model is as shown in the following figure :

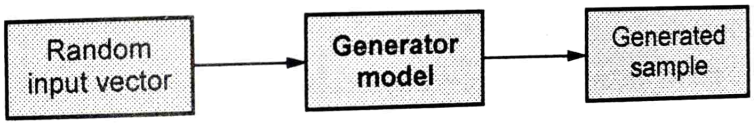


Fig. 7.1.1 Working of generator model

7.1.1.2 Discriminator Model

- The discriminator model, a simple classifier model, takes one of the samples generated by the generator model as an input, which might be either real or generated, and predicts a binary variable for that sample, with a class label of real or fake.
- Real samples are those that are obtained from the original datasets. And these real samples are the outcomes of the generator model.
- Once the classification is done, and the samples are trained accordingly, the discriminator model is discarded, unlike the generator model which is kept for generating more samples in the future.
- The following figure shows the working of the discriminator model :

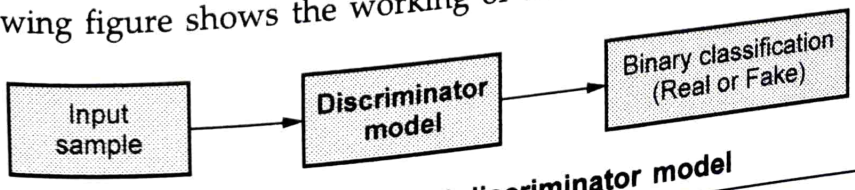


Fig. 7.1.2 Working of discriminator model

7.1.1.3 GAN as a Combination of Generator and Discriminator Models

- Both of the models, generator and discriminator models, are trained in parallel.
- The generator model, after generating a bunch of samples from the realistic examples in the domain, provides these samples as an input to the discriminator model, which in turn provides the data whether the provided samples are real or fake.
- The updates in both the models are then carried so that the generator model can generate the batch of real samples and should not fool the discriminator by providing the fake samples.
- The discriminator model is updated so that it can work with better efficiency for an effective classification of real and fake samples generated by the generator model.
- We can say that these models are competing with each other, and if we can convey this thing in terms of a game, then this can be known as adversarial, and the game can be called as a zero sum game.
- The zero-sum game works somehow like this :
- When a discriminator model do not recognize a fake sample, i.e. generator has fooled the discriminator, no update is needed in the generator model, and larger updates are introduced in the discriminator model.
- If the discriminator succeeds in finding out the fake sample, then there is no update needed in the discriminator model, whereas the generated model has been introduced with large number of updates.
- This is depicted in Fig. 7.1.3.

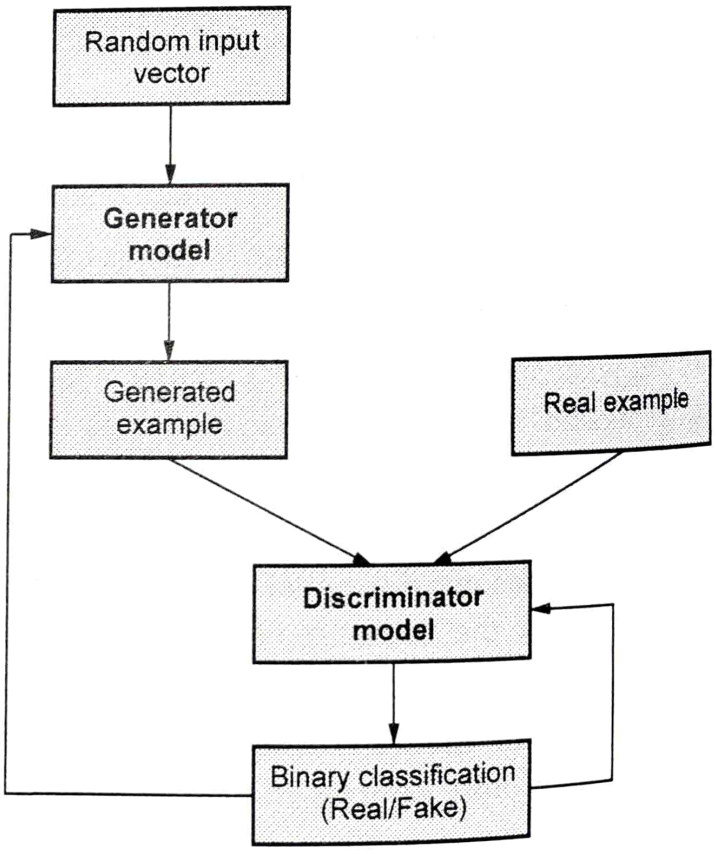


Fig. 7.1.3 Generative adversarial network

7.1.2 Why Generative Adversarial Networks are Used ?

- A technique of computer vision called "Data Augmentation" that is based on the deep learning models widely use the generative adversarial networks.
- Data augmentation enables the creation of new, efficient samples for training the deep learning model. This method not only improves the performance of the models, but also increases its skill and regularization, thus minimizing the generalization errors.
- Mostly, these techniques are used in training the data like images, that particularly involve edits like crop, flip, zoom - in and zoom - out, and other simple transformations.
- The GAN models enable the data augmentation techniques to become more domain specific.
- Domains like deep reinforcement learning also use GANs to train their models.
- In a summary, the best three advantages or areas where GANs are an important factor are as follows :
  1. Image resolution : For The input images, GAN provides high resolution images as an output.
  2. Art creation : The GAN model generates new images, sketches, paintings, etc.
  3. Image translation : Translation of photographs, for example converting a day picture into night picture and so on.

