ANNAMACHARYA UNIVERSITY, RAJAMPET

(ESTD UNDER AP PRIVATE UNIVERSITIES (ESTABLISHMENT AND REGULATION) ACT, 2016
RAJAMPET, Annamayya District, AP, INDIA

Course : Machine Learning

Course Code : 24FMCA032T

Branch : MCA

Prepared by : D. Siva Sanjeev Kumar

Designation : Assistant Professor

Department : Computer Applications

ANNAMACHARYA UNIVERSITY, RAJAMPET

(ESTD UNDER AP PRIVATE UNIVERSITIES (ESTABLISHMENT AND REGULATION) ACT, 2016
RAJAMPET, Annamayya District, AP, INDIA

Title of the Course : Machine Learning
Category : PC
Course Code : 24FMCA032T
Branch : MCA
Semester : III Semester

Lecture Hours	Tutorial Hours	Practice Hours	Credits
3	0	0	3

COURSE OBJECTIVES

- Understand the fundamentals of machine learning.
- Learn how to select, train, evaluate, and improve machine learning models.
- Grasp the concepts of Bayesian learning and common supervised classification algorithms.
- Develop a solid understanding of various supervised regression algorithms.
- Explore and implement unsupervised learning techniques.

UNIT I **11 Hrs**

INTRODUCTION TO MACHINE LEARNING: Introduction, What is Human Learning?, Types of Human Learning, What is Machine Learning?, Types of Machine Learning, Applications of Machine Learning, State-of-The-Art Languages/Tools in Machine Learning.

PREPARING TO MODEL: Introduction, Machine Learning Activities, Basic Types of Data in Machine Learning, Exploring Structure of Data, Data Quality and Remediation, Data Pre-Processing.

UNIT II **11 Hrs**

MODELING AND EVALUATION: Introduction, Selecting a Model, Training a Model (for Supervised Learning):Hold-out Method, k-fold Cross-validation Method, Model Representation and Interpretability, Evaluating Performance of a Model: Supervised Learning–Classification, Supervised Learning – Regression, Unsupervised Learning– Clustering, Improving Performance of a Model

BASICS OF FEATURE ENGINEERING: Introduction, Feature Transformation: Feature Construction, Feature Extraction, Feature Subset Selection: Issues in High-Dimensional Data, Key Drivers of Feature Selection – Feature Relevance and Redundancy, Measures of Feature Relevance and Redundancy, Overall Feature Selection Process, Feature Selection Approaches.

UNIT III **10 Hrs**

BAYESIAN CONCEPT LEARNING: Introduction, Bayes' Theorem: Prior, Posterior, Likelihood, Bayes' Theorem and Concept Learning: Concept of Consistent Learners, Bayes Optimal Classifier, Naïve Bayes Classifier, Applications of Naïve Bayes Classifier.

SUPERVISED LEARNING: CLASSIFICATION: Introduction, Classification Model, Classification Learning Steps, Common Classification Algorithms: k-Nearest Neighbor (k-NN), Decision Tree, Random Forest Model.

UNIT IV **10 Hrs**

SUPERVISED LEARNING: REGRESSION: Introduction, Example of Regression, Common Regression Algorithms: Simple Linear Regression, Multiple Linear Regression, Polynomial Regression Model, Logistic Regression.

UNIT V **9 Hrs**

UNSUPERVISED LEARNING: Introduction, Clustering: Clustering as a Machine Learning Task, Different Types of Clustering Techniques, Partitioning Methods, Hierarchical Clustering, Finding Pattern using Association Rule: Definition of Common Terms, Association Rule, The Apriority Algorithm for Association Rule Learning.

TEXT BOOK:

1. Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das, Machine Learning, Pearson Education, 2019.

REFERENCE BOOKS:

1.Tom M.Mitchell, Machine Learning, McGraw Hill Education, Indian Edition, 2019.
2.Ethern Alpayd in, Introduction to Machine Learning, MIT Press,3 rd Edition, 2014.
3.Stephen Marsland, Machine Learning-An Algorithmic Perspective, CRC Press, 2 nd Edition 2015.

COURSE OUTCOMES:

The Student will be able to

1. Comprehend the fundamental concepts of Machine Learning.
2. Describe the process of model selection and evaluation of process model.
3. Apply the concepts of Bayes' theorem and supervised classification algorithms.
4. Analyze various supervised regression algorithms.
5. Analyze various unsupervised algorithms including clustering.

CO-PO MAPPING:

Course Outcomes	Foundation Knowledge	Problem Analysis	Development of Solutions	Modern Tool Usage	Individual and Teamwork	Project Management and Finance	Ethics	Life-long Learning
24FMCA032T.1	2	2	1	-	-	-	-	1
24FMCA032T.2	2	2	1	-	-	-	-	-
24FMCA032T.3	3	2	1	-	-	-	-	-
24FMCA032T.4	3	3	2	-	-	-	-	-
24FMCA032T.5	3	3	2	-	-	-	-	-

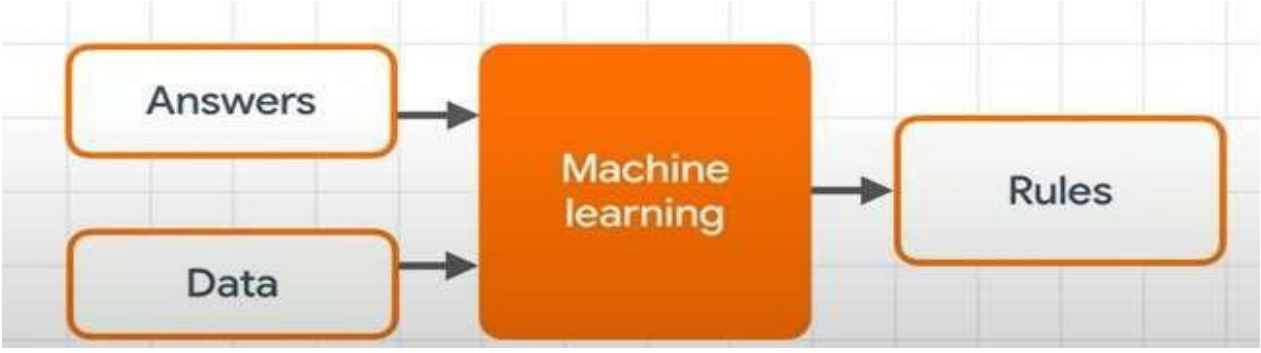
UNIT – I

Introduction: What is Human Learning? Types of Human Learning, what is Machine Learning? Types of Machine Learning, Problems Not to Be Solved Using Machine Learning, Applications of Machine Learning, State-of-The-Art Languages/Tools in Machine Learning, Issues in Machine Learning.

Preparing to Model: Introduction, Machine Learning Activities, Basic Types of Data in Machine Learning, Exploring Structure of Data, Data Quality and Remediation, Data Pre- Processing

---

Traditional programming vs Machine Learning



Human Learning and Importance :-

Learning is typically referred to as the process of gaining information through observation. To do a task in a proper way, we need to have prior information on one or more things related to the task. Also, as we keep learning more or in other words acquiring more information, the efficiency in doing the tasks keep improving.

For example, with more knowledge, the ability to do homework with less number of mistakes increases. In the same way, information from past rocket launches helps in taking the right precautions and makes more successful rocket launch.

TYPES OF HUMAN LEARNING :-

Human Learning happens in one of the three ways –

(1) Learning under expert guidance
(2) Learning guided by knowledge gained from experts
(3) Learning by self

(1) Learning under expert guidance:-

In all phases of life of a human being, there is an element of guided learning. This learning is imparted by someone, purely because of the fact that he/she has already gathered the knowledge by virtue of his/her experience in that field. So guided learning is the process of gaining information from a person having sufficient knowledge due to the past experience.

Example1: In school, baby starts with basic familiarization of alphabets and digits. Then the baby learns how to form words from the alphabets and numbers from the digits. Slowly more complex learning happens in the form of sentences, paragraphs Learning, complex mathematics, science, etc. The baby is able to learn all these things from his teacher who already has knowledge on these areas.

Example2: A grown-up kid can select one odd word from a set of words because it is a verb and other words being all nouns. He could do this because of his ability to label the words as verbs or nouns, taught by his English teacher long back.

(2) Guided by knowledge gained from experts

An essential part of learning also happens with the knowledge which has been imparted by teacher or mentor at some point of time in some other form/context.

In this method, there is NO direct learning. It is some past information shared on some different context, which is used as a learning to make decisions.

Example1: a baby can group together all objects of same color even if his parents have not specifically taught him to do so. He is able to do so because at some point of time or other his parents have told him which colour is blue, which is red, which is green, etc.

Example2: A grown-up kid can select one odd word from a set of words because it is a verb and other words being all nouns. He could do this because of his ability to label the words as verbs or nouns, taught by his English teacher long back.

(3) Learning by Self :-

In many situations, humans are left to learn on their own.

✓A classic example is a baby learning to walk through obstacles. He bumps on to obstacles and falls down multiple times till he learns that whenever there is an obstacle, he needs to cross over it.

✓He faces the same challenge while learning to ride a cycle as a kid or drive a car as an adult. Not all things are taught by others. A lot of things need to be learnt only from mistakes made in the past.

We tend to form a check list on things that we should do, and things that we should not do, based on our experiences.

MACHINE LEARNING:-

Definition 1: Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Definition 2: Tom M. Mitchell, Professor of Machine Learning Department, School of Computer Science, Carnegie Mellon University has defined machine learning as

'A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.'

From the above definition, *a machine can be considered to learn* if it is able to gather experience by doing a certain task and improve its performance in doing the similar tasks in the future. The past experience means past data related to the task. This data is an input to the machine from some source.

Example1: In context of *image classification*,

- ✓ E represents the past data with images having labels or assigned classes (for example whether the

image is of a class cat or a class dog or a class elephant etc.),

- ✓ T is the task of assigning class to new, unlabelled images and
- ✓ P is the performance measure indicated by the percentage of images correctly Classified.

Example2: In context of *the learning to play checkers*,

- ✓ E represents the experience of playing the game,
- ✓ T is the task of playing checkers and
- ✓ P is the performance measure indicated by the percentage of games won by the player.

Type of problems to be solved using Machine Learning

The problems related to

(1) forecast

(2) prediction

(3) analysis of a trend

(4) understanding the different segments or groups of objects, etc.

Type of problems NOT to be solved using Machine Learning

Machine Learning should NOT be applied to

1) tasks in which humans are very effective
2) tasks in which frequent human intervention is needed (Ex: Air traffic control)
3) tasks that are very simple which can be implemented using traditional programming paradigms (Ex: Price calculator engine)
4) the situations where training data is NOT sufficient.

How do machines learn (Process of Machine Learning)?

The basic machine learning process can be divided into three parts.

1. Data Input: Past data or information is utilized as a basis for future decision-making

2. Abstraction (Training the Model): The input data is represented in a broader way through the underlying algorithm

3. Generalization (Future Decisions/Testing the model for accuracy): The abstracted representation is generalized to form a framework for making decisions

Process of machine learning

Explanation

1. Data Input

a. During the machine learning process, knowledge is fed in the form of input data. The vast pool of knowledge is available from the data input.

b. However, the data cannot be used in the original shape and form.

2. Abstraction

a. Machine will perform knowledge abstraction based on the input data. This is called model – it is the summarized knowledge representation of the raw data.

b. The model may be in any one of the following forms –

i. Computational blocks like if/else rules

ii. Mathematical equations

iii. Specific data structures like trees or graphs

iv. Logical groupings of similar observations

Note: The choice of the model used to solve a specific learning problem is a human task. Following are the some of the aspects to be considered for choosing the model –

i. The type of the problem to be solved

ii. Nature of the input data

iii. Domain of the problem

Once the model is choosen, the next task is to fit the model based on the input data. The process of fitting the model based on the input data is known as training. Also, the input data based on which the model is being finalized is known as training data.

3. Generalization

This is the key part and quite difficult to achieve.

In this, we will apply the model to take decision on a set of unknown data, usually called as **test data.**

But, with test data we may encounter two problems –

1. The trained model is aligned with the training data too much, hence may not portray the actual trend.
2. The test data possess certain characteristics apparently unknown to the training data.

Hence, a precise approach of decision making will not work. So, an approximate or heuristic approach, much like gut-feeling-based decision-making in human beings, has to be adopted. This approach has the risk of not making a correct decision.

Define a well-posed learning problem that can be solved using Machine Learning:-

For defining a new problem, which can be solved using machine learning, a simple framework, given below, can be used. This framework also helps in deciding whether the problem is a right candidate to be solved using machine learning. The framework involves answering three questions:

Step 1: What is the problem?

Describe the problem informally and formally and list assumptions and similar problems.

Step 2: Why does the problem need to be solved?

List the motivation for solving the problem, the benefits that the solution will provide and how the solution will be used.

Step 3: How would I solve the problem?

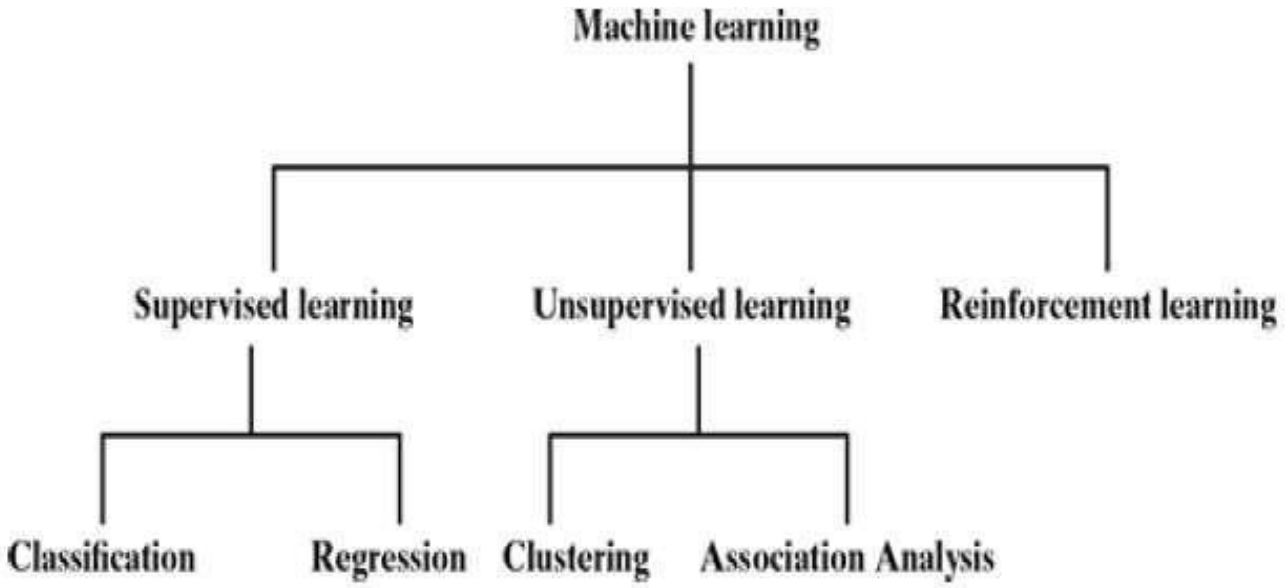
Describe how the problem would be solved manually to flush domain knowledge.

TYPES OF MACHINE LEARNING

Machine learning can be classified into three broad categories:

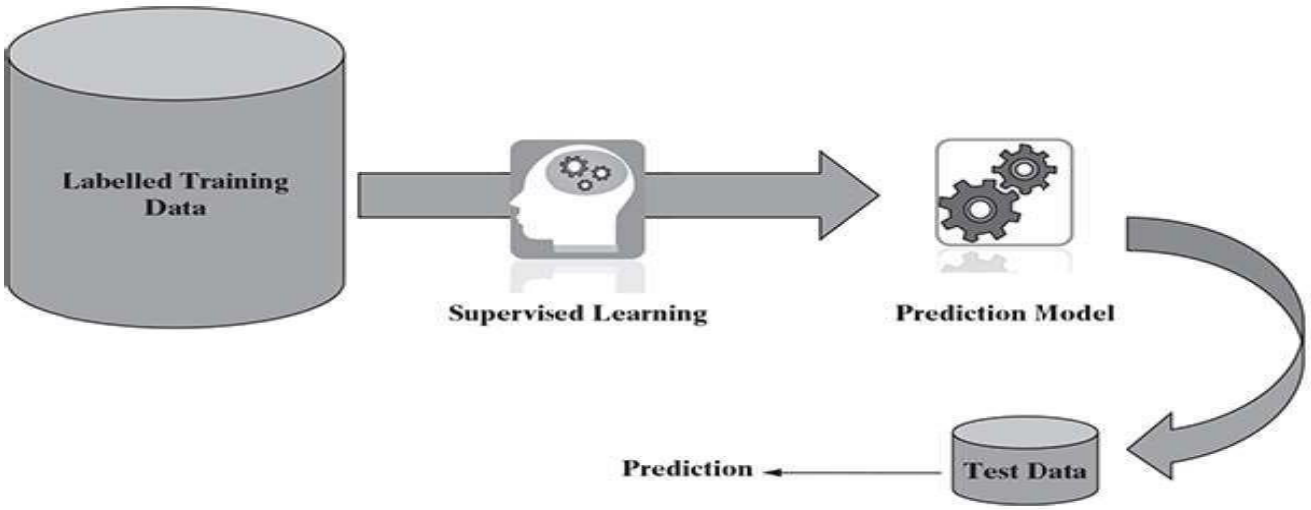
1. **Supervised learning** – Also called **predictive learning**. A machine predicts the class of unknown objects based on prior class related information of similar objects.
2. **Unsupervised learning** – Also called **descriptive learning**. A machine finds patterns in unknown objects by grouping similar objects together.

3. **Reinforcement learning** – A machine learns to act on its own to achieve the given goals.



1) <u>Supervised learning</u>

The major motivation of supervised learning is to learn from past information. It is the information about the task which the machine has to execute. In context of the definition of machine learning, this



past information is the experience.

In supervised learning process,

a) Labeled training data containing past information comes as an input.
b) Based on the training data, the machine builds a predictive model that can be used on test data to assign a label for each record in the test data.

<u>Some examples of supervised learning are</u>

1) Predicting the results of a game

2) Predicting whether a tumor is malignant or benign
3) Predicting the price of domains like real estate, stocks, etc.
4) Classifying texts such as classifying a set of emails as spam or non spam

Types of Supervised Learning :- There are 2 types of Supervised Learning –

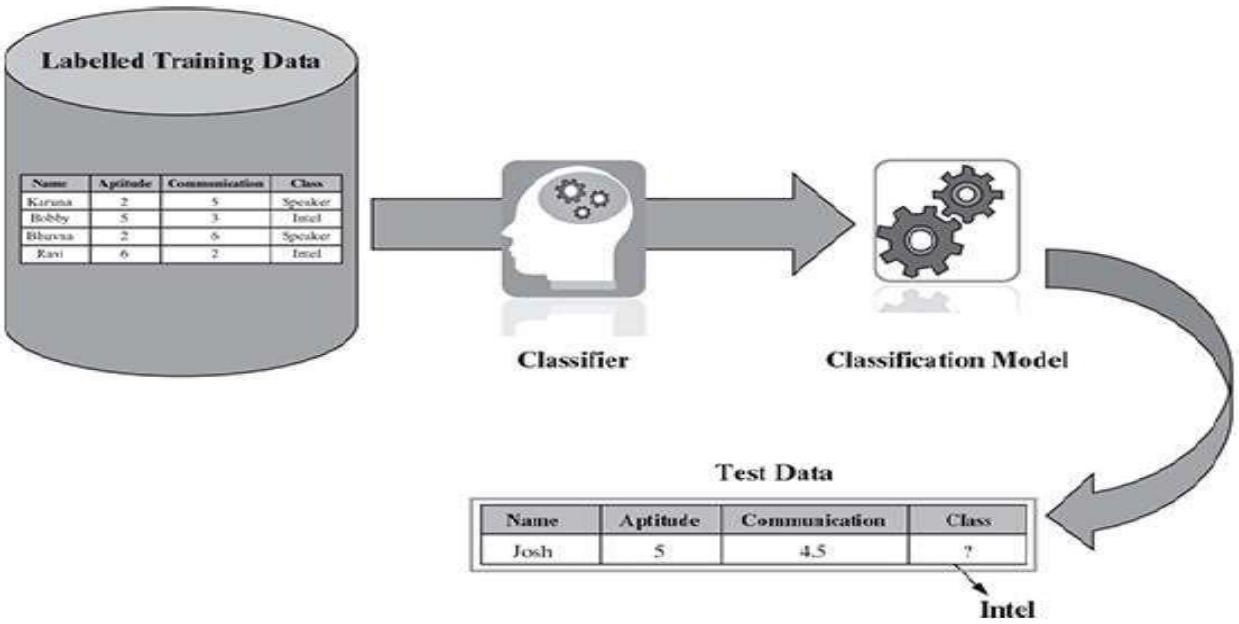
a) Classification
b) Regression

a) Classification

Classification is a type of supervised learning where a target feature, which is of type categorical, is predicted for test data based on the information imparted by training data. The target categorical feature is known as class.

Examples of Typical classification problems

a) Image classification
b) Prediction of disease
c) Win–loss prediction of games
d) Prediction of natural calamity like earthquake, flood, etc.
e) Recognition of handwriting



Few Popular ML algorithms which help in solving classification problems

a) Naïve Bayes

b) Decision Tree

c) K-Nearest Neighbour algorithms

a) Regression

Regression is the process of predicting a continuous value. We can use regression methods to predict a continuous value, such as CO2 emission from a car model, using some other variables.

For example, let us assume that we have access to a dataset that contains data related to the CO2 emissions from different cars. The dataset contains attributes such as car engine size, number of cylinders, fuel consumption and CO2 emission from various automobile models. Now, we are interested in estimating the approximate CO2 emission from a new car model after its production. This is possible using a machine learning regression model.

In regression, there are two types of variables: **a dependent variable** and **one or more independent variables.** The dependent variable is the "state", "target" or "final goal" we study and try to predict, and the independent variables, also known as explanatory variables, are the "causes" of those "states". The independent variables are shown conventionally by X, and the dependent variable is denoted by Y. A regression model relates Y, or the dependent variable, to a function of X, i.e., the independent variables. The key point in regression is that the dependent variable value should be continuous, and not a discrete value. However, the independent variable or variables can be measured on either a categorical or continuous measurement scale.

Types of Regression: Basically, there are 2 types of regression models: simple regression and multiple regression.

Simple regression is when one independent variable is used to estimate a dependent variable. It can be either linear on non-linear.

For example, predicting CO2 emission using the variable 'Engine Size of a Car'. The linearity of regression is based on the nature of the relationship between independent and dependent variables.

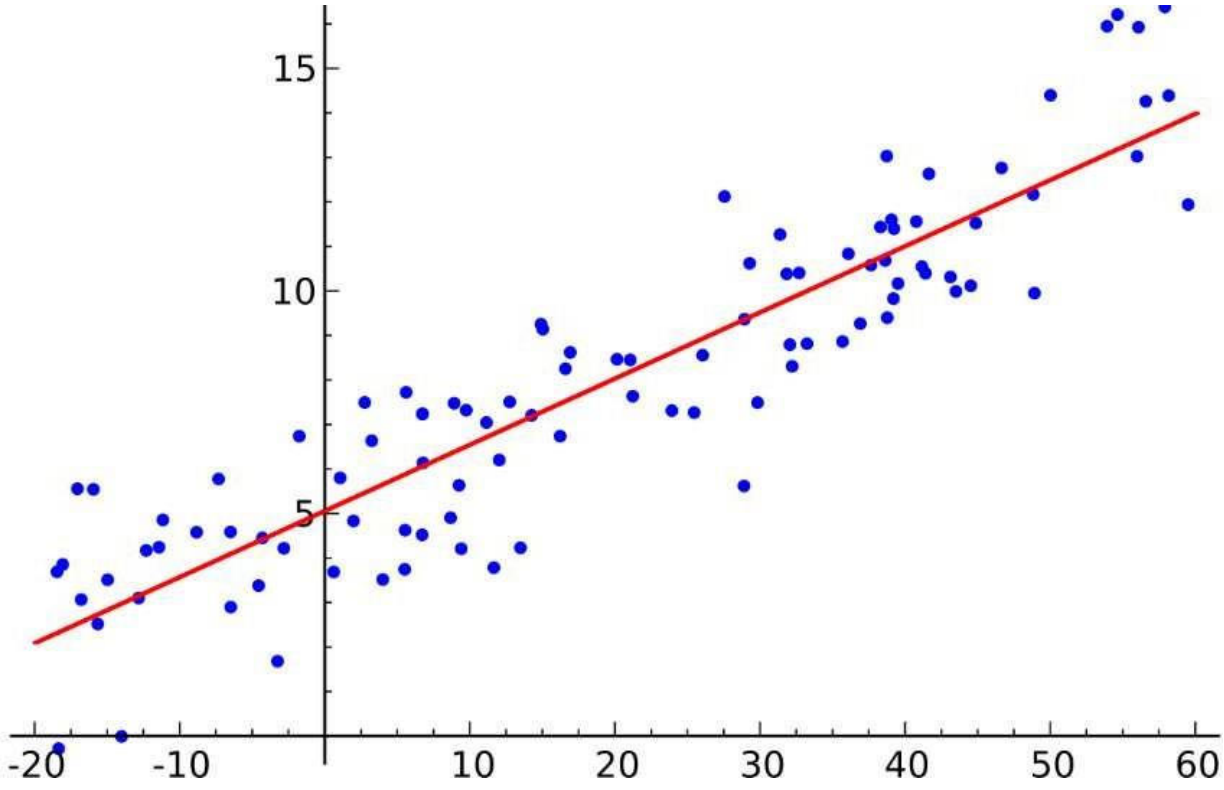
When more than one independent variable is present, the process is called **multiple linear regression**.

For example, predicting CO2 emission using the variables 'Engine Size of a Car' and 'the number of cylinders present in a car'. Again, depending on the relationship between dependent and independent variables, multiple linear regression can be either linear or non-linear regression.

Examples of Typical Regression problems

a) Sales prediction for managers
b) Price prediction in real estate
c) Weather forecast
d) Skill demand forecast in job market

Simple Linear Regression model



2) Unsupervised learning

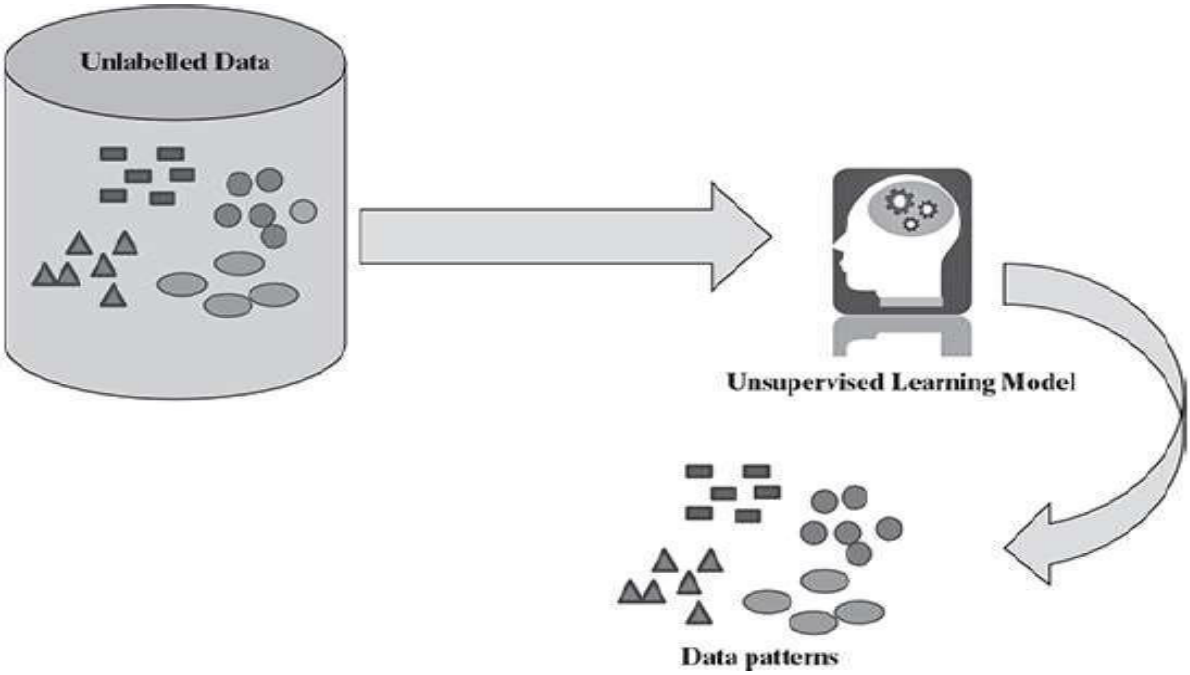
Unlike supervised learning, in unsupervised learning, there is no labeled training data to learn from and no prediction to be made. In unsupervised learning, the objective is to take a dataset as input and try to find natural groupings or patterns within the data elements or records. Therefore, unsupervised learning is often termed as **descriptive model** and the process of unsupervised learning is referred as **pattern discovery** or **knowledge discovery**.

One critical application of unsupervised learning is customer segmentation.

Two types of unsupervised learning are –

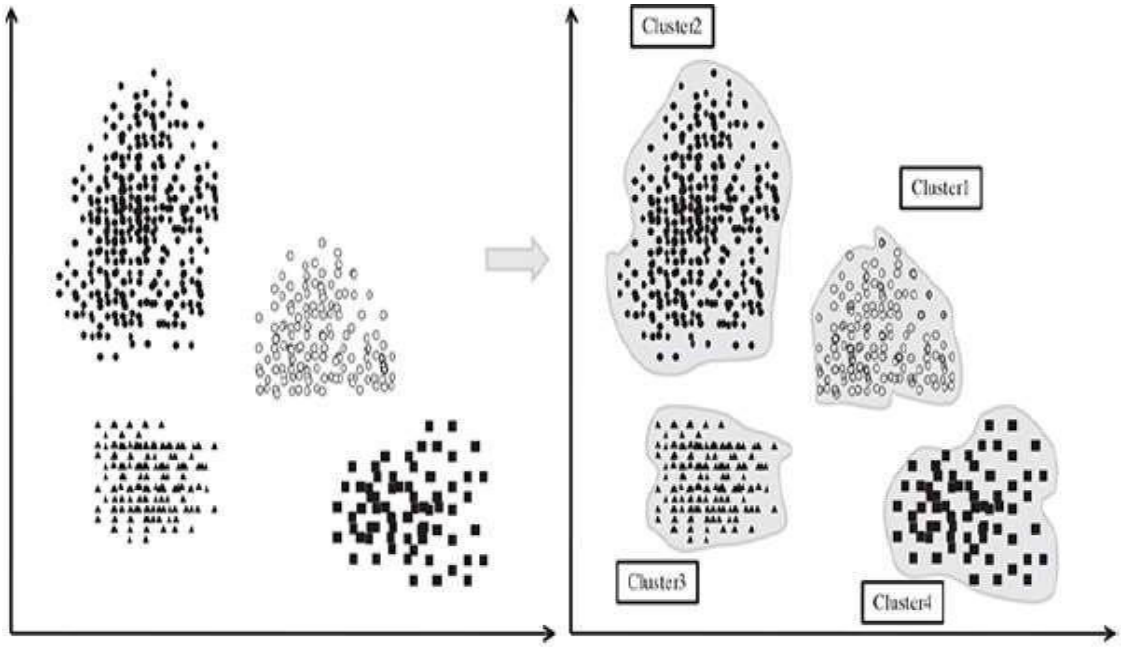
a) Clustering b) Association analysis

Process in Unsupervised learning



a) Clustering

Clustering is the main type of unsupervised learning. It intends to group or organize similar objects together. For that reason, objects belonging to the same cluster are quite similar to each other while objects belonging to different clusters are quite dissimilar.



Distance-based clustering

✓ Different measures of similarity can be applied for clustering. One of the most commonly adopted

similarity measure is **distance**.

✓ Two data items are considered as a part of the same cluster if the distance between them is less.

✓ In the same way, if the distance between the data items is high, the items do not generally belong to the same cluster. This is also known as **distance-based clustering**.

b) **Association analysis**

As a part of association analysis, the association between data elements is identified.

✓ From past transaction data in a grocery store, it may be observed that most of the customers who have bought item A, have also bought item B and item C or at least one of them.

✓ This means that there is a strong association of the event 'purchase of item A' with the event 'purchase of item B', or 'purchase of item C'.

✓ Identifying these sorts of associations is the goal of association analysis. This helps in boosting up sales pipeline, hence a critical input for the sales group.

Critical applications of association analysis include market basket analysis and recommender systems.

TransID	Items Bought
1	{Butter, Bread}
2	{Diaper, Bread, Milk, Beer}
3	{Milk, Chicken, Beer, Diaper}
4	{Bread, Diaper, Chicken, Beer}
5	{Diaper, Beer, Cookies, Ice cream}
...	...

