MATHEMATICAL FOUNDATIONS OF COMPUTER SCIENCE (BCSB01)

I M. Tech I semester (Autonomous IARE R-18)

BY

Mrs. G Sulakshana

Assistant Professor, CSE

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

DUNDIGAL, HYDERABAD - 500 043

UNIT – I
INTRODUCTION

INTRODUCTION TO PROBABILITY MASS

PROBABILITY MASS:

A probability distribution over discrete variables may be described using a probability mass function (PMF).

Probability mass functions can act on many variables at the same time. Such a probability distribution over many variables is know.

P(x=x, y=y) denotes the probability that x=x and y=y simultaneously. We may also write P (x, y) for as a joint probability distribution.

PROBABILITY DENSITY FUNCTION

When working with continuous random variables, we describe probability distributions using a probability density function (PDF) rather than a probability mass function.

To be a probability density function, a function p must satisfy the following properties:

- The domain of p must be the set of all possible states of x.
  - $\forall x \in x$, $p(x) \geq 0$. Note that we do not require $p(x)$
    - $p(x)dx = 1$

EXPECTED VALUE, VARIANCE, CONDITIONA EXPECTATION

The mean, expected value, or expectation of a random variable X is written as E(X) or μX.

If we observe N random values of X, then the mean of the N values will be approximately equal to E(X) for large N.

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

CENTRAL LIMIT THEOREM (CLT):

Analyzing data involves statistical methods like hypothesis testing and constructing confidence intervals.

These methods assume that the population is normally distributed.

In the case of unknown or non-normal distributions, we treat the sampling distribution as normal according to **the central limit theorem**

PROBABILISTIC INEQUALITIES

We often want to bound the probability that X is too far away from its expectation.

In first class, we went in other direction, saying that with reasonable probability, a random walk on n steps reached at least √n distance away from its expectation.

Here are some useful inequalities for showing this:

MARKOV'S INEQUALITY: Let X be a non-negative r.v. Then for any positive k:

UNIT – II
RANDOM SAMPLES

RANDOM SAMPLES

Random sampling is a way of selecting a sample of observations from a population in order to make inferences about the population.

The main forms of random sampling are simple random sampling, stratified sampling, cluster sampling, and multistage sampling.

Samples that are not random are typically called convenience samples.

Simple random sampling is the most straightforward approach for getting a random sample

CLUSTER SAMPLING

Cluster sampling is like stratified random sampling, except that the population is divided into a large number of subgroups.

Then some of these subgroups are selected at random, and simple random samples are then collected within these subgroups.

These subgroups are called clusters.

