LECTURE NOTES

ON

MATHEMATICAL FOUNDATIONS OF COMPUTER SCIENCE

2019 – 2020

M.Tech I Semester (IARE-R18)

G. Sulakshana, **Assistant Professor**

Department of Computer Science and Engineering

INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad - 500 043

UNIT – I
INTRODUCTION

1.1 PROBABILITY MASS:

Probability theory is a fundamental tool of many disciplines of science and engineering. While probability theory allows us to make uncertain statements and to reason in the presence of uncertainty, information theory enables us to quantify the amount of uncertainty in a probability distribution.

Random Variables

A random variable is a variable that can take on different values randomly. We typically denote the random variable itself with a lowercase letter in plain typeface, and the values it can take on with lowercase script letters. For example,x1 and x2 are both possible values that the random variable x can take on. For vector-valued variables, we would write the random variable as x and one of its values as x. On its own, a random variable is just a description of the states that are possible; it must be coupled with a probability distribution that specifies how likely each of these states are.

Random variables may be discrete or continuous. A discrete random variable is one that has a finite or countably infinite number of states. Note that these states are not necessarily the integers; they can also just be named states that are not considered to have any numerical value. A continuous random variable is associated with a real value.

Probability Distributions

A probability distribution is a description of how likely a random variable or set of random variables is to take on each of its possible states. The way we describe probability distributions depends on whether the variables are discrete or continuous.

Discrete Variables and Probability Mass Functions

A probability distribution over discrete variables may be described using a probability mass function (PMF). We typically denote probability mass functions with a capital P. Often we associate each random variable with a different probability mass function and the reader must infer which PMF to use based on the identity of the random variable, rather than on the name of the function; P(x) is usually not the same as P (y).

The probability mass function maps from a state of a random variable to the probability of that random variable taking on that state. The probability that x=x is denoted as P(x), with a probability of 1 indicating that x=x is certain and a probability of 0 indicating that x=x is impossible.

Sometimesto disambiguate which PMF to use, we write the name of the random variable explicitly: P(x=x). Sometimes we define a variable first, then use ∼ notation to specify which distribution it follows later: x ∼ P (x).

Probability mass functions can act on many variables at the same time. Such a probability distribution over many variables is known as a joint probability distribution. P(x=x, y=y) denotes the probability that x=x and y=y simultaneously. We may also write P (x, y) for brevity.

To be a PMF on a random variable x, a function P must satisfy the following properties:

• The domain of P must be the set of all possible states of x.

• ∀x ∈ x, 0≤ P(x)≤1.An impossible event has probability 0, and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring.

•x ∈ x P(x) = 1. We refer to this property as being normalized. Without this property, we could obtain probabilities greater than one by computing the probability of one of many event occurring.

For example, consider a single discrete random variable x with k diff erent states. We can place a uniform distribution on x— that is, make each of its states equally likely—by setting its PMF to

$$P(x = x_i) = 1/k$$

for all i. We can see that this fits the requirements for a probability mass function.The value1kis positive because k is a positive integer. We also see that

$$\sum_i P(x = x_i) = \sum_i 1/k = k/k = 1,$$

so the distribution is properly normalized.

1.2 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 $p(x) \leq 1$.
- $\int p(x)dx = 1$.

A probability density functionp(x) does not give the probability of a specific state directly; instead the probability of landing inside an infinitesimal region with volume δx is given by p(x)δx.

We can integrate the density function to find the actual probability mass of a set of points. Specifically, the probability that x lies in some set S is given by the integral of p (x) over that set. In the univariate example, the probability that x lies in the interval [a, b] is given by $\int_{[a,b]} p(x)dx$

1.3 CUMULATIVE DISTRIBUTION FUNCTION $F_X(.)$

A random variable X is called continuous if there exist a function $f_X(.)$ such that $F_X(x) = x\ R\ -\infty\ f_X(u)du$ for every real number x. In such a case FX(x) is the cumulative distribution and the function $f_X(.)$ is the density function.

Notice that according to the above definition the density function is not uniquely determined. The idea is that if the a function change value if a few points its integral is unchanged. Furthermore, notice that $f_X(x) = dF_X(x)/dx$.

The notations for discrete and continuous density functions are the same, yet they have di ff erent interpretations. We know that for discrete random variables $f_X(x)=P[X = x]$, which is not true for continuous random variables. Furthermore, for discrete random variables $f_X(.)$ is a function with domain the real line and counter domain the interval [0,1], whereas, for continuous random variables $f_X(.)$ is a function with domain the real line and counter domain the interval $[0,\infty)$. Note that for a continuous r.v.

$P(X = x) \leq P(x-\varepsilon \leq X \leq x) = F_X(x) - F_X(x-\varepsilon) \to 0$ as $\varepsilon \to 0$, by the continuity of $F_X(x)$. The set {X = x} is an example of a set of measure (in this case the measure is P or P_X) zero. In fact, any countable set is of measure zero under a distribution which is absolutely continuous with respect to

Lebesgue measure. Because the probability of a singleton is zero

$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X < b)$

for any a, b.

1.4 PARAMETRIC FAMILIES OF DISTRIBUTIONS

A parametric family of density functions is a collection of density functions that are indexed by a quantity called parameter, e.g. let $f(x;\lambda) = \lambda e^{-\lambda x}$ for $x>0$ and some $\lambda>0$. λ is the parameter, and as λ ranges over the positive numbers, the collection $\{f(.;\lambda):\lambda>0\}$ is a parametric family of density functions.

UNIFORM:

Suppose that for $j = 1, 2, 3, ..., n$

$$P(X = x_j | \mathcal{X}) = \frac{1}{n}$$

where $\{x_1, x_2, ...x_n\} = \mathcal{X}$ is the support. Then

$$E(X) = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad Var(X) = \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2.$$

The c.d.f. here is

$$P(X \leq x) = \frac{1}{n}\sum_{j=1}^{n} 1(x_j \leq x)$$

BERNOULLI:

A random variable whose outcome have been classified into two categories, called “success” and “failure”, represented by the letters s and f, respectively, is called a Bernoulli trial. If a random variable X is defined as 1 if a Bernoulli trial results in success and 0 if the same Bernoulli trial results in failure, then X has a Bernoulli distribution with parameter p = P[success]. The definition of this distribution is:

A random variable X has a Bernoulli distribution if the discrete density of X is given by:

$$f_X(x) = f_X(x;p) = \begin{cases} p^x(1-p)^{1-x} & for \quad x = 0, 1 \\ 0 & otherwise \end{cases}$$

where $p = P[X = 1]$. For the above defined random variable X we have that:

$$E[X] = p \qquad and \qquad var[X] = p(1-p)$$

BINOMIAL:

Consider a random experiment consisting of n repeated independent Bernoulli trials with p the probability of success at each individual trial. Let the random variable X represent the number of successes in the n repeated trials. Then X follows a Binomial distribution. The definition of this distribution is: A random variable X has a binomial distribution, X ~ Binomial(n,p), if the discrete density of X is given by:

$$f_X(x) = f_X(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & for \quad x = 0, 1, ..., n \\ 0 & otherwise \end{cases}$$

where $p = P[X = 1]$ i.e. the probability of success in each independent Bernoulli trial and n is the total number of trials. For the above defined random variable X we have that:

$$E[X] = np \qquad and \qquad var[X] = np(1-p)$$

Mgf

$$M_X(t) = \left[pe^t + (1-p)\right]^n .$$

HYPERGEOMETRIC:

Let X denote the number of defective balls in a sample of size n when sampling is done without replacement from a box containing M balls out of which K are defective. The X has a hypergeometric distribution. The definition of this distribution is:

A random variable X has a hypergeometric distribution if the discrete density of X is given by:

$$f_X(x) = f_X(x; M, K, n) = \begin{cases} \dfrac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} & for \quad x = 0, 1, ..., n \\ 0 & otherwise \end{cases}$$

where M is a positive integer, K is a nonnegative that is at most M, and n is a positive integer that is at most M. For this distribution we have that:

$$E[X] = n\frac{K}{M} \qquad and \qquad var[X] = n\frac{K}{M}\frac{M-K}{M}\frac{M-n}{M-1}$$

Notice the difference of the binomial and the hypergeometric i.e. for the binomial distribution we have Bernoulli trials i.e. independent trials with fixed probability of success or failure, whereas in the hypergeometric in each trial the probability of success or failure changes depending on the result.

GEOMETRIC:

Consider a sequence of independent Bernoulli trials with p equal the probability of success on an

individual trial. Let the random variable X represent the number of trials required before the first success. Then X has a geometric distribution. The definition of this distribution is: A random variable X has a geometric distribution, X ~ geometric(p) , if the discrete density of X is given by:

$$f_X(x) = f_X(x;p) = \begin{cases} p(1-p)^x & for \quad x = 0, 1, ..., n \\ 0 & otherwise \end{cases}$$

where p is the probability of success in each Bernoulli trial. For this distribution we have that:

$$E[X] = \frac{1-p}{p} \qquad and \qquad var[X] = \frac{1-p}{p^2}$$

It is worth noticing that the Binomial distribution Binomial(n,p) can be approximated by a Poisson(np) (see below). The approximation is more valid as n→∞,p→0, in such a way so that np = constant.

POISSON:

A random variableX has a Poisson distribution, X ~ Poisson(λ), if the discrete density of X is given by:

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, 3, ...$$

In calculations with the Poisson distribution we may use the fact that

$$e^t = \sum_{j=0}^{\infty} \frac{t^j}{j!} \quad for \quad any \quad t.$$

Employing the above we can prove that

$$E(X) = \lambda, \quad E(X(X-1)) = \lambda^2, \quad Var(X) = \lambda.$$

The Poisson distribution provides a realistic model for many random phenomena. Since the values of a Poisson random variable are nonnegative integers, any random phenomenon for which a count of some sort is of interest is a candidate for modeling in assuming a Poisson distribution. Such a count might be the number of fatal traffic accidents per week in a given place, the number of telephone calls per hour, arriving in a switchboard of a company, the number of pieces of information arriving per hour, etc.

1.5 EXPECTED VALUE, VARIANCE, CONDITIONAL 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. The expectation is defined diff erently for continuous and discrete random variables.

Definition: Let X be a continuous random variable with p.d.f. $f_X(x)$. The expected value of X is

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

Definition: Let X be a discrete random variable with probability function $f_X(x)$. The expected value of X is

$$\mathbb{E}(X) = \sum_x x f_X(x) = \sum_x x \mathbb{P}(X = x).$$

Properties of Expectation

i) Let g and h be functions, and let a and b be constants. For any random variable X (discrete or continuous),

$$\mathbb{E}\Big\{ag(X) + bh(X)\Big\} = a\mathbb{E}\Big\{g(X)\Big\} + b\mathbb{E}\Big\{h(X)\Big\}.$$

In particular,

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

ii) Let X and Y be ANY random variables (discrete, continuous, independent, or non-independent). Then

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

More generally, for ANY random variables $X_1, \ldots, X_n$,

$$\mathbb{E}(X_1 + \ldots + X_n) = \mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n).$$

iii) Let X and Y be independent random variables, and g,h be functions. Then

$$\begin{aligned} \mathbb{E}(XY) &= \mathbb{E}(X)\mathbb{E}(Y) \\ \mathbb{E}\Big(g(X)h(Y)\Big) &= \mathbb{E}\Big(g(X)\Big)\mathbb{E}\Big(h(Y)\Big). \end{aligned}$$

VARIANCE:

The variance of a random variable X is a measure of how spread out it is. Are the values of X clustered tightly around their mean, or can we commonly observe values of X a long way from the mean value? The variance measures how far the values of X are from their mean, on average.

Definition: Let X be any random variable. The variance of X is

$$\mathbf{Var}(X) = \mathbb{E}\Big((X - \mu_X)^2\Big) = \mathbb{E}(X^2) - \Big(\mathbb{E}(X)\Big)^2.$$

The variance is the mean squared deviation of a random variable from its own mean.

If X has high variance, we can observe values of X a long way from the mean.

If X has low variance, the values of X tend to be clustered tightly around the mean value.

Properties of Variance

i) Let g be a function, and let a and b be constants. For any random variable X (discrete or continuous),

$$\mathbf{Var}\Big\{ag(X) + b\Big\} = a^2\,\mathbf{Var}\Big\{g(X)\Big\}.$$

$$\text{In particular, } \mathbf{Var}(aX + b) = a^2\mathbf{Var}(X).$$

ii) Let X and Y be independent random variables. Then

$$\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y).$$

iii) If X and Y are NOT independent, then

$$\mathbf{Var}(X + Y) = \mathbf{Var}(X) + \mathbf{Var}(Y) + 2\mathbf{cov}(X, Y).$$

CONDITIONAL EXPECTATION AND CONDITIONAL VARIANCE:

Suppose that X and Y are discrete random variables, possibly dependent on each other. Suppose that we fix Y at the value y. This gives us a set of conditional probabilities P(X = x|Y = y) for all possible values x of X. This is called the conditional distribution of X , given that Y = y .