Market Basket transactions

Frequent itemsets → (Diaper, Beer)

Possible association: Diaper → Beer

Market basket analysis

3) Reinforcement learning :-

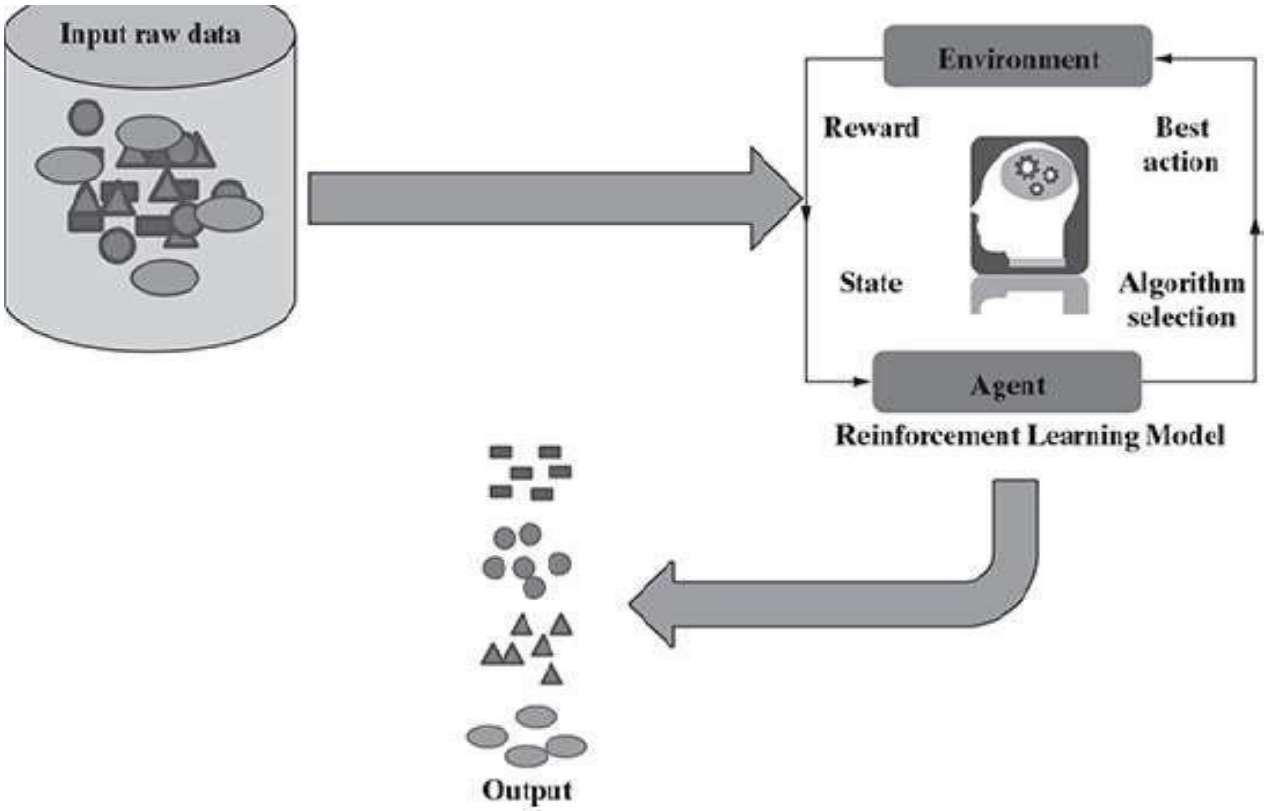
✓ Machines often learn to do tasks autonomously. Let's try to understand in context of the example of the child learning to walk.

✓ The action tried to be achieved is walking, the child is the agent and the place with hurdles on which the child is trying to walk resembles the environment. It tries to improve its performance of doing the task.

✓ When a sub-task is accomplished successfully, a **reward** is given. When a sub-task is not executed correctly, obviously **no reward** is given.

✓ This continues till the machine is able to complete execution of the whole task. This process of learning

✓ known as **reinforcement learning**. One contemporary example of reinforcement learning is **self-driving cars**. The critical information which it needs to take care of are speed and speed limit in different road segments, traffic conditions, road conditions, weather conditions, etc. The tasks that have to be taken care of are start/stop, accelerate/decelerate, turn to left / right, etc.



Differences between SUPERVISED vs UNSUPERVISED vs REINFORCEMENT

Parameters	Supervised	Unsupervised	Reinforcement
When it is used	Used when we know how to classify given data, or in other words classes or labels are available	Used when there is no idea about the class or label of a particular data.	Used when there is no idea about the class or label of a particular data. The model has to find pattern in the data.
Type of work to be performed	The model has to predict the output	The model has to find pattern in the data.	The model has to do the classification - it will get rewarded if the classification is correct, else get punished.
Model building	Labeled training data is needed. Model is built based on training data.	Any unknown and unlabeled data set is given to the model as input and records are grouped.	The model learns and updates itself through reward/punishment.
Model Performance	Performance can be evaluated based on how many misclassifications have been done based on a comparison between predicted and actual values	Difficult to measure whether the model did something useful or interesting. Homogeneity is the only measure.	Model is evaluated by means of the reward function after it had some time to learn
Types	2 Types: 1. classification 2. regression	2 Types: 1. clustering 2. association	No such types
Complexity	Simple one to understand	More difficult to understand and implement than supervised learning	Most complex to understand and apply
Standard algorithms	1. Naïve Bayes 2. K-nearest neighbor (KNN) 3. Decision tree 4. Linear Regression 5. Logistic regression 6. Support vector machine (SVM), etc.	1. K-means 2. Principal Component Analysis (PCA) 3. Self-organizing map (SOM) 4. A priori algorithm 5. DBSCAN, etc.	1. Q-learning 2. Sarsa

Practical Applications	1. Hand writing recognition 2. Stock market prediction 3. Disease prediction 4. Fraud detection, etc.	1. Market basket analysis 2. Recommender systems 3. Customer segmentation, etc.	1. Self-driving cars 2. Intelligent robots 3. AlphaGo Zero (The latest version of DeepMind's AI system playing GO)

Problems that CAN NOT be solved using Machine Learning

❖ Machine learning should not be applied to following tasks -

a) In which humans are very effective or frequent human intervention is needed

Ex: Air traffic control

b) For very simple tasks which can be implemented using traditional programming paradigms

c) For situations where training data is not sufficient, machine learning can not be used effectively.

❖ Machine Learning should be used only when the business process has some lapses.

Applications of Machine Learning

Wherever there is a substantial amount of past data, machine learning can be used to generate actionable insight from the data.

Machine learning is adopted in multiple forms in every business domain.

1) **Banking and Finance**

Following activities will be prevented or reduced using Machine Learning solutions.

a) **Credit card fraudulent transactions**

- ✓ The models work on a real-time basis, i.e. the fraudulent transactions are spotted and prevented right at the time of occurrence. This helps in avoiding a lot of operational hassles in settling the disputes that customers will otherwise raise against those fraudulent transactions.

b) **Reducing the customer churn (attrition rate)**

- ✓ Customers may leave a bank because of
  - o lucrative offers by other competitor banks
  - o poor quality of services
- ✓ Here, both predictive and descriptive models are used to prevent or reduce customer churn.

2) **Insurance**

Insurance industry is extremely data intensive. For that reason, machine learning is extensively used in the insurance industry. Two major areas in the insurance industry where machine learning is used are –

a) **risk prediction during new customer onboarding**

- ✓ During customer onboarding, based on the past information the risk profile of a new customer needs to be predicted. Based on the quantum of risk predicted, the quote is generated for the prospective customer.

b) **claims management.**

When a customer claim comes for settlement, past information related to historic claims along with the adjustor notes are considered to predict whether there is any possibility of the claim to be fraudulent.

3) **Healthcare**

a) Wearable device data form a rich source for applying machine learning and predict the health conditions of the person real time.

- ✓ In case there is some health issue which is predicted by the learning model, immediately the person is alerted to take preventive action.
- ✓ In case of some extreme problem, doctors or healthcare providers in the vicinity of the person can be alerted.
- ✓ **Ex:** Suppose an elderly person goes for a morning walk in a park close to his house. Suddenly, while walking, his blood pressure shoots up beyond a certain limit, which is tracked by the wearable. The wearable data is sent to a remote server and a machine learning algorithm is constantly analysing the streaming data. Alert can be sent to the person to immediately stop walking and take rest. Also, doctors and healthcare providers can be alerted to be on standby.

b) Machine learning along with computer vision also plays a crucial role in disease diagnosis from medical imaging.

STATE-OF-THE-ART LANGUAGES/TOOLS IN MACHINE LEARNING

The algorithms related to different machine learning tasks are known to all and can be implemented

using any language/platform. It can be implemented using a Java platform or C / C++ language or in .NET. However, there are certain languages and tools which have been developed with a focus for implementing machine learning.

a) Python

✓ Python is one of the most popular, open source programming language widely adopted by machine learning community. It was designed by **Guido van Rossum** and was first released in 1991. The reference implementation of Python, i.e.

C Python, is managed by Python Software Foundation, which is a non-profit organization. Python has very strong libraries for advanced mathematical functionalities (numPy), algorithms and mathematical tools (SciPy) and numerical plotting (matplotlib). Built on these libraries, there is a machine learning library named **scikit-learn**, which has various classification, regression, and clustering algorithms embedded in it.

b) R

✓ R is a language for statistical computing and data analysis. It is an open source language, extremely popular in the academic community — especially among statisticians and data miners.

✓ R is a very simple programming language with a huge set of libraries available for different stages of machine learning.

✓ Some of the libraries standing out in terms of popularity are plyr/dplyr (for data transformation), caret ('Classification and Regression Training' for classification), RJava (to facilitate integration with Java), tm (for text mining), ggplot2 (for data visualization).

✓ Other than the libraries, certain packages like Shiny and R Markdown have been developed around R to develop interactive web applications, documents and dashboards on R without much effort.

c) Matlab

✓ MATLAB (matrix laboratory) is a licenced commercial software with a robust support for a wide range of numerical computing.

✓ MATLAB has a huge user base across industry and academia. MATLAB is developed by MathWorks, a company founded in 1984.

✓ Being proprietary software, MATLAB is developed much more professionally, tested rigorously, and has comprehensive documentation.

✓ MATLAB also provides extensive support of statistical functions and has a huge number of machine

learning algorithms in-built. It also has the ability to scale up for large datasets by parallel processing on clusters and cloud.

d) SAS

✓ SAS (earlier known as 'Statistical Analysis System') is another licenced commercial software which provides strong support for machine learning functionalities.

✓ Developed in C by SAS Institute, SAS had its first release in the year 1976

✓ SAS is a software suite comprising different components. The basic data management functionalities are embedded in the Base SAS component whereas the other components like SAS/INSIGHT, Enterprise Miner, SAS/STAT, etc. help in specialized functions related to data mining and statistical analysis.

e) Other languages / tools

i. Owned by IBM, **SPSS** (originally named as Statistical Package for the Social Sciences) is a popular package supporting specialized data mining and statistical analysis. Originally popular for statistical analysis in social science (as the name reflects), **SPSS** is now popular in other fields as well.

ii. Julia is an open source, liberal licence programming language for numerical analysis and computational science. It has baked in all good things of MATLAB, Python, R, and other programming languages used for machine learning for which it is gaining steady attention from machine learning development community. Another big point in favour of Julia is its ability to implement high-performance machine learning algorithms.

ISSUES IN MACHINE LEARNING (Ethical Issues)

✓ Machine learning is a field which is relatively new and still evolving. Also, the level of research and kind of use of machine learning tools and technologies varies drastically from country to country.

✓ The laws and regulations, cultural background, emotional maturity of people differ drastically in different countries. All these factors make the use of machine learning and the issues originating out of machine learning usage are quite different.

✓ The biggest fear and issue arising out of machine learning is related to **privacy and the breach of it**.

✓ The primary focus of learning is on analyzing data, both past and current, and coming up with insight from the data. This insight may be related to people and the facts revealed might be private enough to be kept confidential.

✓ Also, different people have a different preference when it comes to sharing of information. While some people may be open to sharing some level of information publicly, some other people may not want to share it even to all friends and keep it restricted just to family members.

(i) **Breach of Privacy and Sharing information publicly**

Different people have a different preference when it comes to sharing of information. While some people may be open to sharing some level of information publicly, some other people may not want to share it even to all friends and keep it restricted just to family members.

Classic examples are a birth date (not the day, but the date as a whole), photographs of a dinner date with family, educational background, etc. When machine learning algorithms are implemented using those information, inadvertently people may get upset.

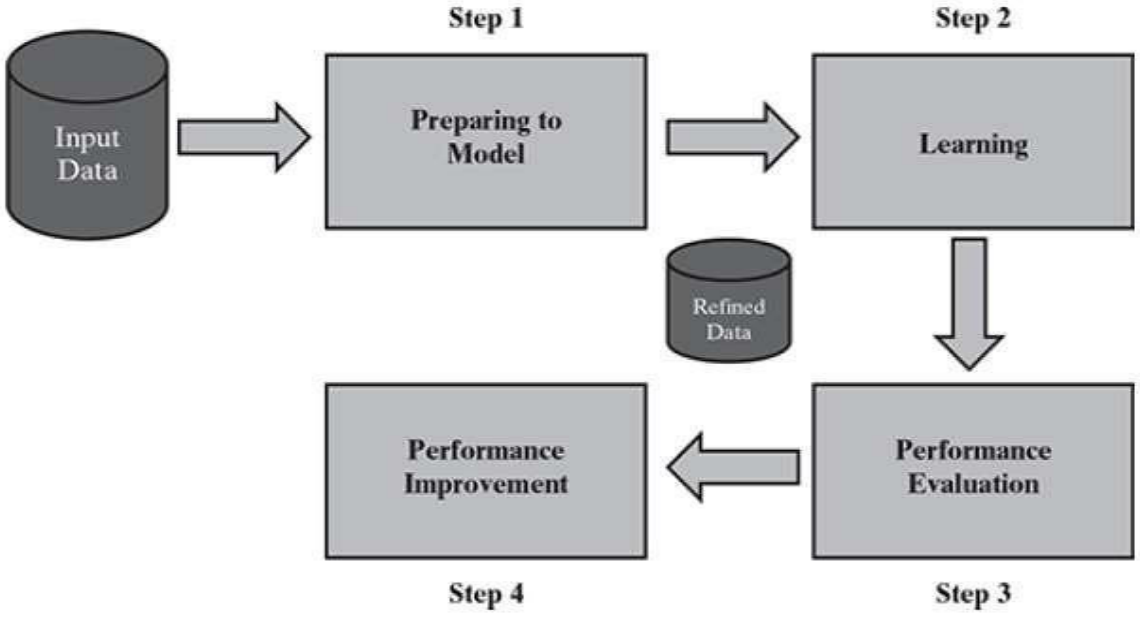
(ii) **Other issues**

Even if there is no breach of privacy, there may be situations where actions were taken based on machine learning may create an adverse reaction.

Example: In the knowledge discovery exercise done before starting an election campaign, If a specific area reveals an ethnic majority or skewness of a certain demographic factor, and the campaign pitch carries a message keeping that in mind, it might actually upset the voters and cause an adverse result.

Machine Learning Activities (OR) Machine Learning Life Cycle

Following figure depicts the four-step process of machine learning.



Detailed process of machine learning

Activities in Machine Learning:

Following are the typical **preparation** activities done once the input data comes into the machine learning system:

1. Understand the type of data in the given input data set.
2. Explore the data to understand the nature and quality.
3. Explore the relationships amongst the data elements, e.g. interfeature relationship.
4. Find potential issues in data.
5. Do the necessary remediation, e.g. impute missing data values, etc., if needed.
6. Apply pre-processing steps, as necessary.
7. Once the data is prepared for modelling, then the learning tasks start off. As a part of it, do the following activities:
    a) The input data is first divided into parts – the training data and the test data (called holdout). This step is applicable for supervised learning only.
    b) Consider different models or learning algorithms for selection.
    c) Train the model based on the training data for supervised learning problem and apply to unknown data.
    d) Directly apply the chosen unsupervised model on the input data for unsupervised learning problem.
8. After the model is selected, trained (for supervised learning), and applied on input data, the performance of the model is evaluated.
9. Based on options available, specific actions can be taken to improve the performance of the model, if possible.

Step#	Step Name	Activities Involved
Step1	Preparing to Model	1. Understand the type of data in the given input data set 2. Explore the data to understand data quality 3. Explore the relationship amongst data elements, Inter-feature relationship 4. Find potential issues in data 5. Remediate data, if needed 6. Apply following pre-processing steps, as necessary: a. Dimensionality reduction b. Feature subset selection

Step2	Learning	1. Data partitioning / holdout 2. Model selection 3. Cross-validation
Step 3	Performance evaluation	1. Examine the model performance, e.g. confusion matrix in case of classification 2. Visualize performance trade-offs using ROC curves
Step 4	Performance Improvement	1. Tuning the model 2. Ensembling 3. Bagging 4. Boosting

BASIC TYPES OF DATA IN MACHINE LEARNING

DATA SET

A data set is a collection of related information or records. The information may be on some entity or some subject area.

Example

Student data set:

Roll Number	Name	Gender	Age
129/011	Mihir Karmarkar	M	14
129/012	Geeta Iyer	F	15
129/013	Chanda Bose	F	14
129/014	Sreenu Subramanian	M	14
129/015	Pallav Gupta	M	16
129/016	Gajanan Sharma	M	15

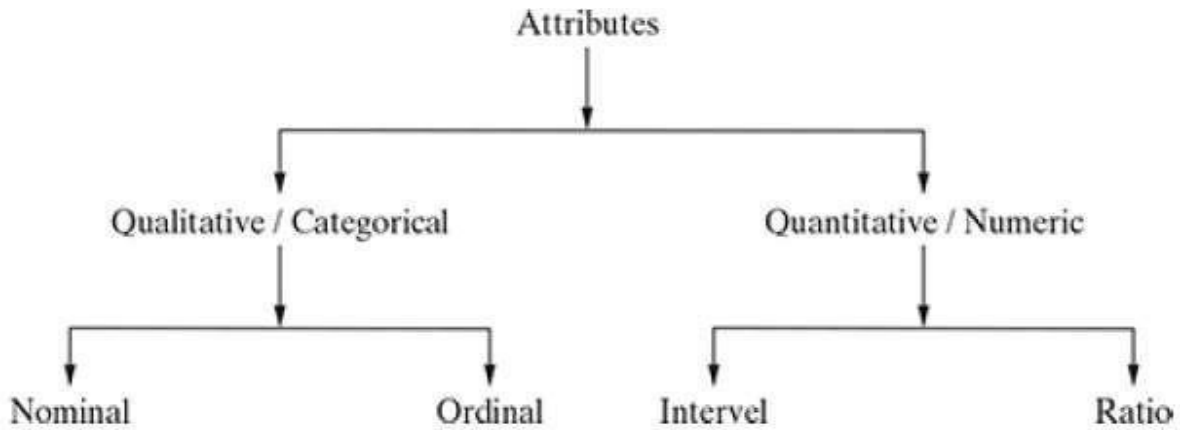
Student performance data set:

Roll Number	Maths	Science	Percentage
129/011	89	45	89.33%
129/012	89	47	90.67%
129/013	68	29	64.67%
129/014	83	38	80.67%
129/015	57	23	53.33%

Each row of a data set is called a **record**. Each data set also has multiple attributes, each

of which gives information on a specific characteristic.

TYPES OF DATA



Data can broadly be divided into following two types:

1. Qualitative data
2. Quantitative data

1. Qualitative data or Categorical data

Qualitative data provides information about the quality of an object or information which cannot be measured.

Ex: a) Quality of performance of students: GOOD, AVERAGE and POOR

b) Name and Roll Number Types

of qualitative data

Qualitative data can be further subdivided into two types as follows:

a. Nominal data
b. Ordinal data

a. Nominal data

Nominal data is one which has no numeric value, but a named value. It is used for assigning named values to attributes. Nominal values cannot be quantified.

Ex: i. **Blood group** : A, B, O, AB, etc.

ii. **Nationality**: India, American, British, etc.

iii. **Gender**: Male, Female, Other

b. Ordinal data

Ordinal data = Nominal data + Natural ordering

Ex: i. Customer satisfaction : Very Happy, Happy, Unhappy

ii. Grades: A,B,C, etc.

iii. Hardness of metal: Very Hard, Hard, Soft, etc.

2. Quantitative data or Numeric data

It relates to information about the quantity of an object – hence it can be measured. Ex: Marks — can be measured using scale of measurement.

Types of quantitative data :- Quantitative data can be further subdivided into two types as follows:

a. Interval data

b. Ratio data

a. Interval data

is numeric data for which not only the order is known, but the exact difference between values is also known.

Ex: Celsius temperature, date and time.

b. Ratio data

Represents numeric data for which exact value can be measured. Ex: height, weight, age, salary, etc.

Other types of attributes

Attributes can also be categorized into 2 types based on a number of values that can be assigned.

a) Discrete Attributes

Discrete attributes can assume a finite or countably infinite number of values.

Ex: Roll number, street number, rank of students, etc.

i) Binary attribute is a special type of discrete attribute which can have only two values — male/female, positive/negative, yes/no, etc.

b) **Continuous attributes**

Continuous attributes can assume any possible value which is a real number. Ex: length, height, weight, price, etc.

EXPLORING STRUCTURE OF DATA

Exploring structure of data is to identify the OUTLIERS, DATA SPREAD, MISSING values, etc.

Outliers are the values which are unusually high or low, compared to the other values.

1) Exploring Numerical data

a) Understanding the central tendency

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

✓ In statistics, measures of central tendency help us understand the central point of a set of data.

Mean

is a sum of all data values divided by the count of data elements.

Ex: Mean of a set of observations – 21, 89, 34, 67, and 96 is calculated as below.

$$\text{Mean} = \frac{21 + 89 + 34 + 67 + 96}{5} = 61.4$$

Median

is the value of the element appearing in the middle of an ordered list of data elements. Ex: Median of a set of observations — 21, 89, 34, 67, and 96 is calculated as below.

The ordered list would be -> 21, 34, 67, 89, and 96. Since there are 5 data elements, **the 3rd element in**

the ordered list is considered as the median. Hence, the median value of this set of data is 67.

MEAN is likely to get shifted drastically even due to the presence of a small number of outliers

b) Understanding the data spread

i) Measuring data dispersion

ii) Measuring different data values position

i) Measuring data dispersion

Variance of the data is used to measure the extent of dispersion of data or to find out how much the different values of a data are spread out.

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

Standard deviation

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

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

Example: Consider the data values of two attributes

1. Attribute 1 values : 44, 46, 48, 45, and 47
2. Attribute 2 values : 34, 46, 59, 39, and 52

For Attribute 1

Mean =44+46+48+45+47 / 5 = 46

Median = 46 (after arranging into sorted order

44,45,46,47,48) Variance =

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

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

$$= \frac{1936 + 2116 + 2304 + 2025 + 2209}{5} - \left(\frac{230}{5}\right)^2 = \frac{10590}{5} - (46)^2 = 2$$

For Attribute 2

Mean =34+46+59+39+52 / 5 = 46

Median = 46 (after arranging into sorted order

34,39,46,52,59) Variance =

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

$$= \frac{34^2 + 46^2 + 59^2 + 39^2 + 52^2}{5} - \left(\frac{34 + 46 + 59 + 39 + 52}{5}\right)^2$$

$$= \frac{1156 + 2116 + 3481 + 1521 + 2704}{5} - \left(\frac{230}{5}\right)^2 = \frac{10978}{5} - (46)^2 = 79.6$$

So it is quite clear from the measure that attribute 1 values are quite concentrated around the mean while attribute 2 values are extremely spread out.

ii) **Measuring different data values position**

Any data set has five values –

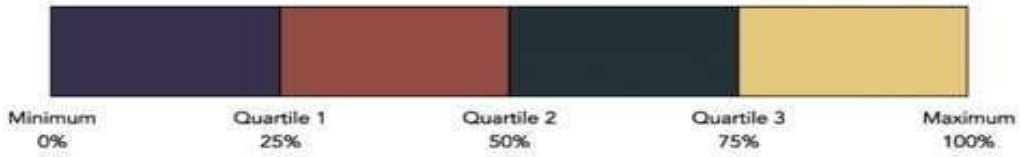
Minimum, first quartile (Q1),median (Q2), third quartile (Q3), and maximum

✓ When the data values of an attribute are arranged in an increasing order, we have seen earlier that median gives the central data value, which divides the entire data set into two halves.

✓ Similarly, if the first half of the data is divided into two halves so that each half consists of one quarter of the data set, then that median of the first half is known as first quartile or Q1.

✓ In the same way, if the second half of the data is divided into two halves, then that median of the second half is known as third quartile or Q3.

✓ The overall median is also known as second quartile or Q2.



Quantiles: Refer to specific points in a data set which divide the data set into equal parts or equally sized quantities.

Quartile: When the entire data set is which is ordered, splitted into 4 parts is known as a quartile.

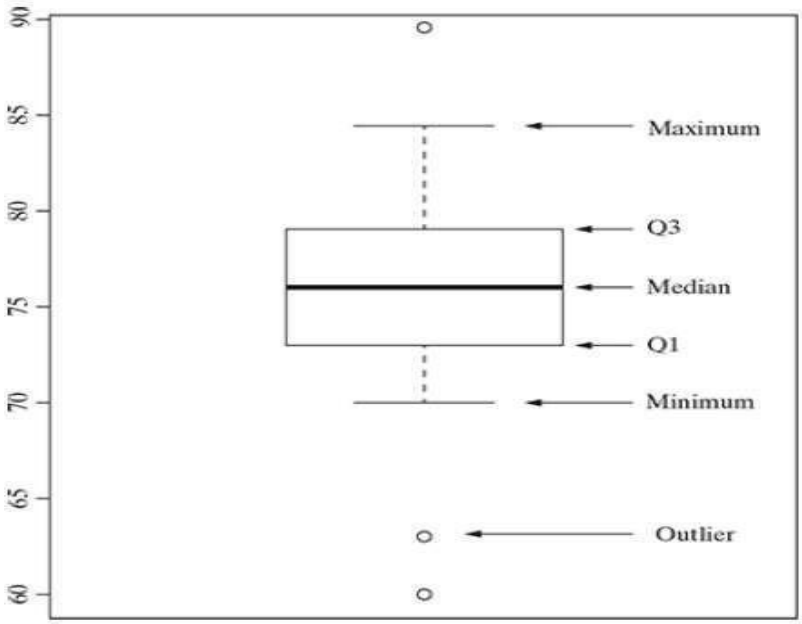
Percentile: When the data set is splitted into 100 parts is known as a percentile.

Plotting and exploring numerical data

Following two techniques are used to plot and explore the numerical data

i) Box plots

❖ A box plot is an extremely effective mechanism to get a one-shot view and understand the nature of the data.



Inter-quartile range (IQR)

- ✓ The central rectangle or the box spans from first to third quartile (i.e. Q1 to Q3), thus giving the inter-quartile range (IQR).

  IQR = Q3-Q1

ii) Histogram

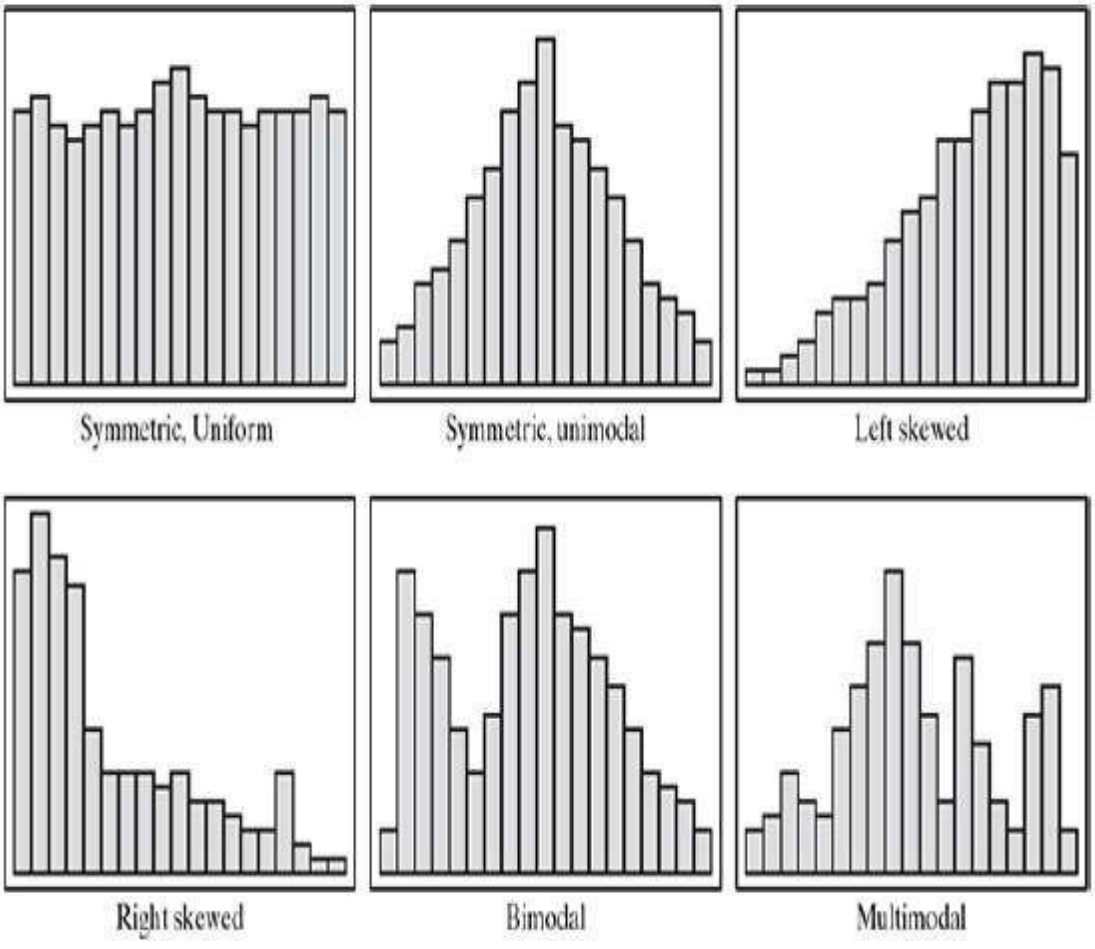
✓Histogram is another plot which helps in effective visualization of numeric attributes. It helps in understanding the distribution of a numeric data into series of intervals, also termed as 'bins'.

Difference between histogram and box plot is

a) The focus of **histogram** is to plot ranges of data values (acting as 'bins'), the number of data elements in each range will depend on the data distribution. Based on that, the size of each bar corresponding to the different ranges will vary.

b) The focus of **box plot** is to divide the data elements in a data set into four equal portions, such that each portion contains an equal number of data elements.

General Histogram shapes



✓ The histogram is composed of a number of bars, one bar appearing for each of the 'bins'. The height of the bar reflects the total count of data elements whose value falls within the specific bin value, or the frequency.

2) <u>Exploring Categorical data</u>

✓ There are not many options for exploring categorical data.

✓ MODE is only the measure we can apply to explore the categorical data. mode is also a statistical measure for central tendency of a data.

Mode of a data is the data value which appears most often.

✓ **Count** and **proportion** (**percentage)** are two parameters used to measure categorical data.

✓ **<u>Ex:</u>** <u>Count of CAR names</u>

For attribute 'car name'

1. Chevrolet chevelle malibu
2. Buick skylark 320
3. Plymouth satellite
4. Amc rebel sst
5. Ford torino
6. Ford galaxie 500
7. Chevrolet impala
8. Plymouth fury iii
9. Pontiac catalina
10. Amc ambassador dpl

Count of Categories for 'car name' Attribute

Attribute Value	amc ambas-sador brougham	amc ambas-sador dpl	amc ambassa-dor sst	amc concord	amc concord d/l	amc con-cord dl 6	amc gremlin	...
Count	1	1	1	1	2	2	4	...

✓ **<u>Ex:</u>** <u>Percentage of count of data elements (for CAR names)</u>

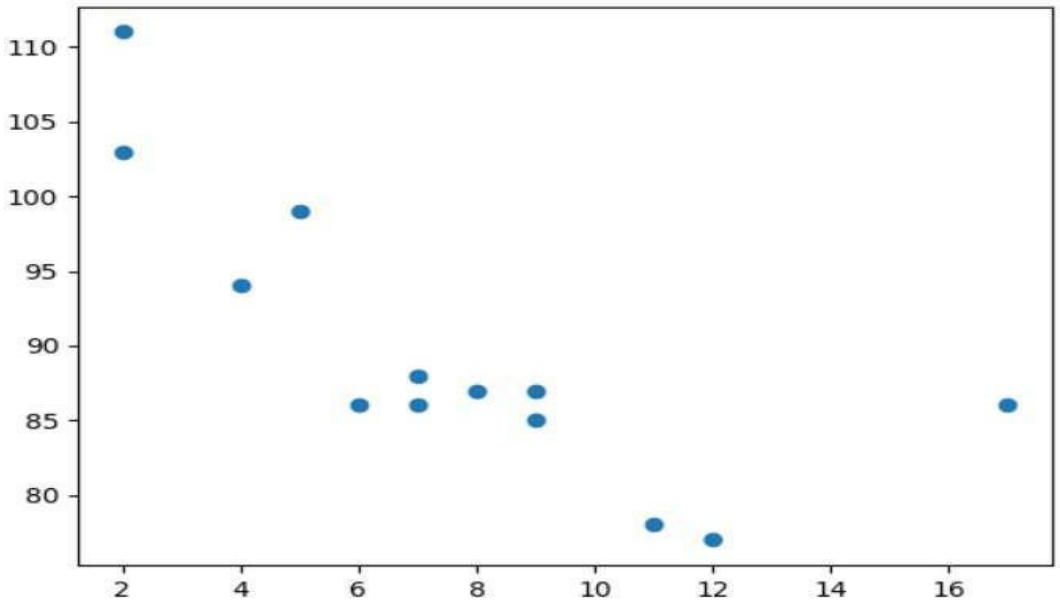
Attribute	Amc	Amc	Amc	Amc	Amc	Amc	Amc	...
Value	ambas-sador brougham	ambassa-dor dpl	ambassa-dor sst	concord	concord d/l	concord dl 6	gremlin	
Count	0.003	0.003	0.003	0.003	0.005	0.005	0.01	...

Exploring relationship between variables

One more important angle of data exploration is to explore relationship between attributes. There are multiple plots to enable us explore the relationship between variables. The basic and most commonly used plot is **scatter plot** and **two-way cross-tabulations.**

a) Scatter plot

- ✓ A scatter plot helps in visualizing bivariate relationships, i.e. relationship between two variables. It is a two dimensional plot in which points or dots are drawn on coordinates provided by values of the attributes.
- ✓ For example, in a data set there are two attributes — attr_1 and attr_2. We want to understand the relationship between two attributes, i.e. with a change in value of one attribute, say attr_1, how does the value of the other attribute, say attr_2, changes.