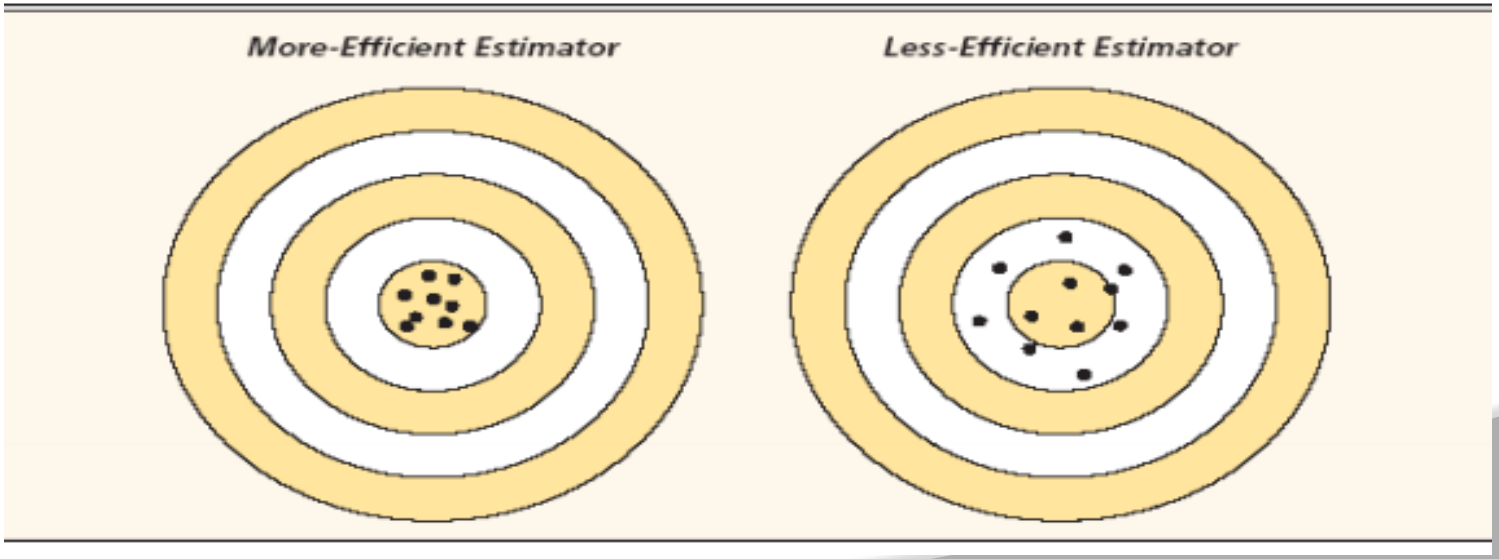
Note: the purpose of cluster sampling is to reduce the costs of data collection.

SAMPLING DISTRIBUTIONS OF ESTIMATOR

In statistics, it is the probability distribution of the given statistic estimated on the basis of a random sample.

It provides a generalized way to statistical inference. The estimator is the generalized mathematical parameter to calculate sample statistics.

An estimate is the result of the estimation.



METHODS OF MOMENTS AND MAXIMUM LIKELIHOOD

Methods of Moments and Maximum Likelihood

Random sample X1, · · · , Xn from the probability distribution f(x|θ) with unknown parameter(s) θ.

f(x|θ) could either be density function or probability mass function depending on the problem at hand.

The purpose is to estimate the unknown parameter(s) θ from the sample X1, · · · , Xn.

We have so far investigated two methods for estimating parameters, namely**, method of moment and maximum likelihood**

UNIT – III
STATISTICAL INTERFACE

STATISTICAL INFERENCE

- Statistical Interference is the process of using data analysis to deduce properties of an underlying probability distribution.
- Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates.
- t is assumed that the observed data set is sampled from a larger population.

- The conclusion of a statistical inference is a statistical proposition. Some common forms of statistical proposition are the following:
- a point estimate, i.e. a particular value that best approximates some parameter of interest;
- an interval estimate, e.g. a confidence interval (or set estimate), i.e. an interval constructed using a dataset drawn from a population so that, under repeated sampling of such datasets, such intervals would contain the true parameter value with the probability at the stated confidence level;
- a credible interval, i.e. a set of values containing, for example, 95% of posterior belief; rejection of a hypothesis clustering or classification of data points into groups.

DEGREE OF MODELS/ASSUMPTIONS

Fully parametric

- The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters.

- The probability distributions describing the data-generation process are assumed to be fully described by a family of probability distributions involving only a finite number of unknown parameters.

Non-parametric

The assumptions made about the process generating the data are much less than in parametric statistics and may be minimal

For example, every continuous probability distribution has a median, which may be estimated using the sample median or the which has good properties.

when the data arise from simple random sampling.

Semi-parametric

This term typically implies assumptions in between fully and non-parametric approaches.

For example, one may assume that a population distribution has a finite mean.

MULTIVARIATE ANALYSIS

Introduction to Multivariate Analysis:

Multivariate analysis is an analysis of statistical technique that analyze the relationship between more than two variables which shows the effect of more than one variable on one variable that is independent

. Multivariate analysis helps the organization in decision making for the future. Applied multivariate analysis refers to the application of multivariate statistical techniques to the problems of market researches; analysts and business researchers use multivariate techniques for analyzing the data variable.