Definition: Let X and Y be discrete random variables. The conditional probability function of X, given that Y = y, is:

$$\mathbb{P}(X = x \,|\, Y = y) = \frac{\mathbb{P}(X = x \;\mathbf{AND}\; Y = y)}{\mathbb{P}(Y = y)}.$$

We write the conditional probability function as:

$$f_{X|Y}(x \,|\, y) = \mathbb{P}(X = x \,|\, Y = y).$$

CONDITIONAL EXPECTATION AS A RANDOM VARIABLE

The unconditional expectation of X, E(X), is just a number: e.g. EX = 2 or EX = 5.8 .

The conditional expectation, E(X |Y = y), is a number depending on y .

If Y has an influence on the value of X, then Y will have an influence on the average value of X. So, for example, we would expect E(X |Y = 2) to be diff erent from E(X |Y = 3).

We can therefore view E(X |Y = y) as a function of y , say E(X |Y=y) = h(y) .

To evaluate this function, h(y) = E(X |Y = y), we:

i) fix Y at the chosen valuey ;

ii) find the expectation of X when Y is fixed at this value.

However, we could also evaluate the function at a random value of Y :

i) observe a random value of Y;

ii) fixYat that observed random value;

iii) evaluate E(X |Y = observed random value) .

We obtain a random variable: E(X |Y) = h(Y) .

The randomness comes from the randomness in Y, not in X

Conditional expectation:

E(X |Y) , is a random variable with randomness inherited from Y , not X .

1.6 APPLICATIONS OF THE UNIVARIATE AND MULTIVARIATE CENTRAL LIMIT THEOREM

Multivariate Random Variables

We now consider the extension to multiple r.v., i.e.,

$$X = (X_1, X_2, .., X_k) \in \mathbb{R}^k$$

The joint pmf, $f_X(x)$, is a function with

$$P(X \in A) = \sum_{x \in A} f_X(x)$$

The joint pdf, $f_X(x)$, is a function with

$$P(X \in A) = \int_{x \in A} f_X(x)dx$$

This is a multivariate integral, and in general difficult to compute. If A is a rectangle $A = [a_1, b_1] \times ... \times [a_k, b_k]$, then

$$\int_{x \in A} f_X(x)dx = \int_{a_k}^{b_k} ... \int_{a_1}^{b_1} f_X(x)dx_1..dx_k$$

The joint c.d.f. is defined similarly

$$F_X(x) = \sum_{z_1 \leq x_1, .., z_k \leq x_k} f_X(z_1, z_2, ..., z_k)$$

$$F_X(x) = P(X_1 \leq x_1, ..., X_k \leq x_k) = \int_{-\infty}^{x_1} ... \int_{-\infty}^{x_n} f_X(z_1, z_2, ..., z_k)dz_1..dz_k$$

The multivariate c.d.f. has similar coordinate-wise properties to a univariate c.d.f.

For continuously differentiable c.d.f.'s

$$f_X(x) = \frac{\partial^k F_X(x)}{\partial x_1 \partial x_2..\partial x_k}$$

CENTRAL LIMIT THEOREM (CLT):

Given a dataset with unknown distribution (it could be uniform, binomial or completely random), the sample means will approximate the normal distribution.

- 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
- If we increase the samples drawn from the population, the standard deviation of sample means will decrease. This helps us estimate the population mean much more accurately

- Also, the sample mean can be used to create the range of values known as a confidence interval (that is likely to consist of the population mean)

1.7 PROBABILISTIC INEQUALITIES

Say we have a random variable X. 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:

$$\mathbf{Pr}[X \geq k\mathbf{E}[X]] \leq 1/k.$$

(No need for k to be integer.) Equivalently, we can write this as:

$$\mathbf{Pr}[X \geq t] \leq \mathbf{E}[X]/t.$$

Proof. $\mathbf{E}[X] \geq \mathbf{Pr}[X \geq t] \cdot t + \mathbf{Pr}[X < t] \cdot 0 = t \cdot \mathbf{Pr}[X \geq t]$.

Defn of Variance: $\mathbf{var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2]$. Standard deviation is square root of variance. Can multiply out variance definition to get:

$$\mathbf{var}[X] = \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2.$$

UNIT –II
RANDOM SAMPLES

2.1 RANDOM SAMPLES

Random sampling is a way of selecting a sample of observations from a population in order to make inferences about the population. For example, exit polls from voters that aim to predict the likely results of an election. Random sampling is also known as probability sampling. 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.

The selection of observations must occur in a 'random' way, meaning that they do not differ in any significant way from observations not sampled. It is typically assumed that statistical tests contain data that has been obtained through random sampling.

2.1.1 SIMPLE RANDOM SAMPLING

Simple random sampling is the most straightforward approach for getting a random sample. It involves picking a desired sample size and selecting observations from a population in such a way that each observation has an equal chance of selection until the desired sample size is achieved

2.1.2 STRATIFIED RANDOM SAMPLING

Stratified random sampling involves breaking a population into key subgroups and obtaining a simple random sample from each group. These subgroups are called strata.

For example, if you want a sample size of 200, then you can pick samples of 50 from each strata. The required sample size for each stratum will be designed either to match known population proportions, or to over represent key subgroups of interest.

Note: The main benefit of stratified sampling over simple random sampling is making sure that you have good sample sizes in key subgroups.

2.1.2 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.

2.1.4 MULTISTAGE SAMPLING

Multistage sampling is where a combination of the above techniques is used. A combination of different sampling techniques

e.g., stratification with cluster sampling within each strata.

2.2 SAMPLING DISTRIBUTIONS OF ESTIMATORS

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.

The sampling distribution of an estimator is the probability distribution of all possible values the statistic may assume when a random sample of size n is taken. An estimator is a random variable since samples vary.

Sampling error = $\hat{\Theta}-\Theta$

Estimator – a statistic derived from a sample to infer the value of a population parameter.

Estimate – the value of the estimator in a particular sample. Population parameters are represented by Greek letters and the corresponding statistic by Roman letters.

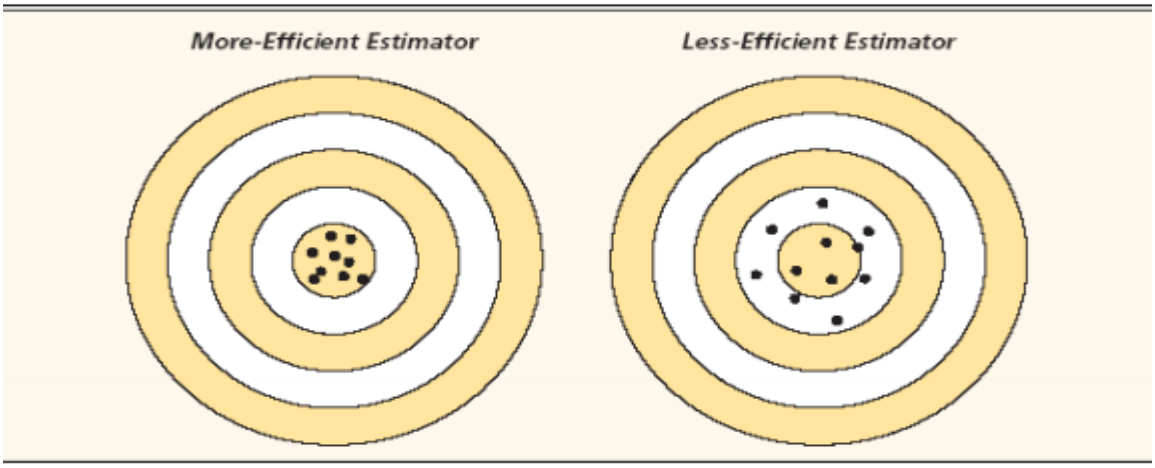
Bias: Bias is the difference between the expected value of the estimator and the true parameter

Bias= $E(\hat{\theta}) - \theta$

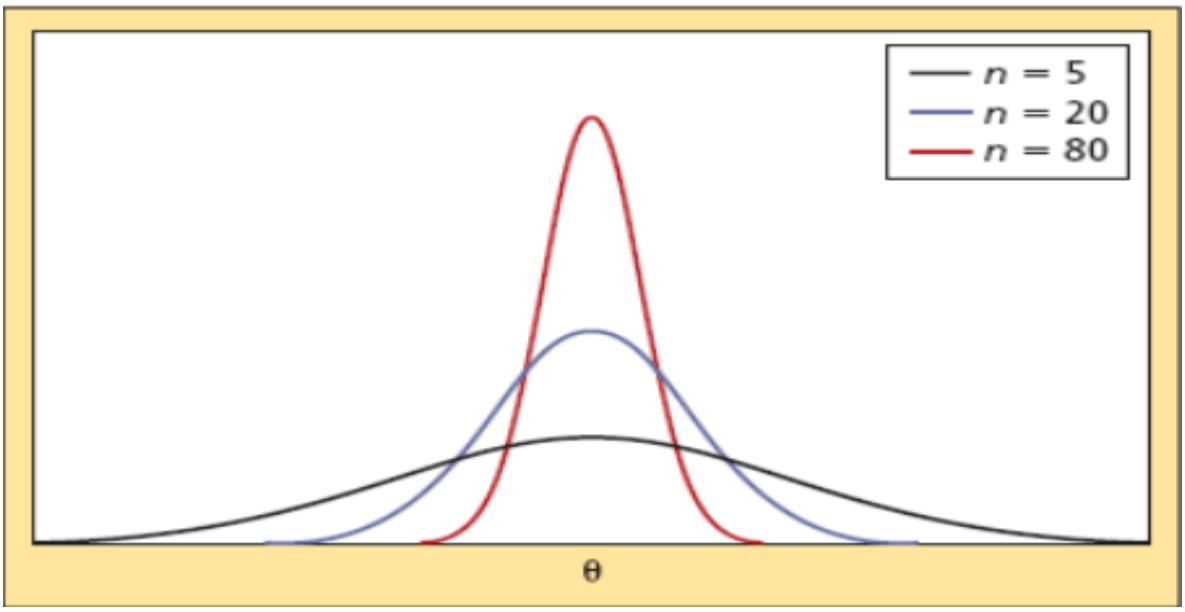
An estimator is unbiased if $E(\hat{\theta}) = \theta$

On average, an unbiased estimator neither overstates nor understates the true parameter

Efficiency: Efficiency refers to the variance of the estimator's sampling distribution. A more efficient estimator has smaller variance



Consistency A consistent estimator converges toward the parameter being estimated as the sample size increases.



The sampling distribution of estimator depends on the sample size. The effect of change of the sample size has to be determined. An estimate has a single numerical value and hence they are called point estimates. There are various estimators like sample mean, sample standard deviation, proportion, variance, range etc.

Sampling distribution of the mean: It is the population mean from which the samples are drawn. For all the sample sizes, it is likely to be normal if the population distribution is normal. The population mean is equal to the mean of the sampling distribution of the mean. Sampling distribution of mean has the standard deviation as follows:

$$\sigma_M = \frac{\sigma}{\sqrt{n}}$$

Where σ_M , is the standard deviation of the sampling mean, σ is the population standard-deviation

As the size of the sample increases, the spread of the sampling distribution of the mean decreases. But the mean of the distribution remains the same and it is not affected by the sample size.

The sampling distribution of the standard deviation is the standard error of the standard deviation. It is defined as:

$$\sigma_S = \frac{\sigma}{\sqrt{2n}}$$

Here, σ_S is the sampling distribution of the standard deviation. It is positively skewed for small *n* but it approximately becomes normal for sample sizes greater than 30.

2.3 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**

The Method of Moments (MoM) consists of equating sample moments and population moments. If a population has t parameters, the MOM consists of solving the system of equations

$$m_k' = \mu_k', \quad k = 1, 2, \ldots, t$$

for the t parameters.

in the method of moment, we estimate the parameter(s) by solving the equation(s):

$$E(X^k) = \mu_k = m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k - \mu_k = 0$$

While in the method of maximum likelihood, we want to maximize the log-likelihood function with respect to the unknown parameter

$$l(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

and this is equivalent to solving the following equation

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta) = 0$$

We can generalize the method of moment in the following way: let h(x) be a real-valued function such that E[h(X)] exists. We define

$$\mu_h = E[h(X)] = \int h(x) f(x|\theta) dx \quad \text{and} \quad m_h = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

Usually, μh is a function of the unknown parameter θ. Solving the equation

$$\mu_h = m_h$$

will give us an estimate of θ. Please note that if we let h(x) = x k − μk we will have the original method of moment because in this case μh = E[h(X)] = E(Xk −μk) = μk −μk = 0; mh = 1 n Pn i=1 h(Xi) = 1 n Pn i=1(Xk i − μk) = 1 n Pn i=1 Xk i − μk. Equating mh to μh, we will have 1 n Pn i=1 Xk i − μk = 0, which is the method of moment equation (1).

