MU

Sem 6

Computer Engineering

2021 EDITION

Course Code (CSDLO6021)

(Department Level Optional Course– II)

Machine Learning

with MCQ's

Dr. Shubhangi Vaikole
Associate. Prof.
Datta Meghe COE,
Airoli, Navi Mumbai

Dr. Sawarkar S.D.
Principal,
Datta Meghe COE,
Airoli, Navi Mumbai

e-books (PDF download)

Download App

SCAN TO VISIT

Sure Marks UNIVERSITY PAPER SOLUTIONS

ISBN 978-81-941546-7-8

M6-14A

Price ₹ 245/-

With all previous Solved University Q. Papers

TECH-NEO PUBLICATIONS
Where Authors Inspire Innovation
A Sachin Shah Venture

FREE DOWNLOAD Sample chapter from our Android App

University of Mumbai

Machine Learning

(Course Code : CSDLO6021)

Department Level Optional Course -II

Semester VI - Computer Engineering

Strictly as per the Revised Syllabus (REV-2016 'CBCGS' Scheme) of Mumbai University w.e.f. academic year 2018-2019

Dr. Vaikole Shubhangi Liladhar

Associate.Prof. (Ph.D)

Datta Meghe College of Engineering Sector-3, Airoli,
Navi Mumbai - 400708

Dr. Sawarkar Sudhirkumar Deoraoji

Professor & Principal (Ph.D)

Datta Meghe College of Engineering Sector-3, Airoli,
Navi Mumbai - 400708

TECH-NEO PUBLICATIONS
Where Authors Inspire Innovation
A Sachin Shah Venture

M6-14A

Machine Learning

(Course Code : CSDLO6021)
(MU - Semester 6/Computer – For Rev Syll. 2018-2019)

Authors : Dr. Vaikole Shubhangi Liladhar,
Dr. Sawarkar Sudhirkumar Deoraoji

First Edition for Rev Syllabus : February 2021

Tech-Neo ID : M6-14

Published by

Mr. Sachin S. Shah
Managing Director, B. E (Industrial Electronics)
An Alumunus of IIM Ahmedabad

Address

Tech-Neo Publications LLP
Sr. No. 38/1, Behind Pari Company, Khedekar Industrial Estate, Narhe, Maharashtra, Pune-411041.

Website : www.techneobooks.com

Printed at

Image Offset (Mr. Rahul Shah)

Dugane Ind. Area, Survey No. 28/25, Dhayari Near Pari Company, Pune - 411041. Maharashtra State, India.

E-mail : rahulshahimage@gmail.com

Dedicated to .....

The Readers of this Book

- Authors

Preface

We are glad to present the New Edition of this book titled "**Machine Learning**". This book covers the revised Syllabus of Computer Engineering for Third year engineering (semester-6) course of "Mumbai University" which has been effective since the academic year 2018-2019.

We have divided the subject into small chapters so that the topics can be arranged and understood properly. The topics within the chapters have been arranged in a proper sequence to ensure smooth flow of the subject.

A large number of solved examples have been included. So, we are sure that this book will cater all your needs for this subject.

We are thankful to Shri. Sachin Shah for the encouragement and support that they have extended to us. We are also thankful to the staff members of Tech-Neo Publications and others for their efforts to make this book as good as it is. We have jointly made every possible efforts to eliminate all the errors in this book. However if you find any, please let us know, because that will help us to improve further.

We are also thankful to my family members and friends for their patience and encouragement.

- Authors

Syllabus...

Course Code	Course Name	Credits Assigned
CSDLO6021	Machine Learning	4

Prerequisites : Data Structures, Basic Probability and Statistics, Algorithms

Course Objectives :

1. To introduce students to the basic concepts and techniques of Machine Learning.
2. To become familiar with regression methods, classification methods, clustering methods.
3. To become familiar with Dimensionality reduction Techniques.

Course Outcome : Students will be able to -

1. Gain knowledge about basic concepts of Machine Learning
2. Identify machine learning techniques suitable for a given problem
3. Solve the problems using various machine learning techniques
4. Apply Dimensionality reduction techniques.
5. Design application using machine learning techniques

Module	Detailed Contents	Hrs.
1	**Introduction to Machine Learning** Machine Learning, Types of Machine Learning, Issues in Machine Learning, Application of Machine Learning, Steps in developing a Machine Learning Application. **(Refer chapter 1)**	6
2	**Introduction to Neural Network** Introduction - Fundamental concept - Evolution of Neural Networks - Biological Neuron, Artificial Neural Networks, NN architecture, Activation functions, McCulloch-Pitts Model. **(Refer chapter 2)**	8
3	**Introduction to Optimization Techniques** Derivative based optimization- Steepest Descent, Newton method. Derivative free optimization- Random Search, Down Hill Simplex. **(Refer chapter 3)**	6
4	**Learning with Regression and trees** Learning with Regression : Linear Regression, Logistic Regression. Learning with Trees : Decision Trees, Constructing Decision Trees using Gini Index, Classification and Regression Trees (CART). **(Refer chapter 4)**	10
5	**Learning with Classification and clustering** 5.1 **Classification :** Rule based classification, classification by Bayesian Belief networks, Hidden Markov Models. **Support Vector Machine :** Maximum Margin Linear Separators, Quadratic Programming solution to finding maximum margin separators, Kernels for learning non-linear functions. 5.2 **Clustering :** Expectation Maximization Algorithm, Supervised learning after clustering, Radial Basis functions. **(Refer chapter 5)**	14
6	**Dimensionality Reduction** Dimensionality Reduction Techniques, Principal Component Analysis, Independent Component Analysis, Single value decomposition. **(Refer chapter 6)**	8
	Total	52

□□□

Index

Module 1

Chapter... 1

Introduction to Machine Learning

University Prescribed Syllabus

Machine Learning, Types of Machine Learning, Issues in Machine Learning, Application of Machine Learning, Steps in developing a Machine Learning Application.

1.1 MACHINE LEARNING

- A machine that is intellectually capable as much as humans, have always attracted writers and early computer scientist who were excited about artificial intelligence and machine learning.
- The first machine learning system was developed in the 1950s. In 1952, Samuel has developed a program to play checkers. The program was able to observe positions at game and learn the model that gives better moves for machine player.
- In 1957, Frank Rosenblatt designed the Perceptron, which is a simple classifier but when it is combined in large numbers, in a network, it became a powerful tool.
- Minsky in 1960, came up with limitation of perceptron. He showed that the X-OR problem could not be represented by perceptron and such inseparable data distribution cannot be handled and following this Minsky's work neural network research went to dormant until 1980s.
- Machine learning became very famous in 1990s,due to the introduction of statistics. Computer science and statistics combination lead to probabilistic approaches in Artificial intelligence. This area is further shifted to data driven techniques. As Huge amount of data is available, scientists started to design intelligent systems that are able to analyze and learn from data.
- Machine learning is a category of Artificial Intelligence. In machine learning computers has the ability to learn themselves, explicit programming is not required.
- Machine focuses on the study and development of algorithms that can learn from data and also make predictions on data.
- Machine learning is defined by Tom Mitchell as "A program learns from experience 'E' with respect to some class of tasks 'T' and performance measure 'P', if its performance on tasks in 'T' as measured by 'P' improves with 'E'." Here 'E' represents the past experienced data and 'T' represents the tasks such as prediction, classification, etc. Example of 'P', we might want to increase accuracy in prediction.
- Machine learning mainly focuses on the design and development of computer programs that can teach themselves to grow and change when exposed to new data.
- Using machine learning we can collect information from a dataset by asking the computer to make some sense from data.Machine learning is turning data into information.

Fig. 1.1.1: Machine Learning

- Fig. 1.1.1 is the schematic representation of the ML system. ML system takes the training data and background knowledge as the input. Background knowledge and data helps the Learner program to provide a solution for a particular task or problem. Performance corresponding to the solution can be also measured. ML system comprises of mainly two components, Learner and a Reasoner. Learner use the training data and background knowledge to build the model and this can be used by reasoner to provide the solution for a task.
- Machine learning can be applied to many applications such as politics to geosciences. It is a tool that can be applied to many problems. Any application which needs to extract some information from data and also takes some action on data,can benefit from machine learning methods.

- Some of the applications arespam filtering in email, face recognition, product recommendations from Amazon.com and handwriting digit recognition.
- In detecting spam email, if you check for the occurrence of single word it will not be very helpful. But checking the occurrences of certain words used together and combined this with the length of the email and other parameters, you could get a much clearer idea of whether the email is spam or not.

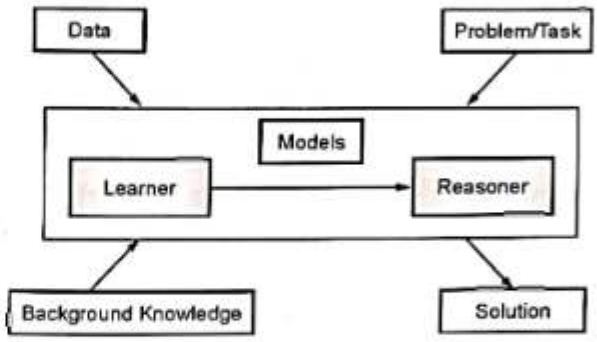


Fig. 1.1.2 : Schematic diagram of Machine Learning

- Machine learning is used by most of the companies toincrease productivity, forecast weather,to improve business decisions, detect diseaseand do many more things.
- Machine learning uses statistics. There are many problems where the solution is not deterministic. There are certain problems for which we don't have that much information and also don't have that much computing power to properly model the problem. For these problems we need statistics, example of such type of problem is prediction of motivation and behavior of humans.The behaviorand motivationof humans is a problem that is currently very difficult to model.

Machine learning = Take data + understand it + process it + extract value from it + visualize it + communicate it

1.2 KEY TERMINOLOGY

Expert System

- Expert system is a system which is developed using some training set, testing set, and knowledge representation, features, algorithm and classification terminology.
  - *(i)* ***Training Set :*** A training set comprises of training examples which will be used to train machine learning algorithms.
  - *(ii)* ***Testing Set :*** To test machine learning algorithms what's usually done is to have a training set of data and a separate dataset, called a test set.
  - *(iii)* ***Knowledge Representation :*** Knowledge representation may be stored in the form of a set of rules. It may be an example from the training set or a probability distribution.
  - *(iv)* ***Features :*** Important properties or attributes.
  - *(v)* ***Classification :*** We classify the data based on features.
- **Process :** Suppose we want to use a machine learning algorithm for classification. The next step is to train the algorithm, or allows it to learn. To train the algorithm we give as a input a quality data called as training set.
- Each training example has some features and one target variable. The target variable is what we will be trying to predict with our machine learning algorithms. In a training dataset the target variable is known. The machine learns by finding some relationship between thetarget variable and the features.In the classification tasks the target variables are known as classes. It is assumed that there will be a limited number of classes.

- The class or target variable that the training example belongs to is then compared to the predicted value, and we can get idea about the accuracy of the algorithm.
- **Example :** First we will see some terminologies that are frequently used in machine learning methods. Let's take an example that we want to design a classification system that will classify the instances in to either Acceptable or Unacceptable. This kind of system is a fascinating topic often related with machine learning called *expert systems.*
- Four features of the various cars are stored inTable 1.2.1. The features or the attributes selected are Buying_Price, Maintenance_Price, Lug_Boot and Safety. Examples belong to Table 1.2.1 represents a record comprises of features.
- In Table 1.2.1 all the features are categorical in nature and takes limited disjoint values. The first two features represent the buying price and maintenance price of a car such as high, medium and low. Third feature shows the luggage capacity of a car as small, medium or big. Fourth feature represents whether the car has safety measures or not, which takes the value as low, medium or high.
- Classification is one of the important task in machine learning. In this application we want to evaluate the car out of a group of other cars. Suppose we have all information about car's Buying_Price, Maintenance_Price, Lug_Boot and Safety. Classification method is used to evaluatea given car as Acceptable or Unacceptable.Many machine learning algorithms are there that can be used for classification. The target or the response variable in this example is the evaluation of a car.
- Suppose we have selected a machine learning algorithm to use for classification. The main task in the classification is to train the algorithm, or allow it to learn. We give the experienced data as the input to train the algorithm which is called as training data.
- Let's assume training dataset contains 14 training records in Table 1.2.1. Suppose each training record has four features and one target or the response variable,as shown in Fig. 1.2.1. The machine learning algorithm is used to predict the target variable.
- In classification task the target variable takes a discrete value, and in the task of regression its value could be continuous.
- In a training dataset we have the value of target variable. The relationship that exists between the features and the target variable is used by machine for learning. The target variable is the evaluation of the car. Classes are the target variables in the classification task. In classification systems it is assumed that classes are to beof limited number.
- Attributes or features are the individual values that, when combined with other features, make up a training example. This is usually columns in a training or test set.
- A training dataset and a testing dataset, is used to test machine learning algorithms. First the training dataset is given as input to the program. Program uses this data to learn. Next, the test set is given to the program. The program decides which instance of test data belongs to which class. The predicted output is compared with the actual output of the program, and we can get anidea about the accuracy of the algorithm. There are best ways to use all the information in the training dataset and test dataset.
- Assume in car evaluation classification system, we have tested the program and it meets the desired level of accuracy. Knowledge representation is used to check what the machine has learned. There are many ways in which knowledge can be represented.
- We can use set of rules or a probability distribution to represent the knowledge.

Module 1

– Many algorithms represent the knowledge which is more interpretable to humans than others. In some situations we may not want to build an expert system but we are interested only in the knowledge representation that's acquired from training a machine learning algorithm.

Table 1.2.1 : Car evaluation classification based on four features.

Buying_Price	Maintenance_ Price	Lug_Boot	Safety	Evaluation?
High	High	Small	High	Unacceptable
High	High	Small	Low	Unacceptable
Medium	High	Small	High	Acceptable
Low	Medium	Small	High	Acceptable
Low	Low	Big	High	Acceptable
Low	Low	Big	Low	Unacceptable
Medium	Low	Big	Low	Acceptable
High	Medium	Small	High	Unacceptable
High	Low	Big	High	Acceptable
Low	Medium	Big	High	Acceptable
High	Medium	Big	Low	Acceptable
Medium	Medium	Small	Low	Acceptable
Medium	High	Big	High	Acceptable
Low	Medium	Small	Low	Unacceptable

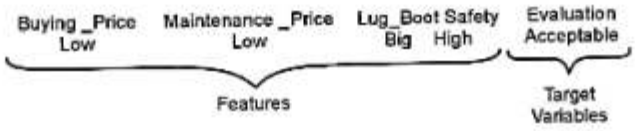


Fig. 1.2.1: Features and target variable identified

1.3 TYPES OF MACHINE LEARNING

Some of the main types of machine learning are :

1. Supervised Learning

In this type of learning we use data which is comprises of input and corresponding output. For every instance of data we can have input 'X' and corresponding output 'Y'. From this ML system will build model so that given an observation 'X', for new observation 'X' it will try to find out what is corresponding 'Y'. In supervised learning training data is labelled with the correct answers, e.g. "spam" or "ham." Two most important types of supervised learning are **classification** (where the outputs are discrete labels, as in spam filtering) and **regression** (where the outputsare real-valued).

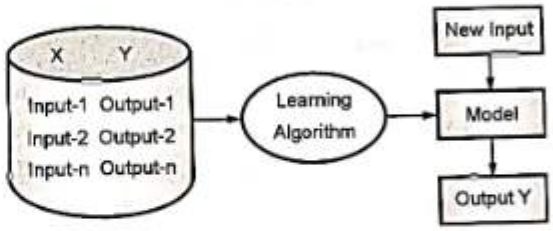


Fig. 1.3.1: Supervised Learning

2. Unsupervised learning

In unsupervised learning you are only given input 'X',there is no label to the data and given the data or different data points, you may want to form clusters or want to find some pattern. Two important unsupervised learning tasks are **dimensionreduction** and **clustering**.

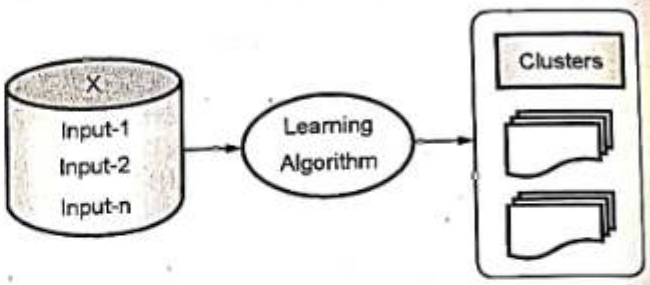


Fig. 1.3.2 : Unsupervised Learning

3. Reinforcement learning

In reinforcement learning you have an agent who is acting in an environment and you want to findout what action the agent must take based on the reward or penalty that the agent gets it.In this an agent (e.g., a robot or controller) seeks to learn theoptimal actions to take based the outcomes of past actions.

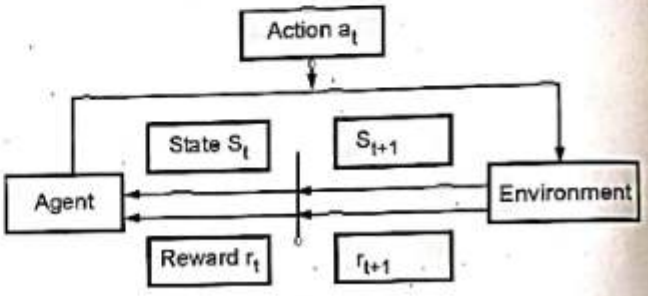


Fig. 1.3.3 : Reinforcement Learning

4. Semi-supervised learning

- It is a combination of supervised and unsupervised learning. In this there is some amount of labeled training data and also you have large amount of unlabeled data and you try to come up with some learning algorithm that convert even when training data is not labeled.
- In classification task, the aim is to predict class of an instance of data. Another method in machine learning is regression.
- Regression is the prediction of a numeric value. Regression's example is to draw a best fit line which passes through some data points in order to generalize the data points.
- Classification and regression are examples of supervised learning. These types of problems are called as supervised because we are asking the algorithm what to predict.
- The exact opposite of supervised is a task called as unsupervised learning. In unsupervised learning, target value or label is not given for the data.
- A problem in which similar items are grouped together is called as clustering. In unsupervised learning, we may also want to find statistical values that describe the data. This is called as density estimation.
- Another task of unsupervised learning may be reducing the huge amount of data from many attributes to a small number so that we can properly visualize it in two or three dimensions.

Table 1.3.1 : Supervised learning tasks

k-Nearest Neighbours	Linear
Naive Bayes	Locally weighted linear
Support Vector Machines	Ridge
Decision Trees	Lasso

Table 1.3.2 : Unsupervised learning tasks

k-Means	Expectation Maximization
DBSCAN	Parzen Window

1.4 ISSUES IN MACHINE LEARNING

1. Which algorithm we have to select to learn general target functions from specific training dataset? What should be the settings for particular algorithms, so as to converge to the desired function, given sufficient training data? Which algorithms perform best for which type of problems and representations?
2. How much training data is sufficient? What should be the general amount of datathat can be found to relate the confidence in learned hypotheses to the amount training experience and the character of the learner's hypothesis space?
3. Prior knowledge held by the learner is used at which time and manner to guide the process of generalizing from examples? If we have approximately correct knowledge, will it helpful even when it is only approximately correct?
4. What is the best strategy for choosing a useful next training experience, and how does the choice of this strategy after the complexity of the learning problem?
5. To reduce the task of learning to one or more function approximation problems, what will be the best approach? What specific functions should the system attempt to learn? Can this process itself be automated?
6. To improve the knowledge representation and to learn the target function, how can the learner automatically alter its representation?

1.5 HOW TO CHOOSE THE RIGHT ALGORITHM?

With all the different algorithms available in machine learning, how can you select which one to use? First you need to focus on your goal. What are you trying to get out of this? What data do you have or can you collect? Secondly you have to consider the data.

1. **Goal :** If you are trying to predict or forecast a target value, then you need to look into supervised learning. Otherwise, you have to use unsupervised learning.

   (a) If you have chosen **supervised** learning, then next you need to focus on what's your **target** value?

   If **target value** is **discrete** (e.g. Yes/ No, 1 /2/3, A/B/C), then use **Classification.**

   If **target value** is continuous i.e. Number of values (e.g. 0 – 100, – 99 to 99), then use **Regression.**

   (b) If you have chosen **unsupervised** learning, then next you need to focus onwhat is your **aim?**

If youwant to **fit your data** into some **discrete groups**, then use **Clustering.**

If you want to **find numerical estimate** of how strong the fit into each group, then use **density estimation algorithm.**

2. **Data :** Are the features continuous or nominal? Are there missing values in features? If yes, what is a reason for missing values? Are there outliers in the data? To narrow the algorithm selection process, all of these features of your data can help you.

Table 1.5.1 : Selection of Algorithm

	Supervised Learning	Unsupervised Learning
Discrete	Classification	Clustering
Continuous	Regression	Density Estimation

1.6 STEPS IN DEVELOPING A MACHINE LEARNING APPLICATION

1. Collection of Data

You could collect the samples from a website and extracting data.

- From RSS feed or an API
- From device to collect wind speed measurement
- Publicly available data.

2. Preparation of the input data

- Once you have the input data, you need to check whether it's in a useable format or not.
- Some algorithm can accept target variables and features as string; some need them to be integers.
- Some algorithm accepts features in a special format.

3. Analyse the input data

- Looking at the data you have passed in a text editor to checkcollection and preparation of input data steps are properly working and you don't have a bunch of empty values.
- You can also check at the data to find out if you can see any patterns or if there is anything obvious, such as a few data points greatly differ from remaining set of the data.
- Plotting data in 1, 2 or 3 dimensions can also help.
- Distil multiple dimensions down to 2/3 so that you can visualize the data.

4. The importance of this step is that it makes you understand that you don't have any garbage value coming in.

5. Train the algorithm

- Good clean data from the first two steps is given as input to the algorithm. The algorithm extracts information or knowledge. This knowledge is mostly stored in a format that is readily useable by machine for next 2 steps.
- In case of unsupervised learning, training step is not there because target value is not present. Complete data is used in the next step.

6. Test the algorithm

- In this step the information learned in the previous step is used. When you are checking an algorithm, you will test it to find out whether it works properly or not. In supervised case, you have some known values that can be used to evaluate the algorithm.
- In case of unsupervised, you may have to use some other matrices to evaluate the success. In either case, if you are not satisfied, you can again go back to step 4, change some things and test again.
- Mostly problem occurs in collection or preparation of data and you will have to go back to step 1.

7. Use it

In this step a real program is developed to do some task, and once again it is checked if all the previous steps worked as you expected. You might encounter some new data and have to revisit step 1-5.

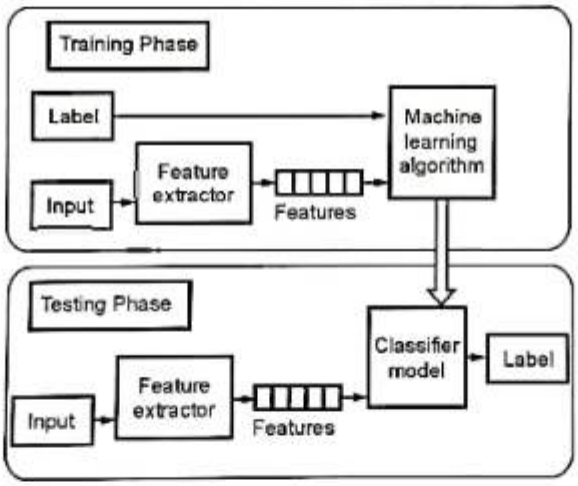


Fig. 1.6.1 : Typical example of Machine Learning Application

1.7 APPLICATIONS OF MACHINE LEARNING

1. Learning Associations

- A supermarket chain-one an example of retail application of machine learning is basket analysis, which is finding associations between products bought by customers:
- If people who buy P typically also buy Q and if there is a customer who buys Q and does not buy P, he or she is a potential P customer. Once we identify such customers, we can target them for cross-selling.

- In finding an association rule, we are interested in learning a conditional probability of the form P (Q/P) where Q is the product we would like to condition on P, which are the product / products which we know that customer has already purchased.

  P (Milk / Bread) = 0.7

- It implies that 70% of customers who buy bread also buy milk

2. Classification

- A credit is an amount of money loaned by a financial institution. It is important for the bank to be able to predict in advance the risk associated with a loan. Which is the probability that the customer will default and not pay the whole amount back?
- In credit scoring, the bank calculates the risk given the amount of credit and the information about the customer. (Income, savings, collaterals, profession, age, past financial history). The aim is to infer a general rule from this data, coding the association between a customer's attributes and his risk.
- Machine Learning system fits a model to the past data to be able to calculate the risk for a new application and then decides to accept or refuse it accordingly.

  If income > Q_1 and savings > Q_2

  Then low – risk ELSE high – risk
- Other classification examples are Optical character recognition, face recognition, medical diagnosis, speech recognition and biometric.

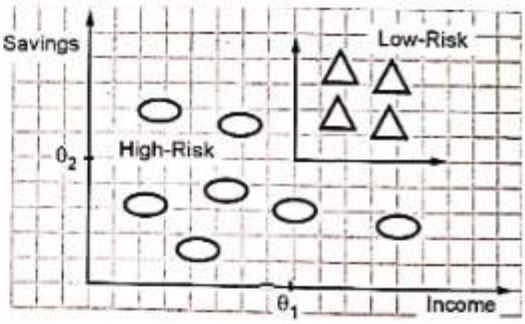


Fig. 1.7.1 : Classification for credit scoring

3. Regression

- Suppose we want to design a system that can predict the price of a flat. Let's take the inputs as the area of the flat, location and purchase year and other information that affects the rate of flat. The output is the price of the flat. The applications where output is numeric are regression problems.
- Let X represents flat features and Y is the price of flat. We can collect training data by surveying past purchased transactions and the Machine Learning algorithm fits a function to this data to learn Y as a function of X for the suitable values of W and W_0.

  $Y = w^*x + w_0$

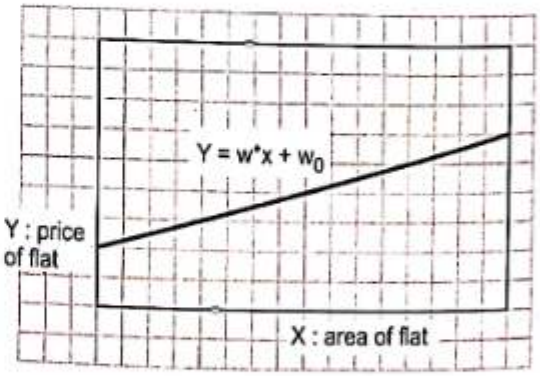


Fig. 1.7.2 : Regression for prediction of price of flat

4. Unsupervised Learning

- One of the important unsupervised learning problem is clustering. In clastering dataset is partitioned in to meaningful sub classes known as clusters. For example, suppose you want to decorate your home using given items.
- Now you will classify them using unsupervised learning (no prior knowledge) and this classification can be on the basis of color of items, shape of items, material used for items, type of items or whatever way you would like.

5. Reinforcement Learning

- There are some of the applications where output of system is a sequence of actions. In such applications the sequence of correct actions instead of single action is important in order to reach goal. An action is said to be good if it is part of good policy. Machine learning program generates a policy by learning previous good action sequences. Such methods are called reinforcement methods
- A good example of reinforcement learning is chess playing. In artificial intelligence and machine learning, one of the most important research area is game playing. Games can be are easily described but at the same time, they arequite difficult to play well. Let's take a example of chess that has limited number of rules, but the game is very difficult because for each statethere can be large number of possible moves.
- Another application of reinforcement learning is robot navigation. The robot can move in all possible directions at any point of time.The algorithm should reach goal state from an initial state by learning the correct sequence of actions after conducting number of trial runs.
- When the system has unreliable and partial sensory information, it makes reinforcement learning complex. Let's take an example of robot with incomplete camera information. Here robot does not know its exact location.

1.8 UNIVERSITY QUESTIONS AND ANSWERS

May 2015

Q.1 What are the issues in Machine learning ? *(Ans. : Refer section 1.4)* **(5 Marks)**

May 2016

Q.2 What are the key tasks of Machine Learning ? *(Ans. : Refer sections 1.2 and 1.3)* **(5 Marks)**

Q.3 Explain the steps required for selecting the right machine learning algorithm. *(Ans. : Refer section 1.5)* **(8 Marks)**

Q.4 Write short note on : Machine learning applications. *(Ans. : Refer section 1.7)* **(10 Marks)**

May 2017

Q.5 What is Machine learning ? Explain how supervised learning is different from unsupervised learning. *(Ans. : Refer sections 1.1 and 1.3)* **(5 Marks)**

Q.6 Write short note on : Machine learning applications. *(Ans. : Refer section 1.7)* **(10 Marks)**

May 2019

Q.7 What is Machine learning ? How it is different from data mining ? *(Ans. : Refer section 1.1).* **(5 Marks)**

Q. 8 Explain the steps of developing Machine Learning applications. *(Ans. : Refer section 1.6)* **(10 Marks)**

➲ **Dec. 2019**

Q. 9 Define Machine learning and explain with example importance of Machine Learning. **(5 Marks)**
(Ans. : Refer section 1.1)

Multiple Choice Questions

Q. 1.1 What is Machine Learning ?
(a) The autonomous acquisition of knowledge through the use of computer programs
(b) The selective acquisition of knowledge through the use of computer programs
(c) The autonomous acquisition of knowledge through the use of manual programs
(d) The selective acquisition of knowledge through the use of manual programs ✓**Ans. : (a)**

Explanation : Definition of machine learning- system automatically learns from previous experience using an algorithm.

Q. 1.2 Different learning methods does not include _____
(a) Memorization (b) Analogy
(c) Deduction (d) Introduction ✓**Ans. : (d)**

Explanation : Memory (previous experience), logic and inference (deduction) is required.

Q. 1.3 Which of these is not the type of machine Learning ?
(a) Supervised Learning
(b) Unsupervised Learning
(c) Reinforcement learning
(d) Semi-unsupervised Learning ✓**Ans. : (d)**

Explanation : Supervised, unsupervised, reinforcement and hybrid are types of learning.

Q. 1.4 Which of the following is not an application of learning ?
(a) Data mining (b) World wide web
(c) Speech recognition (d) Data manipulation
✓**Ans. : (d)**

Explanation : Data manipulation is data preprocessing.

Q. 1.5 Fraud Detection, Image Classification, Diagnostics are applications in _____
(a) Unsupervised Learning (b) Supervised Learning
(c) Reinforcement learning (d) Inductive learning
✓**Ans. : (d)**

Explanation : In inductive learning learner discovers rules by observing examples.

Q. 1.6 Car price prediction is an example of _____
(a) Supervised Learning
(b) Unsupervised Learning
(c) Active Learning
(d) Reinforcement learning ✓**Ans. : (a)**

Explanation : In car price predication car database is used to find relationship between car attributes and car price and using this system is trained.

Q. 1.7 What is the function of Unsupervised Learning?
(a) Find clusters of the data (b) Classifications
(c) Predict time series (d) Regression
✓**Ans. : (a)**

Explanation : In unsupervised based on common attributes grouping is done. Rest three options are examples of supervised learning.

Q. 1.8 In an Unsupervised learning _____
(a) Specific output values are given
(b) Specific output values are not given
(c) No specific Inputs are given
(d) Both inputs and outputs are given ✓**Ans. : (b)**

Explanation : In unsupervised learning since supervisor is not present, we do not have the idea about expected output. System learns only from the input.

Q. 1.9 Spam mail filtering is
(a) Classification problem
(b) Clustering problem
(c) Classification and Clustering Problem
(d) Time Series problem ✓**Ans. : (a)**

Explanation : In spam mail based on training examples classification is done. Training examples contain attributes of mail and output (spam or not).

Q. 1.10 Data used to optimize the parameter settings of a supervised learner model
(a) Training Data (b) Test Data
(c) Verification Data (d) Validation Data
✓**Ans. : (d)**

Explanation : Using validation data it is tested that whether the model is giving correct and optimized output or not.

Q. 1.11 In regression the output is
(a) Continuous (b) Discrete
(c) May be discrete or continuous
(d) Continuous and always lies in a finite range
✓Ans. : (a)
Explanation : In regression the output is numerical.

Q. 1.12 Machine Learning is a branch of _____
(a) Natural Language processing
(b) Artificial Intelligence
(c) Java (d) C ✓Ans. : (b)
Explanation : ML is a category of AI.

Q. 1.13 A shop owner has a store that stores a variety of fabrics. When fabric is brought to the store, various types of fabrics may be mixed together. The shop owner wants a model that will sort the fabric according to type. Which model will be efficient/accurate for this task?
(a) Machine learning model
(b) Feature based classification technique.
(c) Computer vision
(d) Fuzzy Logic ✓Ans. : (a)
Explanation : Machine learning model will be efficient since ones the model is trained can be used for futuristic use.

Q. 1.14 To Understand the role of machine learning in public health and safety and the cultural, societal, and environmental considerations in determining the non-functional requirements of products and processes. Which type of learning can be used?
(a) Supervised Learning
(b) Unsupervised Learning
(c) Competitive Learning
(d) Reinforcement Learning ✓Ans. : (a)
Explanation : Supervised learning as in this cases we know the expected output.

Q. 1.15 Suppose you want to design a system for waste management. In this first the garbage is collected and it is sent to the main server for analysis. Main server compares the categories of garbage and appropriate disposal method is selected. For comparison of garbage type you will use which Machine learning method ?
(a) Regression (b) Classification
(c) Clustering (d) Dimensionality Reduction
✓Ans. : (b)
Explanation : Classification, depending on the attributes waste will be segregated and also we know output.

Q. 1.16 You are given reviews of few movies marked as positive, negative or neutral. Classifying reviews of a new movie is an example of
(a) Supervised learning
(b) Unsupervised learning
(c) Semi supervised learning
(d) Reinforcement learning ✓Ans. : (a)
Explanation : Supervised learning as in this cases we know the expected output.

Q. 1.17 Imagine a newly born starts to learn walking.it will try to find a suitable policy to learn walking after repeated falling and getting up. Specify what type of MI algorithm is best suited to do the same.
(a) Supervised (b) Unsupervised
(c) Reinforcement (d) Semi supervised
✓Ans. : (c)
Explanation : In this case child learns using the concept of punishment (fall down) and reward (walk properly) which are the characteristics of reinforcement learning.

Q. 1.18 Automated vehicle is an example of _____
(a) Supervised Learning
(b) Unsupervised Learning
(c) Active Learning
(d) Reinforcement learning ✓Ans. : (a)
Explanation : Supervised learning as in this case we know the expected output.

Q. 1.19 Real-time decisions, Game AI, Learning task are applications in
(a) Active Learning
(b) Supervised Learning
(c) Reinforcement learning
(d) Unsupervised Learning ✓Ans. : (c)
Explanation : Reinforcement learning since in these cases system learns from punishment and reward.

Q. 1.20 In which of the following learning the teacher returns reward and punishment to learner ?
(a) Active Learning
(b) Reinforcement learning
(c) Unsupervised Learning
(d) Supervised Learning ✓Ans. : (b)
Explanation : Definition and working of Reinforcement learning.

Q. 1.21 Computational learning theory analyses the sample complexity and computational complexity of _____

(a) Unsupervised Learning (b) Inductive learning
(c) Forced based learning (d) Weak learning

✓Ans. : (b)

Explanation : Since in this system learns from observed examples.

Q. 1.22 Which of the following is an example of active learning?

(a) News Recommender system
(b) Dust cleaning machine
(c) Automated vehicle (d) Speech recognition

✓Ans. : (a)

Explanation : In news recommendation items are presented to the user to learn more about their preferences,there like or dislike where active learning comes in.

Q. 1.23 Which of the following is also called as exploratory learning?

(a) Supervised Learning
(b) Reinforcement learning
(c) Unsupervised Learning
(d) Active Learning ✓Ans. : (c)

Explanation : Since unsupervised learning identify structure within data.

Q. 1.24 Supervised Learning and Unsupervised Learning both require at least one?

(a) Hidden attribute (b) Output Attribute
(c) Input Attribute (d) Categorical Attribute

✓Ans. : (a)

Explanation : Since hidden attribute preserves the robustness of semantic attributes and inherits the discrimination ability of visual features.

Q. 1.25 What is the function of Supervised Learning?

(a) Grouping of data
(b) Find centroid of data
(c) Find relationship between input and output
(d) Learn from punishment and reward

✓Ans. : (c)

Explanation : In supervised based on relationship between input and output model is trained.

Q. 1.26 Which modifies the performance element so that it makes better decision?

(a) Performance element (b) Changing element
(c) Learning element (d) Hearing element

✓Ans. : (c)

Explanation : Learning element learns from new evidences so that performance can be improved.

Q. 1.27 Why is Goggle very successful in Machine Learning

(a) They have more data than other companies
(b) They have better algorithms
(c) They have better training sets
(d) They work on low level data features

✓Ans. : (a)

Explanation : For better generalisation and accuracy more training data is required.

Q. 1.28 Which of the following is not Machine learning disciplines

(a) Information theory (b) Neurostatistics
(c) Optimization (d) Physics ✓Ans. : (b)

Explanation : Game theory, control theory, operation research,information theory, optimization, swarm intelligence and genetic algorithm are disciplines of ML.

Q. 1.29 What are the three essential components of a learning system?

(a) Model, gradient descent, learning algorithm
(b) Error function, model, learning algorithm
(c) Accuracy, Sensitivity, Specificity
(d) Model, error function, cost function ✓Ans. : (b)

Explanation : Learning algorithm trains the model using error function.

Q. 1.30 You are reviewing papers for the World's Fanciest Machine Learning Conference, and you see submissions with the following claims. Which ones would you consider accepting?

(a) My method achieves a training error lower than all previous methods.
(b) My method achieves a test error lower than all previous methods.
(c) My method achieves a test error lower than all previous methods
(d) My method achieves a cross-validation error lower than all previous methods. ✓Ans. : (c)

Explanation : Test error lower than all previous methods when regularization parameter is chosen so as to minimize cross-validation error.

Q. 1.31 What is true about Machine Learning?

(a) Machine Learning (ML) is that field of computer science
(b) ML is a type of artificial intelligence that extracts patterns out of raw data by using an algorithm or method

(c) The main focus of ML is to allow computer systems learn from experience without being explicitly programmed or human intervention.

(d) All of the above ✓Ans. : (d)

Explanation : All statement are true about Machine Learning.

Q. 1.32 ML is a field of AI consisting of learning algorithms that?

(a) Improve their performance

(b) At executing some task

(c) Over time with experience

(d) All of the above ✓Ans. :(d)

Explanation : ML is a field of AI consisting of learning algorithms that : Improve their performance (P), At executing some task (T), Over time with experience (E).

Q. 1.33 Which of the following are ML methods?

(a) based on human supervision

(b) supervised Learning

(c) semi-reinforcement Learning

(d) All of the above ✓Ans. : (d)

Explanation : The following are various ML methods based on some broad categories : Based on human supervision, Unsupervised Learning, Semi-supervised Learning and Reinforcement Learning.

Q. 1.34 Which of the following is NOT supervised learning ?

(a) PCA (b) Decision Tree

(c) Linear Regression (d) Naive Bayesian

✓Ans. : (a)

Explanation : Principal Component Analysis (PCA) is not predictive analysis tool. It is a data pre-processing tool. It helps in picking out the most relevant linear combination of variables and use them in our predictive model.PCA is a technique for reducing the dimensionality of large datasets, increasing interpretability but at the same time minimizing information loss.

Q. 1.35 Which of the following is a good test dataset characteristic?

(a) Large enough to yield meaningful results

(b) Is representative of the dataset as a whole

(c) Both (a) and (b)

(d) None of the above ✓Ans. : (c)

Explanation : For better result more records as well as meaningful records are required.

Q. 1.36 How do you handle missing or corrupted data in a dataset ?

(a) Drop missing rows or columns

(b) Replace missing values with mean/median/mode

(c) Assign a unique category to missing values

(d) All of the above ✓Ans. : (d)

Explanation : All above methods are used to handle missing data.

Q. 1.37 When performing regression or classification, which of the following is the correct way to preprocess the data?

(a) Normalize the data → PCA → training

(b) PCA → normalize PCA output → training

(c) Normalize the data → PCA → normalize PCA output → training

(d) None of the above ✓Ans. : (a)

Explanation : First preprocessing is done then data reduction is done and finally training is performed.

Q. 1.38 What is unsupervised learning ?

(a) features of group explicitly stated

(b) number of groups may be known

(c) neither feature nor number of groups is known

(d) none of the mentioned ✓Ans. : (c)

Explanation : Basic definition of unsupervised learning.

Q. 1.39 What's the main point of difference between human and machine intelligence?

(a) human perceive everything as a pattern while machine perceive it merely as data

(b) human have emotions

(c) human have more IQ and intellect

(d) human have sense organs ✓Ans. : (a)

Explanation : Humans have emotions and thus form different patterns on that basis, while a machine(say computer) is dumb and everything is just a data for him.

Q. 1.40 Choose the options that are correct regarding machine learning (ML) and artificial intelligence (AI),

(a) ML is an alternate way of programming intelligent machines.

(b) All options are correct.

(c) ML is a set of techniques that turns a dataset into a software.

(d) AI is a software that can emulate the human mind.

Ans. : (b)

Explanation : Since all three options are correct.

Q. 1.41 Which of the following sentence is FALSE regarding regression?

(a) It relates inputs to outputs.

(b) It is used for prediction.

(c) It may be used for interpretation.

(d) It discovers causal relationships. ✓**Ans. : (d)**

Explanation : Regression method do not have ability to find causal relationships.

Q. 1.42 Which of the following is/are one of the important step(s) to pre-process the text in NLP based projects?

1. Stemming 2. Stop word removal

3. Object standardization

(a) 1 and 2 (b) 1 and 3

(c) 2 and 3 (d) 1,2 and 3 ✓**Ans. : (d)**

Explanation : Stemming is a rudimentary rule-based process of stripping the suffixes ("ing", "ly", "es", "s" etc) from a word. Stop words are those words which will have not relevant to the context of the data for example is/am/are.Object Standardization is also one of the good way to pre-process the text.

Q. 1.43 Which of the factors affect the performance of learner system does not include?

(a) Representation scheme used

(b) Training scenario

(c) Type of feedback

(d) Good data structures ✓**Ans. :(d)**

Explanation : Factors that affect the performance of learner system does not include good data structures.

Q. 1.44 The model will be trained with data in one single batch is known as _____

(a) batch learning (b) offline learning

(c) both(a) and (b) (d) none of above

✓**Ans. : (c)**

Explanation : In both methods data is trained in a single batch.

Q. 1.45 In general, to have a well-defined learning problem, we must identity which of the following :

(a) The class of tasks

(b) The measure of performance to be improved

(c) The source of experience

(d) All of the above ✓**Ans. : (d)**

Explanation : Since all components are necessary to define a learning problem.

Q. 1.46 Successful applications of ML _____

(a) Learning to recognize spoken words

(b) Learning to drive an autonomous vehicle

(c) Learning to classify new astronomical structures

(d) All of the above ✓**Ans. : (d)**

Explanation : All are applications of ML.

Q. 1.47 Designing a machine learning approach involves _____

(a) Choosing the type of training experience

(b) Choosing the target function to be learned

(c) Choosing a representation for the target function

(d) All of the above ✓**Ans. : (d)**

Explanation : Since all actions are necessary to design a ML system.

Q. 1.48 Real-Time decisions, Game AI, Learning Tasks, Skill Acquisition, and Robot Navigation are applications of which of the following

(a) Supervised Learning : Classification

(b) Reinforcement Learning

(c) Unsupervised Learning : Clustering

(d) Unsupervised Learning : Regression ✓**Ans. :(b)**

Explanation : Since these applications learns using punishment and reward.

Q. 1.49 Targetted marketing, Recommended Systems, and Customer Segmentation are applications in which of the following

(a) Supervised Learning: Classification

(b) Unsupervised Learning: Clustering

(c) Unsupervised Learning: Regression

(d) Reinforcement Learning ✓**Ans. : (b)**

Explanation : Sine in this applications data is grouped together based on common characteristics.

Q. 1.50 If we want output as numeric and we have the idea about expected output which method we should use?

(a) Classification (b) Regression

(c) Clustering (d) Density estimation

✓**Ans. : b**

Explanation : Basic definition of regression.

ChapterEnds...

□□□

Module 2

Chapter...2

Introduction to Neural Network

University Prescribed Syllabus

Introduction - Fundamental concept - Evolution of Neural Networks - Biological Neuron, Artificial Neural Networks, NN architecture, Activation functions, McCulloch-Pitts Model.

2.1 INTRODUCTION

- An Artificial Neural Network (ANN) is inspired from the Biological Nervous System. The way by which Biological Nervous System such as Brain processes the information in the same manner ANN also processes the information. ANN is also called as information processing paradigm. It resembles the brain in two respects :
  - Learning process is used to acquire the knowledge from the Environment by the network.
  - Acquired Knowledge is stored using Interneuron connection strengths known as synaptic weights.
- The Information processing system is composed of a large number of highly interconnected processing elements (neurons) which works together to solve specific problems.
- In Biological System learning process involves adjustments to the synaptic connections that exist between the neurons, in the same manner learning is carried out in ANN.
- ANN can be used in many Applications. Pattern extraction and detection of trends is a tedious process for humans and other computer techniques. Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends.
- One of the applications of ANN includes, Adaptive learning in this the data is given for training and the network uses this data to learn how to perform the tasks based on this data. ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability.
- Other applications of ANN include Signal Processing, Pattern Recognition, Speech Recognition, Data Compression, Computer Vision and Game Playing.

2.2 BIOLOGICAL NEURON

- Neuron is the elementary nerve cell and a basic unit of the nervous system. Neuron has an information processing ability.
- Each neuron has three main regions, cell body or Soma, Axon and Dendrites. Soma contains the nucleus and it processes the information. Axon is a long fiber that serves as a transmission line. End part of the Axon splits into fine arborization that ends into small bulb called as a Synapse almost touching the dendrite of the neighboring neuron.

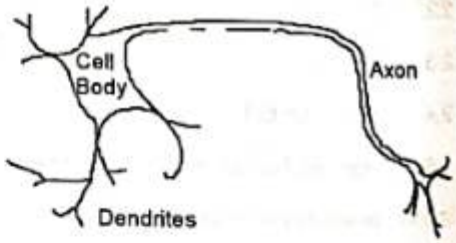


Fig. 2.2.1 : Biological Neuron Model

- Dendrites accept the input from the neighboring neuron through axon. Dendrites, look like a tree structure, receives signals from other neurons. Synapse is the electro-chemical contact between the organs. They do not physically touch because they are separated by a cleft. The electric signals are sent through chemical interaction. The neuron sending the signal is called pre-synaptic cell and the neuron receiving the electrical signal is called postsynaptic cell.
- The electrical signals that the neurons use to convey the information of the brain are all identical. The brain can determine which type of information is being received based on the path of the signal. The brain analyzes all patterns of signals sent, and from that information it interprets the type of information received. There are different types of Biological neurons. When the neurons are classified by the processes they carry out they are classified as unipolar neurons, bipolar neurons and multipolar neurons.

- Unipolar neurons have a single process. Their dendrites and axon are located on the same stem. These neurons are found in invertebrates. Bipolar neurons have two processes. Their dendrites and axon have two separated processes too. Multipolar neurons are commonly found in mammals. Some examples of these neurons are spinal motor neurons, pyramidal cells and purkinje cells.
- When biological neurons are classified by function they fall into three categories. The first group is sensory neurons. These neurons provide all information for perception and motor coordination. The second group provides information to muscles, and glands. There are called motor neurons. The last group, the interneuronal, contains all other neurons and has two subclasses. One group called relay or protection interneuron. They are usually found in the brain and connect different parts of it. The other group called local interneuron's are only used in local circuits.

2.3 BASIC ANN MODEL/ MCCULLOCH – PITTS MODEL

McCulloch and Pitts proposed a computational model that resembles the Biological Neuron model. These neurons were represented as models of biological networks into conceptual components for circuits that could perform computational tasks. The basic model of the artificial neuron is founded upon the functionality of the biological neuron. An artificial neuron is a mathematical function that resembles the biological neuron.

Neuron Model

- A neuron with a scalar input and no bias appears below. Fig. 2.3.1 shows a simple artificial neural net with n input neurons (X_1, X_2..., X_n) and one output neuron (Y). The interconnected weights are given by W_1, W_2, and W_n.

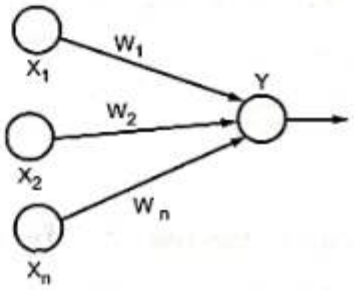


Fig. 2.3.1 : Artificial Neuron Model

- The output of the above model is given as,

$Y = 1$ if $net_i = \Sigma W_{ij} * X_j \geq T$

$= 0$ if $net_i = \Sigma W_{ij} * X_j < T$

Where 'i' represent the output neuron and 'j' represent the input neuron.

- The scalar input X is transmitted through a connection that multiplies its strength by the scalar weight W to form the product W * X, again a scalar. The weighted input W * X is the only argument of the transfer function f, which produces the scalar output Y. The neuron may have a scalar bias, b. You can view the bias as simply being added to the product W * X. The bias is much like a weight, except that it has a constant input of 1.
- The transfer function net input 'n', again a scalar, is the sum of the weighted input W * X and the bias b. This sum is the argument of the transfer function f. Here f is a transfer function, typically a step function or a sigmoid function, that takes the argument n and produces the output Y. Note that W and b are both adjustable scalar parameters of the neuron.
- The central idea of neural networks is that such parameters can be adjusted so that the network exhibits some desired or interesting behaviour. Thus, you can train the network to do a particular job by adjusting the weight or bias parameters, or perhaps the network itself will adjust these parameters to achieve some desired end.
- As previously noted, the bias b is an adjustable (scalar) parameter of the neuron. It is not an input. However, the constant 1 that drives the bias is an input and must be treated as such when you consider the linear dependence of input vectors in Linear Filters.

Example 1 : Simulation of NOT gate using McCulloch - Pitts Model

The truth table of the NOT gate is as follows

Input	Output
X	Y
1	0
0	1

We assume the weight vector as W.

For the first row (i.e. input 1) we may write net value as $W * X = W * 1 = W$ according to the McCulloch - Pitts model if the output is 0 then net value must be less than threshold $W < T$.

For the second row (i.e. input 0) we may write net value as $W * X = W * 0 = 0$ according to the McCulloch - Pitts model if the output is 1 then net value must be greater than or equal to threshold, $0 \geq T$.

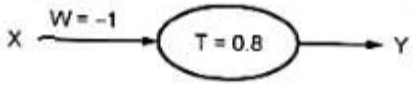
Now we are having two equations

1. $W < T$ 2. $0 \geq T$

Now select the values of W and T such that the above conditions gets satisfied.

One of the possible values are $T = 0.8$, $W = -1$.

Using this values the NOT gate is represented as,



Example 2 : Simulation of AND gate using McCulloch - Pitts Model

The truth table of the AND gate is as follows :

Input		Output
X_1	X_2	Y
0	0	0
0	1	0
1	0	0
1	1	1

We assume the weight vector as W_1 for X_1 and W_2 for X_2.

For the first row, we may write net values as $(W_1 * X_1) + (W_2 * X_2) = (W_1 * 0) + (W_2 * 0) = 0$.

According to the McCulloch - Pitts model if the output is 0 then net value must be less than threshold $0 < T$.

For the second row, we may write net values as $(W_1 * X_1) + (W_2 * X_2) = (W_1 * 0) + (W_2 * 1) = W_2$.

According to the McCulloch - Pitts model if the output is 0 then net value must be less than threshold $W_2 < T$.

For the third row, we may write net values as $(W_1 * X_1) + (W_2 * X_2) = (W_1 * 1) + (W_2 * 0) = W_1$.

According to the McCulloch - Pitts model if the output is 0 then net value must be less than threshold $W_1 < T$.

For the fourth row, we may write net values as $(W_1 * X_1) + (W_2 * X_2) = (W_1 * 1) + (W_2 * 1) = W_1 + W_2$.

According to the McCulloch - Pitts model if the output is 1 then net value must be less than threshold $W_1 + W_2 \geq T$.

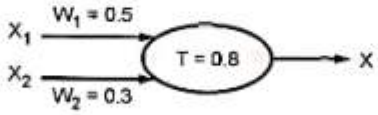
Now we are having four equations :

1. $0 < T$ 2. $W_2 < T$ 3. $W_1 < T$ 4. $W_1 + W_2 \geq T$

Now select the values of W_1, W_2, and T such that the above conditions get satisfied.

Let's assume $T = 0.8$, $W_2 = 0.3$, $W_1 = 0.5$.

Using this values the AND gate is represented as,



Module 2

2.4 DIFFERENCE BETWEEN BIOLOGICAL NEURON AND ARTIFICIAL NEURON

Sr. No.	Points of Difference	Biological NN	Artificial NN
1.	Processing elements	10^{14} synapses	10^{8} transistors
2.	Speed	Slow	Fast
3.	Processing	Parallel Execution	One by one
4.	Size and Complexity of operation	Less	More, difficult to implement complex operation
5.	Fault Tolerance	Exist	Doesn't exist
6.	Storage	If new data is added old is not erased	Erased
7.	Control Mechanism	Every neuron acts independently	CPU

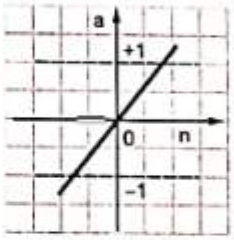
2.5 TYPES OF ACTIVATION FUNCTIONS/ TRANSFER FUNCTIONS

Activation function f (x) is used to give output of a neuron in terms of a local field x or net. The various activation functions are follows.

1. Linear

Output = net

Linear activation function gives the output same as that of the input or net value. The Matlab toolbox has a function, purelin, to realize the mathematical linear transfer function shown above. Neurons of this type are used as linear approximators in Linear Filters



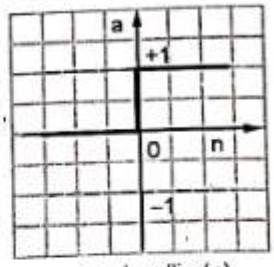
a = purelin(n)

2. Hard limit / Unipolar binary

Output = 0 if net < 0

= 1 if net ≥ 0

The hard-limit transfer function shown above limits the output of the neuron to either 0, if the net input argument n is less than 0, or 1, if n is greater than or equal to 0. This function is used in Perceptron's, to create neurons that make classification decisions.



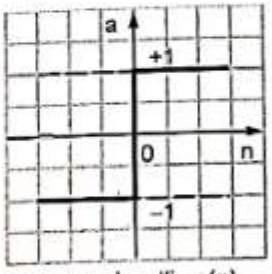
a = hardlim(n)

Hard-Limit Transfer Function

3. Symmetrical Hard limit/Bipolar binary

$$\text{Output} = -1 \quad \text{if net} < 0$$
$$= 1 \quad \text{if net} \geq 0$$

The Symmetrical hard-limit transfer function shown above limits the output of the neuron to either – 1, if the net input argument n is less than 0, or 1, if n is greater than or equal to 0. This function is used in Perceptron's s, to create neurons that make classification decisions.

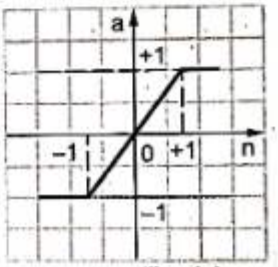


a = hardlims(n)

Symmetric Hard-Limit Transfer Function

4. Saturating linear

$$\text{Output} = 0 \quad \text{if net} < 0$$
$$= \text{net} \quad \text{if } 0 \leq \text{net} < 1$$
$$= 1 \quad \text{if net} \geq 1$$



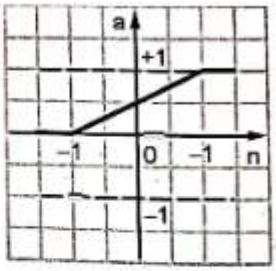
a = satlins (n)

Satlins Transfer Function

5. Symmetrical saturating linear

$$\text{Output} = -1 \quad \text{if net} < 0$$
$$= \text{net} \quad \text{if } 0 \leq \text{net} < 1$$
$$= 1 \quad \text{if net} \geq 1$$

The symbol in the square to the right of each transfer function graph shown above represents the associated transfer function. These icons replace the general f in the boxes of network diagrams to show the particular transfer function being used.



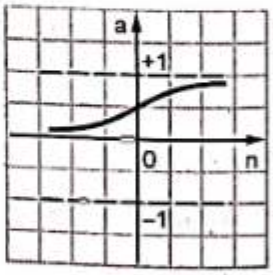
a = satlins(n)

Satlins Transfer Function

6. Unipolar continuous

$$\text{Output} = \frac{1}{1 + e^{-\lambda \text{ net}}}$$

The sigmoid transfer function shown above takes the input, which can have any value between plus and minus infinity, and squashes the output into the range 0 to 1. This transfer function is commonly used in back propagation networks, in part because it is differentiable.



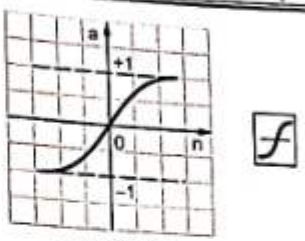
a = logsig(n)

Log-sigmoid Transfer Function

7. Tansig / Bipolar continuous

$$\text{Output} = \frac{2}{1 + e^{-\lambda net}} - 1$$

The sigmoid transfer function shown above takes the input, which can have any value between plus and minus infinity, and squashes the output into the range – 1 to 1.



a = tansig(n)
Tan-sigmoid Transfer Function

2.6 NEURAL NETWORK ARCHITECTURE

Module 2

Neurons are interconnected to each other. The arrangement of neurons is called as Network architectures.