PRINCIPAL COMPONENTS ANALYSIS

- The central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of interrelated variables while retaining as much as possible of the variation present in the data set.

- This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

- **PCA:**
- PCA can be thought of as an unsupervised learning problem. The whole process of obtaining principle components from a raw dataset can be simplified in six parts :
- Take the whole dataset consisting of *d+1 dimensions* and ignore the labels such that our new dataset becomes *d dimensional.*
- Compute the *mean* for every dimension of the whole dataset.
- Compute the *covariance matrix* of the whole dataset.
- Compute *eigenvectors* and the corresponding *eigen values*.
- Sort the eigenvectors by decreasing eigen values and choose k eigenvectors with the largest eigen values to form a $d \times k$ *dimensional* matrix **W.**
- Use this $d \times k$ *eigenvector matrix* to transform the samples onto the new subspace.

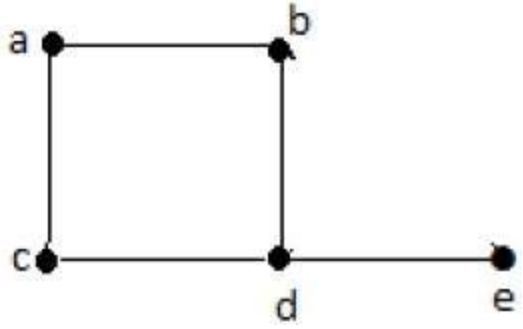
THE PROBLEM OF OVER FITTING MODEL ASSESSMENT

- Over fitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables.
- This problem occurs when the model is too complex.
- In regression analysis, over fitting can produce misleading R-squared values, regression coefficients, and p-values.

UNIT – IV
GRAPH THEORY

GRAPH THEORY

- A graph is a pictorial representation of a set of objects where some pairs of objects are connected by links. The interconnected objects are represented by points termed as **vertices,** and the links that connect the vertices are called **edges**.
- Formally, a graph is a pair of sets **(V, E),** where **V** is the set of vertices and **E** is the set of edges, connecting the pairs of vertices.
- Take a look at the following graph −



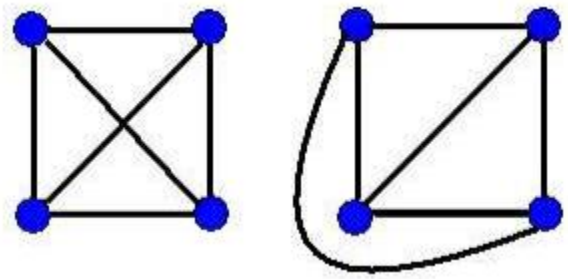
Applications of Graph Theory

- Graph theory has its applications in diverse fields of engineering –
- **Electrical Engineering** – The concepts of graph theory is used extensively in designing circuit connections. The types or organization of connections are named as topologies. Some examples for topologies are star, bridge, series, and parallel topologies.
- **Computer Science** – Graph theory is used for the study of algorithms. For example,
- Kruskal's Algorithm
- Prim's Algorithm
- Dijkstra's Algorithm

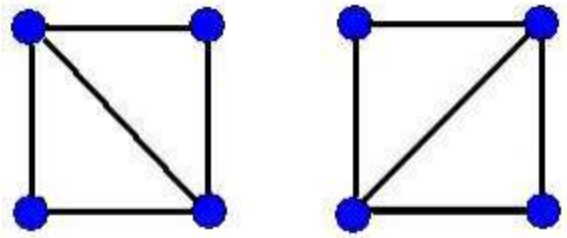
- **Computer Network** – The relationships among interconnected computers in the network follows the principles of graph theory.
- **Science** – The molecular structure and chemical structure of a substance, the DNA structure of an organism, etc., are represented by graphs.
- **Linguistics** – The parsing tree of a language and grammar of a language uses graphs.
- **General** – Routes between the cities can be represented using graphs. Depicting hierarchical ordered information such as family tree can be used as a special type of graph called tree.