Next we will prove that if we use h(x) = ∂ ∂θ log f(x|θ), we will have the method of maximum likelihood. In this case, we have

$$\begin{aligned}\mu_h &= E\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = \int\left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)dx \\ &= \int \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)} f(x|\theta)dx = \int \frac{\partial}{\partial\theta} f(x|\theta)dx \\ &= \frac{\partial}{\partial\theta}\int f(x|\theta)dx = \frac{\partial 1}{\partial\theta} = 0\end{aligned}$$

Here we assume the order of integral and differential can be changed, this is justified under some reasonable conditions. We also have

$$m_h = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log f(X_i|\theta)$$

Equating mh to μh, we will have

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log f(X_i|\theta) = 0$$

which is exactly the equation for maximum likelihood estimation Eqn. 2. In general, if h(x, θ) is selected such that E(h(X, θ)) = 0, then h(X, θ) is called score function. Under this situation, the common framework for estimation equation is

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i, \theta) = 0$$

In method of moments, the score function is h(X, θ) = Xk − μk(θ), and in maximum likelihood method, the score function is h(X, θ) = l 0 (X|θ) = ∂ ∂θ log f(X|θ).

UNIT –III

STATISTICAL INTERFACE

3.1 STATISTICAL INFERENCE

Statistical Interfernce 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. It is assumed that the observed data set is sampled from a larger population.

Inferential statistics can be contrasted with descriptive statistics. Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.

Statistical inference makes propositions about a population, using data drawn from the population with some form of sampling. Given a hypothesis about a population, for which we wish to draw inferences, statistical inference consists of (first) selecting a statistical model of the process that generates the data and (second) deducing propositions from the model[1]

Konishi & Kitagawa state, "The majority of the problems in statistical inference can be considered to be problems related to statistical modeling". Relatedly, Sir David Cox has said, "How [the] translation from subject-matter problem to statistical model is done is often the most critical part of an analysis".

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

Statisticians distinguish between three levels of modeling 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. For example, one may assume that the distribution of population values is truly Normal, with unknown mean and variance, and that datasets are generated by 'simple' random sampling. The family of generalized linear models is a widely used and flexible class of parametric model

- **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 Hodges–Lehmann–Sen estimator, 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. Furthermore, one may assume that the mean response level in the population depends in a truly linear manner on some covariate (a parametric assumption) but not make any parametric assumption describing the variance around that mean (i.e. about the presence or possible form of any heteroscedasticity). More generally, semi-parametric models can often be separated into 'structural' and 'random variation' components. One component is treated parametrically and the other non-parametrically. The well-known Cox model is a set of semi-parametric assumptions.

3.2 INTRODUCTION TO MULTIVARIATE STATISTICAL MODELS: REGRESSION AND CLASSIFICATION PROBLEMS

MULTIVARIATE ANALYSIS

Introduction to Multivariate Analysis:

Multivariate analysis is an analysis of statistical technique that analyse the relationship between more than two variables which shows the effect of more than one variable on one variable that is independent variable. Multivariate analysis helps the organisation 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 analysing the data. It is analytical techniques are applicable in the field of industries, Government sector, Universities and Research centres.

The main benefit of multivariate techniques is to analyse the market scenario which explain what customer's habit, preferences, choices, target market campaigns, new product or serving Design and Refinement existing product. Many years ago, it was very difficult task for analyst and researcher for collecting and analysing the data but nowadays it becomes easy because of availability of computer software, statistical software packages and internet.And it reduces the burden and the task of a common man

Basic Concepts of Multivariate Analysis:

The Variate: A variate is a building block of multivariate analysis. It is a linear combination of variables with empirically determined weights. The variables are specified by the researcher, whereas the weights are determined by the multivariate technique relevant to a specific objective. A variate of n

weighted variables (X1 to Xn Variate value = w)

can be defined mathematically as 1X1+w2X2+…….+wnXn

Where Xn is the observed variable and wn.

Measurement scales: Measurement scale is a technique to measure the variable or to identify the data in selection of the appropriate multivariate method of analysis. Data can be classified into generally two categories.

Non Metric Measurement Scales: It generally measures quality of data which is qualitative in nature. The Non Metric data are data that are not measurable or quantifiable. This Non metric measurement can be made with either a Nominal or an Ordinal scale.

Nominal Scales: Nominal information represents classes or categories associate degree doesn't imply amounts of an attribute or characteristic. Example of nominally scaled information is individual's gender, occupation.

Ordinal Scales: In ordinal scale variables will be ordered or stratified in respect to the number of the attribute possessed. The numbers employed in ordinal scales square measure extremely non quantitative as a result of they indicate solely relative positions in associate degree graduated table. No live of the particular quantity or magnitude is provided by ordinal scales, it denotes solely the order of the values. The analyst cannot perform any arithmetic operations. The analyst should determine all non metric information to confirm that they're used fitly within the variable techniques.

Metric measure Scales: Metric information square measure used once subjects disagree in quantity

or degree on a specific attribute. Metrically measured variables replicate quantity or degree and attributes involving quantity of magnitude. The 2 totally different metric measure scales square measure interval and magnitude relation scales.

Interval Scales: Interval scale provides the best level of measure exactitude, and any mathematical process to be performed. This scale has constant measure unit. Interval scale use associate degree discretionary zero. the foremost acquainted interval scales square measure the physicist and astronomer temperature scales. 3 Ratio Scales: magnitude relation scales represent the best variety of measure exactitude as a result of they possess the benefits of all lower scales and associate degree temperature purpose. All mathematical operations square measure permissible with magnitude relation scale measurements.

Brief Description of Each Multivariate Technique:

Multivariate analysis is an ever expanding set of techniques for data analysis which provides a wide range of possible research situations. The more established and emerging techniques discussed are as follows:

1) Principal Components and Common Factor Analysis

2) Multiple Regression and Multiple Correlation

3) Multiple Discriminant Analysis

4) Multivariate Analysis of Variance and Covariance

5) Conjoint Analysis

6) Canonical Correlation

7) Cluster Analysis

8) Multidimensional Scaling

9) Correspondence Analysis

10) Linear Probability Models such as Logit and Probit

11) Simultaneous/Structural Equation Modelling

Each of the multivariate techniques are briefly described below.

Principal Components and Common Factor Analysis:

Factor analysis together with each principal element and customary factor analysis, could be a statistical approach that may be accustomed to analyse lay relationships among sizable amount of variables and to elucidate these variables in terms of their common underlying factors. Factor analysis could be a variable knowledge reduction technique. All the variables beneath investigation are analysed along to extract the underlying factors. Factor analysis helps in distinguishing underlying structure of the information. Factor analysis makes use of metric knowledge. An element could be a linear combination of variables.

Multiple Regression:

Multiple Regression is that the appropriate technique of research once the matter involves one metric variable and a minimum of two independent variables. The objective of this technique is to predict the changes within the explanatory variables. The estimation of the regression modelling is dispensed by the standard least square technique. The standard least square technique aims at minimizing the error total of the squares whereas estimating the regression model.

Multiple Discriminant Analysis:

Multiple discriminant analysis is that the acceptable technique where the only variable is categorical and explanatory variables is continuous. The essential principle underlying a discriminant model is to settle on linear mixtures of the predictor variables which will maximise between-group variance to withingroup variance. It is applicable in state of affairs within which the whole sample will be divided into teams supported a non metric variable characterising many noted categories. Multiple discriminant analysis will be wont to profile the repose cluster characteristics of the topic and assign them to their acceptable teams.

Multivariate Analysis of Variance and variance (MANOVA):

MANOVA can be defined as a statistical technique that measures the distinction for two or more metric variables supported by a group of non metric variable acting as explanatory variables. MANOVA will be utilized in conjunction with MANCOVA to get rid of the impact of any uncontrolled metric variable on the dependent variables. MANOVA is that the applications of regression procedures among the analysis of variance methodology.

Conjoint Analysis:

Conjoint analysis is best fitted to understanding consumer's reactions and evaluations of pre-set

attribute mixtures that represent potential product or services. for instance, assume a product thought has three attributes (worth, Quality and Colour) every at three possible levels (Blue, Pink and Green) rather than having to gauge all twenty seven doable mixtures, a set (nine or more) will be evaluated for that attractiveness to shoppers and also the researchers will perceive the importance of every level. The results of conjoint analysis also can be utilized in product style simulators which show client acceptance for any range of product formulations and aid within the style of the best product.

Canonical Correlation:

Canonical correlation analysis could be a logical extension of multiple correlation analysis. It involves single metric variable and several other metric independent variables. The target is to develop a linear combination of every set of variables to maximise the correlation between the two sets. The procedure involves getting a group of weights for the dependent and explanatory variables that offer the most straightforward correlation between the set of explanatory variables.

Cluster Analysis: Cluster analysis is an analytical technique for developing meaning subgroups of people or objects. The objective of this technique is to classify a sample of entities into a small number of disjoint groups supported by the similarities among the entities.

This technique is employed in management and promoting field. It is used to section and cluster the client into clearly homogenized teams, which needs specific methods so as to focus on them. The segmentations are often extended to industries, totally different sectors and markets. This technique is effectively wont to cluster individuals into clusters and so devise associate overall career growth arrange within the space of human resources.

Multidimensional Scaling:

The objective of this technique is to remodel consumer's judgement of similarity or preferences into distances presented in Multidimensional space. This system is wide applicable within the space of promoting. It is helpful to review subjective complete perceptions, impact of positioning and advertising methods on complete image. Owing to this technique we are able to notice opportunities for brand spanking new product. Multidimensional scaling is merely one amongst the wide clad of statistical techniques obtainable for getting the item map. This technique classified along is termed as perceptual mapping technique.

Correspondence Analysis:

Correspondence analysis is recently developed interdependence technique that facilitates each dimensional reduction of object ratings (Products and Persons) on a collection of attributes and

therefore the sensory activity mapping of objects relative to those three attributes. Researchers are constantly required to "quantify the qualitative data" found in nominal variables.

Correspondence analysis differs from the mutuality techniques in its ability to accommodate each non metric information and nonlinear relationships. Correspondence analysis employs a contingency table, that is that the cross tabulation of two categorical variables. It then transforms the non metric information to a metric level and performs dimensional reduction and perceptual mapping. This technique provides a multivariate representation of interdependence for non metric information that is not possible with other methods.

Linear probability Models:

This technique is that the combination of multiple regression and multiple discriminant analysis. This system is comparable to multiple correlation analysis within the sense of one or more independent variables are used to predict one dependent variable. Linear probability models are totally different from analysis within the manner that they accommodate every kind of independent variables and don't need the idea of variable normality.

Structural Equation Modelling:

Structural equation modelling is known by covariance structural analysis, latent variable analysis and typically as LISREL (Specialized code Package). Structural equation modelling may be a technique that enables separate relationships for every of dependent variables. Structural equation modelling provides the acceptable and therefore the most effective estimation technique for a series of separate multiple correlation equations calculable at the same time. It is characterized by two basic elements (i) the structural model (ii) the measure model.

The structural model is that the "Path" model, which relates independent to dependent variables. In such state of affairs, theory, previous expertise, or alternative pointers modify the man of science to tell apart that independent variables predict every variable.

The measurement model permits there searcher to use many variables for a single independent or dependent variable

Multivariate Regression

It is a method used to measure the degree at which more than one independent variable (**predictors**) and more than one dependent variable (**responses**), are linearly related. The method is broadly used to

predict the behavior of the response variables associated to changes in the predictor variables, once a desired degree of relation has been established.

This is quite similar to the simple linear regression model we have discussed previously, but with multiple independent variables contributing to the dependent variable and hence multiple coefficients to determine and complex computation due to the added variables. Jumping straight into the equation of multivariate linear regression,

$$Yi=\alpha+\beta1xi(1)+\beta2xi(2)+....+\beta nxi(n)$$

Yi is the estimate of ith component of dependent variable y, where we have **n** independent variables and xij denotes the ith component of the jth independent variable/feature. Similarly cost function is as follows,

$$E(\alpha,\beta1,\beta2,...,\beta n)=12m\sum i=1m(yi-Yi)$$

where we have **m** data points in training data and y is the observed data of dependent variable.
As per the formulation of the equation or the cost function, it is pretty straight forward generalization of simple linear regression. But computing the parameters is the matter of interest here.

Computing parameters

Generally, when it comes to multivariate linear regression, we don't throw in all the independent variables at a time and start minimizing the error function. First one should focus on selecting the best possible independent variables that contribute well to the dependent variable. For this, we go on and construct a correlation matrix for all the independent variables and the dependent variable from the observed data. The correlation value gives us an idea about which variable is significant and by what factor. From this matrix we pick independent variables in decreasing order of correlation value and run the regression model to estimate the coefficients by minimizing the error function. We stop when there is no prominent improvement in the estimation function by inclusion of the next independent feature. This method can still get complicated when there are large no.of independent features that have significant contribution in deciding our dependent variable. Let's discuss the normal method first which is similar to the one we used in univariate linear regression.

Normal Equation

Now let us talk in terms of matrices as it is easier that way. As discussed before, if we have n independent variables in our training data, our matrix X has n+1 rows, where the first row is the 0th term added to each vector of independent variables which has a value of 1 (this is the coefficient of the constant term α). So, X is as follows,

X=[X1..Xm]

Xi contains n entries corresponding to each feature in training data of ith entry. So,
matrix **X** has m rows and n+1 columns (0thcolumn is all 1s and rest for one independent variable each).