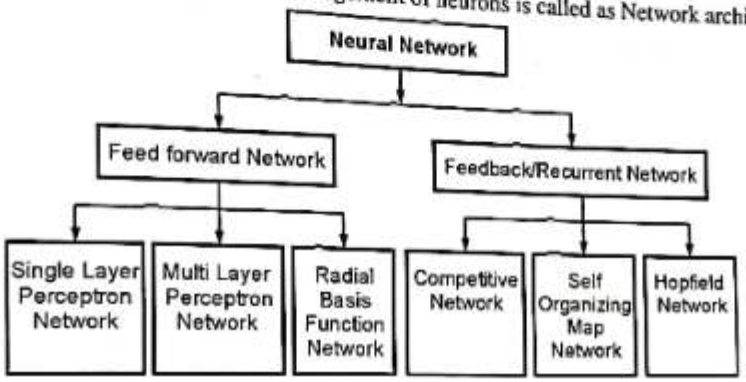


Fig. 2.6.1 : Network Architecture

1. Feed forward Network

- In this type of network the neurons are connected in a forward direction. Connection is allowed from layer i to layer j if and only if, $i < j$
- Input vector, output vector and the weight matrix of the network are represented as follows,

Input vector, $X = [X_1, X_2, X_3 \ldots\ldots\ldots\ldots X_n]$

Output vector, $Y = [Y_1, Y_2, Y_3 \ldots\ldots\ldots\ldots Y_m]$

Weight matrix, $W = \begin{bmatrix} W_{11} & W_{12} & W_{13} & \cdot & \cdot & W_{1n} \\ W_{21} & W_{22} & W_{23} & \cdot & \cdot & W_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ W_{m1} & W_{m2} & W_{m3} & \cdot & \cdot & W_{mn} \end{bmatrix}$

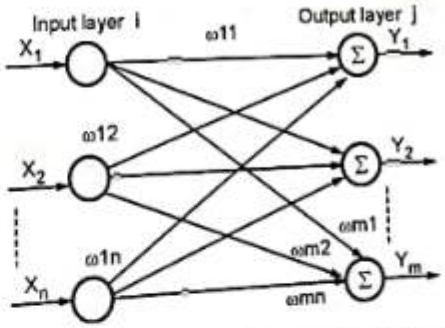


Fig. 2.6.2 : Single Layer Feed forward Network

- The net input to the j^{th} is calculated using the following equation,

$$net_j = \Sigma W_{ji} * X_i \quad \text{for } j = 1 \text{ to } m \text{ and } i = 1 \text{ to } n$$

- Once the net input is calculated, the output of the j^{th} neuron is calculated by applying the activation function on the net input as follows,

$$Y_j = f(net_j)$$

1.1 Single Layer Perceptron Network

- Fig. 2.6.2 represents the single layer Perceptron network. In this type of network there are only input and output layers are present. It consists of single layer, where the inputs are directly connected to the outputs, via a serried of weights.
- The sum of the products of the weights and the inputs is calculated in each neuron node, and if the value is above some threshold (generally, 0) the neuron fires and displays the output (generally, 1) otherwise inhibit the value (generally, – 1).

1.2 Multi Layer Perceptron Network

This type of network besides having the input and output layers also have one or more hidden layers in between input and output layers.

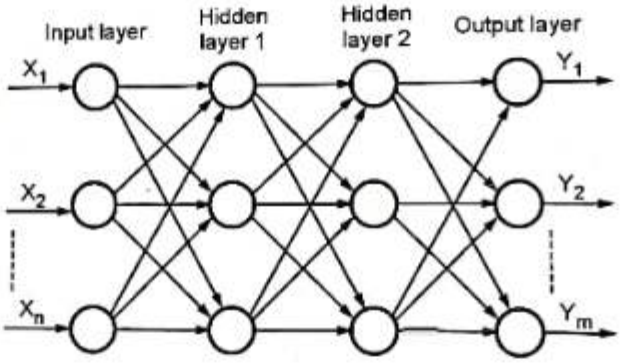


Fig. 2.6.3 : Multi Layer Perceptron Network

1.3 Radial Basis Function Network

In this type of networks a single hidden layer is present.

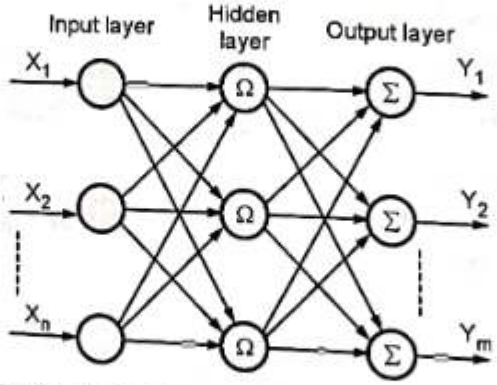


Fig. 2.6.4 : Radial Connected neural network

2. Feedback /Recurrent Network

- Feedback network can be obtained from the feed forward n/w by connecting its output to input. The inputs are applied at the initial instance then the input is removed and the network remains autonomous thereafter (t > 0).
- The recurrent networks differ from feed-forward architecture. A recurrent network has at least one feed back loop.

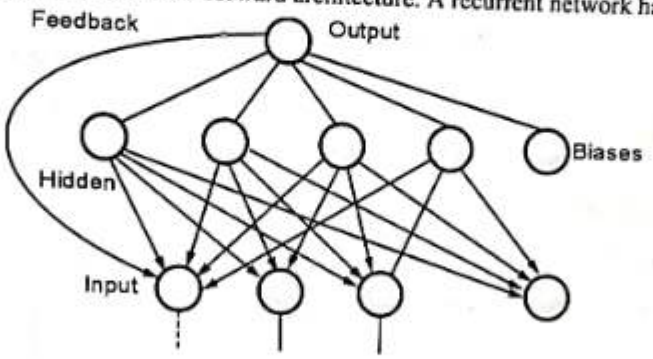


Fig. 2.6.5 : Feedback / Recurrent Network

2.1 Competitive networks

In this type of network the neurons of the output layers compete between themselves to find the maximum output.

2.2 Self Organising Map (SOM)

The input neurons activate the closest output neuron.

2.3 Hopfield Network

Each neuron is connected to every other neuron but not back to itself.

2.7 TYPES OF LEARNING

Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is true of ANNs as well; Learning of neural network means setting or updating the weights.

1. Supervised Learning

In this type of learning when input is applied supervisor provides a desired response. The difference between the actual response (o) and the desired response (d) is calculated called as error measure which is used to correct the network parameters.

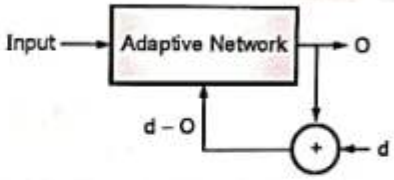


Fig. 2.7.1 : Supervised Learning

2. Unsupervised Learning

In this type of learning supervisor is not present due to this there is no idea or guess of output. Network modifies it's weights based on patterns of input and/or output.

Fig. 2.7.2 : Unsupervised Learning

3. Hybrid Learning

In this type a combination of supervised and unsupervised learning is used.

4. Competitive Learning

In this the output neurons compete between themselves. The neuron having the maximum response is declared as a winner neuron and the weights of winner neurons are modified else remains unchanged.

Example 2.7.1 : A Single layer NN is to be designed with 6 input and 2 output. The outputs are to be limited to and continuous over the range 0 to 1. Answer the following questions.

1. How many neurons are required ?
2. What are the dimensions of weight matrix ?
3. What kind of transfer function should be used ?

☑ **Solution :**

For each input and output one neuron is required, hence total 8 neurons are required.

In weight matrix the number of row reprints the output neurons and number of columns represent the input neurons. The first row reprints all the incoming vectors of the first output neuron and the second row represent all the incoming vectors of the second output neuron. Hence the matrix dimension is 2×6 as follows :

$$W = \begin{bmatrix} W_{11} & W_{12} & W_{13} & W_{14} & W_{15} & W_{16} \\ W_{21} & W_{22} & W_{23} & W_{24} & W_{25} & W_{26} \end{bmatrix}$$

The outputs are to be limited to and continuous over the range 0 to 1. Hence we may use unipolar continuous function.

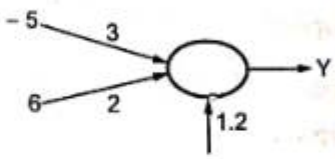
$$f(net) = \frac{1}{(1 + \exp(-\lambda * net))}$$

Example 2.7.2 : Given a 2 input neuron with following parameters, b = 1.2, w = [3, 2], x = [– 5, 6]. Calculate neuron's output for following transfer functions.

1. Hard limit
2. Symmetrical Hard limit
3. Linear
4. Saturating linear
5. Symmetrical saturating linear
6. Unipolar continuous
7. Bipolar continuous

☑ **Solution :**

First we calculate the net value which can be calculated as follows :



$$net_i = \sum_{j=1}^{n} W_{ij} X_j$$

But, in this particular problem bias value is also given, Bias is a value which is added in the net to improve the performance of the network so we will use the following formula

$$net_i = \sum_{j=1}^{n} W_{ij} X_j + b$$

$$net = (-5 * 3) + (6 * 2) + 1.2 = -1.8$$

Now we will calculate the final output for the given activation functions

1. Hard limit

$$\text{Output} = 0 \quad \text{if. net} < 0$$
$$= 1 \quad \text{if. net} \geq 0$$
$$\text{Hence, } Y = 0$$

2. Symmetrical Hard limit

$$\text{Output} = -1 \quad \text{if. net} < 0$$
$$= 1 \quad \text{if. net} \geq 0$$
$$\text{Hence, } Y = -1$$

Module 2

3. Linear

$$\text{Output} = \text{net}$$
$$\text{Hence, } Y = -0.8$$

4. Saturating linear

$$\text{Output} = 0 \quad \text{if. net} < 0$$
$$= \text{net} \quad \text{if. } 0 \leq \text{net} < 1$$
$$= 1 \quad \text{if. net} \geq 1$$
$$\text{Hence, } Y = 0$$

5. Symmetrical saturating linear

$$\text{Output} = -1 \quad \text{if. net} < 0$$
$$= \text{net} \quad \text{if. } 0 \leq \text{net} < 1$$
$$= 1 \quad \text{if. net} \geq 1$$
$$\text{Hence, } Y = -1$$

6. Unipolar continuous

$$\text{Output} = \frac{1}{(1 + \exp(-\lambda * \text{net}))}$$

We assume the value of λ equal to 1(If the value of λ is not given then assume the standard value as equal to 1)

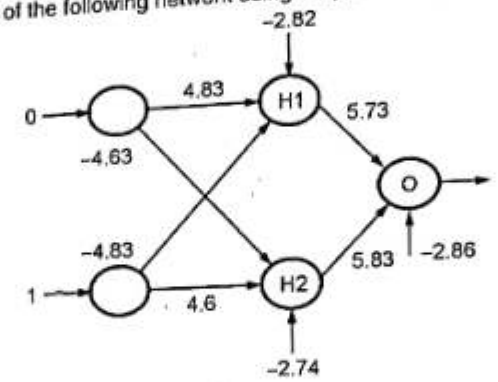
Hence, Y = 0.1418

7. Bipolar continuous

$$\text{Output} = \frac{2}{(1 + \exp(-\lambda * \text{net}))} - 1$$

$$\text{Hence, } Y = -0.71$$

Example 2.7.3 : Compute the output of the following network using Unipolar continuous.



☑ **Solution :** First we will calculate net input and output of hidden nodes,

$$H1_{(net)} = (4.83 * 0) + (-4.83 * 1) - 2.82 = -7.65$$

$$H1_{(output)} = 4.758 \times 10^{-4}$$

$$H2_{(net)} = (-4.63 * 0) + (4.6 * 1) - 2.74 = 1.86$$

$$H2_{(output)} = 0.865$$

Now we will calculate net input and output of output node,

$$O_{(net)} = (4.758 \times 10^{-4} * 5.73) + (0.865 * 5.83) - 2.86 = 2.167$$

$$O_{(output)} = 0.899$$

2.8 PERCEPTRON LEARNING RULE

Perceptron Learning is a supervised type of learning as the desired response is present. It is applicable only for Binary types of neurons (activation functions). Learning signal is the difference between the actual output and desired output of neuron and it is used to update the weight.

Learning signal, $r = d_i - o_i$

Where o_i is the output of the i^{th} neuron and d_i is the desired response.

Weight Increment, $\Delta W_{ij} = C * [d_i - o_i] * X_j$

Where C is a constant, X is input and j = 1 to n.

$$W_{new} = W_{old} + \Delta Wij$$

In the Perceptron learning weights are updated only if, $d_i \neq o_i$

Let's assume

if, $d_i = 1$ and $o_i = -1$

then $\Delta W_{ij} = C^* [d_i - o_i] * X_j = C^* [1 - (-1)] * X_j = 2CX_j$

if $d_i = -1$ and $o_i = 1$

then $\Delta W_{ij} = C^* [d_i - o_i] * X_j = C^* [-1 - 1] * X_j = -2CX_j$

Hence, $\Delta W_{ij} = \pm 2 C X_j$

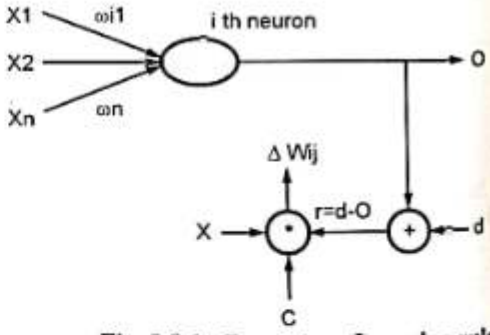


Fig. 2.8.1 : Perceptron Learning rule

Example 2.8.1 : Initial weight vector W_1 needs to be trained using three input vectors, X_1, X_2 and X_3 and their desired responses. Find final weight vector using Perceptron learning rule.

$$W_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0.5 \end{bmatrix}, X_1 = \begin{bmatrix} 1 \\ -2 \\ 0 \\ -1 \end{bmatrix}, X_2 = \begin{bmatrix} 0 \\ 1.5 \\ -0.5 \\ -1 \end{bmatrix}, X_3 = \begin{bmatrix} -1 \\ 1 \\ 0.5 \\ -1 \end{bmatrix}, C = 0.1, d_1 = -1, d_2 = -1, d_3 = 1$$

☑ **Solution :**

Perceptron learning is applicable only for binary function and also in this problem desired responses are given in 1 and – 1 format Hence we solve this problem using bipolar binary function

Module 2

Bipolar Binary

For bipolar binary output $= -1$ if net < 0

$= 1$ if net ≥ 0

Step 1 : When X_0 is applied

$$net_1 = W_1^t * X_1 = [1 \; -1 \; 0 \; 0.5] \begin{bmatrix} 1 \\ -2 \\ 0 \\ -1 \end{bmatrix} = 2.5$$

$$o_1 = f(net_1) = 1 \qquad \text{as } d_1 \neq O_1$$

$$W_2 = W_1 + C * [d_1 - o_1] * X_1$$

$$= \begin{bmatrix} 1 \\ -1 \\ 0 \\ 0.5 \end{bmatrix} + 0.1 (-1 - 1) \begin{bmatrix} 1 \\ -2 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0.8 \\ -0.6 \\ 0 \\ 0.7 \end{bmatrix}$$

Step 2 : When X_2 is applied

$$net_2 = W_2^t * X_2 = [0.8 \; -0.6 \; 0 \; 0.7] \begin{bmatrix} 0 \\ 1.5 \\ -0.5 \\ -1 \end{bmatrix} = -1.6$$

$$o_2 = f(net_2) = -1 \qquad \text{as net} < 0$$

Since, $d_2 = o_2$, weight updation is not required

So, $W_3 = W_2$

Step 3 : When X_3 is applied

$$net_3 = W_3^t * X_3 = [0.8 \; -0.6 \; 0 \; 0.7] \begin{bmatrix} -1 \\ 1 \\ 0.5 \\ -1 \end{bmatrix} = -2.1$$

$$o_3 = f(net_3) = -1 \qquad \text{as } d_3 \neq o_3$$

$$W_4 = W_3 + C * [d_3 - o_3] * X_3$$

$$= \begin{bmatrix} 0.8 \\ -0.6 \\ 0 \\ 0.7 \end{bmatrix} + 1 * (1 - (-1)) \begin{bmatrix} -1 \\ 1 \\ 0.5 \\ -1 \end{bmatrix} = \begin{bmatrix} 0.6 \\ -0.4 \\ 0.1 \\ 0.5 \end{bmatrix}$$

Example 2.8.2 : Implement Perceptron Learning rule.

$$W_1 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, X_1 = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}, X_2 = \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix}, C = 1, d_1 = -1, d_2 = 1$$

Repeat training sequence until two correct responses in a row are achieved.

☑ **Solution :**

Perceptron learning is applicable only for binary function and also in this problem desired responses are given in 1 and −1 format Hence we solve this problem using bipolar binary function

Bipolar Binary

For bipolar binary output $= -1$ if $net < 0$

$= 1$ if $net \geq 0$

Step 1 : When X_1 is applied

$$net_1 = W_1^t * X_1 = [0 \ 1 \ 0] \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = 1$$

$o_1 = f(net_1) = 1$ as $d_1 \neq o_1$

$$W_2 = W_1 + C * [d_1 - o_1] * X_1$$

$$= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 1(-1-1) + \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -4 \\ -1 \\ 2 \end{bmatrix}$$

Step 2 : When X_2 is applied

$$net_2 = W_2^t * X_2 = [-4 \ -1 \ 2] \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix} = -1$$

$o_2 = f(net_2) = -1$ as $net < 0$ $d_2 \neq o_2$

$$W_3 = W_2 + C * [d_2 - o_2] * X_2$$

$$= \begin{bmatrix} -4 \\ -1 \\ 2 \end{bmatrix} + 1(1 - (-1)) \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix} = \begin{bmatrix} -4 \\ -3 \\ 0 \end{bmatrix}$$

In the problem it is given that we have to repeat the training until two correct responses are achieved, so we will again apply X_1.

Step 3 : When X_1 is applied

$$net_3 = W_3^t * X_1 = [-4 \ -3 \ 0] \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix} = -11$$

$$o_3 = f(net_3) = -1$$

As $d_1 = o_3$, weight updation is not required

$$W_4 = W_3$$

Step 4 : When X_2 is applied

$$net_4 = W_4^t * X_2 = [-4 \ -3 \ 0] \begin{bmatrix} 0 \\ -1 \\ -1 \end{bmatrix} = 3$$

$$o_4 = f(net_4) = 1$$

As $d_2 = o_4$, weight updation is not required

$$W_5 = W_4$$

Thus we have obtained the correct response in a row two times.

Example 2.8.3 : A Single neuron network using f (net) = sgn (net) has been trained using the pairs of X_i, d_i as follows. Find Initial weight vector.

$$W_4 = \begin{bmatrix} 3 \\ 2 \\ 6 \\ 1 \end{bmatrix}, X_1 = \begin{bmatrix} 1 \\ -2 \\ 3 \\ -1 \end{bmatrix}, X_2 = \begin{bmatrix} 0 \\ -1 \\ 2 \\ -1 \end{bmatrix}, X_3 = \begin{bmatrix} -2 \\ 0 \\ -3 \\ -1 \end{bmatrix}, C = 1, d_1 = -1, d_2 = 1, d_3 = -1$$

☑ **Solution :** This Problem is different from the previous problems; we have to find the initial weight vector using the final weight vector

Step1 :

$$W_4 = W_3 + \Delta W_3$$

The above equation can be written as, $W_3 = W_4 - \Delta W_3$

As we know $\Delta W_3 = \pm 2 * C * X_3$ and $d_3 = -1$, we consider the negative sign

$$\Delta W_3 = -2 * C * X_3 = \begin{bmatrix} 4 \\ 0 \\ 6 \\ 2 \end{bmatrix}$$

$$W_3 = W_4 - \Delta W_3 = \begin{bmatrix} -1 \\ 2 \\ 0 \\ -1 \end{bmatrix}$$

Step 2 :

$$W_3 = W_2 + \Delta W_2$$

The above equation can be written as, $W_2 = W_3 - \Delta W_2$

As we know $\Delta W_2 = \pm 2 * C * X_2$ and $d_2 = 1$ we consider the positive sign

$$\Delta W_2 = 2 * C * X_2 = \begin{bmatrix} 0 \\ -2 \\ 4 \\ -2 \end{bmatrix}$$

$$W_2 = W_3 - \Delta W_2 = \begin{bmatrix} -1 \\ 4 \\ -4 \\ 1 \end{bmatrix}$$

Step 3 :

$$W_2 = W_1 + \Delta W_1$$

The above equation can be written as, $W_1 = W_2 - \Delta W_1$

As we know $\Delta W_1 = \pm 2 * C * X_1$ and $d_1 = -1$ we consider the negative sign

$$\Delta W_1 = -2 * C * X_1 = \begin{bmatrix} -2 \\ 4 \\ -6 \\ 2 \end{bmatrix}$$

$$W_1 = W_2 - \Delta W_1 = \begin{bmatrix} 1 \\ 0 \\ 2 \\ -1 \end{bmatrix}$$

2.9 SINGLE LAYER PERCEPTRON LEARNING

Perceptron Architecture

One type of NN system is based on the "Perceptron". A Perceptron computes a sum of weighted. Combination of its inputs, if the sum is greater than a certain threshold, then its output is 1 else – 1.

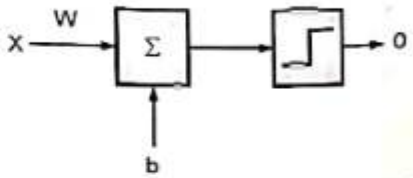


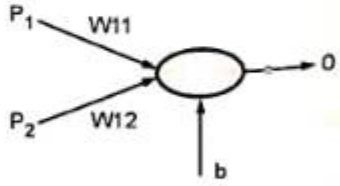
Fig. 2.9.1 : Perceptron Architecture

Output of the neuron is 1/0 or 1/– 1, thus each neuron in the network divides the input space into two regions. This is useful to determine the boundary between these regions.

Let's see the example for this.

Output of the above network is given by,

$$O = \text{hardlim}\,(W_{11} * P_1 + W_{12} * P_2 + b)$$



Input vector for which the net input is zero determines the decision boundary

$W_{11} * P_1 + W_{12} * P_2 + b = 0$

Let's take the value of $W_{11} = W_{12} = 1$ and $b = -1$ and substitute this values in the above equation we will get

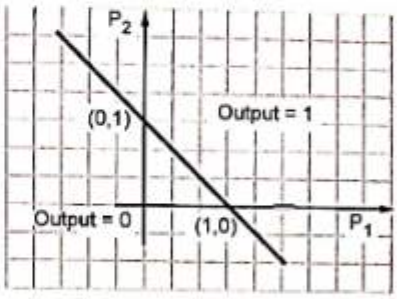
$$P_1 + P_2 - 1 = 0$$

To draw the decision boundary we need to find the intercepting points of P_1 and P_2 axes

$$P_1 + P_2 - 1 = 0$$

Substitute $P_1 = 0$ then we will get $P_2 = 1$ i.e. (0, 1)

Substitute $P_2 = 0$ then we will get $P_1 = 1$ i.e. (1, 0)



Now to find which decision region belongs to output = 1.

Let's pick one point (2, 0) and substitute in the following equation

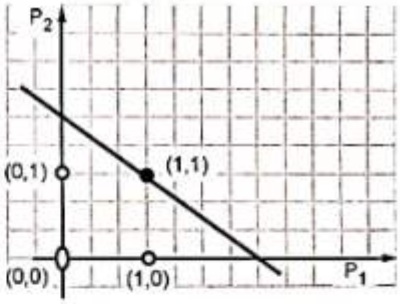
$$O = \text{Hardlim}(W^t P + b) = \text{Hardlim}\left(\begin{bmatrix} 2 \\ 0 \end{bmatrix} [1\ 1] + (-1)\right) = \text{Hardlim}(1) = 1$$

Decision boundary is always orthogonal to the weight vector and it always points towards the region where neuron output is 1.

Example 2.9.1 : Implement Perceptron Network for AND function Using the concept of decision Boundary.
$P_1 = [0\ 0]$, $t_1 = 0$ and $P_2 = [0\ 1]$, $t_2 = 0$ and $P_3 = [1\ 0]$, $t_3 = 0$ and $P_4 = [1\ 1]$, $t_4 = 1$

☑ **Solution :**

Now first we plot the given points. For the P_4 point the target or desired response is given as 1 this is represented as filled circle and for the remaining points it's 0 which is represented as empty circle. After plotting the points the next step is to draw the decision boundary such that it will divide the points into two regions according to their desired responses (i.e. output 1 and 0).



As we know weight vector is orthogonal to the decision boundary and it points towards the region where neuron output is 1 (in this case for P_4 (1, 1) output is 1) Hence we will take $W = [2\ 2]$

To find the value of bias pick one point on decision boundary as $P = [1.5\ 0]$.

By substituting this values in the equation $W^t P + b = 0$ we will get $b = -3$.

We test the network for our calculated values as follows.

Let's take P_1 to test

$$O = \text{Hardlim}(W^t P + b) = \text{Hardlim}\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix} [0\ 0] + (-3)\right) = \text{Hardlim}(-3) = 0$$

Example 2.9.2 : Solve the following classification problem with Perceptron learning rule. Apply each input vector in order for as many repetitions as it takes to ensure that the problem is solved. Draw a graph of the problem only after you found a solution.

$$W_1 = [0\ 0],\ b_1 = 0,\ P_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix} t_1 = 0,\ P_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix} t_2 = 1,\ P_3 = \begin{bmatrix} -2 \\ 2 \end{bmatrix} t_3 = 0,\ P_4 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} t_4 = 1,\ C = 1$$

In this problem the desired responses are given in 1/0 format so we will solve this problem using unipolar function.

☑ **Solution :**

Iteration 1

- **Step 1 : When P_1 is applied**

$$o_1 = \text{hardlim}(W_1 * P_1 + b_1) = \text{hardlim}(0) = 1$$

Error, $e = t_1 - o_1 = 0 - 1 = -1$

$$W_2 = W_1 + e * P_1^t = [0\ 0] + (-1)\ [2\ 2] = [-2\ -2]$$

$$b_2 = b_1 + e = 0 + (-1) = -1$$

- **Step 2 : When P_2 is applied**

$$o_2 = \text{hardlim}(W_2 * P_2 + b_2) = \text{hardlim}(1) = 1$$

Error, $e = t_2 - o_2 = 1 - 1 = 0$

Since the error is 0, weight and bias updation is not required.

$$W_3 = W_2$$

$$b_3 = b_2$$

- **Step 3 : When P_3 is applied**

$$o_3 = \text{hardlim}(W_3 * P_3 + b_3) = \text{hardlim}(-1) = 0$$

Error, $e = t_3 - o_3 = 0 - 0 = 0$

Since the error is 0, weight and bias updation is not required.

$$W_4 = W_3$$

$$b_4 = b_3$$

- **Step 4 : When P_4 is applied**

$$o_4 = \text{hardlim}(W_4 * P_4 + b_4) = \text{hardlim}(-1) = 0$$

Error, $e = t_4 - o_4 = 1 - 0 = 1$

$$W_5 = W_4 + e\,P_4^t = [-2\ -2] + 1\ [-1\ \ 1] = [-3\ -1]$$

$$b_5 = b_4 + e = -1 + 1 = 0$$

We have to repeat the iterations until all input vectors are correctly classified (i.e. error = 0 for all the input vectors)

Iteration 2

- **Step 5 : When P_1 is applied**

$$o_5 = \text{hardlim}(W_5 * P_1 + b_5) = \text{hardlim}(-8) = 0$$

Error, $e = t_1 - o_5 = 0 - 0 = 0$

Since the error is 0, weight and bias updation is not required.

$$W_6 = W_5$$

$$b_6 = b_5$$

- **Step 6 : When P_2 is applied**

$$o_6 = \text{hardlim}(W_6 * P_2 + b_6) = \text{hardlim}(-1) = 0$$

Error, $e = t_2 - o_4 = 1 - 0 = 1$

$$W_7 = W_6 + e\,P_2^{'} = [-3\ -1] + 1\,[1\ -2] = [-2\ -3]$$

$$b_7 = b_6 + e = 0 + 1 = 1$$

- **Step 7 : When P_3 is applied**

$$o_7 = \text{hardlim}(W_7 * P_3 + b_7) = \text{hardlim}(-1) = 0$$

Error, $e = t_3 - o_7 = 0$

Since the error is 0, weight and bias updation is not required.

$$W_8 = W_7$$

$$b_8 = b_7$$

- **Step 8 : When P_4 is applied**

$$o_8 = \text{hardlim}(W_8 * P_4 + b_8) = \text{hardlim}(0) = 0$$

Error, $e = t_4 - o_8 = 0$

Since the error is 0, weight and bias updation is not required.

$$W_9 = W_8$$

$$b_9 = b_8$$

In this iteration for P_1, P_3 and P_4, $e = 0$ Hence we will again go for Iteration 3

Iteration 3

- **Step 9 : When P_1 is applied**

$$o_9 = \text{hardlim}(W_9 * P_1 + b_9) = \text{hardlim}(-9) = 0$$

Error, $e = t_1 - o_9 = 0$

Since the error is 0, weight and bias updation is not required.

$$W_{10} = W_9$$

Module 2

$$b_{10} = b_9$$

Step 10 : When P_2 is applied

$$o_{10} = \text{hardlim}\,(W_{10} * P_2 + b_{10}) = \text{hardlim}\,(5) = 1$$

Error, $e = t_2 - o_{10} = 0$

Since the error is 0, weight and bias updation is not required.

$$W_{11} = W_{10}$$

$$b_{11} = b_{10}$$

Now for P_2 also we are getting the error, e = 0. In this iteration e = 0 for all input vectors. Thus, we can say that we have found a solution

Final weight vector and bias value are as follows

$$W = [-2 \; -3] \; b = 1$$

Now we substitute these values in the following equation to find the equation of decision boundary

$$W_{11} * P_1 + W_{12} * P_2 + b = 0$$

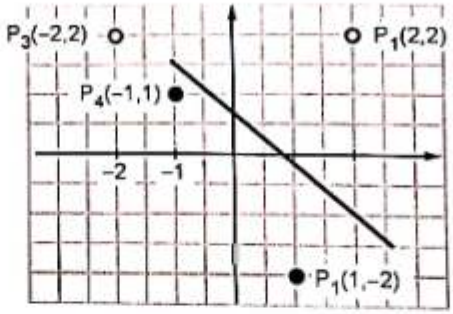
$$-2P_1 - 3P_2 + 1 = 0$$

Now to draw the decision boundary we need to find the intercepting points.

By substituting $P_1 = 0$ we will get $P_2 = 0.3$ i.e. point1 (0, 0.3)

By substituting $P_2 = 0$ we will get $P_1 = 0.5$ i.e. point2 (0.5, 0)

- From the above diagram we can say that the decision boundary classifies the input vectors in to two classes corresponding to output 1 and 0.



2.10 UNIVERSITY QUESTIONS AND ANSWERS

May 2017

Q. 1 Explain hard limit and soft limit activation function. *(Ans. : Refer section 2.5)* **(5 Marks)**

Q. 2 Explain Mc Culloch Pitts neuron model with the help of an example. *(Ans. : Refer section 2.3)* **(5 Marks)**

Q. 3 What is learning in neural networks ? Differentiate between supervised and unsupervised learning. *(Ans. : Refer sections 2.6 and 2.7)* **(10 Marks)**

➲ May 2019

Q. 4 Determine weights and threshold for the given data using McCulloch-Pitts neuron model. **(5 Marks)**
Plot all data points and show separating hyper-plane. *(Ans. : Refer section 2.3)*

Input		Output
X_1	X_2	Y
0	0	0
0	1	0
1	0	1
1	1	0

NOTE : For XOR use (X1 AND (NOT X2)) OR ((NOT X1) AND X2)

➲ Dec. 2019

Q. 5 Write short note on McCulloch-Pitts Neuron Model. *(Ans. : Refer section 2.3)* **(5 Marks)**

Multiple Choice Questions

Q. 2.1 Biological networks are superior to AI networks because of _____
(a) Robustness and fault tolerance
(b) Collective computation
(c) Flexibility
(d) Correctness ✓ **Ans. : (a)**
Explanation : BNN are more robust and also fault tolerance (Information does not lost completely) exist.

Q. 2.2 Soma in a Biological neural network maps to _____ in artificial neural network
(a) Input (b) Node
(c) Output (d) Interconnections ✓ **Ans. : (b)**
Explanation : Node collects all inputs for processing same as that of soma in BNN.

Q. 2.3 _____ are tree-like branches, responsible for receiving the information from other neurons it is connected to
(a) Soma (b) Axon
(c) Dendrites (d) Synapse ✓ **Ans. : (c)**
Explanations: Dendrite is the part of neuron that accepts the input.

Q. 2.4 _____ is just like a cable through which neurons send the information.
(a) Axon (b) Dendrites
(c) Soma (d) Synapse ✓ **Ans. : (a)**
Explanation : Axon is a part of neuron that send the output.

Q. 2.5 _____ is a non-recurrent network having processing units/nodes in layers and all the nodes in a layer are connected with the nodes of the next layers.
(a) Feedback network
(b) Recurrent networks
(c) Feed forward Network
(d) Fully Recurrent networks ✓ **Ans. : (c)**
Explanation : In feed forward network connections are allowed only in forward direction.

Q. 2.6 During the training of ANN under _____ learning, the input vectors of similar type are combined to form clusters
(a) Supervised (b) Unsupervised
(c) Reinforcement
(d) Both Supervised and Unsupervised ✓ **Ans. : (b)**
Explanation : Data is groped together based on common characteristics in unsupervised learning.

Q. 2.7 The maximum time involved in case of layer calculation is in ____
(a) Input layer computation.
(b) Output layer computation.
(c) Hidden layer
(d) Equal effort in each layer ✓ **Ans. : (c)**
Explanation : Main processing is done in hidden layers only.

Q. 2.8 Which one is type of linear activation function ?
(a) Identity (b) Unipolar Continuous
(c) Bipolar Continuous (d) Binary
✓ **Ans. : (a)**
Explanation : In identity input and output both are same.

Q. 2.9 Artificial Neural Network is used in _____

(a) Unsupervised learning model

(b) Supervised learning model

(c) In both Unsupervised and Supervised learning Model

(d) Neither used in Unsupervised nor in Supervised learning Model ✓ **Ans. : (c)**

Explanation : ANN applications are developed using both type of learning methods.

Q. 2.10 Training set of data in supervised learning includes

(a) Input (b) Output

(c) Both input and Output

(d) Neither input nor output ✓ **Ans. : (c)**

Explanation : In training data input and output both are present which is used to learn.

Q. 2.11 _____ is the connection between the axon and other neuron dendrites.

(a) Soma (b) Axon

(c) Dendrites (d) Synapse ✓ **Ans. : (d)**

Explanation : Synapse is the electro chemical contact organ.

Q. 2.12 Which is true for neural networks.

1) It has set of nodes and connections.
2) Each node computes it's weighted input
3) They have the ability to learn by example.
4) The training time does not depends on the size of the network.

(a) 1, 3, 4 (b) 1, 2, 3

(c) 2, 3 ,4 (d) 1, 2, 3, 4 ✓ **Ans. : (b)**

Explanation : Since training time depends on the size of the network.

Q. 2.13 The following Gate cannot be modelled with a single neuron

(a) 3-input AND Gate (b) 3-input XOR Gate

(c) Not Gate

(d) All can be easily modeled ✓ **Ans. : (b)**

Explanation : Since to design XOR gate we require AND, NOT and OR gate.

Q. 2.14 Steps followed in training a perception are listed below, What is the correct sequence of the steps?

1. For a sample input, compute an output
2. Initialize weights of perception randomly
3. Go to the next batch of dataset
4. If the prediction does not match the output, change the weighs

(a) 1, 4, 3, 2 (b) 2, 1, 4, 3

(c) 1, 2, 3, 4 (d) 2, 3, 4, 1 ✓ **Ans. : (b)**

Explanation : In perception output is calculated using input and initialized weights, if output is not same as that of expected output then weights are updated and then next iteration is performed.

Q. 2.15 Processing of ANN depends upon

1) Network Topology
2) Adjustments of Weights or Learning
3) Activation Functions

(a) 1, 2 (b) 2, 3

(c) 1, 3 (d) 1, 2, 3 ✓ **Ans. : (d)**

Explanation : Since all components are used for processing.

Q. 2.16 Suppose you have to design a system where you want to perform word prediction also known as language modelling You are to take output from previous state and also the input at each step to predict the next word The inputs at each step are the words for which the next words are to be predicted. Can we use Recurrent neural network for the design?

(a) Yes (b) No ✓ **Ans. : (a)**

Explanation : Yes, as in case of RNN output is again given to input unless and until we get desired result.

Q. 2.17 Suppose you want to predict the cyber bullying so that the parents can run this system in the background and when the children's are watching any video/audio/site if it contains any offended contents then site is blocked or contents are blurred. According to you which method will give you best result?

(a) Long Short Term Memory

(b) Convolution Neural Network

(c) Recurrent Neural Network

(d) Artificial Neural Network ✓ **Ans. : (a)**

Explanation : LSTM, as audio or video is to be converted in to sentence model then will be given to network.

Q. 2.18 Can you represent the following Boolean function with a single logistic threshold unit?

A	B	F(A,B)
1	1	0
0	0	0
1	0	1
0	1	0

(a) Yes (b) no ✓ **Ans. : (a)**

Explanation : Yes, you can represent this function with a single logistic threshold unit, since it is linearly separable.

Q. 2.19 Consider the neural network below. Find the appropriate weights for w_0, w_1 and w_2 to represent the AND function. Threshold function = {1, if output > 0; 0 otherwise}. x_0 and x_1 are the inputs and $b_1 = 1$ is the bias.

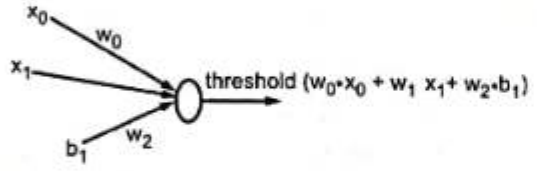


Fig. Q. 2.18

(a) $w_0 = 1, w_1 = 1, w_2 = 1$

(b) $w_0 = 1, w_1 = 1, w_2 = -1$

(c) $w_0 = -1, w_1 = -1, w_2 = -1$

(d) $w_0 = 2, w_1 = -2, w_3 = -1$ ✓ **Ans. : (b)**

Explanation : For $x_0 = 1$, $x_1 = 1$ and $b_1 = 1$,option (b) gives $1*1 + 1*1 + -1 = 1 > 0 = 1$

Q. 2.20 Which of the combination of weights make the network represent OR function ?

(a) $w_0 = 1, w_1 = 1, w_2 = 0$

(b) $w_0 = 1, w_2 = 1, w_3 = 1$

(c) $w_0 = 1, w_1 = 1, w_2 = -1$

(d) $w_0 = -1, w_1 = -1, w_2 = -1$ ✓ **Ans. : (a)**

Explanation : For $x_0 = 1$, $x_1 = 1$ and $b_1 = 1$, option (a) gives $1*1 + 1*1 + 0 = 1 > 0 = 1$

Q. 2.21 Suppose you are to design a system where you want to perform word prediction also known as _____

1) language modelling You are to take output from previous state and also the input at each step to
2) predict the next word. The inputs at each step are the words for which the next words are to be
3) predicted. Which of the following neural network would you use?

(a) Multi Layer Perception

(b) Recurrent Neural Network

(c) Convolution Neural Network

(d) Perception. ✓ **Ans. : (b)**

Explanation : Recurrent Neural Network (RNN) are a type of Neural Network where the output from previous step are fed as input to the current step. Refer to lecture notes for detailed explanation.

Q. 2.22 You are given the task of predicting the price of a house given the various features for a house such as number of rooms, area (sqft), etc. How many neurons should you have at the output ?

(a) 3 (b) 2 (c) 1 (d) 4 ✓ **Ans. : (c)**

Explanation : The price of a house is a single value. Hence, one neuron is enough.

Q. 2.23 Which one of the following activation functions can be used as an activation function at the output layer for the task given in Question 22 ?

(a) Sigmoid (b) Relu

(c) Softmax (d) Identity ✓ **Ans. : (b)**

Explanation : Since we don't have a particular range of value to predict at the output, we can use the Reluactivation function for the best interpretability of the result. Relu(x) = max(0,x) also gives us the same value for positive outputs.

Q. 2.24 Which of the following is true?

(i) Neural networks learn by example

(ii) On average, neural networks have lower computation rates than conventional computers.

(iii) Neural networks imitate the way the human brain works

(a) (i) and (ii) are true (b) (i),(ii) ,(iii) are true

(c) Only (i) and (iii) are true (iv) Only (ii) is true

✓ **Ans. : (b)**

Explanation : All statements are true for ANN.

Q. 2.25 McCulloch Pits model is a form of _____

(a) Unsupervised learning model

(b) Supervised learning model

(c) Stochastic learning Model

(d) Reinforcement learning Model ✓ **Ans. : (b)**

Explanation : McCulloch Pitts model works on basis of supervised learning.

Q. 2.26 In a neural network, which one of the following techniques is NOT useful to reduce overfitting ?

(a) Dropout (b) Regularization

(c) Batch normalizatioh (d) Adding more layers

✓ **Ans. : (d)**

Explanation : Adding more layers does not help to reduce overfitting.

Q. 2.27 One of a possible neuron specification requires a minimum of how many neurons to solve AND problem

(a) Two Neuron (b) Single Neuron

(c) Three neuron (d) Four Neuron

✓ **Ans. : (b)**

Explanation : Single neuron with two inputs and one output is required.

Module 2

Q. 2.28 Those neurons which responds strongly to the stimuli have their weights updated in the following Learning Method.

(a) Competitive Learning
(b) Stochastic Learning
(c) Hebbian Learning
(d) Gradient Descent Learning ✓ **Ans. : (a)**

Explanation : In competitive learning maximum output is considered.

Q. 2.29 We have decided to use a neural network to solve this problem. We have two choices : either to train a separate neural network for each of the diseases or to train a single neural network with one output neuron for each disease, but with a shared hidden layer. Which method do you prefer? (There is dependencies between diseases.)

(a) First (b) Second
(c) Both (d) None ✓ **Ans. : (a)**

Explanation :

1 Neural network with a shared hidden layer can capture dependencies between diseases. It can be shown that in some cases, when there is a dependency between the output nodes, having a shared node in the hidden layer can improve the accuracy.

2 If there is no dependency between diseases (output neurons), then we would prefer to have a separate neural network for each disease.

Q. 2.30 Let's say, you are using activation function X in hidden layers of neural network. At a particular neuron for any given input, you get the output as "– 0.0001". Which of the following activation function could X represent?

(a) ReLU (b) tanh
(c) SIGMOID (d) None of these ✓ **Ans. : (b)**

Explanation : The function is a tanh because the this function output range is between (-1,-1).

Q. 2.31 Why do we need biological neural networks ?

(a) to solve tasks like machine vision and natural language processing
(b) to apply heuristic search methods to find solutions of problem
(c) to make smart human interactive and user friendly system
(d) all of the mentioned ✓ **Ans. : (d)**

Explanation : These are the basic aims that a neural network achieve.

Q. 2.32 What is auto-association task in neural networks ?

(a) find relation between 2 consecutive inputs
(b) related to storage and recall task
(c) predicting the future inputs
(d) none of the mentioned ✓ **Ans. : (b)**

Explanation : This is the basic definition of auto-association in neural network.

Q. 2.33 When the cell is said to be fired ?

(a) if potential of body reaches a steady threshold values
(b) if there is impulse reaction
(c) during upbeat of heart
(d) none of the mentioned ✓ **Ans. : (a)**

Explanation : Cell is said to be fired if and only if potential of body reaches a certain steady threshold values.

Q. 2.34 The cell body of neuron can be analogous to what mathematical operation?

(a) summing (b) differentiator
(c) integrator (d) none of the mentioned ✓ **Ans. : (a)**

Explanation : Because adding of potential(due to neural fluid) at different parts of neuron is the reason of its firing.

Q. 2.35 What is hebb's rule of learning ?

(a) the system learns from its past mistakes
(b) the system recalls previous reference inputs and respective ideal outputs
(c) the strength of neural connection get modified accordingly
(d) none of the mentioned ✓ **Ans. : (c)**

Explanation : The strength of neuron to fire in future increases, if it is fired repeatedly.

Q. 2.36 What is the feature of ANNs due to which they can deal with noisy, fuzzy, inconsistent data?

(a) associative nature of networks
(b) distributive nature of networks
(c) both associative and distributive
(d) none of the mentioned ✓ **Ans. : (c)**

Explanation : General characteristics of ANNs.

Q. 2.37 What was the name of the first model which can perform weighted sum of inputs?

(a) McCulloch-pitts neuron model
(b) Marvin Minsky neuron model
(c) Hopfield model of neuron
(d) None of the mentioned ✓ **Ans. : (a)**

Explanation : McCulloch-pitts neuron model can perform weighted sum of inputs followed by threshold logic operation.

Q. 2.38 Who developed the first learning machine in which connection strengths could be adapted automatically ?

(a) McCulloch-pitts (b) Marvin Minsky

(c) Hopfield (d) none of the mentioned

✓ **Ans. : (b)**

Explanation : In 1954 Marvin Minsky developed the first learning machine in which connection strengths could be adapted automatically and efficiently.

Q. 2.39 What is an activation value ?

(a) weighted sum of inputs (b) threshold value

(c) main input to neuron (d) none of the mentioned

✓ **Ans. : (a)**

Explanation : It is definition of activation value and is basic qanda.

Q. 2.40 The process of adjusting the weight is known as _____

(a) activation (b) synchronization

(c) learning (d) none of the mentioned

✓ **Ans. : (c)**

Explanation : Basic definition of learning in neural nets.

Q. 2.41 Which of the following model has ability to learn?

(a) pitts model

(b) rosenblatt perception model

(c) both rosenblatt and pitts model

(d) neither rosenblatt nor pitts ✓ **Ans. : (b)**

Explanation : Weights are fixed in pitts model but adjustable in rosenblatt.

Q. 2.42 When both inputs are 1, what will be the output of the pitts model nand gate ?

(a) 0 (b) 1

(c) either 0 or 1 (d) z ✓ **Ans. : (a)**

Explanation : According to the truth table of a nand gate.

Q. 2.43 Connections across the layers in standard topologies and among the units within a layer can be organised?

(a) in feedforward manner

(b) in feedback manner

(c) both feedforward and feedback

(d) either feedforward and feedback ✓ **Ans. : (d)**

Explanation : Connections across the layers in standard topologies can be in feedforward manner or in feedback manner but not both.

Q. 2.44 If the change in weight vector is represented by Δw_{ij}, what does it mean ?

(a) describes the change in weight vector for ith processing unit, taking input vector jth into account

(b) describes the change in weight vector for jth processing unit, taking input vector ith into account

(c) describes the change in weight vector for jth and ith processing unit.

(d) none of the mentioned ✓ **Ans. : (a)**

Explanation : According to weight updation formula.

Module 2

Q. 2.45 What are models in neural networks ?

(a) mathematical representation of our understanding

(b) representation of biological neural networks

(c) both way

(d) none of the mentioned ✓ **Ans. : (c)**

Explanation : Model should be close to our biological neural systems, so that we can have high efficiency in machines too.

Q. 2.46 What is competitive learning ?

(a) learning laws which modulate difference between synaptic weight and output signal

(b) learning laws which modulate difference between synaptic weight and activation value

(c) learning laws which modulate difference between actual output and desired output

(d) none of the mentioned ✓ **Ans. : (a)**

Explanation : Competitive learning laws modulate difference between synaptic weight and output signal.

Q. 2.47 What are the advantages of neural networks over conventional computers?

(i) They have the ability to learn

(ii) They are more fault tolerant

(iii) They are more suited for real time operation due to their high computational power

(a) (i) and (ii) (b) (i) and (iii)

(c) Only (i) (d) All ✓ **Ans. : (d)**

Explanation : all statements are true for NN.

Q. 2.48 Which of the following gives non-linearity to a neural network ?

(a) Gradient descent (b) Bias

(c) Relu Activation Function (d) None Correct

✓ **Ans. : (c)**

Explanation : An activation function such as Relu gives a non-linearity to the neural network.

Q. 2.49 For a fully-connected network with one hidden layer, increasing the number of hidden units should have what effect on bias and variance?

(a) Decrease bias, increase variance

(b) Increase bias, increase variance

(c) Increase bias, decrease variance

(d) No change Correct

✓ **Ans. : (a)**

Explanation : Adding more hidden units should decrease bias and increase variance. In general, more complicated models will result in lower bias but larger variance, and adding more hidden units certainly makes the model more complex.

Q. 2.50 Error back propagation uses which learning rule?

(a) Hebbian learning (b) Perception learning

(c) Delta learning (d) Competitive learning

✓ **Ans. : (c**

Explanation : Delta learning rule is used to upda weights of hidden and output layers with the help of cos function.

Chapter Ends

□□□

Module 3

Chapter...3

Introduction to Optimization Techniques

University Prescribed Syllabus

Derivative based optimization - Steepest Descent, Newton method.
Derivative free optimization - Random Search, Down Hill Simplex.

3.1 INTRODUCTION

- Optimization Methods are used to minimize the scalar function of a number of variables. In Unconstrained optimization, the variables are not restricted by inequalities or equality relationships. Scalar function which we want to minimize is called as an objective function.
- The objective function is an error function for feed forward network and energy function for recurrent network. For the further topics we need the terms, Hessian matrix and gradient vector, so first we will see the basic concept of these two terms.
- Gradient Vector - Let's take a scalar function E(x) of a vectorial variable x, defined as an n-elements column vector x is denoted as follows,

$$X = [X_1\ X_2 \ldots\ldots\ldots X_n]^t$$

- The gradient vector of E(X) with respect to column vector x is denoted as $\partial E(X)$

$$\partial E(X) = [dE/dX_1 \quad dE/dX_2 \ldots\ldots\ldots dE/dX_n]$$

- Hessian Matrix – For a scalar function E(x), a matrix of second derivatives called the Hessian matrix is defined as follows

$$\partial^2 E(X) = \partial [\partial E(X)]$$

$$\partial^2 E(X) = \begin{bmatrix} \frac{d^2E}{dx_1^2} & \frac{d^2E}{dx_1\,dx_2} & \cdots & \cdots & \frac{d^2E}{dx_1\,dx_n} \\ \frac{d^2E}{dx_2\,dx_1} & \frac{d^2E}{dx_2^2} & \cdots & \cdots & \frac{d^2E}{dx_2\,dx_n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{d^2E}{dx_n\,dx_1} & \frac{d^2E}{dx_n\,dx_2} & \cdots & \cdots & \frac{d^2E}{dx_n^2} \end{bmatrix}$$

This Hessian matrix is often denoted by H.

- For a objective function E (X) of a single variable X the condition for a minimum at $X = X^*$ are as follows,

$dE(X^*)/dX = 0$ and $d^2E(X^*)/dX^2 > 0$

- For a objective function E (X) of a vectorial variable x the condition for a minimum at $X = X^*$ are as follows,

$\partial E(X^*) = 0$, requires the gradient vector at a minimum of E (X) to be a null vector and $\partial^2 E(X^*)$ is positive definite. requires the Hessian matrix at a minimum of E (x) to be positive definite.

- If X^* is to be minimum of E (X) , any infinitely small change ΔX at X^* should result in $E(X^* + \Delta X) > E(X^*)$. This requires the following :
  1. The gradient vector at X^* vanishes and makes the linear term in the equation of E (X) to zero.
  2. The Hessian matrix at X^* is positive definite.

- Although a wide spectrum of methods exists for unconstrained optimization, methods can be broadly categorized in terms of the derivative information that is, used (Derivative based optimization)or is not used (Derivative free optimization).

3.2 DERIVATIVE BASED OPTIMIZATION

3.2.1 Gradient based Methods

- When the objective function is smooth and if we need efficient local optimization then it is better to use gradient based optimization method.
- Gradient based optimization method determines the search direction using the objective function's derivative information. It is a first-order optimization algorithm.
- At the current point the step is taken in the direction negative of the gradient of the objective function at that point to find a local minimum of a function. In gradient ascent method the step is taken in the positive direction of the gradient to approach a local maximum of that function.

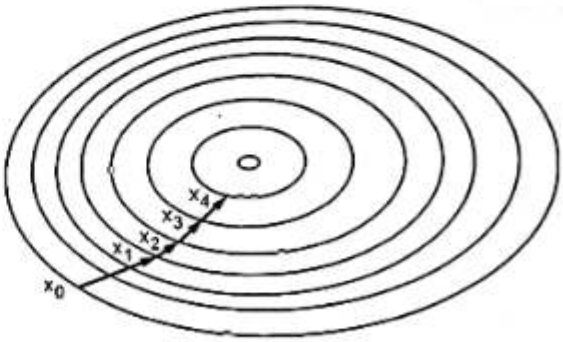


Fig. 3.2.1 : Gradient based method

- When a real valued function E (X) is provided and we are able to differentiate it in a neighborhood of a point 'a' , then E (X) decreases fastest if one goes from 'a' in the direction of the negative gradient of E at 'a', $-\partial E(a)$. This is a basic concept that is used in Gradient descent.
- It follows that, if $b = a - \eta \partial E(a)$

  for $\eta > 0$ a small enough number, then $E(a) \geq E(b)$.
- Due to the complexity of E, we often resort to an iterative algorithm to explore the input space efficiently. With this observation in mind, one starts with a guess X_0 for a local minimum of E, and considers the sequence X_0, X_1, X_2 such that

$$X_{n+1} = X_n - \eta_n \partial E(a), n \geq 0$$

  We have

$$E(X_0) \geq E(X_1) \geq E(X_2) \geq \ldots..$$

- So hopefully the sequence (X_n) converges to the desired local minimum. Note that the value of the step size η is allowed to change at every iteration.
- This process is illustrated in the Fig. 3.2.1. Here E is assumed to be defined on the plane, and that its graph has a bowl shape. The curves are the contour lines, that is, the regions on which the value of E is constant. An arrow originating at a point shows the direction of the negative gradient at that point.
- For minimizing the objective function, the descent procedures are typically repeated until one of the stopping criteria is satisfied such as, the objective function value is sufficiently small. The length of the gradient vector is smaller than a specified value or The specified computing time is exceeded.

Module 3

- Note that the (negative) gradient at a point is orthogonal to the contour line going through that point. We see that gradient descent leads us to the bottom of the bowl, that is, to the point where the value of the function F is minimal.
- Gradient methods are generally more efficient when the function to be minimized is continuous in its first derivative. Higher order methods, such as Newton's method, are only really suitable when the second order information is readily and easily calculated, because calculation of second order information, using numerical differentiation, is computationally expensive. Gradient methods use information about the slope of the function to dictate a direction of search where the minimum is thought to lie.

3.2.2 Method of Steepest Descent

- In the method of steepest descent, the successive adjustments applied to the weight vector w are in the direction of steepest descent i.e., in a direction opposite of gradient vector.

  We can write, $g = \partial E(W)$

- Accordingly, steepest descent method is formally described by,

$$W(n+1) = W(n) - \eta\, g(n)$$

  Where η is positive constant and $g(n)$ is gradient vector evaluated at point $W(n)$. When going from iteration n to n + 1 the algorithm applies the correction as,

$$\Delta W(n) = W(n+1) - W(n) = -\eta\, g(n)$$

- To show that the formulation of the steepest descent algorithm satisfies the condition for iterative descent, we use a first order Taylor series expansion around w(n) to approximate $\varepsilon(W(n+1))$ as,

$$\varepsilon(W(n+1)) = \varepsilon(W(n)) + g(n)\,\Delta W(n)$$

- Substituting value of $\Delta W(n)$ we get,

$$\varepsilon(W(n+1)) = \varepsilon(W(n)) - \eta\, g(n)\, g(n) = \varepsilon(W(n)) - \eta\, g(n)^2$$

  Which shows that for a positive learning parameter η the cost function is decreased as the algorithm progresses from one iteration to the next? The method of steepest descent converges to the optimal solution slowly.

Algorithm

Start with arbitrary point X_1, set iteration number i = 1.

- **Step 1 : Find the search direction S_i as $S_i = -\nabla f_i = \nabla f(x_i)$**
- **Step 2 : Calculate the optimum step length $\lambda_i = \dfrac{S_i^T S_i}{S_i^T H_i S_i}$ and a new point as**

$$X_{i+1} = X_i + \lambda_i S_i$$

- **Step 3 : Test optimality for new point X_{i+1} by $\nabla f(x_{i+1}) \cong 0$**

  If met stop, otherwise repeat step 1 for the new point.

Example 3.2.1 : Minimize $f(X_1, X_2) = X_1 - X_2 + 2X_1^2 + 2X_1X_2 + X_2^2$

Starting from the point $X_1 = (0, 0)$

☑ **Solution :**

We will calculate gradient of f as, $\nabla_f = \begin{bmatrix} \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$

Take a derivative of given equation w.r.t X_1 and X_2

$$\nabla_f = \begin{bmatrix} 1 + 4X_1 + 2X_2 \\ -1 + 2X_1 + 2X_2 \end{bmatrix} \quad ...(1)$$

Now we will calculate Hessian matrix as,

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 x_2} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}$$

Module 3

Iteration 1

At $X_1 = (0, 0)$

- **Step 1 : Find S_1 at X_1, substitute the value(0, 0) in Equation 1 and take a negation of that,**

$$S_1 = -\nabla f(X_1) = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

- **Step 2 : Compute λ_1 at X_1**

$$\lambda_1 = \frac{S_1^T S_1}{S_1^T H S_1} = \frac{[-1\ 1]\begin{bmatrix} -1 \\ 1 \end{bmatrix}}{[-1\ 1]\begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} -1 \\ 1 \end{bmatrix}} = 1$$

Hence the new point is,

$$X_2 = X_1 + \lambda_1 S_1 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

- **Step 3 : Check optimality by substituting value of X_2 in Equation 1.**

$$\nabla f(X_2) = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

So X_2 is not optimum, go to next iteration.

Iteration 2

At $X_2 = (-1, -1)$

- **Step 1 : Find S_2 at X_2, substitute the value(– 1, – 1) in Equation 1 and take a negation of that,**

$$S_2 = -\nabla f(X_2) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

- **Step 2 : Compute λ_2 at X_2**

$$\lambda_2 = \frac{S_2^T S_2}{S_2^T H S_2} = \frac{1}{5}$$

Hence the new point is,

$$X_3 = X_2 + \lambda_2 S_2 = \begin{bmatrix} -0.8 \\ 1.2 \end{bmatrix}$$

- **Step 3 : Check optimality by substituting value of X_3 in Equation 1**

$$\nabla f(X_3) = \begin{bmatrix} 0.2 \\ -0.2 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

So X_3 is not optimum, go to next iteration

Iteration 3

At $X_3 = (-0.8, 1.2)$

- **Step 1 : Find S_3 at X_3, substitute the value(– 0.8, 1.2) in Equation 1 and take a negation of that,**

$$S_3 = -\nabla f(X_3) = \begin{bmatrix} -0.2 \\ 0.2 \end{bmatrix}$$

- **Step 2 : Compute λ_3 at X_3**

$$\lambda_3 = \frac{S_3^T S_3}{S_3^T H S_3} = 1$$

Hence the new point is,

$$X_4 = X_3 + \lambda_3 S_3 = \begin{bmatrix} -1 \\ 1.4 \end{bmatrix}$$

- **Step 3 : Check optimality by substituting value of X_4 in Equation 1**

$$\nabla f(X_4) = \begin{bmatrix} -0.2 \\ -0.2 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

So X_4 is not optimum, go to next iteration

Iteration 4

At $X_4 = (-1, 1.4)$

- **Step 1 : Find S_4 at X_4, substitute the value(– 1, 1.4) in Equation 1 and take a negation of that,**

$$S_4 = -\nabla f(X_4) = \begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix}$$

- **Step 2 : Compute λ_4 at X_4**

$$\lambda_4 = 1/5$$

Hence the new point is,

$$X_5 = X_4 + \lambda_4 S_4 = \begin{bmatrix} -0.96 \\ 1.44 \end{bmatrix}$$

Step 3 : Check optimality by substituting value of X_5 in Equation 1

$$\nabla f(X_5) = \begin{bmatrix} -0.04 \\ -0.04 \end{bmatrix} \approx \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

So X_5 is optimum.