Isomorphism

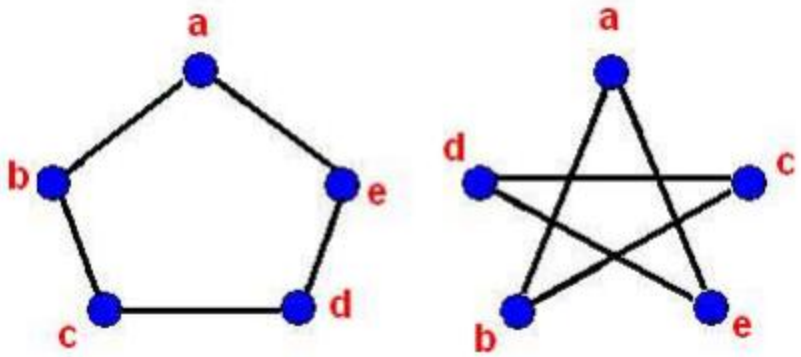
- There are different ways to draw the same graph. Consiser the following two graphs



You probably feel that these graphs do not differ from each other. What about

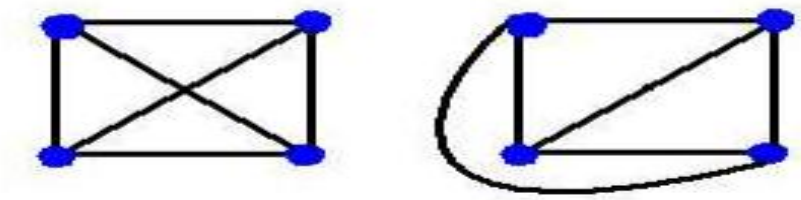


Graphs G1 V1, E1 and G2 V2, E2 are isomorphic if 1. there is a bijection (one-to-one correspondence) f from V1 to V2 and 2. there is a bijection g from E1 to E2 that maps each edge v, Example of non-isomorphic graphs

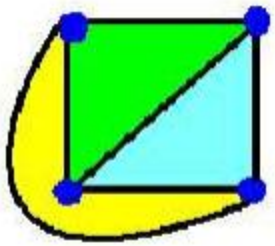


Planar graphs

- An undirected graph is called a planar graph if it can be drawn on a paper without having two edges cross.



- We say that a graph can be embedded in the plane, if it planar. A planar graph divides the plane into regions (bounded by the edges), called faces. The following planar graph has 4 faces.



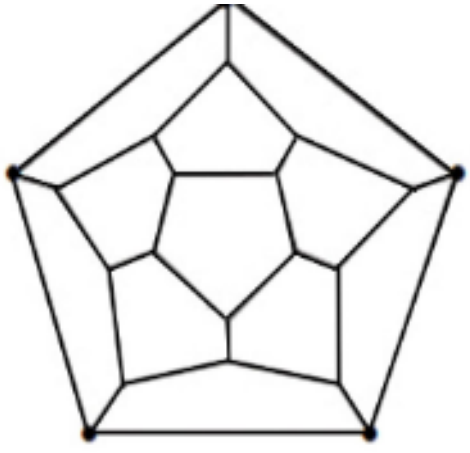
Graph coloring

- If you ever decide to create a map and need to color the parts of it optimally, feel lucky because graph theory is by your side.
- What is the maximum number of colors required to color the regions of a map?
- This question along with other similar ones have generated a lot of results in graph theory.
  First, let us define the constraint of coloring in a formal way-
- **Coloring –** "A coloring of a simple graph is the assignment of a color to each vertex of the graph such that **no two adjacent vertices** are assigned the same color."