- ✓ We can draw a scatter plot, with attr_1 mapped to *x*-axis and attr_2 mapped in *y*-axis.
- ✓ So, every point in the plot will have value of attr_1 in the *x*-coordinate and value of attr_2 in the *y* coordinate.
- ✓ As in a two-dimensional plot, attr_1 is said to be the independent variable and attr_2 as the dependent variable.

***b)* Two-way cross-tabulations**

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

It has a matrix format that presents a summarized view of the bivariate frequency distribution.

A cross-tab, very much like a scatter plot, helps to understand how much the data values of one attribute changes with the change in data values of another attribute.

Cross-tab for 'Model year' *vs.* 'Origin'

Origin \ Model Year	70	71	72	73	74	75	76	77	78	79	80	81	82
1	22	20	18	29	15	20	22	18	22	23	7	13	20
2	5	4	5	7	6	6	8	4	6	4	9	4	2
3	2	4	5	4	6	4	4	6	8	2	13	12	9

'Cylinders' vs. 'Origin'

DATA QUALITY AND REMEDIATION

1) DATA QUALITY

Success of machine learning depends largely on the quality of data. A data which has the right quality helps to achieve better prediction accuracy, in case of supervised learning.

Two types of data quality issues

1. Certain data elements without a value or data with a missing value.
2. Data elements having value surprisingly different from the other elements, which we term as outliers.

Few factors which lead to the above quality issues

a) Incorrect sample set selection

The data may not reflect normal or regular quality due to incorrect selection of sample set.

Example

If we are selecting a sample set of sales transactions from a festive period and trying to use that data to predict sales in future. In this case, the prediction will be far apart from the actual scenario, just because the sample set has been selected in a wrong time.

b) Errors in data collection: resulting in outliers and missing values

- In many cases, a person or group of persons are responsible for the collection of data to be used in a learning activity.
- In this manual process, there is the possibility of wrongly recording data either in terms of value (say 20.67 is wrongly recorded as 206.7 or 2.067) or in terms of a unit of measurement (say cm. is wrongly recorded as m. or mm.).
- This may result in data elements which have abnormally high or low value from other elements. Such records are termed as *outliers*.
- It may also happen that the data is not recorded at all.
- In case of a survey conducted to collect data, it is all the more possible as survey responders may choose not to respond to a certain question. So the data value for that data element in that responder's record is *missing*.

2) DATA REMEDIATION

The issues in data quality need to be remediated, if the right amount of efficiency has to be achieved in the learning activity.

(1) For Incorrect sample set selection – Remedy is "Proper sampling technique"

(2) Handling outliers

Outliers are data elements with an abnormally high value which may impact prediction accuracy, especially in regression models.

- ✓ One of the following approaches are used to handle outliers

i. **Remove outliers:** If the number of records which are outliers is not many, a simple approach may be to remove them.

ii. **Imputation:** One other way is to impute the value with mean or median or mode. The value of the most similar data element may also be used for imputation.

iii. **Capping:** For values that lie outside the 1.5|×| IQR limits, we can cap them by replacing those observations below the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of 95th percentile.

(3) Handling missing values

- In a data set, one or more data elements may have missing values in multiple records.
- There are multiple strategies to handle missing value of data elements. Some of those strategies are:

i. **Eliminate records having a missing value of data elements**

- In case the proportion of data elements having missing values is within a tolerable limit, a simple but effective approach is to remove the records having such data elements.
- This is possible if the quantum of data left after removing the data elements having missing values is sizeable.

ii. Imputing missing values

- Imputation is a method to assign a value to the data elements having missing values. Mean/mode/median is most frequently assigned value.
- For quantitative attributes, all missing values are imputed with the mean, median, or mode of the remaining values under the same attribute.
- For qualitative attributes, all missing values are imputed by the mode of all remaining values of the same attribute.

iii. Estimate missing values

- If there are data points similar to the ones with missing attribute values, then the attribute values from those similar data points can be planted in place of the missing value.
- **Ex:**The weight of a Russian student having age 12 years and height 5 ft. is missing. Then the weight of any other Russian student having age close to 12 years and height close to 5 ft. can be assigned.

DATA PRE-PROCESSING

✓ Two techniques are applied as part of data pre-processing

1) Dimensionality reduction

2) Feature subset selection

1) Dimensionality reduction

- ✓ High-dimensional data sets need a high amount of computational space and time. At the same time, not all features are useful — they degrade the performance of machine learning algorithms. Most of the machine learning algorithms perform better if the dimensionality of data set, i.e. the number of features in the data set, is reduced.
- ✓ **Dimensionality reduction** refers to the techniques of reducing the dimensionality of a data set by creating new attributes by combining the original attributes.

a. The most common approach for dimensionality reduction is known as Principal Component Analysis (PCA).

- ✓ PCA is a statistical technique to convert a set of correlated variables into a set of transformed, uncorrelated variables called principal components.
- ✓ The principal components are a linear combination of the original variables. They are orthogonal to each other.
- ✓ Since principal components are uncorrelated, they capture the maximum amount of variability in the data. However, the only challenge is that the original attributes are lost due to the transformation.

b. Another commonly used technique which is used for dimensionality reduction is Singular Value Decomposition (SVD)

2) Feature subset selection

- Feature subset selection or simply called feature selection, both for supervised as well as unsupervised learning, try to find out the optimal subset of the entire feature set which significantly reduces computational cost without any major impact on the learning accuracy.
- As part of this process, few features will be eliminated which are irrelevant. A feature is considered as irrelevant if it plays an insignificant role (or contributes almost no information) in classifying or grouping together a set of data instances.

UNIT 2

Modeling and Evaluation & Basics of Feature Engineering: Introduction, selecting a Model, training a Model (for Supervised Learning): k-fold Cross-validation Method, Model Representation and Interpretability, Evaluating Performance of a Model: Supervised Learning-Classification, Supervised Learning – Regression, Unsupervised Learning – Clustering, Improving Performance of a Model.

Basics of Feature Engineering: Introduction, Feature Transformation, Feature Construction, Feature Extraction, Feature Subset Selection: Issues in High-Dimensional Data, Key Drivers of Feature Selection – Feature Relevance and Redundancy, Measures of Feature Relevance and Redundancy, Overall Feature Selection Process, Feature Selection Approaches.

SELECTING A MODEL

Multiple factors play a role when we try to select the model for solving a machine learning problem. The most important factors are

(i) the kind of problem we want to solve using machine learning and

(ii) the nature of the underlying data.

Note: There is no one model that works best for every machine learning problem. This is what '**No Free Lunch'** theorem also states.

Machine learning algorithms are broadly of two types:

- models for **supervised learning**, which primarily focus on solving **predictive problems** and
- models for **unsupervised learning**, which solve **descriptive problems**.

1. Predictive Models

- Models for supervised learning or predictive models try to predict certain value using the values in an input data set.
- The learning model attempts to establish a relation between the target feature, i.e. the feature being predicted, and the predictor features.
- The predictive models have a clear focus on what they want to learn and how they want to learn.

Types of Predictive models : - There are two types –

a) Classification models

b) Regression models

a) **Classification models**

The models which are used for prediction of target features of categorical value are known as classification models.

Popular classification models are: K-Nearest Neighbor (KNN), Naïve Bayes, and Decision Tree.

Example

1. Predicting win/loss in a cricket match
2. Predicting whether a transaction is fraud
3. Predicting whether a customer may move to another product

b) **Regression models**

The models which are used for prediction of the numerical value of the target feature of a data instance are known as regression models.

Popular Regression models are: Linear regression and Logistic regression models.

Example

1. Prediction of revenue growth in the succeeding year
2. Prediction of rainfall amount in the coming monsoon
3. Prediction of potential flu patients and demand for flu shots next winter

NOTE:

1. Categorical values can be converted to numerical values and vice versa.
2. Few models like Support Vector Machines and Neural Network can be used for both classifications as well as for regression.

2. **Descriptive Models**

- Models for unsupervised learning or descriptive models are used to describe a data set or gain insight from a data set.
- There is no target feature or single feature of interest in case of unsupervised learning. Based on the value of all features, interesting patterns or insights are derived about the data set.

Types of Descriptive models : - There are two types –

a) Clustering models
b) Pattern discovery (OR) Association analysis models

a) Clustering models

Descriptive models which group together similar data instances, i.e. data instances having a similar value of the different features are called clustering models.

Popular model for clustering is K-means.

Examples

1. Customer grouping or segmentation based on social, demographic, ethnic, etc. factors
2. Grouping of music based on different aspects like genre, language, time- period, etc.
3. Grouping of commodities in an inventory

b) Pattern discovery (OR) Association Analysis models

- Descriptive models related to pattern discovery is used for market basket analysis of transactional data.
- In market basket analysis, based on the purchase pattern available in the transactional data, the possibility of purchasing one product based on the purchase of another product is determined.

Popular model for Pattern discovery is Apriori algorithm

- **Example:** Transactional data may reveal a pattern that generally a customer who purchases milk also purchases biscuit at the same time.
- This can be useful for targeted promotions or in-store set up.

TRAINING A MODEL (FOR SUPERVISED LEARNING)

- In case of supervised learning, a model is trained using the labeled input data.
- In general, 70%–80% of the input data (which is obviously labelled) is used for model training. The remaining 20%–30% is used as test data for validation of the performance of the model.
- However, a different proportion of dividing the input data into training and test data is also acceptable.
- To make sure that the data in both the buckets are similar in nature, the division is done randomly. Random numbers are used to assign data items to the partitions.

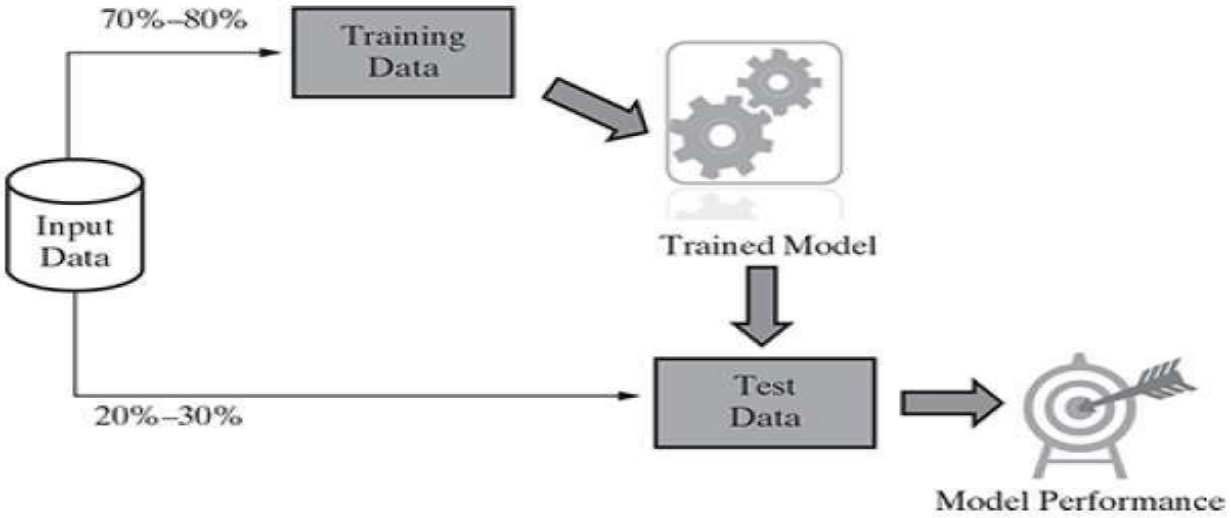
Below are the methods used for training a model for supervised learning.

Methods

1. Holdout method
2. K-fold Cross-validation method
3. Bootstrap sampling
4. Lazy vs. Eager learner

1. Holdout method

- The input data is partitioned into two parts — **training** and **test data**, which is holding back a part of the input data for validating the trained model is known as holdout method.
- In certain cases, the input data is partitioned into three portions – **a training** and **a test data**, and **a third validation data**.
- The validation data is used in place of test data, for measuring the model performance. It is used in iterations and to refine the model in each iteration.
- The test data is used only for once, after the model is refined and finalized, to measure and report the final performance of the model as a reference for future learning efforts.



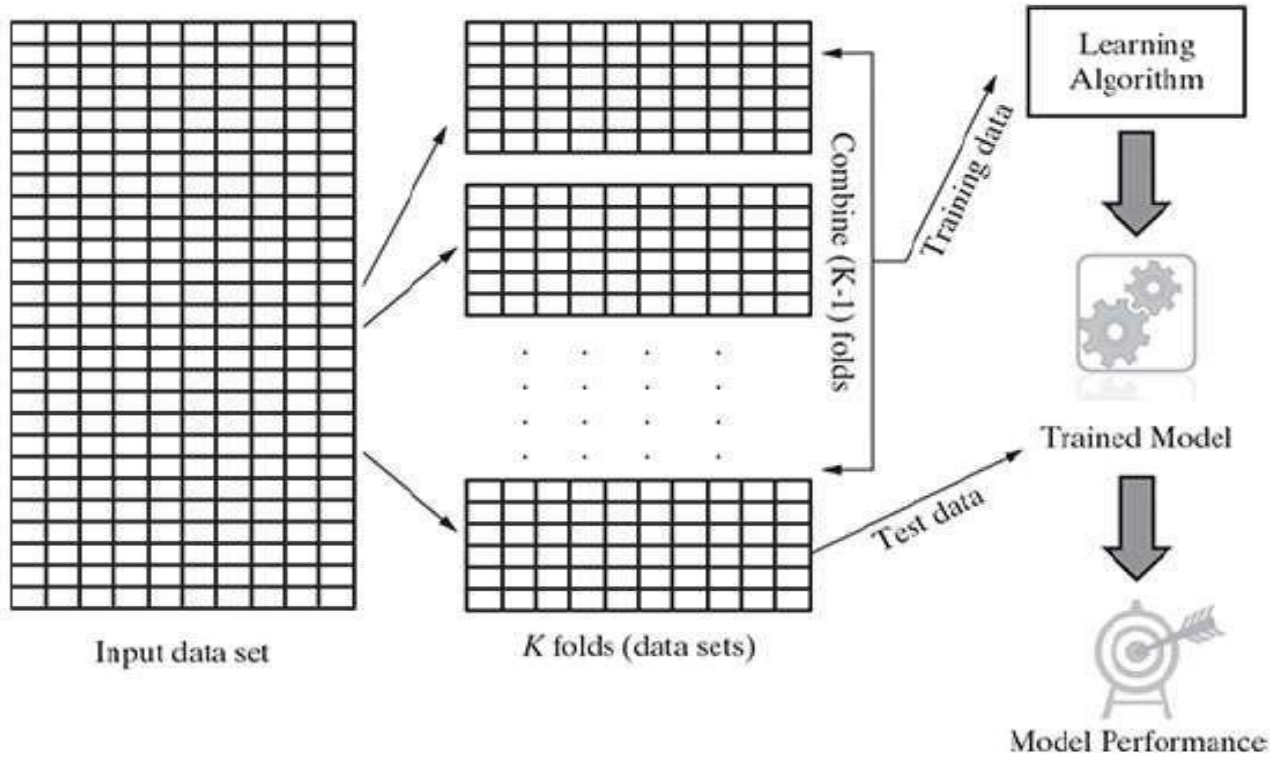
Holdout method - problems

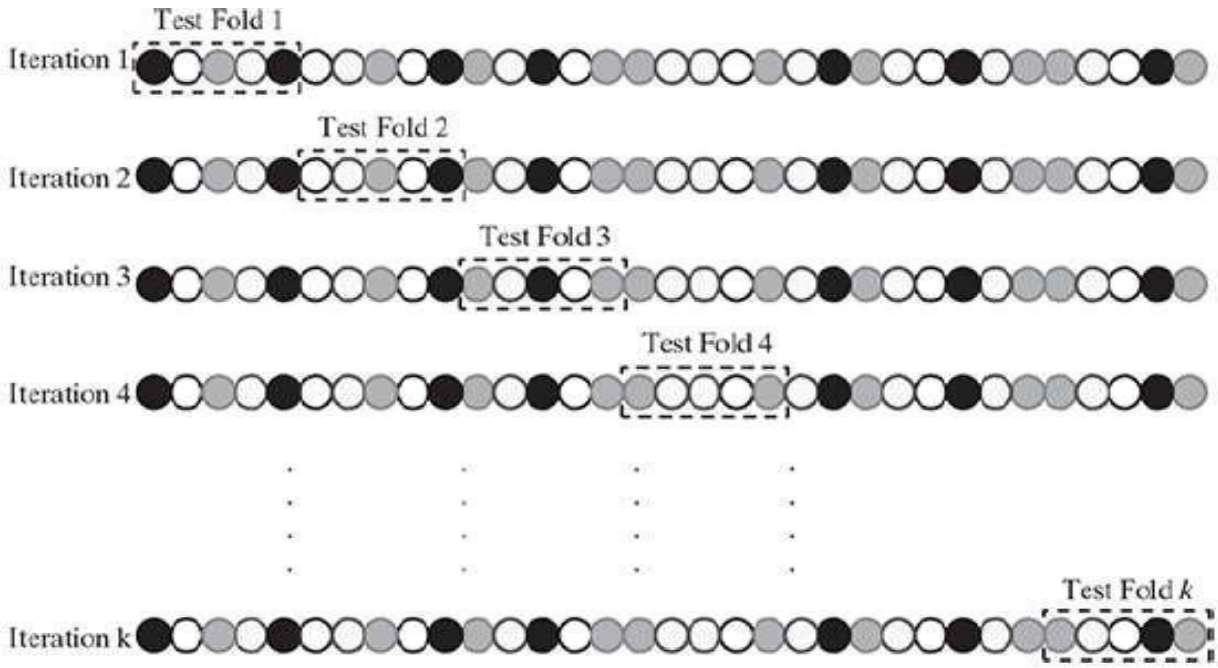
- An obvious problem in this method is that the division of data of different classes into the training and test data may not be proportionate.
- This situation is worse if the overall percentage of data related to certain classes is much less compared to other classes.
- This problem can be addressed to some extent by applying stratified random sampling in place of sampling.

- In case of stratified random sampling, the whole data is broken into several homogenous groups or strata and a random sample is selected from each such stratum.
- This ensures that the generated random partitions have equal proportions of each class.

2. K-fold Cross-validation method

- A special variant of holdout method, called repeated holdout, is sometimes employed to ensure the randomness of the composed data sets.
- In repeated holdout, several random holdouts are used to measure the model performance.
- In the end, the average of all performances is taken. As multiple holdouts have been drawn, the training and test data (and also validation data, in case it is drawn) are more likely to contain representative data from all classes and resemble the original input data closely.
- This process of repeated holdout is the basis of *k*-fold cross-validation technique.
- In *k*-fold cross-validation, the data set is divided into *k* completely distinct or non-overlapping random partitions called folds.





Types of K-fold Cross-validation method

- The value of '*k*' in *k*-fold cross-validation can be set to any number. However, there are two approaches which are extremely popular:

a. 10-fold cross-validation (10-fold CV)

b. Leave-one-out cross-validation (LOOCV)

a. 10-fold cross-validation (10-fold CV)

- In this approach, for each of the 10-folds, each comprising of approximately 10% of the data, one of the folds is used as the test data for validating model performance trained based on the remaining 9 folds (or 90% of the data).
- This is repeated 10 times, once for each of the 10 folds being used as the test data and the remaining folds as the training data. The average performance across all folds is being reported.

b. Leave-one-out cross-validation (LOOCV)

- It is an extreme case of k-fold cross-validation using one record or data instance at a time as a test data.

Lazy learning, on the other hand, completely skips the abstraction and generalization processes, as explained in context of a typical machine learning process.

- In that respect, strictly speaking, lazy learner doesn't 'learn' anything. It uses the training data in exact, and uses the knowledge to classify the unlabeled test data.
- Since lazy learning uses training data as-is, it is also known as rote learning (i.e. memorization technique based on repetition). Due to its heavy dependency on the given training data instance, it is also known as instance learning. They are also called non-parametric learning.
- Lazy learners take very little time in training because not much of training actually happens. However, it takes quite some time in classification as for each tuple of test data, a comparison-based assignment of label happens.
- One of the most popular algorithm for lazy learning is *k*-nearest neighbor

MODEL REPRESENTATION AND INTERPRETABILITY

✓ The goal of supervised machine learning algorithm is to learn or derive a target function which can best determine the target variable from the set of input variables

✓ Fitness of a target function approximated by a learning algorithm determines how correctly it is able to classify a set of data it has never seen.

✓ Following are the methods used to check the fitness of the models

a) Under fitting

b) Over fitting

c) Bias – Variance trade-off

- Errors due to Bias
- Errors due to Variance

a) **Under fitting:** Under fitting is a scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and unseen data.

Reasons for under fitting

- Many times it happens due to unavailability of sufficient training data
- Results in poor performance with training data and test data

Prevention of under fitting

- using more training data
- Reducing features by effective feature selection

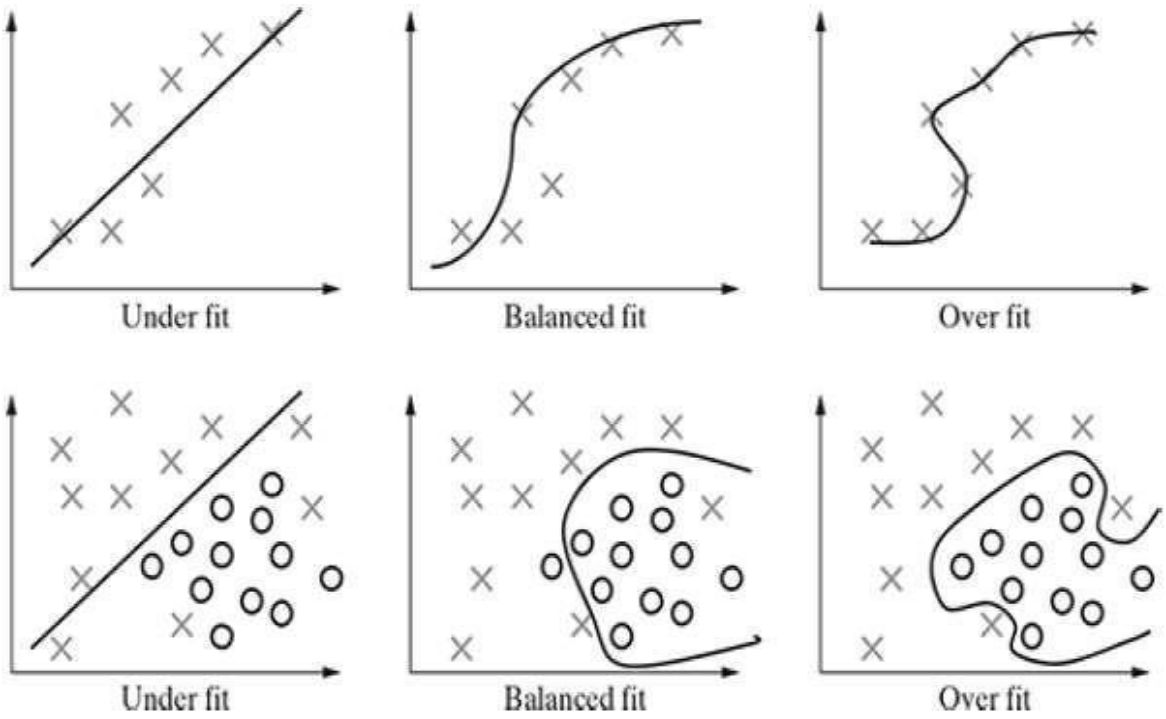
b) Over fitting

✓ Over fitting is a concept in data science, which occurs when a statistical model fits exactly against its training data. When this happens, the algorithm unfortunately cannot perform accurately against unseen data, defeating its purpose.

✓ Over fitting results in good performance with training data set, but poor generalization and hence poor performance with test data set.

Prevention of Over fitting

- using re-sampling techniques like *k*-fold cross validation
- hold back of a validation data set
- remove the nodes which have little or no predictive power for the given machine earning Problem.



Underfitting and Overfitting of models

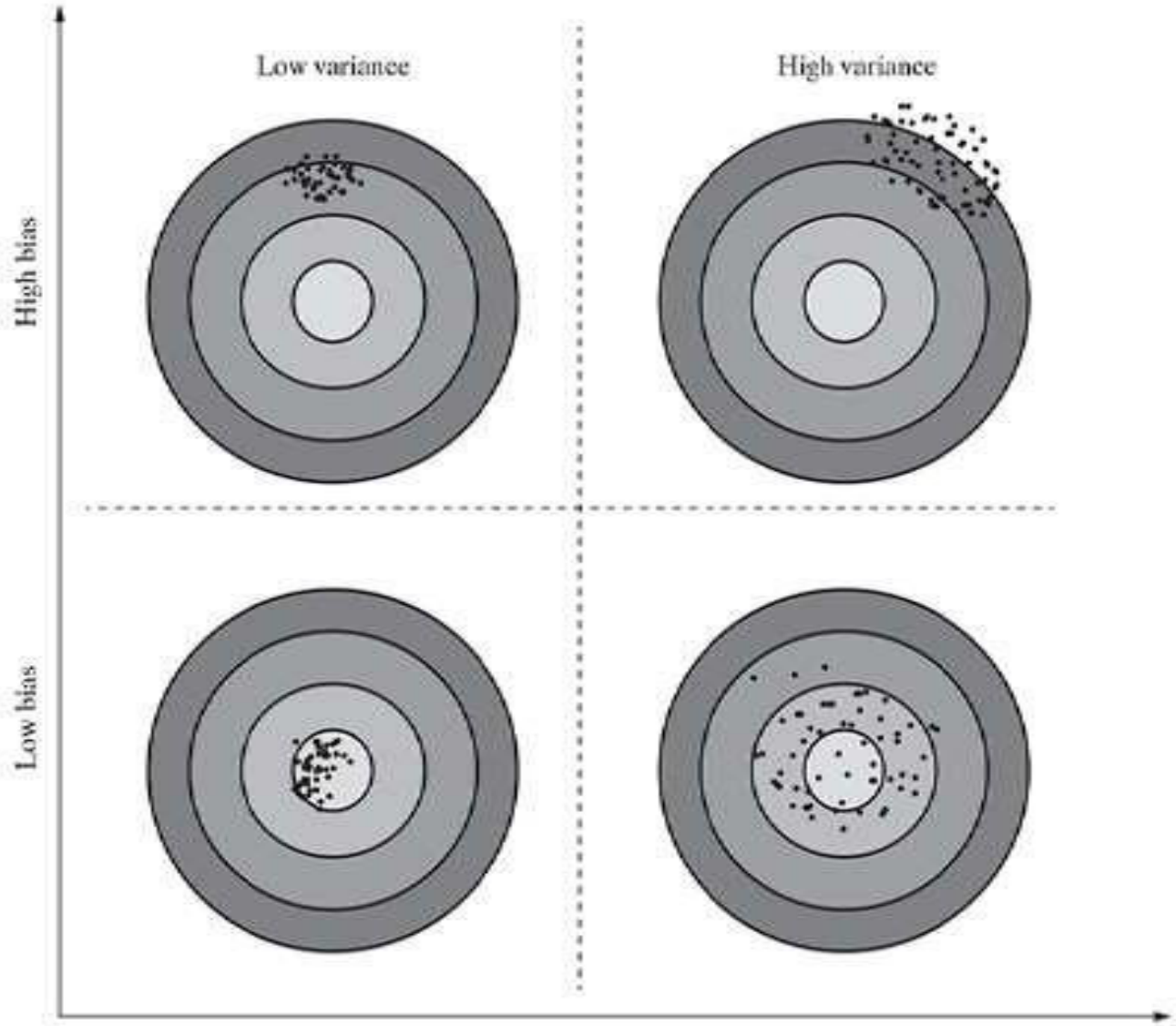
c) Bias-variance trade-off

- In supervised learning, the class value assigned by the learning model built based on the training data may differ from the actual class value.
- This error in learning can be of two types – errors due to 'bias' and error due to

'variance'.

- **Errors due to bias** arise from simplifying assumptions made by the model to make the target function less complex or easier to learn. Bias means — selecting unwanted/irrelevant features from the given data set.
- **Errors due to variance** occur from difference in training data sets used to train the model. Different training data sets (randomly sampled from the input data set) are used to train the model. Ideally the difference in the data sets should not be significant and the model trained using different training data sets should not be too different.

An ideal choice is – Low bias and Low variance.



Evaluating Performance of a Model

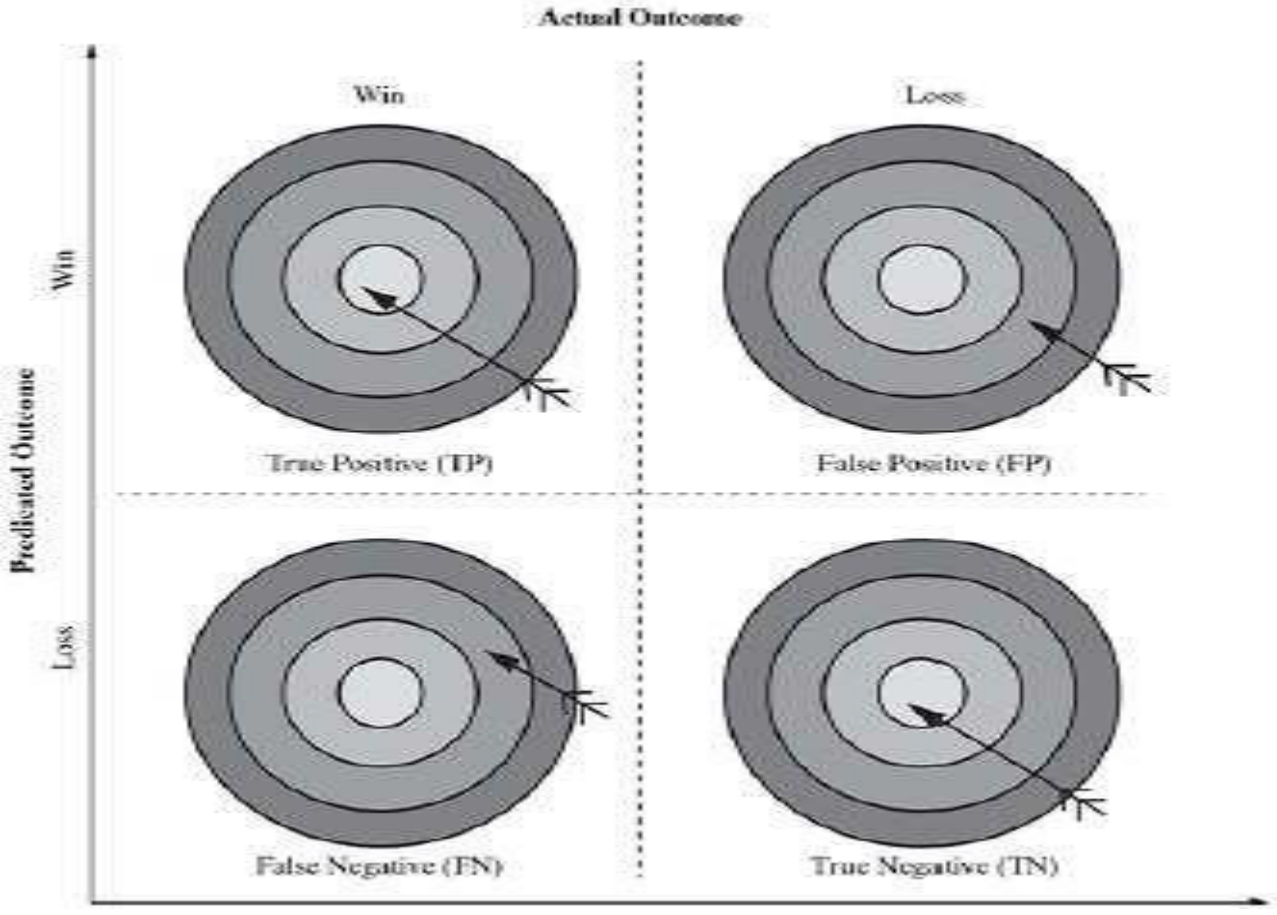
1. Supervised learning – classification

Performance measurement is done via -

a) Accuracy
b) Sensitivity
c) Specificity
d) Precision

There are four possibilities with regards to the cricket match win/loss prediction: CASE 1: the model predicted win and the team won - True Positive (TP) cases.

CASE 2: The model predicted win and the team lost - False Positive (FP) cases. CASE 3: the model predicted loss and the team won - False Negative (FN) cases. CASE 4: the model predicted loss and the team lost — True Negative (TN) cases.
In this problem, the obvious class of interest is 'win'.



model accuracy = total number of correct classifications (either as the class of interest, i.e. True Positive or as not the class of interest, i.e. True Negative) divided by total number of classifications done.

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

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

The win/loss prediction of cricket match has two classes of interest – win and loss. For that reason it will generate a 2 × 2 confusion matrix.

Example: The confusion matrix of the win/loss prediction of cricket match problem is given below. CALCULATE MODEL ACCURACY and ERROR RATE.

	ACTUAL WIN	ACTUAL LOSS
Predicted Win	85	4
Predicted Loss	2	9

In context of the above confusion matrix, total count of TPs = 85, count of FPs = 4, count of FNs

= 2 and count of TNs = 9.

$$\therefore \text{Model accuracy} = \frac{TP + TN}{TP + FP + FN + TN} = \frac{85 + 9}{85 + 4 + 2 + 9} = \frac{94}{100} = 94\%$$

The percentage of misclassifications is indicated using **error rate** which is measured as

$$\text{Error rate} = \frac{FP + FN}{TP + FP + FN + TN}$$

$$\text{Error rate} = \frac{FP + FN}{TP + FP + FN + TN} = \frac{4 + 2}{85 + 4 + 2 + 9} = \frac{6}{100} = 6\%$$

$$= 1 - \text{Model accuracy}$$

Kappa value of a model indicates the adjusted the model accuracy.

$$\therefore P(a) = \frac{TP + TN}{TP + FP + FN + TN} = \frac{85 + 9}{85 + 4 + 2 + 9} = \frac{94}{100} = 0.94$$

$$P(p_r) = \frac{85 + 4}{85 + 4 + 2 + 9} \times \frac{85 + 2}{85 + 4 + 2 + 9} + \frac{2 + 9}{85 + 4 + 2 + 9} \times \frac{4 + 9}{85 + 4 + 2 + 9} \text{ in}$$

$$= \frac{89}{100} \times \frac{87}{100} + \frac{11}{100} \times \frac{13}{100} = 0.89 \times 0.87 + 0.11 \times 0.13 = 0.7886$$

$$\therefore k = \frac{0.94 - 0.7886}{1 - 0.7886} = 0.7162$$

both in case of class of interest as well as the other classes

$$= \frac{TP + FP}{TP + FP + FN + TN} \times \frac{TP + FN}{TP + FP + FN + TN} + \frac{FN + TN}{TP + FP + FN + TN} \times \frac{FP + TN}{TP + FP + FN + TN}$$

NOTE: Kappa value can be 1 at the maximum, which represents perfect agreement between model's prediction and actual values.