3.2.3 Newton's Method

In the gradient based method only, the gradient vector is required whereas in Newton's method Hessian matrix is required.

- In Classical Newton's method the descent direction is determined by using the second order derivative of the available objective function E. In Fig. 3.2.2 gradient descent method is represented using green line and Newton's method is represented using red line for minimizing a function (with small step sizes). Newton's method uses curvature information to take a more direct route.
- Newton's method is an iterative method for finding roots of equations. Generally, Newton's method is used to find critical points of differentiable functions, which are the zero's of the derivative function.

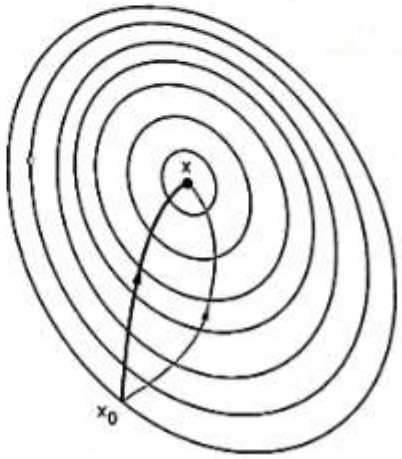


Fig. 3.2.2 : Comparison of gradient descent and Newton's method

- From an initial guess x_0 a sequence x_n is constructed which converges towards x* such that $f'(x^*) = 0$. This x* is called a stationary point of f (.)
- The second order Taylor expansion $f_T(x)$ of function f(.) around x_n (where $\Delta x = x - x_n$) is

$$f_T(X_n + \Delta X) = f_T(X) = f_T(X_n) + f(X_n)\Delta X + 1/2\, f'(X_n)\Delta X^2$$

- Attains its extremism when its derivative with respect to Δx is equal to zero, i.e. when Δx solves the linear equation :

$$f'(X_n) + f'(X_n)\Delta X = 0$$

- Considering the right-hand side of the above equation as a quadratic in Δx, with constant coefficients. Thus, provided that f(x) is a twice-differentiable function well approximated by its second order Taylor expansion and the initial guess is chosen close enough to X_*,

$$\Delta X = X - X_n = -(f'(X_n) / f'(X_n))$$

- The sequence (X_n) defined by :

$$X_{n+1} = X_n - (f'(X_n) / f'(X_n)), \; n = 0, 1, \ldots\ldots$$

will converge towards a root of f', i.e. X_* for which $f'(X_*) = 0$.

Module 3

Algorithm

Start with arbitrary point X_1, set iteration number $i = 1$

- **Step 1 :** **Find the search direction as $\nabla f(X_i)$ and Calculate the inverse of Jaccobain matrix (Newton's step).**
- **Step 2 :** **Calculate the new point as,**

$$X_{i+1} = X_i - J_i^{-1} \nabla f_i$$

- **Step 3 :** **Test optimality for new point X_{i+1} by $\nabla f(x_{i+1}) \cong 0$**

If not stop, otherwise repeat step 1 for the new point.

Example 3.2.2 : Minimize $f(X_1, X_2) = X_1 - X_2 + 2X_1^2 + 2X_1X_2 + X_2^2$

Starting from the point $X_1 = (0, 0)$

☑ **Solution :**

We will calculate gradient of f as, $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_2} \end{bmatrix}$

Take a derivative of given equation w.r.t X_1 and X_2

$$\nabla f = \begin{bmatrix} 1 + 4X_1 + 2X_2 \\ -1 + 2X_1 + 2X_2 \end{bmatrix} \quad \ldots(1)$$

Now we will calculate ∇f_1 by substituting (0, 0) in above equation

$$\nabla f_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Now we will calculate Jaccobian matrix as,

$$J = H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_2^2} & \frac{\partial^2 f}{\partial x_2 x_2} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}$$

Now we will find J^{-1}

$$J^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$$

Iteration 1

At $X_1 = (0, 0)$

- **Step 1 :**

$$X_2 = X_1 - J_1^{-1} \nabla f_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} -1 \\ 1.5 \end{bmatrix}$$

Step 2 : Check optimality by substituting value of X_2 in Equation 1

$$\nabla f(X_2) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

So X_2 is optimum.

3.3 DERIVATIVE FREE OPTIMIZATION

Derivative free optimization is a subject of mathematical optimization. It may refer to problems for which derivative information is unavailable or methods that do not use derivatives.

3.3.1 Random Search

Random search explores the parameter space of an objective function. Sequentially in a seemingly random fashion to find the optimal point that minimizes the objective function.

Algorithm

1. Choose a start point x as the current point. Set initial bias b equal to a zero vector.
2. Add a bias term b and a random vector dx to the current point x in the input space and evaluate the objective function at the new point at x + b + dx.
3. if f (x + b + dx) < f(x), Set the current point x equal to x + b + dx and the bias b equal to 0.2 b + 0.4 dx, go to step 6 otherwise go to next step.
4. if f (x + b − dx) < f(x), Set the current point x equal to x + b − dx and the bias b equal to b − 0.4 dx, go to step 6 otherwise go to next step.
5. Set the bias equal to 0.5b and go to step 6.
6. Stop if the maximum number of function evaluations is reached, otherwise go back to step 2 to find new point.

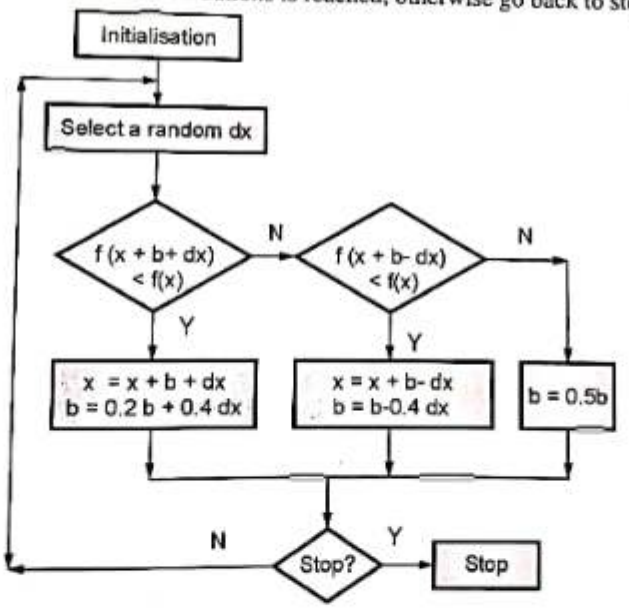


Fig. 3.3.1 : Flowchart of Random Search

Module 3

3.3.2 Down Hill Simplex Algorithm

- The Downhill Simplex algorithm is a multi-dimensional technique of optimization. This technique uses geometric relationships to support in finding function minimums.
- The main superiority of this technique is that in this technique calculation of the derivative of the function is not required. This technique generates its pseudo-derivative by assessing enough points for each independent variable of the function for defining derivative.
- The main component of this technique is a simplex. Simplex is a geometrical object that has n + 1 vertices, here n represents the number of independent variables.
- Example : If there is a problem of a one dimensional optimization then the simplex will contain two points and represent a line segment. In a two dimensional optimization, a three pointed triangle would be used as shown in Fig. 3.3.2.

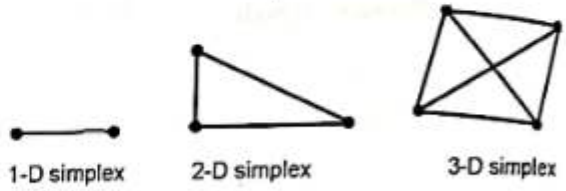


Fig. 3.3.2 : Example of Low order Simplexes

- The downhill simplex optimization technique only function assessing is required instead of derivative calculation. If the space is N-dimensional then a polyhedron with N + 1 vertices is used as a simplex. Initial simplex is define by selecting the N + 1 point. The technique updates the worst point during each iteration. In each iteration four operations are performed as reflection, expansion, one-dimensional contraction, and multiple contractions
- The worst point for which objective function value is maximum is moved to a point which is reflected with the help of remaining N points in Reflection. In expansion method tries to expand the simplex along this line if this point is better as compared to the best point.
- In Contraction the simplex is contracted along one dimension from the maximum point if the new point is not better as compared to the previous point. If the new point is not good as compared to the previous points then the simplex is contracted along all dimensions toward the best point and steps down the valley. In the downhill simplex search technique these operations are applied serially during each iteration till the method finds the optimal solution. In the downhill simplex technique following operations are applied serially :

  Arrange points with the rank and do the relabeling of N + 1 points. Points are ranked according to following condition.

  $f(P_{N+1}) > \ldots > f(P_2) > f(P_1)$

- Create initial P_r point with the help of reflection.

$$P_r = P_c + \alpha^* (P_c - P_{N+1})$$

  Here P_c is calculated by taking the centroid of the N best points in the vertices of the simplex.

- P_{N+1} is replaced by P_r, if following condition satisfied

  $f(P_1) < f(P_r) < f(P_N)$

- A new point P_e is created with the help of expansion if following condition is satisfied

  Condition : $f(P_r) < f(P_1)$

$$P_e = P_r + \beta^* (P_r - P_c)$$

- P_{N+1} is replaced by P_e, if following condition is satisfied, else P_{N+1} is replaced by P_r

  Condition : $f(P_e) < f(P_r)$,
- Create a new point P_t with the help of contraction if following condition is satisfied.

  Condition : $f(P_r) > f(P_N)$

$$P_t = P_c + \gamma^* (P_{N+1} - P_c).$$

- P_{N+1} is replaced by P_t if following condition is satisfied

  Condition : $f(P_t) < f(P_{N+1})$,
- Contract along all dimensions toward P_1 if following condition is satisfied.

  Condition : $f(P_t) > f(P_{N+1})$.
- Assess the objective function values at the N new vertices as,

$$P_i = P_1 + \eta^* (P_i - P_1)$$

3.4 UNIVERSITY QUESTIONS AND ANSWERS

Module 3

May 2018

Q. 1 Write short note on Derivative based optimization. *(Ans. : Refer section 3.2)* **(10 Marks)**

May 2019

Q. 2 Describe Down Hill Simplex method. Why is it called Derivative Free method ? *(Ans. : Refer sections 3.3 and 3.3.2)* **(5 Marks)**

Q. 3 Minimize $f(X_1, X_2) = 4X_1 - 2X_2 + 2X_1^2 + 2X_1X_2 + X_2^2$, Starting from the point $X_1 = (0, 0)$. **(10 Marks)**

Using steepest descent method (Perform only two iterations) *(Ans. : Refer section 3.2.2)*

Q. 4 Differentiate : Derivative Based and Derivative free optimization techniques. *(Ans. : Refer sections 3.2 and 3.3)* **(5 Marks)**

Dec. 2019

Q. 5 List some advantages of derivative-based optimization techniques. Explain Steepest Descent method for optimization. *(Ans. : Refer sections 3.2 and 3.2.2)* **(10 Marks)**

Q. 6 Write short note on Down Hill Simplex Method. *(Ans. : Refer section 3.3.2)* **(5 Marks)**

Multiple Choice Questions

Q. 3.1 What type of constraint in the feasible basic solution must satisfy for simplex method ?

(a) Non-Negativity (b) Negativity

(c) Basic (d) Common ✓ **Ans. : (a)**

Explanation : Non negative points are used for simplex method.

Q. 3.2 For Gradient Decent (GD) and Stochastic Gradient Decent (SGD) which of the following sentence is correct?

(a) You update a set of parameters in an iterative manner to minimize the error function.

(b) You have to run through all the samples in your training set for a single update of a parameter in each iteration for SGD.

(c) You either use the entire data or a subset of training data to update a parameter in each iteration in GD.

(d) You update a set of parameters in parallel manner to minimize the error function. ✓ Ans. : (a)

Explanation : In SGD for each iteration you choose the batch which is generally contain the random sample of data But in case of GD each iteration contain the all of the training observations.

Q. 3.3 Which one of the following is incorrect w.r.t. Derivative based optimization ?

(a) Uses derivative information with objective function

(b) Slow convergence

(c) Follows mathematical methodology

(d) Fast Convergence ✓ Ans. : (b)

Explanation : Derivative based optimization methods converges very fast.

Q. 3.4 In Classical Newton's method the descent direction is determined by which method?

(a) First order derivative of the function

(b) Partial order derivative of the available objective function

(c) Gradient method

(d) Second order derivative of the available objective function ✓ Ans. : (d)

Explanation : According to working of the Newton method.

Q. 3.5 In Derivative free optimization methods points are selected based on which criteria _____

(a) Minimum value (b) Maximum value

(c) Fitness Value (d) Error value ✓ Ans. : (c)

Explanation : Derivative free optimization methods works on the concept of survival of the fittest. It denotes how good the point is.

Q. 3.6 When Gradient information is used it is called as which type of optimization _____

(a) Derivative free optimization

(b) Derivative based optimization

(c) Constrained optimization

(d) Optimization ✓ Ans. : (b)

Explanation : In derivative based optimization derivative information is used.

Q. 3.7 Which one is not a derivative Free optimization method?

(a) Genetic algorithm (b) Downhill simplex method.

(c) Simulated Annealing

(d) Gradient Descent method ✓ Ans. : (d)

Explanation : Gradient based method is a derivative based optimization method.

Q. 3.8 Convergence speed of Derivative Free Optimization method is _____

(a) Fast (b) Slow

(c) Very Fast (d) Medium ✓ Ans. : (b)

Explanation : more number of iterations are required in derivative free methods since they work on random basis.

Q. 3.9 Derivative based Optimization method uses _____

(a) Heuristic Search (b) Best First Search

(c) Breadth First Search (d) Depth First Search

✓ Ans. : (a)

Explanation : Heuristic function (derivative information) is used for optimization.

Q. 3.10 This is being used for evaluation in Derivative Based Optimization method _____

(a) Only Objective Function

(b) Derivative Information with Objective Function

(c) Only Derivative information

(d) Only Objective Information ✓ Ans. : (b)

Explanation : Derivative of objective function is calculated in this method.

Q. 3.11 As the iterations are performed cost to update parameters in method of steepest descent method is _____

(a) increases (b) decreases

(c) constant (d) fluctuates ✓ Ans. : (b)

Explanation : As number of iterations is progressed cost to update weight parameters decreases.

Q. 3.12 Evolutionary concept is used in

(a) Derivative free method

(b) Derivative based method

(c) Both the methods

(d) None of the methods ✓ Ans. : (a)

Explanation : Derivative free methods like genetic algorithm is based on evolutionary concept.

Q. 3.13 If g(z) is the sigmoid function, then its derivative with respect to z may be written in term of g(z) as _____

(a) $g(z)(g(z) - 1)$ (b) $g(z)(1 + g(z))$

(c) $-g(z)(1 + g(z))$ (d) $g(z)(1 - g(z))$ ✓ Ans. : (d)

Explanation : Formula to calculate derivative of sigmoid function.

Q. 3.14 We can get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent.

(a) True (b) False ✓ Ans. : (b)

Explanation : We will not get multiple optimum solution.

Q. 3.15 The error function most suited for gradient descent using logistic regression is _____

(a) The entropy function (b) The squared error

(c) The cross-entropy function

(d) The number of mistakes ✓ **Ans. : (b)**

Explanation : Mean squared error function is used.

Q. 3.16 Suppose your model is overfitting. Which of the following is NOT a valid way to try and reduce the overfitting ?

(a) Increase the amount of training data.

(b) Improve the optimisation algorithm being used for error minimisation.

(c) Decrease the model complexity.

(d) Reduce the noise in the training data. ✓ **Ans. : (b)**

Explanation : Improving optimization algorithm does not help in reducing overfittng.

Q. 3.17 To find the minimum or the maximum of a function, we set the gradient to zero because : _____

(a) The value of the gradient at extrema of a function is always zero

(b) Depends on the type of problem

(c) Both (a) and (b)

(d) None of the above ✓ **Ans. : (a)**

Explanation : According to working of gradient based optimization methods.

Q. 3.18 Which of the following is NOT required for using Newton's method for optimization ?

(a) The lower bound for the search region

(b) Twice differentiable optimization function

(c) The function to be optimized

(d) A good initial estimate that is reasonably close to the optimal Solution ✓ **Ans. : (a)**

Explanation : Newton's method is not a bracketing method but an open method. Only bracketing methods require a lower (or upper) bound for the search region. Therefore a is not required for using Newton's method.

Q. 3.19 Which of the following statements is INCORRECT?

(a) If the second derivative at i x is negative, then i x is a maximum.

(b) If the first derivative at i x is zero, then i x is an optimum.

(c) If i x is a maximum, then the second derivative at i x is positive.

(d) The value of the function can be positive or negative as any optima. ✓ **Ans. : (c)**

Explanation : When the function is at a maximum, its second derivative has a negative value and not a positive value.

Q. 3.20 For what value of x, is the function $x^2 - 6x - 2$ minimized ?

(a) 0 (b) 1 (c) 5 (d) 3 ✓ **Ans. : (b)**

Explanation : Probably the easiest way to solve the problem is to recognize that the second derivative is positive. So the function is a minimum at x = 0, and the correct answer is (b) This is also the absolute minimum because the first derivative is zero only at a single point. Using an initial estimate of 0, using Newton's method, the first iteration would converge to the optimal solution.

Q. 3.21 Which of the following is incorrect ?

(a) Direct search methods are useful when the optimization function is not differentiable

(b) The gradient of f(x,y) is the a vector pointing in the direction of the steepest slope at that point.

(c) The Hessian is the Jacobian Matrix of second-order partial derivatives of a function.

(d) The second derivative of the optimization function is used to determine if we have reached an optimal point. ✓ **Ans. : (d)**

Explanation : The statement "The second derivative of the optimization function is used to determine if we have reached an optimal point" is incorrect. The second derivative tells us if we the optimum we have reached is a maximum or minimum.

Q. 3.22 An initial estimate of an optimal solution is given to be used in conjunction with the steepest ascent method to determine the maximum of the function. Which of the following statements is correct?

(a) The function to be optimized must be differentiable.

(b) If the initial estimate is different than the optimal solution, then the magnitude of the gradient is nonzero.

(c) As more iterations are performed, the function values of the solutions at the end of each subsequent iteration must be increasing.

(d) All 3 statements are correct. ✓ **Ans. : (d)**

Explanation : All statements are correct for steepest ascent method.

Q. 3.23 What are the gradient and the determinant of the Hessian of the function $f(x, y) = x^2 y^2$ at its global optimum?

(a) $\nabla f = 0i + 0j$ and determinant (H) > 0

(b) $\nabla f = 0i + 0j$ and determinant (H) = 0

(c) $\nabla f = 1i + 0j$ and determinant (H) < 0

(d) $\nabla f = 1i + 1j$ and determinant (H) = 0 ✓ **Ans. : (a)**

Module 3

Explanation : When the global optimum is reached, travel in any direction would increase/decrease the function value, therefore the magnitude of the gradient must be 0. At any optimum, the hessian must be positive, therefore the correct answer is a.

Q. 3.24 Determine the gradient of the function $x^2 - 2y^2 - 4y + 6$ at point (0, 0)?

(a) $\nabla f = 2i - 4j$ (b) $\nabla f = 0i - 4j$

(c) $\nabla f = 0i + 0j$ (d) $\nabla f = -4i - 4j$

✓ **Ans. : (b)**

Explanation : At point (0,0), we calculate the gradient at this point as $\frac{df}{dx} = 2x = 2(0) = 0$,

$\frac{df}{dy} = -4y - 4 = -4(0) - 4 = -4$ which are used to determine the gradient as $\nabla f = 0i - 4j$

Q. 3.25 Determine the determinant of hessian of the function $x^2 - 2y^2 - 4y + 6$ at point (0, 0)?

(a) 2 (b) -4 (c) 0 (d) -8 **Ans. : (d)**

Explanation : To determine the Hessian, the second partial derivatives are determined and evaluated as follows $\frac{d^2f}{dx^2} = 2$, $\frac{d^2f}{dy^2} = -4$ $\frac{d^2f}{dydx} = 0$. The resulting Hessian matrix and its determinant are

$H = \begin{bmatrix} 2 & 0 \\ 0 & -4 \end{bmatrix}$ and determinant $(H) = -8$.

Q. 3.26 In descent methods, the particular choice of search direction does not matter so much.

(a) True (b) False. ✓ **Ans. : (b)**

Explanation : Selection of search direction makes an impact.

Q. 3.27 In descent methods, the particular choice of line search does not matter so much.

(a) True (b) False ✓ **Ans. : (a)**

Explanation : Selection of line search does not make any impact.

Q. 3.28 When the gradient descent method is started from a point near the solution, it will converge very quickly.

(a) True (b) False ✓ **Ans. : (b)**

Explanation : Starting point is near or far does not matter in gradient descent method.

Q. 3.29 Newton's method with step size h = 1 always works.

(a) True (b) False ✓ **Ans. : (b)**

Explanation : It will not work for every scenario.

Q. 3.30 When Newton's method is started from a point near the solution, it will converge very quickly.

(a) True (b) False ✓ **Ans. : (a)**

Explanation : Starting point is near or far matter in Newton's method.

Q. 3.31 Newton's method would probably require fewer iterations than the gradient method, but each iteration would be much more costly.

(a) True (b) False ✓ **Ans. : (b)**

Explanation : According to property of Newton's method.

Q. 3.32 Gradient of a continuous and differentiable function ____

(a) is zero at a minimum

(b) is non-zero at a maximum

(c) is zero at a saddle point

(d) decreases as you get closer to the minimum

✓ **Ans. : (a), (c), (d)**

Explanation : Gradient of a continuous and differentiable function is not a non-zero at a maximum.

Q. 3.33 Let us say that we have computed the gradient of our cost function and stored it in a vector g. What is the cost of one gradient descent update given the gradient?

(a) O(D) (b) O(N)

(c) O(ND) (d) $O(ND^2)$ ✓ **Ans. : (a)**

Explanation : According to working of gradient descent method.

Q. 3.34 Which of the following statements is FALSE ?

(a) Multidimensional direct search methods are similar to one-dimensional direct search methods.

(b) Enumerating all possible solutions in a search space and selecting the optimal solutions is an effective method for problems with very high dimensional solution spaces.

(c) Multidimensional direct search methods do not require a twice differentiable function as an optimization function

(d) Genetic Algorithms belong to the family of multidimensional direct search methods.

✓ **Ans. : (b)**

Explanation : Problems with very high dimensional solution spaces are very large and therefore it is computationally difficult to enumerate the search space.

Q. 3.35 Which of the following statements is FALSE?

(a) Multidimensional direct search methods require an upper and lower bound for their search region.

(b) Coordinate cycling method relies on single dimensional search methods to determine an optimal solution along each coordinate direction iteratively.

(c) If the optimization function is twice differentiable, multidimensional direct search methods cannot be used to find an optimal solution.

(d) Multidimensional direct search methods are not guaranteed to find the global optimum.

✓ **Ans. : (c)**

Explanation : Multidimensional direct search methods can be used with any function to find optimal solutions. If the functions are twice differentiable there are more computationally efficient techniques for optimization of these functions.

Q. 3.36 If we want to find the value of the variable 'x' that maximizes a differentiable function f(x), then we should : _____

(a) Find 'x' from $\frac{df}{dx} = 0$

(b) Find 'x' from $\frac{df}{dx} = 0$ and check that that second derivative is positive

(c) Find 'x' from $\frac{df}{dx} = 0$ and check that that second derivative is negative

(d) Find 'x' from second derivative = 0 ✓ **Ans. : (c)**

Explanation : Find 'x' from $\frac{df}{dx} = 0$ and check that that second derivative is negative.

Q. 3.37 Each optimization problem must have certain parameters called _____

(a) linear variables (b) dummy variables

(c) design variables (d) none of the above

✓ **Ans. : (c)**

Explanation : Design variables are required for every optimization problems.

Q. 3.38 A "≤ type" constraint expressed in the standard form is active at a design point if it has _____

(a) zero value (b) more than zero value

(c) less than zero value (d) (a) and (c)

✓ **Ans. : (a)**

Explanation : Standard representation of constraint.

Q. 3.39 Maximization of f(x) is equivalent to minimization of _____

(a) $-f(x)$ (b) $1/f(x)$

(c) $\sqrt{f(x)}$ (d) none of the above ✓ **Ans. : (a)**

Explanation : Standard procedure of constraints.

Q. 3.40 When the optimizaion problem cost functions are differentiable, the problem is referred to as _____

(a) rough (b) nonsmooth

(c) smooth (d) (a) and (b) ✓ **Ans. : (c)**

Explanation : According to definition of cost function of optimization problem.

Q. 3.41 The feasible region for the inequality constraints with respect to equality constraints _____

(a) increases (b) decreases

(c) does not change (d) none of the above

✓ **Ans. : (a)**

Explanation : Standard procedure of constraints.

Q. 3.42 Which of the following is a objective function in derivative based optimization for feed forward network ?.

(a) Error function (b) Energy function

(c) Cost function (d) Fitness function

✓ **Ans. : (a)**

Explanation : Derivative of error function is used in feed forward network optimization.

Q. 3.43 When the objective function is smooth and if we need efficient local optimization then it is better to use _____

(a) Gradient based optimization method.

(b) Derivative based optimization method

(c) Both of above methods

(d) None of the above ✓ **Ans. : (a)**

Explanation : Gradient based optimization method gives better solution if objective function is smooth.

Q. 3.44 Gradient at a point is _____ to the contour line going through that point.

(a) Parallel (b) Orthogonal

(c) None of above (d) All above ✓ **Ans. : (b)**

Explanation : According to property of gradient.

Q. 3.45 In Random search method, if f (x + b + dx) < f(x), Set _____

(a) x = x + b + dx and b = 0.2 b + 0.4 dx

(b) x = x + b – dx and b = b – 0.4 dx

(c) b = 0.5 b

(d) None of above ✓ **Ans. : (a)**

Explanation : According to working of random search method.

Q. 3.46 Down hill simplex method uses which relationship_____

(a) Input-output (b) Geometric

(c) Actual output-desired output (d) None of above

✓ **Ans. : (b)**

Module 3

Explanation : Downhill simplex method is a multi- dimensional technique that uses geometric relationship.

Q. 3.47 Simplex in downhill simplex search method is an object with how many vertices _____

(a) n (b) n + 1

(c) n – 1 (d) n + 2 ✓ **Ans. : (b)**

Explanation : Simplex is a geometrical object that has n + 1 vertices, here n represents the number of independent variables.

Q. 3.48 Out of following which operation is performed in downhill simplex search _____

(a) Reflection (b) Contraction

(c) Expansion (d) All of above ✓ **Ans. : (d)**

Explanation : The technique updates the worst point during each iteration. In each iteration four operations are performed as reflection, expansion, one-dimensional contraction, and multiple contractions.

Chapter Ends...

□□□

Module 4

Chapter... 4

Learning with Regression and Trees

University Prescribed Syllabus

Learning with Regression : Linear Regression, Logistic Regression. Learning with Trees : Decision Trees, Constructing Decision Trees using Gini Index, Classification and Regression Trees (CART).

4.1 LINEAR REGRESSION

- One of the most important supervised learning tasks is regression. In regression set of records are present with X and Y values and this values are used to learn a function, so that if you want to predict Y from an unknown X this learn function can be used. In regression we have to find value of Y, So, a function is required which predicts Y given X. Y is continuous in case of regression.
- Here Y is called as criterion variable and X is called as predictor variable. There are many types of functions or modules which can be used for regression. Linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

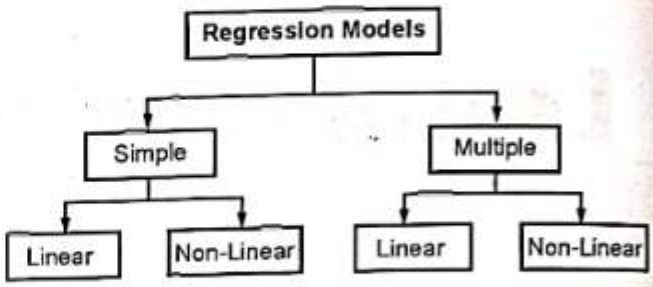
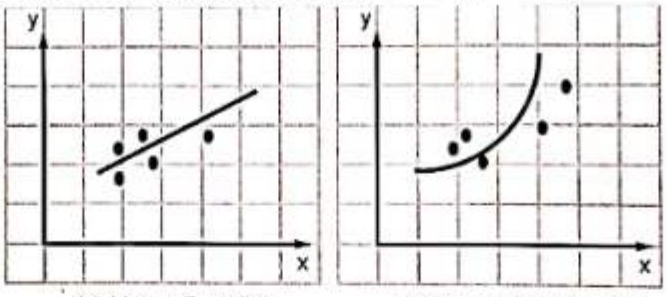


Fig. 4.1.1 : Types of Regression Models

4.1.1 Simple Linear Regression

- Let's see simple regression first, in this X contains a single feature. In multiple regressions, X contains more than one feature. In simple regression training records are plotted as value of X vs. value of Y. Next task is to find a function, so that if a random unknown X value is given we can predict Y. There are different types of functions that can be used. In linear regression, we assume that the function is linear as shown in Fig. 4.1.2(a).



(a) Linear Function (b) Non-Linear Function

Fig. 4.1.2 : Types of regression functions

- In Linear regression a best fitted straight line is drawn which passes through the points called as the regression line. The line shown in Fig. 4.1.2 is the regression line. Regression line shows the calculated values of Y for each possible value of X. Errors in the prediction is shown by the vertical lines drawn from the points to the regression line. Error in prediction is less when the point is very near to the regression line as shown by first and second point in Fig. 4.1.3. When the point is far from the regression line then the error in prediction is high, as shown by the third and fourth point in Fig. 4.1.3.

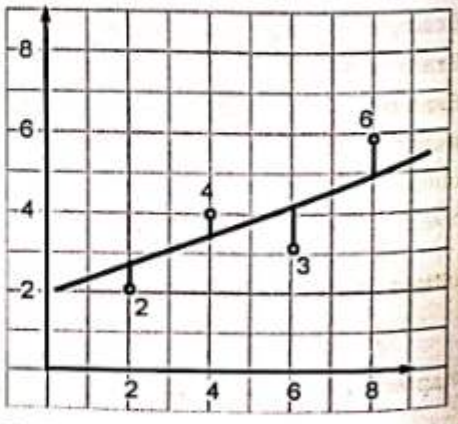


Fig. 4.1.3 : Prediction and the Error in the Prediction

- The difference between the value of point and the predicted value is called as error of prediction. Predicted value is the value of point on the line.
- Let's take an example, the predicted values (Y′) and the errors of prediction (Y-Y′) are shown in Table 4.1.1. From the table we can say that, the second point has Y value as 4 and a predicted Y value as 3.2. The error of prediction is 0.8.

Table 4.1.1 : Regression data.

Sr. No.	X	Y	Y′	Y-Y′	(Y-Y′)²
1	2	2	2.2	– 0.2	0.04
2	4	4	3.2	0.8	0.64
3	6	3	4.3	– 1.3	1.69
4	8	6	5.4	0.6	0.36

- The best-fit line is called as the line that will minimize the sum of the squared errors of prediction. This criterion is used to draw the line in Fig. 4.1.3. The squared errors of prediction are shown in the last column of Table 4.1.1.
- The regression line is represented using the following equation,

$$Y' = a X + b + e$$

- In the above equation Y' represents the predicted value, a represents the slope of the line, b shows the Y intercept and e is the random error. Here we assume that mean value of random error is 0, so the equation becomes,

$$Y' = a X + b$$

Table 4.1.2 : Calculation of Regression Parameters

Sr. No.	X	Y	XY	X^2
1	2	2	4	4
2	4	4	16	16
3	6	3	18	36
4	8	6	48	64
Total	**20**	**15**	**86**	**120**

Module 4

- The regression line equation is,

$$Y' = a X + b$$

$$a = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} = \frac{4 \times 86 - 20 \times 15}{4 \times 120 - 400} = 0.55$$

$$b = \frac{1}{n} (\sum Y - a \times \sum X) = \frac{1}{4} (15 - 0.55 \times 20) = 1$$

- Now the equation for the line becomes,

$$Y' = 0.55 X + 2.2$$

For X = 2,

$$Y' = (0.55)(2) + 1 = 2.2$$

For X = 4,

$$Y' = (0.55)(4) + 1 = 3.2$$

For X = 6,

$$Y' = (0.55)(6) + 1 = 4.3$$

For X = 8,

$$Y' = (0.55)(8) + 1 = 5.4$$

4.1.2 Multiple Linear Regression

- In Multiple linear regressions there are two or more number of features. We can also say it is a extension to simple linear regression.
- The regression line is represented using the following equation,

$$Y' = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \text{----------} + \alpha_n X_n + e$$

- In the above equation Y' is the predicted value, X_1 X_2 ___ X_n are the predictors, e is random error and $\alpha_0, \alpha_1, \alpha_2, \ldots \alpha_n$ are regression coefficients.

4.2 EXAMPLES OF LINEAR REGRESSION

Example 4.2.1 : The expenditure of an organization (in thousand) for every month is shown in table below:

X (Month)	1	2	3	4	5
Y (Expenditure)	12	19	29	37	45

Find regression line, Y = a X + b using least square method.

Estimate the expenditure of company in 6th month using line as a model.

Solution :

Sr. No.	X	Y	XY	X^2
1	1	12	12	1
2	2	19	38	4
3	3	29	87	9
4	4	37	148	16
5	5	45	225	25
Total	**15**	**142**	**510**	**55**

(a) The equation for the regression line is,

$$Y' = aX + b$$

$$a = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{5 \times 510 - 15 \times 142}{5 \times 55 - 225} = 8.4$$

$$b = \frac{1}{n}(\sum Y - a \times \sum X) = \frac{1}{5}(142 - 8.4 \times 15) = 3.2$$

Now the equation for the line becomes,

$$Y' = 8.4\,X + 3.2$$

(b) In 6^{th} month

$$Y' = 8.4 \times 6 + 3.2$$

$$Y' = 53.6 \text{ Thousand}$$

Example 4.2.2 : Consider the set of data as {(– 1, – 1), (2, 2), (3, 2)} (a) Find the equation of regression line. (b) Draw the scatter plot of data and regression line.

☑ **Solution :**

Sr. No.	X	Y	XY	X^2
1	– 1	– 1	1	1
2	2	2	4	4
3	3	2	6	9
Total	4	3	11	14

(a) The equation for the regression line is,

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

$$a = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{3 \times 11 - 4 \times 3}{3 \times 14 - 16} = 0.807$$

$$b = \frac{1}{n}(\sum Y - a \times \sum X) = \frac{1}{3}(3 - 0.807 \times 4) = -0.076$$

Now the equation for the line becomes,

$$Y' = 0.807\,X - 0.076$$

Module 4

(b) Now we can plot the regression line given by equation $Y' = 0.807\,X - 0.076$ and the given data points.

Sr. No.	X	Y'
1	– 1	– 0.883
2	2	1.54
3	3	2.34

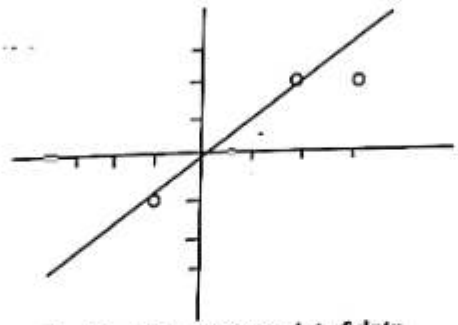


Fig. Ex. 4.2.2 : Scatter plot of data

Example 4.2.3 **MU - May 17, 10 Marks**

Following table shows the midterm and final exam grades obtained for students in a database course. Use the method of least squares using regression to predict the final exam grade of a student who received 86 in the mid term exam.

Midterm exam(X)	72	50	81	74	94	86	59	83	86	33	88	81
Final exam(Y)	84	53	77	78	90	75	49	79	77	52	74	90

☑ **Solution :**

Sr. No.	X	Y	XY	X^2
1	72	84	6048	5184
2	50	53	2650	2500
3	81	77	6237	6561
4	74	78	5772	5476
5	94	90	8460	8836
6	86	75	6450	7396
7	59	49	2891	3481
8	83	79	6557	6889
9	86	77	6622	7396
10	33	52	1716	1089
11	88	74	6512	7744
12	81	90	7290	6561
Total	**887**	**878**	**67205**	**69113**

The equation for the regression line is,

$$Y' = aX + b$$

$$a = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = 0.65$$

$$b = \frac{1}{n}(\sum Y - a \times \sum X) = 25.12$$

Now the equation for the line becomes,

$$Y' = 0.65X + 25.12$$

The final exam grade of a student who received 86 in the mid term exam,

$$Y' = 0.65 \times 86 + 25.12$$

$$Y' = 81.02$$

UExample 4.2.4 **MU - Dec. 19, 10 Marks**

Given the following data for the sales of car of an automobile company for six consecutive years. Predict the sales for next two consecutive years.

Years (x)	2013	2014	2015	2016	2017	2018
Sales (y)	110	100	250	275	230	300

☑ **Solution :**

We will take t = x – 2013

Sr. No.	t	Y	tY	t^2
0	0	110	0	0
1	1	100	100	1
2	2	250	500	4
3	3	275	825	9
4	4	230	920	16
5	5	300	1500	25
Total	15	1265	3845	55

The equation for the regression line is,

$$Y' = at + b$$

$$a = 39$$

$$b = 113.33$$

Now the equation for the line becomes,

$$Y' = 39t + 113.33$$

The sale of company for next two years, X = 2019, t = 6

$$Y' = 39 \times 6 + 113.33$$

$$Y' = 347.33$$

X = 2020, t = 7

$$Y' = 39 \times 7 + 113.33$$

$$Y' = 386.33$$

4.3 LOGISTIC REGRESSION

Module 4

- Suppose we have different training data which belongs to two different classes, and we have to design a system that will identify which data is from which class. The output of this function will be a real value and it is not suitable for classification method. We can use another function on this linear function so that we can use the result for classification. In logistic regression logistic function or the sigmoid can be used.
- Let's first see what is the meaning of Logistic or Sigmoid function.
- In Logistic regression classifier we will take features as the input. These features are multiplied with the logistic coefficients and we add the product. This is called as the net input that will be given to the sigmoid function. The output will be between 0 and 1. Anything above 0.5 is classified as 1and anything below 0.5 is classified as 0.

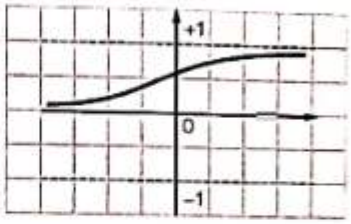


Fig. 4.3.1 : Logistic or Sigmoid Function

- The net input to the sigmoid function is,

$$Z = b_0 + b_1 x_1 + b_2 x_2 + \ldots\ldots\ldots + b_n x_n$$

- In the net input equation x represents the input data or the features. When we use optimized coefficient b, Classifier will be successful. Optimized b can be calculated using the optimization concept.

Gradient Ascent Method

- In gradient ascent method we move in the direction of the gradient to find the maximum point on a function.

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f(x, y)}{\partial x} \\ \frac{\partial f(x, y)}{\partial y} \end{bmatrix}$$

- In the above equation $\partial f(x, y)/\partial x$ represents the amount by which updation is applied in x direction and $\partial f(x, y)/\partial y$ represents the amount by which updation applied in y direction
- The gradient operator will always point towards the direction of gradient increase.

$$b = b + \alpha \times \nabla f(b)$$

- This step is repeated until we reach stopping criterion.

Pseudo code

```
Start using the logistic coefficient, all set to 0
Repeat no. of times
  Find the gradient of complete dataset
  Update, logistic coefficient = logistic coefficient + alpha × gradient
  Return the logistic coefficient
```

Gradient Descent Method

- In gradient descent method we move in the opposite direction of the gradient to find the minimum point on a function
- The gradient operator will always point opposite to the direction of gradient increase.

$$b = b - \alpha \times \nabla f(b)$$

- This step is repeated until we reach stopping criterion.
- Using above mentioned methods optimized b is calculated, net input is calculated and then the prediction is calculated by giving the net input to the function,

$$\text{Prediction} = \frac{1}{1 + e^{-z}}$$

- Finally the data points are classified as,

$$\text{Class} = 1 \quad \text{if Prediction} \geq 0.5$$
$$= 0 \quad \text{else}$$

- The classified data using the decision boundary is as shown in Fig. 4.3.2.
- Logistic regression is used to model a relationship between input variables and a target variable. Let's take an example of a house management system, we can use logistic regression to model the relationship between the parameters such as the total monthly income of the family, various liabilities, monthly expenditure of a family to predict if monthly savings (investment) can be done or not.

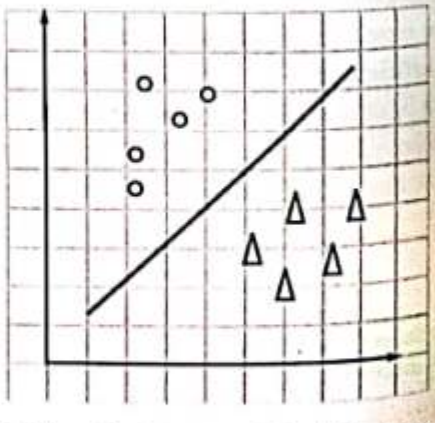


Fig. 4.3.2 : Sample of output of Logistic regression

- There are three types of logistic regression as below.

Binary Logistic Regression

- Binary logistic regression is used if the target variable is binary.
- The application of this method can be above mentioned house management example.

Nominal Logistic Regression

- Nominal Logistic Regression is used if there are three or more classes without any specific orders.
- The examples of this type of regression could be different departments of the engineering college (Computer, IT, Mechanical, Civil, Electronics etc.).

Ordinal Logistic Regression

- Ordinal Logistic Regression is used if there are three or more classes with specific orders.
- The examples of this type of regression could be how customers rate the taste of food using a scale of 1-3 (Bad, Good, and Excellent).

Example 4.3.1 : A Bank has to decide whether to sanction loan or not based on two attributes as person's income and his savings. The data is given in the following table where 1 represents loan is sanctioned and 0 represents loan is not sanctioned. Predict whether a person 3 will get a loan or not having annual income as 12.5 lakhs and savings as 10 lakhs.

Person	Annual income in lakhs (x_1)	Savings in lakhs (x_2)	Loan sanctioned? (y)
1	14.5	12.5	1
2	8.5	4.5	0

Solution :

Module 4

Initially assume logistic regression coefficients $b_0 = b_1 = b_2 = 0$

For 1st row, $x_1 = 14.5$, $x_2 = 12.5$ and $y = 1$

Now we will calculate prediction for the first row,

$$\text{Prediction} = 1 / (1 + e^{\wedge} (- (b_0 + b_1 \times x_1 + b_2 \times x_2)))$$

$$\text{Prediction} = 1 / (1 + e^{\wedge} (- (0 + 0 \times 14.5 + 0 \times 12.5)))$$

$$\text{Prediction} = 0.5$$

Now we will calculate the new coefficient values using a simple update equation. Ideal values for alpha are from 0.1 to 0.3. Let's take alpha as 0.3. For b_0 by default input is 1.

$$b_{new} = b_{old} + \text{alpha} \times (y - \text{prediction}) \times \text{prediction} \times (1 - \text{prediction}) \times \text{input}$$

$$b_{0new} = 0 + 0.3 \times (1 - 0.5) \times 0.5 \times (1 - 0.5) \times 1.0 = 0.0375$$

$$b_{1new} = 0 + 0.3 \times (1 - 0.5) \times 0.5 \times (1 - 0.5) \times 14.5 = 0.54375$$

$$b_{2new} = 0 + 0.3 \times (1 - 0.5) \times 0.5 \times (1 - 0.5) \times 12.5 = 0.46875$$

Now we will calculate prediction for the second row,

$$\text{Prediction} = 1 / (1 + e^{\wedge} (- (b_0 + b_1 \times x_1 + b_2 \times x_2)))$$

Prediction = 1 / (1 + e^ (– (0.0375 + 0.54735 × 8.5 + 0.46875 × 4.5)))

Prediction = 0.99

Now we will calculate the new coefficient values

$$b_{new} = b_{old} + alpha \times (y - prediction) \times prediction \times (1 - prediction) \times input$$

$$b_{0new} = 0.0375 + 0.3 \times (0 - 0.99) \times 0.99 \times (1 - 0.99) \times 1.0 = 0.034$$

$$b_{1new} = 0.54735 + 0.3 \times (0 - 0.99) \times 0.99 \times (1 - 0.99) \times 8.5 = 0.523$$

$$b_{2new} = 0.46875 + 0.3 \times (0 - 0.99) \times 0.99 \times (1 - 0.99) \times 4.5 = 0.456$$

Now we will use these values for prediction

$$Prediction = 1 / (1 + e^{\wedge} (- (b_0 + b_1 \times x_1 + b_2 \times x_2)))$$

Prediction = 1 / (1 + e^ (– (0.034 + 0.523 × 12.5 + 0.456 × 10)))

Prediction = 0.99

Since prediction ≥ 0.5

Prediction for person 3 is, his loan will be sanctioned.

Note : Generally huge amount of data is used for training and number of iterations are also applied to get accuracy. Here for example purpose only two records for training are taken and a single iteration is shown.

4.4 DECISION TREES

- Decision trees are very strong and most suitable tools for classification and prediction. The attractiveness of decision trees is due to the fact that, in contrast to neural network, decision trees represent rules.
- Rules are represented using linguistic variables so that user interpretability may be achieved. By comparing the records with the rules one can easily find a particular category to which the record belongs to.
- In some applications, the accuracy of a classification or prediction is the only thing that matters in such situations we do not necessarily care how or why the model works. In other situations, the ability to explain the reason for a decision is crucial, in marketing one has described the customer segments to marketing professionals, so that they can use this knowledge to start a victorious marketing campaign.
- This domain expert must acknowledge and approve this discovered knowledge and for this we need good descriptions. There are a variety of algorithms for building decision trees that share the desirable quality of interpretability (ID3).

1. Where Decision Tree is applicable?

Decision tree method is mainly used for the tasks that possess the following properties :

- The tasks or the problems in which the records are represented by attribute-value pairs.

  Records are represented by a fixed set of attribute and their value Example : For 'temperature' attribute the value is 'hot'.

  When there are small numbers of disjoint possible values for each attribute, then decision tree learning becomes very simple.

Example: Temperature attribute takes three values as hot, mild and cold.

Basic decision tree algorithm may be extended to allow real valued attributes as well.

Example: we can define floating point temperature.

- An application where the target function takes discrete output values.

  In Decision tree methods an easiest situation exists, if there are only two possible classes.

  Example : Yes or No

  When there are more than two possible output classes then decision tree methods can also be easily extended.

  A more significant extension allows learning target functions with real valued outputs, although the application of decision trees in this area is not frequent.

- The tasks or the problems where the basic requirement is the disjunctive descriptors.

  Decision trees naturally represent disjunctive expressions.

- In certain cases where the training data may contain errors.

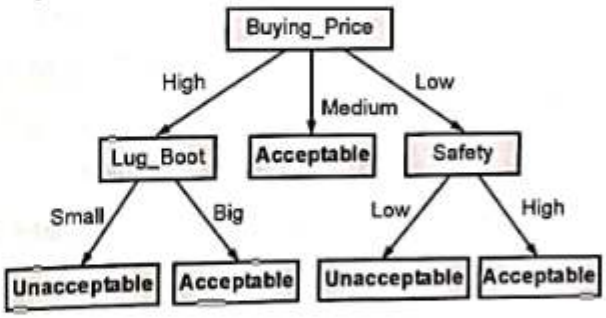
  Decision tree learning methods are tolerant to errors that can be a classification error of training records or attribute-value representation error.

- The training data may be incomplete as there are missing attribute values.

  Although some training records have unknown values, decision tree methods can be used.

2. Decision Tree Representation

- Decision tree is a classifier which is represented in the form of a tree structure where each node is either a leaf node or a decision node.
  - Leaf node represents the value of the target or response attribute (class) of examples.
  - Decision node represents some test to be carried out on a single attribute-value, with one branch and sub tree for each possible outcome of the test.
- Decision tree generates regression or classification models in the form of a tree structure. Decision tree divides a dataset into smaller subsets with increase in depth of tree.
- The final decision tree is a tree with decision nodes and leaf nodes. A decision node (e.g., Buying_Price) has two or more branches (e.g., High, Medium and Low). Leaf node (e.g., Evaluation) shows a classification or decision. The topmost decision node in a tree which represents the best predictor is called root node. Decision trees can be used to represent categorical as well as numerical data.
  - **Root Node :** It represents entire set of records or dataset and this is again divided into two or more similar sets.
  - **Splitting :** Splitting procedure is used to divide a node into two or more sub-nodes depending on the criteria.



Module 4

- **Decision Node :** A decision node is a sub-node which is divided into more sub-nodes.
- **Leaf/ Terminal Node :** Leaf node is a node which is not further divided or a node with no children.
- **Parent and Child Node :** Parent node is a node, which is split into sub-nodes and sub-nodes are called as child of parent node.
- **Branch / Sub-Tree :** A branch or sub-tree is a sub part of decision tree.
- **Pruning :** Pruning method is used to reduce the size of decision trees by removing nodes.

3. Attribute Selection Measure

1. Gini Index

- All attributes are assumed to be continuous valued.
- It is assumed that there exist several possible split values for each attribute.
- Gini index method can be modified for categorical attributes.
- Gini is used in Classification and Regression Tree (CART).

If a data set T contains example from n classes, gini index, gini (T) is defined as,

$$\text{gini}(T) = 1 - \sum_{j}^{n} = 1(P_j)^2 \quad \text{...(4.4.1)}$$

In the above equation P_j represents the relative frequency of class j in T.

After splitting T into two subsets T_1 and T_2 with sizes N_1 and N_2, gini index of split data is,

$$\text{gini}_{\text{split}}(T) = \frac{N_1}{N}\text{gini}(T_1) + \frac{N_2}{N}\text{gini}(T_2) \quad \text{... (4.4.2)}$$

The attribute with smallest $\text{gini}_{\text{split}}(T)$ is selected to split the node.

2. Information Gain (ID3)

In this method all attributes are assumed to be categorical. The method can be modified for continuous valued attributes. Here we select the attribute with highest information gain.

Assume there are 2 classes P and N. Let the set of records S contain p records of class P and n records of class N.

The amount of information required to decide if a random record in S belongs to P or N is defined as,

$$I(p, n) = -\left(\frac{p}{p+n}\right)\log_2\left(\frac{p}{p+n}\right) - \left(\frac{n}{p+n}\right)\log_2\left(\frac{n}{p+n}\right) \quad \text{... (4.4.3)}$$

Assume that using attribute A, a set S will be partitioned in to sets $\{s_1, s_2, ...s_k\}$

If S_i has p_i records of P and n_i records of N, the entropy or the expected information required to classify objects in all subtrees S_i is,

$$E(A) = \sum_{i}^{v} = 1\,\frac{p_i + n_i}{p+n}\,I(p_i, n_i) \quad \text{... (4.4.4)}$$

- **Entropy (E) :** Expected amount of information (in bits) needed to assign a class to a randomly drawn object in S under the optimal shortest length code.

- **Gain (A) :** Measures reduction in entropy achieved because of split. Choose split that achieves most reduction (maximum Gain).

$$\text{Gain (A)} = I(p, n) - E(A)$$

4. Avoid Overfitting in classification (Tree pruning) ...(4.4.5)

The generated tree may overfit the training data.

- If there are too many branches then some may reflect anomalies due to noise or outliers.
- Overfitting result in poor accuracy for unseen samples.

There are two approaches to avoid overfitting, prune the tree so that it is not too specific.

Prepruning (prune while building tree)

Stop tree construction early do not divide a node if this would result in the goodness measure falling below threshold.

Postpruning (prune after building tree)

Fully constructed tree get a sequence of progressively pruned trees.

5. Strengths of Decision Tree Method

- Able to generate understandable rules.
- Performs classification without requiring much computation.
- Able to handle both continuous and categorical variables.
- Decision tree clearly indicates which fields are most important for prediction or classification.

6. Weakness of Decision Tree Method

- Not suitable for prediction of continuous attribute.
- Perform poorly with many class and small data.
- Computationally expensive to train.

Module 4

4.5 CONSTRUCTING DECISION TREES

- The ID3 algorithm starts with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy (or information gain) of that attribute. Algorithm next selects the attribute which has the smallest entropy (or largest information gain) value.
- The set S is then divided by the chosen attribute (e.g. Income is less than 20 K , Income is between 20 K and 40 K, Income is greater than 40 K) to produce subsets of the data. The algorithm is recursively called for each subset, considering the attributes which are not selected before.
- The stopping criteria for recursion can be one of these situations :
  - When all records in the subset belongs to the same class (+ or –), then the node is converted into a leaf node and labelled with the class of the records.

- When we have selected all the attributes, but the records still do not belong to the same class (some are + and some are –), then the node is converted into a leaf node and labelled with the most frequent class of the records in the subset
- When there are no records in the subset, this is due to the non coverage of a specific attribute value for the record in the parent set, for example if there was no record with income = 40 K. Then a leaf node is generated, and labelled with the most frequent class of the record in the parent set.

– Decision tree is generated with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch.

Summary

Entropy of each and every attribute is calculated using the data set

– Divide the set S into subsets using the attribute for which the resulting entropy (after splitting) is minimum (or, equivalently, information gain is maximum)

– Make a decision tree node containing that attribute

– Recurse on subsets using remaining attributes.

Pseudocode

```
ID3 (Records, Target_Attribute, Attributes)
Generate a root node for the tree
If all records are positive, Return the single-node tree Root, with '+' label.
If all records are negative, Return the single-node tree Root, with '-' label.
If number of predicting attributes is empty, then return the single node tree Root, and label with most frequent value of the target attribute in the records.
Otherwise Begin
A ← The Attribute that best classifies records.
Decision Tree attribute for Root 'A'.
For each possible value, v_i, of A,
Add a new tree branch below Root, corresponding to the test A = v_i.
Let Record (v_i) be the subset of records that have the value v_i for A
If Record (v_i) is empty
Then below this new branch add a leaf node and label with most frequent target value in the records
Else below this new branch add the sub tree ID3 (Records (v_i), Target_Attribute,
Attributes – {A})
End
Return Root
```

4.6 EXAMPLE OF CLASSIFICATION TREE USING ID3

Example 4.6.1 : Suppose we want ID3 to evaluate car database as whether the car is acceptable or not. The target classification is "Should we accept car?" which can be acceptable or unacceptable.

Buying_Price	Maintenance_ Price	Lug_Boot	Safety	Evaluation?
High	High	Small	High	Unacceptable
High	High	Small	Low	Unacceptable
Medium	High	Small	High	Acceptable
Low	Medium	Small	High	Acceptable
Low	Low	Big	High	Acceptable
Low	Low	Big	Low	Unacceptable
Medium	Low	Big	Low	Acceptable
High	Medium	Small	High	Unacceptable
High	Low	Big	High	Acceptable
Low	Medium	Big	High	Acceptable
High	Medium	Big	Low	Acceptable
Medium	Medium	Small	Low	Acceptable
Medium	High	Big	High	Acceptable
Low	Medium	Small	Low	Unacceptable

☑ **Solution :**

Class P : Evaluation = "Acceptable"

Class N : Evaluation = "Unacceptable"

Total records = 14

No. of records with Acceptable = 9 and Unacceptable = 5

$$I(p, n) = -\left(\frac{p}{p+n}\right)\log_2\left(\frac{p}{p+n}\right)-\left(\frac{n}{p+n}\right)\log_2\left(\frac{n}{p+n}\right)$$

$$I(9, 5) = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right)-\left(\frac{5}{14}\right)\log_2 = 0.940$$

Step 1

1. Compute entropy for Buying_Price

For Buying_Price = High

$$p_i = 2 \text{ and } n_i = 3$$

$$I(p_i, n_i) = I(2, 3) = \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)-\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) = 0.971$$

Similarly we will calculate $I(p_i, n_i)$ for Medium and Low.

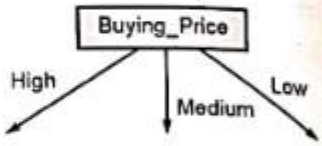
Buying_Price	p_i	n_i	$I(p_i, n_i)$
High	2	3	0.971
Medium	4	0	0
Low	3	2	0.971

Module 4

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Buying\_Price}) = \left(\frac{5}{14}\right) I(2,3) + \left(\frac{4}{14}\right) I(4,0) + \left(\frac{5}{14}\right) I(3,2) = 0.694$$

$$\text{Gain}(S, \text{Buying\_Price}) = I(p, n) - E(\text{Buying\_Price}) = 0.940 - 0.694 = 0.246$$



Similarly, Gain (S, Maintenance_ Price) = 0.029,

Gain (S, Lug_Boot) = 0.151, Gain (S, Safety) = 0.048

Since Buying_Price is the highest we select Buying_Price as the root node.

Step 2

As attribute Buying_Price at root, we have to decide on remaining tree attribute for High branch.