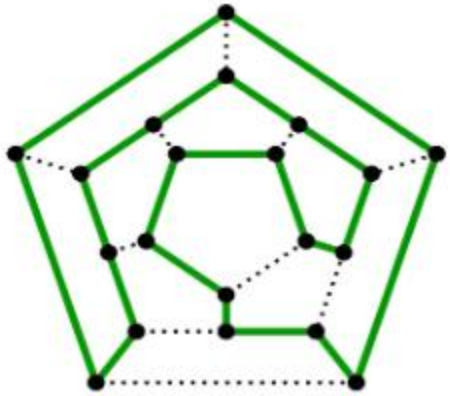
chromatic number

- –“The least number of colors required to color a graph is called its **chromatic number**. It is denoted by
- **Hamilonian Circuit –** A simple circuit in a graph that passes through every vertex exactly once is called a Hamiltonian circuit.
- Unlike Euler paths and circuits, there is no simple necessary and sufficient criteria to determine if there are any Hamiltonian paths or circuits in a graph.
- But there are certain criteria which rule out the existence of a Hamiltonian circuit in a graph, such as- if there is a vertex of degree one in a graph then it is impossible for it to have a Hamiltonian circuit.

- **Example 1-** Does the following graph have a Hamiltonian Circuit?



- **Solution-** Yes, the above graph has a Hamiltonian circuit. The solution is –



- **Euler paths and circuits :**
- An Euler path is a path that uses every edge of a graph **exactly once**.
- An Euler circuit is a circuit that uses every edge of a graph exactly once.
- An Euler path starts and ends at different vertices.

PERMUTATIONS AND COMBINATIONS WITH AND WITHOUT REPETITION.

- **Permutations**
- There are basically two types of permutation:

- **Repetition is Allowed**: such as the lock above. It could be "333".

- **No Repetition**: for example the first three people in a running race. You can't be first and second.

- **Permutations with Repetition**
- These are the easiest to calculate.
- When a thing has **n** different types ... we have **n** choices each time!
- For example: choosing **3** of those things, the permutations are:
- **n × n × n**
  (n multiplied 3 times)
- More generally: choosing **r** of something that has **n** different types, the permutations are:
- **n × n × ... (r times)**

- **Permutations without Repetition**
- In this case, we have to **reduce** the number of available choices each time.

- **Combinations**

- There are also two types of combinations (remember the order does **not** matter now):

- **Repetition is Allowed**: such as coins in your pocket (5,5,5,10,10)

- **No Repetition**: such as lottery numbers (2,14,15,27,30,33)

- **Combinations with Repetition**
- Actually, these are the hardest to explain, so we will come back to this later.

- **Combinations without Repetition**
- This is how lorries work. The numbers are drawn one at a time, and if we have the lucky numbers (no matter what order) we win!

UNIT – V
COMPUTER SCIENCE AND ENGINEERING APPLICATIONS

DATA MINING

- It is defined as the process of discovering patterns in data. The process must be automatic or semiautomatic.
- The pattern discovered must be meaningful, in that they lead to some advantage, usually economic advantage.
- The data is invariably present in substantial quantities.
- Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

