Summer of Science 2021

DATA MINING

July 19, 2021

Aniket Pokle
Mentor: Deepanshu Bagotia

Contents

INTRODUCTION

Data mining is a process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself.

While data analysis is used to test models and hypotheses on the datasets, data mining on the other hand uses machine learning and statistical models to uncover clandestine or hidden patterns in a large volume of data.

In this report I try to understand some of the statistical and machine learning aspects of Data mining such as data management using SQL, data pre-processing, model and inference considerations and machine learning algorithms used in data mining.

PROBABILITY AND STATISTICS

Probability

The extent to which an event is likely to occur, measured by the ratio of the favourable cases to the whole number of cases possible.
Basically Probability is the measure of uncertainity.

Sample space and Event

Any activity for which the outcome is uncertain can be thought of as an "experiment."
The set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted by Ω.
Any subset E of the sample space is known as an event.
An event is a set of possible outcomes of the experiment.

Baye's Theorem

Bayes Theorem provides a principled way for calculating a conditional probability.
$Assume\ that\ A_1, A_2, ...A_m\ are\ events\ such\ that$

- $A_1 \cup A_2 \cup ... \cup A_m = \Omega$
- $A_i \cap A_j = \phi\ (pairwisedisjoint)\ \forall i \neq j = 1, 2, ...m$
- $P(A_i) > 0\ for\ all\ i\ and\ B\ is\ another\ event\ than\ A,\ then$

$$P(A_i|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{m} P(B|A_i)P(A_i)}$$

is known as Bayes' Formula
$P(A_i) : Prior\ probabilities$
$P(A_i|B) : Posterior\ probabilities$
$P(B|A_i) : Model\ probabilities$

Independent Events

Definition: Two random events A and B are called (stochastically) independent If P(AB) = P(A) P(B).
i.e. if the probability of simultaneous occurrence of both events A and B is the product of the individual probabilities of occurrence of A and B.

Need of Random Variable

In any random experiment, we are interested in the value of some numerical quantity determined by the result.
We are not interested in all the details of the experiments.
These quantities of interest that are determined by the result of the experiment are known as random variables.

Random Variable

Let Ω represent the sample space of a random experiment, and let R be the set of real numbers. A random variable is a function X which assigns to each element $\omega \in \Omega$ one and only one number

$$X(\omega) = x, x \in R,\ i.e.\ X : \Omega \rightarrow R$$

Probability Density Function(PDF) for Continuous Random Variable

For a function f(X) to be a probability density function (PDF) of a continuous random variable X, it needs to satisfy the following conditions:

1. f(X) $\leq 0 \forall x \in R$
2. $\int_{-\infty}^{\infty} f(x)\ dx\ =\ 1$

Let X be a random variable with CDF F(x).
If $x_1\ <\ x_2$ where $x_1\ and\ x_2$ are known constants,

$$P(x_1 \leq X \leq x_2) = F(x_2)\ -\ F(x_2)\ = x_2\ -\ x_1\ = \int_{x_1}^{x_2} f(x)\ dx$$

Cumulative Distribution Function(CDF)

The cumulative distribution function, or more simply the distribution function, F of the random variable X is defined for any real number x by

$$F(x) = P(X \leq x)$$

That is, F(x) is the probability that the random variable X takes on a value that is less than or equal to x.

CDF of Continuous Random Variables

A random variable X is said to be continuous if there is a function f(x) such that for all x ∈ R

$$F(x) = \int_{-\infty}^{\infty} f(t)\, dt$$

holds.

- F(x) is the CDF of X, and
- f(x) is the PDF of X.

CDF of Discrete Random Variable

The cumulative distribution function CDF of a discrete random variable as

$$F(X) = \sum_{i=1}^{k} I_{x_i \leq x} p_i$$

where I is an indicator function defined as

$$I_{x_i \leq x} = \begin{cases} 1\ if\ x_i \leq x \\ 0\ otherwise. \end{cases} \tag{1}$$

The CDF of a discrete variable is always a step function.