Buying_Price	Maintenance_ Price	Lug_Boot	Safety	Evaluation?
High	High	Small	High	Unacceptable
High	High	Small	Low	Unacceptable
High	Medium	Small	High	Unacceptable
High	Low	Big	High	Acceptable
High	Medium	Big	Low	Acceptable

No. of records with Acceptable = 2 and Unacceptable = 3

$$I(p, n) = -\left(\frac{p}{p+n}\right)\log_2\left(\frac{p}{p+n}\right) - \left(\frac{n}{p+n}\right)\log_2\left(\frac{n}{p+n}\right)$$

$$I(2, 3) = -\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) = 0.971$$

1. Compute entropy for Maintenance_ Price

Maintenance_ Price	p_i	n_i	$I(P_i, n_i)$
High	0	2	0
Medium	1	1	1
Low	1	0	0

$$E(\text{Maintenance\_ Price}) = \left(\frac{2}{5}\right) I(0,2) + \left(\frac{2}{5}\right) I(1,1) + \left(\frac{1}{5}\right) I(1,0) = 0.4$$

$$\text{Gain}(S_{\text{High}}, \text{Maintenance\_ Price}) = I(p, n) - E(\text{Maintenance\_ Price}) = 0.971 - 0.4 = 0.571$$

2. Compute entropy for Lug_Boot

$P_i = 0$ and $n_i = 3$

$I(P_i, n_i) = I(0, 3) = 0$

Lug_Boot	p_i	n_i	$I(P_i, n_i)$
Small	0	3	0
Big	2	0	0

$$E(\text{Lug\_Boot}) = \left(\frac{3}{5}\right) I(0, 3) + \left(\frac{2}{5}\right) I(2, 0) = 0$$

Gain (S_{High}, Lug_Boot) = I (p, n) – E (Lug_Boot) = 0.971 – 0 = 0.971

3. Compute entropy for Safety

$p_i = 1$ and $n_i = 2$

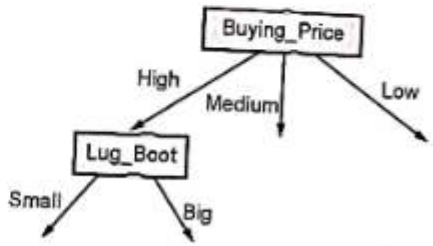
$I(p_i, n_i) = I(1, 2) = 0.918$

Safety	p_i	n_i	$I(P_i, n_i)$
High	1	2	0.918
Low	1	1	1

$$E(\text{Safety}) = \left(\frac{3}{5}\right) I(1, 2) + \left(\frac{2}{5}\right) I(1, 1) = 0.951$$

Gain(S_{High}, Safety) = I (p, n) – E (Safety) = 0.971 – 0.951 = 0.02

Since Lug_Boot is the highest we select Lug_Boot as a next node below High branch.

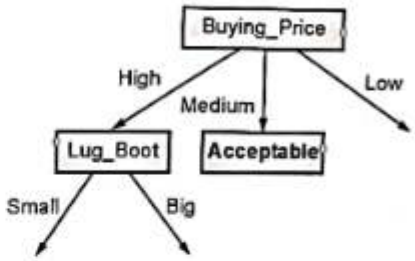


Step 3

Consider now only Maintenance_ Price and Safety for Buying_Price = Medium

Buying_Price	Maintenance_ Price	Lug_Boot	Safety	Evaluation?
Medium	High	Small	High	Acceptable
Medium	Low	Big	Low	Acceptable
Medium	Medium	Small	Low	Acceptable
Medium	High	Big	High	Acceptable

Since for any combination of values of aintenance Price and Safety, Evaluation? value is Acceptable, so we can directly write down the answer as Acceptable.



Step 4

Consider now only Maintenance_ Price and Safety for Buying_Price = Low

Buying_Price	Maintenance_ Price	Lug_Boot	Safety	Evaluation?
Low	Medium	Small	High	Acceptable
Low	Low	Big	High	Acceptable
Low	Low	Big	Low	Unacceptable
Low	Medium	Big	High	Acceptable
Low	Medium	Small	Low	Unacceptable

$P_i = 3$ and $n_i = 2$

$$I(P_i, n_i) = I(3,2) = 0.970$$

1. Compute entropy for Safety

Safety	p_i	n_i	$I(P_i, n_i)$
High	3	0	0
Low	0	2	0

$$E(\text{Safety}) = \left(\frac{3}{5}\right) I(3,0) + \left(\frac{2}{5}\right) I(0,2) = 0$$

$$\text{Gain}(S_{\text{Low}}, \text{Safety}) = I(p,n) - E(\text{Safety}) = 0.970 - 0 = 0.970$$

2. Compute entropy for Maintenance_ Price

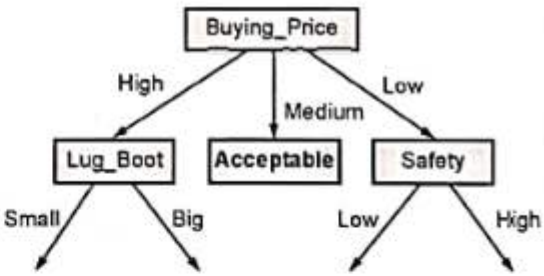
Maintenance_ Price	p_i	n_i	$I(P_i, n_i)$
High	0	0	0
Medium	2	1	0.918
Low	1	1	1

$$E(\text{Maintenance\_ Price}) = \left(\frac{0}{5}\right) I(0,0) + \left(\frac{3}{5}\right) I(2,1) + \left(\frac{2}{5}\right) I(1,1) = 0.951$$

$$\text{Gain}(S_{\text{Low}}, \text{Maintenance\_ Price}) = I(p,n) - E(\text{Maintenance\_ Price}) = 0.970 - 0.951 = 0.019$$

Since, Safety is the highest we select Safety below Low branch.

- Now we will check the value of 'Evaluation?' from the database, for all branches.
  - Buying_Price = High and Lug_Boot = Small -> Evaluation? = Unacceptable
  - Buying_Price = High and Lug_Boot = Big -> Evaluation? = Acceptable
  - Buying_Price = Low and Safety = Low -> Evaluation? = Unacceptable
  - Buying_Price = Low and Safety = High -> Evaluation? = Acceptable



- Final Decision Tree is

ample 4.6.2 : Suppose we want ID3 to decide whether the loan is to be sanctioned or not. The target classification is "Should we sanction loan?" which can be yes or no.

Customer no	Spending_ Habit	Collateral	Income	Credit_Score	Sanction?
1	High	None	Low	Bad	No
2	High	None	Medium	Unknown	No
3	Low	None	Medium	Unknown	No
4	Low	None	Low	Unknown	No
5	Low	None	High	Unknown	Yes
6	Low	Sufficient	High	Unknown	Yes
7	Low	None	Medium	Bad	No
8	Low	Sufficient	High	Bad	No
9	Low	None	High	Good	Yes
10	High	Sufficient	High	Good	Yes
11	High	None	Low	Good	No
12	High	None	Medium	Good	No

Solution :

Class P : Sanction = "Yes"

Class N : Sanction = "No"

Total records = 12

No. of records with Yes = 4 and No = 8

$$I(p, n) = -\left(\frac{p}{p+n}\right)\log_2\left(\frac{p}{p+n}\right)-\left(\frac{n}{p+n}\right)\log_2\left(\frac{n}{p+n}\right)$$

$$I(4, 8) = -\left(\frac{4}{12}\right)\log_2\left(\frac{4}{12}\right)-\left(\frac{8}{12}\right)-\log_2\left(\frac{8}{12}\right) = 0.922$$

Module 4

Step 1

Compute entropy for Spending_Habit

For Spending_Habit = High

$p_i = 1$ and $n_i = 4$

$$I(p_i, n_i) = I(1, 4) = -\left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right)-\left(\frac{4}{5}\right)\log_2\left(\frac{4}{5}\right) = 0.721$$

Similarly, we will calculate $I(p_i, n_i)$ for Low.

Spending_Habit	p_i	n_i	$I(p_i, n_i)$
High	1	4	0.721
Low	3	4	0.985

$$E(A) = \sum_{i=1}^{v}\frac{p_i+n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Spending\_Habit}) = \left(\frac{5}{12}\right)\times 0721 + \left(\frac{7}{12}\right)\times 0.985 = 0.874$$

Gain(S, Spending_Habit) = I (p, n) – E (Spending_Habit) = 0.922 – 0.874 = 0.048

2. Compute entropy for Collateral

Collateral	p_i	n_i	I (p_i, n_i)
None	2	7	0.77
Sufficient	2	1	0.918

$$E(A) = \sum_{i=1}^{v} = 1 \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Collateral}) = \left(\frac{9}{12}\right) \times 0.77 + \left(\frac{3}{12}\right) \times 0.918 = 0.806$$

Gain(S, Collateral) = I (p, n) – E (Collateral) = 0.922 – 0.806 = 0.116

3. Compute entropy for Income

Income	p_i	n_i	I (p_i, n_i)
Low	0	3	0
Medium	0	4	0
High	4	1	0.721

$$E(A) = \sum_{i=1}^{v} = 1 \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Income}) = \left(\frac{9}{12}\right) \times 0 + \left(\frac{4}{12}\right) \times 0 + \left(\frac{5}{12}\right) \times 0.721 = 0.3$$

Gain(S, Income) = I (p, n) – E (Income) = 0.922 – 0.3 = 0.622

4. Compute entropy for Credit_Score

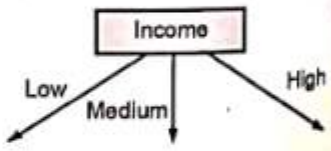
Credit_Score	p_i	n_i	I (p_i, n_i)
Bad	0	3	0
Unknown	2	3	0.97
Good	2	2	1

$$E(A) = \sum_{i=1}^{v} = 1 \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Credit\_Score}) = \left(\frac{3}{12}\right) \times 0 + \left(\frac{5}{12}\right) \times 0.97 + \left(\frac{4}{12}\right) \times 1 = 0.734$$

Gain(S, Credit_Score) = I (p, n) – E (Credit_Score)

= 0.922 – 0.734 = 0.188



Since Income is the highest we select Income as the root node.

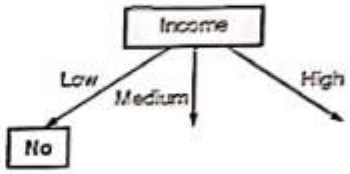
Step 2

As attribute Income at root, we have to decide on remaining tree attribute for Low branch.

Consider only Spending_Habit, Collateral and Credit_Score for Income = Low

Customer no	Spending_ Habit	Collateral	Income	Credit_Score	Sanction?
1	High	None	Low	Bad	No
4	Low	None	Low	Unknown	No
11	High	None	Low	Good	No

Since for any combination of values of Spending_Habit, Collateral and Credit_Score, Sanction? value is No for Income = Low, so we can directly write down the answer as No

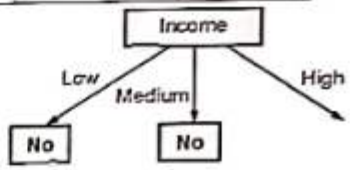


Step 3

Consider only Spending_Habit, Collateral and Credit_Score Income = Medium

Customer no	Spending_ Habit	Collateral	Income	Credit_Score	Sanction?
2	High	None	Medium	Unknown	No
3	Low	None	Medium	Unknown	No
7	Low	None	Medium	Bad	No
12	High	None	Medium	Good	No

Since for any combination of values of Spending_Habit, Collateral and Credit_Score, Sanction? value is No for Income = Medium, so we can directly write down the answer as No



Step 4

Consider only Spending_Habit, Collateral and Credit_Score Income = High

Customer no	Spending_ Habit	Collateral	Income	Credit_Score	Sanction?
5	Low	None	High	Unknown	Yes
6	Low	Sufficient	High	Unknown	Yes
8	Low	Sufficient	High	Bad	No
9	Low	None	High	Good	Yes
10	High	Sufficient	High	Good	Yes

No. of records with Yes = 4 and No = 1

$$I(p, n) = -\left(\frac{p}{p+n}\right)\log_2\left(\frac{p}{p+n}\right) - \left(\frac{n}{p+n}\right)\log_2\left(\frac{n}{p+n}\right)$$

$$I(2, 3) = -\left(\frac{4}{5}\right)\log_2\left(\frac{4}{5}\right) - \left(\frac{1}{5}\right)\log_2\left(\frac{1}{5}\right) = 0.721$$

Module 4

1. Compute entropy for Spending_Habit

Spending_Habit	p_i	n_i	$I(p_i, n_i)$
Low	3	1	0.811
High	1	0	0

$$E(A) = \sum_{i=1}^{v} = 1 \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Spending\_Habit}) = \left(\frac{4}{5}\right) \times 0811 + \left(\frac{1}{5}\right) \times 0 = 0.648$$

$$\text{Gain}(S_{High}, \text{Spending\_Habit}) = I(p, n) - E(\text{Spending\_Habit}) = 0.721 - 0.648 = 0.073$$

2. Compute entropy for Collateral

Collateral	p_i	n_i	$I(p_i, n_i)$
None	2	0	0
Sufficient	2	1	0.918

$$E(A) = \sum_{i=1}^{v} = 1 \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Collateral}) = \left(\frac{2}{5}\right) \times 0 + \left(\frac{3}{5}\right) \times 0.918 = 0.55$$

$$\text{Gain}(S_{High}, \text{Collateral}) = I(p, n) - E(\text{Collateral}) = 0.721 - 0.55 = 0.171$$

3. Compute entropy for Credit_Score

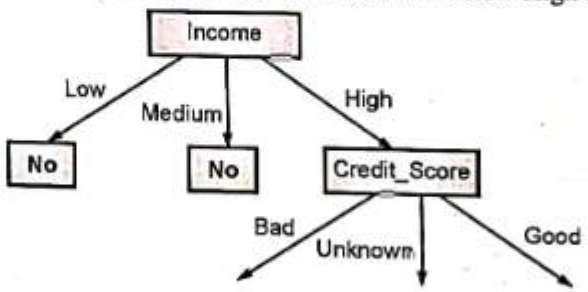
Credit_Score	p_i	n_i	$I(p_i, n_i)$
Unknown	1	0	0
Bad	0	1	0
Good	2	0	0

$$E(A) = \sum_{i=1}^{v} = 1 \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

$$E(\text{Credit\_Score}) = \left(\frac{1}{5}\right) \times 0 + \left(\frac{1}{5}\right) \times 0 + \left(\frac{2}{5}\right) \times 0 = 0$$

$$\text{Gain}(S_{High}, \text{Credit\_Score}) = I(p, n) - E(\text{Credit\_Score}) = 0.721 - 0 = 0.721$$

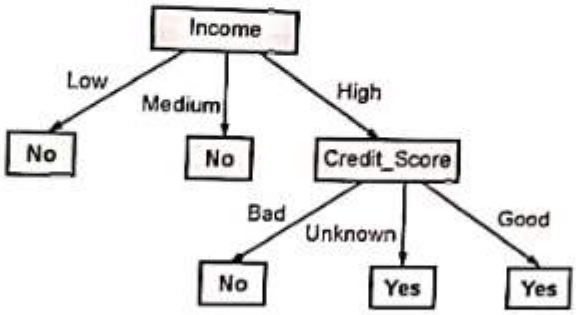
– Since Credit_Score is the highest we select Credit_Score as a next node below High branch.



Now we will check the value of 'Sanction?' from the database, for all branches,

- Income = High and Credit_Score = Bad -> Sanction? = No
- Income = High and Credit_Score = Unknown -> Sanction? = Yes
- Income = High and Credit_Score = Good -> Sanction? = Yes

– Final Decision Tree is



Example 4.6.3 : Suppose we want ID3 to decide whether the car will be stolen or not. The target classification is "car is stolen?" which can be Yes or No.

Car no	Colour	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Module 4

☑ **Solution :**

Class P: Stolen = "Yes" Class N: Stolen = "No"

Total records = 10

No. of records with Yes = 5 and No = 5

$$I(p, n) = \left(\frac{p}{p+n}\right)\log_2\left(\frac{p}{p+n}\right) - \left(\frac{n}{p+n}\right)\log_2\left(\frac{n}{p+n}\right)$$

$$I(5, 5) = -\left(\frac{5}{10}\right)\log_2\left(\frac{5}{10}\right) - \left(\frac{5}{10}\right)\log_2\left(\frac{5}{10}\right) = 1$$

Step 1

1. Compute entropy for Colour

Colour	p_i	n_i	$I(p_i, n_i)$
Red	3	2	0.971
Yellow	2	3	0.971

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Colour}) = \left(\frac{5}{10}\right) \times 0.971 + \left(\frac{5}{10}\right) \times 0.971 = 0.971$$

$$\text{Gain}(S, \text{Colour}) = I(p, n) - E(\text{Colour}) = 1 - 0.971 = 0.029$$

2. Compute entropy for Type

Type	p_i	n_i	$I(p_i, n_i)$
Sports	4	2	0.923
SUV	1	3	0.811

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Type}) = \left(\frac{6}{10}\right) \times 0.923 + \left(\frac{4}{10}\right) \times 0.811 = 0.878$$

$$\text{Gain}(S, \text{Type}) = I(p, n) - E(\text{Type}) = 1 - 0.878 = 0.1218$$

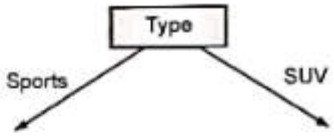
3. Compute entropy for Origin

Origin	p_i	n_i	$I(p_i, n_i)$
Domestic	2	3	0.971
Imported	3	2	0.971

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Origin}) = \left(\frac{5}{10}\right) \times 0.971 + \left(\frac{5}{10}\right) \times 0.971 = 0.971$$

$$\text{Gain}(S, \text{Origin}) = I(p, n) - E(\text{Origin}) = 1 - 0.971 = 0.029$$



Since Type is the highest we select Type as the root node.

Step 2

As attribute Type at root, we have to decide on remaining tree attribute for Sports branch.

Consider only Colour and Origin for Type = Sports

Car no	Colour	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
10	Red	Sports	Imported	Yes

No. Of records with Yes = 4 and No = 2

$$I(p, n) = \left(\frac{p}{p+n}\right)\log_2\left(\frac{p}{p+n}\right) - \left(\frac{n}{p+n}\right)\log_2\left(\frac{n}{p+n}\right)$$

$$I(4, 2) = -\left(\frac{4}{6}\right)\log_2\left(\frac{4}{6}\right) - \left(\frac{2}{6}\right)\log_2\left(\frac{2}{6}\right) = 0.923$$

Compute entropy for Colour

Colour	p_i	n_i	$I(p_i, n_i)$
Red	3	1	0.811
Yellow	1	1	1

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Colour}) = \left(\frac{4}{6}\right) \times 0.811 + \left(\frac{2}{6}\right) \times 1 = 0.873$$

$$\text{Gain}(S_{\text{Sports}}, \text{Colour}) = I(p, n) - E(\text{Colour}) = 0.923 - 0.873 = 0.05$$

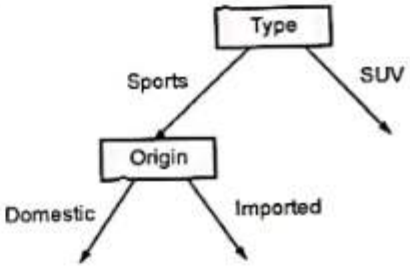
Compute entropy for Origin

Origin	p_i	n_i	$I(p_i, n_i)$
Domestic	2	2	1
Imported	2	0	0

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{Origin}) = \left(\frac{4}{6}\right) \times 1 + \left(\frac{2}{6}\right) \times 0 = 0.666$$

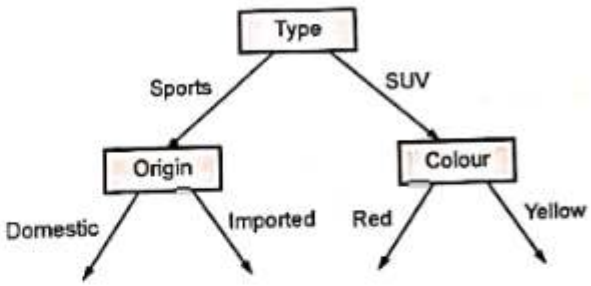
$$\text{Gain}(S_{\text{Sports}}, \text{Origin}) = I(p, n) - E(\text{Origin}) = 0.923 - 0.666 = 0.257$$



Since Origin is the highest we select as a next node below Sports branch.

Step 3

As attribute Type and Origin is already chosen, we have to decide on only remaining Colour attribute for SUV branch.



Now we will check the value of 'Stolen?' from the database, for all branches,

- For, Type = Sports and Origin = Domestic, Stolen? = Yes as well as No

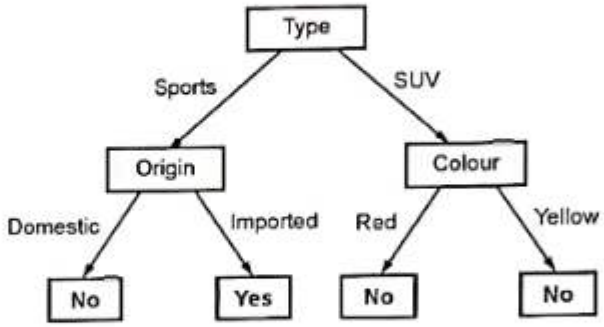
So for this type of case we have to select the most common class. In this example there are 2 instances for Yes as well as No, so we can select any one. Let's we select No.

- For, Type = Sports and Origin = Imported, Stolen? = Yes
- For, Type = SUV and Colour = Red, Stolen? = No

For, Type = SUV and Colour = Yellow, Stolen? = Yes as well as No

Module 4

- So for this type of case we have to select the most common class. In this example there are 2 instances for No and 1 instance of Yes, so we will select No.
- Final Decision Tree is



4.7 EXAMPLE OF DECISION TREE USING GINI INDEX

Example 4.7.1 : Create a decision tree using Gini Index to classify following dataset.

Sr. No.	Income	Age	Own Car
1	Very High	Young	Yes
2	High	Medium	Yes
3	Low	Young	No
4	High	Medium	Yes
5	Very High	Medium	Yes
6	Medium	Young	Yes
7	High	Old	Yes
8	Medium	Medium	No
9	Low	Medium	No
10	Low	Old	No
11	High	Young	Yes
12	Medium	Old	No

☑ **Solution :**

- In this example there are two classes Yes and No.

  No. of records for Yes = 7

  No. of records for No = 5

  Total No. of records = 12

- Now we will calculate Gini of the complete database as,

$$\text{Gini (T)} = 1 - \left(\left(\frac{7}{12}\right)^2 + \left(\frac{5}{12}\right)^2\right) = 0.48$$

- Next we will calculate Split for all attributes, i.e. Income and Age.

Income

$$\text{Split} = \frac{2}{12}\text{gini (Very High)} + \frac{4}{12}\text{gini (High)} + \frac{3}{12}\text{gini (Low)} + \frac{3}{12}\text{gini (Medium)}$$

$$= \frac{2}{12}\left[1-\left(\left(\frac{2}{2}\right)^2+\left(\frac{0}{2}\right)^2\right)\right]+\frac{4}{12}\left[1-\left(\left(\frac{4}{4}\right)^2+\left(\frac{0}{4}\right)^2\right)\right]+\frac{3}{12}\left[1-\left(\left(\frac{0}{3}\right)^2+\left(\frac{3}{3}\right)^2\right)\right]+\frac{3}{12}\left[1-\left(\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2\right)\right]$$

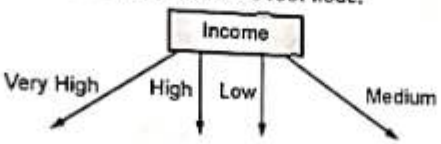
$$= 0.1125$$

$$\text{Age Split} = \frac{4}{12}\text{ gini (Yong)} + \frac{5}{12}\text{gini (Medium)} + \frac{3}{12}\text{gini (Old)}$$

$$= \frac{4}{12}\left[1-\left(\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right)\right]+\frac{5}{12}\left[1-\left(\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right)\right]+\frac{3}{12}\left[1-\left(\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2\right)\right]$$

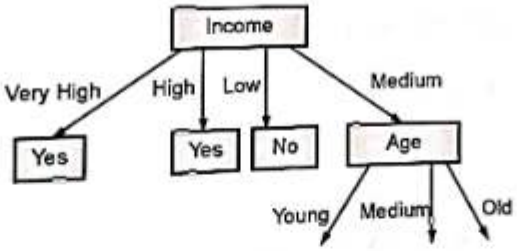
$$= 0.4375$$

- Split value of Income is smallest, so we will select Income as root node.



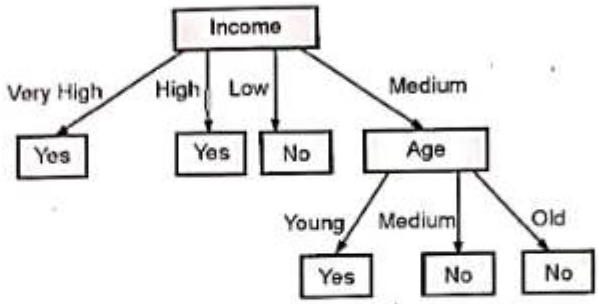
- From the database we can see that,
  - Own Car = Yes for Income = Very High, so we can directly write down 'Yes' for Very High branch.
  - Own Car = Yes for Income = High, so we can directly write down 'Yes' for High branch.
  - Own Car = No for Income = Low, so we can directly write down 'No' for Low branch.
- Since Income is taken as root node, now we have to decide on the Age attribute, so we will take Age as next node below Medium branch.

Module 4



- From the database we can see that,
  - Own Car = Yes for Income = Medium and Age = Young, so we can directly write down 'Yes' for Young branch.
  - Own Car = No for Income = Medium and Age = Medium, so we can directly write down 'No' for medium branch.
  - Own Car = No for Income = Medium and Age = Old, so we can directly write down 'No' for Old branch.

- Final Decision Tree is,



Example 4.7.2 : Stock market involving only Discrete ranges has profit as categorical value {up, down}.Use Gini index method to draw classification tree.

Age	Competition	Type	Profit
old	Yes	software	down
old	No	software	down
old	No	hardware	down
mid	Yes	software	down
mid	Yes	hardware	down
mid	No	hardware	up
mid	No	software	up
new	Yes	software	up
new	No	hardware	up
new	no	software	up

☑ **Solution :**

- In this example there are two classes down and up.

  No. of records for down = 5

  No. of records for up = 5

  Total No. of records = 10

- Now we will calculate Gini of the complete database as,

$$\text{Gini (T)} = \left[1-\left(\left(\frac{5}{10}\right)^2+\left(\frac{5}{10}\right)^2\right)\right]=0.5$$

- Next we will calculate Split for all attributes, i.e. Age, Competition and Type.

Age

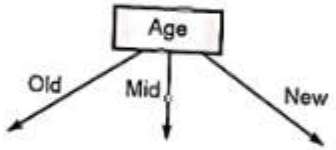
$$\text{Split} = \frac{3}{10}\text{gini (old)} + \frac{4}{10}\text{gini (mid)}\frac{3}{10}\text{ gini (new)}$$

$$= \frac{3}{10}\left[1-\left(\left(\frac{0}{3}\right)^2+\left(\frac{3}{3}\right)^2\right)\right]+\frac{4}{10}\left[1-\left(\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right)\right]+\frac{3}{10}\left[1-\left(\left(\frac{3}{3}\right)^2+\left(\frac{0}{3}\right)^2\right)\right]=0.2$$

Competition

$$\text{Split} = \frac{4}{10}\text{gini (yes)} + \frac{6}{10}\text{gini (no)}$$

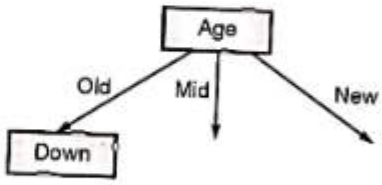
$$= \frac{4}{10}\left[1-\left(\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2\right)\right]+\frac{6}{10}\left[1-\left(\left(\frac{4}{6}\right)^2+\left(\frac{2}{6}\right)^2\right)\right]=0.42$$



Type

$$\text{Split} = \frac{6}{10}\text{gini (software)} + \frac{4}{10}\text{gini (hardware)}$$

$$= \frac{6}{10}\left[1-\left(\left(\frac{3}{6}\right)^2+\left(\frac{3}{6}\right)^2\right)\right]+\frac{4}{10}\left[1-\left(\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right)\right]=0.5$$



- Split value of Age is smallest, so we will select Age as root node.
- From the database we can see that, Profit = Down for Age = Old, so we can directly write down 'Down' for Old branch node.
- Next we will check for Age = mid

Age	Competition	Type	Profit
mid	Yes	software	Down
mid	Yes	hardware	Down
mid	No	hardware	Up
mid	No	software	Up

- Now we will calculate Split for only Competition and Type attributes.

Competition

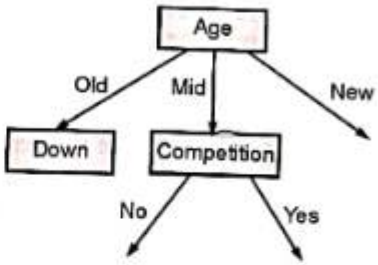
$$\text{Split} = \frac{2}{4}\text{gini (yes)} + \frac{2}{4}\text{gini (no)}$$

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

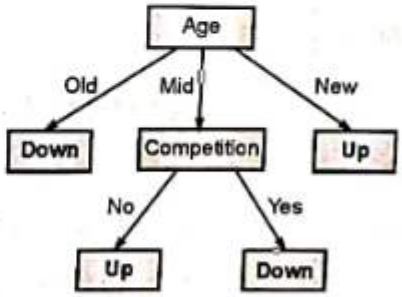
Type

$$\text{Split} = \frac{2}{4}\text{gini (software)} + \frac{2}{4}\text{gini (hardware)}$$

$$= \frac{2}{4}\left[1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)\right]+\frac{2}{4}\left[1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)\right]=0.5$$

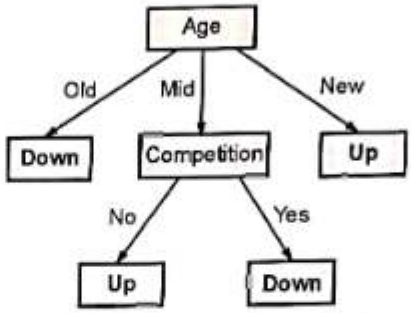


- Split value of Competition is smallest, so we will select Competition as next node below mid branch.
- From the database we can see that, Profit = Up for Age = New, so we can directly write down 'Up' for New branch node.
- Now we will check the value of 'Profit' from the database, for all branches,
  - Age = Mid and Competition = No -> Profit = Up
  - Age = Mid and Competition = Yes -> Profit = Down



Module 4

- Final Decision Tree is,



Example 4.7.3 : Suppose we want Gini index to decide whether the car will be stolen or not. The target classification is "car is stolen?" which can be Yes or No.

Car no	Colour	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

☑ **Solution :**

- In this example there are two classes Yes and No.

  No. of records for Yes = 5

  No. of records for No = 5

  Total No. of records = 10

- Now we will calculate Gini of the complete database as,

$$\text{Gini (T)} = 1-\left(1-\left(\left(\frac{5}{10}\right)^2+\left(\frac{5}{10}\right)^2\right)\right)=0.5$$

- Next we will calculate Split for all attributes, i.e. Colour, Type and Origin.

Colour

$$\text{Split} = \frac{2}{4}$$

$$= \frac{5}{10}\left[1-\left(\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right)\right]+\frac{5}{10}\left[1-\left(\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right)\right]=0.48$$

Type

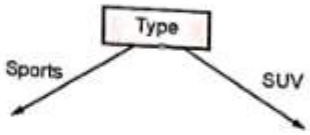
$$\text{Split} = \frac{6}{10}\text{gini (Sports)}+\frac{4}{10}\text{ gini (SUV)}$$

$$= \frac{6}{10}\left[1-\left(\left(\frac{4}{6}\right)^2+\left(\frac{2}{6}\right)^2\right)\right]+\frac{4}{10}\left[1-\left(\left(\frac{1}{4}\right)^2+\left(\frac{3}{4}\right)^2\right)\right]=0.42$$

Origin

$$Split = \frac{5}{10}\,gini\,(Domestic) + \frac{5}{10}\,gini\,(imported)$$

$$= \frac{5}{10}\left[1-\left(\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right)\right]+\frac{5}{10}\left[1-\left(\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right)\right]=0.48$$



- Split value of Type is smallest, so we will select Type as root node.
- Next we will check for Type = Sports
- As attribute Type at root, we have to decide on remaining tree attribute for Sports branch.
- Consider only Colour and Origin for Type = Sports

Car no	Colour	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
10	Red	Sports	Imported	Yes

Colour

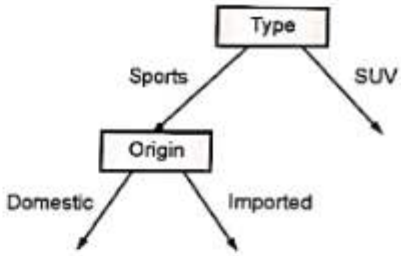
$$Split = \frac{4}{6}\,gini\,(Red) + \frac{2}{6}\,gini\,(Yellow)$$

$$= \frac{4}{6}\left[1-\left(\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right)\right]+\frac{2}{6}\left[1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)\right]=0.417$$

Origin>

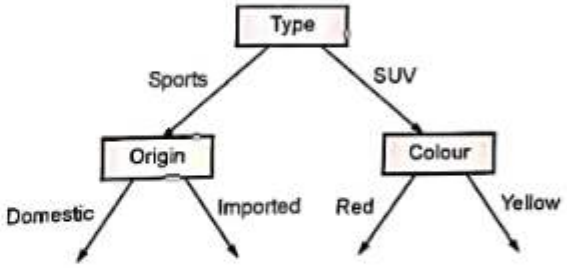
$$Split = \frac{4}{6}\,gini\,(Domestic) + \frac{4}{6}\,gini\,(Imported)$$

$$= \frac{4}{6}\left[1-\left(\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right)\right]+\frac{2}{6}\left[1-\left(\left(\frac{2}{2}\right)^2+\left(\frac{0}{2}\right)^2\right)\right]=0.33$$



- Split value of Origin is smallest, so we will select Origin as next node.
- Next we will check for Type = SUV
- As attribute Type and Origin is already chosen, we have to decide on only remaining Colour attribute for SUV branch.
- Now we will check the value of 'Stolen?' from the database, for all branches,

  For, Type = Sports and Origin = Domestic, Stolen? = Yes as well as No
- So for this type of case we have to select the most common class. In this example there are 2 instances for Yes as well as No, so we can select any one. Let's we select No.



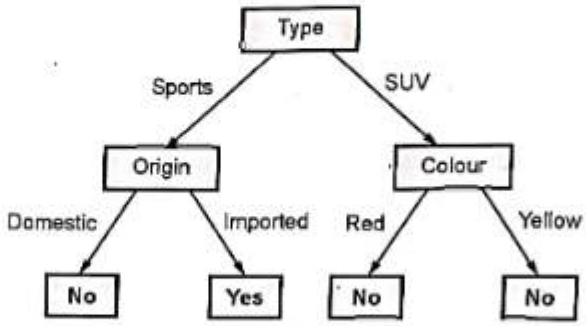
  - For, Type = Sports and Origin = Imported, Stolen? = Yes
  - For, Type = SUV and Colour = Red, Stolen? = No

Module 4

- For, Type = SUV and Colour = Yellow, Stolen? = Yes as well as No

- So for this type of case we have to select the most common class (Since all attributes are already considered). In this example there are 2 instances for No and 1 instance of Yes, so we will select No.

- Final Decision Tree is



UExample 4.7.4 **MU - May 15, 12 Marks**

Create a decision tree for the attribute "class" using the respective values :

Eyecolour	Married	Sex	Hairlength	class
Brown	yes	Male	Long	Football
Blue	yes	Male	Short	Football
Brown	yes	Male	Long	Football
Brown	no	Female	Long	Netball
Brown	no	Female	Long	Netball
Blue	no	Male	Long	Football
Brown	no	Female	Long	Netball
Brown	no	Male	Short	Football
Brown	yes	Female	Short	Netball
Brown	no	Female	Long	Netball
Blue	no	Male	Long	Football
Blue	no	Male	Short	Football

☑ **Solution :**

In this example there are two classes Football and Netball.

No. Of records for Football = 7

No. Of records for Netball = 5

Total No. Of records = 12

Now we will calculate Gini of the complete database as,

$$\text{Gini (T)} = 1 - \left(\left(\frac{7}{12}\right)^2 + \left(\frac{5}{12}\right)^2\right) = 0.48$$

Next we will calculate Split for all attributes, i.e. Eyecolor, Married, Sex and Hairlength.

Eyecolor->

$$\text{Split} = \frac{8}{12}\text{gini (Brown)} + \frac{4}{12}\text{gini (Blue)}$$

$$= \frac{8}{12}\left[1-\left(\left(\frac{3}{8}\right)^2+\left(\frac{5}{8}\right)^2\right)\right]+\frac{4}{12}\left[1-\left(\left(\frac{4}{4}\right)^2+\left(\frac{0}{4}\right)^2\right)\right]=0.31$$

Married->

$$\text{Split} = \frac{4}{12}\text{gini (yes)} + \frac{8}{12}\text{gini (no)}$$

$$= \frac{4}{12}\left[1-\left(\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right)\right]+\frac{8}{12}\left[1-\left(\left(\frac{4}{8}\right)^2+\left(\frac{4}{8}\right)^2\right)\right]=0.458$$

Sex->

$$\text{Split} = \frac{7}{12}\text{gini (male)} + \frac{5}{12}\text{gini (female)}$$

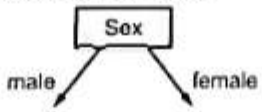
$$= \frac{7}{12}\left[1-\left(\left(\frac{7}{7}\right)^2+\left(\frac{0}{7}\right)^2\right)\right]+\frac{5}{12}\left[1-\left(\left(\frac{0}{5}\right)^2+\left(\frac{5}{5}\right)^2\right)\right]=0$$

Hairlength->

$$\text{Split} = \frac{8}{12}\text{gini (long)} + \frac{4}{12}\text{gini (short)}$$

$$= \frac{8}{12}\left[1-\left(\left(\frac{4}{8}\right)^2+\left(\frac{4}{8}\right)^2\right)\right]+\frac{4}{12}\left[1-\left(\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right)\right]=0.458$$

Split value of Sex is smallest, so we will select Sex as root node.



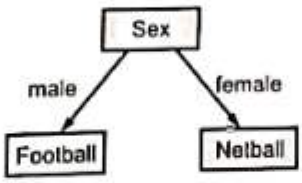
Module 4

From the database we can see that,

class = Football for Sex = male, so we can directly write down 'Football' for male branch.

class = Netball for Sex = female, so we can directly write down 'Netball' for female branch.

Final decision tree is,



UExample 4.7.5 **MU - May 17, 10Marks**

For a Sunburn dataset given below, construct a decision tree

Name	Hair	Height	Weight	Location	Class
Swati	Blonde	Average	Light	No	Yes
Sunita	Blonde	Tall	Average	Yes	No
Anita	Brown	Short	Average	Yes	No
Lata	Blonde	Short	Average	No	Yes
Radha	Red	Average	Heavy	No	Yes
Maya	Brown	Tall	Heavy	No	No
Leena	Brown	Average	Heavy	No	No
Rina	Blonde	Short	Light	Yes	No

☑ **Solution :**

We will calculate Split for all attributes, i.e. Hair, Height, Weight and Location.

Hair->

$$\text{Split} = \frac{4}{8}\text{gini (Blonde)} + \frac{3}{8}\text{gini (Brown)} + \frac{1}{8}\text{gini (Red)}$$

$$= \frac{4}{8}\left[1-\left(\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right)\right]+\frac{3}{8}\left[1-\left(\left(\frac{0}{3}\right)^2+\left(\frac{3}{3}\right)^2\right)\right]+\frac{1}{8}\left[1-\left(\left(\frac{1}{1}\right)^2+\left(\frac{0}{1}\right)^2\right)\right]=0.25$$

Height->

$$\text{Split} = \frac{3}{8}\text{gini (Average)} + \frac{2}{8}\text{gini (Tall)} + \frac{3}{8}\text{gini (Short)}$$

$$= \frac{3}{8}\left[1-\left(\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right)\right]+\frac{2}{8}\left[1-\left(\left(\frac{0}{2}\right)^2+\left(\frac{2}{2}\right)^2\right)\right]+\frac{3}{12}\left[1-\left(\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2\right)\right]=0.40$$

Weight->

$$\text{Split} = \frac{2}{8}\text{gini (Light)} + \frac{3}{8}\text{gini (Average)} + \frac{3}{8}\text{gini (Heavy)}$$

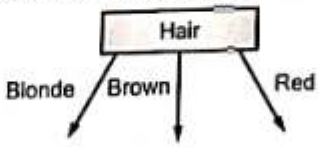
$$= \frac{2}{8}\left[1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)\right]+\frac{3}{8}\left[1-\left(\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2\right)\right]+\frac{3}{8}\left[1-\left(\left(\frac{1}{3}\right)^2+\left(\frac{2}{3}\right)^2\right)\right]=0.525$$

Location->

$$\text{Split} = \frac{5}{8}\text{gini (No)} + \frac{3}{8}\text{gini (Yes)}$$

$$= \frac{5}{8}\left[1-\left(\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right)\right]+\frac{3}{8}\left[1-\left(\left(\frac{0}{3}\right)^2+\left(\frac{3}{3}\right)^2\right)\right]=0.3$$

Split value of Hair is smallest, so we will select Hair as root node.



Now we will split the remaining attributes considering Blonde data.

Height->

$$\text{Split} = \frac{0}{4}\text{gini (Average)} + \frac{1}{4}\text{gini (Tall)} + \frac{2}{4}\text{gini (Short)}$$

$$= \frac{1}{4}\left[1-\left(\left(\frac{1}{1}\right)^2+\left(\frac{0}{1}\right)^2\right)\right]+\frac{1}{4}\left[1-\left(\left(\frac{0}{1}\right)^2+\left(\frac{1}{1}\right)^2\right)\right]+\frac{2}{4}\left[1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)\right]=0.25$$

Weight->

$$\text{Split} = \frac{2}{4}\text{gini (Light)} + \frac{2}{4}\text{gini (Average)}$$

$$= \frac{2}{4}\left[1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)\right]+\frac{2}{4}\left[1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right)\right]=0.5$$

Location->

$$\text{Split} = \frac{2}{4}\text{gini (No)} + \frac{2}{4}\text{gini (Yes)}$$

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

Split value of Location is smallest, so we will select Location node below Blonde branch.

From the database we can see that,

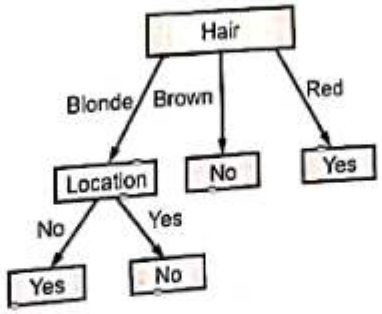
class = Yes for Hair = Blonde and Location = No , so we can directly write down 'Yes' for No branch.

class = No for Hair = Blonde and Location = Yes, so we can directly write down 'No' for Yes branch.

class = Yes for Hair = Red , so we can directly write down 'Yes' for Red branch.

class = No for Hair = Brown, so we can directly write down 'No' for Hair branch.

Final Decision Tree is,



Module 4

UExample 4.7.6 **MU - May 19, 10 Marks**

For a Sunburn dataset given below, construct a decision tree For the following data, Calculate Gini indexes and determines which attribute is root attribute and generate two level deep decision tree.

Sr. No.	Income	Defaulting	Credit score	Location	Give Loan?
1	Low	High	High	bad	No
2	Low	High	High	good	No
3	High	High	High	bad	Yes
4	Medium	Medium	High	bad	Yes
5	Medium	Low	Low	bad	No
6	Medium	Low	Low	good	Yes
7	High	Low	Low	good	Yes
8	Low	Medium	High	bad	No
9	Low	Low	Low	bad	No
10	Medium	Medium	Low	bad	No
11	Low	Medium	Low	good	Yes
12	High	Medium	High	good	Yes
13	High	High	Low	bad	No
14	Medium	Medium	High	good	Yes

☑ **Solution :**

We will calculate Split for all attributes, i.e. Income, Defaulting, Creditscore and Location.

Income-> $\text{Split} = \frac{5}{14}\text{gini (Low)} + \frac{4}{14}\text{gini (High)} + \frac{5}{14}\text{gini (Medium)}$

$$= \frac{5}{14}\left[1-\left(\left(\frac{1}{5}\right)^2+\left(\frac{4}{5}\right)^2\right)\right]+\frac{4}{14}\left[1-\left(\left(\frac{3}{4}\right)^2+\left(\frac{1}{4}\right)^2\right)\right]+\frac{5}{14}\left[1-\left(\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right)\right]=0.392$$

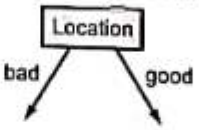
Defaulting-> $\text{Split} = \frac{4}{14}\text{gini (High)} + \frac{6}{14}\text{gini (Medium)} + \frac{4}{14}\text{gini (Low)} = 0.438$

Creditscore-> $\text{Split} = \frac{7}{14}\text{gini (High)} + \frac{7}{14}\text{gini (Low)} = 0.493$

Location-> $\text{Split} = \frac{8}{14}\text{gini (bad)} + \frac{6}{14}\text{gini (good)}$

$$= \frac{5}{8}\left[1-\left(\left(\frac{3}{5}\right)^2+\left(\frac{2}{5}\right)^2\right)\right]+\frac{3}{8}\left[1-\left(\left(\frac{0}{3}\right)^2+\left(\frac{3}{3}\right)^2\right)\right]=0.336$$

Split value of Location is smallest, so we will select Location as root node.



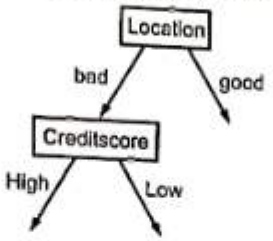
Now we will split the bad branch considering remaining attributes

Income-> $\text{Split} = \frac{3}{8}\text{gini (Low)} + \frac{2}{8}\text{gini (High)} + \frac{3}{8}\text{gini (Medium)} = 0.295$

Defaulting-> $\text{Split} = \frac{3}{8}\text{gini (High)} + \frac{3}{8}\text{gini (Medium)} + \frac{2}{8}\text{gini (Low)} = 0.34$

Creditscore-> $\text{Split} = \frac{4}{8}\text{gini (High)} + \frac{4}{8}\text{gini (Low)} = 0.25$

Split value of Creditscore is smallest, so we will select Creditscore node below bad branch.



Now we will split the good branch considering remaining attributes

Income-> Split $= \frac{2}{6}$ gini (Low) $+ \frac{2}{6}$ gini (High) $+ \frac{2}{6}$ gini (Medium) $= 0.295$

Defaulting-> Split $= \frac{1}{6}$ gini (High) $+ \frac{2}{6}$ gini (Medium) $+ \frac{3}{6}$ gini (Low) $= 0$

Split value of Defaulting is smallest, so we will select Defaulting node below good branch

Since only one attribute is remaining, we can directly select Income below creditscore= High branch

For Location = bad and creditscore = High and Income = Low, Giveloan= No

For Location = bad and creditscore = High and Income = Medium, Giveloan= Yes

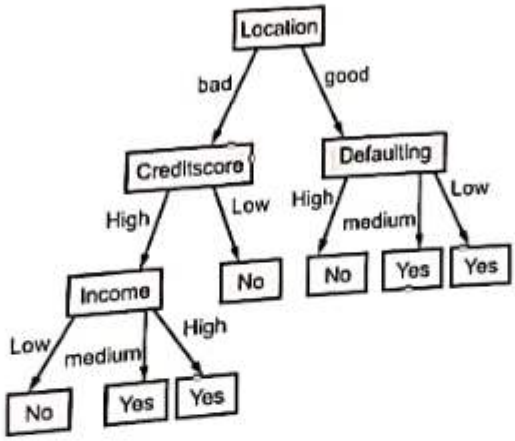
For Location = bad and creditscore = High and Income = High, Giveloan= Yes

For Location = bad and creditscore = Low, Giveloan= No

For Location = good and Defaulting = High, Giveloan= No

For Location = good and Defaulting = Low, Giveloan= Yes

For Location = good and Defaulting = Medium, Giveloan= yes



Module 4

4.8 CLASSIFICATION AND REGRESSION TREE (CART)

- Classification trees are used to divide the dataset into classes belonging to the target variable. Mainly the target variable has two classes that can be yes or no. When the target variable type is categorical classification trees are used.
- In certain applications the target variable is numeric or continuous in that case regression trees are used. Let's take an example of prediction of price of a flat. Hence regression trees are used for problems or tasks where we want to predict some data instead of classifying the data.
- Based on the similarity of the data the records are classified in a standard classification tree. Let's take an example of an Income tax evades. In this example we have two variables, Income and marital status that predict if a person is going to evade the income tax or not. In our training data if it showed that 85% of people who are married does not evade the income tax, we split the data here and Marital status becomes a root node in tree. Entropy or Ginny index is used in classification trees.
- The main basic working of regression tree is to fit a model. The target or response variable does not have classes so a regression model is fit using each independent variable to the target variable. Then the data is split at various split points for each independent variable. At each split point sum of squared errors (SSE) is calculated by taking the square of the difference between predicted and actual value. The criteria for root node is to select the node which is having minimum SSE among all split point errors. The further tree is built using the recursive procedure.

4.9 EXAMPLE OF REGRESSION TREE

Example 1

Buying_Price	Lug_Boot	Safety	Maintenance_ Price? (in thousand)
Low	Small	High	25
Low	Small	Low	30
Medium	Small	High	46
High	Small	High	45
High	Big	High	52
High	Big	Low	23
Medium	Big	Low	43
Low	Small	High	35
Low	Big	High	38
High	Big	High	46
Low	Big	Low	48
Medium	Small	Low	52
Medium	Big	High	44
High	Small	Low	30

Standard deviation

A decision tree is built up top down from root node and involved partitioning the data into subsets that contain instances with similar values. We use SD to calculate homogeneity of a numerical sample.

$$SD, S = \sqrt{\frac{\sum (x-\mu)^2}{n}} = 9.32$$

SD Reduction

It is based on the decrease in SD after a dataset is split on an attribute. Constructing a tree is all about finding attribute that returns highest SDR.

- **Step 1 :**

SD (Maintenance_Price?) = 9.32

- **Step 2 :**

The dataset is then split on the different attribute.SD for each branch is calculated. The resulting SD is subtracted from SD before split.

		Maintenance_ Price(SD)
Buying_Price	Low	7.78
	Medium	3.49
	High	10.87

SD (Maintenance_Price, Buying_Price) = P (Low) SD (Low) + P (Medium) SD (Medium) + P (High) SD (High)

$= \frac{5}{14} \times 7.78 + \frac{4}{14} \times 3.49 + \frac{5}{14} \times 10.87 = 7.66$

SDR = SD (Maintenance_ Price) – SD (Maintenance_ Price, Buying_Price) = 9.32 – 7.66 = 1.66

		Maintenance_ Price (SD)
Lug_Boot	Small	9.36
	Big	8.37

SD (Maintenance_ Price, Lug_Boot) = P (Small) SD (Small) + P (Big) SD (Big)

$= \frac{7}{14} \times 9.36 + \frac{7}{14} \times 8.37 = 8.86$

SDR = SD (Maintenance_ Price) – SD (Maintenance_ Price, Lug_Boot) = 9.32 – 8.86 = 0.46

		Maintenance_ Price (SD)
Safety	High	7.87
	Low	10.59

SD (Maintenance_ Price, Safety) = P (High) SD (High) + P (Low) SD (Low)

$= \frac{8}{14} \times 7.87 + \frac{6}{14} \times 10.59 = 9.02$

SDR = SD (Maintenance_ Price) – SD (Maintenance_ Price, Safety) = 9.32 – 9.02 = 0.3

Module 4

SDR of Buying_Price is highest so we select Buying_Price as our root node.

To avoid over fitting we should terminate unnecessary building branches. For example if there are less than five instances in the sub data set or standard deviation can be less than 5% of the entire data set. I prefer to apply the first one. I will terminate the branch if there are less than 5 instances in current sub data set. If this termination condition is satisfied then i will calculate average of sub data set.

Step 2(a) :

Now we will consider the records of 'High'.

Buying_Price	Lug_Boot	Safety	Maintenance_ Price? (in thousand)
High	Small	High	45
High	Big	High	52
High	Big	Low	23
High	Big	High	46
High	Small	Low	30

For Buying_Price = High, SD = 10.87

We will calculate SDR of only Lug_Boot and Safety

		Maintenance_ Price (SD)
Lug_Boot	Small	7.5
	Big	12.49

SD(High, Lug_Boot) = P (Small) SD (Small) + P (Big) SD (Big)

$= \frac{2}{5} \times 7.5 + \frac{3}{5} \times 12.49 = 10.49$

SDR = SD(High) – SD (Maintenance_ Price, Lug_Boot) = 10.87 – 10.49 = 0.38

		Maintenance_ Price (SD)
Safety	High	3.09
	Low	3.50

SD (High, Safety) = P (High) SD (High) + P (Low) SD (Low)

$= \frac{3}{5} \times 3.09 + \frac{2}{6} \times 3.5 = 3.25$

SDR = SD (High) – SD (Maintenance_ Price, Safety) = 10.87 – 3.25 = 7.62

SDR of Safety is highest so we select Safety as next node below High branch.

- For Buying_Price = High and Safety = High, we can directly write down the answer.
- For Buying_Price = High and Safety = Low, we can directly write down the answer.

– To write down the answer we take average of values of following records,

- For Buying_Price = High and Safety = High,

  Maintenance _Price = (45 + 52 + 46) /3 = 47.7

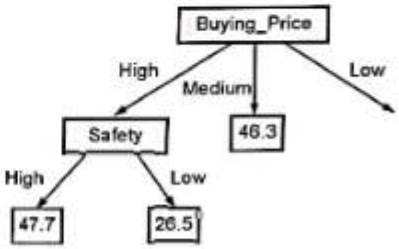
- For Buying_Price = High and Safety = Low, Maintenance _Price = (23 + 30) /2 = 26.5

Step 3:

Now we will consider the records of 'Medium'

Buying_Price	Lug_Boot	Safety	Maintenance_ Price? (in thousand)
Medium	Small	High	46
Medium	Big	Low	43
Medium	Small	Low	52
Medium	Big	High	44

For Buying_Price = Medium, we can directly write down the answer as 46.3. The answer is calculated by taking the average of values of Maintenance_ Price for Medium records (average of 46, 43, 52, and 44).



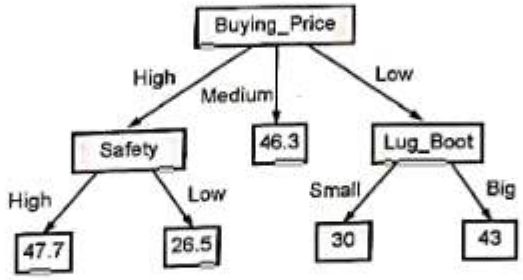
Step 4 :

Now we will consider the records of 'Low'.

Buying_Price	Lug_Boot	Safety	Maintenance_ Price? (in thousand)
Low	Small	High	25
Low	Small	Low	30
Low	Small	High	35
Low	Big	High	38
Low	Big	Low	48

For Buying_Price = Low, SD = 7.78

- Now only Lug_Boot attribute is remaining, so we can directly take Lug_Boot as a next node below Low branch.
- To write down the answer we take average of values of following records,
  - For Buying_Price = Low and Lug_Boot = Small, Maintenance_Price = (25 + 30 + 35) /3 = 30
  - For Buying_Price = Low and Lug_Boot = Big, values of Maintenance_Price = (38 + 48) /2 = 43.
- Final Regression Tree is,



Module 4

Example 2

Day	Outlook	Temp	Humidity	Windy	Hours Play?
1	Rainy	Hot	High	False	25
2	Rainy	Hot	High	True	30
3	Overcast	Hot	High	False	46
4	Sunny	Mild	High	False	45
5	Sunny	Cool	Normal	False	52
6	Sunny	Cool	Normal	True	23
7	Overcast	Cool	Normal	True	43
8	Rainy	Mild	High	False	35
9	Rainy	Cool	Normal	False	38
10	Sunny	Mild	Normal	False	46
11	Rainy	Mild	Normal	True	48
12	Overcast	Mild	High	True	52
13	Overcast	Hot	Normal	False	44
14	Sunny	Mild	High	True	30

Standard deviation

A decision tree is built up top down from root node and involved partitioning the data into subsets that contain instances with similar values. We use SD to calculate homogeneity of a numerical sample.

$$SD, S = \sqrt{\frac{\sum (x-\mu)^2}{n}} = 9.32$$

SD Reduction

It is based on the decrease in SD after a dataset is split on an attribute. Constructing a tree is all about finding attribute that returns highest SDR.

- **Step 1 :**

SD (Hours Play?) = 9.32

- **Step 2 :**

The dataset is then split on the different attribute.SD for each branch is calculated. The resulting SD is subtracted from SD before split.

		Hours Play(SD)
Outlook	Rainy	7.78
	Overcast	3.49
	Sunny	10.87

SD (HoursPlay, Outlook) = P (Rainy) SD (Rainy) + P (Overcast) SD (Overcast) + P (Sunny) SD (Sunny)

$= \frac{5}{14} \times 7.78 + \frac{4}{14} \times 3.49 + \frac{5}{14} \times 10.87 = 7.66$

SDR = SD (Hours Play) – SD (Hours Play, Outlook) = 9.32 – 7.66 = 1.66

Temp		Hours Play(SD)
	Cool	10.51
	Hot	8.95
	Mild	7.65

SD(Hours Play, Temp) = P (Cool) SD (Cool) + P (Hot) SD (Hot) + P (Mild) SD (Mild)

$= \frac{4}{14} \times 10.51 + \frac{4}{14} \times 8.95 + \frac{6}{14} \times 7.65 = 8.84$

SDR = SD (Hours Play) – SD(Hours Play, Temp) = 9.32 – 8.84 = 0.48

Humidity		Hours Play (SD)
	High	9.36
	Normal	8.37

SD (Hours Play, Humidity) = P (High) SD (High) + P (Normal) SD (Normal)

$= \frac{7}{14} \times 9.36 + \frac{7}{14} \times 8.37 = 8.86$

SDR = SD (Hours Play) – SD (Hours Play, Humidity) = 9.32 – 8.86 = 0.46

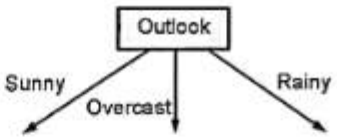
Windy		Hours Play(SD)
	False	7.87
	True	10.59

SD (HoursPlay, Windy) = P (False) SD (False) + P (True) SD (True)

$= \frac{8}{14} \times 7.87 + \frac{6}{14} \times 10.59 = 9.02$

SDR = SD (Hours Play) – SD (Hours Play, Windy) = 9.32 – 9.02 = 0.3

SDR of outlook is highest so we select outlook as our root node.



Step 2(a) :

Now we will consider the records of 'sunny'.