Y=[Y1Y2..Ym]

and coefficient matrix C,

C=[αβ1..βn]

and our final equation for our hypothesis is,

Y=XC

To calculate the coefficients, we need **n+1** equations and we get them from the minimizing condition of the error function. Equating partial derivative of E(α,β1,β2,...,βn) with each of the coefficients to 0 gives a system of n+1 equations. Solving these is a complicated step and gives the following nice result for matrix C,

C=(XTX)−1XTy

where y is the matrix of the observed values of dependent variable.
This method seems to work well when the n value is considerably small (approximately for 3-digit values of n). As n grows big the above computation of matrix inverse and multiplication take large amount of time. In future tutorials lets discuss a different method that can be used for data with large no. of features.

3.3 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 × k dimensional matrix **W.**
- Use this d × k eigenvector matrix to transform the samples onto the new subspace.

So, let's unfurl the maths behind each of this one by one.

1. **Take the whole dataset consisting of d+1 dimensions and ignore the labels such that our new dataset becomes d dimensional.**

Let's say we have a dataset which is d+1 dimensional. Where d could be thought as X_train and 1 could be thought as y_train (labels) in modern machine learning paradigm. So, X_train + y_train makes up our complete train dataset.

So, after we drop the labels we are left with d dimensional dataset and this would be the dataset we will use to find the principal components. Also, let's assume we are left with a three-dimensional dataset after ignoring the labels i.e d = 3.

we will assume that the samples stem from two different classes, where one-half samples of our dataset are labeled class 1 and the other half class 2.

Let our data matrix **X** be the score of three students :

Student	Math	English	Art
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

2. Compute the mean of every dimension of the whole dataset.

The data from the above table can be represented in matrix **A,** where each column in the matrix shows scores on a test and each row shows the score of a student.

$$\mathbf{A} = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

Matrix A

So, The mean of matrix **A** would be

$$\overline{\mathbf{A}} = [\,66 \quad 60 \quad 60\,]$$

Mean of Matrix A

3. Compute the covariance matrix of the whole dataset (sometimes also called as the variance-covariance matrix)

So, we can compute the covariance of two variables **X** and **Y** using the following formula

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n} (Xi - \bar{x})(Yi - \bar{y})$$

Using the above formula, we can find the covariance matrix of **A.** Also, the result would be a square matrix of d ×d dimensions.

Let’s rewrite our original matrix like this

	Math	English	Arts
1	90	60	90
2	90	90	30
3	60	60	60
4	60	60	90
5	30	30	30

Matrix A

Its covariance matrix would be

	Math	English	Art
Math	504	360	180
English	360	360	0
Art	180	0	720

Covariance Matrix of A

Few points that can be noted here is :

- Shown in Blue along the diagonal, we see the variance of scores for each test. The art test has the biggest variance (720); and the English test, the smallest (360). So we can say that art test scores have more variability than English test scores.
- The covariance is displayed in black in the off-diagonal elements of the matrix **A**

a) The covariance between math and English is positive (360), and the covariance between math and art is positive (180). This means the scores tend to covary in a positive way. As scores on math go up, scores on art and English also tend to go up; and vice versa.

b) The covariance between English and art, however, is zero. This means there tends to be no predictable relationship between the movement of English and art scores.

4. Compute Eigenvectors and corresponding Eigenvalues

Intuitively, an eigenvector is a vector whose direction remains unchanged when a linear transformation is applied to it.

Now, we can easily compute eigenvalue and eigenvectors from the covariance matrix that we have above.

Let **A** be a square matrix, **v** a vector and **λ** a scalar that satisfies **A**v = **λ**v, then **λ** is called eigenvalue associated with eigenvector **v** of **A**.

The eigenvalues of **A** are roots of the characteristic equation

$$\det(A-\lambda I) = 0$$

Calculating det(A-λI) first, I is an identity matrix :

$$\det\left(\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \lambda\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right)$$

Simplifying the matrix first, we can calculate the determinant later,

$$\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

$$\begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix}$$

Now that we have our simplified matrix, we can find the determinant of the same :

$$\det\begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix}$$

$$-\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800$$

We now have the equation and we need to solve for λ, so as to get the eigenvalue of the matrix. So, equating the above equation to zero :

$$-\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800 = 0$$

After solving this equation for the value of **λ,** we get the following value

$$\lambda \approx 44.81966\ldots, \lambda \approx 629.11039\ldots, \lambda \approx 910.06995\ldots$$

Eigen values

Now, we can calculate the eigenvectors corresponding to the above eigenvalues. I would not show how to calculate eigenvector here, visit this link to understand how to calculate eigenvectors.

So, after solving for eigenvectors we would get the following solution for the corresponding eigenvalues

$$\begin{pmatrix} -3.75100\ldots \\ 4.28441\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} -0.50494\ldots \\ -0.67548\ldots \\ 1 \end{pmatrix}, \begin{pmatrix} 1.05594\ldots \\ 0.69108\ldots \\ 1 \end{pmatrix}$$

5. Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a d × k dimensional matrix W.

We started with the goal to reduce the dimensionality of our feature space, i.e., projecting the feature space via PCA onto a smaller subspace, where the eigenvectors will form the axes of this new feature subspace. However, the eigenvectors only define the directions of the new axis, since they have all the same unit length 1.

So, in order to decide which eigenvector(s) we want to drop for our lower-dimensional subspace, we have to take a look at the corresponding eigenvalues of the eigenvectors. Roughly speaking, the eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data, and those are the ones we want to drop.

The common approach is to rank the eigenvectors from highest to lowest corresponding eigenvalue and choose the top k eigenvectors.

So, after sorting the eigenvalues in decreasing order, we have

$$\begin{pmatrix} 910.06995 \\ 629.11039 \\ 44.81966 \end{pmatrix}$$

For our simple example, where we are reducing a 3-dimensional feature space to a 2-dimensional feature subspace, we are combining the two eigenvectors with the highest eigenvalues to construct our d×k dimensional eigenvector matrix **W.**

So, eigenvectors corresponding to two maximum eigenvalues are :

$$\mathbf{W} = \begin{bmatrix} 1.05594 & -0.50494 \\ 0.69108 & -0.67548 \\ 1 & 1 \end{bmatrix}$$

6. Transform the samples onto the new subspace

In the last step, we use the 2×3 dimensional matrix **W** that we just computed to transform our samples onto the new subspace via the equation $\mathbf{y = W' \times x}$ where $\mathbf{W'}$ is the transpose of the matrix **W.**

3.4 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. In this post, I explain how over fitting models is a problem and how you can identify and avoid it.

Over fit regression models have too many terms for the number of observations. When this occurs, the regression coefficients represent the noise rather than the genuine relationships in the population.

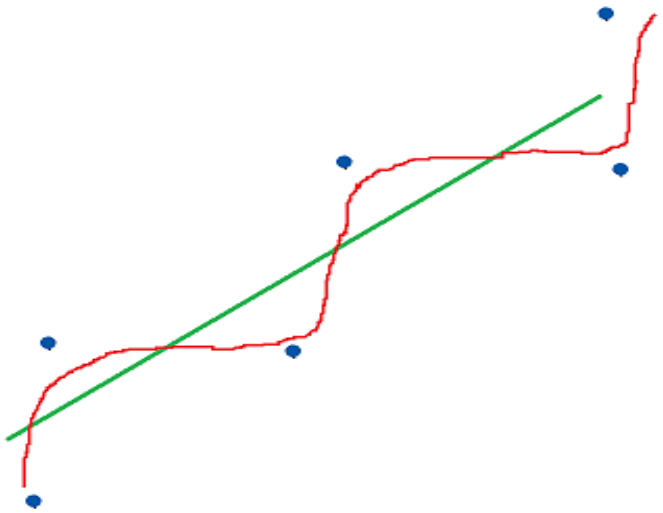
That's problematic by itself. However, there is another problem. Each sample has its own unique quirks. Consequently, a regression model that becomes tailor-made to fit the random quirks of one sample is unlikely to fit the random quirks of another sample. Thus, over fitting a regression model reduces its generalizability outside the original dataset.

Taking the above in combination, an over fit regression model describes the noise, and it's not applicable outside the sample. That's not very helpful, right? I'd really like these problems to sink in because over fitting often occurs when analysts chase a high R-squared. In fact, inflated R-squared values are a symptom of over it models! Despite the misleading results, it can be difficult for analysts to give up that nice high R-squared value.

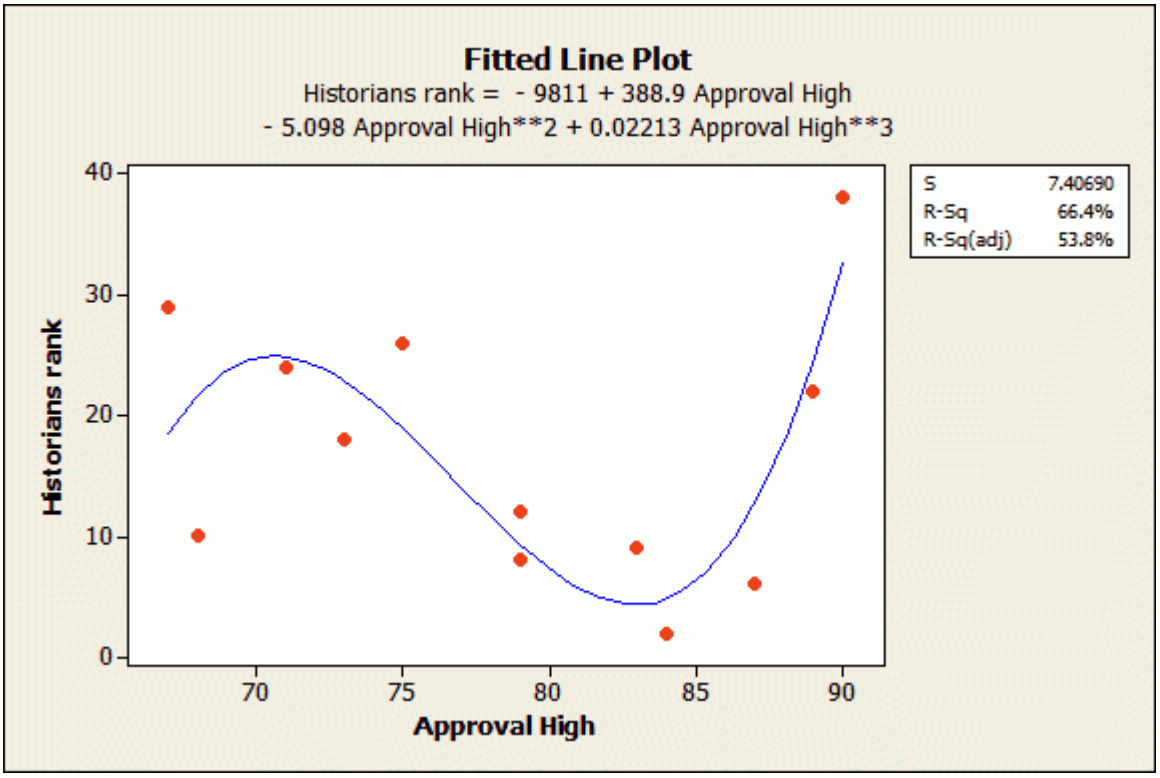
When choosing a regression model, our goal is to approximate the true model for the whole population. If we accomplish this goal, our model should fit most random samples drawn from that population. In other words, our results are more generalizable—we can expect that the model will fit other samples.

Graphical Illustration of Over fitting Regression Models

The image below illustrates an over fit model. The green line represents the true relationship between the variables. The random error inherent in the data causes the data points to fall randomly around the green fit line. The red line represents an over fit model. This model is too complex, and it attempts to explain the random error present in the data.



The example above is very clear. However, it's not always that obvious. Below, the fitted line plot shows an overfit model. In the graph, it appears that the model explains a good proportion of the dependent variable variance. Unfortunately, this is an overfit model, and I'll show you how to detect it shortly.



If you have more than two independent variables, it's not possible to graph them in this manner, which makes it harder to detect.

UNIT –IV

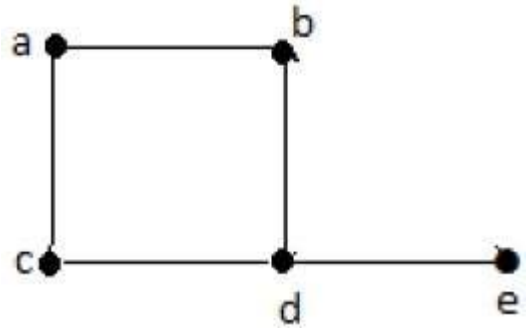
GRAPH THEORY

4.1 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 −



In the above graph,

V = {a, b, c, d, e}

E = {ab, ac, bd, cd, de}

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.

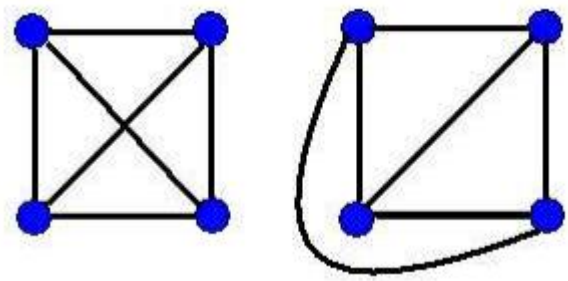
Science − The molecular structure and chemical structure of a substance, the DNA structure of an organism, etc., are represented by graphs.