Sensitivity

- The sensitivity of a model measures the proportion of TP examples or positive cases which were correctly classified. It is measured as —

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

- In the context of the above confusion matrix for the cricket match win prediction problem,

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{85}{85 + 2} = \frac{85}{87} = 97.7\%$$

Specificity is also another good measure to indicate a good balance of a model being excessively conservative or excessively aggressive. Specificity of a model measures the

proportion of negative examples which have been correctly classified.

- In the context, of malignancy prediction of tumors, specificity gives the proportion of benign tumors which have been correctly classified.
- In the context of the above confusion matrix for the cricket match win prediction problem,

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{9}{9 + 4} = \frac{9}{13} = 69.2\%$$

Precision Precision tells us how many of the correctly predicted cases actually turned out to be positive.

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

✓ Precision indicates the reliability of a model in predicting a class of interest. When the model is related to win / loss prediction of cricket, precision indicates how often it predicts the win correctly.

✓ In context of the above confusion matrix for the cricket match win prediction problem,

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{85}{85 + 4} = \frac{85}{89} = 95.5\%$$

Recall

- Recall tells us how many of the actual positive cases we were able to predict correctly with our model.
- In case of win/loss prediction of cricket, recall resembles what proportion of the total wins were predicted correctly.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- In the context of the above confusion matrix for the cricket match win prediction problem,

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{85}{85 + 2} = \frac{85}{87} = 97.7\%$$

F-measure

- *F*-measure is another measure of model performance which combines the precision

and recall. It takes the harmonic mean of precision and recall as calculated as

$$F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

In context of the above confusion matrix for the cricket match win prediction problem,

$$F\text{-measure} = \frac{2 \times 0.955 \times 0.977}{0.955 + 0.977} = \frac{1.866}{1.932} = 96.6\%$$

"Basics of Feature Engineering"

Feature: A feature is an attribute data set that is used in a machine learning process. There is a view amongst certain machine learning Practitioners that only those attributes which are meaningful to a machine learning problem are to be called as "**features"** selection of the subset of features which are meaningful for machine learning is a sub-area of feature engineering. The features in a dataset are also called it's "dimensions So, a dataset having 'n' features is called an 'n' dimensional data set

Ex:- Let's take the example of a famous Machine learning data set, Iris introduced by the "British statistician" and biologist "Ronald Fisher". It has 5 attributes (or) Features namely "sepal Length, Sepal width, Petal Length, Petal Length width an Length and species" out of these, the feature "species" represent the class variable and the remaining features are the predictor Variables. It is a 5-dimensional data set.

Sepal Length	Sepal width	Petal Length	Petal width	species
6.7	3.3	5.7	2.5	virginica
4.9	3	1.4	0.2	Setosa
5.5	2.6	4.4	1.2	virginica
6.8	3.2	5.9	2.3	virginica
5.5	2.5	4	1.3	versicolor

Feature Engineering:- It refers to the process of translating a data set into features such that these features are able to represent the data set more effectively and result in a better learning performance. Feature Engineering is an important pre-Processing Step for

machine learning. It has 2 elements.

(1) Feature Transformation

(2) Feature subset selection

(1) Feature Transformation:- It Transforms the data. Structured (or) un-structured, into a new set of features can represent the underlying problem which machine learning is trying to solve. There are 2 Variants co Feature Transformation.

(1) Feature Construction

(2) Feature Extraction.

Feature Construction:- This process discovers missing information about the relationships between features and augments, the feature space by Creating additional features. If there are 'n' features (or) dimensions in a dataset, after feature construction 'm' more features (or) dimensions may get added. The dataset will become 'n+m" dimensional.

Feature extractions:- It is the process of extracting (or) creating a new set of features from the original Set of features using some functional mapping. unlike feature transformation, in case of "feature sub-set selection" no new feature is generated.

Feature Trans formation:-

Feature Transformation is used as an effective tool for dimensionality reduction and hence for boosting learning motel Performance. There are 2 goals for used & feature Transformation.

(1) Achieving best reconstruction of the original features in the data set.

(2) Achieving highest efficiency in the learning task.

Feature Construction:- Feature construction involves transforming a given set of input features to generate a new set of more powerful features.

Feature construction is an essential activity before we can start with the Machine learning task. These Situations are:

(1) when features have categorical value and Machine Learning needs Numeric value inputs.

(2) when features having Numeric (Continuous) values and need to be Converted to

81dinal Values

(3) when text-specific feature construction needs to be done.

EnCoding Categorical (Nominal) Variables:-

The data set has features "age, city & origin, Parents athlete and chance of win". The feature chance of a win isa Class Variable and others are predictor variables. Any Machine Learning algorithm, it's a classification (KNN) (a Regression algorithm requires Numerical figures to leave from So, there are 3 features - "City of origin, Parents athlete and chance & win"

Feature construction can be used to create new dummy features which are usable by M.L algorithms. The feature "city of origin" has 3 unique values names City-A, City-B, City-C is created. In the Same way, dummy features Parents-athlete-y and Parents-athlete-N are Created for feature "parents athlete" and win-Chance-y and win-chance-N are created for feature chance & win'. The dummy features have Valve o' (or)' based on the categorical value for the original feature in that row.

Age (years)	City of origin	Parents athelete	Change of win
18	CITY A	YES	Y
20	CITY B	NO	Y
23	CITY B	YES	Y
19	CITY A	NO	N
118	CITY C	YES	N
22	CITY A	YES	Y

AGE (YEARS)	ORIGIN CITY A	ORIGIN CITY B	ORIGIN CITY C	PARENTS ATHLETE-Y	PARENTS ATHLETE-N	WIN CHANGE-Y	WIN CHANGE-N
18	1	0	0	1	0	1	0
20	0	1	0	0	1	1	0
23	0	1	0	1	0	1	0
19	1	0	0	0	1	0	1
18	0	0	1	1	0	0	1

22	0	1	0	1	0	1	0

AGE (YEARS)	**ORIGIN CITY A**	**ORIGIN CITY B**	**ORIGIN CITY C**	**PARENTS ATHLETE - Y**	**WIN CHANGE -Y**
18	1	0	0	1	1
20	0	1	0	0	1
23	0	1	0	1	1
19	1	0	0	0	0
18	0	0	1	1	0
22	0	1	0	1	1

Ex:- The second row had a Categorical Valve city-B for the feature "city of origin". So, the newly created features in place & "city & origin", origin-city-A, origin -city-B and origin-city- c will have values 0,1 and O.

In the same way, Parents-athlete-y and Parents-athlete-N will have Values ‘o’ and ‘1’. In row 2 98 the original feature "Parents athlete" had a Categorical "Value 'No' in row 2.

Encoding Categorical (ordinal) Variables :-

Let's Assume that there are 3 Variables - "Science Marks, Maths Marks and Grade". we can see the grade is an ordinal variable with values A, B, C and D. To transform this variable in to a numerical variable. we can create a feature Num-grade mapping a numeric value against Cach ordinal Value.

Ex- A, B, C and D Grades is mapped to Values 1, 2,3,4 in the transformed Variables.

Mark in science	Mark in maths	Grade
78	75	B
56	62	C
87	90	A
91	95	A
45	42	D
62	57	B

(A)

Marks in science	Marks in maths	Num-grade
78	75	2
56	62	3
87	90	1
91	95	1
45	42	4
62	57	2

(B)

Transforming Numeric (continuous) features to Categorical features:-

we want to treat the real estate Price Prediction Problem, which is a Regression problem, as a real estate price category prediction is a Classification problem.

In this case bin the numerical data into multiple Categories based on the data range. Presently, in this example, the Original data set has a numerical feature apartment Price. It can be transformed to a Categorical variable price-grade.

Apartment-Area	Apartment Price
4,720	23,60,000
2430	12,15,000
4368	21,84,000
3969	19,84,500
6142	30,71,000
7912	39,56,000

Apartment-Area	Apartment- grade
4,720	Medium
2430	Low
4368	Medium
3969	Low
6142	High
7912	high

Apartment-Area	Apartment- grade
4,720	2
2430	1
4368	2
3969	1
6142	3
7912	2

Text-specific Feature Construction:-

Text data is the more Popular Keyword is converted to a numerical representation following a process is known as "Vectorization". In this process, word occurrences in all documents belonging to the corpus are consolidated in the form of bag-of-words. There are 3 steps followed.

(1) Tokenize

(2) Count

(3) Normalize

In order to tokenize a Corpus, the blank spaces and Punctuations are used as delimiters to separate out the words (or) tokens. Then the no. of occurrences of each token is counted, for each document. Tokens are weighted with reducing importance when they occur in the majority of the documents.

A Matrix is then formed with each token representing a Column and a specific document of the Corpus representing each row. Each Cell Contains the count & occurrence of the token in a specific document-This matrix is known as a document-term-matrix".

Ex:

This	House	Build	Feeling	Well	Theatre	Move	Good	Lonely
2	01	1	0	0	1	1	1	0
0	0	0	1	1	0	0	0	0
1	0	0	2	1	1	0	0	1
0	0	0	0	0	0	1	1	0

Feature Extraction:- In feature extraction, new features are created from a Combination of original. features. Some of the commonly used operatory for Combining the original features include:

(1) **For Boolean features:** - Conjunctions, Disjunctions, Negation.

(2) **For Nominal features:** Cartesian Product, M of N

(3) For Numerical features:- Min, Max, Addition, Subtraction, Multiplication, Division, Average, Equivalence, Inequality..

Suppose, we have a data set with a feature set Fi (F1, F2---Fn). After feature extraction using. a mapping function f(F1, F2 --- Fn).

we will have a set of features Fi (F1, F2... Fm)

Such that F; = f(Fi) and man vil

Ex:- Fi = Kifi+k2F2

FEAT-A	FEAT-B	FEAT-C	FEAT -D
34	34.5	23	233
44	45.56	11	344
78	22.59	21	45
22	65.22	11	322.3
22	33.8	355	45.2
11	122.32	63	23.2

Feat 1	Feat 2
41.25	185.80
54.20	53.12
43.73	35.79
65.30	264.10
37.02	238.42
113.39	167.74

Feat-1 = 0.3* Feat A+0.9*Feat A

Feat-2 = Feat A + 0.5 Feat B +0.6 * Feat C

PCA (Principle Component Analysis):

In PCA, a new set of features are extracted from the original features which are quite dissimilar in nature. So, an n-dimensional feature space gets transformed to an im-dimensional feature space, where the dimensions are orthogonal to each other is Completely independent of each other.

PCA-Objectives:-

1) The new features are distinct is the Covariance between the new features is that the

principle Component is 'o'.

.The Principal Components are generated in order of the Variability in the data that it Captures. The 1st Principal Component should capture the maximum Variability, the second principal Component should Capture the next highest variability.

2) The sum of variance of the new features equal to the sum of variance of the original features.

PCA-working Process:- PCA works based on a process Called eigenvalue decomposition of a covariance matrix of a data set.

Steps:-

1. First, Calculate the covariance matrix oba data set.

2. Calculate the eigenvalues of the covariance matrix.

3. The eigenvector having highest eigenvalue represents the direction in which there is the highest Variance It is used for to identify 1st principal Component.

4. similarlly, to identify next 2nd principal Component

5. Identify the top 'K eigenvector having top 'K' eigen valves. So, as to get the 'K' Principal Components-

Singular Valve Decomposition:- (SVD)

It is a Matrix factorization technique, commonly used in linear algebra. SVD of a matrix A (m*n) isa factorization of the form. A=UEV

where u, v are orthogonal normal matrices. U is (m*m) and vis (n*n) unitary matrix.
E is (m*n) rectangular diagonal matrix.
Diagonal entries do & are as singular values de matrix A. The Columns of U and Vis called left, right singular vectors of Matrix "A

SVO is old generally in PCA, once the mean do each variable has been removed. since it is not always advisable to remove the mean do a data attribute, especially, when the data

set is sparse, SVD is a good choice for Dimensionality Reduction in those situations.

SVD of a Data matrix is expected Properties!-

1) Patterns in the attributes are captured by the Right-singular vectors is the columns do 'v'

2) Pattering is among the instances are captured by the Left- singular is the columns of 0.

3) Larger a singular value, larger is the part of the matrix-A that it accounts for and it's associated vectors.

4) New data matrix with K attributes is obtained using the equation. D = DX[V1,V2---VK] The Dimensionality gets reduced to `k'. SVD is often used in the context of text data.

Linear Dis-criminent Analysis: (LDA) : It is another LDA feature extraction technique like PCA (or) SVD. objective of LOA is similar to the sense that it intends to transform a data set into a lower dimensional feature Space.

LDA Calculates eigen Valves and eigen values within a class and inter-class scatter matrices. Following steps is!

(1) Calculates the me an vectors for the individual classes

(2) Calculates intra-Class and inter-class scatter matrices

(3) Calculate eigen values and eigen vectors for Sot and SB. whore Sw is intra class and SB is inter class Satter matrix.

$$Sw = \sum_{i=1}^{c} Si$$

$$Si = n\sum xedi\ (x-mi)\ (x-mi)^r$$

Mi= is the mean vector of i^{th} class.

$$S_B = \sum_{i=1}^{c} Ni\ (m_i-m)\ (m_i-m)^r$$

where mi = mean for each class m = mean of overall.

(4) Identify the top 'K' eigen vectors having top k eigen values.

Feature subset selections:- It is a most critical Pre-Processing activity in any M.L. To

select a subset of system attributes (or) features, which makes a most meaningful contribution in a M.L activity.

Ex: In present problem given that Student "weight"," data set has features as "Roll Numbers, Age, Height, weigh Rill Number Can have no bearing, in predicting student weight. So, we can eliminate feature of roll number and build a feature subset.

Roll no	Age	Height	Weight
12	12	1.1	23
14	11	1.05	21.6
19	13	1.2	24.7
32	11	1.07	21.3

Age	Height	Weight
12	1.1	23
11	1.05	21.6
13	1.2	24.7
11	1.07	21.3

Key drivers of Feature selection - Feature relevance and Redundancy:-

Feature relevance:- It refers to how useful (or) important a feature (Variable is for Predicting the target (output). Not all features Contribute equally, Some may be strongly Predictive, others weakly predictive and Some may even add noise.

strongly relevant features:- Directly impact the prediction & the target. Removing them significantly reduces model performance..

Ex:- In Predicting house prices, "square footage" is strongly relevant.

weakly relevant features:- Provides some Predictive power but are not essential. May be redundant if other Strongly relevant Feature exist.

Ex:- "Number of rooms" may be weakly relevant if "square footage" is already included. Irrelevant features:- Do not contribute to predicting the target at all. Ex House color for Predicting price.

Feature Redundancy:- It refers to the situation where 2 (or) more features provide the same or overlapping information about the target variable.

(1) A redundant feature is not useless, but it does not add much new information. Because another feature already Captures it.

(2) Redundancy increases the dimensionality of the dataset unnecessarily, which can lead to

(1) over fitting. (2) Longer training times .(3) Difficulty in model interpretation.

Measures of Feature relevance and redundancy:-

Measures of Feature Relevance:- Feature relevance relevance is to be gauged by the amount of information Contributed by a feature.

For Supervised learning, Mutual information is considered as a good measure of information contribution & a feature to decide the value of the class label. It is a good indicator of the relevance of a feature with respect to the class Variable. Higher the valve of mutual information of a feature, more relevant is that feature.

Mutual information can be calculated as follows: MI (c,f) = H (C) + H (f) -H (C, f) where Marginal entropy of class $H(C)=-k\sum P(Ci)\ log_2 P(Ci)$

i=1 Marginal entropy of feature $H(f)_{=-c} \sum P(f=x)\ log2$

K=no.& classes; C=Class variable

f: feature set that take discrete Valves.

For un-supervised learning, there is no class variable. Hence, feature-to-Class mutual information cannot be used to measure the information contribution of the features.

Measures of Feature redundancy:-

Feature redundancy is based on similar information Contribution by multiple features. There are multiple measures & Similarity & information Contribution.

(1) Correlation - based measures 2)Distance- based measures

(2) other Coefficient measures

Correlation baled similarity measure:-

It is a measure of linear dependency between 2 random Variables. Suppose F1, F2 is 2 Random Variables. The Pearson's correlation coefficient is

$$\alpha = \frac{Cov(F_1, F_2)}{\sqrt{Var(F_1) \cdot Var(F_2)}}$$

$$Cov(F_1, F_2) = \Sigma \left(F_1 - \bar{F_1}\right) \cdot \left(F_2 - \bar{F_2}\right)$$

$$Var(F_1) = \Sigma (F_1 - \bar{F_1})^2 \cdot , \quad \bar{F} = \frac{1}{n} \cdot \Sigma F_{1i}$$

$$Var(F_2) = \Sigma (F_2 - \bar{F_2})^2 , \quad \bar{F} = \frac{1}{n} \cdot \Sigma F_{2i}$$

Correlation Values range between +1 and 1.

Suppose A Correlation 1 (t/-) indicates Perfect, otherwise. '0' & Correlation is , then the features to have no linear relationship.

Distance bated Similarity Measure:-

Most Common distance measure is "Euclidean distance". between F1, F2 random variables is

$$D(f1,f2) = \sqrt{\sum_{i=0}^{n} (f1i - f2i)}$$

Aptitude (f1)	Communication (f2)	(f1-f2)	(f1-f2) ^2
2	6	-4	16
3	5.5	-2.5	6.25
6	4	2	4
7	2.5	4.5	20.25
8	3	5	25
6	5.5	0.5	0.25
6	7	-1	1
7	6	1	1
8	6	2	4
9	7	2	4
			81.75

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

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

other similarity Measures-

Jaccard index / Coefficient is used as a measure de similarity between 2 features.

Jaccard distance is a measure of dissimilarity between 2 features, is complementary of Jaccard index.

For 2 features having Binary Values.

Jaccard index is J = **n11** ,

n01+ n10+n11

where n11 = no of cases where both features have value '1' n01= no of

cases where both features 1 have value '0'

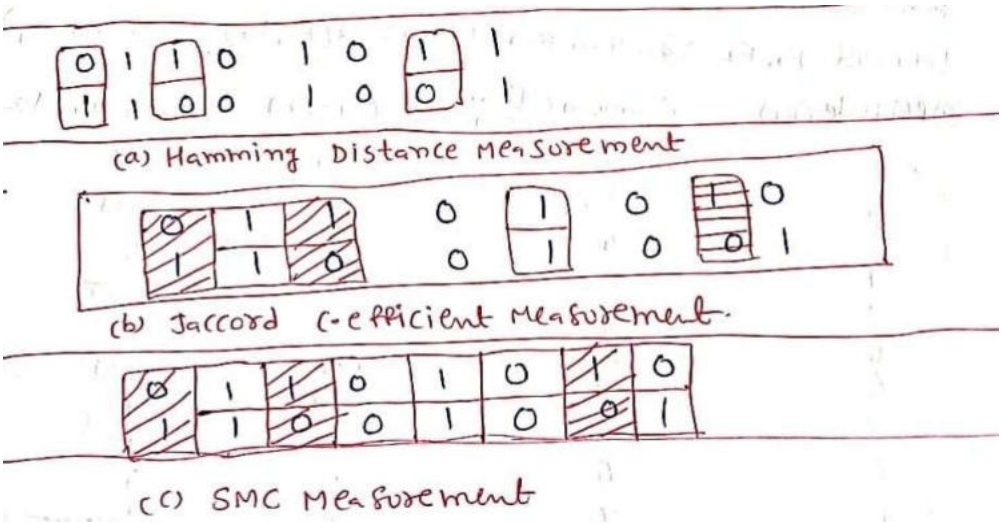
.....

.....

.....

.....

n10 = no of cases where both features 2 have value '0' Binary Numbers are →01101011 and 1100/001



Jaccard Coefficient of F_1 and F_2, $J = \frac{n_{11}}{n_{01}+n_{10}+n_{11}} = \frac{2}{1+2+2} = \frac{2}{5}$

Jaccard distance between F_1 and F_2, $d_J = 1 - J = \frac{1}{2} = 0.6$

SMC (Simple matching co-efficient):-

It is also same as Jaccord Coefficient, except the fact that it includes a no. of Cases, where both features

$$SMC = \frac{n_{11}+n_{00}}{n_{00}+n_{01}+n_{10}+n_{11}}$$

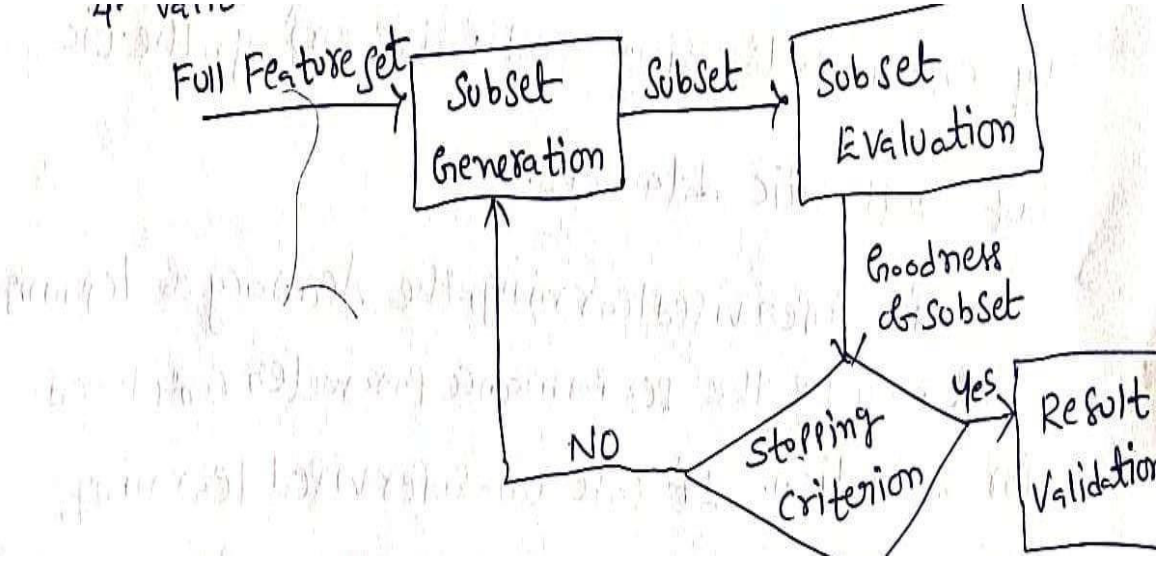
$$SMC = \frac{2+3}{3+1+2+2} = \frac{5}{8} = \frac{1}{2} \text{ (or) } 0.5$$

have a Value of 'o'

Overall Feature selection Process:

It is the process of selecting a Subset of features in a data set. In feature selection Process 4 steps available.

1. Generation of Possible subsets
2. Subset Evaluation
3. Stop Searching based on some stopping criterion
4. Validation of the result.



1. Subset Generation'- A Search Procedure which ideally should Produce all possible Candidate subsets. 'n'-dimensional data set, 20 Subsets can be Generated. Some cases "sequential. forward

selection and sometimes sequential backward elimination Process Conducted. So, this strategy termed as Bi-directional selection.

2. Evaluation Criteria- Each Candidate subset is then evaluated and Compared with the previous best Performing subset based on certain "evaluation criterion. If new subset performs better, it replaces Previous one.

3. Stop searching criteria! This cycle of subset generations and evaluation continues till a pre-defined stopping Criterion is fulfilled. Some commonly used stopping Criteria are:

(1) The Search Completes

(2) Some given bound is reached (no.co iterations. Specified)

(3) Subsequent addition of the feature is not producing a better subset

(4) A sufficiently good subset.

4. Validate the result: The selected best subset is Validated either against prior benchmarks (or) by experiments using real life (or) synthetic but authentic data sets.

If Supervised learning, the Accuracy of learning motel may be the performance parameter considered for validation. If Cafe un-supervised learning, the cluster quality may be the Parameter for Validation.

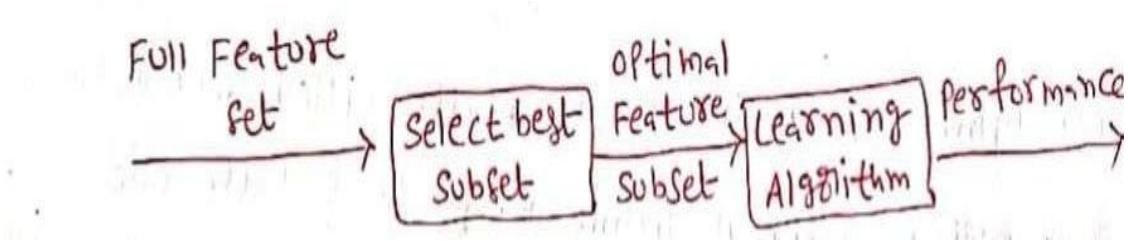
Feature selection Approaches These are 4 types

1.Filter Approach

2 Wrapper Approach

3. Hybrid Approach

4. Embedded Approach

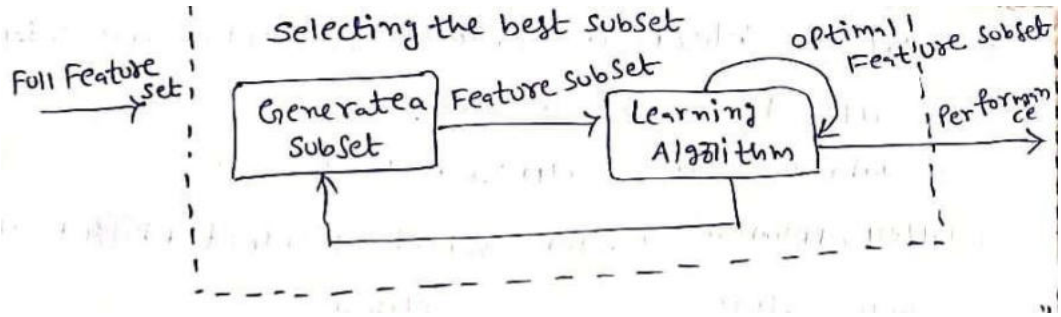


1. Filter Approach- Feature subset is selected based on Statistical measures done to assess the merits of the features from the data perspective. No Learning algorithm is employed to evaluate the goodness of feature selected

Ex 'n' Statistical tests conducted on features as part co filter approach. Pearson Correlation, Chi-square test, ANOVA...

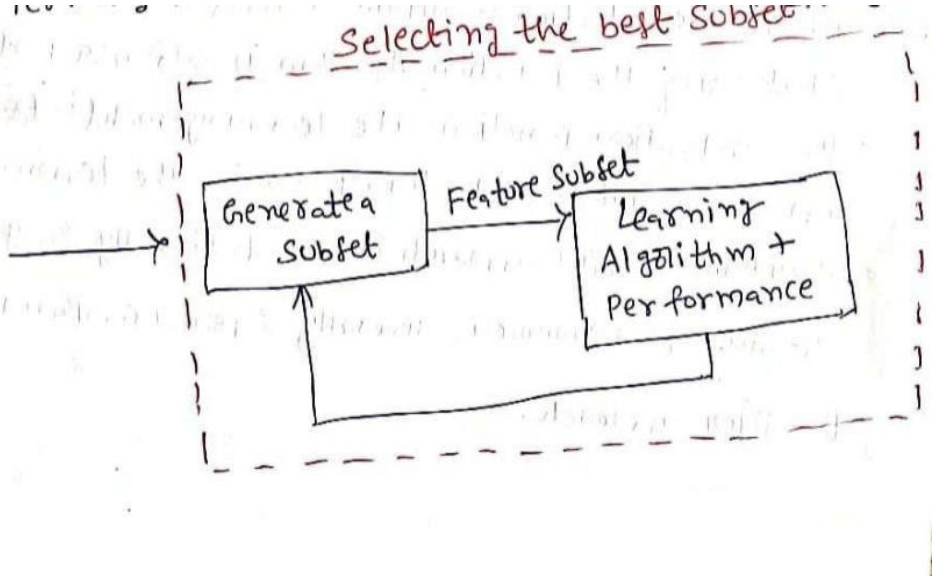
2 Wrapper Approach: Identification of best feature Subset is done using the induction algorithm as a

block box The feature selection algorithm searches for a good feature subset using the induction algorithm it self as a Part of the evaluation function. The learning motel is trained and the result is evaluated by running the learning algorithm, wrapper approach is Computationally very expensive. Performance is generally superior compared to filter approach.



3. Hybrid Approach:- It is used both "Filter and wrapper approach A typical Hybrid algorithm makes use of both the Statistical tests as used in filter approach to recite the best subsets for a given Cardinality and a learning Algorithm. to select the final best subset among the best sub sets across different Cardinalities.

4. Embedded Approach: It is similar to wrapper approach as it also uses and inductive algorithm to evaluate the generated feature subsets. The difference is it Performs feature selection and Classification Simultaneously.



----------END----------

UNIT – III

Bayesian Concept Learning

Introduction, Bayes Theorem: Prior, Posterior , Likelihood, Bayes Theorem Concept Learning : Concept of Consistent Learners , Bayes Optimal Classifier, Naïve Bayes Classifier , Applications of Naïve Bayes Classifier.

Supervised Learning: Classification: Introduction, Classification Model, Classification Learning Steps, Common Classification Algorithms-k-Nearest Neighbor (k-NN), Decision tree, Random forest model.

Introduction

- ✓ Bayes developed the foundational mathematical principles, known as Bayesian methods, which describe the probability of events, and more importantly, how probabilities should be revised when there is additional information available.

Hypothesis

- ✓ A supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.
- ✓ A hypothesis is a proposed explanation for a phenomenon

Features of Bayesian learning methods

Some of the features of Bayesian learning methods that have made them popular are as follows:

1. Text-based classification such as spam or junk mail filtering, author identification, or topic categorization
2. Medical diagnosis such as given the presence of a set of observed symptoms during a disease, identifying the probability of new patients having the disease.
3. Prior knowledge of the candidate hypothesis is combined with the observed data for arriving at the final probability of a hypothesis
4. The Bayesian approach to learning is more flexible than the other approaches because each observed training pattern can influence the outcome of the hypothesis by increasing or decreasing the estimated probability about the hypothesis, whereas most of the other algorithms

tend to eliminate a hypothesis if that is inconsistent with the single training pattern.

5. Bayesian methods can perform better than the other methods while validating the hypotheses that make probabilistic predictions.
6. Through the easy approach of Bayesian methods, it is possible to classify new instances by combining the predictions of multiple hypotheses, weighted by their respective probabilities.
7. In some cases, when Bayesian methods cannot compute the outcome deterministically, they can be used to create a standard for the optimal decision against which the performance of other methods can be measured.

BAYES' THEOREM

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' probability rule

where A and B are conditionally related events and p(A|B) denotes the probability of event A occurring when event B has already occurred.

Let us assume that we have a training data set D where we have noted some observed data. Our task is to determine the best hypothesis in space H by using the knowledge of D.

Prior

The prior knowledge or belief about the probabilities of various hypotheses in H is called Prior in context of Bayes' theorem.

P(h) is the initial probability of a hypothesis 'h' that the patient has a malignant tumor based only on the malignancy test, without considering the prior knowledge of the correctness of the test process or the so-called training data.

P(T) is the prior probability that the training data will be observed or, in this case, the probability of positive malignancy test results.

P(T|H) as the probability of observing data T in a space where 'h' holds true, which means the probability of the test results showing a positive value when the tumor is actually malignant

Posterior

The probability that a particular hypothesis holds for a data set based on the Prior is called the posterior probability or simply Posterior.

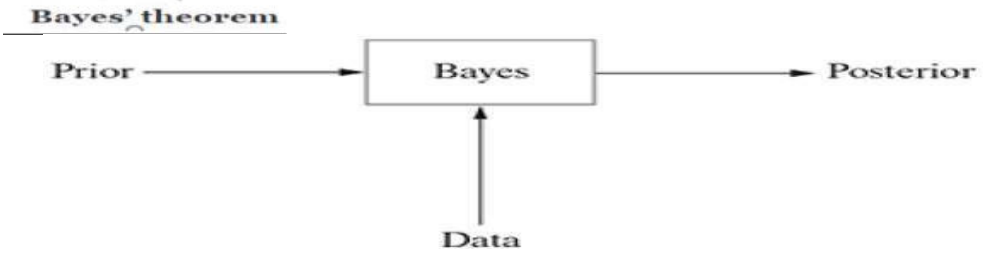
In our notation, we will say that we are interested in finding out P(H|T), which means whether the hypothesis holds true given the observed training data T.