Day	Outlook	Temp	Humidity	Windy	Hours Play?
4	Sunny	Mild	High	False	45
5	Sunny	Cool	Normal	False	52
6	Sunny	Cool	Normal	True	23
10	Sunny	Mild	Normal	False	46
14	Sunny	Mild	High	True	30

For Outlook = Sunny, SD = 10.87

We will calculate SDR of only Temp, Humidity and Windy

		Hours Play (SD)
Temp	Cool	14.5
	Mild	7.31

SD (Sunny, Temp) = P (Cool) SD (Cool) + P (Mild) SD (Mild)

$= \frac{2}{5} \times 14.5 + \frac{3}{5} \times 7.31 = 10.18$

SDR = SD(Sunny) – SD (Hours Play, Temp) = 10.87 – 10.18 = 0.69

		Hours Play (SD)
Humidity	High	7.5
	Normal	12.49

SD (Sunny, Humidity) = P (High) SD (High) + P (Normal) SD (Normal)

$= \frac{2}{5} \times 7.5 + \frac{3}{5} \times 12.49 = 10.49$

SDR = SD (Sunny) – SD (Hours Play, Humidity) = 10.87 – 10.49 = 0.38

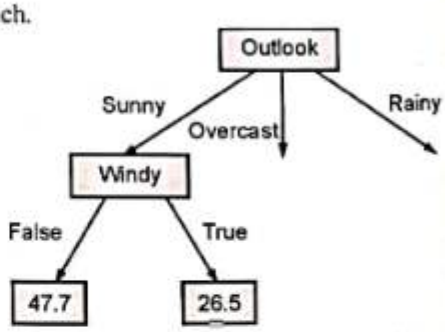
		Hours Play (SD)
Windy	False	3.09
	True	3.50

SD (Sunny, Windy) = P (False) SD (False) + P (True) SD (True)

$= \frac{3}{5} \times 3.09 + \frac{2}{5} \times 3.5 = 3.25$

SDR = SD (Sunny) – SD (Hours Play, Windy) = 10.87 – 3.25 = 7.62

- SDR of windy is highest so we select windy as next node below sunny branch.
  - For Outlook = sunny and Windy = false, we can directly write down the answer.
  - For Outlook = sunny and Windy = true, we can directly write down the answer



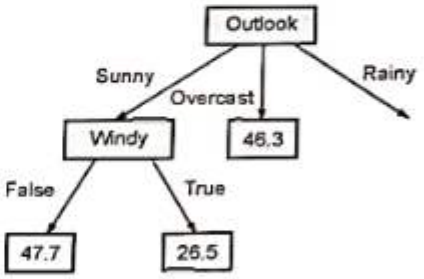
- To write down the answer we take average of values of following records,
  - For Outlook = sunny and Windy = false, Hours Play = (45 + 52 + 46) /3 = 47.7
  - For Outlook = sunny and Windy = true, Hours Play = (23 + 30) /2 = 26.5

Step 3 :

Now we will consider the records of 'overcast'. For Overcast we will directly write down the answer.

Day	Outlook	Temp	Humidity	Windy	Hours Play?
3	Overcast	Hot	High	False	46
7	Overcast	Cool	Normal	True	43
12	Overcast	Mild	High	True	52
13	Overcast	Hot	Normal	False	44

The answer is calculated by taking the average of values of Hours Play for overcast records (average of 46, 43, 52, and 44).



Step 4 :

Now we will consider the records of 'Rainy'.

Day	Outlook	Temp	Humidity	Windy	Hours Play?
1	Rainy	Hot	High	False	25
2	Rainy	Hot	High	True	30
8	Rainy	Mild	High	False	35
9	Rainy	Cool	Normal	False	38
11	Rainy	Mild	Normal	True	48

For Outlook = Rainy, SD = 7.78

Now we will calculate SDR of only Humidity and Temp

		Hours Play (SD)
Temp	Hot	2.5
	Cool	0
	Mild	6.5

SD (Rainy, Temp) = P (Hot) SD (Hot) + P (Cool) SD (Cool) + P (Mild) SD (Mild)

$= \frac{2}{5} \times 2.5 + \frac{1}{5} \times 0 + \frac{2}{5} \times 6.5 = 3.6$

SDR = SD (Rainy) – SD (Rainy Play, Temp) = 7.78 – 3.6 = 4.18

		Hours Play(SD)
Humidity	High	5
	Normal	5

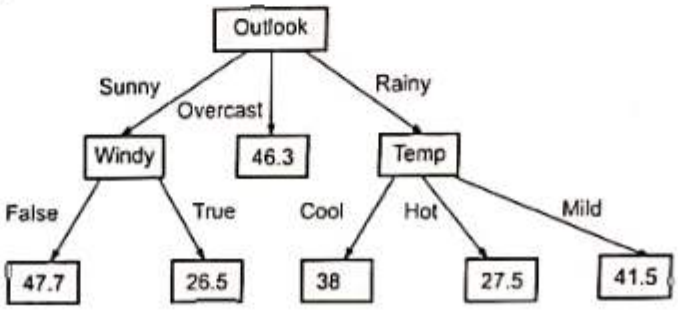
SD (Sunny, Humidity) = P (High) SD (High) + P (Normal) SD (Normal)

$= \frac{3}{5} \times 5 + \frac{2}{5} \times 5 = 5$

SDR = SD (Rainy) – SD (Hours Play, Humidity) = 7.78 – 5 = 2.28

Module 4

- SDR of Temp is highest so we select Temp as next node below rainy branch.
  - For Outlook = Rainy and Temp = Cool, we can directly write down the answer.
  - For Outlook = Rainy and Temp = Hot, we can directly write down the answer.
  - For Outlook = Rainy and Temp = Mild, we can directly write down the answer.
- To write down the answer we take average of values of following records,
  - For Outlook = Rainy and Temp = Hot, Hours Play = (25 + 30) /2 = 27.5
  - For Outlook = Rainy and Temp = Mild, Hours Play = (35 + 48) /2 = 41.5
  - For Outlook = Rainy, Temp = Cool, Hours Play = 38
- Final Regression Tree is,



4.10 UNIVERSITY QUESTIONS AND ANSWERS

May 2015

Q. 1 Create a decision tree for the attribute "class" using the respective values. *(Ans. : Refer Example 4.7.4)* **(12 Marks)**

Eyecolour	Married	Sex	Hairlength	class
Brown	yes	male	Long	Football
Blue	yes	male	Short	Football
Brown	yes	male	Long	Football
Brown	no	female	Long	Netball
Brown	no	female	Long	Netball
Blue	no	male	Long	Football
Brown	no	female	Long	Netball
Brown	no	male	Short	Football
Brown	yes	female	Short	Netball
Brown	no	female	Long	Netball
Blue	no	male	Long	Football
Blue	no	male	Short	Football

Q. 2 Write short note on Issues in Decision Tree *(Ans. : Refer section 4.4)* **(10 Marks)**

Q.3 Explain Regression line, Scatter plot, Error in prediction and Best fitting line. *(Ans. : Refer section 4.1)* **(4 Marks)**

➲ **May 2016**

Q.4 Explain in brief Linear Regression Technique. *(Ans. : Refer section 4.1)* **(5 Marks)**

➲ **May 2017**

Q.5 For a Sunburn dataset given below, construct a decision tree. *(Ans. : Refer Example 4.7.5)* **(10 Marks)**

Name	Hair	Height	Weight	Location	Class
Swati	Blonde	Average	Light	No	Yes
Sunita	Blonde	Tall	Average	Yes	No
Anita	Brown	Short	Average	Yes	No
Lata	Blonde	Short	Average	No	Yes
Radha	Red	Average	Heavy	No	Yes
Maya	Brown	Tall	Heavy	No	No
Leena	Brown	Average	Heavy	No	No
Rina	Blonde	Short	Light	Yes	No

Q.6 Following table shows the midterm and final exam grades obtained for students in a database course. Use the method of least squares using regression to predict the final exam grade of a student who received 86 in the midterm exam. *(Ans. : Refer Example 4.2.3)* **(10 Marks)**

Midterm exam (X)	72	50	81	74	94	86	59	83	86	33	88	81
Final exam (Y)	84	53	77	78	90	75	49	79	77	52	74	90

Q.7 Write short note on : Logistic Regression. *(Ans. : Refer section 4.3)* **(10 Marks)**

➲ **May 2019**

Module 4

Q.8 For a Sunburn dataset given below, construct a decision tree For the following data, Calculate Gini indexes and determines which attribute is root attribute and generate two level deep decision tree. **(10 Marks)**

(Ans. : Refer Example 4.7.6)

Sr. No.	Income	Defaulting	Credit score	Location	Give Loan?
1	Low	High	High	bad	No
2	Low	High	High	good	No
3	High	High	High	bad	Yes
4	Medium	Medium	High	bad	Yes
5	Medium	Low	Low	bad	No
6	Medium	Low	Low	good	Yes
7	High	Low	Low	good	Yes
8	Low	Medium	High	bad	No
9	low	Low	Low	bad	No
10	Medium	Medium	Low	bad	No
11	Low	Medium	Low	good	Yes
12	High	Medium	High	good	Yes
13	High	High	Low	bad	No
14	medium	Medium	High	good	Yes

Q. 9 Explain how regression problem can be solved using Steepest descent method. Write down the steps. *(Ans. : Refer section 4.3)* **(5 Marks)**

Dec. 2019

Q. 10 Given the following data for the sales of car of an automobile company for six consecutive years. Predict the sales for next two consecutive years. *(Ans. : Refer Example 4.2.4)* **(10 Marks)**

Years	2013	2014	2015	2016	2017	2018
Sales	110	100	250	275	230	300

Multiple Choice Questions

Q. 4.1 Linear Regression is represented by following equation _____

(a) Y = a + bX where a is X-intercept and b is Slope of the line

(b) Y = a + bX where a is the slope of the line and b is X-Intercept

(c) Y = a + bX where a is the Y-Intercept and b is the slope of the line

(d) Y = a + bX where a is the slope of the line and b is the Y-Intercept ✓ **Ans. : (c)**

Explanation : Definition of linear regression

Q. 4.2 A correlation between age and percentage of getting infected with COVID-19 is – 1. What you can interpret from this _____

(a) Age is a good predictor and is negatively correlated

(b) Age is a good predictor and it is positively correlated

(c) Age is not good precitor and it is negatively correlated

(d) Age is not good predictor and it is positively related

✓ **Ans. : (a)**

Explanation : Age is a important factor in COVID so, we can say age is a good predictor and it is negatively correlated.

Q. 4.3 In linear regression, which plot should be used to represent the relationship between dependent variable and independent variable?

(a) Box plot (b) Bar graph

(c) Scatter (d) Plot Histogram ✓ **Ans. : (c)**

Explanation : Scatter plot is used to plot predicted output vs input.

Q. 4.4 How many coefficients do you need to estimate in a simple linear regression model ?

(a) 0 (b) 3 (c) 1 (d) 2 ✓ **Ans. : (d)**

Explanation : Y = a + bX where a is the Y-Intercept and b is the slope of the line. a and b coefficients are required.

Q. 4.5 Linear regression belongs to which type of machine learning algorithm ?

(a) Supervised Learning (b) Hybrid Learning

(c) Unsupervised learning

(d) Reinforcement Learning ✓ **Ans. : (a)**

Explanation : Best fitting line is drawn from input and output points which is used for prediction.

Q. 4.6 Choose the appropriate method used to find best fit the data in logistic regression.

(a) Jaccard Distance (b) Least Square error

(c) Maximum likelihood (d) Pearson coefficient

✓ **Ans. : (c)**

Explanation : Maximum likelihood(probability) is used to decide the class in logistic regression.

Q. 4.7 Choose the appropriate method used to find best fit the data in linear regression.

(a) Least Square error (b) Maximum likelihood

(c) R-square (d) Pearson coefficient

✓ **Ans. : (a)**

Explanation : Least square error is used to draw regression line.

Q. 4.8 On given data logistic regression model is build. It has training accuracy as X and testing accuracy as Y. If new features are added in this already build model how it will effects on accuracy. Choose appropriate one.

(a) Testing accuracy will decrease

(b) Training accuracy will decrease

(c) Testing accuracy will increase

(d) Training accuracy will increase or remain same

✓ **Ans. : (d)**

Explanation : If we add more features to a trained model then training accuracy will increase or remain same.

Q. 4.9 To find the minimum or the maximum of a function, we set the gradient to zero because _____

(a) The value of the gradient at extrema of the function is always equal to zero

(b) Depends upon the type of problem

(c) The value of the gradient at extrema of the function is always equal to maximum

(d) The value of the gradient at extrema of the function is always equal to minimum

✓ **Ans. : (a)**

Explanation : Property of gradient.

Q. 4.10 Which of the following is a disadvantage of decision trees ?

(a) Factor analysis

(b) Decision trees are robust to outliers

(c) Decision trees are prone to overfit

(d) Decision trees are not prone to overfit

✓ **Ans. : (c)**

Explanation : If we continue to split all the branches then it will lead to overfitting.

Q. 4.11 Find the statement which is not true in case of gini index.

(a) Thegini index is biased towards multivalued attribute

(b) Gini index does not favour equal sized partitions

(c) Gini index favour test when purity is there in both partitions.

(d) When the number of classes is large Gini index is not a good choice.

✓ **Ans. : (b)**

Explanation : Gini index favours equal size partitions.

Q. 4.12 Which one of these is not a tree based learner?

(a) CART (b) ID3

(c) Bayesian Classifier (d) Random forest

✓ **Ans. : (c)**

Explanation : Bayesian classifier is based on baye's theorem.

Q. 4.13 Finding the number of Covid Patient cases is _____

(a) Classification model (b) Regression Model

(c) Clustering Model (d) Astrological Model

✓ **Ans. : (b)**

Explanation : Since we are finding number of patients (numerical output).

Q. 4.14 Logistic regression is _____

(a) Supervised Regression

(b) Supervised Classification

(c) Unsupervised Learning

(d) Reinforcement learning

✓ **Ans. : (b)**

Explanation : Logistic regression is extended for deciding class. In this expected output is also known.

Q. 4.15 Find the statement which is true in case of gini index.

(a) The gini index is biased towards multivalued attribute

(b) Gini index does not favours equal sized partitions

(c) Gini index favour test when impurity is there in both partitions.

(d) When the number of classes is large Gini index is a good choice.

✓ **Ans. : (a)**

Explanation : Property of gini index method.

Q. 4.16 The Family of Decision Tree learning algorithm is _____

(a) Unsupervised learning model

(b) Supervised learning model

(c) Stochastic learning Model

(d) Reinforcement learning Model

✓ **Ans. : (b)**

Explanation : In decision tree training data contains input as well as output.

Q. 4.17 Temperature Prediction is _____

(a) Classification problem (b) Regression Problem

(c) Clustering problem (d) Astrological Problem

✓ **Ans. : (b)**

Explanation : Since we are predicting temperature (numerical value).

Q. 4.18 How many Coefficients are needed to estimate a simple linear regression model with one independent variable ?

(a) 1 (b) 2 (c) 3 (d) 4

✓ **Ans. : (b)**

Explanation : Along with one independent variable we will require Y-Intercept and slope of the line.

Q. 4.19 You want to design a system to predict the price of house and after prediction if predicted and actual is nearby same then deal is registered and processed for the loan approval. Once loan is approved sale deed is done. The selling price of a house depends on the following factors. For eg. It depends on the number of bedrooms, number of kitchen, number of bathrooms, the year the house was built and the square footage of the lot. Also the facilities that are available near the house. Loan sanction depends on the credit history of customer. Given these factors, predicting the selling price of the house is an example of which task and also what process you will follow to design?

(a) Locality survey, Binary Classification, deal registration, credit scoring, loan approval, sale deed

(b) Locality survey, Maltilevel classification, deal registration, credit scoring, loan approval, sale deed

(c) Locality survey, Simple linear regression,, deal registration, credit scoring, loan approval, sale deed

(d) Locality survey, Multiple linear regression, deal registration, credit scoring, loan approval, sale deed

✓ **Ans. : (d)**

Module 4

Explanation : Locality survey, Multiple linear regression, deal registration, credit scoring, loan approval, sale deed As price (numeric) and loan sanction (Y/1) is predicted based on more than one parameter.

Q. 4.20 Suppose there are five instances, 1, 2, 3, 4, 5 in a dataset having three features, p, q and r as shown in the table below :

Instances	p	q	r
1	1.6	2.3	5.1
2	2.4	2.5	4.6
3	3.9	3.6	3.7
4	4.1	3.7	2.5
5	5.6	8.3	1.8

In order to find the dependence between two variables we use the Pearson's Correlation

Coefficient. Based on your understanding of Correlation Coefficient, choose the correct

(a) A strong positive correlation between p and q

(b) A strong negative correlation between p and q.

(c) A weak positive correlation between p and r.

(d) A weak negative correlation between p and r.

✓ **Ans. : (a)**

Explanation : Correlation coefficient value of exactly 1.0 means there is a perfect positive relationship between the two variables. For a positive increase in one variable, there is also a positive increase in the second variable. A value of -1 means there is a perfect negative relationship between the two variables. This shows that the variables move in opposite directions for a positive increase in one variable, there is a decrease in the second variable. While the strength of the relationship varies in degree based on the absolute value of the correlation coefficient.

Q. 4.21 The sales of a company (in thousands) for each month are shown in table below :

Find least square regression line, $y = ax + b$. Use line as a model to estimate the sales of company in the 6^{th} month.

X(month)	1	2	3	4	5
Y(sales)	12	19	29	37	45

(a) 84 (b) 60

(c) 74 (d) 80 ✓ **Ans. : (a)**

Explanation : 84, from calculation of regression line $Y = aX + b$

Q. 4.22 Suppose you want to develop multi classification problem and you are only allowed to use binary logistic classifiers to solve a multi-class classification problem. Given a training set with 2 classes, this classifier can learn a model, which can then be used to classify a new test point to one of the 2 classes in the training set. For enhancement or better experiment we are now given a 6 class problem along with the training set, and have to use more than one binary logistic classifier to solve the problem, as mentioned before. Propose the following scheme : you will first train a binary logistic classifier for every pair of classes. Now, for a new test point, you will run through each of these models, and the class which wins the maximum number of pairwise contests, is the predicted label for the test point. How many binary logistic classifiers will you need to solve the problem using your proposed scheme?

(a) 2 (b) 3 (c) 4 (d) 6 ✓ **Ans. : (d)**

Explanation : 6 as we need 3 to execute this for inputs (first 3 then 2 then 1)

Q. 4.23 The selling price of a house depends on the following factors. For example, it depends on the number of bedrooms, number of kitchen, number of bathrooms, the year the house was built and the square footage of the lot. Given these factors, predicting the selling price of the house is an example of which task?

(a) Binary classification

(b) Multilabel classification

(c) Simple linear regression

(d) Multiple linear regression ✓ **Ans. : (d)**

Explanation : Since we are prediction selling price it is an example of regression. Selling price depends on multiple factors so it is multiple linear regression.

Q. 4.24 A feature F1 can take certain value: A, B, C, D, E and F represents grade of students from a college. Which of the following statement is true in following case?

(a) Feature F1 is an example of nominal variable.

(b) Feature F1 is an example of ordinal variable.

(c) It doesn't belong to any of the above category.

(d) Both of these ✓ **Ans. : (b)**

Explanation : Ordinal variables are the variables which have some order in their categories. For example, grade A should be considered as high grade than grade B.

Q. 4.25 Which of the following is true for a decision tree ?

(a) A decision tree is an example of a linear classifier.

(b) The entropy of a node typically decreases as we go down a decision tree.

(c) Entropy is a measure of purity.

(d) An attribute with lower mutual information should be preferred to other attributes. ✓ **Ans. : (b)**

Explanation : Property of a decision tree.

Q. 4.26 Suppose you got a situation where you find that linear egression model is underfitting the data in such situation which of the following options woluld you consider?

(a) Will add more features

(b) Will strat intoducing higher degree features

(c) Will remove some features

(d) None of above ✓ **Ans. : (a)**

Explanation : In case of under-fitting, you need to induce more features in variable space or you can add some polynomial degree variables to make the model more complex to be able to fit the data better.

Q. 4.27 Consider the dataset, S given below :

Elevation	Road Type	Speed Limit	Speed
steep	Uneven	Yes	Slow
steep	Smooth	Yes	Slow
flat	Uneven	No	Fast
steep	Smooth	No	Fast

Elevation, Road Type and speed Limit are the features and Speed is the target label that we want to predict.

Find the entropy of the dataset, S as given above:

(a) 0.5 (b) 0 (c) 1 (d) 0.7 ✓ **Ans. : (c)**

Explanation : For a dataset, S with C many classes, the entropy of the set, S given by H(S) is defined as : $H(S) = \sum p_c \log p_c$ where p_c is the probability of an element of S belonging to a class. In this case, P(slow) = 0.5; P(fast) = 0.5 and hence H(S) = 1

Q. 4.28 Find the information Gain if the dataset is split at the feature "Elevation" :

(a) 1 (b) 0 (c) 0.675 (d) 0.325 ✓ **Ans. : (d)**

Explanation : The feature, Elevation has 2 values = {Steep, Flat}. For a split on the feature Elevation, one subtree would be of inputs having feature Steep and the other having feature, Flat. For the feature, Steep, there are 3 examples, out of which P(slow) = 2/3 and P(fast) = 1/3 Thus, Entropy (Steep) = 0.9 and Entropy(Flat) = 0.

Thus Information Gain = 1 – ((3/4) * 0.9 + (1/4) *0)
= 1 – 0.675 = 0.325

Q. 4.29 Find the feature on which the parent node must be chosen to split the dataset, S based on information gain

(a) Speed Limit (b) Road Type

(c) Elevation ✓ **Ans. : (a)**

Explanation : Using the Information Gain formula, we can find the information gain for each of the features.

The values should be IG(S, Elevation) = 0.325,
IG(S, Road Type) = 0, IG(S, Speed Limit) = 1

Since the decision tree is constructed on the feature having maximum information gain, (a) is the correct answer.

Q. 4.30 Consider a simple linear regression model with One independent variable (X). The output variable is Y. The equation is : Y = aX + b where a is the slope and b is the intercept. If we change the input variable (X) by 1 unit, by how much output variable (Y) will change?

(a) 1 unit (b) By slope

(c) By intercept (d) None ✓ **Ans. : (c)**

Explanation : Equation for simple linear regression: Y = a + bX. Now if we increase the value of X by 1 then the value of Y would be a + b (x + 1) i.e. value of Y will get incremented by b.

Q. 4.31 The following table shows the results of a recently conducted study on the correlation of the number of hours spent driving with the risk of developing acute backache. Find the equation of the best fit line for this data.

No of hrs(x)	Risk on a scale (y)
10	95
9	80
2	10
15	50
10	45
16	98
11	38
16	93

Choose which of the options is correct?

(a) y = 3.39x + 11.62 (b) Y = 4.69x + 12.58

(c) Y = 4.59x + 12.58 (d) Y = 3.59x + 10.58

✓ **Ans. : (c)**

Explanation : For each x calculate the value of Y using the given equations. Then calculate error for each equation. Equation with the lowest error is the desired answer.

Q. 4.32 Pruning is a technique that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. This is done in order to avoid _____

(a) overfitting (b) underfitting

(c) Both (d) None ✓ **Ans. : (a)**

Explanation : Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting.

Module 4

Q. 4.33 We define a concept as the conjunction of literals. For the given instance space, X, find the specific concept using Find-S algorithm.

Sky	Airtemp	Humidity	Wind	Enjoys port
sunny	warm	High	Strong	Yes
sunny	warm	High	Strong	Yes
rainy	cold	Normal	Strong	No
sunny	warm	High	Weak	Yes
cloudy	cold	Normal	Strong	No

'?' means don't care and '0' indicates no value

(a) <Sunny,Cold,High,Strong>

(b) <Sunny, 0, ?, Strong>

(c) <Sunny, Warm, High, ?>

(d) <Sunny, 0, Warm, Strong> ✓ **Ans. : (c)**

Explanation : For Sunny,Warm,High,? The class is Yes.

Q. 4.34 Find the number of instances possible in X using the possible values that can be seen in the table given in above Question.

(a) 12 (b) 48

(c) 36 (d) 24 ✓ **Ans. : (d)**

Explanation :we count possible values for each attribute and multiply them, 3*2*2*2 = 24.

Q. 4.35 Find the number of syntactically distinct hypotheses given in the table Question 33.

Hint : Each attribute can have 2 more values : ? and 0

(a) 160 (b) 320

(c) 80 (d) 40 ✓ **Ans. : (b)**

Explanation : For each attribute possible values as 3*2*2*2 and each attribute can have 2 more values so, 5*4*4*4 = 320

Q. 4.36 We can get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent.

(a) True (b) False ✓ **Ans. : (b)**

Explanation : we do not get multiple optimum solution.

Q. 4.37 When a decision tree is grown to full depth, it is more likely to fit the noise in the data.

(a) True (b) False. ✓ **Ans. : (a)**

Explanation : As tree grows to full depth, possibility of redundant and noisy data gets increased.

Q. 4.38 As the number of training examples goes to infinity, your model trained on that data will have _____

(a) Lower variance (b) Higher variance

(c) Same variance ✓ **Ans. : (a)**

Explanation : When model is trained using infinite number of training examples then the model exhibits lower variance in data.

Q. 4.39 As the number of training examples goes to infinity, your model trained on that data will have _____

(a) Lower bias (b) Higher bias

(c) Same bias ✓ **Ans. : (c)**

Explanation : There will be no changes in bias.

Q. 4.40 The error function most suited for gradient descent using logistic regression is _____

(a) The entropy function

(b) The squared error

(c) The cross-entropy function

(d) The number of mistakes ✓ **Ans. : (c)**

Explanation : For logistic regression, the cross-entropy function (loss function or cost function) is convex. A convex function has just one minimum; there are no local minima to get stuck in, so gradient descent starting from any point is guaranteed to find the minimum. Since the Cross Entropy cost function is convex a variety of local optimization schemes can be more easily used to properly minimize it. For this reason the Cross Entropy cost is used more often in practice for logistic regression than is the logistic Least Squares cost.

Q. 4.41 Regarding bias and variance, which of the following statements are true?

(a) Models which overfit are more likely to have high bias

(b) Models which overfit are more likely to have low bias

(c) Models which overfit are more likely to have high variance

(d) Models which overfit are more likely to have low variance ✓ **Ans. : (b), (c)**

Explanation : The bias of a classifier gets reduced when the training set error lowers down to zero causing low bias, while due to overfitting the gap between the training error and test error becomes higher, causing high variance.

Q. 4.42 Consider a binary classification problem. Suppose I have trained a model on a linearly separable training set, and now I get a new labeled data point which is correctly classified by the model, and far away from the decision boundary. If I now add this new point to my earlier training set and re-train, in which cases is the learnt decision boundary likely to change?

(a) When my model is a perceptron.

(b) When my model is logistic regression.

(c) When my model is an SVM.
(d) When my model is HMM. ✓ **Ans. : (b)**

Explanation : If we add new point to already trained model and we are using logistic regression then the decision boundary will change.

Q. 4.43 When doing least-squares regression with regularisation (assuming that the optimisation can be done exactly), increasing the value of the regularisation parameter ____

(a) will never decrease the training error.
(b) will never increase the training error.
(c) will never decrease the testing error.
(d) will never increase the testing error. ✓ **Ans. : (a)**

Explanation : Increasing the value of regularization parameter will not decrease training error.

Q. 4.44 High entropy means that the partitions in classification are ____

(a) pure (b) not pure
(c) useful (d) useless ✓ **Ans. : (b)**

Explanation : Entropy is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information. It is a measure of disorder or purity or unpredictability or uncertainty. Low entropy means less uncertain and high entropy means more uncertain.

Q. 4.45 A machine learning problem involves four attributes plus a class. The attributes have 3, 2, 2, and 2 possible values each. The class has 3 possible values. How many maximum possible different examples are there?

(a) 12 (b) 24
(c) 48 (d) 72 ✓ **Ans. : (d)**

Explanation : Maximum possible different examples are the products of the possible values of each attribute and the number of classes; $3 * 2 * 2 * 2 * 3 = 72$

Q. 4.46 Adding a non-important feature to a linear regression model may result in.

1. Increase in R-square 2. Decrease in R-square

(a) Only 1 is correct (b) Only 2 is correct
(c) Either 1 or 2 (d) None of these ✓ **Ans. : (a)**

Explanation : After adding a feature in feature space, whether that feature is important or unimportant features the R-squared always increase.

Q. 4.47 For which of the following hyperparameters, higher value is better for decision tree algorithm?

1. Number of samples used for split
2. Depth of tree
3. Samples for leaf

(a) 1 and 2 (b) 2 and 3
(c) 1 and 3 (d) Can't say ✓ **Ans. : (d)**

Explanation : For all three options a, b, and c, it is not necessary that if you increase the value of parameter the performance may increase. For example, if we have a very high value of depth of tree, the resulting tree may overfit the data, and would not generalize well. On the other hand, if we have a very low value, the tree may underfit the data. So, we can't say for sure that "higher is better".

Q. 4.48 In Regression tree which method is used ____

(a) Gini index (b) ID3
(c) Standard deviation reduction
(d) None of above ✓ **Ans. : (c)**

Explanation : Standard deviation reduction is calculated for all attributes and based on that values tree is constructed.

Q. 4.49 In Regression tree output attribute is ____

(a) discrete (b) categorical
(c) continuous (d) all of above ✓ **Ans. : (c)**

Explanation : In regression tree output is continuous whereas in classification tree it is categorical.

Q. 4.50 Another name for output attribute ____

(a) predictive variable
(b) Independent variable
(c) estimated variable
(d) dependent variable ✓ **Ans. : (a)**

Explanation : Since we predict the output attribute.

Chapter Ends...

❑❑❑

UNIT 5

Chapter...5

Learning with Classification and Clustering

University Prescribed Syllabus

Classification : Rule based classification, classification by Bayesian Belief networks, Hidden Markov Models. Support Vector Machine : Maximum Margin Linear Separators, Quadratic Programming solution to finding maximum margin separators, Kernels for learning non-linear functions.
Clustering : Expectation Maximization Algorithm, Supervised learning after clustering, Radial Basis functions.

5.1 RULE BASED CLASSIFICATION

Rule-based classifier classifies records using a set of IF-THEN rules. The rules can be expressed in the following form

(Condition) → Y

Where LHS of above rule is known as an antecedent or condition

RHS of above rule is known as rule consequent

Condition is a conjunction of attribute. Here condition consists of one or more attribute tests which are logically ANDed.

Y represents the class label

Assume a rule R1,

R1: IF Buying_Price = high AND Maintenance_Price = high AND Safety = low THEN Car_evaluation = unacceptable

Rule R1 can also be rewritten as:

R1: (Buying_Price = high) ^ (Maintenance_Price = high) ^ (Safety = low) → Car_evaluation = unacceptable

If the antecedent is true for a given record, then the consequent is given as output.

5.1.1 Example of Rule based Classifier

Buying_Price	Maintenance_ Price	Lug_Boot	Safety	Evaluation?
High	High	Small	High	Unacceptable
High	High	Small	Low	Unacceptable
Medium	High	Small	High	Acceptable
Low	Medium	Small	High	Acceptable
Low	Low	Big	High	Acceptable
Low	Low	Big	Low	Unacceptable
Medium	Low	Big	Low	Acceptable
High	Medium	Small	High	Unacceptable
High	Low	Big	High	Acceptable
Low	Medium	Big	High	Acceptable
High	Medium	Big	Low	Acceptable
Medium	Medium	Small	Low	Acceptable
Medium	High	Big	High	Acceptable
Low	Medium	Small	Low	Unacceptable

Module 5

- R1 : (Buying-Price = high) ^ (Maintenance_Price = high) ^ (Safety = low) → Car_evaluation = unacceptable
- R2 : (Buying-Price = medium) ^ (Maintenance_Price = high) ^ (Lug_Boot = big) → Car_evaluation = acceptable
- R3 : (Buying-Price = high) ^ (Lug_Boot = big) → Car_evaluation = unacceptable
- R4 : (Maintenance_Price = medium) ^ (Lug_Boot = big) ^ (Safety = high) → Car_evaluation = acceptable

5.1.2 Application of Rule based Classifier

A record x is said to be covered by a rule, if all the attributes present in the record satisfy the antecedent of the rule.

Car	Buying_Price	Maintenance_ Price	Lug_Boot	Safety	Evaluation?
1	High	High	Small	low	?
2	medium	High	big	Low	?
3	High	medium	big	high	?
4	High	medium	small	low	?

- Car 1 triggers rule R1 → unacceptable
- Car 2 triggers rule R2 → acceptable
- Car 3 triggers both R3 and R4
- Car 4 triggers none of the rules

5.1.3 Characteristics of Rule based Classifier

1. **Mutually Exclusive Rules :** Rule based Classifier comprises of mutually exclusive rules where the rules are independent of each other. Each and every record is covered by at most one rule.

Solution : Arrange rules in the order

Arrangement of Rules in the Order

Rules are assigned a priority and based on this they are arranged and ranks are associated. When a test record is given as a input to the classifier, a label of the class with highest priority triggered rule is assigned. If the test record does not trigger any of the rules then a default class is assigned.

Rule-based ordering

In rule based ordering individual rules are ranked based on their quality.

- R1 : (Buying-Price = high) ^ (Maintenance_Price = high) ^ (Safety = low) → Car_evaluation = unacceptable
- R2 : (Buying-Price = medium) ^ (Maintenance_Price = high) ^ (Lug_Boot = big) → Car_evaluation = acceptable
- R3 : (Buying-Price = high) ^ (Lug_Boot = big) → Car_evaluation = unacceptable
- R4 : (Maintenance_Price = medium) ^ (Lug_Boot = big) ^ (Safety = high) → Car_evaluation = acceptable

Car	Buying_Price	Maintenance_ Price	Lug_Boot	Safety	Evaluation?
3	High	medium	big	high	?

- Car 3 triggers rule R3 first → unacceptable

Class-based ordering

In class based ordering rules which belong to the same class are grouped together.

- R2 : (Buying-Price = medium) ^ (Maintenance_Price = high) ^ (Lug_Boot = big) → Car_evaluation = acceptable

R4 : (Maintenance_Price = medium) (Lug_Boot = big) ^ (Safety = high) → Car_evaluation = acceptable

R1 : (Buying-Price = high) ^ (Maintenance_Price = high) ^ (Safety = low) → Car_evaluation = unacceptable

R3 : (Buying-Price = high) ^ (Lug_Boot = big) → Car_evaluation = unacceptable

Car	Buying_Price	Maintenance_ Price	Lug_Boot	Safety	Evaluation?
3	High	medium	big	high	?

Car 3 triggers rule R4 first → acceptable

2. **Exhaustive Rules :** It said to have a complete coverage for the rule based Classifier if it accounts for each doable attribute values combination. Every instance is roofed with a minimum of one rule.

Solution : Use a default class

5.1.4 Building Classification Rules

1. Direct Method : Sequential Covering Algorithm

- Training data is used to extract IF-THEN rules in Sequential Covering Algorithm. In sequential covering algorithm it is not required to construct a decision tree initially. In sequential covering algorithm, every rule that belongs to a given class covers maximum number of the records of that class.
- AQ, CN2, and RIPPER are some of the sequential Covering Algorithms. According to the general strategy the rules are trained one at a time. For every time rules are trained, a record covered by the rule is eliminated and therefore the method continues for the remainder of the records. This is often as a result of the path to every leaf in a decision tree corresponds to a rule.
- The Decision tree induction will be thought of as learning a group of rules at the same time.

Algorithm

(a) Start using an empty rule.

(b) Grow a rule using the, learn one rule at a time method. Either we can use general to specific or specific to general strategy.

(c) Remove training records covered by the rule, otherwise the next rule will be exactly same as that of the previous rule.

(d) Repeat above two steps until stopping criteria is reached. Stopping criteria can be computation of significant gain.

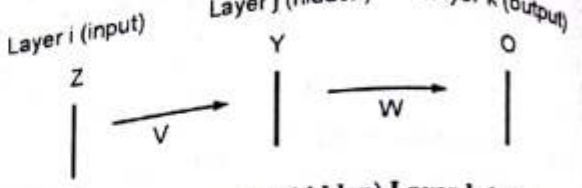
Module 5

2. Indirect Method : From Decision Trees

- In the Indirect method we will learn how to generate a rule-based classifier by extracting IF-THEN rules from a decision tree.
- Points that we need to remember while extracting a rule from a decision tree :
  - For each path from the root to the leaf node one rule is created.
  - Each splitting criterion is logically ANDed to form a rule condition or antecedent.
  - The leaf node represents the class prediction, forming the rule consequent.

5.2 CLASSIFICATION BY BACKPROPAGATION

5.2.1 Generalised Delta Learning Rule



Layer i (input) Layer j (hidden) Layer k (output)

- We will derive a general expression for the weight increment ΔV_{ji} for any layer of neurons that is not an output layer
- The above is a simplified diagram for the multi layer perceptron network which represents the three layer input (i), hidden (j) and output (k).
- At the input layer input Z is applied, Y represents the output of the hidden layer and O represents the output of the output layer. V is the weight vector present between the input and the hidden layer and W represents the weight vector present between the hidden and the output layer.
- According to –ve gradient descent formula for hidden layer ...(5.2.1)

$$\Delta V_{ji} = -\eta \, dE / d V_{ji}$$

- We may write the term $dE / d V_{ji}$ as ...(5.2.2)

$$dE / d V_{ji} = dE / dnet_j * dnet_j / dV_{ji}$$

$(dE / dnet_j)$ is the error signal for hidden layer which can be represented as dY_j

$dnetj/dV_{ji}$ represent the input applied at the input layer i.e., Zi

- By substituting this values in Equation (5.2.2) and Equation (5.2.1) we will get ...(5.2.3)

$$\Delta V_{ji} = \eta * dY_j * Zi$$

- As we are saying that $dY_j = - dE/dnet_j$
- We can write this term as ...(5.2.4)

$$dY_j = - dE/dY_j * dY_j/dnet_j \cdots$$

- In the above equation the second term is nothing but the $fj'(net_j)$ and the first term is the differentiation of error w.r.t Y
- As we know error is given as $E = (d - O)^2$

$$dE/dY_j = d/dY_j \, (\tfrac{1}{2} \sum \{d_k - f(net_k (y)^2 \})$$

$$= - \sum \{d_k - O_k) \, d/dY_j \, f(net_k (y))$$

$$= - \sum (d_k - O_k) * f'(net_k) * (d(net_k)/dY_j)$$

- We can write $(d_k - O_k) * f'(net_k) = d_{ok}$ which is the error present in the output layer and $d(net_k)/dY_j$ as Wk_j which is the weight vector present between the hidden and output layer.
- Thus we will get

$$dE/dY_j = - \sum d_{ok} W_{kj}$$

- By substituting this in Equation (4) we will get

$$dY_j = (\sum d_{ok} Wk_j) \, fj'(net_j)$$

By substituting this in Equation (3) we will get

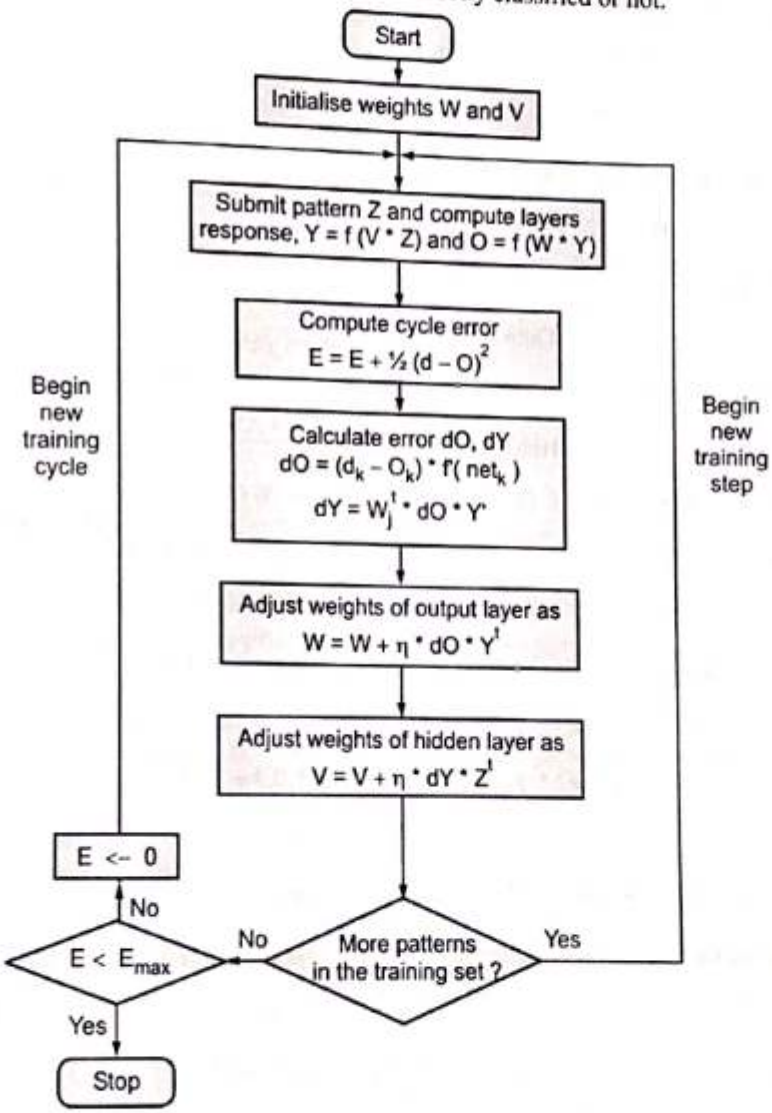
$$\Delta V_{ji} = \eta * Zi * f_j'(net_j) * (\sum d_{ok} Wkj)$$

Where $f_j'(net_j) = \frac{1}{2}(1 - O^2)$ for bipolar and $O(1 - O)$ for unipolar function. This is the Generalised Delta learning rule.

5.2.2 Error Back Propagation Training

In Error back propagation, the network learns during a training phase. The steps followed during learning are :

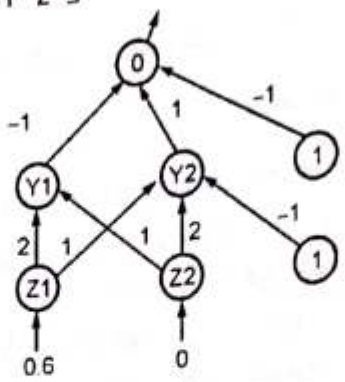
1. First the input is applied to the input layer to calculate the output of the hidden layer. The output of the hidden layer becomes the input of the next layer. Finally, the output of the output layer is calculated.
2. The desired and the actual output at the output layer are compared with each other and an error signal is generated.
3. The error of the output layer is back propagated to the hidden layer so that the weights connected in each layer of the network can be properly adjusted.
4. When Back propagation network is trained for the correct classification for a training data, then a testing data is applied to the network in order to check if the unseen patterns are correctly classified or not.



Module 5

Example 5.2.1 : Classify the input of following network using unipolar continuous function and EBPTA.

$$W = [-1\ 1]\ W_0 = -1,\ V = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix},\ V_0 = [0\ -1]\ t = 0.9,\ \eta = 0.3$$



☑ **Solution :**

- **Feedforward stage**

First, we will calculate the output of Hidden Layer as,

$$Y_1 = f(0.6 * 2 + 1 * 0) = 0.76$$

$$Y_2 = f(-1 + 0.6 * 1 + 2 * 0) = 0.4$$

Now we will calculate the output of Output Layer as,

$$O = f(0.76 * (-1) + 0.4 * 1 + (-1)) = 0.2$$

- **Back propagation of error**

First, we will calculate the Error of the Output Layer as,

$$dO = (t - O) f'(O) = (0.9 - 0.2) * 0.2 * (1 - 0.2) = 0.112$$

Now we will calculate the Error of the Hidden Layer as,

$$dY_1 = dO * (-1) * f'(Y_1) = 0.112 * (-1) * 0.76 (1 - 0.76) = -0.02$$

$$dY_2 = dO * (1) * f'(Y_2) = 0.112 * (1) * 0.4 (1 - 0.4) = 0.026$$

Updation of weights

First we will update the weights between the Hidden and the Output Layer as,

$$W_{11} = W_{11} + \eta * dO * Y_1 = -1 + 0.3 * 0.112 * 0.76 = -0.975$$

$$W_{12} = W_{12} + \eta * dO * Y_2 = 1 + 0.3 * 0.112 * 0.4 = 1.013$$

Bias is updated as

$$W_0 = W_0 + \eta * dO = -1 + 0.3 * 0.112 = -0.996$$

Now we will update the weights between the Hidden and the Input Layer as,

$$V_{11} = V_{11} + \eta * dY_1 * Z_1 = 2 + 0.3 * (-0.02) * 0.6 = 2.0036$$

$$V_{12} = V_{12} + \eta * dY_1 * Z_2 = 1 + 0.3 * (-0.02) * 0 = 1$$

$$V_{21} = V_{21} + \eta * dY_2 * Z_1 = 1 + 0.3 * 0.026 * 0.6 = 1.3$$

$$V_{22} = V_{22} + \eta * dY_2 * Z_2 = 2 + 0.3 * 0.026 * 0 = 2$$

Bias is updated as

$$V_{02} = V_{02} + \eta * dY_2 = -1 + 0.3 * 0.026 = -0.992$$

5.3 BAYESIAN BELIEF NETWORK

5.3.1 Bayes Theorem

- To understand the Bayesian belief network first we will revise the concepts of Bayes theorem.
- The conditional probability is represented in the following way.

$$P(C/A) = \frac{P(A, C)}{P(A)}$$

$$P(A/C) = \frac{P(A, C)}{P(C)}$$

- The Bayes theorem is defined using the following equation.

$$P(C/A) = \frac{P\left(\frac{A}{C}\right) P(C)}{P(A)}$$

- Now we will see the simple Bayes Theorem example.
- Given : A shop owner understands that due to Rain there is increase in sale of umbrella 50% of the time. Prior probability of rainy weather is 1/50,000. Prior probability of any customer purchasing umbrella is 1/20. If a customer has purchased umbrella, what's the probability the weather is rainy?

$$P(R/U) = \frac{P\left(\frac{U}{R}\right) P(R)}{P(U)} = \frac{0.5 * 1/50{,}000}{1/20} = \mathbf{0.0002}$$

5.3.2 Bayesian Classifiers

- In Bayesian classifier each attribute and class label is considered as a random variable. When the records with attributes $(A_1, A_2, \ldots, A_n)$ are given as a input to the classifier the goal is to predict the class C. Specifically we can say that, we want to find the value of C that maximizes $P(C \mid A_1, A_2, \ldots, A_n)$.
- Now the question is, can we estimate $P(C \mid A_1, A_2, \ldots, A_n)$ directly from data. The approach used is to compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem.

$$P(C \mid A_1, A_2, \ldots, A_n) = \frac{P(A_1, A_2, \ldots, A_n \mid C)\, P(C)}{P(A_1, A_2, \ldots, A_n)}$$

- Now we have to select the value of C that maximizes $P(C \mid A_1, A_2, \ldots, A_n)$. It will be as good as selecting the C that gives a maximum value for $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$. Next question is calculation of $P(A_1, A_2, \ldots, A_n \mid C)$, that we will see in the following sections.

Module 5

5.3.3 Naïve Bayes Classifier

When class is given assume attributes are independent of each other. So $P(A_1, A_2, \ldots, A_n | C)$ can be represented in the following way.

$$P(A_1, A_2, \ldots, A_n | C) = P(A_1 | C_j) P(A_2 | C_j) \ldots P(A_n | C_j)$$

$P(A_i | C_j)$ is calculated for all classes C_j and attributes A_i. The test record is assigned as a class C_j,

if $P(C_j) * P(A_i | C_j)$ is maximum.

Example 1: Naïve Bayes Classifier

Person id	Loan(Y/N)	Status(S:Single,M:Married,D:Divorced)	Salary	Savings? (Y/N)
1	Y	S	125k	N
2	N	M	100k	N
3	N	S	70k	N
4	Y	M	120k	N
5	N	D	95k	Y
6	N	M	60k	N
7	Y	D	220k	N
8	N	S	85k	Y
9	N	M	75k	N
10	N	S	90k	Y

Class: $P(C) = N_c / N$

$P(N) = 7/10$, $P(Y) = 3/10$

If the attributes are discrete then the probability is calculated as:

$$P(A_i | C_k) = |A_{ik}| / N_c$$

Here $|A_{ik}|$ represents the number of records having attribute A_i and belongs to class C_k

Examples : $P(\text{Status} = M|N) = 4/7$, $P(\text{Loan} = Y|Y) = 0$

If the attributes are continuous then the probability is calculated using Normal distribution for each combination of A_i and C_j

$$P(A_i / C_j) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For (Salary, N): mean = 110 and variance = 2975

$$P(\text{Salary} = 120/N) = \frac{1}{\sqrt{2\pi * 2975}} e^{-\frac{(120-110)^2}{2*2975}} = 0.0072$$

Suppose a record X = (Loan = N, M, Salary = 120K) is given for testing

$P(\text{Loan} = Y|N) = 3/7$

$P(\text{Loan} = N|N) = 4/7$

P (Loan = Y|Y) = 0

P (Loan = N|Y) = 1

P (Status = S|N) = 2/7

P (Status = D|N) = 1/7

P (Status = M|N) = 4/7

P (Status = S|Y) = 2/7

P (Status = D|Y) = 1/7

P (Status = M|Y) = 0

For Salary : If Savings = N : mean = 110, variance = 2975

If Savings = Y : mean = 90, variance = 25

P(X | Savings = N)

= P (Loan = N| Savings = N) × P (Status = M | Savings = N) × P (Salary = 120K | Savings = N)

= 4/7 × 4/7 × 0.0072 = 0.0024

P (X| Savings = Y)

= P (Loan = N | Savings = Y) × P (Status = M | Savings = Y) × P (Salary = 120K | Savings = Y)

= $1 \times 0 \times 1.2 \times 10^{-9} = 0$

Since P (X | N) P (N) > P (X|Y) P(Y)

Therefore P (N|X) > P (Y|X) = > Savings = N for the given record.

Example 2 : Naïve Bayes Classifier

We want to classify record < Red, Domestic, SUV>

Car no	Colour	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Module 5

P (yes) = 5/10, P (no) = 5/10

Color-> P (Red/yes) = 3/5, P (Red/no) = 2/5

P (Yellow/yes) = 2/5, P (Yellow/no) = 3/5

Type-> P (SUV/yes) = 3/5, P (SUV/no) = 2/5

P (Sports/yes) = 2/5, P (Sports/no) = 3/5

Origin-> P (Domestic/yes) = 3/5, P (Domestic/no) = 2/5

P (Imported/yes) = 2/5, P (Imported/no) = 3/5

P (Red, Domestic, SUV/Y)*P(Y)

= P (Red/yes)* P (Domestic/yes)* P (SUV/yes)*P(Y) = 3/5*2/5*1/5*5/10

= 0.024

P (Red, Domestic, SUV/N)*P (N)

= P (Red/no)* P (Domestic/no)* P (SUV/no)*P (N) = 2/5*3/5*3/5*5/10

= 0.072

P (Red, Domestic, SUV /N)*P (N)> P (Red, Domestic, SUV /Y)*P(Y)

So, record <Red, Domestic, SUV> is classified as Stolen = No

UExample 5.3.1 **MU - Dec. 19, 10 Marks**

For a unknown tuple t =<Outllook =Sunny, Temperature =Cool, Wind= Strong> use naïve Bayes classifier to find whether the class for PlayTennis is yes or no. The dataset is given below

Outlook	Temperature	Wind	Play Tennis
Sunny	Hot	Weak	No
Sunny	Hot	Strong	No
Overcast	Hot	Weak	Yes
Rain	Mild	Weak	Yes
Rain	Cool	Weak	Yes
Rain	Cool	Strong	No
Overcast	Cool	Strong	Yes
Sunny	Mild	Weak	No
Sunny	Cool	Weak	Yes
Rain	Mild	Weak	Yes
Sunny	Mild	Strong	Yes
Overcast	Mild	Strong	Yes
Overcast	Hot	Weak	Yes
Rain	Mild	Strong	No

☑ **Solution :**

P (Sunny, Cool, Strong /Yes)*P(Yes)

= P (Sunny/yes)* P (Cool/yes)* P (Strong/yes)*P(Yes) = 3/5*1/5*3/5*5/14

= 0.0257

P (Sunny, Cool, Strong /No)*P(No)

= P (Sunny/no)* P (Cool/no)* P (Strong/no)*P(no) = 2/9*3/9*3/9*9/14

= 0.0158

P (Sunny, Cool, Strong /Y)*P (Y) > P (Sunny, Cool, Strong /N)*P(N)

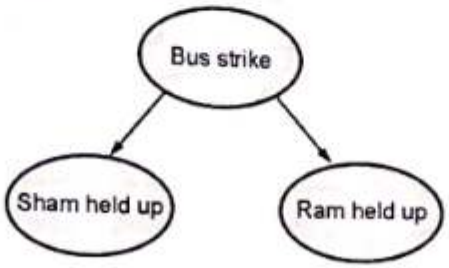
So record < Sunny, Cool, Strong > is classified as Play Tennis = Yes

5.3.4 Bayesian Belief Network

A Bayesian Belief Network is an uncommon sort of chart (called a coordinated diagram) together with a related arrangement for table of probabilities. The graph comprises of arcs and nodes. The discrete or continuous variables are represented by the nodes. The causal relationships between the variables are represented by the arcs.

Bayesian Belief Networks empower us to model and reason about *uncertainty*. Bayesian Belief Networks suits probabilities based on ***objective*** data as well as ***subjective*** probabilities. The important utilization of Bayesian Belief Networks is the use of ***revising probabilities*** based on the observation of actual events.

Example 5.3.2 : We need to find the probability that 'Ram is held up'.



Sham held up	Bus strike	
	T	F
T	0.2	0.8
F	0.3	0.7

Ram held up	Bus strike	
	T	F
T	0.9	0.1
F	0.4	0.6

Bus strike	
T	F
0.2	0.8

Module 5

Solution :

P (Ram held up) = P (Ram held upl Bus strike) * P (Bus strike) + P (Ram held upl No bus strike) * P (No bus strike)

= (0.9 * 0.2) + (0.1 * 0.8) = 0.26

This is called the marginal probability

The most vital utilization of Bayesian Belief Networks is in *revising probabilities* in the light of actual observations of events.

Let's, for example, we *know* there is a bus strike. In this case **evidence** 'bus strike = T' can be used.

The conditional probability tables already tell us the revised probabilities for Ram being held up (0.9) and Sham being held up (0.2).

Suppose, however, that we do not know if there is a bus strike but do know that Ram is held up.

Then we can enter the evidence that 'Ram held up = true' and we can use this observation to determine:

(a) The (revised) probability that there is a bus strike, and

(b) The (revised) probability that Sham will held up

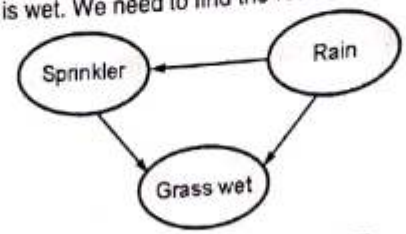
(c) P (bus strikel Ram held up) = P (Ram held upl bus strike) * P (bus strike) / P (Ram held up) = (0.9 * 0.2) / 0.26 = 0.69

(d) P (Sham held up) = P (Sham held upl bus strike) * P (bus strike) + P (Sham held upl No bus strike) * P (No bus strike)

= (0.2 * 0.69) + (0.8 * 0.31) = 0.38

Probabilities are updated based on the arrival of the evidences, this is called as **propagation.**

Example 5.3.3 : It is observed that grass is wet. We need to find the reason that it is due to rain or sprinkler.



Rain	Sprinkler	
	True	False
False	0.4	0.6
True	0.01	0.99

Rain	
True	False
0.2	0.8

Sprinkler	Rain	Grasswet	
		True	False
False	False	0	1
False	True	0.8	0.2
True	False	0.9	0.1
True	True	0.99	0.01

☑ **Solution :**

First we will write the joint probability distribution equation using the given graph.

Joint probability distribution: P (G, S, R) = P (G | S, R) P (S | R) P (R)

We need to find the reason of wet grass (Rain or Sprinkler)

First we will find the probability of wet grass due to rain,

$$P(R = T \mid G = T) = \frac{P(G = T, R = T)}{P(G = T)}$$

$$= \frac{\Sigma_{S \in (T, F)} P(G = T, S, R = T)}{\Sigma_{S, R \in (T, F)} P(G = T, S, R)}$$

Now each probability term is to be written in the form of joint probability distribution as follows.

$$P(G = T, S = T, R = T) = P(G = T \mid S = T, R = T)\, P(S = T \mid R = T)\, P(R = T)$$
$$= 0.99 * 0.01 * 0.2 = 0.00198$$

Similarly we will calculate all probabilities and final probability is calculated as below,

$$P(R = T \mid G = T) = \frac{0.00198_{TTT} + 0.1584_{TFT}}{0.00198_{TTT} + 0.288_{TTF} + 0.1584_{TFT} + 0_{TFF}} = 35.77$$

Now we will find the probability of wet grass due to sprinkler.

$$P(S = T \mid G = T) = \frac{P(G = T, S = T)}{P(G = T)} = \frac{\Sigma_{R \in (T,F)} P(G = T, S = T, R)}{\Sigma_{S, R \in (T,F)} P(G = T, S, R)} = 64.23$$

Since $P(S = T \mid G = T) > P(R = T \mid G = T)$ the reason for the wet grass is "sprinkler is on".

5.4 HIDDEN MARKOV MODEL

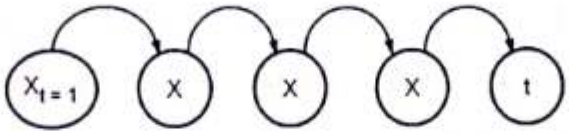
5.4.1 Markov Models

- A Markov model is a discrete finite system that has N distinct states. The model starts at time t = 1 called as initial state. As the time step increases the system moves from present state to next state with the transition probabilities that are assigned to present state. Such type of system is known as a discrete, or finite Markov model.
- In Discrete Markov Model every a_{ij} indicates the probability of transition to state j from state i. The a_{ij} are stored in $A = \{a_{ij}\}$ matrix. p_i is the probability to begin from a given state i. These start probabilities are indicated by vector p.

Markov Model Property

At t + 1 time the state of the system depends only on the state of the system at time t.

$$P(X_{t+1} = x_{t+1} \mid X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_1 = x_1, X_0 = x_0)$$
$$= P(X_{t+1} = x_{t+1} \mid X_t = x_t)$$



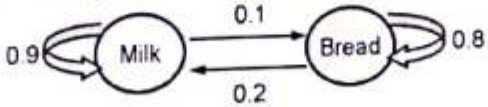
Markov Chains : Probabilities independent of t when process is "stationary"

So, for all t, $P(X_{t+1} = x_{t+1} \mid X_t = x_t) = p_{ij}$

This can be inferred as p_{ij} represents the probability with which the system will be present in the next system without depending on the value of t, if the present system is in state *i*.