DATA MINING

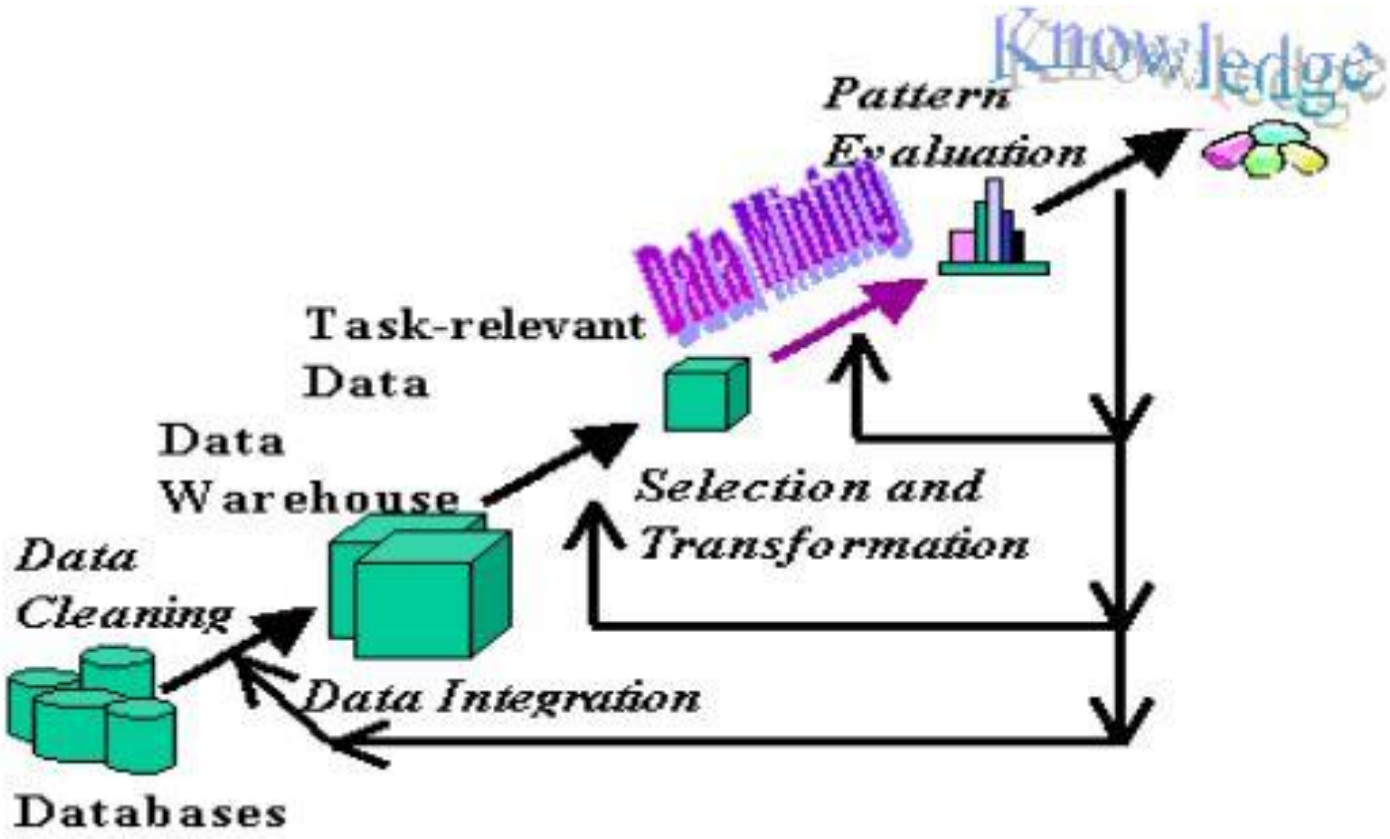


Figure 1.1: Data Mining is the core of Knowledge Discovery process

DATA MINING APPLICATION AREAS:

- Business Transactions
- E-commerce
- Scientific Study
- Health Care Study
- Web Study
- Crime Detection
- Loan Delinquency
- Banking

DATA MINING TECHNIQUES:

- Association Rule Mining
- Cluster Analysis
- Classification Rule Mining
- Frequent Episodes
- Deviation Detection/ Outlier Analysis
- Genetic Algorithms
- Rough Set Techniques
- Support Vector Machines

NETWORK PROTOCOLS:

- Network protocols are formal standards and policies comprised of rules, procedures and formats that define communication between two or more devices over a network.

- Network protocols govern the end-to-end processes of timely, secure and managed data or network communication.