Linguistics − The parsing tree of a language and grammar of a language uses graphs.

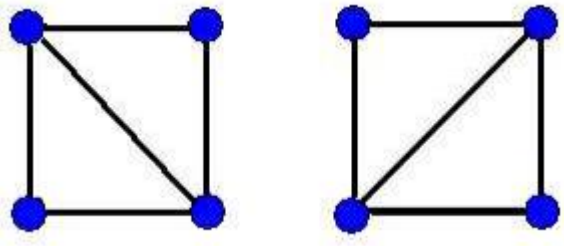
General − Routes between the cities can be represented using graphs. Depicting hierarchical ordered information such as family tree can be used as a special type of graph called tree.

4.1.1 Isomorphism

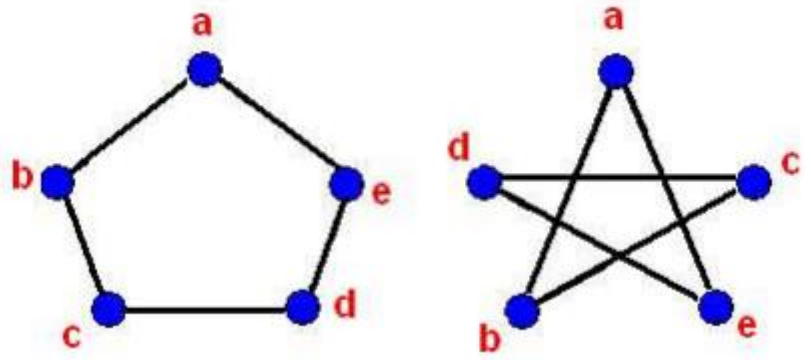
There are different ways to draw the same graph. Consiser the following two graphs



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



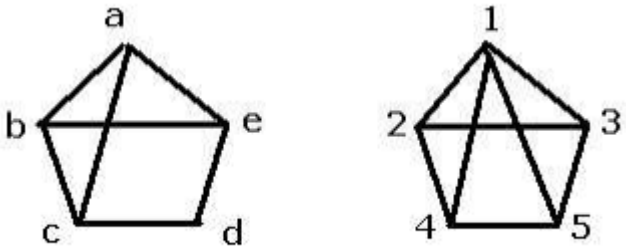
Another pair



We discuss two models - an adjacency list and an adjacency matrix. So, we can say that a notion about which pair of vertices are adjacent and which are not is the most important property. Base on this, we can say that two graphs above must be identical - check the pair of adjacent vertices

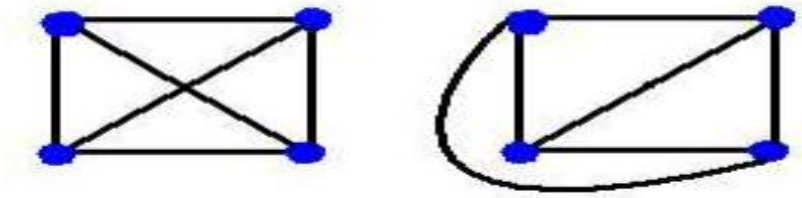
Definition.

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



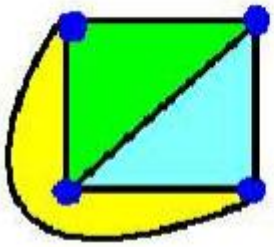
4.1.2 Planar graphs

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



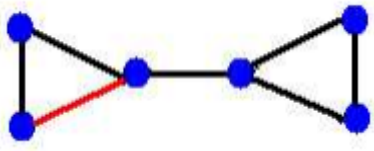
cross.

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.



Theorem (Euler's formula)

For any connected planar graph G V,



Choose an edge e connecting two different faces of G, and remove it. The graph remains connected. This removal decreases both the number of faces and edges by one. So, we can use inductive hypothesis. $V - E + F = V - (E - 1) + (F - 1) = 2.$ QED.

4.1.3 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."

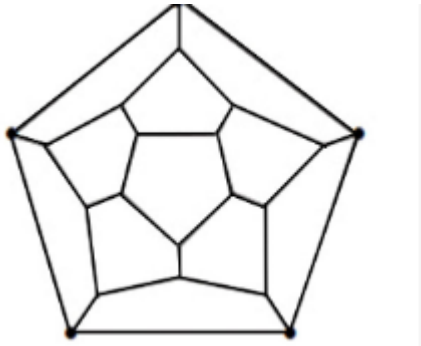
A simple solution to this problem is to color every vertex with a different color to get a total of colors. But in some cases, the actual number of colors required could be less than this.

4.1.4 chromatic number –"The least number of colors required to color a graph is called its **chromatic number**. It is denoted by $\chi(G)$.

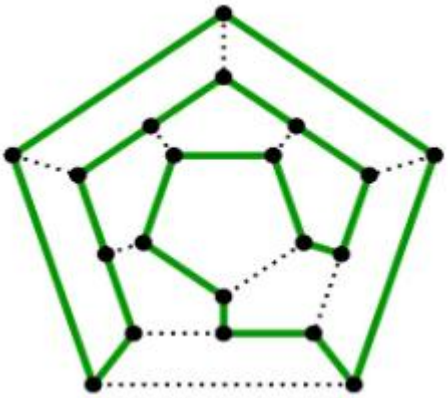
4.1.5 Hamilonian Circuit – A simple circuit in a graph that passes through every vertex exactly once is called a Hamiltonian circuit.

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.
There are certain theorems which give sufficient but not necessary conditions for the existence of Hamiltonian graphs.

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



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



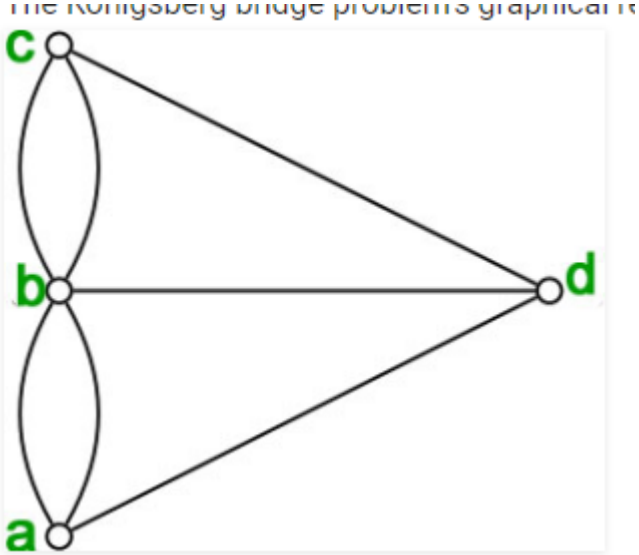
Euler path problem was first proposed in the 1700's.

4.1.6 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.

- An Euler circuit starts and ends at the same vertex.

The Konigsberg bridge problem's graphical representation



4.2 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.

4.2.1 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)

(In other words, there are **n** possibilities for the first choice, THEN there are **n** possibilites for the second choice, and so on, multplying each time.)

Which is easier to write down using an exponent of **r**:

$n \times n \times ...$ (r times) = n^r

Example: in the lock above, there are 10 numbers to choose from (0,1,2,3,4,5,6,7,8,9) and we choose 3 of them:

$10 \times 10 \times ...$ (3 times) = 10^3 = 1,000 permutations

So, the formula is simply:

N^r – where n is the number of things to choose from, and we choose r of them, repetition is allowed and order matters.

4.2.2. Permutations without Repetition

In this case, we have to **reduce** the number of available choices each time.

Example: what order could 16 pool balls be in?

After choosing, say, number "14" we can't choose it again.

So, our first choice has 16 possibilites, and our next choice has 15 possibilities, then 14, 13, 12, 11, ... etc. And the total permutations are:

$16 \times 15 \times 14 \times 13 \times ...$ = 20,922,789,888,000

But maybe we don't want to choose them all, **just 3** of them, and that is then:

$16 \times 15 \times 14$ = 3,360

In other words, there are 3,360 different ways that 3 pool balls could be arranged out of 16 balls.

Without repetition our choices get reduced each time.

Example Our "order of 3 out of 16 pool balls example" is:

$$\frac{16!}{(16-3)!} = \frac{16!}{13!} = \frac{20{,}922{,}789{,}888{,}000}{6{,}227{,}020{,}800} = 3{,}360$$

(which is just the same as: **16 × 15 × 14 = 3,360**)

Example: How many ways can first and second place be awarded to 10 people?

$$\frac{10!}{(10-2)!} = \frac{10!}{8!} = \frac{3{,}628{,}800}{40{,}320} = 90$$

(which is just the same as: **10 × 9 = 90**)

Notation

Instead of writing the whole formula, people use different notations such as these:

$$P(n,r) = {}^{n}P_r = {}_{n}P_r = \frac{n!}{(n-r)!}$$

Example: **P(10,2) = 90**

4.2.2 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)

4.2.1. Combinations with Repetition

Actually, these are the hardest to explain, so we will come back to this later.

4.2.2. Combinations without Repetition

This is how lotteries work. The numbers are drawn one at a time, and if we have the lucky numbers (no matter what order) we win!

The easiest way to explain it is to:

- assume that the order does matter (ie permutations),
- then alter it so the order does **not** matter.

Going back to our pool ball example, let's say we just want to know which 3 pool balls are chosen, not the order.

We already know that 3 out of 16 gave us 3,360 permutations.

But many of those are the same to us now, because we don't care what order!

For example, let us say balls 1, 2 and 3 are chosen. These are the possibilites:

Order does matter	Order doesn't matter
1 2 3 1 3 2 2 1 3 2 3 1 3 1 2 3 2 1	1 2 3

So, the permutations have 6 times as many possibilites.

In fact there is an easy way to work out how many ways "1 2 3" could be placed in order, and we have already talked about it. The answer is:

3! = 3 × 2 × 1 = 6

(Another example: 4 things can be placed in **4! = 4 × 3 × 2 × 1 = 24** different ways, try it for yourself!)

So we adjust our permutations formula to **reduce it** by how many ways the objects could be in order (because we aren't interested in their order any more):

$$\frac{n!}{(n-r)!} \times \frac{1}{r!} = \frac{n!}{r!(n-r)!}$$

That formula is so important it is often just written in big parentheses like this:

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

where **n** is the number of things to choose from,
and we choose **r** of them,
no repetition,
order doesn't matter.

It is often called "n choose r" (such as "16 choose 3")

And is also known as the Binomial Coefficient .

Notation

As well as the "big parentheses", people also use these notations:

$$C(n,r) = {}^{n}C_r = {}_{n}C_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Just remember the formula:

n!**r!(n − r)!**

Example: Pool Balls (without order)

So, our pool ball example (now without order) is:

16!**3!(16−3)!** = 16!**3!** × **13!**

= 20,922,789,888,000**6** × **6,227,020,800**

= 560

Or we could do it this way:

16×15×14**3×2×1** = 3360**6** = 560

It is interesting to also note how this formula is nice and **symmetrical**:

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

In other words choosing 3 balls out of 16, or choosing 13 balls out of 16 have the same number of combinations.

16!3!(16−3)! = **16!**13!(16−13)! = **16!**3! × 13! = **560**

4.3 Specialized techniques to solve combinatorial enumeration problems.

Combinatorics is the study of finite sets of objects defined by certain specified properties – combinatorial structures – such as:

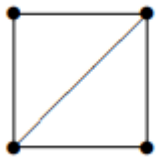
I Subsets of a finite set

∅, {1}, {2}, {3}, {1, 2}, {1, 3}, {2, 3}, {1, 2, 3}

I Partitions of a number 4 = 1 + 1 + 1 + 1 = 1 + 1 + 2 = 1 + 3 = 2 + 2

I Words over a finite alphabet aaa, aab, aba, abb, baa, bab, bba, bbb

Graphs:



Latin square

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

:

For any particular combinatorial structure, a number of (related) questions can be asked:

I Existence Are there any combinatorial structures of this type?

I Enumeration How many combinatorial structures of this type are there?

I Generation List all the combinatorial structures of this type.

Existence An existence question can be answered in different ways:

I Constructively, by giving I An explicit example, or I An algorithm to construct an example.

I Non-constructively, by giving an existence proof that does not yield an actual example. A constructive proof is regarded as being “better” than a non-constructive one.

Enumeration There are even more ways in which an enumeration question can be answered, including:

I Exactly A set of size n has n k subsets of size k.

I Approximately The number L(n) of Latin squares of order n satisfies log L(n) = n 2 log n + O(n 2).

I Implicitly The n’th Fibonacci number Fn is the coefficient of x n in the expansion of $\frac{1}{1-x-x^2}$ as a power series (about 0).

Bijectively The number of switching classes of graphs is equal to the number of Eulerian graphs.

I Computationally If all else fails, it may only be possible to produce a short list of the numbers of small combinatorial structures by direct computer generation. For example, the number of Steiner triple systems on n points for n = 1,. . ., 19 is 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 2, 0, 80, 0, 0, 0, 11084874829.

Generation Algorithms for the generation of combinatorial structures are often called combinatorial algorithms. Such an algorithm should generate every combinatorial structure of a particular type – for

example, we might want to generate all the subsets of a given set. Other considerations include:

I Efficiency The algorithm should be efficient in both space and time terms.