Mod 5

Example 5.4.1 : Suppose a person has purchased milk, then there is a 90% chance that his next purchase will also be milk. If the same person purchased bread then there is an 80% chance that his next purchase will also be bread. Let's assume that a person currently purchased milk, what is the probability that he will purchase bread two purchases from now and three purchases from now?



☑ **Solution :**

$$A = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}$$

$$A^2 = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}\begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} = \begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix}$$

The probability of he will purchase bread two purchases from now is 0.17.

$$A^3 = \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix}\begin{bmatrix} 0.83 & 0.17 \\ 0.34 & 0.66 \end{bmatrix} = \begin{bmatrix} 0.781 & 0.219 \\ 0.438 & 0.562 \end{bmatrix}$$

The probability of he will purchase bread three purchases from now is 0.219

Example 5.4.2 : Given that the weather on day 1 (t = 1) is sunny, what is the probability for the observation O = sunny, sunny, sunny, rainy, rainy, sunny, cloudy, sunny. The states are represented as, Rainy: 1, Cloudy: 2, Sunny: 3. The transition matrix is given as A.

$$A = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

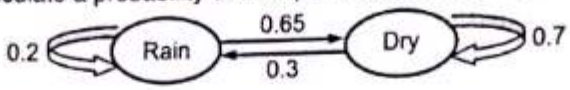
☑ **Solution :**

$$P(O \mid Model) = P(3, 3, 3, 1, 1, 3, 2, 3 \mid Model) = P(3)\,P(3|3)\,P(3|3)\,P(1|3)\,P(1|1)\,P(3|1)\,P(2|3)\,P(3|2)$$

$$= \pi_3 a_{33} a_{33} a_{31} a_{11} a_{13} a_{32} a_{23} = 1.0 * 0.8 * 0.8 * 0.1 * 0.4 * 0.3 * 0.1 * 0.2 = 1.536 \times 10^{-4}$$

UExample 5.4.3 **MU - May 19, 10 Marks**

Consider Markov chain model for 'Rain' and 'Dry' is shown in following figure Two states: 'Rain' and 'Dry'. Transition probabilities : P('Rain'|'Rain') = 0.2, P('Dry'|'Rain') = 0.65, P('Rain'|'Dry') = 0.3, P('Dry'|'Dry') = 0.7, Initial probabilities: say P('Rain') = 0.4, P('Dry') = 0.6.Calculate a probability of a sequence of states {'Dry', 'Rain', 'Rain', 'Dry'}.



☑ :

$$P(O \mid Model) = P(Dry, Rain, Rain, Dry \mid Model) = P(Dry)\,P(Rain|Dry)\,P(Rain|Rain)\,P(Dry|Rain)$$

$$= 0.6 * 0.3 * 0.2 * 0.65 = 0.0234$$

5.4.2 Main issues using HMMs

1. **Evaluation Problem :** Given observation sequence $O = O_1, O_2, \ldots, O_T$ and λ how to compute $P(O|\lambda)$.

   For example, calculation of probability of observing HTTHHHT.

   $A = \begin{bmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{bmatrix}$, $b_F(H) = 0.5$, $b_F(T) = 0.5$, $b_B(H) = 0.8$, $b_B(T) = 0.2$, $\pi_F = 0.6$, $\pi_B = 0.4$

   Here A represents transition probability matrix and b represents emission probabilities

   Suppose state sequence Q = FFBFFBB

   Then, $P(O|\lambda) = b_F(H)\,b_F(T)\,b_B(T)\,b_F(H)\,b_F(H)\,b_B(H)\,b_B(T)$

   $= 0.5 * 0.5 * 0.2 * 0.5 * 0.5 * 0.8 * 0.2$

Suppose state sequence Q = BFBFBBB

Then, $P(O \mid \lambda) = b_B(H)\, b_F(T)\, b_B(T)\, b_F(H)\, b_B(H)\, b_B(H)\, b_B(T)$

$= 0.8 * 0.5 * 0.2 * 0.5 * 0.8 * 0.8 * 0.2$

For each given state sequence Q the probability of O will be different.

Now the question is how likely are we to get this path Q.

$Q = FFBFFBB = \pi_F\, a_{FF}\, a_{FB}\, a_{BF}\, a_{FF}\, a_{FB}\, a_{BB} = 0.6 * 0.9 * 0.1 * 0.3 * 0.9 * 0.1 * 0.7$

$Q = BFBFBBB = \pi_B\, a_{BF}\, a_{FB}\, a_{BF}\, a_{FB}\, a_{BB}\, a_{BB} = 0.4 * 0.3 * 0.1 * 0.3 * 0.1 * 0.7 * 0.7$

Each given path Q has its own probability. So the question is how to find $P(Q \mid \lambda)$

Solution :

1. **Forward Procedure :** The forward algorithm is mostly used in applications that need us to determine the probability of being in a specific state when we know about the sequence of observations

Initialisation : $\alpha_1(i) = \pi_i\, b_i(O_1)$

$$\alpha_{t+1}(J) = \left[\sum_{i=1}^{N} \alpha_t(i)\, a_{ij}\right] b_j(O_{t+1})$$

Induction :

$O = HTTHHHT$

$\alpha_1(F) = \pi_F\, b_F(H) = 0.6*0.5 = 0.2$

$\alpha_1(B) = \pi_B\, b_B(H) = 0.4*0.8 = 0.48$

Maximum probability is of B, so B is selected.

$\alpha_2(F) = [\alpha_1(F)\, a_{FF} + \alpha_1(B)\, a_{BF}]\, b_F(T) = (0.2 * 0.9 + 0.48 * 0.3) * 0.5 = 0.162$

$\alpha_2(B) = [\alpha_1(F)\, a_{FB} + \alpha_1(B)\, a_{BB}]\, b_B(T) = (0.2 * 0.1 + 0.48 * 0.7) * 0.2 = 0.0172$

Maximum probability is of F, so F is selected.

Similarly α_3 to α_7 are calculated. Based on this path Q is decided. Once we get Q we can easily calculate $P(O \mid \lambda)$.

2. **Backward procedure**

Initialisation : $\beta_T(i) = 1$

Induction : $\beta_T(i) = \sum_{j=1}^{N} a_{ij}\, b_j\, O_{t+1}\, \beta_{t+1}(j)$

$O = HTTHHHT$

$\beta_T(F) = 1$

$\beta_T(B) = 1$

$\beta_{-1}(F) = a_{FF}\, b_F(T) * 1 + a_{FB}\, b_B(T) * 1 = 0.9 * 0.5 + 0.1 * 0.2 = 0.47$

$\beta_{-1}(B) = a_{BF}\, b_F(T) * 1 + a_{BB}\, b_B(T) * 1 = 0.3 * 0.5 + 0.7 * 0.2 = 0.29$

Module 5

Maximum probability is of F, so F is selected.

$$\beta_{-2}(F) = a_{FF} b_F(H) * \beta_{-1}(F) + a_{FB} b_B(H) * \beta_{-1}(B)$$

$$= 0.9 * 0.5 * 0.47 + 0.1 * 0.8 * 0.29 = 0.234$$

$$\beta_{-2}(F) = a_{BF} b_F(H) * \beta_{-1}(F) + a_{BB} b_B(H) * \beta_{-1}(B)$$

$$= 0.3 * 0.5 * 0.47 + 0.7 * 0.8 * 0.29 = 0.2329$$

Maximum probability is of F, so F is selected.

Similarly, β_{-3} to β_{-7} are calculated, Based on this path Q is decided. Once we get Q we can easily calculate $P(Q \mid \lambda)$.

2. **Decoding Problem :** Given observation sequence O and λ how to choose state sequence $Q = q_1 q_1 \ldots q_7$

For example, what is hidden coin behind each flip.

☑ **Solution :**

Forward-Backward algorithm

$$y_t(i) = \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{j=1}^{N} \alpha_t(i)\,\beta_t(i)}$$

$y_t(F)$ and $y_t(B)$ is calculated and maximum is selected.

Based on this path sequence Q is decided

3. **Learning Problem :** How to estimate $\lambda = (A, B, M)$ so as to maximize $P(Q \mid \lambda)$

For example, How to estimate coin parameter λ.

☑ **Solution :**

Use EM algorithm

Guess initial HMM parameters

E step : Compute distribution over paths

M step : Compute max likelihood parameters

But how do we do this efficiently

Also known as the Baum-Welch algorithm

Compute probability of each state at each position using forward and backward probabilities

→ (Expected) observation counts

Compute probability of each pair of states at each pair of consecutive positions *i* and *i + 1* using *forward (i)* and *backward (i + 1)*

→ (Expected) transition counts

5.5 SUPPORT VECTOR MACHINE

5.5.1 Maximum Margin Linear Separators

- Support Vector machine is a type of supervised learning that can be used for classification or regression. Even if the data points are unseen (not from the training dataset), support vector machine classifies the data properly.
- Let's take an example of dataset that belongs to two different categories, and the distribution of data is proper means the data is separated from each other properly. In this case we can draw a straight line (decision boundary) on the graph in such a way that the input space is divided in to two regions.
- Data points that belongs to one category lies on one side of the decision boundary and the data points of other category lies on the opposite side. Such type of data is called as linearly separable data.

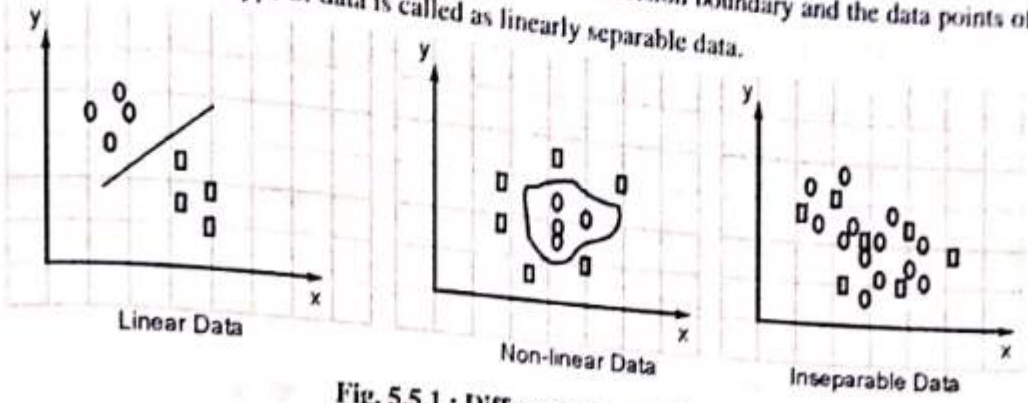


Fig. 5.5.1 : Different types of data

- Separating hyperplane is the line which is used to separate the dataset. If we are using simple 2- dimensional plots then it's just a line. We require a plane to separate the data if data is 3 dimensional. So we can say that if data is N dimensional, we require N-1 dimensional hyperplane.
- We want that our classifier should be designed in a manner that if a data point is far away from the decision boundary then we will be more confident about the prediction we have made.
- We would like to find the data point near to the separating hyperplane and also make sure that this point should be far away from the separating line as possible. This is called as margin. We would like to find the greatest possible margin, because if we trained our classifier on limited data or made a mistake, we would want it to be as robust as possible.
- Support vectors are the points which are nearest to the separating hyperplane. We have to maximize the distance between the support vectors and the separating line.

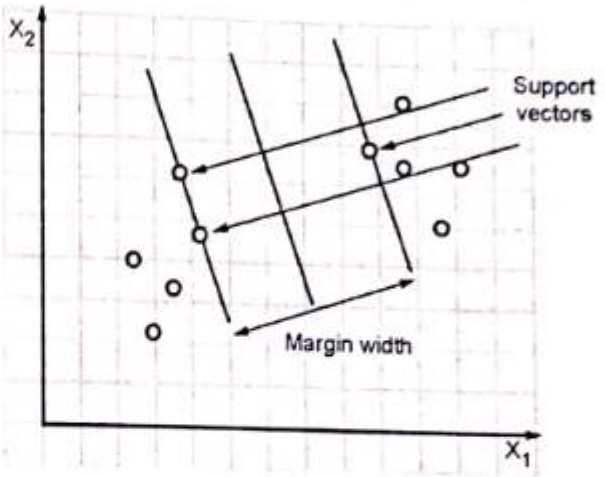


Fig. 5.5.2 : Support vectors and Margin

Distance between a hyperplane (w, b) and a point x is calculated as,

$$\text{Distance} = \frac{|w^t x + b|}{\|w\|}$$

Module 5

5.5.2 Quadratic Programming Solution to Find Maximum Margin Separator

- A hyperplane is formally defined by the following notation as,

$$F(x) = w^t x + b$$

In above equation, w represents the weight vector and b represents the bias.

- By scaling the values of w and b we can represent the optimal hyperplane in many ways. As a matter of convention among all the possible notations of the hyperplane the one selected is

$$|w^t x + b| = 1$$

- Here x represents the training records closest to the hyperplane. In general the training records that are closest to the hyperplane are called as support vectors. This notation is called as the canonical hyperplane.
- The distance between a point x and a hyperplane (w, b) is given by the result of geometry as follows,

$$\text{Distance} = \frac{|w^t x + b|}{\|w\|}$$

- In general, the numerator is equal to one for the canonical hyperplane and distance to the support vector is given as,

$$\text{Distance}_{sv} = \frac{|w^t x + b|}{\|w\|} = \frac{1}{\|w\|}$$

- Margin is twice the distance to nearest samples

$$M = \frac{2}{\|w\|}$$

- Ultimately, the task of maximizing M is same as compared to the task of minimizing a function L (w) subject to some conditions. The conditions used to model the requirements for correct classification of all training samples x_i by the hyperplane are formally stated as,

$$\min_{w, b} L(w) = \frac{1}{2}\|w\|^2 \text{ subject to } y_i (w^t x + b) \geq 1 \text{ for all i.}$$

Where y_i represents the labels of training

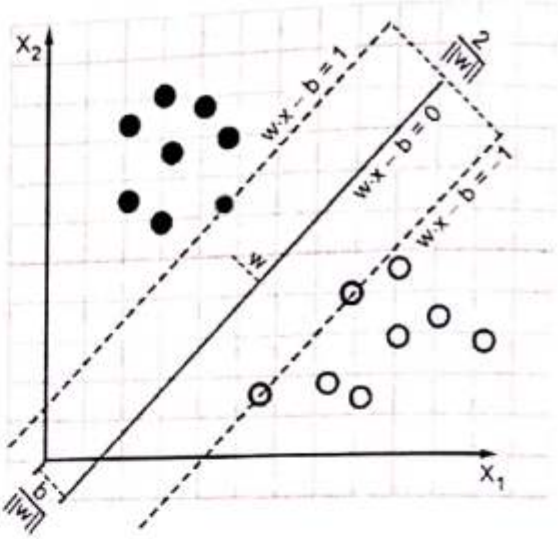


Fig. 5.5.3 : Solution to find maximum margin

This is a problem of Lagrangian optimization that can be solved using Lagrange multiplier to calculate weight vector 'w' and the bias 'b' of the optimal hyperplane.

Let's assume that we have 2 classes of 2 dimensional data to separate. Let's also assume that each class consist of only one point

These points are

$$X_1 = A_1 = (3, 3)$$

$$X_2 = B_1 = (6, 6)$$

Find the hyper plane that separates theses 2 classes

$$f(w) = \frac{1}{2}\|w\|^2$$

The constraints are,

$$c_1(w, b) = y_i | wx_1 + b | - 1 \geq 0$$

$$c_1(w, b) = 1 | wx_1 + b | - 1 \geq 0$$

$$c_2(w, b) = -1 | wx_2 + b | - 1 \geq 0$$

Next, we put equation into form of Lagrangian

$$L(w, b, m) = f(w) - m_1 c_1(w, b) - m_2 c_2(w, b)$$

$$= \frac{1}{2}\|w\|^2 - m_1((wx_1 + b) - 1) - m_2(-(wx_2 + b) - 1)$$

$$= \frac{1}{2}\|w\|^2 - m_1((wx_1 + b) - 1) + m_2((wx_2 + b) + 1)$$

We solve for the gradient of Lagrangian

$$\nabla L(w, b, m) = \nabla f(w) - m_1 \nabla c_1(w, b) + m_2 \nabla c_2(w, b) = 0$$

$$\frac{\partial}{\partial w} L(w, b, m) = w - m_1 x_1 + m_2 x_2 = 0 \quad \ldots(5.5.1)$$

$$\frac{\partial}{\partial b} L(w, b, m) = -m_1 + m_2 = 0 \quad \ldots(5.5.2)$$

$$\frac{\partial}{\partial \lambda_1} L(w, b, m) = (wx_1 + b) - 1 = 0 \quad \ldots(5.5.3)$$

$$\frac{\partial}{\partial \lambda_2} L(w, b, m) = (wx_2 + b) + 1 = 0 \quad \ldots(5.5.4)$$

Equating Equation (5.5.3) and (5.5.4), we get

$$(wx_1 + b) - 1 = (wx_2 + b) + 1$$

$$(wx_1) - 1 = (wx_2) + 1$$

$$(wx_1) - (wx_2) = 2$$

$$w(x_1 - x_2) = 2$$

w is divided into parts as,

$$w = (w_1, w_2)$$

$$W(x_1 - x_2) = 2$$

$$(w_1, w_2)[(3, 3) - (6, 6)] = 2$$

$$(w_1, w_2)[(-3, -3)] = 2$$

$$-3w_1 - 3w_2 = 2$$

$$w_1 = -(0.67 + w_2) \quad \ldots(5.5.5)$$

Module 5

Adding values to Equation (5.5.1) and combining with Equation (5.5.2)

$(w_1, w_2) - m_1 (1, 1) + m_2 (2, 2) = 0$

From Equation (5.5.2)

$$m_1 = m_2$$

$(w_1, w_2) - m_1 (3, 3) + m\lambda_1 (6, 6) = 0$

$$(w_1, w_2) + m_1 (3, 3) = 0 \quad \ldots(5.5.6)$$

$$w_1 + 3 m_1 = 0 \quad \ldots(5.5.7)$$

$$w_2 + 3 m_1 = 0$$

Equating these we get,

$$w_1 = w_2$$

Putting this in Equation (5.5.5)

$$w_1 = w_2 = -0.34$$

Putting this in either Equation (5.5.6) or Equation (5.5.7) will give

$$m_1 = m_2 = 0.11$$

And finally, using this in Equation (5.5.3) and Equation (5.5.4)

$$b = 1 - (wx_1) \text{ or } = -1 - (w x_2)$$

$$= 1 - ((-0.34, -0.34), (3, 3)) \text{ or } = 1 - ((-0.34, -0.34), (6, 6)) = 3.04$$

5.5.3 Kernels for Learning Non-Linear Functions

- Linear classifiers are able to separate only linearly separable data. Support vector machine provides the solution to this problem by transforming an input space into a feature space that contains non linear features. A hyperplane is constructed in the feature space so that other equations remain the same.
- This is also known as non-linear support vector machine. Here we separate the data linearly using a high dimensional space. Kernel functions with its own set of variables are used for this purpose. The result is going to be non linear if we convert this back to the original feature space.

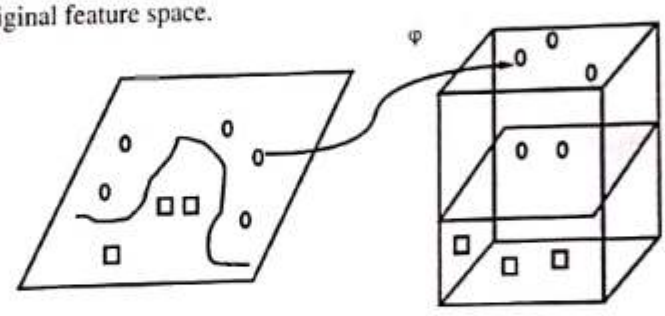


Fig. 5.5.4 : Mapping of Input space to Feature space

Particularly, the data is preprocessed with

$$X \rightarrow \Phi (x)$$

And then $\Phi(x)$ is mapped to y

$$F(x) = w\,\Phi(x) + b$$

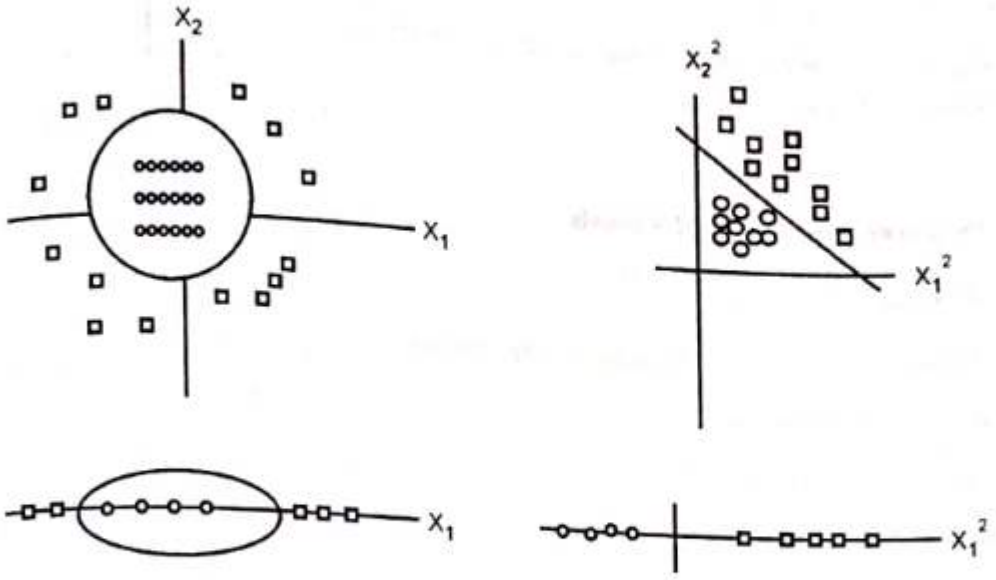


Fig. 5.5.5 : Mapping of one feature space to another feature space

- Kernel function transforms the data in to an easily understandable form. This is done via mapping input space to another feature space. In support vector machine inner product is calculated of two vectors and the result of this is always a single number. When we replace this inner product by a kernel it is called as kernel trick.
- There are different algorithms that use different kinds of kernel functions. Among the different types of kernel functions such as nonlinear, linear, radial basis function (RBF), polynomial, and sigmoid, RBF is mostly used. The reason for this is along the X-axis radial basis function gives the localized and finite response.
- The number of support vectors will be determined based on the different criteria's such as what is the complexity of the model, how much slack is allowed. One or more than one support vectors need to be defined for every complications in the final model from the input space. Support vector machines output compromises of support vectors and alpha. This is used to specify the effect of support vectors on the final decision.
- If we select the model with high complexity it will result in to over fitting. For better generalization if large margin is selected then it may lead to incorrect classification. And accuracy depends on the trade-off between these two selections criteria. If we over fit the data then the range of support vectors may vary from very less to each single point. This tradeoff is controlled through the selection of kernel and its parameters.
- In support vector machine the data points are tested by taking the dot product of each support vector with the test point. Hence the computational complexity increases, if we increase the number of support vectors. Classification of test points will be faster if we have less number of support vectors.

Module 5

5.5.4 Rules for the Kernel Function

Kernel function or a window is defined as follows:

If $\|\bar{x}\| \leq 1$ then $K(\bar{x}) = 1$ else 0.

This kernel function is shown by the Fig. 5.5.6,

$$K((z - x_i)/h) = 1$$

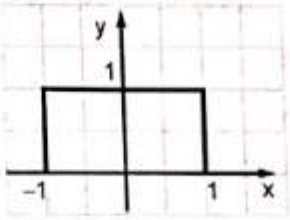


Fig. 5.5.6

For a fixed value of x_i, the function takes the value as 1 as shown in Fig. 5.5.7.

By selecting the argument of K (·), window can be moved to be centerd at the point x_i and to be of radius h.

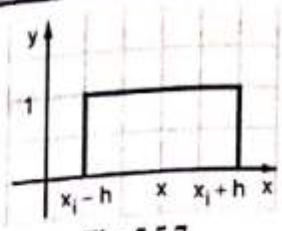


Fig. 5.5.7

5.5.5 Different Types of SVM Kernels

1. Polynomial kernel

Polynomial kernel is mostly used in image processing methods.

Polynomial Kernel is represented as,

$$K(x_i, x_k) = (x_i \cdot x_k + 1)^p$$

Here p represents the degree of the polynomial.

2. Gaussian kernel

There are some applications where prior knowledge is not available. For this type of applications Gaussian kernel is used.

Gaussian kernel is defined as,

$$K(x, y) = \exp\left(-\frac{\| x - y \|^2}{2\sigma^2}\right)$$

3. Gaussian radial basis function (RBF)

This is also used for the applications where prior knowledge is not available.

Gaussian radial basis function is defined as,

$$K(x_i, x_j) = \exp(-\gamma \| x - y \|^2) \text{ for } \gamma > 0$$

Sometimes it is parametrized using the value of γ as $1/2\sigma^2$

4. Laplace RBF kernel

Laplace RBF kernel is defined as,

$$K(x, y) = \exp\left(-\frac{\| x - y \|}{\sigma}\right)$$

5. Hyperbolic tangent kernel

Hyperbolic tangent kernel is used in neural networks.

It is defined as,

$$K(x_i, x_j) = \tanh(kx_i * x_j + c), \text{ for some (not every) } k > 0 \text{ and } c < 0.$$

6. Sigmoid kernel

Sigmoid kernel can be used as a proxy for neural networks.

It is defined as,

$$K(x, y) = \tanh(\alpha x^d y + c)$$

7. Bessel function of the first kind Kernel

Cross terms in mathematical functions can be removed by using this type of kernel function.

It is defined as,

$$K(x, y) = \frac{J_{v+1}(\sigma \| x - y \|)}{\| x - y \|^{-n(v+1)}}$$

Here j represents the Bessel function of first type.

8. ANOVA radial basis kernel

In regression problems this kernel can be used.

It is defined as,

$$K(x, y) = \sum_{k=1}^{n} \exp(-\sigma (x^k - y^k)^2)^d$$

5.6 CLUSTERING

- In unsupervised learning the most important task is the Clustering. Clustering is used to store data points in to related groups. In clustering advance knowledge is not present about the group definitions.

Definition : "Clustering is a process of partitioning a set of data in a set of meaningful sub-classes, called as clusters".

- In clustering we group the "similar" objects in one cluster and "dissimilar" objects in another cluster.

5.6.1 K- means Clustering

- To solve the well known clustering problem K-means is used, which is one of the simplest unsupervised learning algorithms. Given data set is classified assuming some prior number of clusters through a simple and easy procedure. In k- means clustering for each cluster one centroid is defined. Total there are k centroids.
- The centroids should be defined in a tricky way because result differs based on the location of centroids. To get the better results we need to place the centroids far away from each other as much as possible. Next, each point from the given data set is stored in a group with closest centroid. This process is repeated for all the points.
- The first step is finished when all points are grouped. In the next step new k centroids are calculated again from the result of the earlier step.
- After finding these new k centroids, a new grouping is done for the data points and closest new centroids. This process is done iteratively.

Module 5

- The process is repeated unless and until no data point moves from one group to another. The aim of this algorithm is to minimize an objective function such as sum of a squared error function. The objective function is defined as follows :

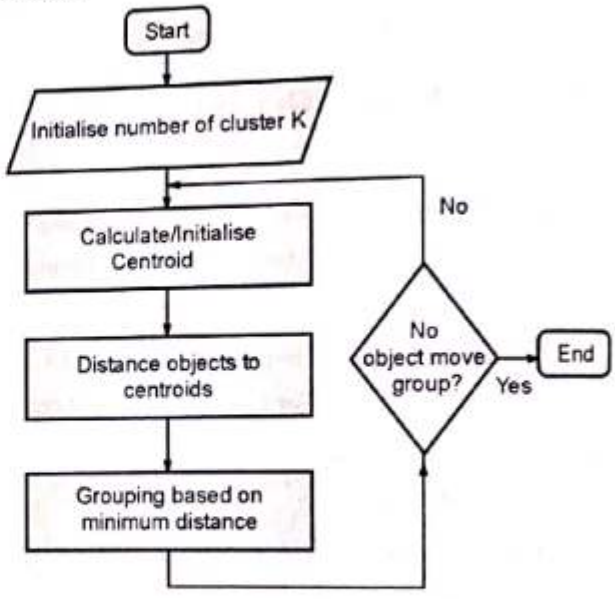
$$J = \sum_{j=1}^{k} \sum_{i=1}^{x} \| x_i^j - C_j \|^2$$

- Here $\| x_i^j - C_j \|^2$ shows the selected distance measure between a data point x_i^j and the cluster centre C_j. It is a representation of the distance of the n data points from their respective cluster centers.
- The algorithm is comprises of the following steps :
  1. Identify the K centroids for the given data points that we want to cluster.
  2. Store each data point in the group that has the nearest centroid.
  3. When all data points have been stored, redefine the K centroids.
  4. Repeat Steps 2 and 3 until the no data points move from one group to another. The result of this process is the clusters from which the metric to be minimized can be calculated.
- The k-means algorithm does not guarantee the most optimal solution corresponding to global minimum objective function, although it can be proved that the process will always terminate. Initial random selection of cluster centers affects the performance of the algorithm. The k-means algorithm is applied for a number of times to reduce this effect.
- Let's assume that n sample data points $x_1, x_2, ..., x_n$ of the same class are present, and we know that the data points belongs to k clusters, $k < n$. Let m_i represents the mean of the data points in cluster i. x can be stored in cluster i, if $\| x - m_i \|$ is the minimum of all the k distances.
- The k-means procedure is shown below:
- Select initial values for the means $m_1, m_2, ..., m_k$

  Until no data point moves from one group to another
  - Use the calculated means to group the data points into clusters
  - For i from 1 to k

    Mean of all of the samples for cluster i is used to replace m_i with the
  - end_for

  end_until
- The K-means algorithm is implemented in three steps.
- Iterate until stable (= no data point move group)
  1. Determine the centroid coordinate
  2. Determine the distance of each data point to the centroid
  3. Group the data points based on minimum distance.



Example 5.6.1 : Given { 2, 4, 10, 12, 3, 20, 30, 11, 25}. Assume number of clusters i.e. K = 2

Solution :

Randomly assign means: $m_1 = 3, m_2 = 4$

The numbers which are close to mean $m_1 = 3$ are grouped into cluster k_1 and others in k_2

Again calculate new mean for new cluster group.

$K_1 = \{2, 3\}, k_2 = \{4, 10, 12, 20, 30, 11, 25\}$ $m_1 = 2.5, m_2 = 16$

$K_1 = \{2, 3, 4\}, k_2 = \{10, 12, 20, 30, 11, 25\}$ $m_1 = 3, m_2 = 18$

$K_1 = \{2, 3, 4, 10\}, k_2 = \{12, 20, 30, 11, 25\}$ $m_1 = 4.75, m_2 = 19.6$

$K_1 = \{2, 3, 4, 10, 11, 12\}, k_2 = \{20, 30, 25\}$ $m_1 = 7, m_2 = 25$

Final clusters

$K_1 = \{2, 3, 4, 10, 11, 12\}, k_2 = \{20, 30, 25\}$

Example 5.6.2 : Given {10, 4, 2, 12, 3, 20, 30, 11, 25, 31} Assume number of clusters i.e. K = 2

Solution :

Randomly assign alternative values to each cluster

$K_1 = \{10, 2, 3, 30, 25\}, k_2 = \{4, 12, 20, 11, 31\}$ $m_1 = 14, m_2 = 15.6$

Re assign

$K_1 = \{2, 3, 4, 10, 11, 12\}, k_2 = \{20, 25, 30, 31\}$ $m_1 = 7, m_2 = 26.5$

Re assign

$K_1 = \{2, 3, 4, 10, 11, 12\}, k_2 = \{20, 25, 30, 31\}$ $m_1 = 7, m_2 = 26.5$

Final clusters

$K_1 = \{2, 3, 4, 10, 11, 12\}, k_2 = \{20, 25, 30, 31\}$

Example 5.6.3 : Let's assume that we have 4 types of items and each item has 2 attributes or features. We need to group these items in to k = 2 groups of items based on the two features.

Object	Attribute 1(x) Number of parts	Attribute 2(y) Colour code
Item 1	1	1
Item 2	2	1
Item 3	4	3
Item 4	5	4

Module 5

Solution :

Initial value of centroid

Suppose we use item 1 and 2 as the first centroids, $c_1 = (1, 1)$ and $c_2 = (2, 1)$

The distance of item 1 = (1, 1) to $c_1 = (1, 1)$ and with $c_2 = (2, 1)$ is calculated as,

$$D = \sqrt{(1-1)^2 + (1-1)^2} = 0$$

$$D = \sqrt{(1-2)^2 + (1-1)^2} = 1$$

The distance of item 2 = (2, 1) to c_1 = (1, 1) and with c_2 = (2, 1) is calculated as,

$$D = \sqrt{(2-1)^2 + (1-1)^2} = 1$$

$$D = \sqrt{(2-2)^2 + (1-1)^2} = 0$$

The distance of item 3 = (4, 3) to c_1 = (1, 1) and with c_2 = (2, 1) is calculated as,

$$D = \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

$$D = \sqrt{(4-2)^2 + (3-1)^2} = 2.83$$

The distance of item 4 = (5, 4) to c_1 = (1, 1) and with c_2 = (2, 1) is calculated as,

$$D = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$D = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

Objects-centroids distance

$$D^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \begin{matrix} c_1 = (1, 1) & \text{group 1} \\ c_2 = (2, 1) & \text{group 2} \end{matrix}$$

To find the cluster of each item we consider the minimum Euclidian distance between group1 and group 2.

From the above object centroid distance matrix we can see,

- Item 1 has minimum distance for group1, so we cluster item 1 in group 1.
- Item 2 has minimum distance for group 2, so we cluster item 2 in group 2.
- Item 3 has minimum distance for group 2, so we cluster item 3 in group 2.
- Item 4 has minimum distance for group 2, so we cluster item 4 in group 2.

Object Clustering

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

Iteration 1 : Determine centroids

C_1 has only one member thus c_1 = (1, 1) remains same.

$$C_2 = (2 + 4 + 5/3, 1 + 3 + 4/3) = (11/3, 8/3)$$

The distance of item 1 = (1, 1) to c_1 = (1, 1) and with c_2 = (11/3, 8/3) is calculated as,

$$D = \sqrt{(1-1)^2 + (1-1)^2} = 0$$

$$D = \sqrt{(1-11/3)^2 + (1-8/3)^2} = 3.41$$

The distance of item 2 = (2, 1) to c_1 = (1, 1) and with c_2 = (11/3, 8/3) is calculated as,

$$D = \sqrt{(2-1)^2 + (1-1)^2} = 1$$

$$D = \sqrt{(2-11/3)^2 + (1-8/3)^2} = 2.36$$

The distance of item 3 = (4, 3) to c_1 = (1, 1) and with c_2 = (2, 1) is calculated as,

$$D = \sqrt{(4-1)^2 + (3-1)^2} = 3.61$$

$$D = \sqrt{(4-11/3)^2 + (3-8/3)^2} = 0.47$$

The distance of item 4 = (5, 4) to c_1 = (1, 1) and with c_2 = (2, 1) is calculated as,

$$D = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$D = \sqrt{(5-11/3)^2 + (4-8/3)^2} = 1.89$$

Objects-centroids distance

$$D^2 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.41 & 2.36 & 0.47 & 1.89 \end{bmatrix} \begin{matrix} c_1 = (1, 1) & \text{group 1} \\ c_2 = \left(\frac{11}{3}, \frac{8}{3}\right) & \text{group 2} \end{matrix}$$

From the above object centroid distance matrix we can see,

- Item 1 has minimum distance for group1, so we cluster item 1 in group 1.
- Item 2 has minimum distance for group 1, so we cluster item 2 in group 1.
- Item 3 has minimum distance for group 2, so we cluster item 3 in group 2.
- Item 4 has minimum distance for group 2, so we cluster item 4 in group 2.

Object Clustering

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Iteration 2 : Determine centroids

$$C_1 = (1 + 2/2, 1 + 1/2) = (3/2, 1)$$

$$C_2 = (4 + 5/2, 3 + 4/2) = (9/2, 7/2)$$

The distance of item 1 = (1, 1) to c_1 = (3/2, 1) and with c_2 = (9/2, 7/2) is calculated as,

$$D = \sqrt{(1-3/2)^2 + (1-1)^2} = 0.5$$

$$D = \sqrt{(1-9/2)^2 + (1-7/2)^2} = 4.3$$

The distance of item 2 = (2, 1) to c_1 = (3/2, 1) and with c_2 = (9/2, 7/2) is calculated as,

$$D = \sqrt{(2-3/2)^2 + (1-1)^2} = 0.5$$

$$D = \sqrt{(2-9/2)^2 + (1-7/2)^2} = 3.54$$

The distance of item 3 = (4, 3) to c_1 = (3/2, 1) and with c_2 = (9/2, 7/2) is calculated as,

$$D = \sqrt{(4-3/2)^2 + (3-1)^2} = 3.20$$

$$D = \sqrt{(4-9/2)^2 + (3-7/2)^2} = 0.71$$

The distance of item 4 = (5, 4) to c_1 = (3/2, 1) and with c_2 = (9/2, 7/2) is calculated as,

$$D = \sqrt{(5-3/2)^2 + (4-1)^2} = 4.61$$

Module 5

$$D = \sqrt{(5-9/2)^2 + (4-7/2)^2} = 0.71$$

Objects-centroids distance

$$D^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.3 & 3.54 & 0.71 & 0.71 \end{bmatrix} \begin{matrix} c_1 = \left(\frac{3}{2}, 1\right) & \text{group 1} \\ c_2 = \left(\frac{9}{2}, \frac{7}{2}\right) & \text{group 2} \end{matrix}$$

From the above object centroid distance matrix we can see,

- Item 1 has minimum distance for group1, so we cluster item 1 in group 1.
- Item 2 has minimum distance for group 1, so we cluster item 2 in group 1.
- Item 3 has minimum distance for group 2, so we cluster item 3 in group 2.
- Item 4 has minimum distance for group 2, so we cluster item 4 in group 2.

Object Clustering

$$G^2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

$G^2 = G^1$, Objects does not move from group any more. So, the final clusters are as follows:

- Item 1 and 2 are clustered in group 1
- Item 3 and 4 are clustered in group 2

Example 5.6.4 : Suppose we have eight data points and each data point has 2 features. Cluster the data points into 3 clusters using k-means algorithm.

Data points	Attribute 1(x)	Attribute 2(y)
1	2	10
2	2	5
3	8	4
4	5	8
5	7	5
6	6	4
7	1	2
8	4	9

☑ **Solution :**

Initial value of centroid

Suppose we use data points 1, 4 and 7 as the first centroids, $c_1 = (2, 10)$, $c_2 = (5, 8)$ and $c_3 = (1, 2)$

The distance of data point 1 = (2, 10) to $c_1 = (2, 10)$, $c_2 = (5, 8)$ and with $c_3 = (1, 2)$ is,

$$D = \sqrt{(2-2)^2 + (10-10)^2} = 0$$

$$D = \sqrt{(2-5)^2 + (10-8)^2} = 3.61$$

$$D = \sqrt{(2-1)^2 + (10-2)^2} = 8.06$$

The distance of data point 1 = (2, 5) to c_1 = (2, 10), c_2 = (5, 8) and with c_3 = (1, 2) is,

$$D = \sqrt{(2-2)^2 + (5-10)^2} = 5$$

$$D = \sqrt{(2-5)^2 + (5-8)^2} = 4.24$$

$$D = \sqrt{(2-1)^2 + (5-2)^2} = 3.16$$

The distance of data point 1 = (8, 4) to c_1 = (2, 10), c_2 = (5, 8) and with c_3 = (1, 2) is,

$$D = \sqrt{(8-2)^2 + (4-10)^2} = 8.48$$

$$D = \sqrt{(8-5)^2 + (4-8)^2} = 5$$

$$D = \sqrt{(8-1)^2 + (4-2)^2} = 7.28$$

The distance of data point 1 = (5, 8) to c_1 = (2, 10), c_2 = (5, 8) and with c_3 = (1, 2) is,

$$D = \sqrt{(5-2)^2 + (8-10)^2} = 3.61$$

$$D = \sqrt{(5-5)^2 + (8-8)^2} = 0$$

$$D = \sqrt{(5-1)^2 + (8-2)^2} = 7.21$$

The distance of data point 1 = (7, 5) to c_1 = (2, 10), c_2 = (5, 8) and with c_3 = (1, 2) is,

$$D = \sqrt{(7-2)^2 + (5-10)^2} = 7.07$$

$$D = \sqrt{(7-5)^2 + (5-8)^2} = 3.61$$

$$D = \sqrt{(7-1)^2 + (5-2)^2} = 6.71$$

The distance of data point 1 = (6, 4) to c_1 = (2, 10), c_2 = (5, 8) and with c_3 = (1, 2) is,

$$D = \sqrt{(6-2)^2 + (4-10)^2} = 7.21$$

$$D = \sqrt{(6-5)^2 + (4-8)^2} = 4.12$$

$$D = \sqrt{(6-1)^2 + (4-2)^2} = 5.39$$

The distance of data point 1 = (1, 2) to c_1 = (2, 10), c_2 = (5, 8) and with c_3 = (1, 2) is,

$$D = \sqrt{(1-2)^2 + (2-10)^2} = 8.06$$

$$D = \sqrt{(1-5)^2 + (2-8)^2} = 7.21$$

$$D = \sqrt{(1-1)^2 + (2-2)^2} = 0$$

The distance of data point 1 = (4, 9) to c_1 = (2, 10), c_2 = (5, 8) and with c_3 = (1, 2) is,

$$D = \sqrt{(4-2)^2 + (9-10)^2} = 2.24$$

$$D = \sqrt{(4-5)^2 + (9-8)^2} = 1.4$$

$$D = \sqrt{(4-1)^2 + (9-2)^2} = 7.62$$

Objects-centroids distance

$$D^0 = \begin{bmatrix} 0 & 5 & 8.48 & 3.61 & 7.07 & 7.21 & 8.06 & 2.24 \\ 3.61 & 4.24 & 5 & 0 & 3.61 & 4.12 & 7.21 & 1.4 \\ 8.06 & 3.16 & 7.28 & 7.21 & 6.71 & 5.39 & 0 & 7.62 \end{bmatrix} \begin{matrix} c_1 = (2, 10) & \text{group 1} \\ c_2 = (5, 8) & \text{group 2} \\ c_3 = (1, 2) & \text{group 3} \end{matrix}$$

Module 5

From the above object centroid distance matrix we can see,

- Data point 1 has minimum distance for group1, so we cluster data point 1 in group 1.
- Data point 2 has minimum distance for group3, so we cluster data point 2 in group 3.
- Data point 3 has minimum distance for group 2, so we cluster data point 3 in group 2.
- Data point 4 has minimum distance for group 2, so we cluster data point 4 in group 2.
- Data point 5 has minimum distance for group 2, so we cluster data point 5 in group 2.
- Data point 6 has minimum distance for group 2, so we cluster data point 6 in group 2.
- Data point 7 has minimum distance for group 3, so we cluster data point 7 in group 3.
- Data point 8 has minimum distance for group 2, so we cluster data point 8 in group 2.

Object Clustering

$$G^0 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Iteration 1 : Determine centroids

C1has only one member thus $c_1 = (2, 10)$ remains same.

$$C_2 = (8 + 5 + 7 + 6 + 4/5, 4 + 8 + 5 + 4 + 9/5) = (6, 6)$$

$$C_3 = (2 + 1/2, 5 + 2/2) = (1.5, 3.5)$$

Objects-centroids distance

$$D^1 = \begin{bmatrix} 0 & 5 & 8.48 & 3.61 & 7.07 & 7.21 & 8.06 & 2.24 \\ 5.66 & 4.12 & 2.83 & 2.24 & 1.41 & 2 & 6.40 & 3.16 \\ 6.52 & 1.58 & 6.25 & 5.7 & 5.7 & 4.52 & 1.58 & 6.04 \end{bmatrix} \begin{matrix} c_1 = (2, 10) & \text{group 1} \\ c_2 = (6, 6) & \text{group 2} \\ c_3 = (1.5, 3.5) & \text{group 3} \end{matrix}$$

Object Clustering

$$G^1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Iteration 2 : Determine centroids

$$C_1 = (2 + 4/2, 10 + 9/2) = (3, 9.5)$$

$$C_2 = (8 + 5 + 7 + 6/4, 4 + 8 + 5 + 4/4) = (6.5, 5.25)$$

$$C_3 = (2 + 1/2, 5 + 2/2) = (1.5, 3.5)$$

$$D^2 = \begin{bmatrix} 1.12 & 2.35 & 7.43 & 2.5 & 6.02 & 6.26 & 7.76 & 1.12 \\ 6.54 & 4.51 & 1.95 & 3.13 & 0.56 & 1.35 & 6.38 & 7.68 \\ 6.52 & 1.58 & 6.52 & 5.7 & 5.7 & 4.52 & 1.58 & 6.04 \end{bmatrix} \begin{matrix} c_1 = (3, 9.5) & \text{group 1} \\ c_2 = (6.5, 5.25) & \text{group 2} \\ c_3 = (1.5, 3.5) & \text{group 3} \end{matrix}$$

Object Clustering

$$G^2 = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Iteration 3 : Determine centroids

$C_1 = (2 + 5 + 4/3, 10 + 9 + 8/3) = (3.67, 9)$

$C_2 = (8 + 7 + 6/3, 4 + 5 + 4/3) = (7, 4.33)$

$C_3 = (2 + 1/2, 5 + 2/2) = (1.5, 3.5)$

$$D^2 = \begin{bmatrix} 1.95 & 4.33 & 6.61 & 1.66 & 5.2 & 5.52 & 7.49 & 0.33 \\ 6.01 & 5.04 & 1.05 & 4.17 & 0.67 & 1.05 & 6.44 & 5.55 \\ 6.52 & 1.58 & 6.52 & 5.7 & 5.7 & 4.52 & 1.58 & 6.04 \end{bmatrix} \begin{matrix} c_1 = (3.67, 9) & \text{group 1} \\ c_2 = (7, 4.33) & \text{group 2} \\ c_3 = (1.5, 3.5) & \text{group 3} \end{matrix}$$

Object Clustering

$$G^3 = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$G^3 = G^2$. Objects does not move from group any more. So, the final clusters are as follows:

- Data points 1, 4 and 8 are clustered in group 1
- Data points 3, 5 and 6 are clustered in group 2
- Data points 2 and 7 are clustered in group 3

Example 5.6.5 **MU - May 15, 10 Marks**

Apply K-means algorithm on given data for k = 3. Use c_1 (2), c_2 (16) and c_3 (38) as initial cluster centres.

Data : 2, 4, 6, 3, 31, 12, 15, 16, 38, 35, 14, 21, 23, 25, 30

Solution :

$c_1 = 2, c_2 = 16, c_3 = 38$

The numbers which are close to mean are grouped into respective clusters.

$k_1 = \{2, 4, 6, 3\}, k_2 = \{12, 15, 16, 14, 21, 23, 25\}, k_3 = \{31, 35, 30\}$

Again calculate new mean for new cluster group.

$c_1 = 3.75, c_2 = 18, c_3 = 32$

New clusters

$K_1 = \{2, 4, 6, 3\}, k_2 = \{12, 15, 16, 14, 21, 23, 25\}, k_3 = \{31, 35, 30\}$ $c_1 = 3.75, c_2 = 18, c_3 = 32$

Clusters remains unchanged

Final clusters

$K_1 = \{2, 4, 6, 3\}, k_2 = \{12, 15, 16, 14, 21, 23, 25\}, k_3 = \{31, 35, 30\}$

Modul 5

UExample 5.6.6 **MU - May 16, 10 Marks**

Apply K-means algorithm on given data for k = 3. Use c_1 (2), c_2 (16) and c_3 (38) as initial cluster centres.

Data : 2, 4, 6, 3, 31, 12, 15, 16, 38, 35, 14, 21, 23, 25, 30

☑ **Solution :**

$c_1 = 2, c_2 = 16, c_3 = 38$

The numbers which are close to mean are grouped into respective clusters.

$k_1 = \{2, 4, 6, 3\}, k_2 = \{12, 15, 16, 14, 21, 23, 25\}, k_3 = \{31, 35, 30\}$

Again calculate new mean for new cluster group.

$c_1 = 3.75, c_2 = 18, c_3 = 32$

New clusters

$K_1 = \{2, 4, 6, 3\}, k_2 = \{12, 15, 16, 14, 21, 23, 25\}, k_3 = \{31, 35, 30\}$ $c_1 = 3.75, c_2 = 18, c_3 = 32$

Clusters remains unchanged

Final clusters

$K_1 = \{2, 4, 6, 3\}, k_2 = \{12, 15, 16, 14, 21, 23, 25\}, k_3 = \{31, 35, 30\}$

5.6.2 Hierarchical Clustering

Agglomerative Hierarchical Clustering

- In agglomerative clustering initially each data point is considered as a single cluster. In the next step, pairs of clusters are merged or agglomerated. This step is repeated until all clusters have been merged in to a single cluster. At the end a single cluster remains that contains all the data points.
- Hierarchical clustering algorithms works in top-down manner or bottom-up manner. Hierarchical clustering is known as Hierarchical agglomerative clustering.
- In agglomerative, clustering is represented as a dendogram as in Fig. 5.6.1 where each merge is represented by a horizontal line

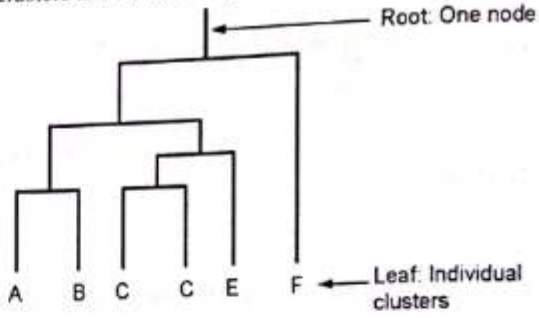


Fig. 5.6.1 : Dendogram

- A clustering of the data objects is obtained by cutting the dendogram at the desired level, then each connected forms a cluster.
- The basic steps of Agglomerative hierarchical clustering are as follows :
  1. Compute the proximity matrix (distance matrix)
  2. Assume each data point as a cluster.
  3. Repeat
  4. Merge the two nearest clusters.
  5. Update the proximity matrix
  6. Until only a single cluster remains
- In Agglomerative hierarchical clustering proximity matrix is symmetric i.e., the number on lower half will be same as the numbers on top half.

Different approaches to defining the distance between clusters distinguish the different algorithm's i.e., Single linkage, Complete linkage and Average linkage clusters.

In single linkage, the distance between two clusters is considered to be equal to shortest distance from any member of one cluster to any member of other cluster.

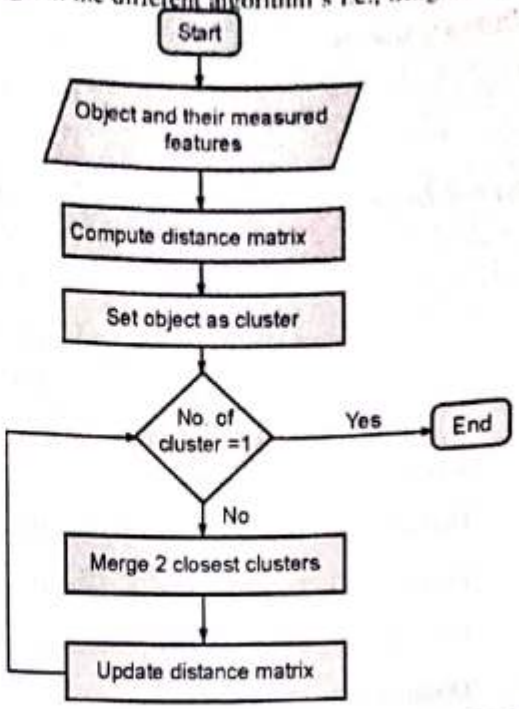
$p(r, s)$ = Min {d (i, j), object i → cluster r and object j→ cluster s

In complete linkage, the distance between two clusters is considered to be equal to greatest distance from any member of one cluster to any member of other cluster.

$p(r, s)$ = Max {d (i, j), object i→ cluster r and object j→ cluster s

In average linkage, we consider the distance between any two clusters A and B is taken to be equal to average of all distances between pairs of object i in A and j in B.ie., mean distance between elements of each other.

$p(r, s)$ = Mean {d (i, j), object i→ cluster r and object j→ cluster s



Example 5.6.7 : The table below shows the six data points. Use all link methods to find clusters. Use Euclidian distance measure.

	X	y
D_1	0.4	0.53
D_2	0.22	0.38
D_3	0.35	0.32
D_4	0.26	0.19
D_5	0.08	0.41
D_6	0.45	0.30

Solution :

First we will solve using single linkage

The distance of data point $D_1 = (0.4, 0.53)$ to $D_2 = (0.22, 0.38)$ is,

$$D = \sqrt{(0.4 - 0.22)^2 + (0.53 - 0.38)^2} = 0.24$$

The distance of data point $D_1 = (0.4, 0.53)$ to $D_3 = (0.35, 0.32)$ is,

$$D = \sqrt{(0.4 - 0.35)^2 + (0.53 - 0.32)^2} = 0.22$$

The distance of data point $D_1 = (0.4, 0.53)$ to $D_4 = (0.26, 0.19)$ is,

$$D = \sqrt{(0.4 - 0.26)^2 + (0.53 - 0.19)^2} = 0.37$$

The distance of data point $D_1 = (0.4, 0.53)$ to $D_5 = (0.08, 0.41)$ is,

$$D = \sqrt{(0.4 - 0.08)^2 + (0.53 - 0.41)^2} = 0.34$$

The distance of data point $D_1 = (0.4, 0.53)$ to $D_6 = (0.45, 0.30)$ is,

$$D = \sqrt{(0.4 - 0.45)^2 + (0.53 - 0.30)^2} = 0.23$$

Module 5

Similarly we will calculate all distances.

Distance matrix

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	0					
D_2	0.24	0				
D_3	0.22	0.15	0			
D_4	0.37	0.20	0.15	0		
D_5	0.34	0.14	0.28	0.29	0	
D_6	0.23	0.25	0.11	0.22	0.39	0

0.11 is smallest. D_3 and D_6 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance $((D_3, D_6), D_1)$ = min (distance (D_3, D_1), distance (D_6, D_1)) = min (0.22, 0.23) = 0.22

Distance $((D_3, D_6), D_2)$ = min (distance (D_3, D_2), distance (D_6, D_2)) = min (0.15, 0.25) = 0.15

Distance $((D_3, D_6), D_4)$ = min (distance (D_3, D_4), distance (D_6, D_4)) = min (0.15, 0.22) = 0.15

Distance $((D_3, D_6), D_5)$ = min (distance (D_3, D_5), distance (D_6, D_5)) = min (0.28, 0.39) = 0.28

Similarly we will calculate all distances.

Distance matrix

	D_1	D_2	(D_3, D_6)	D_4	D_5
D_1	0				
D_2	0.24	0			
(D_3, D_6)	0.22	0.15	0		
D_4	0.37	0.20	0.15	0	
D_5	0.34	0.14	0.28	0.29	0

0.14 is smallest. D_2 and D_5 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance $((D_3, D_6), (D_2, D_5))$ = min (distance (D_3, D_2), distance (D_6, D_2), distance (D_3, D_5), distance (D_6, D_6))

= min (0.15, 0.25, 0.28, 0.29) = 0.15

Similarly, we will calculate all distances.

Distance matrix

	D_1	(D_2, D_5)	(D_3, D_6)	D_4
D_1	0			
(D_2, D_5)	0.24	0		
(D_3, D_6)	0.22	0.15	0	
D_4	0.37	0.20	0.15	0

0.15 is smallest. (D_2, D_5) and (D_3, D_6) as well as D_4 and (D_3, D_6) have smallest distance. We can pick either one.

Distance matrix

	D_1	(D_2, D_5, D_3, D_6)	D_4
D_1	0		
(D_2, D_5, D_3, D_6)	0.22	0	
D_4	0.37	0.15	0

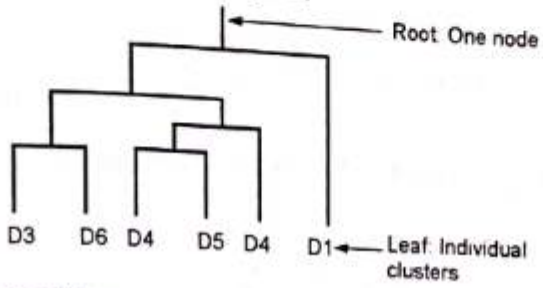
0.15 is smallest. (D_2, D_5, D_3, D_6) and D_4 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

	D_1	$(D_2, D_5, D_3, D_6, D_4)$
D_1	0	
$(D_2, D_5, D_3, D_6, D_4)$	0.22	0

Now a single cluster remains $(D_2, D_5, D_3, D_6, D_4, D_1)$

Next, we represent the final dendogram for single linkage as,



Now we will solve using complete linkage

Distance matrix

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	0					
D_2	0.24	0				
D_3	0.22	0.15	0			
D_4	0.37	0.20	0.15	0		
D_5	0.34	0.14	0.28	0.29	0	
D_6	0.23	0.25	0.11	0.22	0.39	0

0.11 is smallest. D_3 and D_6 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance $((D_3, D_6), D_1)$ = max (distance (D_3, D_1), distance (D_6, D_1)) = max (0.22, 0.23) = 0.23

Similarly, we will calculate all distances.

Module 5

Distance matrix

	D_1	D_2	(D_3, D_6)	D_4	D_5
D_1	0				
D_2	0.24	0			
(D_3, D_6)	0.23	0.25	0		
D_4	0.37	0.20	0.22	0	
D_5	0.34	0.14	0.39	0.29	0

0.14 is smallest. D_2 and D_5 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

	D_1	(D_2, D_5)	(D_3, D_6)	D_4
D_1	0			
(D_2, D_5)	0.34	0		
(D_3, D_6)	0.23	0.39	0	
D_4	0.37	0.29	0.22	0

0.22 is smallest. Here (D_3, D_6) and D_4 have smallest distance. So, we combine these two in one cluster and recalculate distance matrix.

Distance matrix

	D_1	(D_3, D_6, D_4)	(D_3, D_6, D_4)
D_1	0		
(D_2, D_5)	0.34	0	
(D_3, D_6, D_4)	0.37	0.39	0

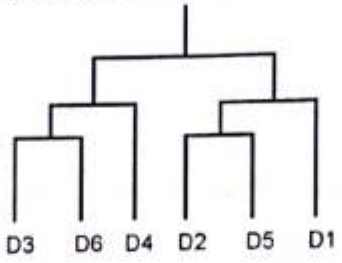
0.34 is smallest. (D_2, D_5) and D_1 have smallest distance so, we combine these two in one cluster and recalculate distance matrix.