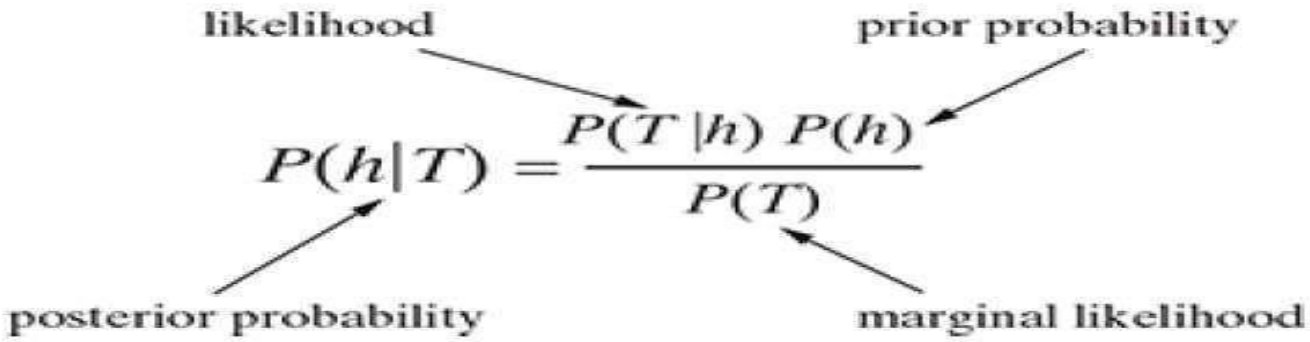
This is called the posterior probability or simply Posterior in machine learning language. According to

Bayes' theorem, combines the prior and posterior probabilities together.

$$P(h|T) = \frac{P(T|h)P(h)}{P(T)}$$

Likelihood:-





Bayes optimal classifier

we will discuss the use of the MAP hypothesis to answer the question what is the most probable

So, extending the above example,

The set of possible outcomes for the new instance x is within the set $C = \{\text{True}, \text{False}\}$ and

$$P(h_1 \mid T) = 0.4,\ P(\text{False} \mid h_1) = 0,\ P(\text{True} \mid h_1) = 1$$
$$P(h_2 \mid T) = 0.3,\ P(\text{False} \mid h_2) = 1,\ P(\text{True} \mid h_2) = 0$$
$$P(h_3 \mid T) = 0.3,\ P(\text{False} \mid h_3) = 1,\ P(\text{True} \mid h_3) = 0$$

Then,

$$\sum_{h_i \in H} P(\text{True} \mid h_i) P(h_i \mid T) = 0.4$$

classification of the new instance given the training data. To illustrate the concept, let us assume three hypotheses h1, h2, and h3 in the hypothesis space H. Let the posterior probability of these hypotheses be 0.4, 0.3, and 0.3, respectively. There is a new instance x, which is classified as true by h1 , but false by

h2 and h3. Then the most probable classification of the new instance (x) can be obtained by combining the predictions of all hypotheses weighed by their corresponding posterior probabilities. By denoting the possible classification of the new instance as c_i from the set C, the probability $P(c_i|T)$ that the correct classification for the new instance is c_i is

$$P(c_i|T) = \sum_{h_i \in H} P(c_i|h_i)P(h_i|T)$$

The optimal classification is for which $P(c_i|T)$ is maximum is

$$\text{Bayes optimal classifier} = \underset{c_i \in C}{\text{argmax}} \sum_{h_i \in H} P(c_i|h_i)P(h_i|T)$$

Naïve Bayes classifier

- ✓ A Naïve Bayes classifier is a primary probabilistic classifier based on a view of applying Bayes' theorem (from Bayesian inference with strong naive) independence assumptions.
- ✓ The prior probabilities in Bayes' theorem that are changed with the help of newly available information are classified as posterior probabilities.

$$\underset{c_i \in \{\text{True, False}\}}{\text{argmax}} \sum_{h_i \in H} P(c_i|h_i)P(h_i|T) = \text{False}$$

- ✓ A key benefit of the naive Bayes classifier is that it requires only a little bit of training information (data) to gauge the parameters (mean and differences of the variables) essential for the classification (arrangement).
- ✓ In the Naïve Bayes classifier, independent variables are always assumed, and only the changes (variances) of the factors/variables for each class should be determined and not the whole covariance matrix.
- ✓ Because of the rather naïve assumption that all features of the dataset are equally important and

$$\sum_{h_i \in H} P(\text{False}|h_i)P(h_i|T) = 0.6$$

and

independent, this is called Naïve Bayes classifier.

Naïve Bayes is a simple technique for building classifiers: models that assign class labels to problem instances. The basic idea of Bayes rule is that the outcome of a hypothesis can be predicted on the basis

$$\textbf{Posterior probability} = \frac{\textbf{(Prior probability} \times \textbf{Conditional Probability)}}{\textbf{Evidence}}$$

Posterior Probability is of the format 'What is the probability that a particular object belongs to class i given its observed feature values?'

$$C_{NB} = \underset{c_i \in C}{\operatorname{argmax}} \sum_{h_i \in H} P(c_i) \prod_i P(a_i|c_j)$$

of some evidence (E) that can be observed.

Strengths and Weaknesses of Bayes Classifiers

Strengths	Weakness
Simple and fast in calculation but yet effective in result	The basis assumption of equal importance and independence often does not hold true
In situations where there are noisy and missing data, it performs well	If the target dataset contains large numbers of numeric features, then the reliability of the outcome becomes limited
Works equally well when smaller number of data is present for training as well as very large number of training data is available	Though the predicted classes have a high reliability, estimated probabilities have relatively lower reliability
Easy and straightforward way to obtain the estimated probability of a prediction	

Naive Bayes classifier steps

Step 1: First construct a frequency table. A frequency table is drawn for each attribute against the target outcome. For example, in Figure 6, the various attributes are (1) Weather Condition,

(2) How many matches won by this team in last three matches, (3) Humidity Condition, and (4) whether they won the toss and the target outcome is will they win the match or not?

Step 2: Identify the cumulative probability for 'Won match = Yes' and the probability for 'Won match = No' on the basis of all the attributes. Otherwise, simply multiply probabilities of all favourable conditions to derive 'YES' condition. Multiply probabilities of all non-favourable conditions to derive 'No' condition.

Step 3: Calculate probability through normalization by applying the below formula

$$P(\text{Yes}) = \frac{P(\text{Yes})}{P(\text{Yes}) + P(\text{No})}$$

$$P(\text{No}) = \frac{P(\text{No})}{P(\text{Yes}) + P(\text{No})}$$

P(Yes) will give the overall probability of favourable condition in the given scenario.

P(No) will give the overall probability of non-favourable condition in the given scenario.

Solving the problem with Naïve Bayes Classifier

Refer below table for the training dataset.

Step 1: Construct a frequency table. The posterior probability can be easily derived by constructing a frequency table for each attribute against the target. For example, frequency of Weather Condition variable with values 'Sunny' when the target value Won match is 'Yes', is, 3/(3+4+2) = 3/9.

Training dataset

	Won Match			Won Match	
Wins in last 3 matches	Yes	No	Win toss	Yes	No
3 wins	2	2	FALSE	6	2
1 win	4	2	TRUE	3	3
2 wins	3	1			
Total	9	5	Total	9	5

	Won Match			Won Match	
Weather condition	Yes	No	**Humidity**	Yes	No
Sunny	3	2	High	3	4
OverCast	4	0	Normal	6	1
Rainy	2	3			
Total	9	5	Total	9	5

Frequency Table :

Weather Condition	Wins in last 3 matches	Humidity	Win toss	Won match?
Rainy	3 wins	High	FALSE	No
Rainy	3 wins	High	TRUE	No
OverCast	3 wins	High	FALSE	Yes
Sunny	2 wins	High	FALSE	Yes
Sunny	1 win	Normal	FALSE	Yes
Sunny	1 win	Normal	TRUE	No
OverCast	1 win	Normal	TRUE	Yes
Rainy	2 wins	High	FALSE	No
Rainy	1 win	Normal	FALSE	Yes
Sunny	2 wins	Normal	FALSE	Yes
Rainy	2 wins	Normal	TRUE	Yes
OverCast	2 wins	High	TRUE	Yes
OverCast	3 wins	Normal	FALSE	Yes
Sunny	2 wins	High	TRUE	No

Step 2:

To predict whether the team will win for given weather conditions (a1) = Rainy, Wins in last three matches (a2) = 2 wins, Humidity (a3) = Normal and Win toss (a4) = True, we need to choose 'Yes' from the above table for the given conditions.

From Bayes' theorem, we get

$$P(\text{Win match}|a_1 \cap a_2 \cap a_3 \cap a_4) = \frac{P(a_1 \cap a_2 \cap a_3 \cap a_4|\text{Win match})P(\text{Win match})}{P(a_1 \cap a_2 \cap a_3 \cap a_4)}$$

$$P(\text{Win match}|a_1 \cap a_2 \cap a_3 \cap a_4) =$$

$$= \frac{P(a_1|\text{Win match})P(a_2|\text{Win match})P(a_3|\text{Win match})P(a_4|\text{Win match})P(\text{Win match})}{P(a_1)P(a_2)P(a_3)P(a_4)}$$

$$= 2/9*4/9*6/9*9/14$$

$$= 0.014109347$$

This should be compared with

$$P(!\text{Win match}|a_1 \cap a_2 \cap a_3 \cap a_4)$$

$$= \frac{P(a_1|!\text{Win match})P(a_2|!\text{Win match})P(a_3|!\text{Win match})P(a_4|!\text{Win match})P(!\text{Win match})}{P(a_1)P(a_2)P(a_3)P(a_4)}$$

$$= 3/5*2/5*1/5*5/14$$

$$= 0.010285714$$

Step 3: by normalizing the above two probabilities, we can ensure that the sum of these two probabilities is 1.

$$P(\text{Win match}) = \frac{P(\text{Win match})}{P(\text{Win match}) + P(!\text{Win match})}$$

$$= \frac{0.014109347}{0.014109347 + 0.010285714}$$

$$= 0.578368999$$

$$P(!\text{Win match}) = \frac{P(!\text{Win match})}{P(\text{Win match}) + P(!\text{Win match})}$$

$$= \frac{0.010285714}{0.014109347 + 0.010285714}$$

$$= 0.421631001$$

Applications of Naïve Bayes classifier

1. **Text classification**: Naïve Bayes classifier is among the most successful known algorithms for learning to classify text documents.
2. **Spam filtering:** Spam filtering is the best known use of Naïve Bayesian text classification. Presently, almost all the email providers have this as a built-in functionality, which makes use of a Naïve Bayes classifier to identify spam email on the basis of certain conditions and also the probability of classifying an email as 'Spam'.
3. **Hybrid Recommender System:** It uses Naïve Bayes classifier and collaborative filtering. Recommender systems (used by e-retailors like eBay, Alibaba, Target, Flipkart, etc.) apply machine learning and data mining techniques for filtering unseen information and can predict whether a user would like a given resource. For example, when we log in to these retailer websites, on the basis of the usage of texts used by the login and the historical data of purchase, it automatically recommends the product for the particular login persona**.**
4. **Online Sentiment Analysis:** The online applications use supervised machine learning (Naïve Bayes) and useful computing. In the case of sentiment analysis, let us assume there are three sentiments such as nice, nasty, or neutral, and Naïve Bayes classifier is used to distinguish between them. Simple emotion modelling combines a statistically based classifier with a dynamical model.

Supervised Learning: Classification

Some examples of supervised learning are as follows

a) Prediction of results of a game based on the past analysis of results
b) Predicting whether a tumor is malignant or benign on the basis of the analysis of data
c) Price prediction in domains such as real estate, stocks, etc.

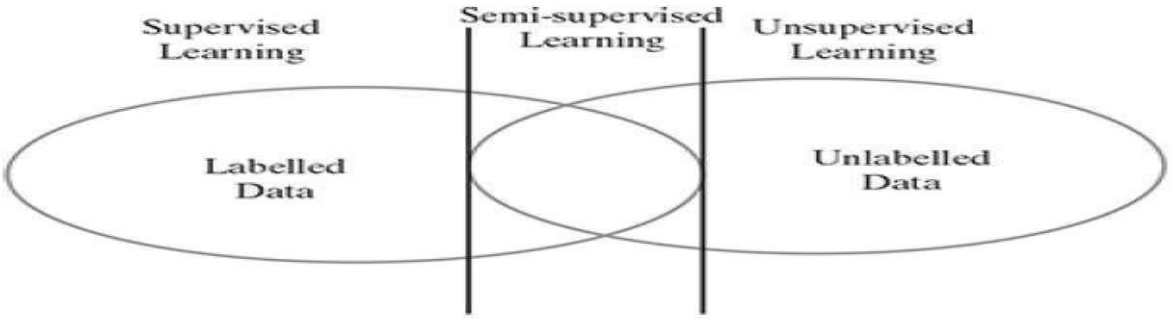
CLASSIFICATION MODEL

We can observe that in classification, the whole problem centres around assigning a label or category or class to a test data on the basis of the label or category or class information that is imparted by the training data. Because the target objective is to assign a class label, we call this type of problem as a classification problem.

Classification is a type of supervised learning where a target feature, which is of categorical type, is predicted for test data on the basis of the information imparted by the training data. The target categorical feature is known as class.

Some typical classification problems include the following:

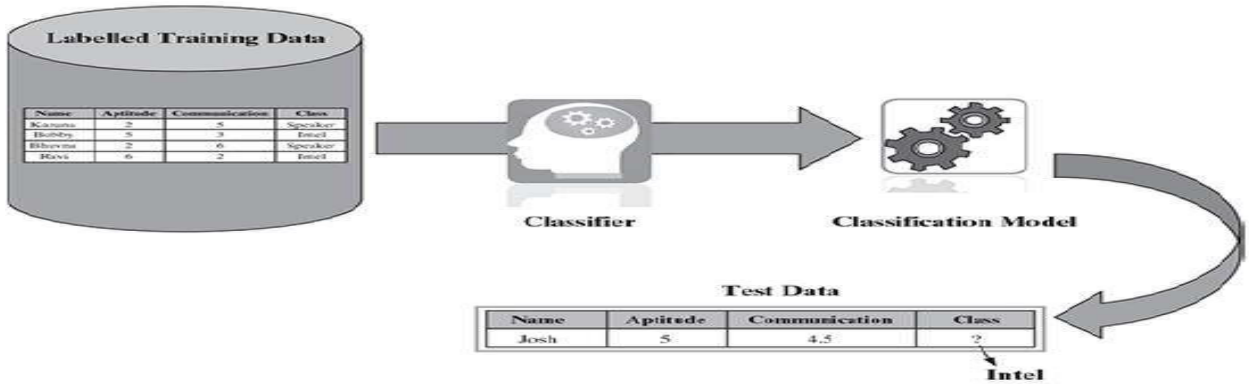
- ✓ Image classification
- ✓ Disease prediction



Supervised learning vs. unsupervised learning

- ✓ Win–loss prediction of games
- ✓ Prediction of natural calamity such as earthquake, flood, etc.
- ✓ Handwriting recognition

Classification Model



CLASSIFICATION LEARNING STEPS

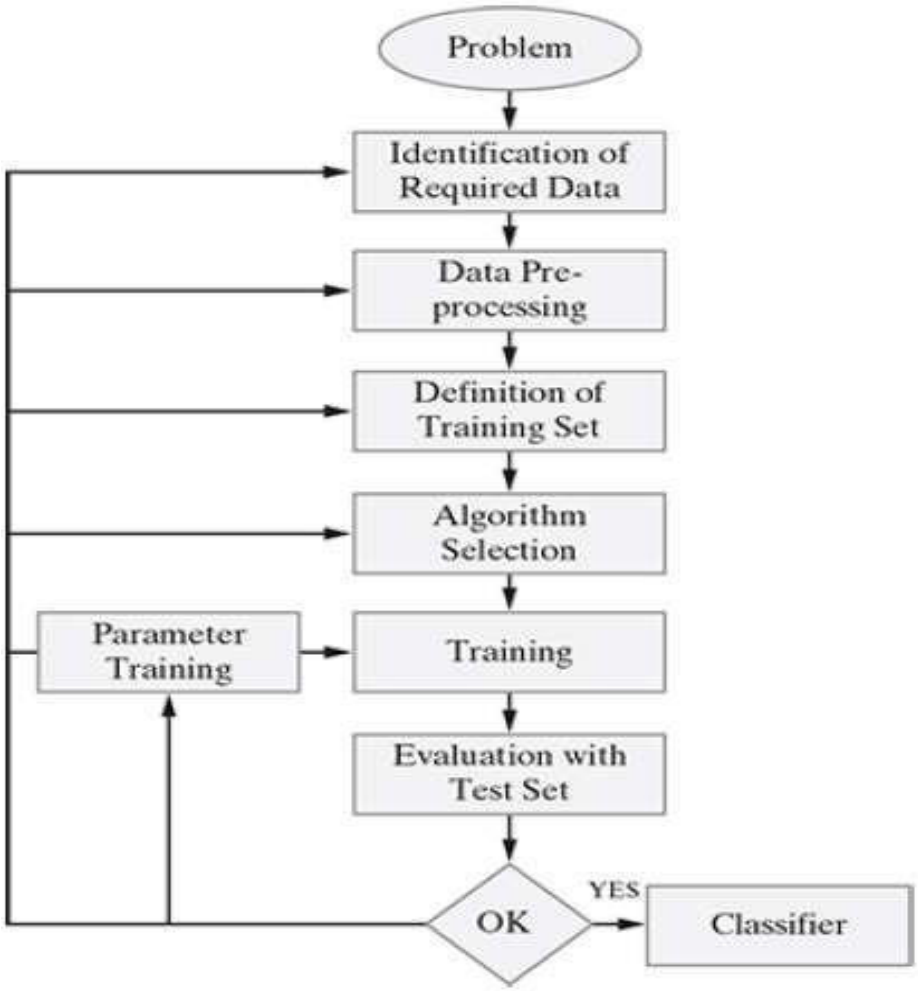
Following steps are followed in classification learning model.

1) **Problem Identification:** Identifying the problem is the first step in the supervised learning model. The problem needs to be a well-formed problem.
2) **Identification of Required Data:** On the basis of the problem identified above, the required data set that precisely represents the identified problem needs to be identified/ evaluated.
3) **Data Pre-processing:** This is related to the cleaning/transforming the data set. This step ensures that all the unnecessary/irrelevant data elements are removed. Data pre- processing refers to the transformations applied to the identified data before feeding the same into the algorithm.

4) **Definition of Training Data Set:** Before starting the analysis, the user should decide what kind of data set is to be used as a training set.

5) Algorithm Selection: This involves determining the structure of the learning function and the corresponding learning algorithm. This is the most critical step of supervised learning model. On the basis of various parameters, the best algorithm for a given problem is chosen.

6) Training: The learning algorithm identified in the previous step is run on the gathered training set for further fine tuning. Some supervised learning algorithms require the user to determine specific control parameters (which are given as inputs to the algorithm). These parameters (inputs given to algorithm) may also be adjusted by optimizing performance on a subset (called as validation set) of the training set.



7) Evaluation with the Test Data Set: Training data is run on the algorithm, and its performance is measured here. If a suitable result is not obtained, further training of parameters may be required.

COMMON CLASSIFICATION ALGORITHMS

Following are the common classification algorithms –

1. k-Nearest Neighbor (k-NN)
2. Decision tree
3. Random forest
4. Support Vector Machine (SVM)
5. Naïve Bayes classifier

a) k-Nearest Neighbor (k-NN)

- The k-NN algorithm is a simple but extremely powerful classification algorithm.
- The name of the algorithm originates from the underlying philosophy of k-NN — i.e. people having similar background or mindset tend to stay close to each other. In other words, neighbors in a locality have a similar background.
- In the same way, as a part of the k-NN algorithm, the unknown and unlabeled data which comes for a prediction problem is judged on the basis of the training data set elements which are similar to the unknown element.

How KNN Works?

Input: Training data set, test data set (or data points), value of '*k*' (i.e. number of nearest neighbours to be considered)

Steps:

Do for all test data points

Calculate the distance (usually Euclidean distance) of the test data point from the different training data points.
Find the closest '*k*' training data points, i.e. training data points whose distances are least from the test data point.

If $k = 1$

Then assign class label of the training data point to the test data point

Else

Whichever class label is predominantly present in the training data points, assign that class label to the test data point

End do

Working factor of choose the Factor K

- It is often a tricky decision to decide the value of k. The reasons are as follows:
    - If the value of k is very large (in the extreme case equal to the total number of records in the training data), the class label of the majority class of the training data set will be assigned to the test data regardless of the class labels of the neighbors nearest to the test data.
    - If the value of k is very small (in the extreme case equal to 1), the class value of a noisy data or outlier in the training data set which is the nearest neighbor to the test data will be assigned to the test data.
- The best k value is somewhere between these two extremes.

Few strategies, highlighted below, are adopted by machine learning practitioners to arrive at a value for k.

1) One common practice is to set k equal to the square root of the number of training records.
2) An alternative approach is to test several k values on a variety of test data sets and choose the one that delivers the best performance.
3) Another interesting approach is to choose a larger value of k, but apply a weighted voting process in which the vote of close neighbors is considered more influential than the vote of distant neighbors.

Example

Student dataset

Name	Aptitude	Communication	Class
Karuna	2	5	Speaker
Bhuvna	2	6	Speaker
Gaurav	7	6	Leader
Parul	7	2.5	Intel
Dinesh	8	6	Leader
Jani	4	7	Speaker
Bobby	5	3	Intel
Parimal	3	5.5	Speaker
Govind	8	3	Intel
Susant	6	5.5	Leader
Gouri	6	4	Intel
Bharat	6	7	Leader
Ravi	6	2	Intel
Pradeep	9	7	Leader
Josh	5	4.5	Intel

Segregated student data set

	Name	Aptitude	Communication	Class
Training Data	Karuna	2	5	Speaker
	Bhuvna	2	6	Speaker
	Gaurav	7	6	Leader
	Parul	7	2.5	Intel
	Dinesh	8	6	Leader
	Jani	4	7	Speaker
	Bobby	5	3	Intel
	Parimal	3	5.5	Speaker
	Govind	8	3	Intel
	Susant	6	5.5	Leader
	Gouri	6	4	Intel
	Bharat	6	7	Leader
	Ravi	6	2	Intel
	Pradeep	9	7	Leader
Test Data →	Josh	5	4.5	Intel

Modeling

- In the *k*-NN algorithm, the class label of the test data elements is decided by the class label of the training data elements which are neighboring, i.e. similar in nature.

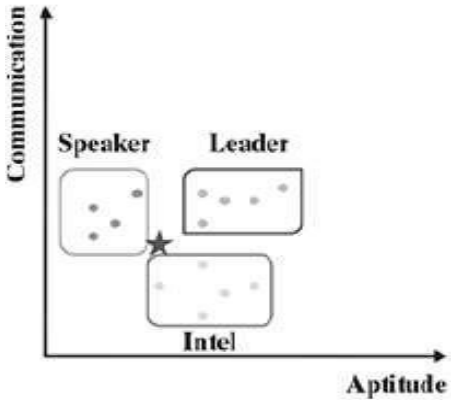
But there are two challenges:

1. What is the basis of this similarity or when can we say that two data elements are similar?
2. How many similar elements should be considered for deciding the class label of each test data element?

- The most common approach adopted by *k*-NN to measure similarity between two data elements is **Euclidean distance**.

Euclidean Distance assumes that an object can travel freely in space and reach point $B(x_1, y_1)$ from point $A(x_2, y_2)$ in a straight line.

The formula for the euclidian distance is : $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$



Name	Aptitude	Communication	Class
Karuna	2	5	Speaker
Bhuvna	2	6	Speaker
Gaurav	7	6	Leader
Parul	7	2.5	Intel
Dinesh	8	6	Leader
Jani	4	7	Speaker
Bobby	5	3	Intel
Parimal	3	5.5	Speaker
Govind	8	3	Intel
Susant	6	5.5	Leader
Gouri	6	4	Intel
Bharat	6	7	Leader
Ravi	6	2	Intel
Pradeep	9	7	Leader
Josh	5	4.5	???

- ✓ In the *k*NN algorithm, the value of '*k*' indicates the number of neighbors that need to be considered.

Name	Aptitude	Communication	Class	Distance	*k* = 1	*k* = 2	*k* = 3
Karuna	2	5	Speaker	3.041			
Bhuvna	2	6	Speaker	3.354			
Parimal	3	5.5	Speaker	2.236			
Jani	4	7	Speaker	2.693			
Bobby	5	3	Intel	1.500			1.500
Ravi	6	2	Intel	2.693			
Gouri	6	4	Intel	1.118	1.118	1.118	1.118
Parul	7	2.5	Intel	2.828			
Govind	8	3	Intel	3.354			
Susant	6	5.5	Leader	1.414			
Bharat	6	7	Leader	2.693			
Gaurav	7	6	Leader	2.500			
Dinesh	8	6	Leader	3.354			
Pradeep	9	7	Leader	4.717			
Josh	5	4.5	???				

- ✓ For example, if the value of *k* is 3, only three nearest neighbors or three training data elements closest to the test data element are considered. Out of the three data elements, the class which is predominant is considered as the class label to be assigned to the test data.
- ✓ In case the value of k is 1, only the closest training data element is considered. The class label of that data element is directly assigned to the test data element.

Why KNN algorithm is called as Lazy Learner?

- ✓ It only stores a training dataset versus undergoing a training stage.
- ✓ This also means that all the computation occurs when a classification or prediction is being made.
- ✓ Since it heavily relies on memory to store all its training data, it is also referred to as an instance-based or memory-based learning method.

Strength and Weaknesses of KNN algorithm

Strengths

- ✓ Extremely simple algorithm – easy to understand
- ✓ Very effective in certain situations, e.g. for recommender system design
- ✓ Very fast or almost no time required for the training phase

Weakness

- ✓ Does not learn anything in the real sense. Classification is done completely on the basis of the training data.
- ✓ Does not scale well: Since KNN is a lazy algorithm, it takes up more memory and data storage compared to other classifiers. This can be costly from both a time and money perspective.
- ✓ Curse of dimensionality: The KNN algorithm tends to fall victim to the curse of dimensionality, which means that it doesn't perform well with high-dimensional data inputs.

Applications

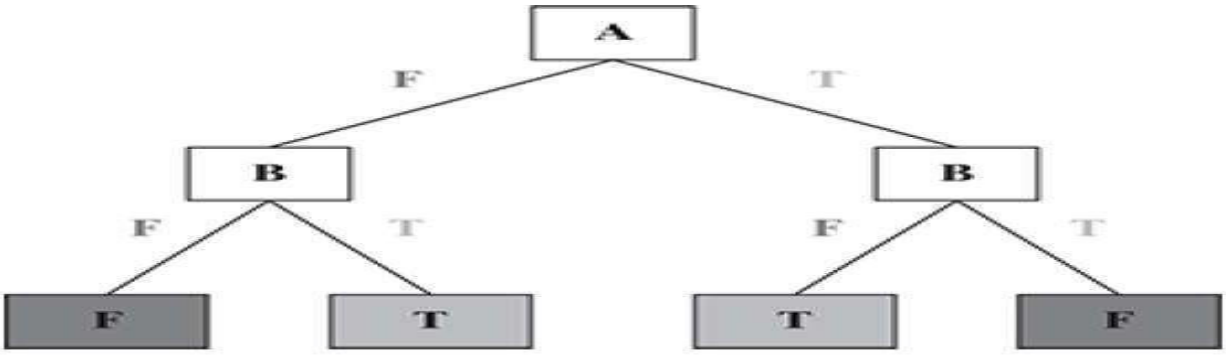
1) Recommendation Engines
2) Healthcare
3) Searching documents/ contents similar to a given document/content.

b) Decision tree algorithm

- ✓ Decision tree learning is one of the most widely adopted algorithms for classification. As the name indicates, it builds a model in the form of a tree structure.
- ✓ A decision tree is used for multi-dimensional analysis with multiple classes. It is characterized by fast execution time and ease in the interpretation of the rules.
- ✓ Each node (or decision node) of a decision tree corresponds to one of the feature vector. From every node, there are edges to children, wherein there is an edge for each of the possible values

(or range of values) of the feature associated with the node.

- ✓ The tree terminates at different leaf nodes (or terminal nodes) where each leaf node represents a possible value for the output variable. The output variable is determined by following a path that starts at the root and is guided by the values of the input variables.



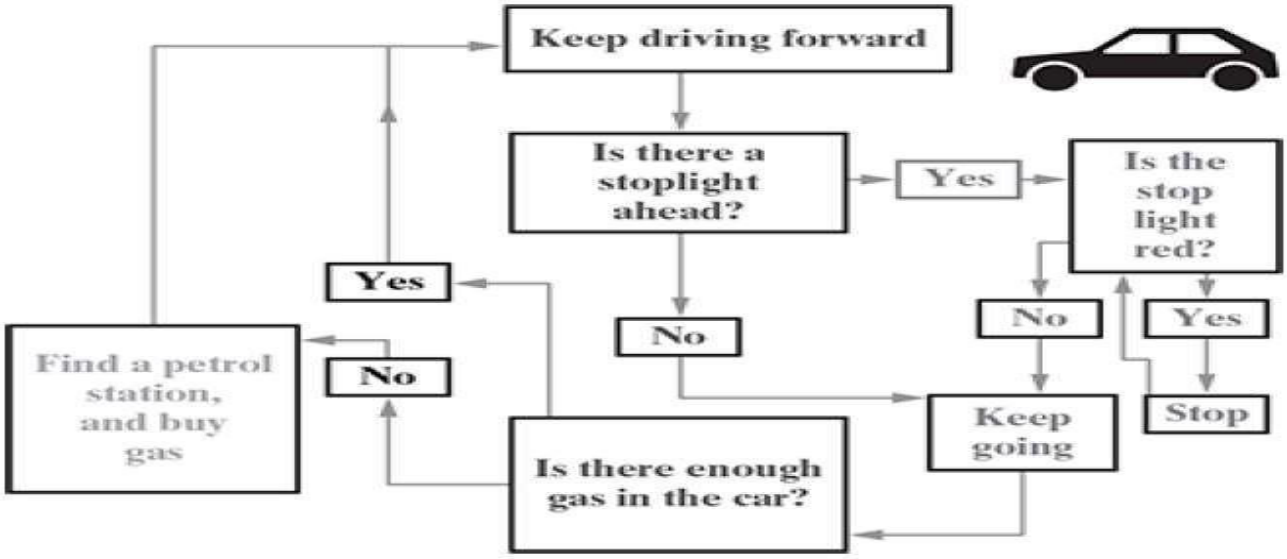
Decision tree structure

A decision tree consists of three types of nodes:

- Root Node
- Branch Node
- Leaf Node

Example

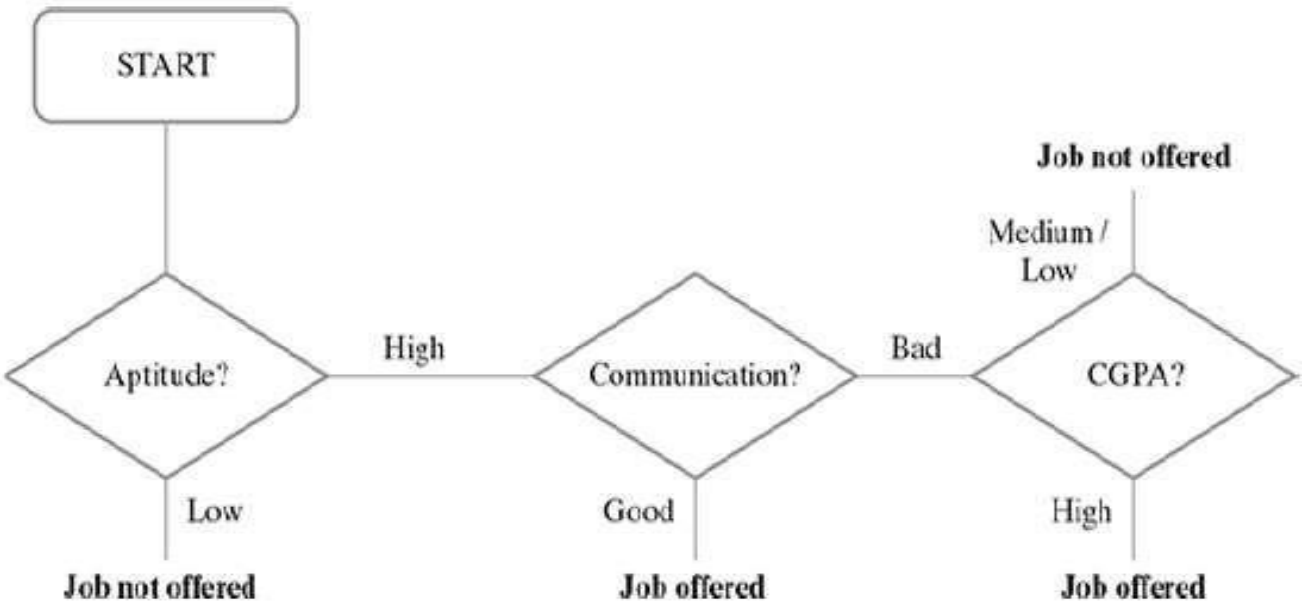
An example decision tree for a car driving – the decision to be taken is whether to 'Keep Going' or to 'Stop', which depends on various situations as depicted in the figure. If the signal is RED in colour, then the car should be stopped. If there is not enough gas (petrol) in the car, the car should be stopped at the next available gas station.



Building a Decision Tree

Training data set

CGPA	Communication	Aptitude	Programming Skill	Job offered?
High	Good	High	Good	Yes
Medium	Good	High	Good	Yes
Low	Bad	Low	Good	No
Low	Good	Low	Bad	No
High	Good	High	Bad	Yes
High	Good	High	Good	Yes
Medium	Bad	Low	Bad	No
Medium	Bad	Low	Good	No
High	Bad	High	Good	Yes
Medium	Good	High	Good	Yes
Low	Bad	High	Bad	No
Low	Bad	High	Bad	No
Medium	Good	High	Bad	Yes
Low	Good	Low	Good	No
High	Bad	Low	Bad	No
Medium	Bad	High	Good	No
High	Bad	Low	Bad	No
Medium	Good	High	Bad	Yes



Decision tree based on the training data

Algorithm for decision tree

Input: Training data set, test data set (or data points)

Steps:

Do for all attributes

Calculate the entropy E_i of the attribute F_i

if $E_i < E_{min}$

then $E_{min} = E_i$ and $F_{min} = F_i$

end if

End do

Split the data set into subsets using the attribute F_{min}

Draw a decision tree node containing the attribute F_{min} and split the data set into subsets

Repeat the above steps until the full tree is drawn covering all the attributes of the original table.

Implementations of Decision tree

There are many implementations of decision tree, the most prominent ones being C5.0, CART (Classification and Regression Tree), CHAID (Chi-square Automatic Interaction Detector) and ***ID3 (Iterative Dichotomiser 3) algorithms***.