I Ordering The order in which the structures are output may be significant — a specific order may be required by the user, or a cleverly chosen order may reduce the amount of work required.

UNIT –V

COMPUTER SCIENCE AND ENGINEERING APPLICATIONS

5.1 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 is about extracting useful knowledge from large databases.

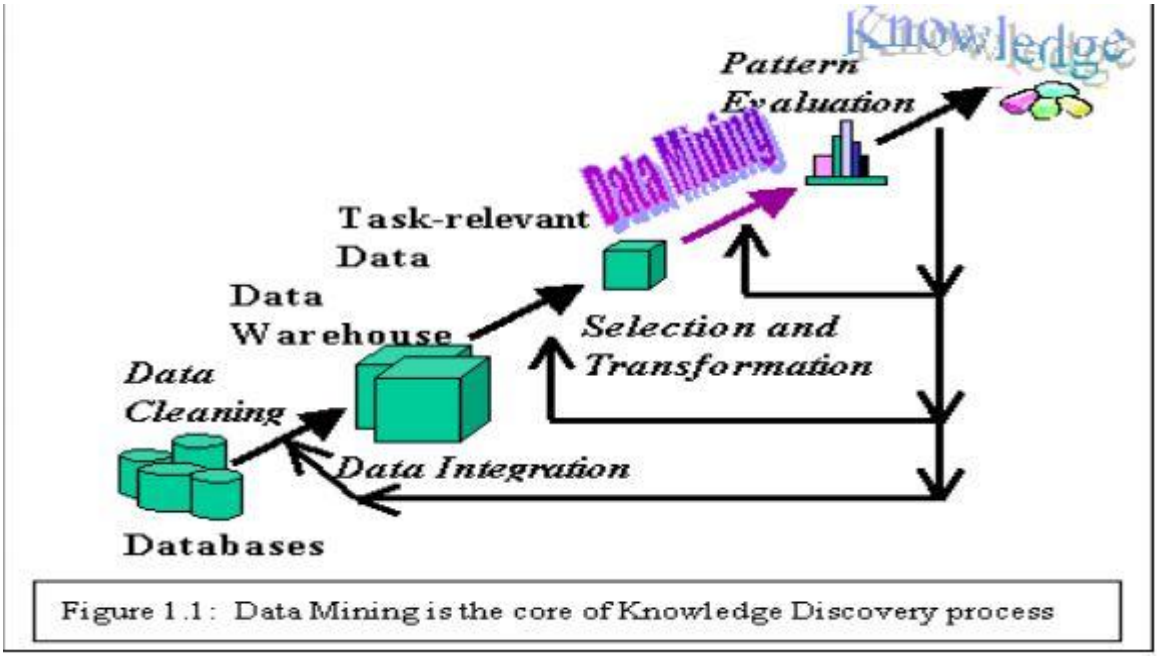


Figure 1.1: Data Mining is the core of Knowledge Discovery process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

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

5.2 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.

Network protocols incorporate all the processes, requirements and constraints of initiating and accomplishing communication between computers, servers, routers and other network-enabled devices. Network protocols must be confirmed and installed by the sender and receiver to ensure network/data communication and apply to software and hardware nodes that communicate on a network.

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.

5.3 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.

Web traffic is the amount of data sent and received by visitors to a website. This necessarily does not include the traffic generated by bots. Since the mid-1990s, web traffic has been the largest portion of Internet traffic.This is determined by the number of visitors and the number of pages they visit. Sites monitor the incoming and outgoing traffic to see which parts or pages of their site are popular and if there are any apparent trends, such as one specific page being viewed mostly by people in a particular country. There are many ways to monitor this traffic and the gathered data is used to help structure sites, highlight security problems or indicate a potential lack of bandwidth.

Not all web traffic is welcomed. Some companies offer advertising schemes that, in return for increased web traffic (visitors), pay for screen space on the site. There is also "fake traffic", which is bot traffic generated by a third party. This type of traffic can damage a website's reputation, its visibility on Google, and overall domain authority.

Web analytics is the measurement of the behavior of visitors to a website. In a commercial context, it especially refers to the measurement of which aspects of the website work towards the business objectives of Internet marketing initiatives; for example, which landing pages encourage people to make a purchase. Notable vendors of web analytics software and services include Google Analytics, IBM Digital Analytics (formerly Coremetrics) and Adobe Omniture.

5.4 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.

- Often people confuse computer security with other related terms like information security and cyber security. One way to ascertain the similarities and differences among these terms is by asking what is being secured. For example,
- Information security is securing information from unauthorized access, modification & deletion
- Computer Security means securing a standalone machine by keeping it updated and patched
- Cyber security is defined as protecting computer systems, which communicate over the computer networks

It’s important to understand the distinction between these words, though there isn’t necessarily a clear consensus on the meanings and the degree to which they overlap or are interchangeable.

So, Computer security can be defined as controls that are put in place to provide confidentiality, integrity, and availability for all components of computer systems. Let’s elaborate the definition.

Components of 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

Computer security is mainly concerned with three main areas:

- Confidentiality is ensuring that information is available only to the intended audience
- Integrity is protecting information from being modified by unauthorized parties
- Availability is protecting information from being modified by unauthorized parties

Computer security threats

Computer security threats are possible dangers that can possibly hamper the normal functioning of your computer. In the present age, cyber threats are constantly increasing as the world is going digital. The most harmful types of computer security are:

Viruses

A computer virus is a malicious program which is loaded into the user's computer without user's knowledge. It replicates itself and infects the files and programs on the user's PC. The ultimate goal of a virus is to ensure that the victim's computer will never be able to operate properly or even at all.

Computer Worm

A computer worm is a software program that can copy itself from one computer to another, without human interaction. The potential risk here is that it will use up your computer hard disk space because a worm can replicate in greate volume and with great speed.

Cyber Security Training

Phishing :

Disguising as a trustworthy person or business, phishers attempt to steal sensitive financial or personal information through fraudulent email or instant messages. Phishing in unfortunately very easy to execute. You are deluded into thinking it's the legitimate mail and you may enter your personal information.

Botnet :

A botnet is a group of computers connected to the internet, that have been compromised by a hacker using a computer virus. An individual computer is called 'zombie computer'. The result of this threat is the victim's computer, which is the bot will be used for malicious activities and for a larger scale attack like DDoS.

Rootkit :

A rootkit is a computer program designed to provide continued privileged access to a computer while actively hiding its presence. Once a rootkit has been installed, the controller of the rootkit will be able to remotely execute files and change system configurations on the host machine

Keylogger :

 Also known as a keystroke logger, keyloggers can track the real-time activity of a user on his computer. It keeps a record of all the keystrokes made by user keyboard. Keylogger is also a very powerful threat to steal people's login credential such as username and password.

5.5 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.

Description: A software product is judged by how easily it can be used by the end-user and the features it offers to the user. An application must score in the following areas:-

1) Operational: -This tells how good a software works on operations like budget , usability, efficiency, correctness ,functionality , dependability , security and safety.

2) Transitional: - Transitional is important when an application is shifted from one platform to another. So, portability, reusability and adaptability come in this area.

3) Maintenance: - This specifies how good a software works in the changing environment. Modularity, maintainability, flexibility and scalability come in maintenance part.

Software Development Lifecycle or SDLC is a series of stages in software engineering to develop proposed software application, such as:

1) Communication

2) Requirement Gathering

3) Feasibility Study

4) System Analysis

5) Software Design

6) Coding

7) Testing

8) Integration

9) Implementation

10) Operations and maintenance

11) Disposition

Software engineering generally begins with the first step as a user-request initiation for a specific task or an output. He submits his requirement to a service provider organization. The software development team segregates user requirement, system requirement and functional requirements. The requirement is collected by conducting interviews of a user, referring to a database, studying the existing system etc. After requirement gathering, the team analyses if the software can be made to fulfil all the requirements of the user. The developer then decides a roadmap of his plan. System analysis also includes an understanding of software product limitations. As per the requirement and analysis, a software design is made. The implementation of software design starts in terms of writing program code in a suitable programming language. Software testing is done while coding by the developers and thorough testing is conducted by testing experts at various levels of code such as module testing, program testing, product testing, in-house testing and testing the product at user's engagement and feedback.

5.6 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.

As with other contexts and meanings of the word architecture, computer architecture is likened to the art of determining the needs of the user/system/technology, and creating a logical design and standards based on those requirements.

A very good example of computer architecture is von Neumann architecture, which is still used by most types of computers today. This was proposed by the mathematician John von Neumann in 1945. It describes the design of an electronic computer with its CPU, which includes the arithmetic logic unit, control unit, registers, memory for data and instructions, an input/output interface and external storage functions.

There are three categories of computer architecture:

- 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. It also includes memory controllers, data paths and miscellaneous things like multiprocessing and virtualization.
- 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.
- Microarchitecture: 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.

5.7 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.

The OS helps you to communicate with the computer without knowing how to speak the computer's language. It is not possible for the user to use any computer or mobile device without having an operating system.

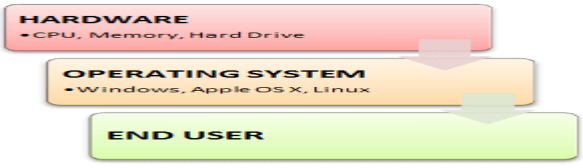
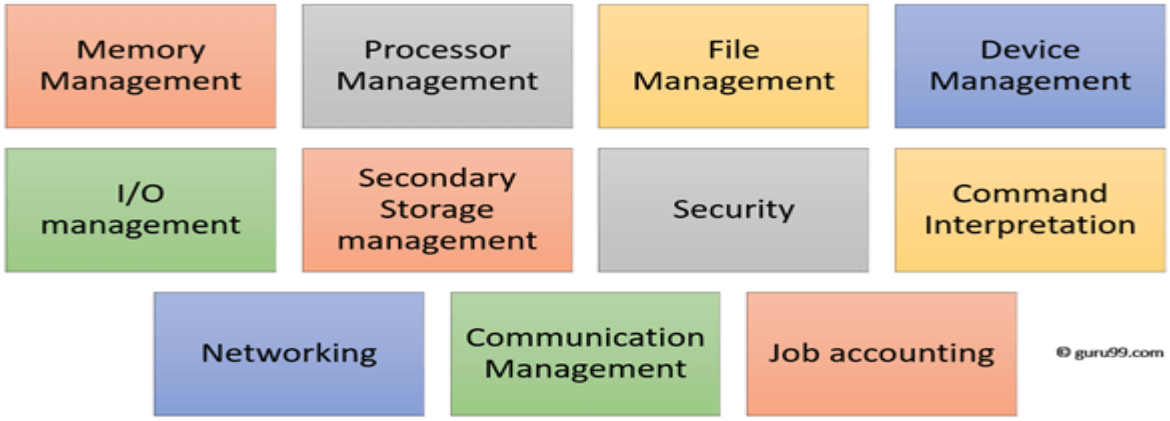


Fig: Operating System

Functions of operating Systems:



In an operating system software performs each of the function:

1. Process management:- Process management helps OS to create and delete processes. It also provides mechanisms for synchronization and communication among processes.

2. Memory management:- Memory management module performs the task of allocation and de-allocation of memory space to programs in need of this resources.

3. File management:- It manages all the file-related activities such as organization storage, retrieval, naming, sharing, and protection of files.

4. Device Management: Device management keeps tracks of all devices. This module also responsible for this task is known as the I/O controller. It also performs the task of allocation and de-allocation of the devices.

5. I/O System Management: One of the main objects of any OS is to hide the peculiarities of that hardware devices from the user.

6. Secondary-Storage Management: Systems have several levels of storage which includes primary storage, secondary storage, and cache storage. Instructions and data must be stored in primary storage or cache so that a running program can reference it.

7. Security:- Security module protects the data and information of a computer system against malware threat and authorized access.

8. Command interpretation: This module is interpreting commands given by the and acting system resources to process that commands.

9. Networking: A distributed system is a group of processors which do not share memory, hardware devices, or a clock. The processors communicate with one another through the network.

10. Job accounting: Keeping track of time & resource used by various job and users.

11. Communication management: Coordination and assignment of compilers, interpreters, and another software resource of the various users of the computer systems.

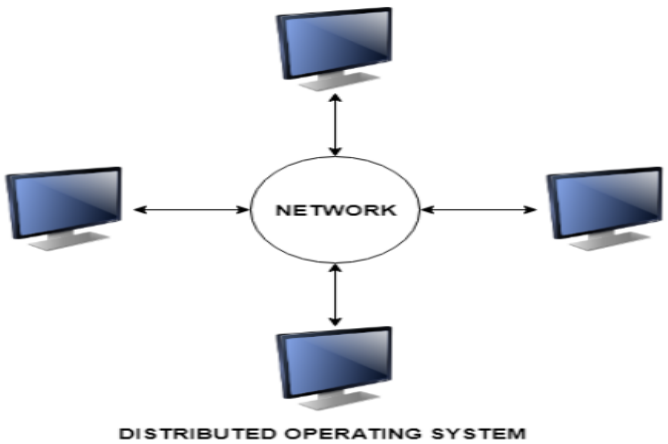
Types of Operating system

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

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

Types of Distributed Systems

The nodes in the distributed systems can be arranged in the form of client/server systems or peer to peer systems. Details about these are as follows:

Client/Server Systems

In client server systems, the client requests a resource and the server provides that resource. A server may serve multiple clients at the same time while a client is in contact with only one server. Both the client and server usually communicate via a computer network and so they are a part of distributed systems.