- There are several broad types of networking protocols, including:
- Network communication protocols: Basic data communication protocols, such as TCP/IP and HTTP.
- Network security protocols: Implement security over network communications and include HTTPS, SSL and SFTP.
- Network management protocols: Provide network governance and maintenance and include SNMP and ICMP.

Analysis of Web Traffic :

- When creating website content or changing the keywords in your website, measuring those changes is necessary to know if the changes made an impact to your web traffic which you are targeting.

- With web analytics tools that track directly the visitors and its interaction to every page, measuring of web traffic is easy. If you will not analyze and measure your web traffic, you will not be able to expand and effectively grow your presence in the web.

COMPUTER SECURITY:

- Computer security basically is the protection of computer systems and information from harm, theft, and unauthorized use.
- It is the process of preventing and detecting unauthorized use of your computer system.

- The components of a computer system that needs to be protected are:
- Hardware, the physical part of the computer, like the system memory and disk drive

- Firmware, permanent software that is etched into a hardware device's nonvolatile memory and is mostly invisible to the user

- Software, the programming that offers services, like operating system, word processor, internet browser to the user

SOFTWARE ENGINEERING:

- Software engineering is a detailed study of engineering to the design, development and maintenance of software.
- Software engineering was introduced to address the issues of low-quality software projects.
- Problems arise when a software generally exceeds timelines, budgets, and reduced levels of quality.
- It ensures that the application is built consistently, correctly, on time and on budget and within requirements.
- The demand of software engineering also emerged to cater to the immense rate of change in user requirements and environment on which application is supposed to be working.

DATA MINING

- It is defined as the process of discovering patterns in data. The process must be automatic or semiautomatic.
- The pattern discovered must be meaningful, in that they lead to some advantage, usually economic advantage.
- The data is invariably present in substantial quantities.
- Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

COMPUTER ARCHITECTURE :

- Computer architecture is a specification detailing how a set of software and hardware technology standards interact to form a computer system or platform.

- In short, computer architecture refers to how a computer system is designed and what technologies it is compatible with.

DATA MINING

- It is defined as the process of discovering patterns in data. The process must be automatic or semiautomatic.
- The pattern discovered must be meaningful, in that they lead to some advantage, usually economic advantage.
- The data is invariably present in substantial quantities.
- Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

- System Design: This includes all hardware components in the system, including data processors aside from the CPU, such as the graphics processing unit and direct memory access.
- Instruction Set Architecture (ISA): This is the embedded programming language of the central processing unit. It defines the CPU's functions and capabilities based on what programming it can perform or process. This includes the word size, processor register types, memory addressing modes, data formats and the instruction set that programmers use.
- Micro architecture: Otherwise known as computer organization, this type of architecture defines the data paths, data processing and storage elements, as well as how they should be implemented in the ISA.

DATA MINING

- It is defined as the process of discovering patterns in data. The process must be automatic or semiautomatic.
- The pattern discovered must be meaningful, in that they lead to some advantage, usually economic advantage.
- The data is invariably present in substantial quantities.
- Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

Operating Systems:

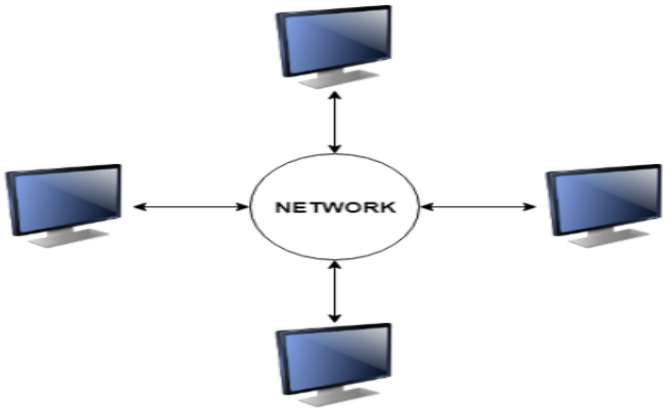
- An operating system is a software which acts as an interface between the end user and computer hardware.
- Every computer must have at least one OS to run other programs.
- An application like Chrome, MS Word, Games, etc needs some environment in which it will run and perform its task