7.2 Deep Reinforcement Learning

- A subfield of Machine learning, deep reinforcement learning is the combination of Reinforcement learning and Deep Learning.
- The basic working of reinforcement learning is based on making the decisions for the computational problems on the basis of trial and error method.
- Hence, deep reinforcement learning adds the deep learning techniques to this method, and allows the computer to make decisions on the unstructured data, without any manual engineering of creating the state space. The relation between deep learning and reinforcement learning can be as shown in the Fig. 7.2.1.
- Deep RL algorithms can take a large number of inputs at a time, for instance, in a video game playing on screen, each and every pixel is taken as an input at a time, and hence decides the actions to be performed on these inputs to provide efficient outcomes.

- The algorithms of deep reinforcement learning are used in various emerging fields such as - robotics, computer vision, natural language processing, video games, transportation, healthcare, education, etc.

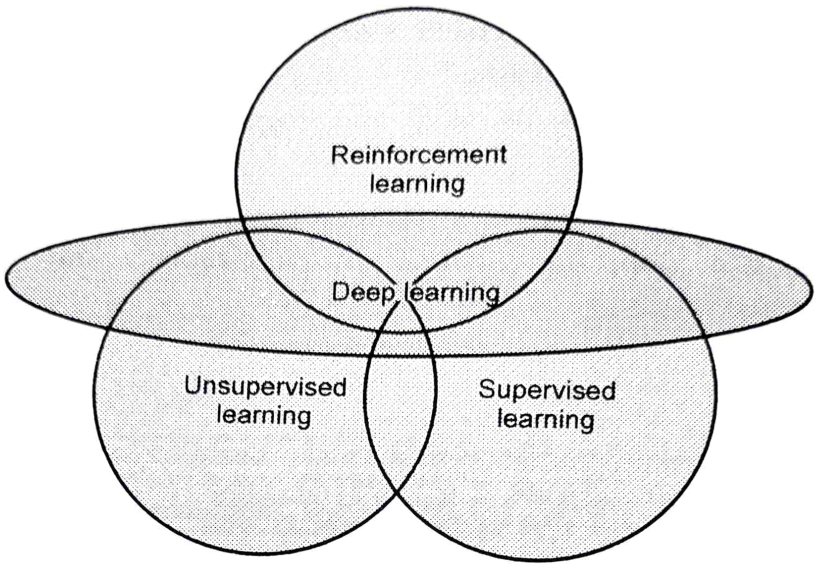


Fig. 7.2.1 Relationship between deep learning and reinforcement learning

7.2.1 Applications of Deep Reinforcement Learning

- The current evolution in deep reinforcement learning has brought up many advancements and transformations in information technology sector, with the applications widely accepted in the field of clinical analysis, marketing, finance, driving, robotics, smart grid and so on.
- The detailed applications of deep reinforcement learning are listed as below :

1. **AI toolkits :**

   AI toolkits, like DeepMind Lab, Psychlab, OpenAI Gym, etc. provide the environment for training the large data sets for innovation in deep reinforcement learning algorithms.

2. **Process of manufacturing :**

   The installation of intelligent robots in the warehouse and stockyards can manage the inwards and outwards of the millions of quantities of raw materials used in the manufacturing process. The delivery of the products to the right person can also be ensured.

   When a robot picks a product, deep reinforcement learning algorithm will enable the robot, based on the knowledge provided, whether it has picked the right product or not.

3. **Automotives :**

The use of autonomous vehicles, i.e. self driven vehicles make a use of deep reinforcement learning model for getting the correct route, maintenance of the vehicle, etc.

The deep reinforcement learning model that consist the data of the customers, dealers, warranties given and their expiries, etc. will provide a satisfactory customer assistance in automotive industry.

4. **Finance :**

The techniques of deep reinforcement learning applied on the financial sector can prove better investment managers than the human beings, and also evaluate trading strategies .

5. **Healthcare :**

Diagnosis for critical diseases, treatment plans, new drug development, etc. can be carried out on the basis of the techniques of deep reinforcement learning.

6. **Bots :**

The bots on the UI platforms of a web application are developing rapidly with the convergence of deep reinforcement learning techniques to gain the grip over languages used by a common user.