A Decision tree is built based on two properties

1. Entropy
2. Information gain

Entropy

Entropy is an information theory metric that measures the impurity or uncertainty in a group of observations. It determines how a decision tree chooses to split data.

Entropy (S) measuring the impurity of S is defined as

$$\textbf{Entropy}(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

where c is the number of different class labels and p refers to the proportion of values falling into the i-th class label.

Example:

Let's have a dataset made up of three colors; red, purple, and yellow. If we have one red, three purple, and four yellow observations in our set, our equation becomes:

$$E = -(p_r log_2 p_r + p_p log_2 p_p + p_y log_2 p_y)$$

Where p_r, p_p and p_y are the probabilities of choosing a red, purple and yellow example respectively. We have $p_r = \frac{1}{8}$ because only $\frac{1}{8}$ of the dataset represents red. $\frac{3}{8}$ of the dataset is purple hence $p_p = \frac{3}{8}$. Finally, $p_y = \frac{4}{8}$ since half the dataset is yellow. As such, we can represent p_y as $p_y = \frac{1}{2}$. Our equation now becomes:

$$E = -(\frac{1}{8} log_2(\frac{1}{8}) + \frac{3}{8} log_2(\frac{3}{8}) + \frac{4}{8} log_2(\frac{4}{8}))$$

Our entropy would be: 1.41

- ✓ what happens when all observations belong to the same class? In such a case, the entropy will always be zero.

$$E = -(1 log_2 1)$$

$$= 0$$

- ✓ Such a dataset has no impurity. This implies that such a dataset would not be useful for learning.
- ✓ However, if we have a dataset with say, two classes, half made up of yellow and the other half being purple, the entropy will be one.

$$E = -((0.5 log_2 0.5) + (0.5 log_2 0.5))$$

$$= 1$$

- ✓ This kind of dataset is good for learning.

Information gain

- ✓ The information gain is created on the basis of the decrease in entropy (S) after a data set is split according to a particular attribute (A).

- ✓ Constructing a decision tree is all about finding an attribute that returns the highest information gain (i.e. the most homogeneous branches).

Information gain for a particular feature A is calculated by the difference in entropy before a split (or S_{bs}) with the entropy after the split (S_{as}).

$$\textbf{Information Gain (S, A) = Entropy } (S_{bs}) - \textbf{Entropy } (S_{as})$$

For calculating the entropy after split, entropy for all partitions needs to be considered. Then, the weighted summation of the entropy for each partition can be taken as the total entropy after split. For performing weighted summation, the proportion of examples falling into each partition is used as weight.

$$\textbf{Entropy}(S_{as}) = \sum_{i=1}^{n} w_i \text{ Entropy } (p_i)$$

Example

Suppose we have a dataset with two classes. This dataset has 5 purple and 5 yellow examples. The initial value of entropy will be given by the equation below. Since the dataset is balanced, we expect the answer to be 1.

$$E_{initial} = -((0.5log_20.5) + (0.5log_20.5))$$

$$= 1$$

Say we split the dataset into two branches. One branch ends up having four values while the other has six. The left branch has four purples while the right one has five yellows and one purple.

We mentioned that when all the observations belong to the same class, the entropy is zero since the dataset is pure. As such, the entropy of the left branch $E_{left} = 0$. On the other hand, the right branch has five yellows and one purple. Thus:

$$E_{right} = -(\frac{5}{6}log_2(\frac{5}{6}) + \frac{1}{6}log_2(\frac{1}{6}))$$

A perfect split would have five examples on each branch. This is clearly not a perfect split, but we can determine how good the split is. We know the entropy of each of the two branches. We weight the entropy of each branch by the number of elements each contains.

The entropy before the split, which we referred to as initial entropy $E_{initial} = 1$. After splitting, the current value is 0.39. We can now get our information gain, which is the entropy we "lost" after splitting.

$$Gain = 1 - 0.39$$

$$= 0.61$$

The more the entropy removed, the greater the information gain. The higher the information gain, the better the split.

This helps us calculate the quality of the split. The one on the left has 4, while the other has 6 out of a total of 10. Therefore, the weighting goes as shown below:

$$E_{split} = 0.6 * 0.65 + 0.4 * 0$$

$$= 0.39$$

Avoiding overfitting in decision tree – pruning

There are two approaches of pruning:

a) Pre-pruning: Stop growing the tree before it reaches perfection.

b) Post-pruning: Allow the tree to grow entirely and then post-prune some of the branches from it.

- ✓ In the case of pre-pruning, the tree is stopped from further growing once it reaches a certain number of decision nodes or decisions. Hence, in this strategy, the algorithm avoids over fitting as well as optimizes computational cost. However, it also stands a chance to ignore important information contributed by a feature which was skipped, thereby resulting in miss out of certain patterns in the data.
- ✓ In the case of post-pruning, the tree is allowed to grow to the full extent. Then, by using certain pruning criterion, e.g. error rates at the nodes, the size of the tree is reduced. This is a more effective approach in terms of classification accuracy as it considers all minute information available from the training data. However, the computational cost is obviously more than that of pre-pruning.

Strengths and weaknesses of Decision trees

Strengths of decision tree

- ✓ It produces very simple understandable rules. For smaller trees, not much mathematical and computational knowledge is required to understand this model.
- ✓ Works well for most of the problems.
- ✓ It can handle both numerical and categorical variables.
- ✓ Can work well both with small and large training data sets.
- ✓ Decision trees provide a definite clue of which features are more useful for classification.

Weaknesses of decision tree

- ✓ Decision tree models are often biased towards features having more number of possible values, i.e. levels.
- ✓ This model gets over fitted or under fitted quite easily.
- ✓ Decision trees are prone to errors in classification problems with many classes and relatively small number of training examples.
- ✓ A decision tree can be computationally expensive to train.
- ✓ Large trees are complex to understand.

Application of decision tree

Here are some applications of decision trees:

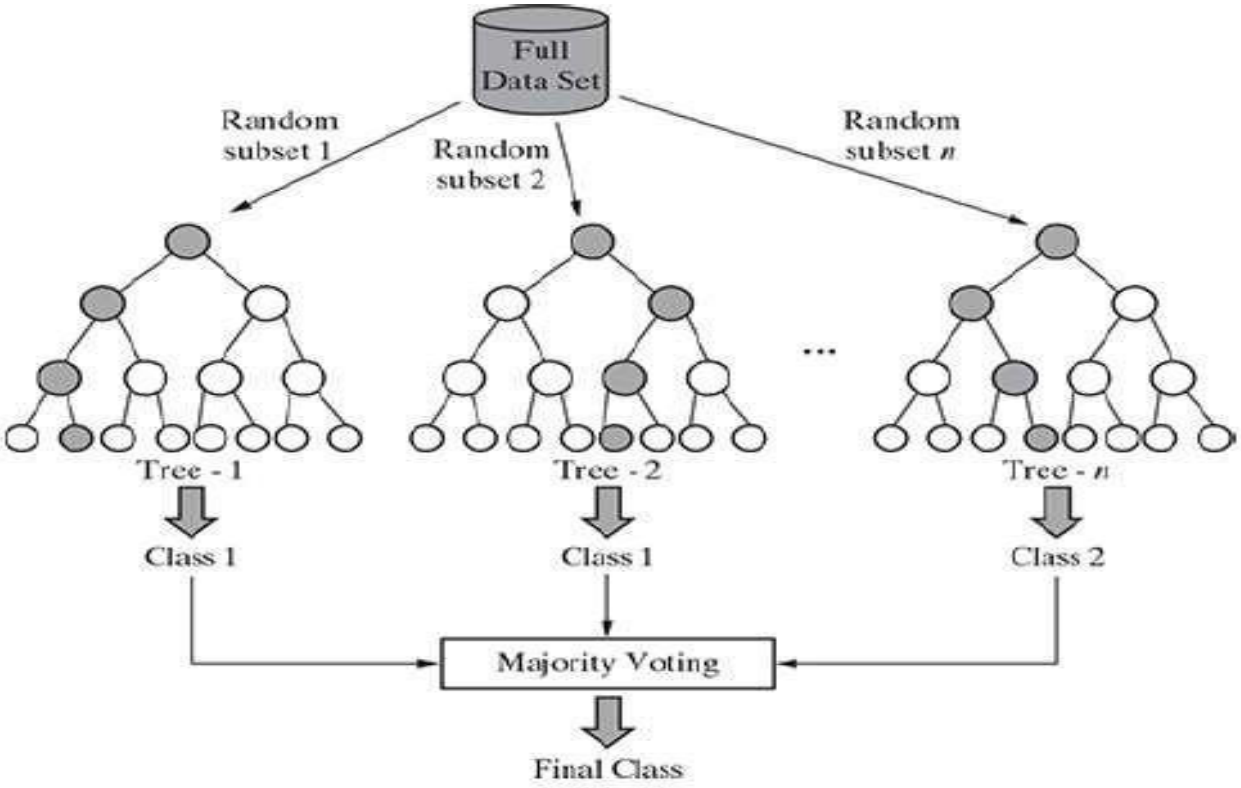
1. Marketing
2. Retention of Customers
3. Diagnosis of Diseases and Ailments
4. Detection of Frauds

1) Random forest model

- ✓ Random forest is an ensemble classifier, i.e. a combining classifier that uses and combines many decision tree classifiers.
- ✓ Ensembling is usually done using the concept of bagging with different feature sets. The reason for using large number of trees in random forest is to train the trees enough such that contribution from each feature comes in a number of

models.

- ✓ After the random forest is generated by combining the trees, majority vote is applied to combine the output of the different trees.



Random forest working Process

1. If there are N variables or features in the input data set, select a subset of 'm' (m < N) features at random out of the N features. Also, the observations or data instances should be picked randomly.
2. Use the best split principle on these 'm' features to calculate the number of nodes 'd'.
3. Keep splitting the nodes to child nodes till the tree is grown to the maximum possible extent.
4. Select a different subset of the training data 'with replacement' to train another decision tree following steps (1) to (3). Repeat this to build and train 'n' decision trees.
5. Final class assignment is done on the basis of the majority votes from the 'n' trees.

Out-of-bag (OOB) error in random forest

✓ In random forests, we have seen, that each tree is constructed using a different bootstrap sample from the original data. The samples left out of the bootstrap and not used in the construction of the i-th tree can be used to measure the performance of the model.

✓ At the end of the run, predictions for each such sample evaluated each time are tallied, and the final prediction for that sample is obtained by taking a vote.

✓ The total error rate of predictions for such samples is termed as out-of-bag (OOB) error rate.

✓ The error rate shown in the confusion matrix reflects the OOB error rate. Because of this reason, the error rate displayed is often surprisingly high.

Strengths and Weaknesses of Random forest

Strengths of random forest

1) It runs efficiently on large and expansive data sets.
2) It has a robust method for estimating missing data and maintains precision when a large proportion of the data is absent.
3) It has powerful techniques for balancing errors in a class population of unbalanced data sets.
4) It gives estimates (or assessments) about which features are the most important ones in the overall classification.
5) It generates an internal unbiased estimate (gauge) of the generalization error as the forest generation progresses.
6) Generated forests can be saved for future use on other data.
7) Lastly, the random forest algorithm can be used to solve both classification and regression problems.

Weaknesses of random forest

1) This model, because it combines a number of decision tree models, is not as easy to understand as a decision tree model.
2) It is computationally much more expensive than a simple model like decision tree.

Application of random forest

Random forest is a very powerful classifier which combines the versatility of many decision tree models into a single model. Because of the superior results, this ensemble model is gaining wide adoption and popularity amongst the machine learning practitioners to solve a wide range of classification problems.

UNIT – IV

Supervised Learning: Regression

Introduction, Example of Regression, Common Regression Algorithms-Simple linear regression, Multiple linear regression, Polynomial Regression Model, Logistic Regression,

---

Regression

- ✓ Regression focuses on solving problems such as predicting value of real estate, demand forecast in retail, weather forecast, etc.
- ✓ Regression is essentially finding a relationship (or) association between the dependent variable (Y) and the independent variable(s) (X), i.e. to find the function 'f ' for the association Y = f (X).

Example of Regression – Real estate price prediction

New City is the primary hub of the commercial activities in the country. In the last couple of decades, with increasing globalization, commercial activities have intensified in New City.
Together with that, a large number of people have come and settled in the city with a dream to achieve professional growth in their lives. As an obvious fall-out, a large number of housing projects have started in every nook and corner of the city. But the demand for apartments has still outgrown the supply.

To get benefit from this boom in real estate business, Karen has started a digital market agency for buying and selling real estates (including apartments, independent houses, town houses, etc.). Initially, when the business was small, she used to interact with buyers and sellers personally and help them arrive at a price quote — either for selling a property (for a seller) or for buying a property (for a buyer). Her long experience in real estate business helped her develop an intuition on what the correct price quote of a property could be — given the value of certain standard parameters such as area (sq. m.) of the property, location, floor, number of years since purchase, amenities available, etc.

However, with the huge surge in the business, she is facing a big challenge. She is not able to manage personal interactions as well as setting the correct price quote for the properties all alone. She hired an assistant for managing customer interactions. But the assistant, being new in the real estate business, is struggling with price quotations. How can Karen solve this problem?

Fortunately, Karen has a friend, Frank, who is a data scientist with in-depth knowledge in machine learning models. Frank comes up with a solution to Karen's problem. He builds a model which can predict the correct value of a real estate if it has certain standard inputs such as area (sq. m.) of the

property, location, floor, number of years since purchase, amenities available, etc. Wow, that sounds to be like Karen herself doing the job! Curious to know what model Frank has used? Yes, you guessed it right. He used a regression model to solve Karen's real estate price prediction problem.

COMMON REGRESSION ALGORITHMS

The most common regression algorithms are:

1) Simple linear regression
2) Multiple linear regression
3) Polynomial regression
4) Multivariate adaptive regression splines
5) Logistic regression
6) Maximum likelihood estimation (least squares)

Simple linear regression:-

As the name indicates, simple linear regression is the simplest regression model which involves only one predictor. This model assumes a linear relationship between the dependent variable and the predictor variable as shown in Figure below.

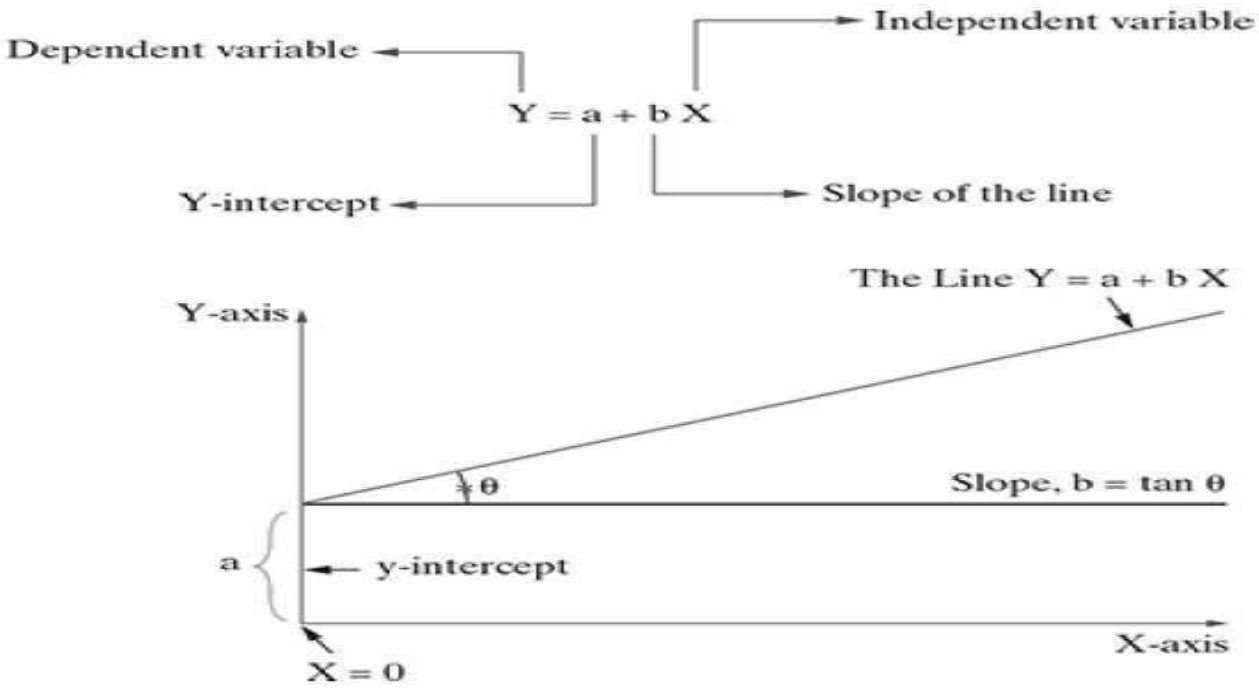
In the context of Karen's problem, if we take Price of a Property as the dependent variable and the Area of the Property (in sq. m.) as the predictor variable, we can build a model using simple linear regression.

$$\text{Price}_{\text{Property}} = f(\text{Area}_{\text{Property}})$$

Assuming a linear association, we can reformulate the model as –

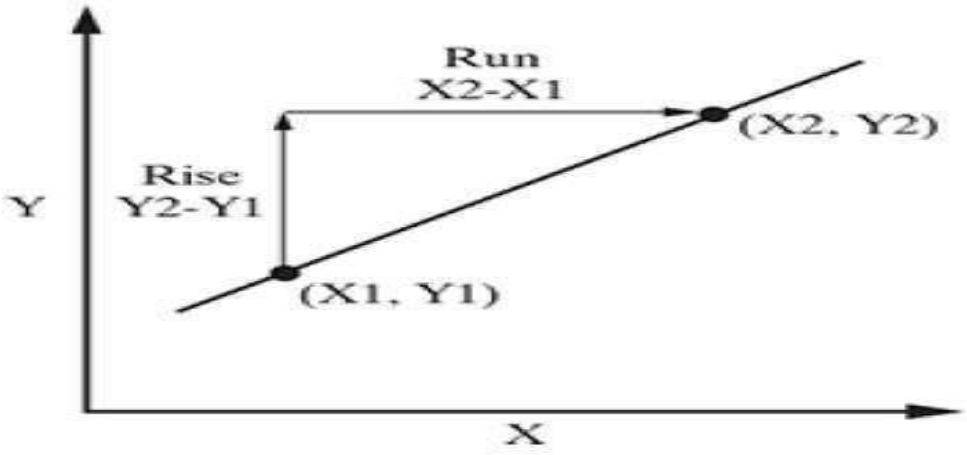
$$\text{Price}_{\text{Property}} = a + b.\,\text{Area}_{\text{Property}}$$

where 'a' and 'b' are intercept and slope of the straight line, respectively.



Slope of the simple linear regression model

Slope of a straight line represents how much the line in a graph changes in the vertical direction (Y-axis) over a change in the horizontal direction **(X-axis).**



Rise and run representation

Slope = Change in Y/Change in X

$$\text{Slope} = \frac{\text{Rise}}{\text{Run}} = \frac{Y2 - Y1}{X2 - X1}$$

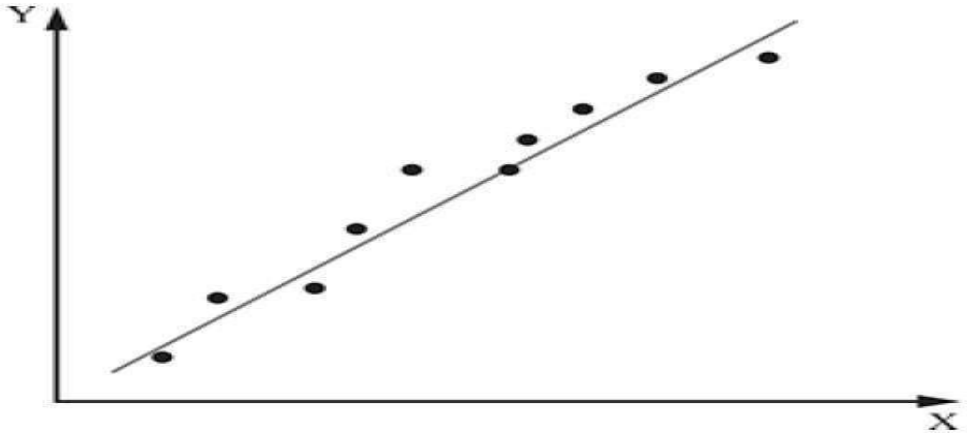
Types of slopes in a Linear Regression

There can be two types of slopes in a linear regression model: positive slope and negative slope. Different types of regression lines based on the type of slope include

1) Linear positive slope
2) Curve linear positive slope
3) Linear negative slope
4) Curve linear negative slope

1) Linear Positive Slope

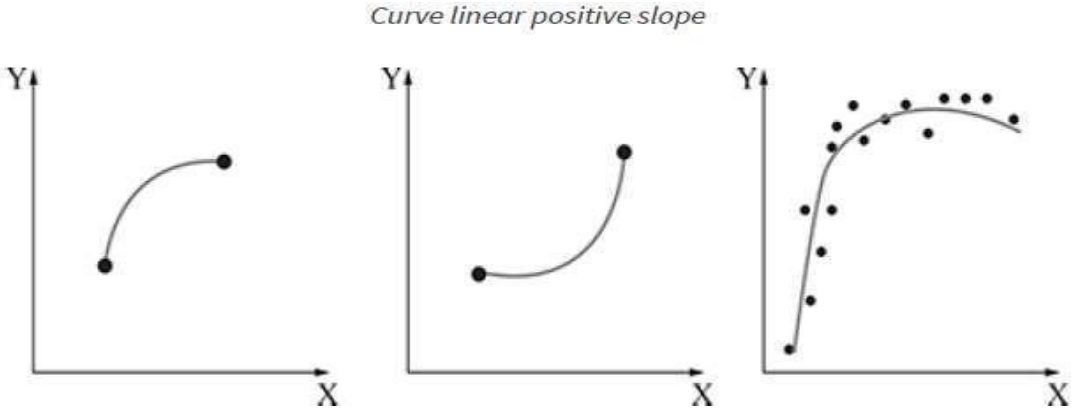
A positive slope always moves upward on a graph from left to right



Slope = Rise/Run = (Y2 − Y1) / (X2 − X1) = Delta (Y) / Delta(X)

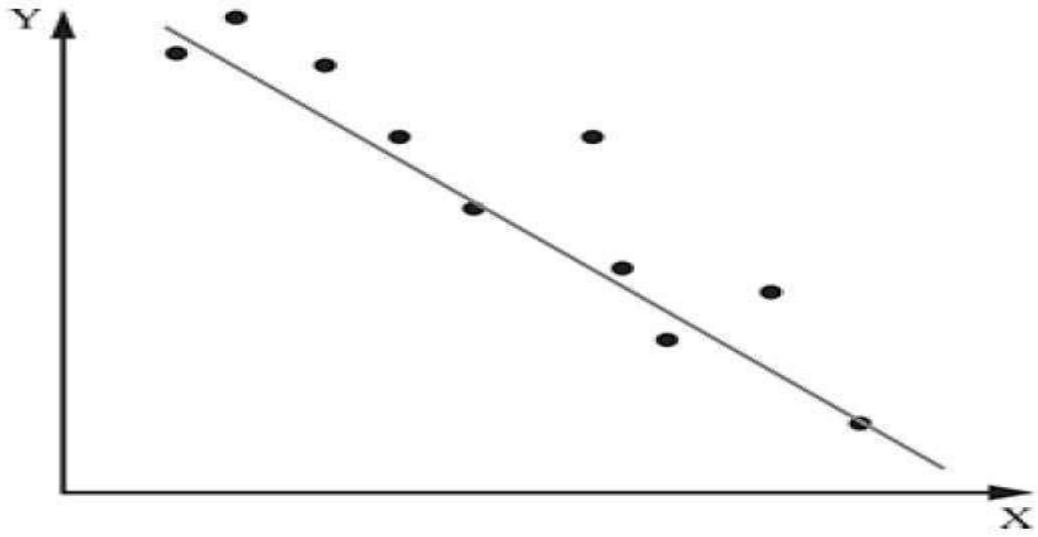
***2)* Curve Linear Positive Slope**

Curves in these graphs slope upward from left to right.



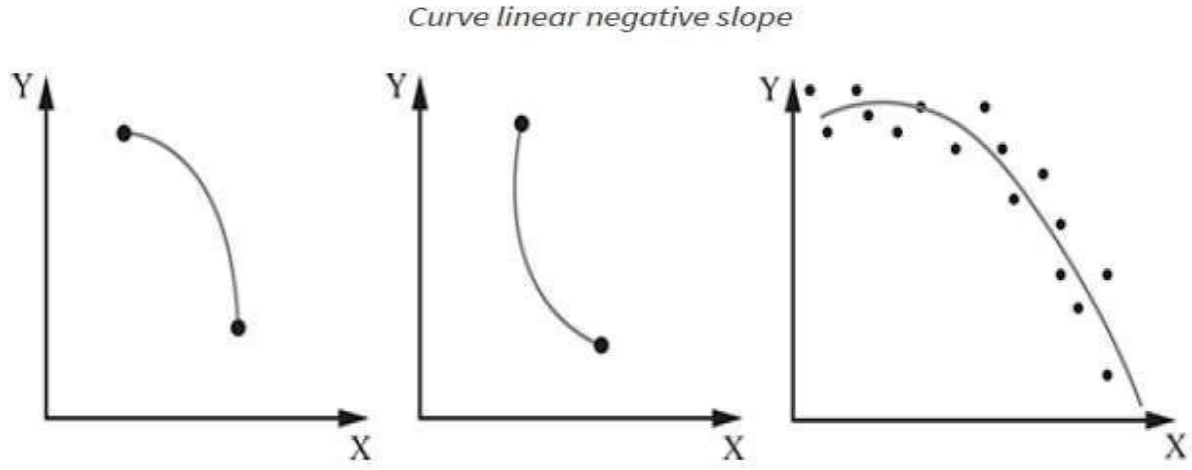
3) **Linear Negative Slope**

A negative slope always moves downward on a graph from left to right. As X value (on X-axis) increases, Y value decreases



4) **Curve Linear Negative Slope**

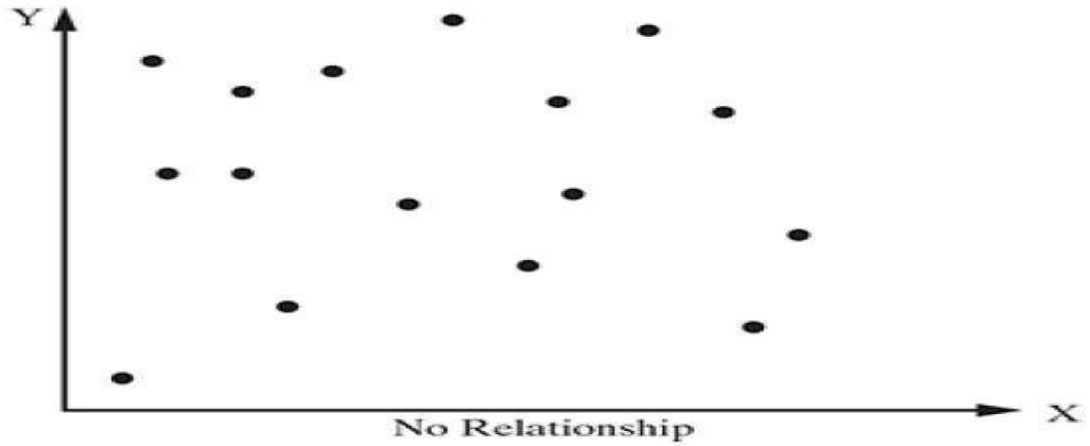
Curves in these graphs slope downward from left to right.



No relationship graph

Scatter graph shown in below Figure indicates 'no relationship' curve as it is very difficult to conclude

whether the relationship between X and Y is positive or negative.

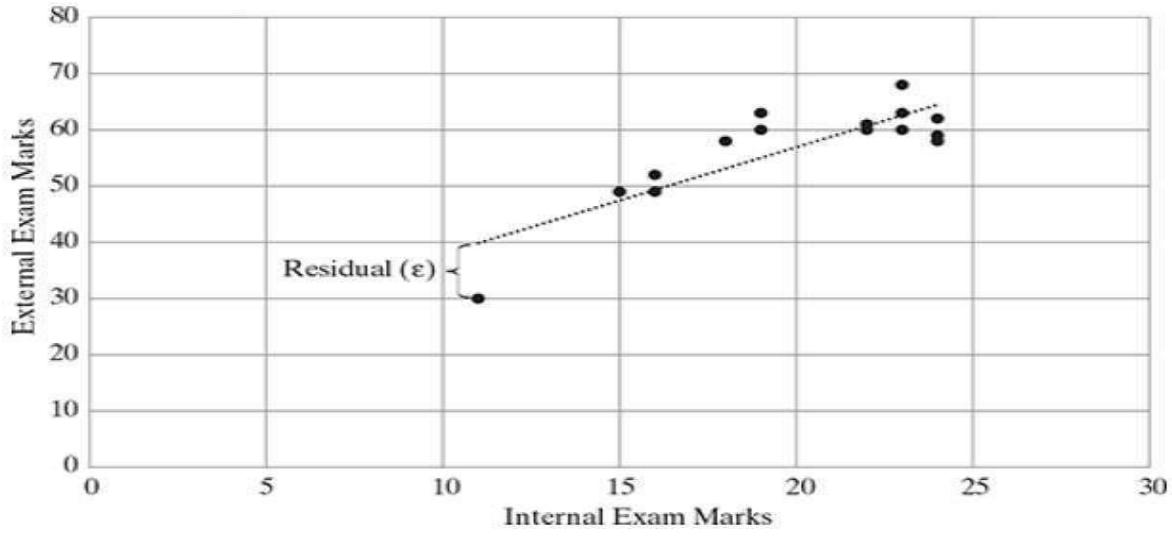


Error in simple regression (Marginal or Residual error)

- ✓ The regression equation model in machine learning uses the above slope–intercept format in algorithms. X and Y values are provided to the machine, and it identifies the values of a (intercept) and b (slope) by relating the values of X and Y.
- ✓ However, identifying the exact match of values for a and b is not always possible. There will be some error value (ε) associated with it. This error is called marginal or residual error.
- ✓ Residual is the distance between the predicted point (on the regression line) and the actual point

$$Y = (a + bX) + \varepsilon$$

Example for Residual error



Ordinary Linear Squares (OLS) Algorithm

Ordinary Least Squares (OLS) is the technique used to

- build a simple linear regression model for a given problem
- estimate a line that will minimize the error (ε)

In Y = a + bX, b value can be calculated with the below formula so that Sum of the Squares of the Errors is least.

$$b = \frac{\Sigma_i(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma_i(X_i - \overline{X})^2} = \frac{\text{cov}(X, Y)}{\text{Var}(X)}$$

The corresponding value of 'a' calculated using the above value of 'b' is

$$a = \overline{Y} - b\overline{X}$$

Where $\overline{Y}$ is the mean of Y and $\overline{X}$ is the mean of X,

Steps in OLS algorithm

- Step 1: Calculate the mean of X and Y
- Step 2: Calculate the errors of X and Y
- Step 3: Get the product
- Step 4: Get the summation of the products
- Step 5: Square the difference of X
- Step 6: Get the sum of the squared difference
- Step 7: Divide output of step 4 by output of step 6 to calculate 'b'
- Step 8: Calculate 'a' using the value of 'b'

Example

A college professor believes that if the grade for internal examination is high in a class, the grade for external examination will also be high. A random sample of 15 students in that class was selected, and the data is given below:

Internal Exam	15	23	18	23	24	22	22	19	19	16	24	11	24	16	23
External Exam	49	63	58	60	58	61	60	63	60	52	62	30	59	49	68

Step 2 Step 3 Step 5

X	Y	X- mean (X)	Y- Mean (Y)	$(Xi-\overline{X})(Yi-\overline{Y})$	$(Xi-\overline{X})^2$
15	49	-4.93	-7.8	38.454	24.3049
23	63	3.07	6.2	19.034	9.4249
18	58	-1.93	1.2	-2.316	3.7249
23	60	3.07	3.2	9.824	9.4249
24	58	4.07	1.2	4.884	16.5649
22	61	2.07	4.2	8.694	4.2849
22	60	2.07	3.2	6.624	4.2849
19	63	-0.93	6.2	-5.766	0.8649
19	60	-0.93	3.2	-2.976	0.8649
16	52	-3.93	-4.8	18.864	15.4449
24	62	4.07	5.2	21.164	16.5649
11	30	-8.93	-26.8	239.324	79.7449
24	59	4.07	2.2	8.954	16.5649
16	49	-3.93	-7.8	30.654	15.4449
23	68	3.07	11.2	34.384	9.4249
19.9	**56.8**		$\Sigma(Xi-\overline{X})(Yi-\overline{Y})$	**429.8**	**226.9335**

Step 1 Step 4 Step 6

The Calculation summary is

Step 7: Divide (step4 / step6)

b = 429.28 / 226.93 = 1.89

Step 8: Calculate a using the value of b

$a = \overline{Y} - b\overline{X}$

$a = 56.8 - 1.89 \times 19.9$

a = 19.05

Sum of $X = 299$

Sum of $Y = 852$

Mean X, $M_X = 19.93$

Mean Y, $M_Y = 56.8$

Sum of squares (SS_X) = 226.9333

Sum of products (SP) = 429.8

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

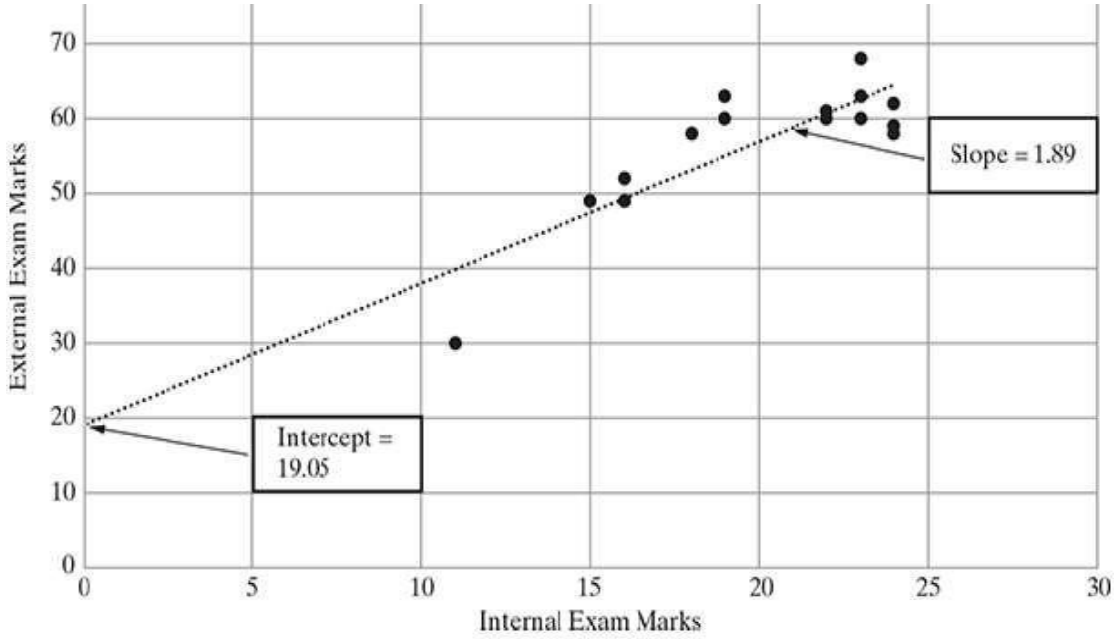
$$b = \frac{SP}{SS_X} = \frac{429.8}{226.93} = 1.89395$$

$$a = M_Y - bM_X = 56.8 - (1.89 \times 19.93) = 19.0473$$

$$\hat{y} = 1.89395X + 19.0473$$

Linear Regression model is

$$M_{Ext} = 19.04 + 1.89 \times M_{Int}$$



Multiple Linear Regression

In a multiple regression model, two or more independent variables, i.e. predictors are involved in the model.