Types of Operating system

- Batch Operating System
- Multitasking/Time Sharing OS
- Multiprocessing OS
- Real Time OS
- Distributed OS
- Network OS
- Mobile OS

Distributed Systems:

- **5.8 Distributed Systems:**
- A distributed system contains multiple nodes that are physically separate but linked together using the network. All the nodes in this system communicate with each other and handle processes in tandem. Each of these nodes contains a small part of the distributed operating system software.
- A diagram to better explain the distributed system is:



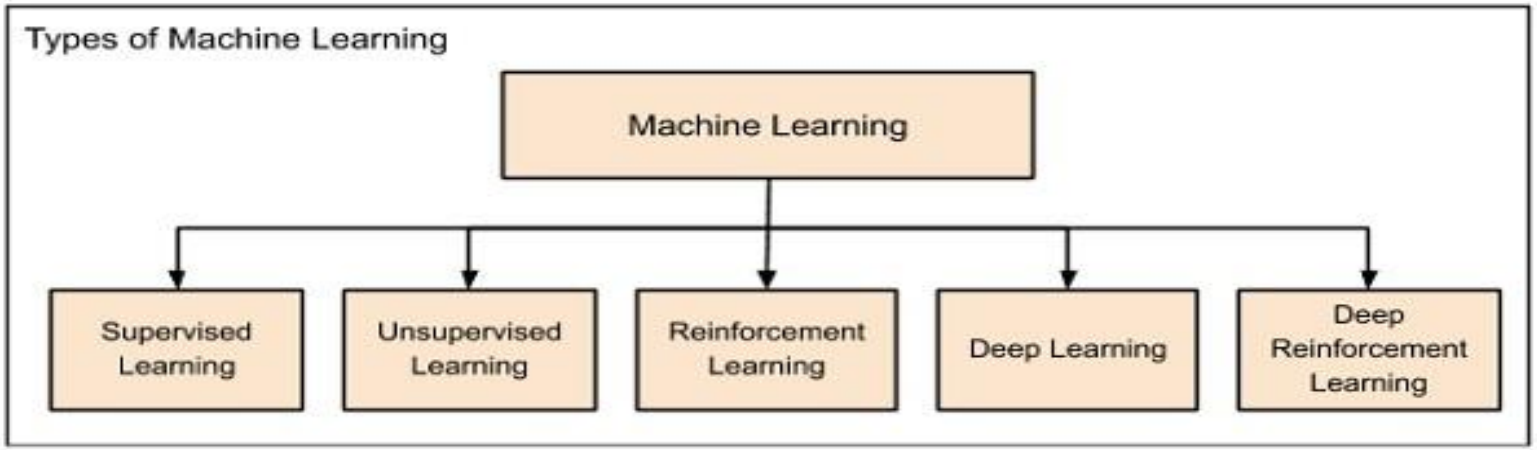
DISTRIBUTED OPERATING SYSTEM

MACHINE LEARNING

- Today's Artificial Intelligence (AI) has far surpassed the hype of blockchain and quantum computing.
- This is due to the fact that huge computing resources are easily available to the common man.
- The developers now take advantage of this in creating new Machine Learning models and to re-train the existing models for better performance and results.
- The easy availability of High Performance Computing (HPC) has resulted in a sudden increased demand for IT professionals having Machine Learning skills.

Some of the examples of statistical techniques that are used for developing AI applications in those days and are still in practice are listed here −

- Regression
- Classification
- Clustering
- Probability Theories
- Decision Trees



Deep Learning

- The deep learning is a model based on Artificial Neural Networks (ANN), more specifically Convolutional Neural Networks (CNN)s.
- There are several architectures used in deep learning such as deep neural networks, deep belief networks, recurrent neural networks, and convolutional neural networks.