Distance matrix

	(D_2, D_5, D_1)	(D_3, D_6, D_4)
(D_2, D_5, D_1)	**0**	**0**
(D_3, D_6, D_4)	**0.39**	**0**

Now a single cluster remains $(D_2, D_5, D_1, D_3, D_6, D_4)$

Next, we represent the final dendogram for complete linkage as.



Now we will solve using average linkage

Distance matrix

	D_1	D_2	D_3	D_4	D_5	D_6
D_1	0					
D_2	0.24	0				
D_3	0.22	0.15	0			
D_4	0.37	0.20	0.15	0		
D_5	0.34	0.14	0.28	0.29	0	
D_6	0.23	0.25	0.11	0.22	0.39	0

0.11 is smallest. D_3 and D_6 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance $((D_3, D_6), D_1)$ = 1/2 (distance (D_3, D_1) + distance (D_6, D_1)) = 1/2 (0.22 + 0.23) = 0.23

Similarly, we will calculate all distances.

Distance matrix

D_1	0				
D_2	0.24	0			
(D_3, D_6)	0.23	0.2	0		
D_4	0.37	0.20	0.19	0	
D_5	0.34	0.14	0.34	0.29	0
	D_1	D_2	(D_3, D_6)	D_4	D_5

0.14 is smallest. D_2 and D_5 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

D_1	0			
(D_2, D_5)	0.29	0		
(D_3, D_6)	0.22	0.27	0	
D_4	0.37	0.22	0.15	0
	D_1	(D_2, D_5)	(D_3, D_6)	D_4

(D_3, D_6) and D_4 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

D_1	0		
(D_2, D_5)	0.24	0	
(D_3, D_6, D_4)	0.27	0.26	0
	D_1	(D_2, D_5)	(D_3, D_6, D_4)

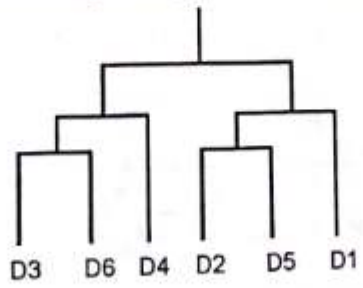
0.24 is smallest. (D_2, D_5) and D_1 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

(D_2, D_5, D_1)	**0**	**0**
(D_3, D_6, D_4)	**0.26**	**0**
	(D_2, D_5, D_1)	(D_3, D_6, D_4)

Now a single cluster remains $(D_2, D_5, D_1, D_3, D_6, D_4)$

Next, we represent the final dendogram for average linkage as,



Module 5

Example 5.6.8 : Apply single linkage, complete linkage and average linkage on the following distance matrix and draw dendogram.

☑ **Solution :**

First we will solve using single linkage

Distance matrix

	P_1	P_2	P_3	P_4	P_5
P_1	0				
P_2	2	0			
P_3	6	3	0		
P_4	10	9	7	0	
P_5	9	8	5	4	0

2 is smallest. P_1 and P_2 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance $((P_1, P_2), P_3)$ = min (distance (P, P_3), distance (P_2, P_3)) = min (6, 3) = 3

Similarly, we will calculate all distances.

Distance matrix

	(P_1, P_2)	P_3	P_4	P_5
(P_1, P_2)	0			
P_3	3	0		
P_4	9	7	0	
P_5	8	5	4	0

3 is smallest. (P_1, P_2) and P_3 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance $((P_1, P_2, P_3), P_4)$ = min (distance (P_1, P_4), distance (P_2, P_4), distance (P_3, P_4)) = min (9, 7) = 7

Similarly, we will calculate all distances.

Distance matrix

	(P_1, P_2, P_3)	P_4	P_5
(P_1, P_2, P_3)	0		
P_4	7	0	
P_5	5	4	0

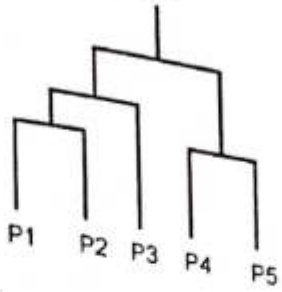
4 is smallest. P_4 and P_5 have smallest distance.

Distance matrix

	(P_1, P_2, P_3)	(P_4, P_5)
(P_1, P_2, P_3)	0	
(P_4, P_5)	5	0

Now a single cluster remains (P_1, P_2, P_3, P_4, P_5)

Next, we represent the final dendogram for single linkage as,



Now we will solve using complete linkage

Distance matrix

P_1	0				
P_2	2	0			
P_3	6	3	0		
P_4	10	9	7	0	
P_5	9	8	5	4	0
	P_1	P_2	P_3	P_4	P_5

2 is smallest. P_1 and P_2 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance $((P_1, P_2), P_3)$ = max (distance (P_1, P_3), distance (P_2, P_3)) = max (6, 3) = 6

Similarly, we will calculate all distances.

Distance matrix

(P_1, P_2)	0			
P_3	6	0		
P_4	10	7	0	
P_5	9	5	4	0
	(P_1, P_2)	P_3	P_4	P_5

4 is smallest. P_4 and P_5 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

(P_1, P_2)	0		
P_3	6	0	
(P_4, P_5)	10	7	0
	(P_1, P_2)	P_3	(P_4, P_5)

6 is smallest. (P_1, P_2) and P_3 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

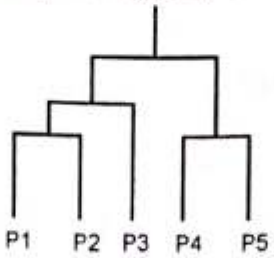
Module 5

Distance matrix

(P_1, P_2, P_3)	**0**	
(P_4, P_5)	**10**	**0**
	(P_1, P_2, P_3)	(P_4, P_5)

Now a single cluster remains $(P_1, P_2, P_3, P_4, P_5)$

Next, we represent the final dendogram for complete linkage as,



Now we will solve using average linkage

Distance matrix

P_1	0				
P_2	2	0			
P_3	6	3	0		
P_4	10	9	7	0	
P_5	9	8	5	4	0
	P_1	P_2	P_3	P_4	P_5

2 is smallest. P_1 and P_2 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance $((P_1, P_2), P_3)$ = 1/2 (distance (P_1, P_3), distance (P_2, P_3)) = 1/2 (6, 3) = 4.5

Similarly, we will calculate all distances.

Distance matrix

(P_1, P_2)	**0**			
P_3	**4.5**	**0**		
P_4	**9.5**	**7**	**0**	
P_5	**8.5**	**5**	**4**	**0**
	(P_1, P_2)	P_3	P_4	P_5

4 is smallest. P_4 and P_5 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

(P_1, P_2)	**0**		
P_3	**4.5**	**0**	
(P_4, P_5)	**9**	**6**	**0**
	(P_1, P_2)	P_3	(P_4, P_5)

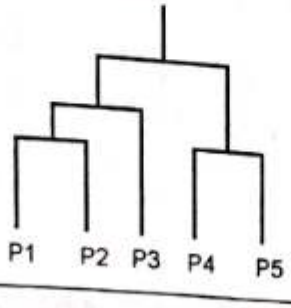
4.5 is smallest. (P_1, P_2) and P_3 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

(P_1, P_2, P_3)	0	
(P_4, P_5)	8	0
	(P_1, P_2, P_3)	(P_4, P_5)

Now a single cluster remains $(P_1, P_2, P_3, P_4, P_5)$

Next, we represent the final dendogram for average linkage as,



Example 5.6.9 **MU - May 16, 10 Marks**

Apply Agglomerative clustering algorithm on given data and draw dendrogram. Show three clusters with its allocated points. Use single link method.

	A	B	C	D	E	F
A	0	$\sqrt{2}$	$\sqrt{10}$	$\sqrt{17}$	$\sqrt{5}$	$\sqrt{20}$
B	$\sqrt{2}$	0	$\sqrt{8}$	3	1	$\sqrt{18}$
C	$\sqrt{10}$	$\sqrt{8}$	0	$\sqrt{5}$	$\sqrt{5}$	2
D	$\sqrt{17}$	1	$\sqrt{5}$	0	2	3
E	$\sqrt{5}$	1	$\sqrt{5}$	2	0	$\sqrt{13}$
F	$\sqrt{20}$	$\sqrt{18}$	2	3	$\sqrt{13}$	0

Solution :

Distance matrix

A	0					
B	1.414	0				
C	3.162	2.828	0			
D	4.123	1	2.236	0		
E	2.236	1	2.236	2	0	
F	4.472	4.242	2	3	3.6	0
	A	B	C	D	E	F

1 is smallest. B, D and B, E have smallest distance. We can select anyone. So, we combine B, D in one cluster and recalculate distance matrix using single linkage.

Modu 5

Distance matrix

	A	B,D	C	E	F
A	0				
B,D	1.414	0			
C	3.162	2.236	0		
E	2.26	1	2.236	0	
F	4.472	3	2	3.6	0

1 is smallest. B, D and E have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

	A	B,D,E	C	F
A	0			
B,D,E	1.414	0		
C	3.162	1	0	
F	4.472	3	2	0

1 is smallest. B, D, E and C are combined together.

Distance matrix

	A	B,D,E,C	F
A	0		
B,D,E,C	1.414	0	
F	4.472	2	0

In the questions three clusters are asked with their allocated points. Three clusters are A, (B, D, E, C) and F.

UExample 5.6.10 **MU - May 17, 10 Marks**

For the given set of points identify clusters using complete link and average link using Agglomerative clustering.

	A	B
P_1	1	1
P_2	1.5	1.5
P_3	5	5
P_4	3	4
P_5	4	4
P_6	3	3.5

☑ **Solution :**

First we will solve using complete linkage

Distance matrix

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.707	0				
P3	5.656	4.949	0			
P4	3.605	2.915	2.236	0		
P5	4.242	3.535	1.414	1	0	
P6	5.201	2.5	1.802	0.5	1.118	0

0.5 is smallest. P4 and P6 have smallest distance. We can select anyone. So, we combine this in one cluster and recalculate distance matrix using complete linkage.

Distance matrix

	P1	P2	P3	P4,P6	P5
P1	0				
P2	0.707	0			
P3	5.656	4.949	0		
P4,P6	5.201	2.5	1.802	0	
P5	4.242	3.535	1.414	1.118	0

0.707 is smallest. P1 and P2 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

	P1,P2	P3	P4,P6	P5
P1,P2	0			
P3	5.656	0		
P4,P6	5.201	2.236	0	
P5	4.242	1.414	1.118	0

1.118 is smallest. P4, P6 and P5 are combined together.

Distance matrix

	P1,P2	P3	P4,P5,P6
P1,P2	0		
P3	5.656	0	
P4,P5,P6	5.201	2.236	0

2.236 is smallest. P4, P5, P6 and P3 are combined together.

	P1,P2	P3,P4,P5,P6
P1,P2	0	
P3,P4,P5,P6	5.656	0

Next we will combine all clusters in a single cluster.

Now we will solve using average linkage.

Module 5

Distance matrix

P1	0					
P2	0.707	0				
P3	5.656	4.949	0			
P4	3.605	2.915	2.236	0		
P5	4.242	3.535	1.414	1	0	
P6	5.201	2.5	1.802	0.5	1.118	0
	P1	P2	P3	P4	P5	P6

0.5 is smallest. P4 and P6 have smallest distance. We can select anyone .So, we combine this in one cluster and recalculate distance matrix using complete linkage.

Distance matrix

P1	0				
P2	0.707	0			
P3	5.656	4.949	0		
P4,P6	4.403	2.707	2.019	0	
P5	4.242	3.535	1.414	1.059	0
	P1	P2	P3	P4,P6	P5

0.707 is smallest. P1 and P2 have smallest distance. So, we combine this two in one cluster and recalculate distance matrix.

Distance matrix

P1,P2	0			
P3	5.302	0		
P4,P6	3.55	2.019	0	
P5	3.888	1.414	1.059	0
	P1,P2	P3	P4,P6	P5

1.059 is smallest. P4, P6 and P5 are combined together.

Distance matrix

P1,P2	0		
P3	5.302	0	
P4,P5,P6	3.66	1.817	0
	P1,P2	P3	P4,P5,P6

2.236 is smallest. P4,P5,P6 and P3 are combined together.

P1,P2	0	
P3,P4,P5,P6	4.07	0
	P1,P2	P3,P4,P5,P6

Next we will combine all clusters in a single cluster.

5.6.3 Expectation - Maximization Algorithm

The Expectation maximization algorithm is an unsupervised clustering method. This method is based on the concept of mixture models. The method is executed in an iterative way. In the presence of missing data the probability distribution parameters are calculated that has the maximum likelihood of its attributes.

We give the input to the EM algorithm as the data set D, the total number of clusters/models K, the accepted error to converge and the maximum number of iterations. In each iteration , first Expectation step is executed. In Expectation step, the probability of each point belonging to each model is calculated.

Next Maximization step is executed, in which the parameters of the probability distribution of each model are again calculated.

The stopping criteria of the algorithm is, when the distribution parameters converge or reach the maximum number of iterations. Convergence is guaranteed as the algorithm increases the likelihood at each iteration until it reaches the local maximum.

The EM algorithm is executed in the following way :

- **Initialisation-step :** Model's parameters are assigned to random values.
- **Expectation-step :** Assign points to the model that fits each one best
- **Maximization-step :** Update the parameters of the model using the points assigned in the earlier step
- Iterate until parameter values converge
- Consider a set of starting parameters given a set of incomplete (observed) data. Assume observed data come from a specific model.
- Use these to "estimate" the missing data. Formulate some parameters for that model. Use this to guess the missing value (E step).
- Use "Complete" data to update parameters. From missing data and observed data find the most likely parameters (M step).
- Repeat step 2 and 3 until convergence.

Now let's understand EM algorithm with the help of example. In example 1 we will see what a use of EM algorithm is and in the example 2 we will see how to use EM algorithm.

Example 5.6.11 : Suppose Coin A and B is used for tossing. Each coin is tossed 10 times. Following table shows the observation sequence of getting H and T when coin A and B is used. What is the probability of getting H if coin A and B is used?

	Number of Toss									
Coin Used for Toss	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
B	H	H	H	H	H	T	T	T	T	T
A	H	H	H	H	H	H	H	H	T	T
A	H	H	H	H	H	H	H	H	H	T
B	H	H	H	H	T	T	T	T	T	T
A	H	H	H	H	H	H	H	T	T	T

Module 5

☑ **Solution :**

Let's first calculate number of H and T for each coin as follows

	Number of Toss											
Coin Used for Toss	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Coin A**	**Coin B**
B	H	H	H	H	H	T	T	T	T	T		5H, 5T
A	H	H	H	H	H	H	H	H	T	T	8H, 2T	
A	H	H	H	H	H	H	H	H	H	T	9H, 1T	
B	H	H	H	H	T	T	T	T	T	T		4H, 6T
A	H	H	H	H	H	H	H	T	T	T	7H, 3T	
Total											24H, 6T	9H, 11T

Probability of getting Head when coin A is used, $P_A = \frac{24}{24+6} = 0.8$

Probability of getting Head when coin B is used, $P_B = \frac{9}{9+11} = 0.45$

In this example if the coin state is hidden i.e. whether coin A or B is used is not given then how we will calculate the probability?

Example 5.6.12 : Suppose Coin A and B is used for tossing. Each coin is tossed 10 times. Following table shows the observation sequence of getting H and T for each round. But which coin is used for which round is not known. Then how to calculate the probability of getting H for coin A and B?

	Number of Toss									
Round number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0	H	H	H	H	H	T	T	T	T	T
1	H	H	H	H	H	H	H	H	T	T
2	H	H	H	H	H	H	H	H	H	T
3	H	H	H	H	T	T	T	T	T	T
4	H	H	H	H	H	H	H	T	T	T

☑ **Solution :**

When only observation sequence is known, but the state is not known (Coin A or B) then EM algorithm is used.

Now Let's solve this example using EM algorithm.

Assume $P_A = 0.6$ and $P_B = 0.5$

Round 0 : In round 0 there are 5 H, 5 T and total tosses are 10.

Now we will calculate probability of using A and B coin as,

$$A = (P_A)^H (1-P_A)^{N-H} = (0.6)^5 (1-0.6)^{10-5} = 0.00079626$$

$$B = (P_B)^H (1-P_B)^{N-H} = (0.5)^5 (1-0.5)^{10-5} = 0.0009765$$

Now we will apply Normalization,

$$A_N = \frac{A}{A+B} = 0.45$$

$$B_N = \frac{B}{A+B} = 0.55$$

Now we will calculate probability of getting H and T for Coin A and B for round 0 as,

$$A_H = A_N * \text{Number of H} = 0.45 * 5 = 2.25$$

$$A_T = A_N * \text{Number of T} = 0.45 * 5 = 2.25$$

$$B_H = B_N * \text{Number of H} = 0.55 * 5 = 2.75$$

$$A_H = B_N * \text{Number of T} = 0.55 * 5 = 2.75$$

Round 1 : In round 1 there are 8 H, 2 T and total tosses are 10.

Now we will calculate probability of using A and B coin as,

$$A = (P_A)^H (1 - P_A)^{N-H} = (0.6)^8 (1 - 0.6)^{10-8} = 0.002687$$

$$B = (P_B)^H (1 - P_B)^{N-H} = (0.5)^8 (1 - 0.5)^{10-8} = 0.0009755$$

Now we will apply Normalization,

$$A_N = \frac{A}{A+B} = 0.73$$

$$B_N = \frac{B}{A+B} = 0.27$$

Now we will calculate probability of getting H and T for Coin A and B for round 1 as,

$$A_H = A_N * \text{Number of H} = 0.73 * 8 = 5.84$$

$$A_T = A_N * \text{Number of T} = 0.73 * 2 = 1.46$$

$$B_H = B_N * \text{Number of H} = 0.27 * 8 = 2.16$$

$$A_H = B_N * \text{Number of T} = 0.27 * 2 = 0.54$$

Round 2 : In round 2 there are 9 H, 1 T and total tosses are 10.

Now we will calculate probability of using A and B coin as,

$$A = (P_A)^H (1 - P_A)^{N-H} = (0.6)^9 (1 - 0.6)^{10-9} = 0.004031$$

$$B = (P_B)^H (1 - P_B)^{N-H} = (0.5)^9 (1 - 0.5)^{10-9} = 0.0009765$$

Now we will apply Normalization,

$$A_N = \frac{A}{A+B} = 0.80$$

$$B_N = \frac{B}{A+B} = 0.20$$

Now we will calculate probability of getting H and T for Coin A and B for round 2 as,

$$A_H = A_N * \text{Number of H} = 0.80 * 9 = 7.2$$

$$A_T = A_N * \text{Number of T} = 0.80 * 1 = 0.8$$

$$B_H = B_N * \text{Number of H} = 0.2 * 9 = 1.8$$

$$A_H = B_N * \text{Number of T} = 0.2 * 1 = 0.2$$

Mod 5

Round 3 : In round 3 there are 4 H, 6 T and total tosses are 10.

Now we will calculate probability of using A and B coin as,

$$A = (P_A)^H (1 - P_A)^{N-H} = (0.6)^4 (1 - 0.6)^{10-4} = 0.0005308$$

$$B = (P_B)^H (1 - P_B)^{N-H} = (0.5)^4 (1 - 0.5)^{10-4} = 0.0009765$$

Now we will apply Normalization,

$$A_N = \frac{A}{A + B} = 0.35$$

$$B_N = \frac{B}{A + B} = 0.65$$

Now we will calculate probability of getting H and T for Coin A and B for round 3 as,

$$A_H = A_N * \text{Number of H} = 0.35 * 4 = 1.4$$

$$A_T = A_N * \text{Number of T} = 0.35 * 6 = 2.1$$

$$B_H = B_N * \text{Number of H} = 0.65 * 4 = 2.6$$

$$A_H = B_N * \text{Number of T} = 0.65 * 6 = 3.9$$

Round 4 : In round 4 there are 7 H, 3 T and total tosses are 10.

Now we will calculate probability of using A and B coin as,

$$A = (P_A)^H (1 - P_A)^{N-H} = (0.6)^7 (1 - 0.6)^{10-7} = 0.001792$$

$$B = (P_B)^H (1 - P_B)^{N-H} = (0.5)^7 (1 - 0.5)^{10-7} = 0.0009765$$

Now we will apply Normalization,

$$A_N = \frac{A}{A + B} = 0.65$$

$$B_N = \frac{B}{A + B} = 0.35$$

Now we will calculate probability of getting H and T for Coin A and B for round 4 as,

$$A_H = A_N * \text{Number of H} = 0.65 * 7 = 4.55$$

$$A_T = A_N * \text{Number of T} = 0.65 * 3 = 1.95$$

$$B_H = B_N * \text{Number of H} = 0.35 * 7 = 2.45$$

$$A_H = B_N * \text{Number of T} = 0.35 * 3 = 1.05$$

Now we will calculate P_A and P_B by summarizing the results of round 0 to round 4.

	Coin A		Coin B	
	A_H	A_T	B_H	B_T
0	2.25	2.25	2.75	2.75
1	5.84	1.46	2.16	0.54
2	7.2	0.8	1.8	0.2
3	1.4	2.1	2.6	3.9
4	4.55	1.95	2.45	1.05
Total	21.24	8.56	11.76	8.44

Probability of getting H when Coin A is used, $P_A = \frac{\Sigma H}{\Sigma H + \Sigma T} = 0.71$

Probability of getting H when Coin B is used, $P_B = \frac{\Sigma H}{\Sigma H + \Sigma T} = 0.58$

Now with these new values of P_A and P_B second iteration is applied. This process is repeated unless and until there will be no change in the values of P_A and P_B and then that will be the final probabilities.

5.6.4 Supervised Learning after Clustering

In unsupervised learning the target variable is not known. The relationship between input and output pattern is studied to update the network parameters.

Unsupervised learning gives the output as a label or a target which can be used by supervised learning. For example, based on the purchased history of customers, we can divide the customers in to different groups such as customers who purchases particular items frequently. Then this result can be used for cross selling purpose.

Let's see another example of document clustering. Suppose we have a set of documents that contains different news such as related to sports, fashion and education. This type of problem can be solved by grouping the documents together that contains the common keywords. After using the cluster learning, a number of clusters are created based on keyword similarity.

- Each cluster will contain similar documents terms. After creating the clusters, semantic features can be used to identify these clusters depend on supervised model like SVM to make accurate categorizations.
- In this our aim is to create more accurate categorizations because if we want to test a new document we should know if this document can be related to these categorizations or not.

5.6.5 Radial Basis Function

- Radial Basis Function network is a multiplayer feed forward network. In RBF 'n' number of input neurons and 'm' number of output neurons are present. Single hidden layer exists between input and output layer.
- Hypothetical connection is present between the input layer and the hidden layer, whereas weighted connections are present between hidden layer and output layer.
- Weights present in all interconnections are updated using the training algorithm.
- Radial basis function network, represented in Fig. 5.6.2, can be shown as a mapping : $\Re^T \rightarrow \Re^S$
- Let the input vector is $P \in \Re^T$ and prototype of the input vector is $C_i \in \Re^T$ $(1 \leq i \leq u)$ be the. The output of each RBF unit is as follows:

$$R_i(P) = R_i(\| P - C_i \|) \; i = 1 \ldots\ldots u \quad (5.6.1)$$

Where $(\| - \|)$ indicates the Euclidean norm on the input space.

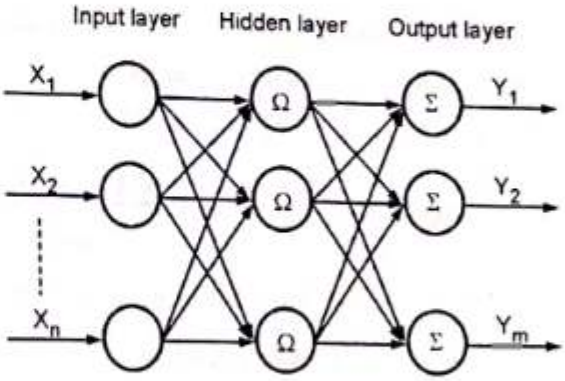


Fig. 5.6.2 : Radial Connected neural network

Module 5

- Generally, the Gaussian function is mostly used among all possible radial basis functions due to the reason that it is factorizable. Hence ...(5.6.2)

$$R_i(P) = \exp[-(\|P - C_i\|)^2 / \sigma_i^2]$$

Where σ_i^2 represents the width of the i^{th} RBF unit.

- RBF neural network's j^{th} output $y_i(P)$ is, ...(5.6.3)

$$y_i(P) = \sum_{i=1}^{u} R_i(P) \times w(j, i)$$

where, Bias of the j^{th} output is $w(j, 0)$, $R_0 = 1$, and weight or strength of the i^{th} receptive field to the j^{th} output is $w(j, i)$.

- Network complexity is reduced by not taking bias in to consideration in the following analysis. We can say from Equation (5.6.2) and Equation(5.6.3) that the outputs of radial basis function classifier are characterized by a linear discriminant function.
- They create linear decision boundaries in the output space. Consequently, separability of classes in the k-dimensional space strongly affects the performance of RBF classifier. This seperability is created by the nonlinear transformation carried out by the RBF units.
- Theorem on the separability of patterns also called as Cover's Theorem states that a complex pattern classification problem represented in a high-dimensional space nonlinearly is more likely to be linearly separable than in a low-dimensional space, the number of Gaussian nodes $u \geq r$, where r is the dimension of input space. If we increase the number of Gaussian units then it may result in poor generalization because of overfitting, especially, in the case of small training sets .
- RBF neural network classifies the input space into a number of subspaces which are in the form of hyperspheres. Clustering algorithms are widely used in RBF neural networks to solve the problems .In this clustering approaches category information about the patterns is not used hence they are also called as unsupervised learning algorithms.

5.6.6 RBF Learning Strategies

Learning strategies are followed in the design of RBF network.

1. Fixed Centers selected at random

Assume fixed radial basis function defining the activation function of hidden units. The location of centers may be chosen randomly from training data set.

RBF centered at t_i is defined as,

$$G(\|x - t_i\|^2) = \exp \frac{-m(\|x - t_i\|^2)}{d_{i(max)}^2}$$

Where m is number of centres and $d_{i(max)}$ represents maximum Euclidian distance between selected centers

Gaussian matrix is constructed as

$$G = (\psi_1(x), \psi_2(x), b)$$

Where $\psi_1(x)$ represents the hidden function corresponding to first selected center and $\psi_2(x)$ represents the hidden function corresponding to second selected center and b represents the bias.

The pseudo inverse matrix is constructed as

$$G^+ = (G^T G)^{-1} G^T$$

The weight matrix is calculated as

$$W = G^+ d$$

ple 5.6.13 : Design a XOR problem with given data set as (1,1) , (0,1) , (0,0) , (1,0) and also find weight vector.

Solution :

As we are designing the problem for XOR the input and desired responses are as follows.

Input pattern	Desired output
(1,1)	0
(0,1)	1
(0,0)	0
(1,0)	1

We will calculate the Euclidian distance between all the centres,

Euclidian distance between (1, 1) and (0, 1) centres

$$d = \sqrt{(1-0)^2 + (1-1)^2} = \sqrt{1}$$

Euclidian distance between (1, 1) and (0, 0) centres

$$d = \sqrt{(1-0)^2 + (1-1)^2} = \sqrt{2}$$

Euclidian distance between (1, 1) and (1, 0) centres

$$d = \sqrt{(1-1)^2 + (1-0)^2} = \sqrt{1}$$

Euclidian distance between (0, 1) and (0, 0) centres

$$d = \sqrt{(0-0)^2 + (1-0)^2} = \sqrt{1}$$

Euclidian distance between (0, 1) and (1, 0) centres

$$d = \sqrt{(0-1)^2 + (1-0)^2} = \sqrt{2}$$

Euclidian distance between (0, 0) and (1, 0) centres

$$d = \sqrt{(0-0)^2 + (1-0)^2} = \sqrt{1}$$

Maximum Euclidian distance is between (1, 1), (0, 0) centres and (0, 1), (1, 0)

We will select the two centres as $t_1 = (1, 1)$ and $t_2 = (0, 0)$

$$G\left(\|x - t_i\|^2\right) = \exp\frac{-m\left(\|x - t_i\|^2\right)}{d_{i(max)}^2}$$

$$G\left(\|x - t_i\|^2\right) = \exp\frac{-2\left(\|x - t_i\|^2\right)}{\sqrt{2}^2}$$

Module 5

$$G(\|x - t_i\|^2) = \exp(-\|x - t_i\|^2)$$

Now we will construct a G matrix

For the first column we use, $\exp(-(X - t_1)^2)$ and for the second column, $\exp(-(X - t_2)^2)$ and third column is for bias (1)

Now we will see how the values are calculated

For example for first row and first column X = (1, 1) and t_1 = (1, 1)

$= \exp(-((1-1)^2 + (1-1)^2)) = \exp(0) = 1$

For example for second row and first column X = (0, 1) and t_1 = (1, 1)

$= \exp(-((0-1)^2 + (1-1)^2)) = 0.3678$

For example for third row and first column X = (0, 0) and t_1 = (1, 1)

$= \exp(-((0-1)^2 + (0-1)^2)) = 0.1353$

For example for fourth row and first column X = (1, 0) and t_1 = (1, 1)

$= \exp(-((1-1)^2 + (0-1)^2)) = 0.3678$

For example for first row and second column X = (1, 1) and t_2 = (0, 0)

$= \exp(-((1-0)^2 + (1-0)^2)) = 0.1353$

For example for second row and second column X = (0, 1) and t_2 = (0, 0)

$= \exp(-((0-0)^2 + (1-0)^2)) = 0.3678$

For example for third row and second column X = (0, 0) and t_2 = (0, 0)

$= \exp(-((0-0)^2 + (0-0)^2)) = 1$

For example for fourth row and second column X = (1, 0) and t_2 = (0, 0)

$= \exp(-((1-0)^2 + (0-0)^2)) = 0.3678$

$$G = \begin{bmatrix} 1 & 0.1353 & 1 \\ 0.3678 & 0.3678 & 1 \\ 0.1353 & 1 & 1 \\ 0.3678 & 0.3678 & 1 \end{bmatrix}$$

Now by multiplying G^T by G we will get

$$G^T G = \begin{bmatrix} 1.2889 & 0.5412 & 1.8709 \\ 0.5412 & 1.2889 & 1.8709 \\ 1.8709 & 1.8709 & 4 \end{bmatrix}$$

Now we will find the inverse of this matrix by dividing the adjacency matrix by determinant

Determinant = $1.2889(1.2889 * 4 - 1.8709 * 1.8709) - 0.5412(0.5412 * 4 - 1.8709 * 1.8709) + 1.8709(0.5412 * 1.8709 - 1.2889 * 1.8709) = 0.2392$

Now to find the adjacency matrix we find the values row wise and we consider the alternate + and – sign for the above matrix

For first row first column = $+(1.2889 * 4 - 1.8709 * 1.8709) = 1.6553$

For first row second column = $-(0.5412 * 4 - 1.8709 * 1.8709) = 1.3355$

For first row third column = $+(0.5412 * 1.8709 - 1.2889 * 1.8709) = -1.3989$

Similarly all the values are calculated and we will get the following matrix

$$\text{Adjacency} = \begin{bmatrix} 1.6553 & 1.3355 & -1.3985 \\ 1.3355 & 1.6553 & -1.3989 \\ -1.3989 & 1.3989 & 1.3684 \end{bmatrix}$$

$$(G^T G)^{-1} = \text{adjacency/determinant}$$

$$G^+ = (G^T G)^{-1} G^T = \begin{bmatrix} 1.8273 & -1.2496 & 0.6727 & -1.2496 \\ 0.6727 & -1.2496 & 1.8292 & -1.2496 \\ 0.9202 & 1.4202 & -0.9202 & 1.42 \end{bmatrix}$$

$$W = G^+ d = \begin{bmatrix} 1.8273 & -1.2496 & 0.6727 & -1.2496 \\ 0.6727 & -1.2496 & 1.8292 & -1.2496 \\ 0.9202 & 1.4202 & -0.9202 & 1.42 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

$$W = \begin{bmatrix} -2.5018 \\ -2.5018 \\ 2.8404 \end{bmatrix}$$

Strict Interpolation with regularisation

In this all centers are selected. RBF centered at it is defined as,

$$\phi(\|X - t_i\|^2) = \exp[-(\|X - t_i\|)^2]$$

Interpolation matrix is constructed as

$$\phi = (\psi_1(x), \psi_2(x), \psi_3(x), \psi_4(x))$$

Where $\psi(x)$ represents the hidden function corresponding to centers.

The weight matrix is calculated as

$$W = \phi^{-1} d$$

Example 5.6.14 : Design a XOR problem with given data set as (1,1) , (0,1) , (0,0) , (1,0) and also find Weight vector and interpolation matrix.

Solution :

As we are designing the problem for XOR the input and desired responses are as follows :

Input pattern	Desired output
(1,1)	0
(0,1)	1
(0.0)	0
(1,0)	1

As we have to select all the centers $t_1 = (1, 1)$, $t_2 = (0, 1)$, $t_3 = (0,0)$ and $t_4 = (1, 0)$

The Interpolation matrix is given as

$$\phi = \begin{bmatrix} 1 & 0.3678 & 0.1353 & 0.3678 \\ 0.3678 & 1 & 0.3678 & 0.1353 \\ 0.1353 & 0.3678 & 1 & 0.3678 \\ 0.3678 & 0.1353 & 0.3678 & 1 \end{bmatrix}$$

Module 5

$$\phi^{-1} = \begin{bmatrix} -0.4918 & 0.13373 & -0.4918 & 0.1809 \\ 0.1809 & -0.4918 & 1.3373 & -0.4918 \\ 0.4918 & 0.1809 & -0.4918 & 1.3373 \end{bmatrix}$$

$$W = \phi^{-1} d = \begin{bmatrix} -0.9837 \\ 1.5182 \\ -0.9837 \\ 1.5182 \end{bmatrix}$$

5.7 UNIVERSITY QUESTIONS AND ANSWERS

May 2015

Q. 1 Describe the essential steps of K-means algorithm for clustering analysis. *(Ans. : Refer section 5.6.1)* **(5 Marks)**

Q. 2 Apply K-means algorithm on given data for k = 3. Use c_1 (2), c_2 (16) and c_3 (38) as initial cluster centres. **(10 Marks)**

Data : 2, 4, 6, 3, 31, 12, 15, 16, 38, 35, 14, 21, 23, 25, 30 *(Ans. : Refer Example 5.6.5)*

Q. 3 What is SVM ? Explain the following terms: hyperplane, separating hyperplane, margin and support vectors with suitable example. *(Ans. : Refer section 5.5.1)* **(5 Marks)**

May 2016

Q. 4 Apply K-means algorithm on given data for k = 3. Use c_1 (2), c_2 (16) and c_3 (38) as initial cluster centres. **(10 Marks)**

Data : 2, 4, 6, 3, 31, 12, 15, 16, 38, 35, 14, 21, 23, 25, 30 *(Ans. : Refer Example 5.6.6)*

Q. 5 What are the key terminologies of Support Vector Machine ? *(Ans. : Refer section 5.5.1)* **(5 Marks)**

Q. 6 Write detail notes on: Quadratic Programming solution for finding maximum margin separation in support vector machine. *(Ans. : Refer sections 5.5.1 and 5.5.2)* **(10 Marks)**

Q. 7 Apply Agglomerative clustering algorithm on given data and draw dendrogram. Show three clusters with its allocated points. Use single link method. *(Ans. : Refer Example 5.6.9)* **(10 Marks)**

	a	B	c	d	E	F
a	0	$\sqrt{2}$	$\sqrt{10}$	$\sqrt{17}$	$\sqrt{5}$	$\sqrt{20}$
b	$\sqrt{2}$	0	$\sqrt{8}$	3	1	$\sqrt{18}$
c	$\sqrt{10}$	$\sqrt{8}$	0	$\sqrt{5}$	$\sqrt{5}$	2
d	$\sqrt{17}$	1	$\sqrt{5}$	0	2	3
e	$\sqrt{5}$	1	$\sqrt{5}$	2	0	$\sqrt{13}$
f	$\sqrt{20}$	$\sqrt{18}$	2	3	$\sqrt{13}$	0

May 2017

Q. 8 What is Support Vector Machine ? How to compute the margin ? *(Ans. : Refer sections 5.5.1 and 5.5.2)* **(10 Marks)**

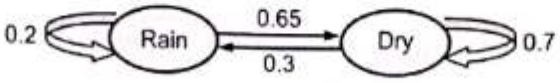
Q. 9 For the given set of points identify clusters using complete link and average link using Agglomerative clustering. **(10 Marks)**

(Ans. : Refer Example 5.6.10)

	A	B
P_1	1	1
P_2	1.5	1.5
P_3	5	5
P_4	3	4
P_5	4	4
P_6	3	3.5

May 2019

Q. 10 Consider Markov chain model for 'Rain' and 'Dry' is shown in following figure Two states: 'Rain' and 'Dry'. Transition probabilities: P('Rain'|'Rain') = 0.2, P('Dry'|'Rain') = 0.65, P('Rain'|'Dry') = 0.3, P('Dry'|'Dry') = 0.7, Initial probabilities : say P('Rain') = 0.4, P('Dry') = 0.6.Calculate a probability of a sequence of states {'Dry', 'Rain', 'Rain', 'Dry'}.. *(Ans. : Refer Example 5.4.3)* **(10 Marks)**



Q. 11 Explain following terms Initial hypothesis, Expectation step and Maximization step w.r.t E-M algorithm. Explain How Initial hypothesis converges to optimal solution? (You may explain it with an example) *(Ans. : Refer section 5.6.3)* **(10 Marks)**

Q. 12 Explain following terms w.r.t Bayes' theorem with proper examples : (a) Independent probabilities (b) Dependent Probabilities (c) Conditional Probability (d) Prior and Posterior probabilities Define Bays theorem based on these Probabilities. *(Ans. : Refer section 5.3.1)* **(10 Marks)**

Q. 13 Draw and discuss the structure of Radial Basis Function Network. How RBFN can be used to solve non linearly separable pattern ? *(Ans. : Refer section 5.6.5)* **(10 Marks)**

Q. 14 Illustrate Support Vector machine with neat labeled sketch and also show how to derive optimal hyper-Plane? *(Ans. : Refer sections 5.5.1 and 5.5.2)* **(10 Marks)**

Dec. 2019

Q. 15 Why is SVM more accurate than logistic regression ? *(Ans. : Refer section 5.5)* **(5 Marks)**

Q. 16 Explain Radial Basis Function with example. *(Ans. : Refer section 5.6.5)* **(5 Marks)**

Q. 17 Explain various basic evaluation measures of supervised learning Algorithm for Classification. *(Ans. : Refer sections 5.1, 5.2 and 5.3)* **(10 Marks)**

Q. 18 Define Support Vector Machine. Explain how margin is computed and optimal hyper-plane is decided ? *(Ans. : Refer sections 5.5.1 and 5.5.2)* **(10 Marks)**

Q. 19 Write short note on : Hidden Markov Model. *(Ans. : Refer section 5.4)* **(5 Marks)**

Q. 20 Write short note on : EM algorithm. *(Ans. : Refer section 5.6.3)* **(5 Marks)**

Module 5

Q. 21 For a unknown tuple t =<Outllook =Sunny, Temperature =Cool, Wind= Strong> use naïve Bayes classifier to find whether the class for Play Tennis is yes or no. The dataset is given below *(Ans. : Refer Example 5.3.1)* **(10 Marks)**

Outlook	Temperature	Wind	Play Tennis
Sunny	Hot	Weak	No
Sunny	Hot	Strong	No
Overcast	Hot	Weak	Yes
Rain	Mild	Weak	Yes
Rain	Cool	Weak	Yes
Rain	Cool	Strong	No
Overcast	Cool	Strong	Yes
Sunny	Mild	Weak	No
Sunny	Cool	Weak	Yes
Rain	Mild	Weak	Yes
Sunny	Mild	Strong	Yes
Overcast	Mild	Strong	Yes
Overcast	Hot	Weak	Yes
Rain	Mild	Strong	No

Multiple Choice Questions

Q. 5.1 Which is not a desirable property of a logical rule based system ?

(a) Locality (b) Attachment

(c) Truth Functionality (d) Global attribute

✓ **Ans. : (b)**

Explanation : Remaining three are the properties of Rule based system.

Q. 5.2 A rule-based system consists of a bunch of IF-THEN rules.

(a) TRUE (b) NO

(c) MAYBE (d) CANT SAY ✓ **Ans. : (a)**

Explanation : Rule base classifier classifies the records by matching if and then conditions.

Q. 5.3 SVM can be used to solve _____ problems.

(a) Classification (b) Regression

(c) Clustering

(d) Both Classification and Regression ✓ **Ans. : (d)**

Explanation : With the help of SVM we can get categorical as well as numerical output.

Q. 5.4 SVM is a _____ learning algorithm.

(a) Supervised (b) Unsupervised

(c) Semisupervised (d) Reinforcement ✓ **Ans. : (a)**

Explanation : Expected output is known.

Q. 5.5 In Clustering which is not true ?

(a) We do not have idea about output

(b) We group data in to different groups

(c) It is unsupervised method

(d) We have the idea about output ✓ **Ans. : (d)**

Explanation : We do not know the expected output.

Q. 5.6 Which of the following clustering algorithms suffers from the problem of convergence at local optima ?

(a) K-Means clustering algorithm

(b) K-Means clustering algorithm and Expectation-Maximization clustering algorithm

(c) Agglomerative clustering algorithm and Diverse clustering algorithm

(d) Agglomerative clustering algorithm and Diverse clustering algorithm and K- Means clustering algorithm ✓ **Ans. : (b)**

Explanation : Out of the options given, only K-Means clustering algorithm and EM clustering algorithm has the drawback of converging at local minima.

Q. 5.7 Which of the following algorithm is most sensitive to outliers ?

(a) K-means clustering algorithm

(b) K-medians clustering algorithm

(c) K-modes clustering algorithm

(d) K-medoids clustering algorithm ✓ **Ans. : (a)**

Explanation : Out of all the options, K-Means clustering algorithm is most sensitive to outliers as it uses the mean of cluster data points to find the cluster center.

Q. 5.8 Which algorithm is used for solving temporal probabilistic reasoning ?

(a) Hill-climbing search

(b) Hidden markov model

(c) Depth-first search

(d) Breadth-first search ✓ **Ans. : (b)**

Explanation : HMM model uses the probability concept along with time dimension.

Q. 5.9 How does the state of the process is described in HMM ?

(a) Literal (b) Single random variable

(c) Single discrete random variable

(d) Multi discrete random variable ✓ **Ans. : (c)**

Explanation : States are represented in HMM model using Single discrete random variable.

Q. 5.10 Select the false statement related to Support vector machine.

(a) SVM can be used as binary classifier

(b) SVM can be used as Multi-class Classifier

(c) SVM can not perform non-linear classification

(d) SVM can be used for linear and non-linear classification ✓ **Ans. : (c)**

Explanation : SVM can perform non-linear classification using the non liner kernels.

Q. 5.11 In Regression tree output attribute is _____

(a) Categorical (b) Discrete

(c) Numerical (d) Range ✓ **Ans. : (c)**

Explanation : In regression tree at the leaf node we get the numerical output.

Q. 5.12 Which of the following is/are valid iterative strategy for treating missing values before clustering analysis?

(a) Nearest Neighbor assignment

(b) Imputation with mean

(c) Imputation with Expectation Maximization algorithm

(d) Imputation with outlier ✓ **Ans. : (c)**

Explanation : When we perform imputation with EM algorithm missing values can be handled before applying clustering.

Q. 5.13 In Radial basis function network the connection between input and hidden layers is called as _____

(a) Weighted connection

(b) Intra connection

(c) Inter connection

(d) Hypothetical connection ✓ **Ans. : (d)**

Explanation : In RBF, hidden layer weight values are calculated using the learning strategies. They are not user defined.

Q. 5.14 Which of the following is a clustering algorithm in machine learning ?

(a) CART

(b) Expectation Maximization

(c) Gaussian Naïve Bayes

(d) Apriori ✓ **Ans. : (b)**

Explanation : EM is clustering. CART and Naive bayes are classification algorithms whereas Apriori is used as Association algorithm.

Q. 5.15 SVMs are less effective when _____

(a) The data is linearly seperable

(b) The data is clean and ready to use

(c) The data is noisy and contains overlapping points

(d) The data is clean **Ans. : (c)**

Explanation : If data is noisy and distribution od data is not proper then we will not be able to draw proper decision boundary.

Q. 5.16 Calculate the Sensitivity from given data TP = 30, TN = 930, FP = 30, FN = 10

(a) 0.75 (b) 1

(c) 0.86 (d) 0.99 ✓ **Ans. : (a)**

Explanation :

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} = \frac{30}{40} = 0.75$$

Q. 5.17 Calculate the Accuracy from given data TP = 30, TN = 930, FP = 30, FN = 10

(a) 0.96 (b) 1

(c) 0.86 (d) 0.99 ✓ **Ans. : (a)**

Explanation : Accuracy $= \frac{(TP + TN)}{(TP + TN + FP + FN)}$

$$= \frac{960}{1000} = 0.96$$

Q. 5.18 How the bayesian network can be used to answer any query?

(a) Full distribution (b) Joint distribution

(c) Partial distribution (d) Zero distribution

✓ **Ans. : (b)**

Explanation : In Bayesian network first we have to find joint probability distribution then from this we can answer any query.

Module 5

Q. 5.19 For clustering, model does not require _____

(a) Unlabeled data (b) Labeled data

(c) Numerical data (d) Categorical data

✓ **Ans. : (b)**

Explanation : In clustering labelled data is not present. In this based on common properties data is grouped together.

Q. 5.20 In EM algorithm that finds maximum likelihood estimates for a model with latent variables. You are supposed to modify the algorithm so that it finds MAP estimates instead. Which step or steps do you need to modify?

(a) Expectation (b) Maximization

(c) No modification required

(d) Sorting ✓ **Ans. : (b)**

Explanation : In maximization step we reinitialize the parameters.

Q. 5.21 The main use of "kernel trick" is/are _____

(a) Used to perform binary classification

(b) Used to perform Multi-class classification

(c) Used to perform linear classification

(d) Implicitly mapping the inputs into high-dimensional feature spaces ✓ **Ans. : (d)**

Explanation : For non linear classification inputs are mapped to high-dimensional feature spaces.

Q. 5.22 The goal of the SVM is to _____

(a) Find the optimal separating hyperplane which minimizes the margin of training data

(b) Find the optimal separating hyperplane which maximizes the margin of training data

(c) Finds hyperplane without any criteria

(d) Does not find hyperplane for classification

✓ **Ans. : (b)**

Explanation : If margin is maximum then we can gain more confidence in our classification.

Q. 5.23 Probabilities in Bayes theorem that are changed with the help of new available information are classified as _____

(a) Independent probabilities

(b) Posterior probabilities

(c) Interior probabilities

(d) Dependent probabilities ✓ **Ans. : (b)**

Explanation : Based on new evidences original probabilities are updated.

Q. 5.24 The basic purpose of clustering is to _____

(a) Combine the data points into single group

(b) Classify the data point into different classes

(c) Predict the output values of input data points

(d) Search data from group ✓ **Ans. : (b)**

Explanation : In clustering based on common properties data is grouped in to different clusters.

Q. 5.25 In EM algorithm, Out of the two repeated steps, the step 2 is _____

(a) minimization step (b) optimization step

(c) normalization step (d) maximization step

✓ **Ans. : (d)**

Explanation : As per working of EM algorithm.

Q. 5.26 The training examples closest to the separating hyperplane are called as _____

(a) Training vectors (b) Test vectors

(c) Support vectors (d) validation vectors

✓ **Ans. : (c)**

Explanation : The points which are closest to decision boundary are called as support vectors.

Q. 5.27 Identify the false statement regarding soft margin of SVM.

(a) It is a modified maximum margin idea that allows for mislabeled examples.

(b) If there exists no hyperplane that can split the "yes" and "no" examples, the Soft Margin method will choose a hyperplane that splits the examples as cleanly as possible

(c) It uses the concept of slack variables

(d) It does not use kernels. ✓ **Ans. : (d)**

Explanation : SVM uses the kernels.

Q. 5.28 If you are using Multinomial mixture models with the expectation-maximization algorithm for clustering a set of data points into two clusters, which of the assumptions are important ?

(a) All the data points follow two Gaussian distribution

(b) All the data points follow n Gaussian distribution ($n > 2$)

(c) All the data points follow two multinomial distribution

(d) All the data points follow n multinomial distribution ($n > 2$) ✓ **Ans. : (c)**

Explanation : For Multinomial mixture models data points should follow two multinomial distributions.

Q. 5.29 The method of finding hidden structure from unlabeled data is called _____

(a) Supervised learning

(b) Unsupervised Learning

(c) Reinforcement learning

(d) Instructional Learning

✓ **Ans. : (b)**

Explanation : In unsupervised learning unlabeled data is used.

Q. 5.30 Classification problems are distinguished from estimation problems in that _____

(a) classification problems require the output attribute to be numeric.

(b) classification problems require the output attribute to be categorical.

(c) classification problems do not allow an output attribute.

(d) classification problems are designed to predict future outcome.

✓ **Ans. : (b)**

Explanation : In classification output attribute should be discreet.

Q. 5.31 Which statement is true about prediction problems?

(a) The output attribute must be categorical.

(b) The output attribute must be numeric.

(c) The resultant model is designed to determine future outcomes.

(d) The resultant model is designed to classify current behaviour.

✓ **Ans. : (c)**

Explanation : In prediction problems based on historical data model is trained to predict the output of future.

Q. 5.32 Unlike traditional production rules, association rules _____

(a) allow the same variable to be an input attribute in one rule and an output attribute in another rule.

(b) allow more than one input attribute in a single rule

(c) require input attributes to take on numeric values.

(d) require each rule to have exactly one categorical output attribute.

✓ **Ans. : (a)**

Explanation : As per working of association rules like apriori algorithm.

Q. 5.33 Given desired class C and population P, lift is defined as _____

(a) the probability of class C given population P divided by the probability of C given a sample taken from the population.

(b) the probability of population P given a sample taken from P.

(c) the probability of class C given a sample taken from population P

(d) the probability of class C given a sample taken from population P divided by the probability of C within the entire population P

✓ **Ans. : (d)**

Explanation : Standard definition of lift.

Q. 5.34 Instead of representing knowledge in a relatively declarative, static way (as a bunch of things that are true), rule based system represent knowledge in terms of _____ that tell you what you should do or what you could conclude in different situations.

(a) Raw Text (b) A bunch of rules

(c) Summarized Text

(d) Collection of various Texts

✓ **Ans. : (b)**

Explanation : In rule based system knowledge is represented using IF-THEN rules.

Q. 5.35 Autonomous Question / Answering systems are _____

(a) Expert Systems

(b) Rule Based Expert Systems

(c) Decision Tree Based Systems

(d) All of the mentioned

✓ **Ans. : (d)**

Explanation : Above all methods can be used to implement autonomous Q & A system.

Q. 5.36 Where does the Hidden Markov Model is used?

(a) Speech recognition

(b) Understanding of real world

(c) Both Speech recognition & Understanding of real world

(d) None of the mentioned

✓ **Ans. : (a)**

Explanation : In speech recognition hidden signals are present for this HMM is used.

Q. 5.37 Where does the baye's rule can be used?

(a) Solving queries

(b) Increasing complexity

(c) Decreasing complexity

(d) Answering probabilistic query

✓ **Ans. : (d)**

Explanation : Based on probability concept queries are answered in baye's system.

Q. 5.38 What does the bayesian network provides?

(a) Complete description of the domain

(b) Partial description of the domain

(c) Complete description of the problem

(d) None of the mentioned

✓ **Ans. : (a)**

Explanation : Bayesian belief network provides the complete scenario of a particular event.

Module 5

Q. 5.39 What are the main components of the expert systems?

(a) Inference Engine

(b) Knowledge Base

(c) Inference Engine & Knowledge Base

(d) None of the mentioned ✓ **Ans. : (c)**

Explanation : Knowledge base is used to store the knowledge and inference engine is used to provide the inference.

Q. 5.40 Three components of Bayes decision rule are class prior, likelihood and _____

(a) Evidence (b) Instance

(c) Confidence (d) Salience ✓ **Ans. : (a)**

Explanation : Based on new evidences rules are updated.

Q. 5.41 In Bays theorem, unconditional probability is called as _____

(a) Evidence (b) Likelihood

(c) Prior (d) Posterior ✓ **Ans. : (a)**

Explanation : Basic definition

Q. 5.42 In Bays theorem, class conditional probability is called as _____

(a) Evidence (b) Likelihood

(c) Prior (d) Posterior ✓ **Ans. : (b)**

Explanation : Basic definition

Q. 5.43 One person is tossing a coin inside a closed room and he tells only the output (Head or Tail) to the person who is standing outside the room .The type of coin which he is using whether fair coin or biased coin is not known. Suppose you want to find out the type of coin which algorithm will help you?

(a) Hidden Markov Model

(b) Discrete Markov Model

(c) Prediction Model

(d) Classification Model ✓ **Ans. : (a)**

Explanation : HMM, Since it has the ability to predict the underlying model for which the output is known.

Q. 5.44 K-Nearest Neighbor is a _____, _____ algorithm

(a) Non-parametric, eager (b) Parametric, eager

(c) Non-parametric, lazy (d) Parametric, lazy

✓ **Ans. : (c)**

Explanation : KNN is non-parametric because it does not make any assumption regarding the underlying data distribution. It is a lazy learning technique because during training time it just memorizes the data and finally computes the distance during testing.

Q. 5.45 You have been given the following 2 statements. Find out which of these options is/are true in case of k-NN?

1. In case of very large value of k, we may include points from other classes into the neighborhood.
2. In case of too small value of k, the algorithm is very sensitive to noise.

(a) 1 is True and 2 is False

(b) 1 is False and 2 is True

(c) Both are True

(d) Both are False ✓ **Ans. : (c)**

Explanation : Both the options are true and are self explanatory.

Q. 5.46 State whether the statement is True/False : k-NN algorithm does more computation on test time rather than train time.

(a) True (b) False ✓ **Ans. : (a)**

Explanation : The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the testing phase, a test point is classified by assigning the label which are most frequent among the k training samples nearest to that query point – hence higher computation.

Q. 5.47 Suppose, you have given the following data where x and y are the 2 input variables and Class is the dependent variable

x	y	class
-1	1	-
0	1	+
0	2	-
1	-1	-
1	0	+
1	2	+
2	2	-
2	3	+

Suppose, you want to predict the class of new data point x=1 and y=1 using Eucledian distance in 3-NN. In which class this data point belong to?

(a) + Class (b) – Class

(c) Can't Say (d) None of these

✓ **Ans. : (a)**

Explanation : All three nearest point are of + class so this point will be classified as + class

5.48 What is the naive assumption in a Naive Bayes Classifier?

(a) All the classes are independent of each other

(b) All the features of a class are independent of each other

(c) The most probable feature for a class is the most important feature to be considered for classification

(d) All the features of a class are conditionally dependent on each other.

✓ **Ans. : (b)**

Explanation : Naive Bayes Assumption is that all the features of a class are independent of each other which is not the case in real life. Because of this assumption, the classifier is called Naive Bayes Classifier.

5.49 Consider the following dataset. a, b, c are the features and K is the class(1/0) :

a	b	c	k
1	0	1	1
1	1	1	1
0	1	1	0
1	1	0	0
1	0	1	0
0	0	0	1

Classify the test instance given below into class 1/0 using a Naive Bayes Classifier.

a	b	c	k
0	0	1	?

(a) 0 (b) 1

✓ **Ans. : (b)**

Explanation :

P(K=1|a=0,b=0,c=1)=3/6*1/3*2/3*2/3 = 0.7407

P(K=0|a=0,b=0,c=1)= 3/6*1/3*1/3*2/3 = 0.03703

P(K=1|a=0,b=0,c=1) > P(K=1|a=0,b=0,c=1)

Q. 5.50 A patient goes to a doctor with symptoms S_1, S_2 and S_3. The doctor suspects disease D_1 and D_2 and constructs a Bayesian network for the relation among the disease and symptoms as the following :

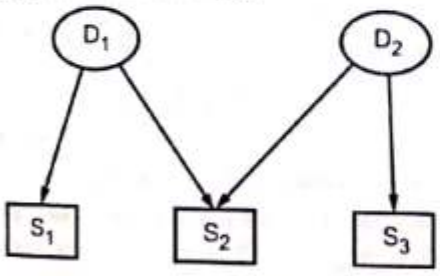


Fig. Q. 5.50

What is the joint probability distribution in terms of conditional probabilities?

(a) $P(D_1)*P(D_2/D_1)*P(S_1/D_1)*P(S_2/D_1)*P(S_3/D_2)$

(b) $P(D_1)*P(D_2)*P(S_1/D_1)*P(S_2/D_1)*P(S_3/D_2)$

(c) $P(D_1)*P(D_2)*P(S_1/D_2)*P(S_2/D_2)*P(S_3/D_2)$

(d) $P(D_1)*P(D_2)*P(S_1/D_1)*P(S_2/D_1,D_2)*P(S_3/D_2)$

✓ **Ans. : (d)**

Explanation : From the figure, we can see that D_1 and D_2 are not dependent on any variable as they don't have any incoming directed edges. S_1 has an incoming edge from D_1, hence S_1 depends on D_1. S_2 has 2 incoming edges from D_1 and D_2, hence S_2 depends on D_1 and D_2. S_3 has an incoming edge from D_2. S_3 depends on D_2. Hence, d is the answer.

Q. 5.51 What is the Markov blanket of variable, S_3

(a) D_1 (b) D_2

(c) D_1 and D_2 (d) None

✓ **Ans. : (b)**

Explanation : In a Bayesian Network, the Markov blanket of node, X is the set consisting of X's parents, X's children and Parents of X's children. In the given diagram, variable, S_2 has a parent D_2 and no children. Hence, the correct answer is (b).

Q. 5.52 Suppose you are using a Linear SVM classifier with 2 class classification problem. Consider the following data in which the points circled red represent support vectors. Will the decision boundary change if any of the red points are removed?

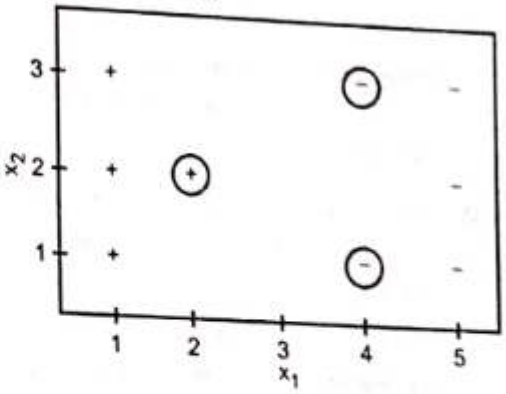


Fig. Q. 5.52

(a) Yes (b) No

✓ **Ans. : (a)**

Explanation : These three examples are positioned such that removing any one of them introduces slack in the constraints. So the decision boundary would completely change.