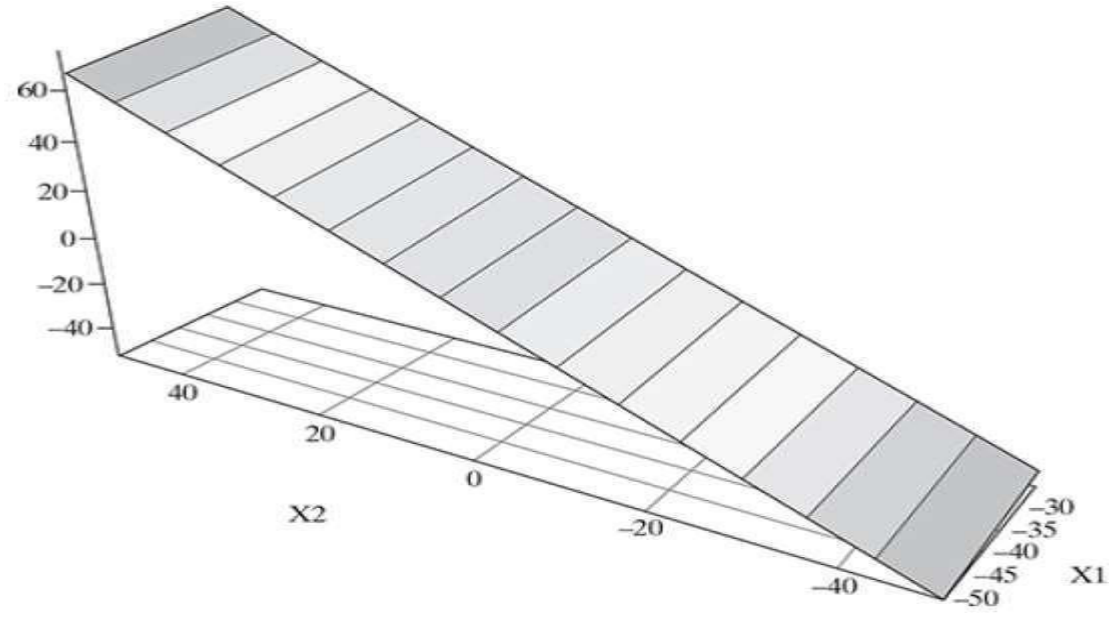
For Example, we consider Price of a Property (in $) as the dependent variable and Area of the Property (in sq. m.), location, floor, number of years since purchase and amenities available as the independent variables, we can form a multiple regression equation as shown below:

$$\text{Price}_{\text{Property}} = f(\text{Area}_{\text{Property}}, \text{location, floor, Ageing, Amenities})$$

The following expression describes the equation involving the relationship with two predictor variables, namely X_1 and X_2

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

- ✓ The model describes a plane in the three-dimensional space of $\hat{Y}$, X_1, and X_2. Parameter 'a' is the intercept of this plane. Parameters 'b1' and 'b2' are referred to as partial regression coefficients.
- ✓ Parameter b_1 represents the change in the mean response corresponding to a unit change in X_1 when X_2 is held constant.
- ✓ Parameter b_2 represents the change in the mean response corresponding to a unit change in X_2 when X_1 is held constant.



$$\hat{Y} = 22 + 0.3X_1 + 1.2X_2$$

Above figure shows the sample graph for

Assumptions in Regression Analysis

1. The dependent variable (Y) can be calculated / predicated as a linear function of a specific set of independent variables (X's) plus an error term (ε).
2. The number of observations (n) is greater than the number of parameters (k) to be estimated,

i.e. n > k.

3. Relationships determined by regression are only relationships of association based on the data set and not necessarily of cause and effect of the defined class.
4. Regression line can be valid only over a limited range of data. If the line is extended (outside the range of extrapolation), it may only lead to wrong predictions.
5. If the business conditions change and the business assumptions underlying the regression model are no longer valid, then the past data set will no longer be able to predict future trends.
6. Variance is the same for all values of X (**homoskedasticity**).

7. The error term (ε) is normally distributed. This also means that the mean of the error (ε) has an expected value of 0.
8. The values of the error (ε) are independent and are not related to any values of X. This means that there are no relationships between a particular X, Y that are related to another specific value of X, Y.

Main Problems in Regression Analysis

In multiple regressions, there are two primary problems: **multicollinearity** and heteroskedasticity.

1) **Multicollinearity**

- ✓ Two variables are perfectly collinear if there is an exact linear relationship between them.
- ✓ Multi collinearity is the situation in which the degree of correlation is not only between the dependent variable and the independent variable, but there is also a strong correlation within (among) the independent variables themselves.
- ✓ A multiple regression equation can make good predictions when there is multicollinearity.

2) **Heteroskedasticity**

Heteroskedasticity refers to the changing variance of the error term. If the variance of the error term is not constant across data sets, there will be erroneous predictions. In general, for a regression equation to make accurate predictions, the error term should be independent, identically (normally) distributed (iid).

Mathematically, this assumption is written as

$$\text{var}(u_i) = \sigma^2 \quad \text{and}$$
$$\text{cov}(u_i u_j) = 0 \quad \text{for } i \neq j.$$

- ✓ '*u*' represents the error terms
- ✓ 'var' represents the variance
- ✓ 'cov' represents the covariance

Improving the accuracy of Linear Regression Model

Bias and Variance

Accuracy refers to how close the estimation is near the actual value, whereas prediction refers to continuous estimation of the value.

High bias = low accuracy (not close to real value)

High variance = low prediction (values are scattered)

Low bias = high accuracy (close to real value)

Low variance = high prediction (values are close to each other)

- ✓ Let us say we have a regression model which is highly accurate and highly predictive; therefore, the overall error of our model will be low, implying a low bias (high accuracy) and low variance (high prediction). This is highly preferable.
- ✓ Similarly, we can say that if the variance increases (low prediction), the spread of our data points increases, which results in less accurate prediction.
- ✓ As the bias increases (low accuracy), the error between our predicted value and the observed values increases.
- ✓ Therefore, balancing out bias and accuracy is essential in a regression model.

Methods to improve accuracy of Linear Regression Model

Accuracy of linear regression can be improved using the following three methods:

1) Shrinkage Approach
2) Subset Selection
3) Dimensionality (Variable) Reduction

1) Shrinkage (Regularization) approach

By limiting (shrinking) the estimated coefficients, we can try to reduce the variance at the cost of a negligible increase in bias. This can in turn lead to substantial improvements in the accuracy of the

model.

The two best-known techniques for shrinking the regression coefficients towards zero are

a) ridge regression

b) lasso (Least Absolute Shrinkage Selector Operator)

a) Ridge regression

- ✓ Ridge regression performs L2 regularization, i.e. it adds penalty equivalent to square of the magnitude of coefficients

Minimization objective of ridge = LS Obj + α × (sum of square of coefficients)

- ✓ Ridge regression (include all k predictors in the final model) is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity

b) Lasso regression

Lasso regression performs L1 regularization, i.e. it adds penalty equivalent to the absolute value of the magnitude of coefficients.

Minimization objective of ridge = LS Obj + α × (absolute value of the magnitude of coefficients)

- ✓ The lasso yields sparse models (involving only subset) that are simpler as well as more interpretable.
- ✓ The lasso can be expected to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or equal to zero.

2) Subset selection

Identify a subset of the predictors that is assumed to be related to the response and then fit a model using OLS on the selected reduced subset of variables. There are two methods in which subset of the regression can be selected:

a) Best subset selection (considers all the possible (2^k))

b) Stepwise subset selection

   i. Forward stepwise selection (0 to k)

   ii. Backward stepwise selection (k to 0)

a) Best subset selection (considers all the possible (2k))

In best subset selection, we fit a separate least squares regression for each possible subset of the k predictors. For computational reasons, best subset selection cannot be applied with very large value of

predictors (k). The best subset selection procedure considers all the possible (2^k) models containing subsets of the p predictors.

b) Stepwise subset selection

i. **Forward stepwise selection (0 to k)**

- ✓ Forward stepwise selection begins with a model containing no predictors, and then, predictors are added one by one to the model, until all the k predictors are included in the model.
- ✓ In particular, at each step, the variable (X) that gives the highest additional improvement to the fit is added.

ii. Backward stepwise selection (k to 0)

Backward stepwise selection begins with the least squares model which contains all k predictors and then iteratively removes the least useful predictor one by one.

3) Dimensionality reduction (Variable reduction)

- ✓ The earlier methods, namely subset selection and shrinkage, control variance either by using a subset of the original variables or by shrinking their coefficients towards zero. In dimensionality reduction, predictors (X) are transformed, and the model is set up using the transformed variables after dimensionality reduction.
- ✓ The number of variables is reduced using the dimensionality reduction method. Principal component analysis is one of the most important dimensionality (variable) reduction techniques.

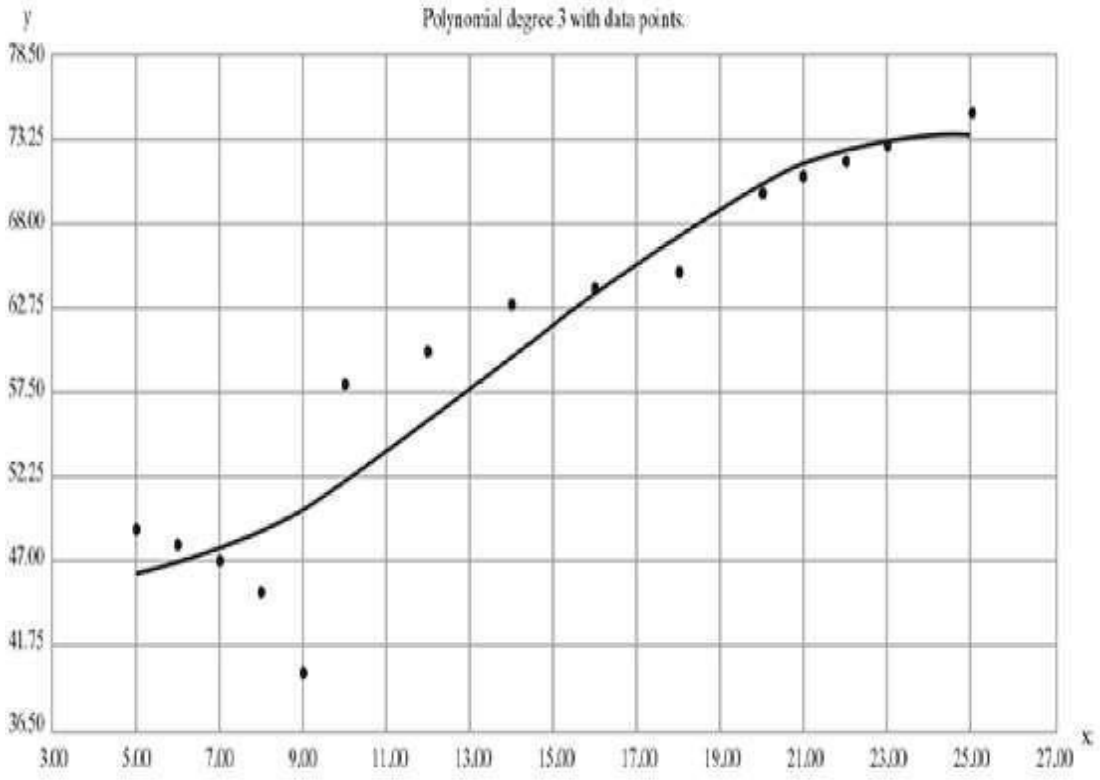
Polynomial Regression Model

- ✓ Polynomial regression model is the extension of the simple linear model by adding extra predictors obtained by raising (squaring) each of the original predictors to a power.
- ✓ For example, if there are three variables, X, X^2, and X^3 are used as predictors. This approach provides a simple way to yield a non-linear fit to data.

$$f(x) = c_0 + c_1 X^1 + c_2 X^2 + c_3 X^3$$

In the above equation, c_0 , c_1, c_2 , and c_3 are the coefficients.

Example: Let us use the below data set of (*X*, *Y*) for degree 3 polynomial.



Polynomial regression degree 3

Logistic Regression

- ✓ Logistic regression is both classification and regression technique depending on the scenario used.
- ✓ Logistic regression (logit regression) is a type of regression analysis used for predicting the outcome of a categorical dependent variable similar to OLS regression.
- ✓ In logistic regression, dependent variable (Y) is binary (0,1) and independent variables (X) are continuous in nature.
- ✓ The goal of logistic regression is to predict the likelihood that Y is equal to 1 (probability that Y = 1 rather than 0) given certain values of X. That is, if X and Y have a strong positive linear relationship, the probability that a person will have a score of Y = 1 will increase as values of X increase.
- ✓ So, we are predicting probabilities rather than the scores of the dependent variable.

Formula for Logistic Regression

- ✓ An explanation of logistic regression begins with an explanation of the logistic function, which always takes values between zero and one. The logistic formulae are stated in terms of the

probability that Y = 1, which is referred to as P. The probability that Y is 0 is 1 − P.

$$\ln\left(\frac{P}{1-P}\right) = a + bX$$

$$\ln(p/1-p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

- ✓ Probability (P) can also be computed from the regression equation. So, if we know the regression equation, we could, theoretically, calculate the expected probability that Y = 1 for a given value of X.

$$P = \frac{\exp(a + bX)}{1 + \exp(a + bx)} = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

'exp' is the exponent function, which is sometimes also written as e.

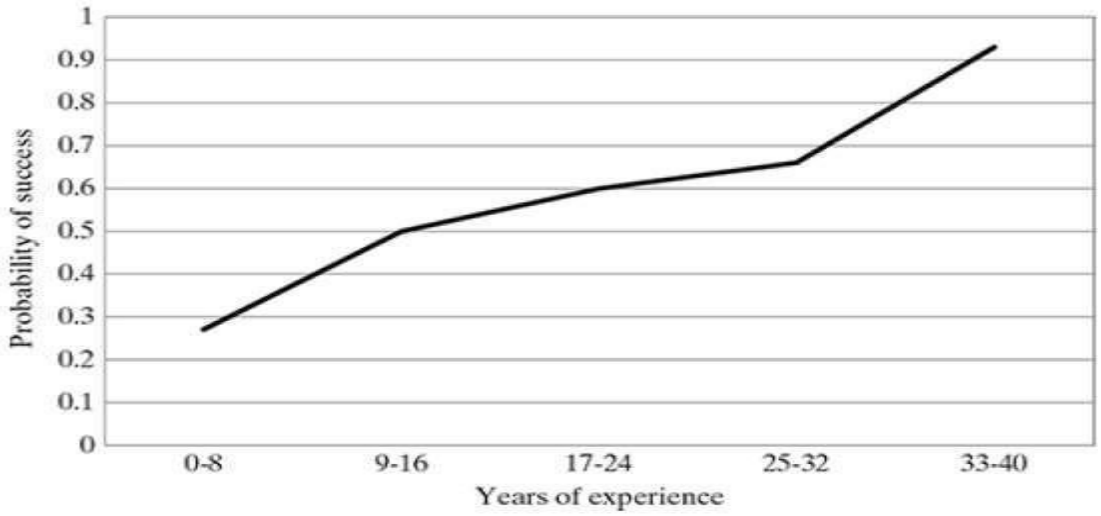
Example

We might try to predict whether or not a small project will succeed or fail on the basis of the number of years of experience of the project manager handling the project. We presume that those project managers who have been managing projects for many years will be more likely to succeed. This means that as X (the number of years of experience of project manager) increases, the probability that Y will be equal to 1 (success of the new project) will tend to increase.

To illustrate this, it is convenient to segregate years of experience into categories (i.e. 0–8, 9– 16, 17–24, 25–32, 33–40). If we compute the mean score on Y (averaging the 0s and 1s) for each category of years of experience, we will get something like

X	Y
0–8	0.27
9–16	0.5
17–24	0.6
25–32	0.66
33–40	0.93

When the graph is drawn for the above values of X and Y, it appears like the graph in below Figure



Logistic regression

Assumptions in logistic regression

- ✓ The following assumptions must hold when building a logistic regression model:
- ✓ There exists a linear relationship between logit function and independent variables
- ✓ The dependent variable Y must be categorical (1/0) and take binary value, e.g. if pass then Y = 1; else Y = 0
- ✓ The data meets the 'iid' criterion, i.e. the error terms, ε, are independent from one another and identically distributed
- ✓ The error term follows a binomial distribution [n, p]
    - o n = # of records in the data
    - o p = probability of success (pass, responder)

Maximum Likelihood Estimation

The coefficients in a logistic regression are estimated using a process called Maximum Likelihood Estimation (MLE).

what is likelihood function

A fair coin outcome flips equally heads and tails of the same number of times. If we toss the coin 10 times, it is expected that we get five times Head and five times Tail.

Let us now discuss about the probability of getting only Head as an outcome; it is 5/10 = 0.5 in the above case. Whenever this number (P) is greater than 0.5, it is said to be in favour of Head. Whenever P

is lesser than 0.5, it is said to be against the outcome of getting Head.

Let us represent 'n' flips of coin as $X1$, $X2$, $X3$,…, Xn . Now Xi can take the value of 1 or 0.

$Xi = 1$ if Head is the outcome $Xi = 0$ if Tail is the outcome

When we use the Bernoulli distribution represents each flip of the coin:

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}$$

Each observation X is independent and also identically distributed (iid), and the joint distribution simplifies to a product of distributions.

$$f(x_1, ..., x_n|\theta)\prod_{i=1}^{n} f(x_i|\theta) = \theta^{x_1}(1-\theta)^{1-x_i} ... \theta^{x_n}(1-\theta)^{1-x_n} = \theta^{\#H}(1-\theta)^{n-\#H},$$

where $\#H$ is the number of flips that resulted in the expected outcome (heads in this case).

The likelihood equation is

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

MLE is about predicting the value for the parameters that maximizes the likelihood function.

$$\log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

----------END----------

UNIT – V

Unsupervised Learning : Introduction, Clustering: Clustering as a machine learning task, Different types of clustering techniques, Partitioning methods, Hierarchical clustering, Finding Pattern using Association Rule: Definition of common terms, Association rule, The Apriority Algorithm for association rule learning.

Unsupervised Learning

Unsupervised learning is a machine learning concept where the unlabeled and unclassified information is analysed to discover hidden knowledge. The algorithms work on the data without any prior training, but they are constructed in such a way that they can identify patterns, groupings, sorting order, and numerous other interesting knowledge from the set of data.

Unsupervised VS Supervised Learning

Category	Supervised Learning	Unsupervised Learning
Data	Labeled data is supplied	Unlabeled data is supplied
Training	Training will happen	No training will happen
Output	Try to learn the probability of outcome Y for particular input X and Predict the output	Find out the association between the features or their grouping to understand the nature of the data
Types of Algorithms	Classification and Regression	Clustering and Association Analysis

Applications of Unsupervised Learning

Because of its flexibility that it can work on uncategorized and unlabeled data, there are many domains where unsupervised learning finds its application. Few examples of such applications are as follows:

a. Segmentation of target consumer populations by an advertisement consulting agency on the basis of few dimensions such as demography, financial data, purchasing habits, etc. so that the advertisers can reach their target consumers efficiently.

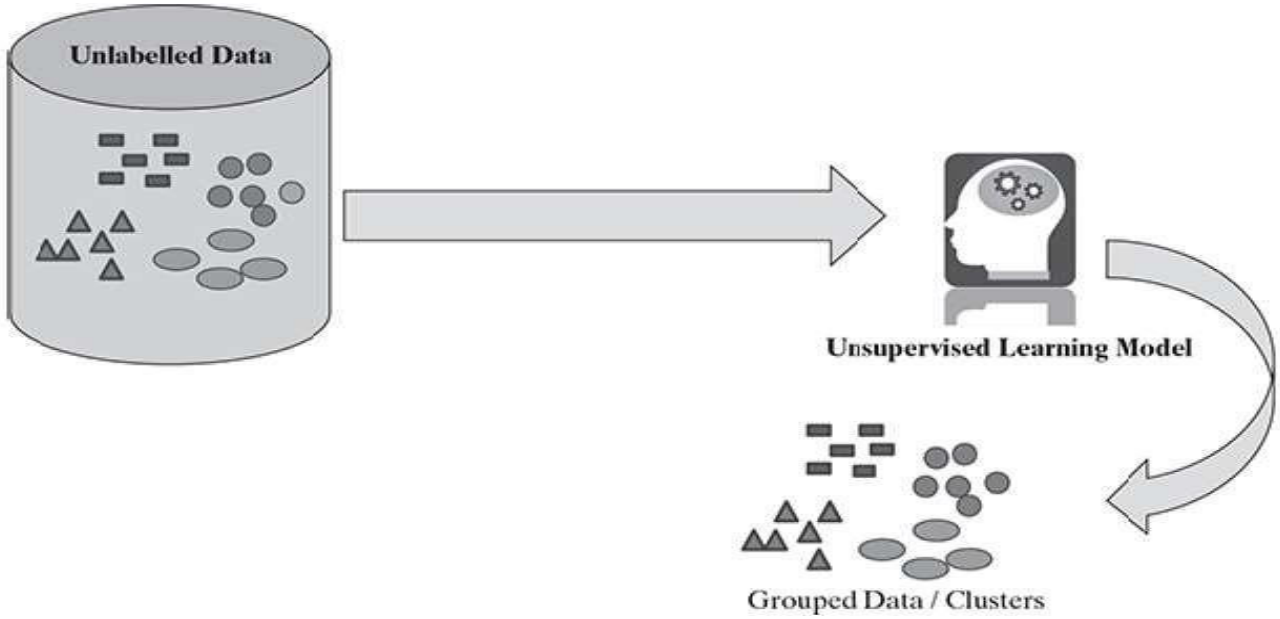
b. Anomaly or fraud detection in the banking sector by identifying the pattern of loan defaulters

c. Image processing and image segmentation such as face recognition, expression identification.

d. Grouping of important characteristics in genes to identify important influencers in new areas of genetics
e. Utilization by data scientists to reduce the dimensionalities in sample data to simplify modeling Document clustering and identifying potential labeling options

Chat bots, self-driven cars, and many more recent innovations are results of the combination of unsupervised and supervised learning.

CLUSTERING:



Clustering refers to a broad set of techniques for finding subgroups, or clusters, in a data set on the basis of the characteristics of the objects within that data set in such a manner that the objects within the group are similar (or related to each other) but are different from (or unrelated to) the objects from the other groups.

Clustering is defined as an unsupervised machine learning task that automatically divides the data into **clusters** or groups of similar items.

The effectiveness of clustering depends on how similar or related the objects within a group are or how different or unrelated the objects in different groups are from each other.

Uses of Cluster Analysis: There are many different fields where cluster analysis is used effectively, such as —

a. **Text data mining:** this includes tasks such as text categorization, text clustering, document summarization, concept extraction, sentiment analysis, and entity relation modeling
b. **Customer segmentation:** creating clusters of customers on the basis of parameters such as demographics, financial conditions, buying habits, etc., which can be used by retailers and advertisers to promote their products in the correct segment

c. **Anomaly checking:** checking of anomalous behaviors such as fraudulent bank transaction, unauthorized computer intrusion, suspicious movements on a radar scanner, etc.

d. **Data mining:** simplify the data mining task by grouping a large number of features from an extremely large data set to make the analysis manageable

Different types of clustering techniques

The major clustering techniques are

1) Partitioning methods,

2) Hierarchical methods, and

3) Density-based methods.

1) Partitioning methods

Two of the most important algorithms for partitioning based clustering are

k-means and **k-medoid.**

(a) K-means - A centroid-based technique

✓ This is one of the oldest and most popularly used algorithm for clustering.

✓ The principle of the k-means algorithm is to assign each of the 'n' data points to one of the K clusters where 'K' is a user-defined parameter as the number of clusters desired.

✓ The objective is to maximize the homogeneity within the clusters and also to maximize the differences between the clusters.

Simple algorithm of K-means:-

Step 1: Select *K* points in the data space and mark them as initial centroids

loop

Step 2: Assign each point in the data space to the nearest centroid to form K clusters

Step 3: Measure the distance of each point in the cluster from the centroid

Step 4: Calculate the Sum of Squared Error (SSE) to measure the quality of the clusters *(described later in this chapter)*

Step 5: Identify the new centroid of each cluster on the basis of distance between points

Step 6: Repeat Steps 2 to 5 to refine until centroids do not change

end loop

Example :-

In the above figure, let's assume K=4 implying that we want to create four clusters out of this data set.

Step1: we assign four random points from the data set as the centroids, as represented by the * signs, and we assign the data points to the nearest centroid to create four clusters.

Step2: on the basis of the distance of the points from the corresponding centroids, the centroids are updated and points are reassigned to the updated centroids.

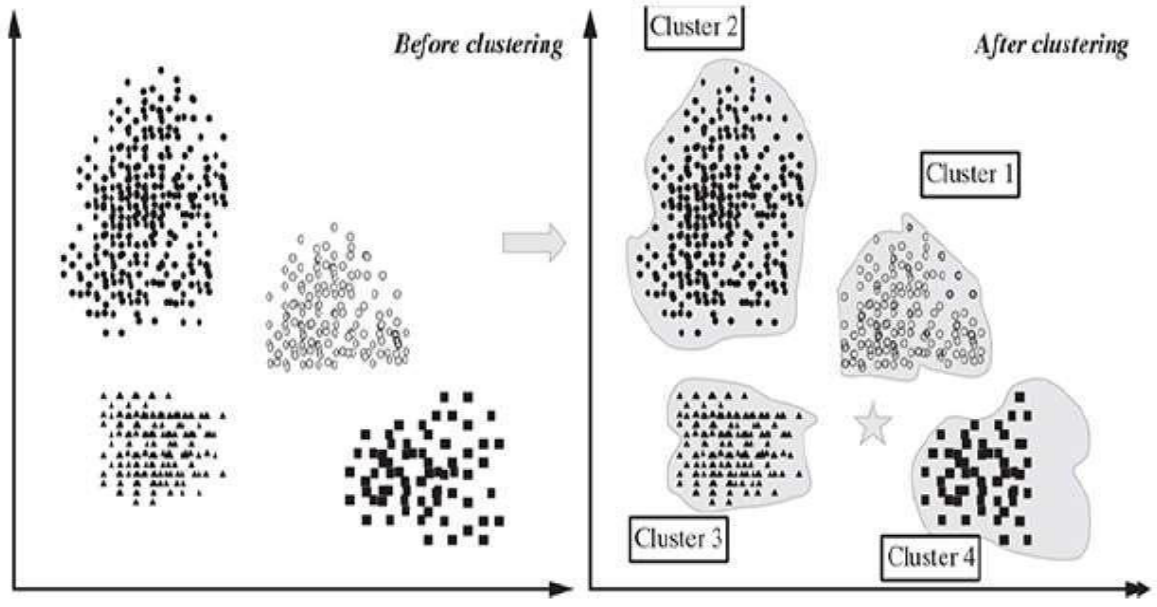
Step3: The iterative step is to recalculate the centroids of the data set after each iteration. The proximities of the data points from each other within a cluster is measured to minimize the distances. The measure of quality of clustering uses the SSE technique.

$$\text{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} \text{dist}\,(c_i, x)^2$$

Where, **dist()** calculates the ***Euclidean distance*** between the centroid c of the cluster C and the data points x in the cluster.

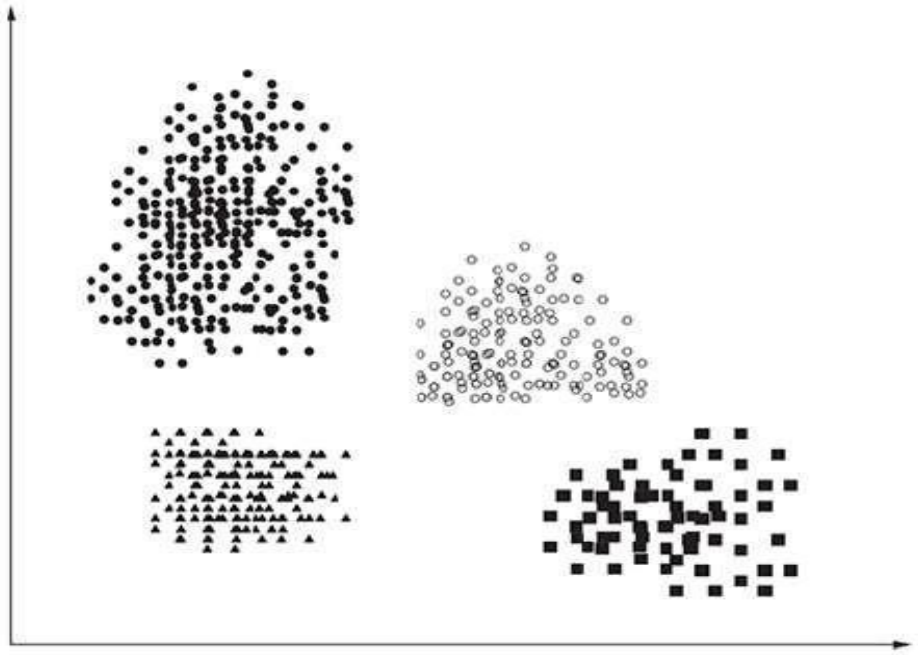
$$\text{dist}(x, y) = \sqrt{\sum_{1}^{n} (x_i - y_i)^2}$$

The lower the SSE for a clustering solution, the better is the representative position of the centroid.

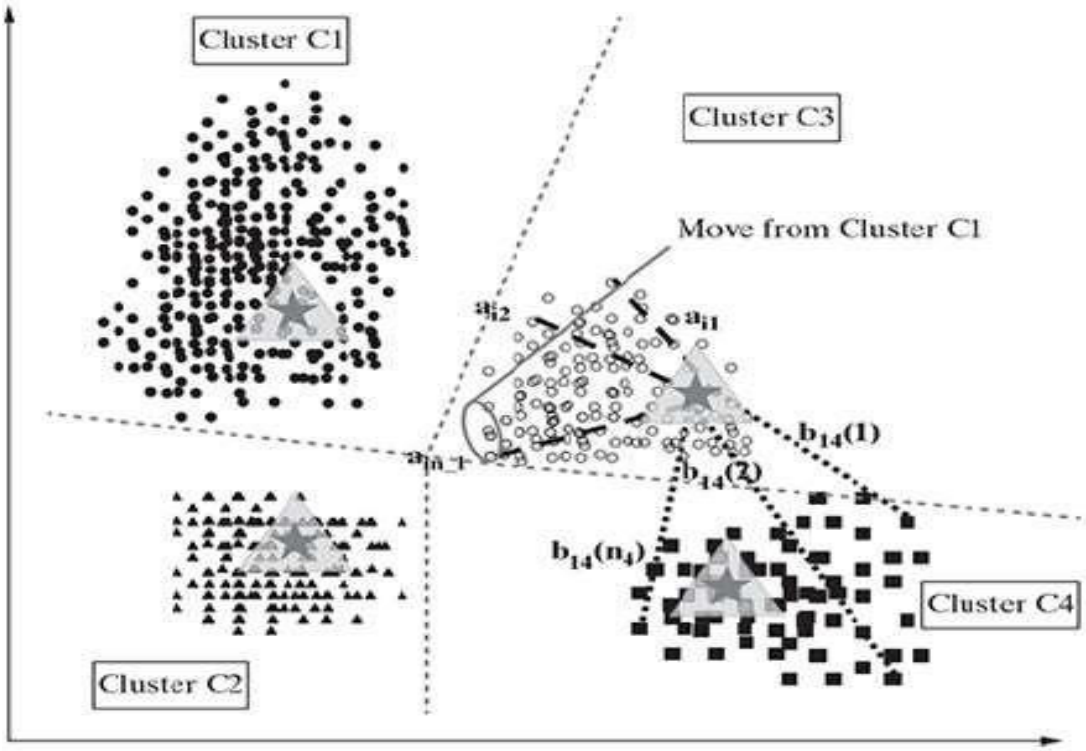


Clustering concept – before and after clustering

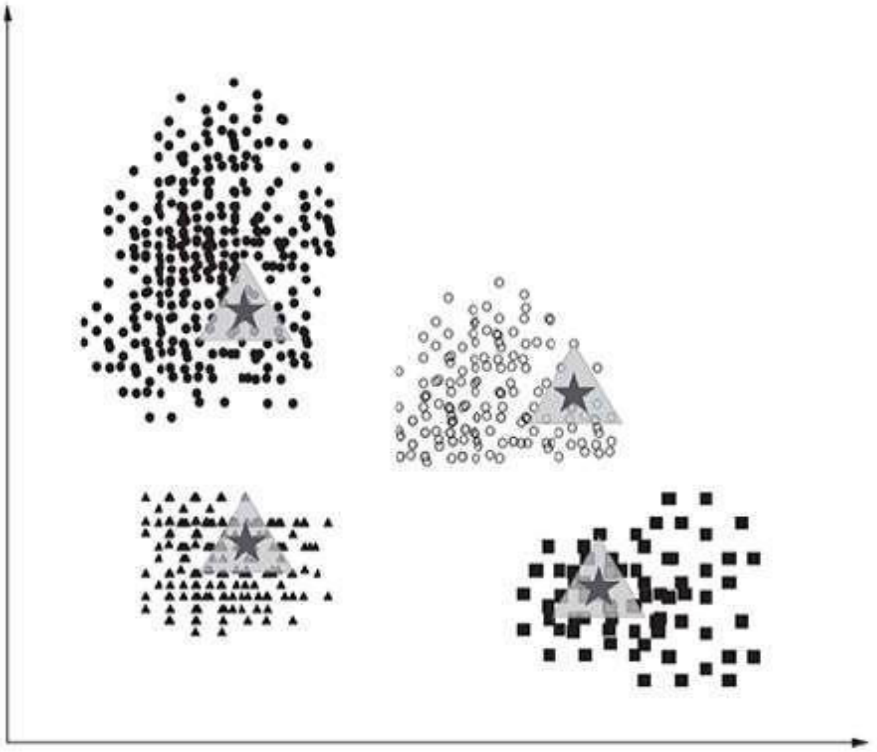
Limitation of SSE: One limitation of the squared error method is that in the case of presence of outliers in the data set, the squared error can distort the mean value of the clusters.