7.2.2 Future Development of Deep Reinforcement Learning

- The newer evolutions in deep reinforcement learning will enable the embedding of explicit algorithms into neural networks by making them differentiable.
- This will create more abstract algorithms which would be suited for better reasoning for the inputs to be trained, which would be helpful in covering the wider range of applications using deep RL methods.
- The deep RL methods can be also used in the fields of meta learning and lifelong learning, where there would be a need of previous knowledge extracted to gain the present knowledge.
- Another development required in the deep RL field is to transfer the model that is based on simulated inputs, to that of real world inputs.
- This will allow the implementations of complex problems on simulations, and then implementing them on real world problems.

7.3 Adversarial Attacks

- Adversarial attacks in machine learning as well as deep learning is a technique to fool the models by providing the deceptive inputs.

- Most of the models are designed to address the problems in which the input dataset is to be trained. When these models are finally implemented in real world applications, these adversaries might supply the fake data that may violate the model assumptions.
- These adversaries might affect the overall performance of the model, in addition to the final outcome.
- The following example gives the clear picture of these adversarial attacks :

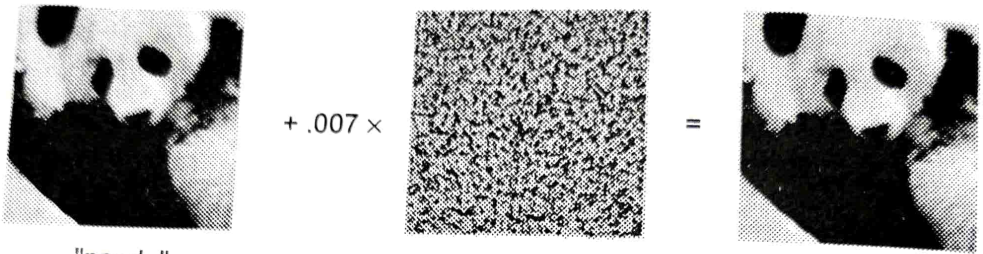


Fig. 7.3.1

- In the above figure, the panda is recognized correctly in the image initially. When a slight noise has been introduced in the model, the results of the model predicting panda has now been changed to gibbon, also with high rate of confidence.
- By looking at both the images, we might not find most of the differences between the two, whereas for the model predicting the animal in the picture, it seems to have major differences.
- This shows the effect of adversarial threats to the models to compute the input data. This happens because in vision, we might not notice any change in the data, hence we cannot predict whether the data is affected by noise or not.
- This leads to the disability of proving the correctness of the output given by the model, as we don't know whether the data given is valid or not.

7.3.1 Types of Adversarial Attacks

- There are two types of Adversarial attacks possible - Targeted and Untargeted attacks.
- Talking about the targeted attack, let us say there is a class Y, there is a model M on which the attack is targeted, which classifies the image I, which is of class X.

- Hence, the targeted attack will be on model M, which will misclassify the image I by providing an adversarial sample of data to M. This will lead to the resulting class Y instead of class X, provided as the output of the model M.
- Unlike the targeted attack, the untargeted attack does not intend to attack on a specific class of data input. It simply provides miscellaneous data samples to the model that can produce the adversarial outputs.
- Although untargeted attacks are not much effective as compared to the targeted attacks, still the untargeted attacks are proved faster and easier to implement.
- Targeted attacks, although being highly effective on the outputs of the model, come at a cost of more time consumption.

7.3.2 Black Box Attacks

- One thing can be concluded from the definition of targeted or untargeted attacks, is that the training model is known to the one initiating the attack.
- Only if the model to be affected is known, then there can be adversarial inputs to it to generate false outcomes.
- Generally it may also happen that the attacker may not have the access to the targeted model, still he succeeds to feed the adversarial inputs to the model.
- Black box attacks are basically working by transferring the adversarial examples from one model to another.
- For example, if the targeted attack is on model G, along with that, the same inputs are provided to model H also and the latter model also gives the undesired outcomes. Hence, we can say that the adversarial input is transferred to model H after being input on model G, providing the desired results to the attacker.
- This type of attack can be launched in the training dataset of the model, which can be known or unknown to the attacker.
- In case the dataset of the model is known, a dummy dataset can be created that mimics the original one and that mimicked dataset can be fed to the model.

7.3.2.1 Types of Black Box Attacks

- In this section, let us discuss various tactics used by the adversaries for black box attacks :

1. Physical attacks :

- One simple way of changing the input, say x, to x' is to simply add something manually or physically, for example, changing the color combination in the image to disturb the training process of the model.

- The example of such attacks is the addition of sunglasses in the image of a person which is to be used as an input data in the face recognition system.
- Another tactic that can be used to distort the accuracy of the trained model is to introduce certain objects in the image, such that the object to be extracted is faded out from the focus.
- For example, in an image of banana, you introduce watermelon, which is much larger in size as compared to banana, and the feature extraction does not extract the features of banana, but will extract the features of the watermelon, which will provide undesired output.

2. Out of Distribution (OOD) attacks

- There is a traditional and old school assumption in the machine learning models that the dataset provided to train the model are obtained from the same source and data distribution.
- This assumption is exploited by the OOD attack and feeds the dataset for the model with random inputs from varying distributions. Hence, the name justifies itself - out of distribution attacks.

❑❑❑