Module 5

Q. 5.53 Consider the data-points in the figure below.

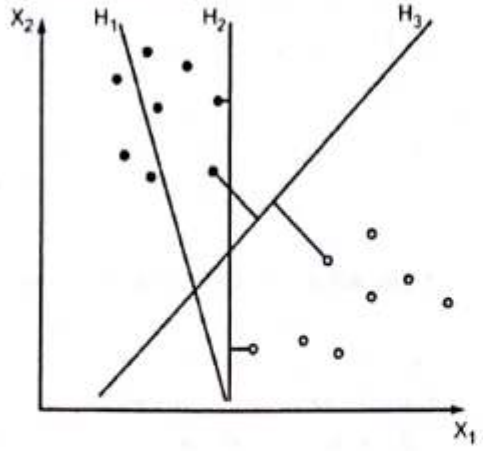


Fig. Q. 5.53

Let us assume that the black-colored circles represent positive class whereas the white colored circles represent negative class. Which of the following among H_1, H_2 and H_3 is the maximum-margin hyperplane?

(a) H_1 (b) H_2

(c) H_3 (d) None of the above. ✓ **Ans. : (c)**

Explanation : H_1 does not separate the classes. H_2 does, but only with a small margin. H_3 separates them with the maximal margin.

Q. 5.54 The soft margin SVM is more preferred than the hard-margin svm when :

(a) The data is linearly separable

(b) The data is noisy and contains overlapping point

✓ **Ans. : (b)**

Explanation : When the data has noise and overlapping points, there is a problem in drawing a clear hyperplane without misclassifying.

Q. 5.55 After training an SVM, we can discard all examples which are not support vectors and can still classify new examples?

(a) True (b) False ✓ **Ans. : (a)**

Explanation : Since the support vectors are only responsible for the change in decision boundary.

Q. 5.56 Suppose that we use a RBF kernel with appropriate parameters to perform classification on a particular two class data set where the data is not linearly separable. In this scenario

(a) the decision boundary in the transformed feature space is non-linear

(b) the decision boundary in the transformed feature space is linear

(c) the decision boundary in the original feature space is linear

(d) the decision boundary in the original feature space is non-linear ✓ **Ans. : (b), (d)**

Explanation : As per working of RBF.

Q. 5.57 Which of the following statements is/are true about kernel in SVM?

1. Kernel function map low dimensional data to high dimensional space
2. It's a similarity function

(a) 1 is True but 2 is False (b) 1 is False but 2 is True

(c) Both are True (d) Both are False ✓ **Ans. : (c)**

Explanation : As per working of SVM

Q. 5.58 What is true about K-Mean Clustering?

1. K-means is extremely sensitive to cluster center initializations
2. Bad initialization can lead to Poor convergence speed
3. Bad initialization can lead to bad overall clustering

(a) 1 and 3 (b) 1 and 2

(c) 2 and 3 (d) 1, 2 and 3 ✓ **Ans. : (c)**

Explanation : All three of the given statements are true. K-means is extremely sensitive to cluster center initialization. Also, bad initialization can lead to Poor convergence speed as well as bad overall clustering.

Q. 5.59 If in the following figure we draw a horizontal line on y axis for y = 2 how many number of clusters we will get?

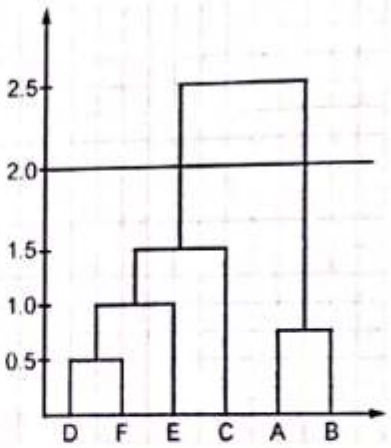


Fig. Q. 5.59

(a) 1 (b) 2 (c) 3 (d) 4 ✓ **Ans. : (b)**

Explanation : Since the number of vertical lines intersecting the red horizontal line at y = 2 in the dendrogram are 2, therefore, two clusters will be formed.

Q. 5.60 Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration the clusters : C_1, C_2, C_3 has the following observations :

1. C_1 : {(1,1), (4,4), (7,7)}
2. C_2 : {(0,4), (4,0)}
3. C_3 : {(5,5), (9,9)}

What will be the cluster centroids if you want to proceed for second iteration?

(a) C_1 : (4,4), C_2 : (2,2), C_3 : (7,7)

(b) C_1 : (2,2), C_2 : (0,0), C_3 : (5,5)

(c) C_1 : (6,6), C_2 : (4,4), C_3 : (9,9)

(d) None of these

✓ **Ans. : (a)**

Explanation : Finding centroid for data points in cluster

$$C_1 = \left(\frac{(2+4+6)}{3}, \frac{(2+4+6)}{3}\right) = (4, 4)$$

Finding centroid for data points in cluster

$$C_2 = \left(\frac{(0+4)}{2}, \frac{(4+0)}{2}\right) = (2, 2)$$

Finding centroid for data points in cluster

$$C_3 = \left(\frac{(5+9)}{2}, \frac{(5+9)}{2}\right) = (7, 7)$$

Hence, C_1 : (4,4), C_2 : (2,2), C_3 : (7,7)

Q. 5.61 If two variables V_1 and V_2 are used for clustering. Which of the following are true for K means clustering with k = 3 ?

1. If V_1 and V_2 has a correlation of 1, the cluster centroids will be in a straight line
2. If V_1 and V_2 has a correlation of 0, the cluster centroids will be in straight line

Choose the correct answer?

(a) 1 Only

(b) 2 Only

(c) Both 1 and 2

(d) None of the above

✓ **Ans. : (a)**

Explanation : If the correlation between the variables V_1 and V_2 is 1, then all the data points will be in a straight line. Hence, all the three cluster centroids will form a straight line as well.

Chapter Ends...

❑❑❑

Module 6

Chapter...6 Dimensionality Reduction

University Prescribed Syllabus

Dimensionality Reduction Techniques, Principal Component Analysis, Independent Component Analysis, Single value decomposition.

6.1 DIMENSIONALITY REDUCTION TECHNIQUES

- Dimension Reduction alludes to the way toward changing over an arrangement of information having tremendous dimensions into information with lesser dimensions guaranteeing that it passes on comparable data briefly. These methods are normally utilized while tackling machine learning issues to acquire better properties for classification or regression problem.
- We should take a gander at the picture demonstrated as follows. It indicates 2 dimensions P_1 and P_2, which are given us a chance to state estimations of a few objects in cm (P_1) and inches (P_2). Presently, if we somehow managed to utilize both the measurements in machine learning, they will pass on comparable data and present a lot of noise in the system, so it is better to simply utilizing one dimension. Dimension of information present here is changed over from 2D (from P_1 and P_2) to 1D (Q_1), to make the information moderately less demanding to clarify.

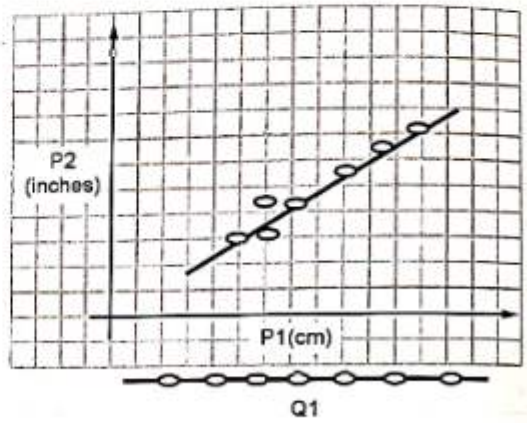


Fig. 6.1.1 : Example of Dimension Reduction

- There are number of methods using which, we can get k dimensions by reducing n dimensions (k < n) of informational index. These k dimensions can be specifically distinguished (sifted) or can be a blend of dimensions (weighted midpoints of dimensions) or new dimension(s) that speak to existing numerous dimensions well. A standout amongst the most well-known use of this procedure is Image preparing.
- Now we will see the significance of applying Dimension Reduction method :
  - It assists in information packing and diminishing the storage room required
  - It decreases the time which is required for doing same calculations. If the dimensions are less then processing will be less, added advantage of having less dimensions is permission to use calculations unfit for countless.
  - It handles with multi-collinearity that is used to enhance the execution of the model. It evacuates excess highlights. For example: it makes no sense to put away an incentive in two distinct units (meters and inches).

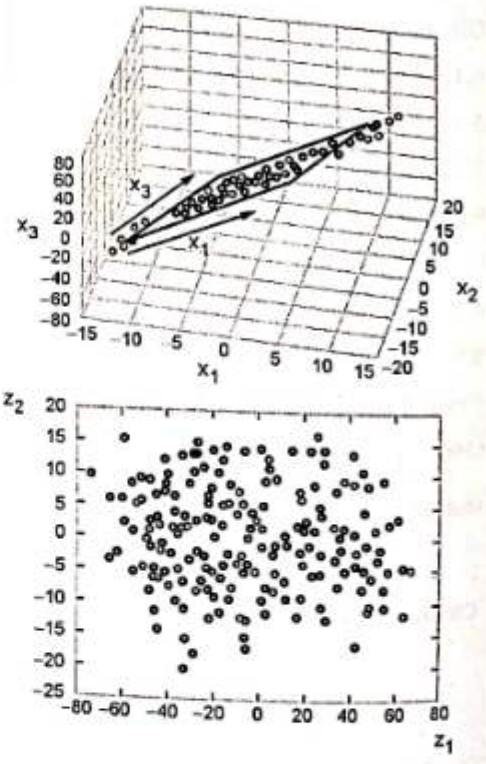


Fig. 6.1.2 : Example of Dimension Reduction

- Lessening the dimensions of information to 3D or 2D may enable us to plot and visualize exactly. You would then be able to watch designs all the more unmistakably. Beneath we can see that, how a 3D information is changed over into 2D. It has distinguished the 2D plane at that point spoke to the focuses on these two new axis z_1 and z_2.
- It is helpful in noise evacuation additionally and because of this we can enhance the execution of models.
- The classifier's performance usually will degrade for a large number of features.

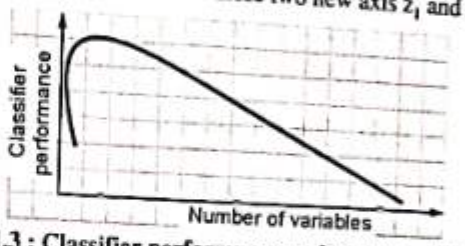


Fig. 6.1.3 : Classifier performance and amount of data

6.1.1 Dimension Reduction Techniques in ML

6.1.1(A) Feature Selection

Given a set of features $F = \{X_1, \ldots, X_n\}$. The feature selection problem is to find a subset $F' \subset F$ that maximizes the learner's ability to classify the patterns. Finally F' should maximize some scoring function.

$$\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_N \end{bmatrix} \xrightarrow{\text{Feature selection}} \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ . \\ . \\ x_{i_M} \end{bmatrix}$$

Feature Selection Steps

Feature selection is an **optimization** problem.

- **Step 1 : Search the space of possible feature subsets.**
- **Step 2 : Pick the subset that is optimal or near-optimal with respect to some objective function.**

Subset selection

- There are d initial features and 2^d possible subsets.
- Criteria to decide which subset is the best :
  - Classifier based on these m features has the lowest probability of error of all such classifiers.
- It is not possible to go over all 2^d possibilities, so we need some heuristics.
- Here we select uncorrelated features.
- Forward search
  - Start from empty set of features.
  - Try each of remaining features.
  - Estimate classification/regression error for adding specific feature.
  - Select feature that gives maximum improvement in validation error.

- Stop when no significant improvement.

– Backward search

- Start with original set of size *d*.
- Drop features with smallest impact on error.

6.1.1(B) Feature Extraction

Suppose a set of features F = {X_1, ..., X_N} is given. The Feature Extraction task is to map F to some feature set F" that will maximize the learner's ability to classify patterns.

$$\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_N \end{bmatrix} \xrightarrow[-\quad -]{\text{Feature extraction}} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_m \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_N \end{bmatrix}\right)$$

A projection matrix w is computed from N-dimensional to M-dimensional vectors to achieve low error.

$Z = w^T x$. Principle component analysis and Independent component analysis are the feature extraction methods.

6.2 PRINCIPLE COMPONENT ANALYSIS

– Dimensional reduction is regularly the best beginning stage when dealing with high dimensional information. It is utilized for an assortment of reasons, from perception to denoising, and in a wide range of uses, from signal processing to bioinformatics.

– A standout amongst the most broadly utilized dimensional reduction tools is Principal Component Analysis (PCA).

– PCA verifiably accept that the dataset under thought is typically dispersed, furthermore, chooses the subspace which expands the anticipated difference. We consider a centered data set, and develop the sample covariance matrix, at that point q-dimensional PCA is identical to anticipating onto the q-dimensional subspace spread over by the q eigenvectors of S with biggest eigenvalues.

– In this system, variables are changed into another arrangement of variables, which are straight blend of unique variables. These new arrangement of variables are known as **principle components**. They are calculated so that first **principle component** s represents a large portion of the conceivable variety of unique information after which each succeeding component has the most noteworthy conceivable variance.

– The second principal component should be symmetrical to the primary principal component. As it were, it does its best to catch the difference in the information that isn't caught by the primary principal component. For two-dimensional dataset, there can be just two principal components. The following is a depiction of the information and its first and second principal component. You can see that second principal component is symmetrical to first principal component.

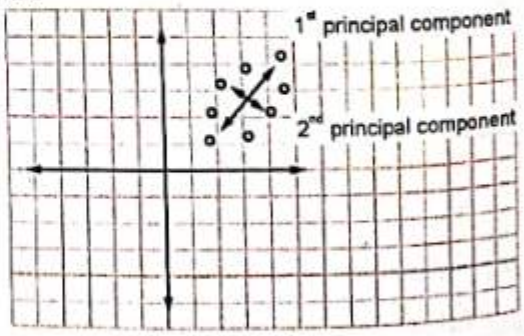


Fig. 6.2.1

- The principal components are sensitive to the size of estimation, now to settle this issue we ought to dependably institutionalize factors before applying PCA. Applying PCA to your informational collection loses its importance. In the event that interpretability of the outcomes is critical for your investigation, PCA isn't the correct system for your task.

Example 6.2.1 : Apply PCA on the following data and find the principle component.

X	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
Y	2.4	0.7	2.9	2.2	3	2.7	1.6	1.1	1.6	0.9

Solution :

- First we will find the mean values

$$X_m = \sum \frac{X}{N} \quad \text{..... N = number of data points = 10}$$

$$X_m = 1.81$$

$$Y_m = \sum \frac{Y}{N}$$

$$Y_m = 1.91$$

Now we will find the covariance matrix, $C = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix}$

$$C_{XX} = \sum \frac{(X - X_m)^2}{N - 1} = 0.6165$$

$$C_{XY} = C_{YX} = \sum \frac{(X - X_m)(Y - Y_m)}{N - 1} = 0.61544$$

$$C_{YY} = \sum \frac{(Y - Y_m)^2}{N - 1} = 0.7165$$

$$C = \begin{bmatrix} 0.6165 & 0.61544 \\ 0.61544 & 0.7165 \end{bmatrix}$$

- Now to find the eigen values following equation is used.

$$|C - \lambda I| = 0$$

$$\left| \begin{bmatrix} 0.6165 & 0.61544 \\ 0.61544 & 0.7165 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

- By solving the above determinant we will get quadratic equation and solving that equation we will get two eigen values of λ as $\lambda_1 = 0.0489$ and $\lambda_2 = 1.284$.
- Now we will find the eigen vectors corresponding to eigen values.
- First we will find the first eigen vector corresponding to first eigen value, $\lambda 1 = 0.0489$.

$$C \times V_1 = \lambda_1 \times V_1 = \begin{bmatrix} 0.6165 & 0.61544 \\ 0.61544 & 0.7165 \end{bmatrix} \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix} = 0.0489 \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix}$$

From the above equation we will get two equations as

$$0.6165\, V_{11} + 0.61544\, V_{12} = 0.0489\, V_{11}$$

$$0.61544\, V_{11} + 0.7165\, V_{12} = 0.0489\, V$$

- To find the eigen vector we can take either Equation (1) or (2), for both the equations answer will be the same, let's take the first equation

$$= 0.6165\ V_{11} + 0.61544\ V_{12} = 0.0489\ V_{11}$$

$$= 0.5676\ V_{11} + 0.61544\ V_{12}$$

$$V_{11} = -0.61544\ V_{12}$$

$$V_{11} = -1.0842\ V_{12}$$

- Now we will assume $V_{12} = 1$ then $V_{11} = -1.0842$

$$V_1 = \begin{bmatrix} -1.0842 \\ 1 \end{bmatrix}$$

- As we are assuming the value of V_{12} we have to normalized V_1 as follow

$$V_{1N} = \frac{1}{\sqrt{1.0842^2 + 1^2}} \begin{bmatrix} -1.0842 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.735 \\ 0.677 \end{bmatrix}$$

- Similarly we will find the second eigen vector corresponding to second eigen value, $\lambda_2 = 1.284$

$$C \times V_2 = \lambda_2 \times V_2 = \begin{bmatrix} 0.6165 & 0.61544 \\ 0.61544 & 0.7165 \end{bmatrix} \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix} = 1.284 \begin{bmatrix} V_{11} \\ V_{12} \end{bmatrix}$$

- From the above equation we will get two equations as

$$0.6165\ V_{21} + 0.61544 V_{22} = 0.0489\ V_{21}$$

$$0.61544\ V_{21} + 0.7165 V_{22} = 0.0489\ V_{22}$$

- To find the eigen vector we can take either Equation (1) or (2), for both the equations answer will be the same, let's take the first equation

$$= 0.6165\ V_{21} + 0.61544\ V_{22} = 1.284\ V_{21} - 0.6675\ V_{21} = -0.61544 V_{22}$$

$$V_{21} = 0.922\ V_{22}$$

- Now we will assume $V_{22} = 1$ then $V_{21} = 0.922$

$$V_2 = \begin{bmatrix} 0.922 \\ 1 \end{bmatrix}$$

- As we are assuming the value of V_{22} we have to normalized V_2 as follows :

$$V_{2N} = \frac{1}{\sqrt{0.922^2 + 1^2}} \begin{bmatrix} 0.922 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.677 \\ 0.735 \end{bmatrix}$$

- Now we have to find the principle component, it is equal to the eigen vector corresponding to maximum Eigen value, in this λ_2 is maximum, hence principle component is Eigen vector V_{2N}.

Principle component $= \begin{bmatrix} 0.677 \\ 0.735 \end{bmatrix}$

6.3 INDEPENDENT COMPONENT ANALYSIS

- Independent component analysis breaks a multivariate signal in to autonomous non-Gaussian signals. For instance, we take a example of sound which is a typical a signal consist of the numerical expansion. In sound at each time, signals from some sources are included.

- The inquiry at that point is whether we can isolate these contributing sources from the watched aggregate signal or not. At the point when the factual freedom presumption is right, blind ICA partition of a blended signal leads in great results. This is additionally utilized for signals that shouldn't be created by blending for investigation purposes.
- A straightforward use of ICA is the "Get-together party issue", in which the basic discourse signals are separated from example information consisting of individuals talking at the same time in a room. When we do not consider the echoes and the delay in time, the issue is improved.
- A separated and deferred signal represents a duplicate of a reliant part, and accordingly the measurable autonomy supposition isn't violated.
- ... point to take in ... consideration that if 'N' sources are available, in any event N perceptions (for example amplifiers, if the observed signal is sound) are expected to recuperate the first signals. This comprises the situation when the matrix is square (J = D, here D indicates the dimension of the input and model's measurement is J). Different instances of underdetermined (i.e. J > D) and overdetermined (i.e. J < D) have been examined.
- Independent component analysis separates blended signals that leads to quality outcomes depending on two beliefs and three effects of blending source signals.
- Two beliefs :
  1. The source signals are autonomous of one another.
  2. Each source signal values have non-Gaussian conveyances.
- Three effects of blending source signals :
  1. **Autonomy :** Considering to belief 1, ICA comprises of the autonomous source signals; their signal blends are most certainly not. This is based on the fact that the signal blends share a similar source signals.
  2. **Ordinariness :** "Considering the Central Limit Theorem, the dispersion of an aggregate of free irregular factors having limited fluctuation approaches towards a Gaussian circulation". Freely, an aggregate of two autonomous irregular factors as a rule has a dissemination nearer to Gaussian than any of the two unique variables. At this point we think about the calculation of each signal as the random variable.
  3. **Multifaceted nature :** The transient complexity of any of the signal blend is more noteworthy than that of its least difficult constituent source signal.
- Those standards add to the essential foundation of ICA. On the off chance that the signals we happen to extricate from an arrangement of blends are autonomous like sources signals, and have non-Gaussian histograms or have low complexity like source signals, at that point they should be source signals.
- ICA finds the independent components by amplifying the factual autonomy of the evaluated components. We may pick one of numerous approaches to characterize an intermediary for autonomy, and this decision oversees the type of the Independent component analysis calculation. The two general meanings of autonomy for ICA are :
  1. Mutual information minimization
  2. Non-Gaussianity maximization
- The Minimization is achieved for Mutual data groups of ICA by doing the calculations that utilizes measures such as maximum entropy or Kullback-Leibler Divergence. The non-Gaussianity group of ICA calculations, persuaded by as far as possible hypothesis, uses measures such as negentropy and kurtosis.

- Ordinary calculations for Independent component analysis utilize centering, whitening, and dimension reduction as preprocessing ventures with the end goal to streamline and diminish the complexity of the issue for the real iterative calculation.
- Dimension reduction and whitening can be accomplished with main part examination or solitary esteem disintegration. Whitening guarantees that complete measurements are dealt with similarly from the earlier before the calculation is run. Surely understood calculations for ICA incorporate Fast ICA, infomax, kernel-independent component analysis, JADE, and among others. ICA can't recognize the real number of source signals, an interestingly right requesting of the source signals, nor the correct scaling (counting sign) of the source signals.
- Independent component analysis is vital to dazzle signal detachment and has numerous down to real time applications. It is firmly identified with (or even an exceptional instance of) the scan for a factorial code of the information, i.e., another vector-esteemed portrayal of every datum vector to such an extent that it gets interestingly encoded by the subsequent code vector (misfortune free coding), yet the code segments are factually autonomous.
- To thoroughly characterize ICA, we can utilize a measurable "latent factors" demonstrate. Expect that we watch n direct blends $x_1 \ldots x_n$ of n independent segments

$$x_j = a_{j1} s_1 + a_{j2} s_2 + \ldots\ldots\ldots + a_{jn} s_n, \text{ for all } j$$

- In Independent component analysis demonstrate, we have now dropped the time record t, we expect that every blend x_j and also every free part s_k is a random variable, rather than a legitimate time signal. The observed values $x_j(t)$, e.g., the mouthpiece motions in the get-together party issue, are then an example of this arbitrary variable. Without loss of sweeping statement, we can expect that both the blend variables and the free segments have zero mean: If this isn't valid, at that point the noticeable variables x_i can simply be focused by subtracting the example mean, which makes the model zero-mean.
- It is helpful to utilize vector-matrix documentation rather than the aggregates like in the past condition. Lets indicate by x the arbitrary vector whose components are the blends $x_1 \ldots x_n$, and in like manner by s the random vector with components $s_1, \ldots, s_n$. Lets represent mean by A the matrix with components a_{ij}. All vectors are comprehended as column vectors; subsequently x^T, or the transpose of x, is a column vector. Utilizing this vector-matrix documentation, the above blending model is composed as x= As
- Some of the time we require the matrix A columns; represent columns by a_j and the model is composed as

$$x = \sum_{i=1}^{n} a_i s_i$$

- The factual model x = As is known as independent component analysis. The ICA is a generative model, which implies that it depicts how the observed information are created by a procedure of blending the parts s_i. The independent components are inactive variables, implying that they can't be specifically observed. Additionally the mixing matrix is thought to be obscure. All we observe is the arbitrary vector x, and we should estimate both A and s utilizing it. This must be done under as general suppositions as could reasonably be expected.
- The beginning stage for ICA is the plain straightforward suspicion that the parts s_i are factually autonomous. It will be seen underneath that we should likewise expect that the autonomous component must have non Gaussian appropriations.

- Nonetheless, in the essential model we don't accept these dispersions known (on the off chance that they are known, the issue is extensively disentangled.) For effortlessness, we are additionally expecting that the obscure blending grid is square, yet this presumption can be now and then loose. At that point, subsequent to evaluating the matrix A, we can Fig. its opposite, say W, and acquire the autonomous component just by : $s = Wx$
- ICA is firmly identified with the technique called *blind source separation* (BSS) or blind signal partition. A "source" implies here a unique signal, i.e. independent component, similar to the speaker in a cocktail party issue.
- "Blind" implies that we know practically nothing, in the event that anything, on the mixing matrix, and make little suppositions on the source signals. ICA is one technique, maybe the most generally utilized, for performing blind source partition.
- In numerous applications, it would be more sensible to accept that there is some noise in the estimations, which would mean including a noise term in the model. For straightforwardness, we exclude any noise terms, since the estimation of the commotion free model is sufficiently troublesome in itself, and is by all accounts adequate for some applications.

6.3.1 Preprocessing for ICA

Preprocessing is exceptionally helpful before actually applying an ICA calculation on the information. Now we will see some preprocessing methods that lead into less complex ICA estimation and a better model.

Centering

- The essential and vital preprocessing method is to center x. Centering is done by subtracting its mean vector $m = E\{x\}$ from x, so that x becomes a zero-mean variable. This means s is zero-mean too.
- This preprocessing is made exclusively to improve the ICA calculations : It doesn't imply that the mean couldn't be assessed. In the wake of assessing A is the mixing matrix that contains centered information, Calculations can be finished by including the mean vector of s again back to the focused evaluations of s. $A^{-1}m$ represents the mean vector of s , where m is the mean that was subtracted in the preprocessing.

Whitening

- Initially whitening the observed variables is another important preprocessing procedure in ICA. This indicates that before actually doing the calculation of ICA, we need to change the observed vector x straightly with the desired goal that we get another white vector $\tilde{x}$. The parts of the white vectors are uncorrelated and their differences measure up to solidarity. As it were, the covariance matrix of $\tilde{x}$ meets the identity matrix :

$$E\left\{\tilde{x}\,\tilde{x}^T\right\} = I$$

- The whitening change is constantly conceivable. One prevalent technique of whitening uses the eigen- value decay of the covariance matrix $E\{xx^T\} = EDE^T$, here E represents the orthogonal matrix of eigenvectors of $E\{xx^T\}$ and corner to corner network of its eigenvalues is D. $D = diag(d_1, \ldots, d_n)$. $E\{xx^T\}$ can be assessed uniformly from the available example x(1),...., x(T) . whitening should now be possible by

$$\tilde{x} = E D^{-1/2} E^T x$$

where the matrix $D^{-1/2} D$ is calculated by a simple component-wise operation as $D^{-1/2} = diag\,(d_1^{-1/2}, \ldots, d_n^{-1/2})$. It is easy to check that now $E\left\{\tilde{x}\,\tilde{x}^T\right\} = I$

where the grid D – 1/2 D is registered by a straightforward segment shrewd activity as
D – 1/2 = diag ([d_1] ^ (– 1/2),... , [d_n] ^ (– 1/2)). It is anything but difficult to watch that now E{ $\tilde{x}$ $\tilde{x}$ T} = I

- Whitening transforms the mixing matrix into a new one, A. We have from (4) and (34) :

$$\tilde{x} = E D^{-1/2} E^T A s = \tilde{A} s$$

- Brightening changes the blending lattice into another one, A. We have from (4) and (34) :

The utility of whitening resides in the fact that the new mixing matrix $\tilde{A}$ is orthogonal. This can be seen from

$$E\left\{\tilde{x}\tilde{x}^T\right\}^T) = \tilde{A} E\{ss^T\} \tilde{A}^T = \tilde{A}\tilde{A}^T = I$$

The utility of brightening lives in the way that the new blending lattice $\tilde{A}$ is symmetrical.

- Whitening decreases the quantity of parameters to be evaluated. The new, orthogonal mixing matrix A needs to be calculated instead of assessing the n^2 parameters which are the components of the original matrix A. An orthogonal matrix consists of n(n – 1)/2degrees of opportunity. For example, in two measurements, orthogonal transformation is dictated by a solitary angle parameter. In bigger measurements, orthogonal matrix consist of a portion of the quantity of parameters of a discretionary matrix. Consequently we can say that whitening handles half of the problems of ICA. Because whitening is an extremely basic and standard methodology, it is significantly less difficult than any ICA computations, it is a good thought to lessen the multifaceted nature of the issue along these lines.

- It will be very useful to lessen the measurement of the information in the meantime as we do the whitening. In the case of eigenvalues d_j of E{ xx^T}, we dispose those that are too little, as is regularly done in the factual method of essential part examination. This has frequently the impact of decreasing commotion. Besides, measurement decrease anticipates overlearning, which can once in a while be seen in ICA.

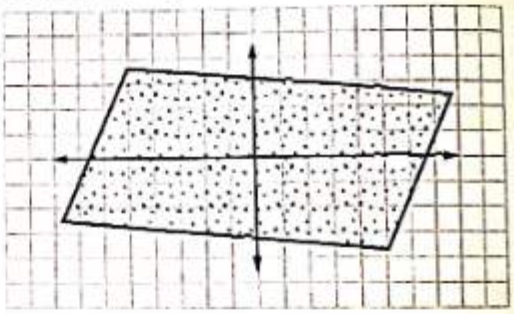


Fig. 6.3.1 : The joint distribution of the whitened mixtures

- In whatever remains of this instructional exercise, we accept that the information has been preprocessed by centering and whitening. For effortlessness of documentation, we signify the preprocessed information just by x, and the changed mixing matrix by A, discarding the tildes.

6.3.2 The Fast ICA Algorithm

In the first segments, we presented diverse proportions of nongaussianity, i.e. target capacities for ICA estimation. By and by, one additionally needs a calculation for expanding the difference work. In this area, we present an extremely productive strategy for augmentation. Here it is accepted that the information is preprocessed by centering and whitening as examined in the first area.

Fast ICA for one unit

- Now, we will demonstrate the one-unit adaptation of Fast ICA. By a "unit" we allude to a computational unit, in the long run a fake neuron, with a weight vector w that the neuron can be updated by a learning method. The Fast ICA

learning principle finds a bearing, i.e. a unit vector w such that the projection $w^T x$ augments non gaussianity. Non gaussianity is here estimated by the guess of negentropy $J(w^T x)$. Review that the fluctuation of $w^T x$ should here be compelled to solidarity; for brightened information this is equal to obliging the standard of w to be unity.

- The Fast ICA depends on a $w^T x$ emphasis conspire for finding a greatest of the non gaussianity of $w^T x$. It tends to be likewise inferred as an approximative Newton cycle. Indicate by g the subordinate of the nonquadratic work G.

$$g_1(u) = \tanh(a_1 u),$$

$$g_2(u) = u \exp(-u^2/2)$$

where $1 \le a_1 \le 2$ is some proper constant, frequently chosen as $a_1 = 1$. The fundamental type of the Fast ICA calculation is as per the following :

1. Select an initial weight vector w.
2. Calculate, $w^+ = E\{xg(w^T x)\} - = E\{g'(w^T x)\}w$
3. Calculate, $w = w^+ / \|W^+\|$
4. If not converged, go back to step 2.

Note that convergence implies that the old and new estimations of w point a similar way, i.e. their dot-product is (nearly) equivalent to 1. It isn't essential that the vector converges to a solitary point, since w and - w characterize a same direction. This is again in light of the fact that the independent components can be characterized just up to a multiplicative sign. Note likewise that it is here expected that the information is prewhitened.

The induction of Fast ICA is as per the following. First note that the maxima of the estimate of the negentropy of $w^T x$ are gotten at certain optima of $E\{G(w^T x)\}$. As indicated by the Kuhn-Tucker conditions, the optima of $E\{G(w^T x)\}$ under the requirement $E\{(w^T x)^2\} = \|W\|^2 = 1$ are acquired at points where

$$E\{xg(w^T x)\} - \beta w = 0$$

Let us try to solve this equation by Newton's method. Jacobian matrix JF(w) as $JF(w) = E\{xx^T g'(w^T x)\} - \beta I$

1. To make easy the inversion of this matrix, we have to approximate the first term. Since the data is sphered, a reasonable approximation seems to be $E\{xx^T g'(w^T x)\} \approx E\{xx^T\} E\{g'(w^T x)\} = E\{g'(wTx)\} I$.

   Thus, the Jacobian matrix becomes diagonal, and can easily be inverted. Thus we obtain the following approximative Newton iteration :

$$w^+ = w - [E\{xg(w^T x)\} - \beta w] / [E\{g'(w^T x)\} - \beta]$$

6.4 SINGLE VALUE DECOMPOSITION

- In singular value decomposition method a matrix is decomposed into three other matrices:

$$A = USV^T$$

- Here, A represents m × n matrix. U represents m × n orthogonal matrix. S is a n × n diagonal matrix and V is a n × n orthogonal matrix.
- Matrix U has the left singular vectors as columns; S is a diagonal matrix which contains singular values; and V^T has right singular vectors as rows. In singular value decomposition original data present in a coordinate system is expanded. Here the covariance matrix is diagonal.
- To calculate singular value decomposition we need to find the eigenvalues and eigenvectors of AA^T and A^TA. The columns of V consists of eigenvectors of A^T. The columns of U consists of the eigenvectors of AA^T.

Module 6

- The square roots of eigenvalues from AA^T or A^TA represents the singular values in S.
- The singular values are arranged in descending order and stored as the diagonal entries of the S matrix. The singular values are always real numbers. If the matrix A is a real matrix, then U and V are also real.

Example 1 : Find SVD for $A = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}$

First we will calculate $A^TA = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$

Now we will calculate eigen vectors V_1 and V_2 using the method that we have seen in PCA

$$V_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \qquad V_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

Next we will calculate AV_1 and AV_2

$$AV_1 = \begin{bmatrix} 2\sqrt{2} \\ 0 \end{bmatrix} \qquad AV_2 = \begin{bmatrix} 0 \\ \sqrt{2} \end{bmatrix}$$

Next we will calculate U_1 and U_2

$$U_1 = \frac{AV_1}{AV_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad U_2 = \frac{AV_2}{|AV_2|} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

SVD is written as,

$$\therefore A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2\sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

6.5 UNIVERSITY QUESTIONS AND ANSWERS

May 2015

Q. 1 Explain in detail Principal Component Analysis for Dimension Reduction. *(Ans. : Refer section 6.2)* **(10 Marks)**

May 2016

Q. 2 Explain in detail Principal Component Analysis for Dimension Reduction. *(Ans. : Refer section 6.2)* **(10 Marks)**

May 2017

Q. 3 Use Principal Component analysis (PCA) to arrive at the transformed matrix for the given data.

$A = \begin{bmatrix} 2 & 1 & 0 & -1 \\ 4 & 3 & 1 & 0.5 \end{bmatrix}$ *(Ans. : Refer section 6.2)* **(10 Marks)**

May 2019

Q. 4 Why Dimensionality Reduction is very Important step in Machine Learning ? *(Ans. : Refer section 6.1)* **(5 Marks)**

Q. 5 Why Dimensionality reduction is an important issue? Describe the steps to reduce dimensionality using Principal Component Analysis method by clearly stating mathematical formulas used.

(Ans. : Refer sections 6.1 and 6.2) **(10 Marks)**

Q. 6 Write Short note on ISA and compare it with PCA. *(Ans. : Refer sections 6.2 and 6.3)* **(5 Marks)**

Dec. 2019

Q. 7 What is Dimensionality reduction? Describe how Principal Component Analysis is carried out to reduce dimensionality of data sets. *(Ans. : Refer sections 6.1 and 6.2)* **(10 Marks)**

Q. 8 Find the singular value decomposition of **(10 Marks)**

$$A = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}$$ *(Ans. : Refer section 6.4)*

Multiple Choice Questions

Q. 6.1 Which of the following method is a dimensionality reduction method?

(a) Principal Component Analysis
(b) Regression
(c) Classification
(d) Clustering ✓ **Ans. : (a)**

Explanation : PCA is a DR technique whereas regression, classification are supervised learning examples and clustering is example of unsupervised learning.

Q. 6.2 Which of the following property is true for PCA Algorithm ?

(a) Data used for PCA is having Less variance
(b) Maximum number of principal components are greater than number of features
(c) All principal components are orthogonal to each other
(d) PCA is a Supervised learning method **Ans. : (c)**

Explanation : Property of PCA algorithm.

Q. 6.3 Single Value Decomposition(SVD) is which type of method?

(a) Regression (b) Classification
(c) Clustering Dimension Reduction ✓ **Ans. : (d)**

Explanation : SVD is used to decompose matrix in to its components.

Q. 6.4 If eigenvalues are roughly equal then ____

(a) PCA will perform outstandingly
(b) PCA will perform badly
(c) LDA will perform outstandingly
(d) LDA will perform badly ✓ **Ans. : (b)**

Explanation : When all eigen vectors are same in such case you won't be able to select the principal components because in that case all principal components are equal.

Q. 6.5 In PCA principal component is a eigen vector for which eigen value is ____

(a) maximum (b) minimum
(c) zero (d) one ✓ **Ans. : (a)**

Explanation : Property of eigenvalue and eigenvector.

Q. 6.6 Feature selection is a process of ____

(a) selecting best k features from original d features
(b) extracting any k features from original d features
(c) selecting any k features from original d features
(d) extracting best k features from original d features ✓ **Ans. : (a)**

Explanation : Out of all features best features are selected.

Q. 6.7 Out of following which method is used for dimension reduction?

(a) Classification (b) Clustering
(c) Regression
(d) Independent component analysis(ICA) ✓ **Ans. : (d)**

Explanation : ICA is a DR technique whereas regression, classification are supervised learning examples and clustering is example of unsupervised learning.

Q. 6.8 Dimensionality reduction algorithm ____

(a) Reduce Time complexity
(b) Increase Memory complexity
(c) Increase Time Complexity
(d) Increase overfitting problem ✓ **Ans. : (a)**

Explanation : Since data dimension is reduced, time required to process the data will also be reduced.

Q. 6.9 Which following is an example of Supervised dimensionality reduction algorithm ?

(a) Naïve Bayes (b) SVM
(c) PCA (d) LDA ✓ **Ans. : (d)**

Module 6

Explanation : Naïve bayes and SVM are used for classification. PCA is unsupervised whereas LDA is supervised.

Q. 6.10 Which of the following algorithms cannot be used for reducing the dimensionality of data ?

(a) t-SNE (b) PCA

(c) LDA (d) Random Forest ✓ **Ans. : (d)**

Explanation : Random Forest algorithm is used for classification.

Q. 6.11 Which following is an example of Unsupervised dimensionality reduction algorithm ?

(a) Naïve Bayes (b) SVM

(c) PCA (d) LDA ✓ **Ans. : (c)**

Explanation : Naïve bayes and SVM are used for classification. PCA is unsupervised whereas LDA is supervised.

Q. 6.12 The key difference between feature selection and extraction is ____

(a) feature selection keeps a subset of the original features while feature extraction creates brand new ones.

(b) feature selection keeps original features while feature extraction creates brand new ones.

(c) feature selection keeps original features while feature extraction keeps subset of original ones.

(d) feature selection creates new brand features while feature extraction keeps original ones.

✓ **Ans. : (a)**

Explanation : Out of all features best features are selected in feature selection whereas new features are calculated from original features in feature extraction.

Q. 6.13 Which of the following statement is correct in Dimensionality reduction?

(a) Removing columns with dissimilar data trends

(b) Removing columns which have high variance in data

(c) Removing columns with dissimilar data trends

(d) Removing columns which have too many missing values ✓ **Ans. : (d)**

Explanation : Coloumns which have too many missing values are not important for processing hence they are removed.

Q. 6.14 Which of the following Property of LDA and PCA is true?

(a) PCA maximize the variance of the data, whereas LDA maximize the separation between different class

(b) LDA is Unsupervised whereas PCA is supervised

(c) PCA minimize the variance of the data, whereas LDA minimize the separation between different classes

(d) Both LDA and PCA are linear transformation techniques ✓ **Ans. : (d)**

Explanation : PCA and LDA transform data to a new coordinate system.

Q. 6.15 What happens when you get features in lower dimensions using PCA?

(a) The features will still have interpretability

(b) The features will lose interpretability

(c) The features must carry all information present in data

(d) The features carry all information present in data

✓ **Ans. : (b)**

Explanation : When you get the features in lower dimension then you will lose some information of data most of the times and you won't be able to interpret the lower dimension data.

Q. 6.16 PCA is a deterministic algorithm because, ____

(a) it doesn't have local minima problem

(b) You need to initialize parameters in PCA

(c) PCA can be trapped into local minima problem

(d) PCA can be trapped into global minima problem

✓ **Ans. : (a)**

Explanation : PCA is a deterministic algorithm which does not have parameters to initialize and it does not have local minima problem.

Q. 6.17 When you are applying PCA on a image dataset ____

(a) It can be used to effectively detect deformable objects

(b) It is invariant to affine transforms.

(c) It can be used for lossy image compression

(d) It is invariant to shadows. ✓ **Ans. : (c)**

Explanation : When we apply PCA on image some information may be lost.

Q. 6.18 To produce the same projection result in SVD and PCA which following condition has to satisfy ?

(a) When data has zero median

(b) When data has zero mean

(c) Both are always same

(d) Both are not always same ✓ **Ans. : (b)**

Explanation : Mean of zero is needed for finding a basis that minimizes the mean square error of approximation of data in both SVD and PCA.

Q. 6.19 Which of the following statement is true?

(a) LDA explicitly attempts to model the difference between the classes of data.

(b) Both attempt to model the difference between the classes of data.

(c) PCA explicitly attempts to model the difference between the classes of data.

(d) LDA on the other hand does not take into account any difference in class. ✓ **Ans. : (a)**

Explanation : Property of LDA.

Q. 6.20 Which of the following offset, do we consider in PCA?

(a) Vertical (b) Perpendicular offset

(c) Horizontal offset (d) Parallel offset

✓ **Ans. : (b)**

Explanation : Eigen vectors generated in PCA are orthogonal to each other.

Q. 6.21 Which of the following necessitates feature reduction in machine learning?

(a) Irrelevant and redundant features

(b) Limited training data

(c) Limited computational resources.

(d) All of the above ✓ **Ans. : (d)**

Explanation : If we are having limited training data, computational resources and presence of irrelevant and redundant features then we have to apply dimension reduction techniques.

Q. 6.22 What are the optimum number of principal components in the below Fig. Q. 6.22 ?

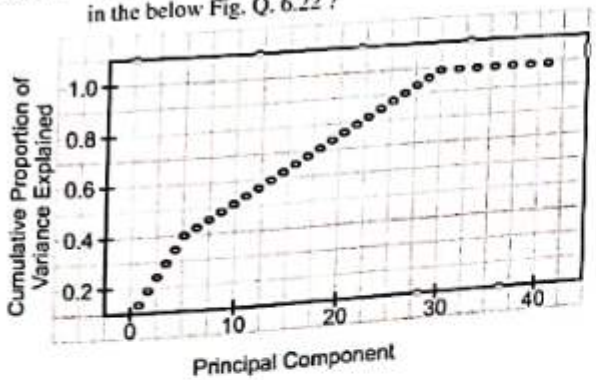


Fig. Q. 6.22

(a) 10 (b) 20 ✓ **Ans. : (c)**

(c) 30 (d) 40

Explanation : We can see in the figure that after 30 principal components the curve remains the constant.

Q. 6.23 Suppose we are using dimensionality reduction as preprocessing technique, i.e., instead of using all the features, we reduce the data to k dimensions with PCA. And then use these PCA projections as our features. Which of the following statements is correct ? Choose which of the options is correct?

(a) Higher value of 'k' means more regularization

(b) Higher value of 'k' means less regularization

(c) Both of above

(d) None of above ✓ **Ans. : (b)**

Explanation : Higher k would lead to less smoothening as we would be able to preserve more characteristics in data, hence less regularization.

Q. 6.24 When performing regression or classification, which of the following is the correct way to preprocess the data?

(a) Normalize the data → PCA → training PCA → normalize PCA output → training

(b) Normalize the data → PCA → normalize PCA output → training

(c) None of the above

(d) All of above **Ans. : (a)**

Explanation : First we normalize the data then we apply dimension reduction method. After that we train PCA and normalize the eigen vectors. Finally we go for training.

Q. 6.25 Which of the following is an example of feature extraction?

(a) Constructing bag of words vector from an email

(b) Applying PCA projects to a large high-dimensional data

(c) Removing stopwords in a sentence

(d) All of the above ✓ **Ans. : (d)**

Explanation : All of above methods are used for feature extraction.

Q. 6.26 What is PCA components in Sklearn?

(a) Set of all eigen vectors for the projection space

(b) Matrix of principal components

(c) Result of the multiplication matrix

(d) None of the above options ✓ **Ans. : (a)**

Explanation : In Sklearn (python) PCA components are represented by set of all eigen vectors for the projection space.

Q. 6.27 Which of the following is a reasonable way to select the number of principal components "k"?

(a) Choose k to be the smallest value so that at least 99% of the variance is retained.

Module 6

(b) Choose k to be 99% of m (k = 0.99*m, rounded to the nearest integer).

(c) Choose k to be the largest value so that 99% of the variance is retained.

(d) Use the elbow method ✓ **Ans. : (a)**

Explanation : Standard procedure to select the number of principal components.

Q. 6.28 Imagine, you have 1000 input features and 1 target feature in a machine learning problem. You have to select 100 most important features based on the relationship between input features and the target features. Do you think, this is an example of dimensionality reduction?

(a) Yes (b) No ✓ **Ans. : (a)**

Explanation : Since we have to select 100 important features out of 1000 features this is a example of feature selection.

Q. 6.29 It is not necessary to have a target variable for applying dimensionality reduction algorithms.

(a) TRUE (b) FALSE ✓ **Ans. : (b)**

Explanation : LDA is an example of supervised dimensionality reduction algorithm.

Q. 6.30 I have 4 variables in the dataset such as – A, B, C & D. I have performed the following actions :

Step 1 : Using the above variables, I have created two more variables, namely E = A + 3 * B and F = B + 5 * C + D.

Step 2 : Then using only the variables E and F I have built a Random Forest model.

Could the steps performed above represent a dimensionality reduction method?

(a) True (b) False ✓ **Ans. : (a)**

Explanation : Yes, Because Step 1 could be used to represent the data into 2 lower dimensions.

Q. 6.31 Which of the following techniques would perform better for reducing dimensions of a data set?

(a) Removing columns which have too many missing values

(b) Removing columns which have high variance in data

(c) Removing columns with dissimilar data trends

(d) None of these ✓ **Ans. : (a)**

Explanation : If columns have too many missing values, (say 99%) then we can remove such columns.

Q. 6.32 Dimensionality reduction algorithms are one of the possible ways to reduce the computation time required to build a model.

(a) TRUE (b) FALSE ✓ **Ans. : (a)**

Explanation : Reducing the dimension of data will take less time to train a model.

Q. 6.33 Which of the following algorithms cannot be used for reducing the dimensionality of data ?

(a) t-SNE (b) PCA

(c) LDA (d) None of these ✓ **Ans. : (d)**

Explanation : All of the algorithms are the example of dimensionality reduction algorithm.

Q. 6.34 PCA can be used for projecting and visualizing data in lower dimensions.

(a) TRUE (b) FALSE **Ans. : (a)**

Explanation : Sometimes it is very useful to plot the data in lower dimensions. We can take the first 2 principal components and then visualize the data using scatter plot.

Q. 6.35 The most popularly used dimensionality reduction algorithm is Principal Component Analysis (PCA). Which of the following is/are true about PCA?

1. PCA is an unsupervised method
2. It searches for the directions that data have the largest variance
3. Maximum number of principal components <= number of features
4. All principal components are orthogonal to each other

(a) 1 and 2 (b) 1 and 3

(c) 2 and 3 (d) 1, 2 and 3

(e) 1,2 and 4 (f) All of the above ✓ **Ans. : (f)**

Explanation : All options are self explanatory.

Q. 6.36 Which of the following statement is correct for t-SNE and PCA?

(a) t-SNE is linear whereas PCA is non-linear

(b) t-SNE and PCA both are linear

(c) t-SNE and PCA both are nonlinear

(d) t-SNE is nonlinear whereas PCA is linear

✓ **Ans. : (d)**

Explanation : t-SNE applies non linear transformation whereas PCA applies linear transformation.

Q. 6.37 What is of the following statement is true about t-SNE in comparison to PCA?

(a) When the data is huge (in size), t-SNE may fail to produce better results.

(b) T-NSE always produces better result regardless of the size of the data

(c) PCA always performs better than t-SNE for smaller size data.

(d) None of these ✓ **Ans. : (a)**

Explanation : As per working of method.

Q. 6.38 Xi and Xj are two distinct points in the higher dimension representation, where as Yi & Yj are the representations of Xi and Xj in a lower dimension.

1. The similarity of datapoint Xi to datapoint Xj is the conditional probability p (j|i)
2. The similarity of datapoint Yi to datapoint Yj is the conditional probability q (j|i)

Which of the following must be true for perfect representation of xi and xj in lower dimensional space?

(a) p (j|i) = 0 and q (j|i) = 1

(b) p (j|i) < q (j|i)

(c) p (j|i) = q (j|i)

(d) p (j|i) > q (j|i) ✓ **Ans. : (c)**

Explanation : The conditional probabilities for similarity of two points must be equal because similarity between the points must remain unchanged in both higher and lower dimension for them to be perfect representations.

Q. 6.39 Which of the following comparison(s) are true about PCA and LDA?

Both LDA and PCA are linear transformation techniques

LDA is supervised whereas PCA is unsupervised.

PCA maximize the variance of the data, whereas LDA maximize the separation between different classes,

(a) 1 and 2 (b) 2 and 3

(c) 1 and 3 (d) Only 3

(e) 1, 2 and 3 ✓ **Ans. : (e)**

Explanation : All of the options are correct

Q. 6.40 PCA works better if there is?

1. A linear structure ix the data
2. If the data lies on a curved surface and not on a flat surface
3. If variables are scaled in the same unit

(a) 1 and 2 (b) 2 and 3

(c) 1 and 3 (d) 1,2 and 3 ✓ **Ans. : (c)**

Explanation : Option c is correct.

Q. 6.41 Imagine, you are given the following scatter plot between height and weight.

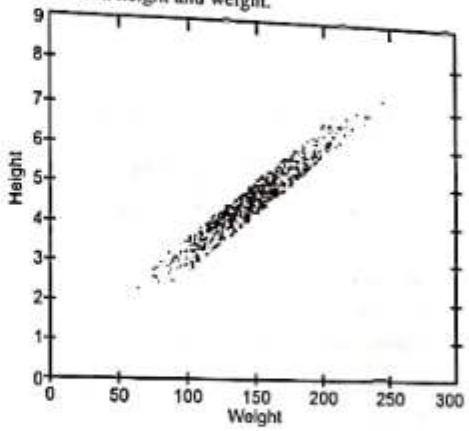


Fig. Q. 6.41

Select the angle which will capture maximum variability along a single axis?

(a) ~ 0 degree (b) ~ 45 degree

(c) ~ 60 degree (d) ~ 90 degree ✓ **Ans. : (b)**

Explanation : Option b has largest possible variance in data.

Q. 6.42 The below snapshot shows the scatter plot of two features (X1 and X2) with the class information (Red, Blue). You can also see the direction of PCA and LDA.

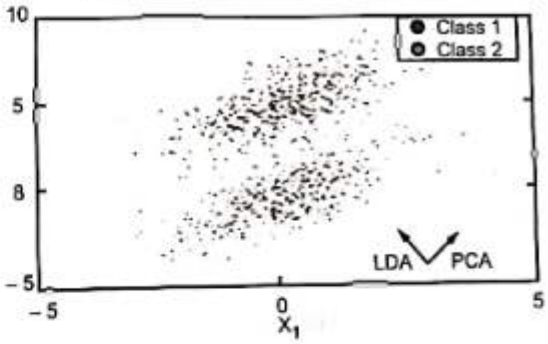


Fig. Q. 6.42

Which of the following method would result into better class prediction?

(a) Building a classification algorithm with PCA (A principal component in direction of PCA)

(b) Building a classification algorithm with LDA

(c) Can't say

(d) None of these ✓ **Ans. : (b)**

Explanation : If our goal is to classify these points, PCA projection does only more harm than good. The majority of blue and red points would land overlapped on the first

principal component. Hence PCA would confuse the classifier.

Q. 6.43 Which of the following options are correct, when you are applying PCA on a image dataset?

1. It can be used to effectively detect deformable objects.
2. It is invariant to affine transforms.
3. It can be used for lossy image compression.
4. It is not invariant to shadows.

(a) 1 and 2 (b) 2 and 3
(c) 3 and 4 (d) 1 and 4 ✓ **Ans. : (c)**

Explanation : Option c is correct

Q. 6.44 Under which condition SVD and PCA produce the same projection result?

(a) When data has zero median
(b) When data has zero mean
(c) Both are always same
(d) None of these ✓ **Ans. : (b)**

Explanation : When the data has a zero mean vector, otherwise you have to center the data first before taking SVD.

Q. 6.45 Consider 3 data points in the 2-d space : (– 1, – 1), (0,0), (1,1).

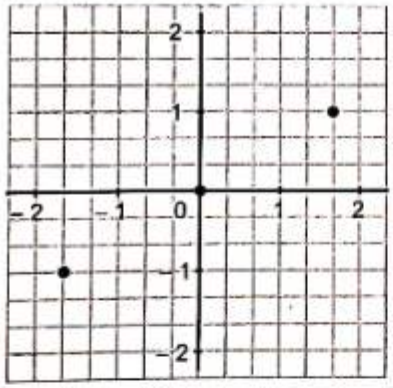


Fig. Q. 6.45

What will be the first principal component for this data?

1. $\left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right]$ 2. $\left[\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right]$
3. $\left[\frac{-\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right]$ 4. $\left[\frac{-1}{\sqrt{3}}, \frac{-1}{\sqrt{3}}\right]$

(a) 1 and 2 (b) 3 and 4
(c) 1 and 3 (d) 2 and 4 ✓ **Ans. : (c)**

Explanation : The first principal component is $v = \left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right]^T$ (you shouldn't really need to solve any SVD or eigen problem to see this). Note that the principal component should be normalized to have unit length. (The negation $v = \left[\frac{-\sqrt{2}}{2}, \frac{-\sqrt{2}}{2}\right]^T$ is also correct.)

Q. 6.46 If we project the original data points into the 1-d subspace by the principal component $\left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right]^T$. What are their coordinates in the 1-d subspace?

(a) $\sqrt{-2}, (0), \sqrt{2}$ (b) $\sqrt{2}, (0), \sqrt{2}$
(c) $\sqrt{2}, (0), \sqrt{-2}$ (d) $\sqrt{-2}, (0), \sqrt{-2}$

✓ **Ans. : (a)**

Explanation : The coordinates of three points after projection should be

$z_1 = x\,T_1\,v = [-1, -1]\left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right]^T = \sqrt{-2}, z_2$

$= x\,T_1\,v = 0, z_3 = x\,T_3\,v = \sqrt{2}$

Q. 6.47 For the projected data you just obtained projections $\sqrt{-2}$, (0), $\sqrt{2}$. Now if we represent them in the original 2-d space and consider them as the reconstruction of the original data points, what is the reconstruction error ? (Context : 45-47)

(a) 0% (b) 10% (c) 30% (d) 40% ✓ **Ans. : (a)**

Explanation : The reconstruction error is 0, since all three points are perfectly located on the direction of the first principal component. Or, you can actually calculate the reconstruction : $z_1 \cdot v$.

$\hat{x_1} = -\sqrt{2} \cdot \left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right]^T = [-1, -1]^T$

$\hat{x_2} = 0 * [0, 0]^T = [0,0]$

$\hat{x_3} = \sqrt{2} * [1, 1]^T = [1,1]$

which are exactly x_1, x_2, x_3.

Q. 6.48 Which above graph shows better performance of PCA? Where M is first M principal components and D is total number of features.

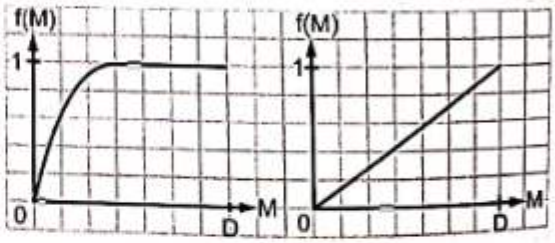


Fig. Q. 6.48

(a) Left
(b) Right
(c) Any of A and B
(d) None of these

✓ **Ans. : (a)**

Explanation : PCA is good if f(M) asymptotes rapidly to 1. This happens if the first eigenvalues are big and the remainder are small. PCA is bad if all the eigenvalues are roughly equal. See examples of both cases in Fig. Q. 6.48.

Q. 6.49 Which of the following can be the first 2 principal components after applying PCA?

(0.5, 0.5, 0.5, 0.5) and (0.71, 0.71, 0, 0)

(0.5, 0.5, 0.5, 0.5) and (0, 0, – 0.71, – 0.71)

(0.5, 0.5, 0.5, 0.5) and (0.5, 0.5, – 0.5, – 0.5)

(0.5, 0.5, 0.5, 0.5) and (– 0.5, – 0.5, 0.5, 0.5)

(a) 1 and 2 (b) 1 and 3

(c) 2 and 4 (d) 3 and 4 ✓ **Ans. : (d)**

Explanation : For the first two choices, the two loading vectors are not orthogonal.

Q. 6.50 Which of the following offset, do we consider in PCA?

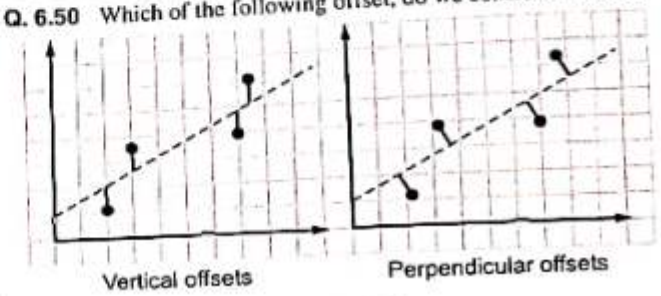


Fig. Q. 6.50

(a) Vertical offset

(b) Perpendicular offset

(c) Both

(d) None of these ✓ **Ans. : (b)**

Explanation : We always consider residual as vertical offsets. Perpendicular offset are useful in case of PCA.

Chapter Ends...

❑❑❑