Expectation of a Continuous Random Variable

Let X be a continuous random variable having the probability density function f (x).
Suppose g(X) is a real valued function of X.
Obviously g(X) will also be a random variable.
Then expectation of g(X) is defined as is defined as

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx$$

provided with

$$\int |g(x)| f(x)\, dx < \infty$$

Expectation of a Discrete Random Variable

Let X be a discrete random variable having the probability mass function P(X = x_i) = p_i . Suppose g(X) is a real valued function of X. Obviously g(X) will also be a random variable. Thus X takes the values $x_1, x_2, ..., x_k$ with respective probabilities $p_1, p_2, ..., p_k$ Then expectation of g(X) exists and is defined as

$$E[g(X)] = \sum_{i=1}^{\infty} g(x_i)P(X = x_i) = \sum_{i=1}^{\infty} g(x_i)p_i$$

provided

$$\sum_{i=1}^{\infty} |g(x_i)|p_i < \infty$$

Moments

Moments are used to describe different characteristics and features of a probability distribution, viz., central tendency, disperson, symmetry and peakedness (hump) of probability curve.

1. g(X) = X^r where r is nonnegative integer,
   then E[g(X)] = E(X^r) = $\mu_r^{'}$
   $\mu_r^{'}$ is called as the r^{th} moment of X about origin.

2. g(X) = $(X - A)^r$ where r is non-negative integer,
   then E[g(X)] = E$(X - A)^r$
   is called as r^{th} moment of X about the point "A".

Central Moment

The moments of a variable X about the arithmetic mean $\bar{x}$ are called central moments.
For E[g(X)] = E$(X - A)^r$
If A = E(X) : Mean
then E$(X - A)^r = E[X - E(X)]^r = \mu_r$
μ_r is called as the r^{th} central moment of X.

Quantiles

We define quantiles in terms of the distribution function.
The value x_p for which the cumulative distribution function is

$F(x_p) = p$ $(0 \leq p \leq 1)$
is called the p-quantile.
x_p is the value which divides the CDF into two parts:

- the probability of observing a value left of x_p is p
- the probability of observing a value right of x_p is 1 - p.

Tschebyschev's Inequality

If we do not know the distribution of a random variable X, we can still make statements about the probability using Tschebyschev's inequality that X takes values in a certain interval (which has to be symmetric around the expectation μ) if the mean μ and the variance σ^2 of X are known.
Let X be a random variable with E(X) = μ and Var(X) = σ^2. It holds that

$$P(|X - \mu| \geq c) \leq \frac{Var(X)}{c^2}$$

This is equivalent to

$$P(|X - \mu| < c) \leq 1 - \frac{Var(X)}{c^2}$$

Need of Probability Distributions

Probability distribution functions have special properties and describe how probabilities are distributed under different conditions.
The form of such functions depends upon the nature and complexity of the phenomenon under consideration.
We have probability distributions for discrete and continuous random variables.

Degenerate Distribution

A random variable X has a degenerate distribution at c, if c is the only possible outcome.
The probability mass function (PMF) of X is given by

$$P(X = x) = \begin{cases} 1 \text{ if } x = c \\ 0 \; if \; x \neq c \end{cases} \tag{2}$$

The CDF in such case is given by

$$F(X) = \begin{cases} 0 \text{ if x < c} \\ 1 \; if \; x \geq c. \end{cases} \tag{3}$$

Further, The mean (expectation) and variance of X are **E(X) = c** and **Var(X) = 0**.
The degenerate distribution indicates that there is only one possible fixed outcome, and therefore, no randomness is involved.

Discrete Uniform Distribution

Consider a situation where the probability of all the outcomes are the same.
In such situations, the discrete uniform distribution can be used to describe the probabilities and the phenomenon.
The discrete uniform distribution assumes that all possible outcomes have equal probability of occurrence.

PMF

A discrete random variable X with k possible outcomes $x_1, x_2, ..., x_k$ is said to follow a discrete uniform distribution if the probability mass function (PMF) of X is given by

$$P(X = x_i) = \frac{1}{k} \; \forall \; i \; = \; 1, 2, ...k.$$

Mean and Variance

If the outcomes are the natural numbers x_i = i (i = 1, 2, . . . , k), then the mean and variance of X are as follows:

$$E(X) \; = \; \frac{k+1}{2}$$

$$Var(X) \; = \; \frac{k^2 \; - \; 1}{2}$$

Bernoulli Distribution

A Bernoulli experiment is a random experiment, the outcome of which can be classified in but one of two mutually exclusive and exhaustive ways, e.g. success or failure.
If we let X = 1 when the outcome is a success and X = 0 when it is a failure, then a random

variable X has a Bernoulli distribution if the probability mass function (PMF) of X is given by

$$P(X = x) = \begin{cases} p \; if \; x = 1 \\ 1 - p \; if \; x = 0. \end{cases} \tag{4}$$