Peer to Peer Systems

The peer to peer systems contains nodes that are equal participants in data sharing. All the tasks are equally divided between all the nodes. The nodes interact with each other as required as share resources. This is done with the help of a network.

Advantages of Distributed Systems

Some advantages of Distributed Systems are as follows:

- All the nodes in the distributed system are connected to each other. So nodes can easily share data with other nodes.
- More nodes can easily be added to the distributed system i.e. it can be scaled as required.
- Failure of one node does not lead to the failure of the entire distributed system. Other nodes can still communicate with each other.
- Resources like printers can be shared with multiple nodes rather than being restricted to just one.

Disadvantages of Distributed Systems

Some disadvantages of Distributed Systems are as follows:

- It is difficult to provide adequate security in distributed systems because the nodes as well as the connections need to be secured.
- Some messages and data can be lost in the network while moving from one node to another.
- The database connected to the distributed systems is quite complicated and difficult to handle as compared to a single user system.

- Overloading may occur in the network if all the nodes of the distributed system try to send data at once.

5.9 Bioinformatics:

a hybrid science that links biological data with techniques for information storage, distribution, and analysis to support multiple areas of scientific research, including biomedicine. Bioinformatics is fed by high-throughput data-generating experiments, including genomic sequence determinations and measurements of gene expression patterns. Database projects curate and annotate the data and then distribute it via the World Wide Web. Mining these data leads to scientific discoveries and to the identification of new clinical applications. In the field of medicine in particular, a number of important applications for bioinformatics have been discovered. For example, it is used to identify correlations between gene sequences and diseases, to predict protein structures from amino acid sequences, to aid in the design of novel drugs, and to tailor treatments to individual patients based on their DNA sequences (pharmacogenomics).

The Data Of Bioinformatics

The classic data of bioinformatics include DNA sequences of genes or full genomes; amino acid sequences of proteins; and three-dimensional structures of proteins, nucleic acids and protein–nucleic acid complexes. Additional "-omics" data streams include: transcriptomics, the pattern of RNA synthesis from DNA; proteomics, the distribution of proteins in cells; interactomics, the patterns of protein-protein and protein–nucleic acid interactions; and metabolomics, the nature and traffic patterns of transformations of small molecules by the biochemical pathways active in cells. In each case there is interest in obtaining comprehensive, accurate data for particular cell types and in identifying patterns of variation within the data. For example, data may fluctuate depending on cell type, timing of data collection (during the cell cycle, or diurnal, seasonal, or annual variations), developmental stage, and various external conditions. Metagenomics and metaproteomics extend these measurements to a comprehensive description of the organisms in an environmental sample, such as in a bucket of ocean water or in a soil sample.

Bioinformatics has been driven by the great acceleration in data-generation processes in biology. Genome sequencing methods show perhaps the most dramatic effects. In 1999 the nucleic acid sequence archives contained a total of 3.5 billion nucleotides, slightly more than the length of a single human genome; a decade later they contained more than 283 billion nucleotides, the length of about 95 human genomes. The U.S. National Institutes of Health has challenged researchers by setting a goal to reduce the cost of sequencing a human genome to $1,000; this would make DNA

sequencing a more affordable and practical tool for U.S. hospitals and clinics, enabling it to become a standard component of diagnosis.

Storage And Retrieval Of Data

In bioinformatics, data banks are used to store and organize data. Many of these entities collect DNA and RNA sequences from scientific papers and genome projects. Many databases are in the hands of international consortia. For example, an advisory committee made up of members of the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL-Bank) in the United Kingdom, the DNA Data Bank of Japan (DDBJ), and GenBank of the National Center for Biotechnology Information (NCBI) in the United States oversees the International Nucleotide Sequence Database Collaboration (INSDC). To ensure that sequence data are freely available, scientific journals require that new nucleotide sequences be deposited in a publicly accessible database as a condition for publication of an article. (Similar conditions apply to nucleic acid and protein structures.) There also exist genome browsers, databases that bring together all the available genomic and molecular information about a particular species.

Goals of Bioinformatics:

The development of efficient algorithms for measuring sequence similarity is an important goal of bioinformatics. The Needleman-Wunsch algorithm, which is based on dynamic programming, guarantees finding the optimal alignment of pairs of sequences. This algorithm essentially divides a large problem (the full sequence) into a series of smaller problems (short sequence segments) and uses the solutions of the smaller problems to construct a solution to the large problem. Similarities in sequences are scored in a matrix, and the algorithm allows for the detection of gaps in sequence alignment.

Although the Needleman-Wunsch algorithm is effective, it is too slow for probing a large sequence database. Therefore, much attention has been given to finding fast information-retrieval algorithms that can deal with the vast amounts of data in the archives. An example is the program BLAST (Basic Local Alignment Search Tool). A development of BLAST, known as position-specific iterated- (or PSI-) BLAST, makes use of patterns of conservation in related sequences and combines the high speed of BLAST with very high sensitivity to find related sequences.

Another goal of bioinformatics is the extension of experimental data by predictions. A fundamental goal of computational biology is the prediction of protein structure from an amino acid sequence. The spontaneous folding of proteins shows that this should be possible. Progress in the development of

methods to predict protein folding is measured by biennial Critical Assessment of Structure Prediction (CASP) programs, which involve blind tests of structure prediction methods.

Bioinformatics is also used to predict interactions between proteins, given individual structures of the partners. This is known as the "docking problem." Protein-protein complexes show good complementarity in surface shape and polarity and are stabilized largely by weak interactions, such as burial of hydrophobic surface, hydrogen bonds, and van der Waals forces. Computer programs simulate these interactions to predict the optimal spatial relationship between binding partners. A particular challenge, one that could have important therapeutic applications, is to design an antibody that binds with high affinity to a target protein.

Initially, much bioinformatics research has had a relatively narrow focus, concentrating on devising algorithms for analyzing particular types of data, such as gene sequences or protein structures. Now, however, the goals of bioinformatics are integrative and are aimed at figuring out how combinations of different types of data can be used to understand natural phenomena, including organisms and disease.

5.10 MACHINE LEARNING

Machine Learning - Introduction

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.

In this tutorial, you will learn in detail about −

What is the crux of machine learning?

- What are the different types in machine learning?
- What are the different algorithms available for developing machine learning models?
- What tools are available for developing these models?
- What are the programming language choices?
- What platforms support development and deployment of Machine Learning applications?
- What IDEs (Integrated Development Environment) are available?
- How to quickly upgrade your skills in this important area?

Machine Learning - What Today's AI Can Do?

When you tag a face in a Facebook photo, it is AI that is running behind the scenes and identifying faces in a picture. Face tagging is now omnipresent in several applications that display pictures with human faces. Why just human faces? There are several applications that detect objects such as cats, dogs, bottles, cars, etc. We have autonomous cars running on our roads that detect objects in real time to steer the car. When you travel, you use Google **Directions** to learn the real-time traffic situations and follow the best path suggested by Google at that point of time. This is yet another implementation of object detection technique in real time.

Let us consider the example of Google **Translate** application that we typically use while visiting foreign countries. Google's online translator app on your mobile helps you communicate with the local people speaking a language that is foreign to you.

There are several applications of AI that we use practically today. In fact, each one of us use AI in many parts of our lives, even without our knowledge. Today's AI can perform extremely complex jobs with a great accuracy and speed. Let us discuss an example of complex task to understand what capabilities are expected in an AI application that you would be developing today for your clients.

Example

We all use Google **Directions** during our trip anywhere in the city for a daily commute or even for inter-city travels. Google Directions application suggests the fastest path to our destination at that time instance. When we follow this path, we have observed that Google is almost 100% right in its suggestions and we save our valuable time on the trip.

You can imagine the complexity involved in developing this kind of application considering that there are multiple paths to your destination and the application has to judge the traffic situation in every possible path to give you a travel time estimate for each such path. Besides, consider the fact that Google Directions covers the entire globe. Undoubtedly, lots of AI and Machine Learning techniques are in-use under the hoods of such applications.

Considering the continuous demand for the development of such applications, you will now appreciate why there is a sudden demand for IT professionals with AI skills.

In our next chapter, we will learn what it takes to develop AI programs.

Machine Learning - Traditional AI

The journey of AI began in the 1950's when the computing power was a fraction of what it is today. AI started out with the predictions made by the machine in a fashion a statistician does predictions using his calculator. Thus, the initial entire AI development was based mainly on statistical techniques.

In this chapter, let us discuss in detail what these statistical techniques are.

Statistical Techniques

The development of today's AI applications started with using the age-old traditional statistical techniques. You must have used straight-line interpolation in schools to predict a future value. There are several other such statistical techniques which are successfully applied in developing so-called AI programs. We say "so-called" because the AI programs that we have today are much more complex and use techniques far beyond the statistical techniques used by the early AI programs.

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

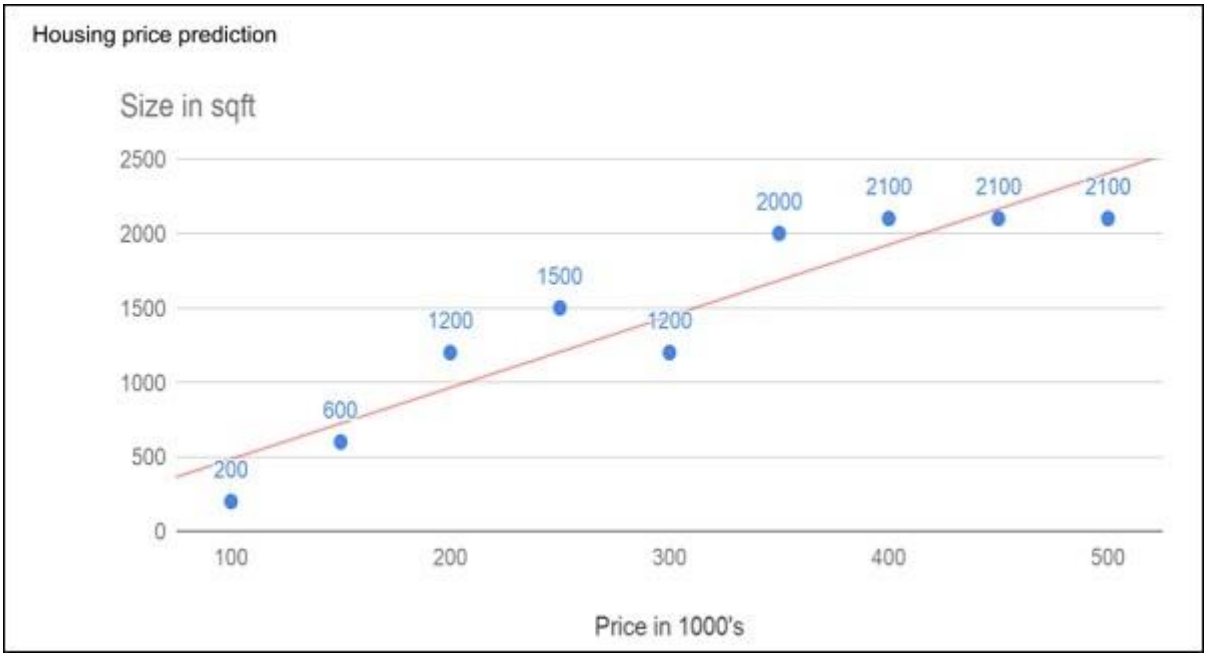
Here we have listed only some primary techniques that are enough to get you started on AI without scaring you of the vastness that AI demands. If you are developing AI applications based on limited data, you would be using these statistical techniques.

However, today the data is abundant. To analyze the kind of huge data that we possess statistical techniques are of not much help as they have some limitations of their own. More advanced methods such as deep learning are hence developed to solve many complex problems.

As we move ahead in this tutorial, we will understand what Machine Learning is and how it is used for developing such complex AI applications.