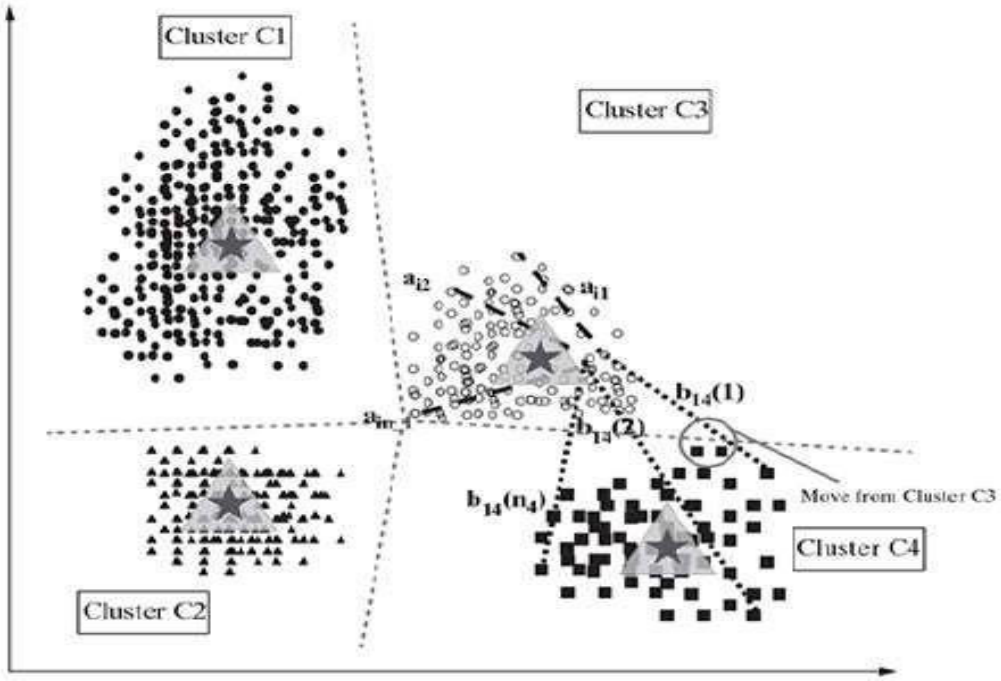
Clustering of data set



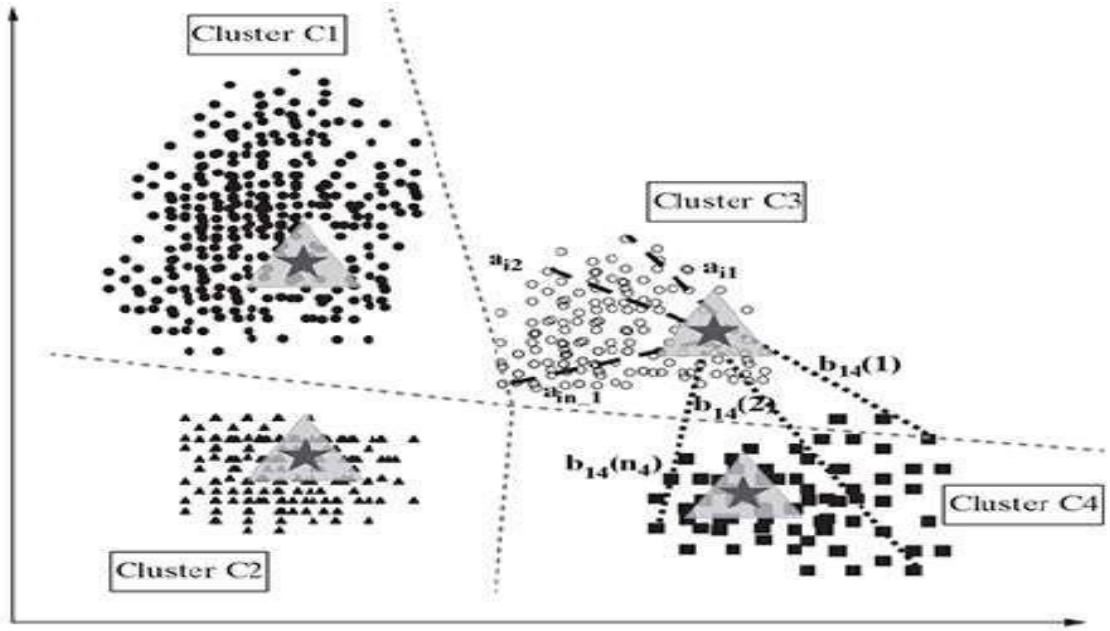
Iteration 2: Centroids recomputed and points redistributed among the clusters according to the nearest centroid



Clustering with initial centroids



Iteration 3: Final cluster arrangement: Centroids recomputed and points redistributed among the clusters according to the nearest centroid



Iteration 1: Four clusters and distance of points from the centroids

K-means : Strengths and Weaknesses

Strengths	Weaknesses
• The principle used for identifying the clusters is very simple and involves very less complexity of statistical terms • The algorithm is very flexible and thus can be adjusted for most scenarios and complexities • The performance and efficiency are very high and comparable to those of any sophisticated algorithm in term of dividing the data into useful clusters	• The algorithm involves an element of random chance and thus may not find the optimal set of cluster in some cases • The starting point of guessing the number natural clusters within the data requires some experience of the user, so that the final outcome is efficient

(b) K-Medoids: a representative object-based technique

The k-means algorithm is sensitive to outliers in the data set and inadvertently produces skewed clusters when the means of the data points are used as centroids.

Ex:- Take an example of eight data points, and for simplicity, we can consider them to be 1-D data

with values 1, 2, 3, 5, 9, 10, 11, and 25. Point 25 is the outlier, and it affects the cluster formation negatively when the mean of the points is considered as centroids.

With K = 2, the initial clusters we arrived at are {1, 2,3, 6} and {9, 10, 11, 25}.

$$\text{The mean of the cluster } \{1, 2, 3, 6\} = \frac{12}{4} = 3,$$

and the mean of the cluster

$$\{9, 10, 12, 25\} = \frac{56}{4} = 14.$$

So, the SSE within the clusters is

$$(1 - 3)^2 + (2 - 3)^2 + (3 - 3)^2 + (6 - 3)^2 + (9 - 14)^2 + (10 - 14)^2 + (12 - 14)^2 + (25 - 14)^2 = 179$$

If we compare this with the cluster {1, 2, 3, 6, 9} and {10, 11, 25},

$$\text{the mean of the cluster } \{1, 2, 3, 6, 9\} = \frac{21}{5} = 4.2,$$

Because the SSE of the second clustering is lower, k-means tend to put point 9 in the same cluster with 1, 2, 3, and 6 though the point is logically nearer to points 10 and 11. This skewedness is introduced due to the outlier point 25, which shifts the mean away from the centre of the cluster.

k-medoids provides a solution to this problem. Instead of considering the mean of the data points

and the mean of the cluster

$$\{10, 12, 25\} = \frac{47}{3} = 15.67.$$

So, the SSE within the clusters is

$$(1 - 4.2)^2 + (2 - 4.2)^2 + (3 - 4.2)^2 + (6 - 4.2)^2 + (9 - 4.2)^2 + (10 - 15.67)^2 + (12 - 15.67)^2 + (25 - 15.67)^2 = 113.84$$

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist\,(o_i, x)^2 \qquad (9.3)$$

where o_i is the representative point or object of cluster C_i.

in the cluster, k-medoids considers k representative data points from the existing points in the data set as the centre of the clusters. It then assigns the data points according to their distance from these centres to form k clusters. Note that the medoids in this case are actual data points or objects from the data set and not an imaginary point as in the case when the mean of the data sets within cluster is used as the centroid in the k-means technique. The SSE is calculated as Thus, the k-medoids method groups n objects in k clusters by minimizing the SSE. Because of the use of medoids from the actual representative data points, k-medoids is less influenced by the outliers in the data.

One of the practical implementation of the k-medoids principle is the **Partitioning Around Medoids (PAM) algorithm.**

Algorithm PAM

Step 1: Randomly choose k points in the data set as the initial representative points loop

Step 2: Assign each of the remaining points to the cluster which has the nearest representative point

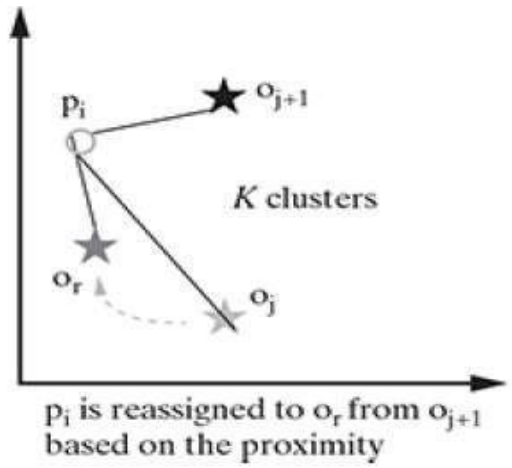
Step 3: Randomly select a non-representative point o in each cluster

Step 4: Swap the representative point o with o and compute the new SSE after swapping

Step 5: If SSE new < SSE old, then swap o with o to form the new set of k representative objects;

Step 6: Refine the k clusters on the basis of the nearest representative point. Logic continues until there is no change

end loop



. PAM algorithm: Reassignment of points to different clusters

2) Hierarchical methods

The hierarchical clustering methods are used to group the data into hierarchy or tree-like structure. For example, in a machine learning problem of organizing employees of a university in different departments, first the employees are grouped under the different departments in the university, and then within each department, the employees can be grouped according to their roles such as professors, assistant professors, supervisors, lab assistants, etc. This creates a hierarchical structure of the employee data and eases visualization and analysis.

Types of Hierarchical Clustering Methods

There are two main techniques –

a) Agglomerative clustering (Bottom-up technique)

b) Divisive clustering (Top-down)

a) Agglomerative clustering (Bottom-up technique)

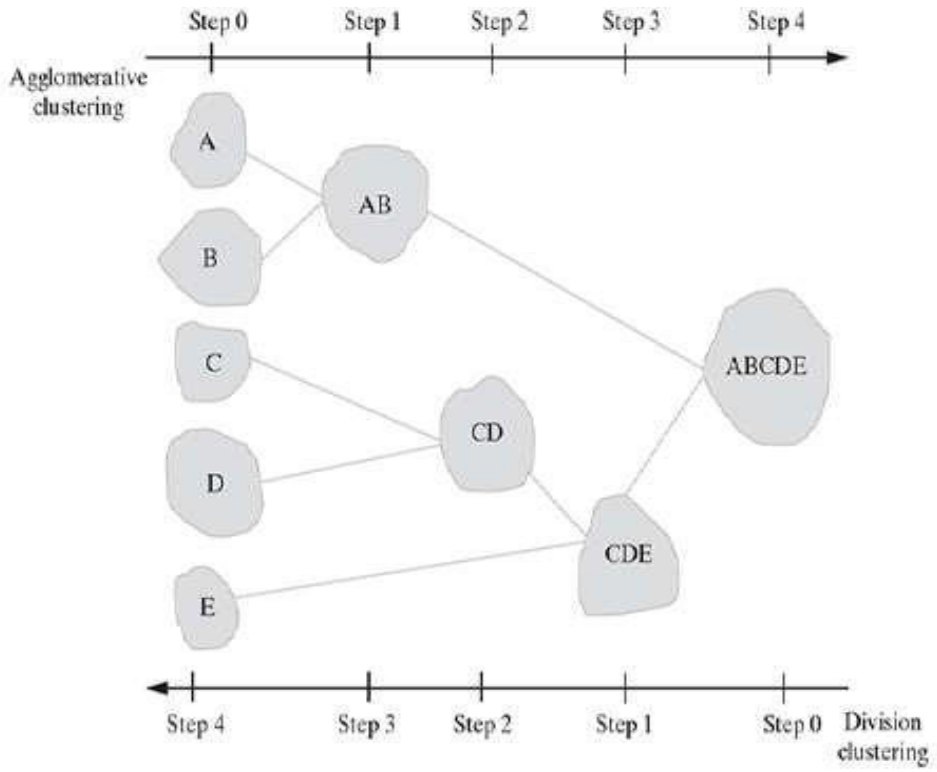
The agglomerative hierarchical clustering method uses the bottom-up strategy. It starts with each object forming its own cluster and then iteratively merges the clusters according to their similarity to form larger clusters. It terminates either when a certain clustering condition imposed by the user is achieved or all the clusters merge into a single cluster.

b) Divisive clustering (Top-down)

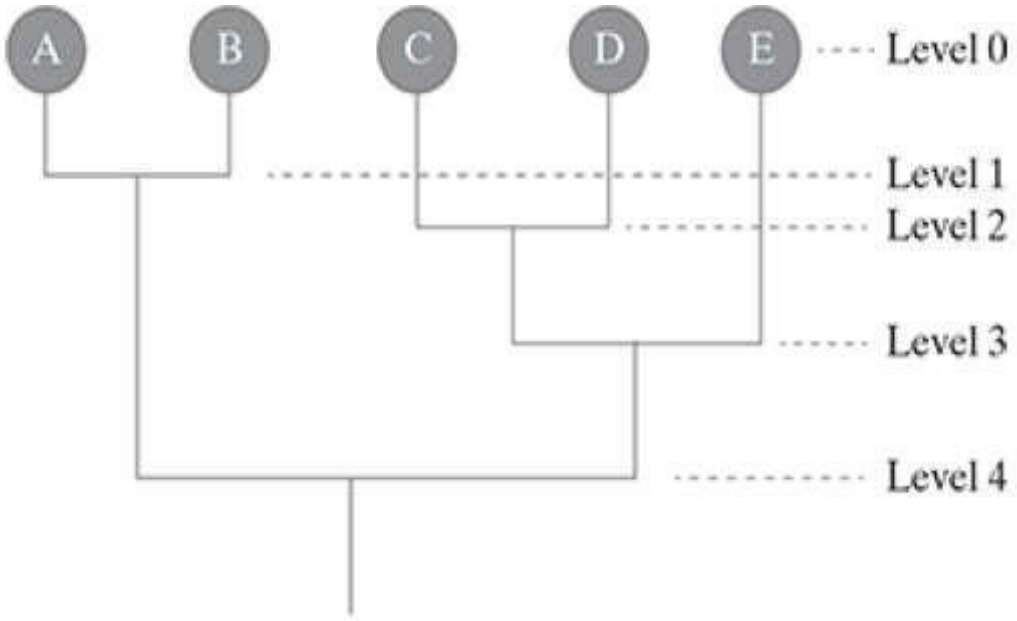
The divisive hierarchical clustering method uses a top-down strategy. The starting point is the largest cluster with all the objects in it, and then, it is split recursively to form smaller and smaller clusters, thus forming the hierarchy. The end of iterations is achieved when the objects in the final clusters are sufficiently homogeneous to each other or the final clusters contain only one object or the user-defined clustering condition is achieved.

Dendogram technique :-

A dendrogram is a commonly used tree structure representation of step-by-step creation of hierarchical clustering. It shows how the clusters are merged iteratively (in the case of agglomerative clustering) or split iteratively (in the case of divisive clustering) to arrive at the optimal clustering solution.



Agglomerative and divisive hierarchical clustering



Dendrogram representation of hierarchical clustering

Distance Measure

There are four standard methods to measure the distance between clusters:

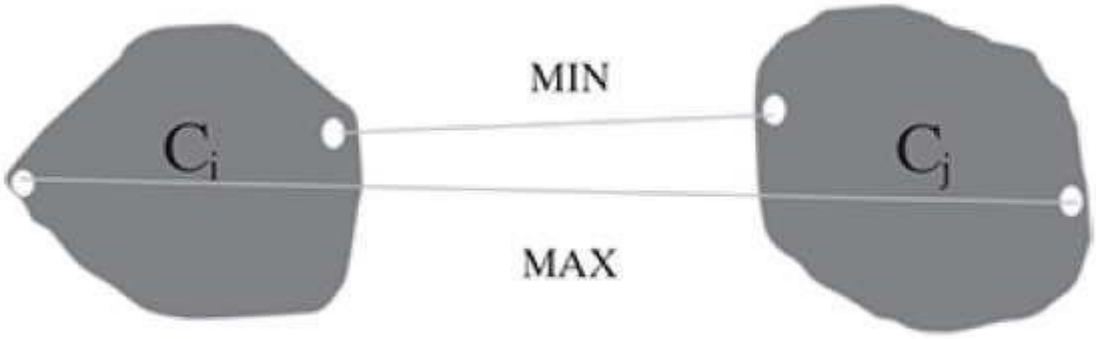
Let C_i and C_j be the two clusters with n_i and n_j respectively. p_i and p_j represents the points in clusters C_i and C_j respectively. We will denote the mean of cluster C_i as m_i.

Minimum distance $D_{\min}(C_i, C_j) = \min_{p_i \in C_i,\ p_j \in C_j}\{|p_i - p_j|\}$

Maximum distance $D_{\max}(C_i, C_j) = \max_{p_i \in C_i,\ p_j \in C_j}\{|p_i - p_j|\}$

Mean distance $D_{\text{mean}}(C_i, C_j) = \{|m_i - m_j|\}$

Average distance $D_{\text{avg}}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p_i \in C_i,\ p_j \in C_j} |pi - pj|$



Distance measure in algorithmic methods

3) **Density-based methods – DBSCAN**

- In the above two approaches, the resulting clusters are spherical or nearly spherical in nature.
- The density-based clustering approach provides a solution to identify clusters of arbitrary shapes.
- The principle is based on identifying the dense area and sparse area within the data set and then run the clustering algorithm.

- DBSCAN is one of the popular density-based algorithm which creates clusters by using connected regions with high density.

Differences between different CLUSTERING Methods

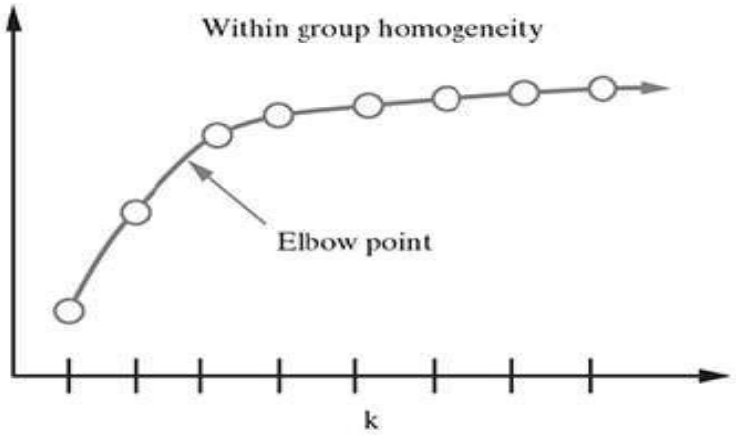
Method	Characteristics
Partitioning methods	• Uses mean or medoid (etc.) to represent cluster centre • Adopts distance-based approach to refine clusters • Finds mutually exclusive clusters of spherical or nearly spherical shape • Effective for data sets of small to medium size
Hierarchical methods	• Creates hierarchical or tree-like structure through decomposition or merger • Uses distance between the nearest or furthest points in neighbouring clusters as a guideline for refinement • Erroneous merges or splits cannot be corrected at subsequent levels
Density-based methods	• Useful for identifying arbitrarily shaped clusters • Guiding principle of cluster creation is the identification of dense regions of objects in space which are separated by low-density regions • May filter out outliers

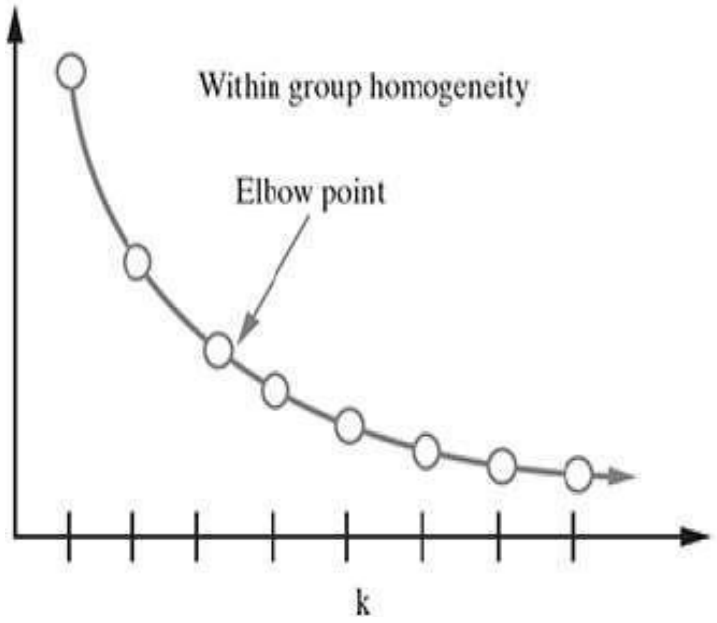
Elbow method and Elbow point

This method tries to measure the homogeneity or heterogeneity within the cluster and for various values of 'K' and helps in arriving at the optimal 'K'.

From the below figure, we can see the homogeneity will increase or heterogeneity will decrease with increasing 'K' as the number of data points inside each cluster reduces with this increase.

The point at which optimal clustering performance is produced is called as Elbow point.





FINDING PATTERN USING ASSOCIATION RULE

Association rule presents a methodology that is useful for identifying interesting relationships hidden in large data sets. It is also known as association analysis, and the discovered relationships can be represented in the form of association rules comprising a set of frequent items.

A common application of this analysis is the **Market Basket Analysis** that retailers use for cross-selling of their products.

The application of association analysis is also widespread in other domains such as bioinformatics, medical diagnosis, scientific data analysis, and web data mining.

For example, by discovering the interesting relationship between food habit and patients developing breast cancer, a new cancer prevention mechanism can be found which will benefit thousands of people in the world.

Few common terminologies used in association analysis

Item set

One or more items are grouped together and are surrounded by brackets to indicate that they form a set, or more specifically, an item set that appears in the data with some regularity.

{Bread, Milk, Egg} can be grouped together to form an item set as those are frequently bought together.

A collection of zero or more items is called an **item set**.

A null item set is the one which does not contain any item.

k-Item set: **In the association analysis**, an item set is called **k-item set** if it contains k number of items.

Thus, the item set {Bread, Milk, Egg} is a three-item set.

Support Count

Support count denotes the number of transactions in which a particular item set is present. This is a very important property of an item set as it denotes the frequency of occurrence for the item set. This is expressed as

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \varepsilon T\}|$$

where |{}| denotes the number of elements in a set

Market Basket Transaction Data

Transaction Number	Purchased Items
1	{Bread, Milk, Egg, Butter, Salt, Apple}
2	{Bread, Milk, Egg, Apple}
3	{Bread, Milk, Butter, Apple}
4	{Milk, Egg, Butter, Apple}
5	{Bread, Egg, Salt}
6	{Bread, Milk, Egg, Apple}

The item set {Bread, Milk, Egg} occurs together three times and thus have a support count of 3.

ASSOCIATION RULE

- ✓ The result of the market basket analysis is expressed as a set of association rules that specify patterns of relationships among items.
- ✓ A typical rule might be expressed as {Bread, Milk}→{Egg}, which denotes that if Bread and Milk are purchased, then Egg is also likely to be purchased. Thus, association rules are learned from subsets of item sets.

  It should be noted that an association rule is an expression of X → Y where X and Y are disjoint item sets, i.e. X ∩ Y = 0.

Measuring the strength of an association rule

Support and **confidence** are the two concepts that are used for measuring the strength of an association rule.

Support denotes how often a rule is applicable to a given data set.

Confidence indicates how often the items in Y appear in transactions that contain X in a total transaction of N.

$$\text{Support}, s(X \rightarrow Y) = \frac{\sigma(X \bigcup Y)}{N}$$

$$\text{Confidence}, c(X \rightarrow Y) = \frac{\sigma(X \bigcup Y)}{\sigma(X)}$$

Support s({Bread, Milk} -> {Egg}) = support count of {Bread, Milk, Egg} / Total transactions

$$= 3/6$$

$$= 0.5$$

$$\text{Confidence } c(\{\text{Bread, Milk}\} \rightarrow \{\text{Egg}\}) = \frac{\text{support count of } \{\text{Bread, Milk, Egg}\}}{\text{support count of } \{\text{Bread, Milk}\}}$$

$$= \frac{3}{4}$$

$$= 0.75$$

A low support may indicate that the rule has occurred by chance. Also, from its application perspective, this rule may not be a very attractive business investment as the items are seldom bought together by the customers. Thus, support can provide the intelligence of identifying the most interesting rules for analysis.

Similarly, **confidence** provides the measurement for reliability of the inference of a rule. Higher confidence of a rule X → Y denotes more likelihood of to be present in transactions that contain X as it is the estimate of the conditional probability of Y given X.

The APRIORI algorithm for association rule learning

The main challenge of discovering an association rule and learning from it is the large volume of transactional data and the related complexity. Because of the variation of features in transactional data, the number of feature sets within a data set usually becomes very large. This leads to the problem of handling a very large number of item sets, which grows exponentially with the number of features.

Following steps are followed for generating association rules:

1. Decide the minimum support and minimum confidence of the association rules. From a set of transaction T, let us assume that we will find out all the rules that have support ≥ min S and confidence ≥ min C, where min S and min C are the support and confidence thresholds, respectively, for the rules to be considered acceptable.
2. Generate Frequent
3. Generate Rules

Build the a priori principle rules

One of the most widely used algorithm to reduce the number of item sets to search for the association rule is known as A priori.

a) If an item set is frequent, then all of its subsets must also be frequent.
b) If an item set is frequent, then all the supersets must be frequent too.

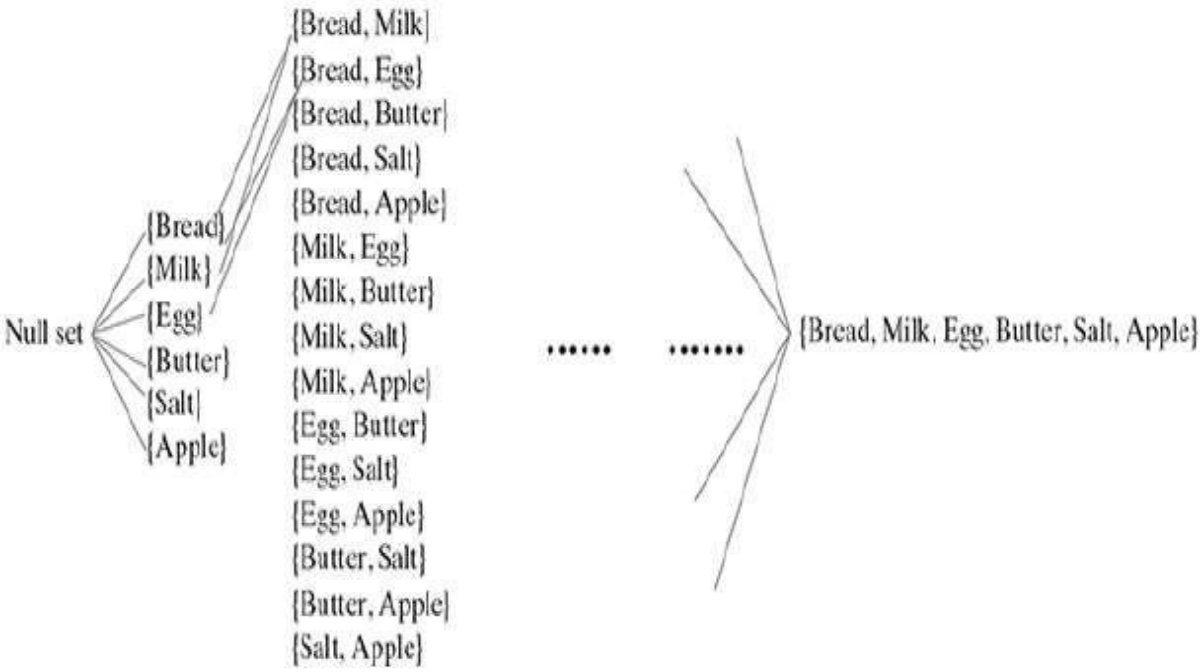
Demonstration of A priori principle

a. Consider the dataset given below

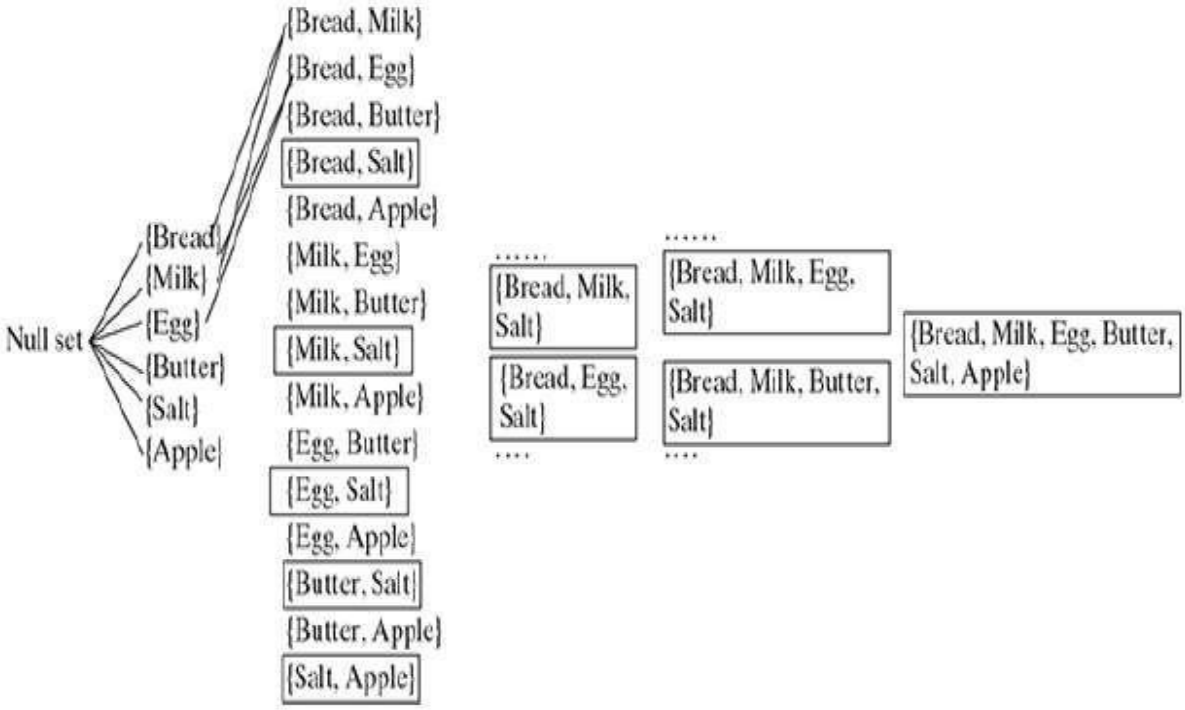
Market Basket Transaction Data

Transaction Number	Purchased Items
1	{Bread, Milk, Egg, Butter, Salt, Apple}
2	{Bread, Milk, Egg, Apple}
3	{Bread, Milk, Butter, Apple}
4	{Milk, Egg, Butter, Apple}
5	{Bread, Egg, Salt}
6	{Bread, Milk, Egg, Apple}

From the full item set of six items {Bread, Milk, Egg, Butter, Salt, Apple}, there are 2^6 ways to create baskets or item sets (including the null item set) as shown in the below



Sixty-four ways to create itemsets from 6 items



Discarding the itemsets consisting of Salt

The actual process of creating rules involves two phases:

a. Identifying all item sets that meet a minimum support threshold set for the analysis
b. Creating rules from these item sets that meet a minimum confidence threshold which identifies the strong rules

Strengths and Weaknesses of apriori algorithm

Strengths	Weaknesses
• Provides reasonable accuracy while working with very large amounts of transactional data • Discovers rules that are easy to understand • Provides valuable insight into the unexpected knowledge in data sets, which is a key aspect of learning	• Not very accurate in the case the data set is small as the smaller occurrences of itemsets may not be due to chance • Some effort is involved to separate the insight from the common sense • In the case of widespread presence of random patterns, the principle can draw spurious conclusions

----------END----------