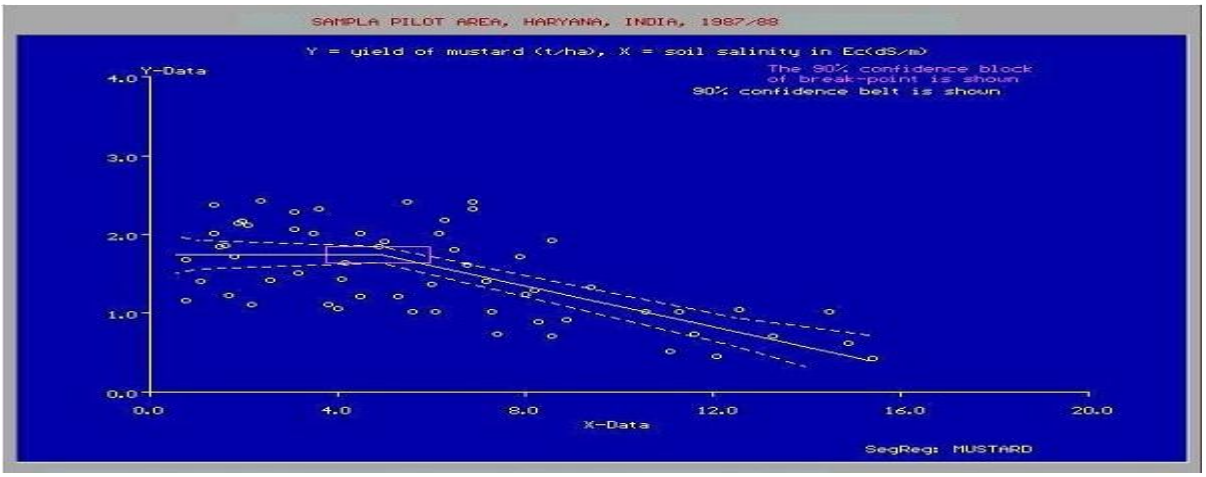
Machine Learning - What is Machine Learning?

Consider the following figure that shows a plot of house prices versus its size in sq. ft.



After plotting various data points on the XY plot, we draw a best-fit line to do our predictions for any other house given its size. You will feed the known data to the machine and ask it to find the best fit line. Once the best fit line is found by the machine, you will test its suitability by feeding in a known house size, i.e. the Y-value in the above curve. The machine will now return the estimated X-value, i.e. the expected price of the house. The diagram can be extrapolated to find out the price of a house which is 3000 sq. ft. or even larger. This is called regression in statistics. Particularly, this kind of regression is called linear regression as the relationship between X & Y data points is linear.

In many cases, the relationship between the X & Y data points may not be a straight line, and it may be a curve with a complex equation. Your task would be now to find out the best fitting curve which can be extrapolated to predict the future values. One such application plot is shown in the figure below.

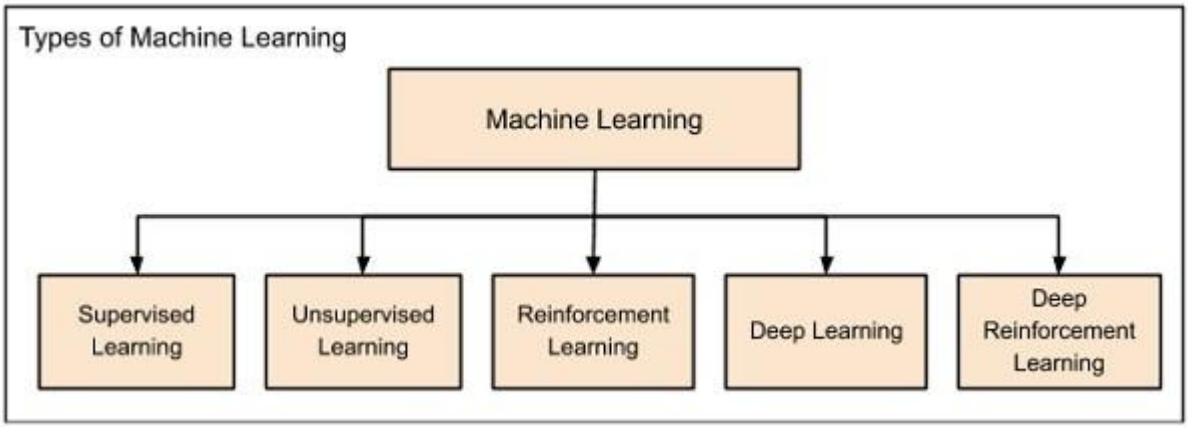


You will use the statistical optimization techniques to find out the equation for the best fit curve here. And this is what exactly Machine Learning is about. You use known optimization techniques to find the best solution to your problem.

Next, let us look at the different categories of Machine Learning.

Machine Learning - Categories

Machine Learning is broadly categorized under the following headings −



Machine learning evolved from left to right as shown in the above diagram.

- Initially, researchers started out with Supervised Learning. This is the case of housing price prediction discussed earlier.
- This was followed by unsupervised learning, where the machine is made to learn on its own without any supervision.
- Scientists discovered further that it may be a good idea to reward the machine when it does the job the expected way and there came the Reinforcement Learning.
- Very soon, the data that is available these days has become so humongous that the conventional techniques developed so far failed to analyze the big data and provide us the predictions.
- Thus, came the deep learning where the human brain is simulated in the Artificial Neural Networks (ANN) created in our binary computers.
- The machine now learns on its own using the high computing power and huge memory resources that are available today.
- It is now observed that Deep Learning has solved many of the previously unsolvable problems.

- The technique is now further advanced by giving incentives to Deep Learning networks as awards and there finally comes Deep Reinforcement Learning.

Let us now study each of these categories in more detail.

Supervised Learning

Supervised learning is analogous to training a child to walk. You will hold the child's hand, show him how to take his foot forward, walk yourself for a demonstration and so on, until the child learns to walk on his own.

Regression

Similarly, in the case of supervised learning, you give concrete known examples to the computer. You say that for given feature value x1 the output is y1, for x2 it is y2, for x3 it is y3, and so on. Based on this data, you let the computer figure out an empirical relationship between x and y.

Once the machine is trained in this way with a sufficient number of data points, now you would ask the machine to predict Y for a given X. Assuming that you know the real value of Y for this given X, you will be able to deduce whether the machine's prediction is correct.

Thus, you will test whether the machine has learned by using the known test data. Once you are satisfied that the machine is able to do the predictions with a desired level of accuracy (say 80 to 90%) you can stop further training the machine.

Now, you can safely use the machine to do the predictions on unknown data points, or ask the machine to predict Y for a given X for which you do not know the real value of Y. This training comes under the regression that we talked about earlier.

Classification

You may also use machine learning techniques for classification problems. In classification problems, you classify objects of similar nature into a single group. For example, in a set of 100 students say, you may like to group them into three groups based on their heights - short, medium and long. Measuring the height of each student, you will place them in a proper group.

Now, when a new student comes in, you will put him in an appropriate group by measuring his height. By following the principles in regression training, you will train the machine to classify a student based on his feature – the height. When the machine learns how the groups are formed, it will

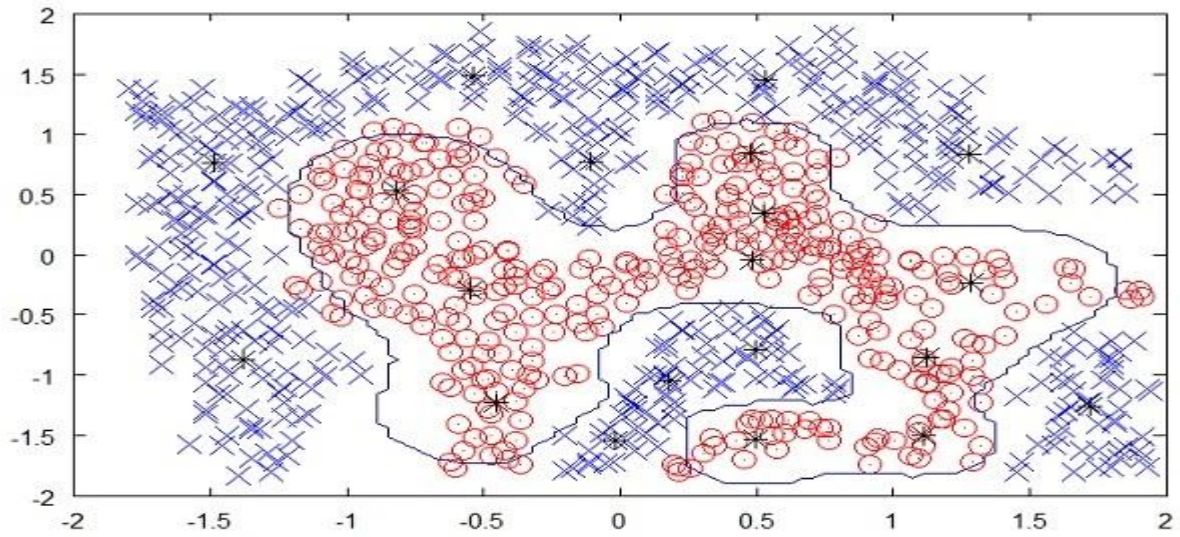
be able to classify any unknown new student correctly. Once again, you would use the test data to verify that the machine has learned your technique of classification before putting the developed model in production.

Supervised Learning is where the AI really began its journey. This technique was applied successfully in several cases. You have used this model while doing the hand-written recognition on your machine. Several algorithms have been developed for supervised learning. You will learn about them in the following chapters.

Unsupervised Learning

In unsupervised learning, we do not specify a target variable to the machine, rather we ask machine "What can you tell me about X?". More specifically, we may ask questions such as given a huge data set X, "What are the five best groups we can make out of X?" or "What features occur together most frequently in X?". To arrive at the answers to such questions, you can understand that the number of data points that the machine would require to deduce a strategy would be very large. In case of supervised learning, the machine can be trained with even about few thousands of data points. However, in case of unsupervised learning, the number of data points that is reasonably accepted for learning starts in a few millions. These days, the data is generally abundantly available. The data ideally requires curating. However, the amount of data that is continuously flowing in a social area network, in most cases data curation is an impossible task.

The following figure shows the boundary between the yellow and red dots as determined by unsupervised machine learning. You can see it clearly that the machine would be able to determine the class of each of the black dots with a fairly good accuracy.

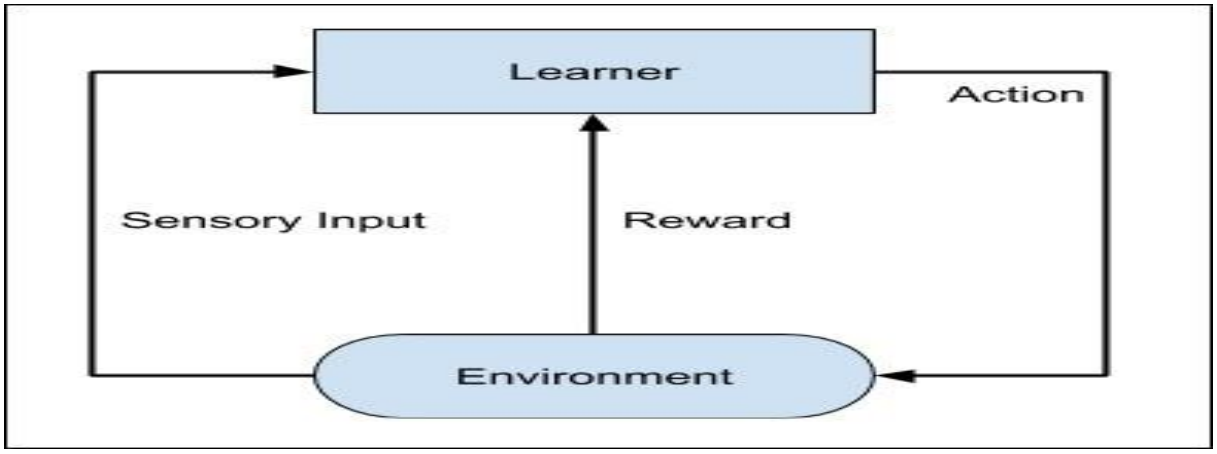


The unsupervised learning has shown a great success in many modern AI applications, such as face detection, object detection, and so on.

Reinforcement Learning

Consider training a pet dog, we train our pet to bring a ball to us. We throw the ball at a certain distance and ask the dog to fetch it back to us. Every time the dog does this right, we reward the dog. Slowly, the dog learns that doing the job rightly gives him a reward and then the dog starts doing the job right way every time in future. Exactly, this concept is applied in "Reinforcement" type of learning. The technique was initially developed for machines to play games. The machine is given an algorithm to analyze all possible moves at each stage of the game. The machine may select one of the moves at random. If the move is right, the machine is rewarded, otherwise it may be penalized. Slowly, the machine will start differentiating between right and wrong moves and after several iterations would learn to solve the game puzzle with a better accuracy. The accuracy of winning the game would improve as the machine plays more and more games.

The entire process may be depicted in the following diagram −



This technique of machine learning differs from the supervised learning in that you need not supply the labelled input/output pairs. The focus is on finding the balance between exploring the new solutions versus exploiting the learned solutions.

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.

These networks have been successfully applied in solving the problems of computer vision, speech recognition, natural language processing, bioinformatics, drug design, medical image analysis, and games. There are several other fields in which deep learning is proactively applied. The deep learning requires huge processing power and humongous data, which is generally easily available these days.

We will talk about deep learning more in detail in the coming chapters.

Deep Reinforcement Learning

The Deep Reinforcement Learning (DRL) combines the techniques of both deep and reinforcement learning. The reinforcement learning algorithms like Q-learning are now combined with deep learning to create a powerful DRL model. The technique has been with a great success in the fields of robotics, video games, finance and healthcare. Many previously unsolvable problems are now solved by creating DRL models. There is lots of research going on in this area and this is very actively pursued by the industries.

So far, you have got a brief introduction to various machine learning models, now let us explore slightly deeper into various algorithms that are available under these models.