The CDF in such case is given by

$$F(X) = \begin{cases} 0 \; if \; x < 0 \\ 1 - p \; if \; 0 \leq x \leq 1 \\ 1 \; if \; x \leq 1. \end{cases} \tag{5}$$

The mean(expectation) and variance of a Bernoulli random variable are calculated as **E(X) = p** and **Var(X) = p(1 – p)**.

Binomial Distribution

Consider n independent trials or repetitions of a Bernoulli experiment with probability of success p in each trial so that p remains constant in each trial.
In each trial or repetition, we may observe either A or $\bar{A}$.
At the end of the experiment, we have thus observed A between 0 and n times.
Suppose we are interested in the probability of A occurring k times, then the binomial distribution is useful.
A discrete random variable X is said to follow a binomial distribution with parameters n and p if its PMF is given by

$$P(X = k) = \begin{cases} \binom{n}{k} p^k \; q^{n-k} \; if \; k = 0, 1, ...n \\ 0 \; otherwise. \end{cases} \tag{6}$$

We also write X B(n,p).
The mean and variance of a binomial random variable X are given by **E(X) = np**, **Var(X) = np(1 - p)**.
Remark: A Bernoulli random variable is therefore B(1, p) distributed.

Poisson Distribution

A discrete random variable X is said to follow a Poisson distribution with parameter $\lambda > 0$ if its PMF is given by

$$P(X = x) = \frac{\lambda^x \, exp(-\lambda)}{x!} \; x = 0, 1, 2, ..$$

We also write X P(λ).
The mean and variance of a Poisson random variable are identical: E(X) = Var(X) = λ

Geometric Distribution

The geometric distribution can be used to determine the probability that the event of interest happens at the k^{th} trial for the first time.
A discrete random variable X is said to follow a geometric distribution with parameter p if its PMF is given by

$$P(X = k) = p(1-p)^{k-1}, \; k = 1, 2, 3, ...$$

The mean (expectation) and variance are given by

$$E(X) = \frac{1}{p}$$

$$Var(X) = \frac{1-p}{p^2}$$

Continuous Uniform Distribution

A continuous random variable X is said to follow a (continuous) uniform distribution in the interval [a, b], if its probability density function (PDF) is given by

$$f_x(x) \equiv f(x) = \begin{cases} \frac{1}{b-a} \; if \; a \leq x \leq b, \; (a < b) \\ 0 \; otherwise. \end{cases} \tag{7}$$

We also write X U(a,b).
The mean and variance of a X U(a,b) are given by

$$E(X) = \frac{a+b}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

Normal Distribution

The normal distribution is also often called a **Gaussian distribution**.
The most widely used model for the distribution of a random variable is a normal distribution.
A random variable X is said to follow a normal distribution with parameters μ and σ^2 if its PDF is given by

$$f(x;\mu,\sigma^2) \equiv f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(\frac{-(x-\mu)^2}{2\sigma^2}) - \infty < x < \infty; -\infty < \mu < \infty; \sigma^2 > 0. \quad (8)$$

We write X N(μ,σ^2).
The mean and variance of X are **E(X) =** μ and **Var(X) =** σ^2 respectively.
The value of E(X) = μ determines the center of the probability density function and the value of Var(X) = σ^2 determines the width.

Standard Normal Distribution

If $\mu = 0$ and $\sigma^2 = 1$, then X is said to follow a standard normal distribution.
We write X N(0,1).

Exponential Distribution

The exponential distribution is useful in many situations, for example when one is interested in the waiting time, or lifetime, until an event of interest occurs.
A continuous random variable X is said to follow an exponential distribution with parameter $\lambda > 0$, if its probability density function (PDF) is given by

$$f_x(x) \equiv f(x) = \begin{cases} \lambda exp(-\lambda x), \ if \ 0 \leq x \leq \infty \\ 0 \ otherwise \end{cases} \quad (9)$$

We also write X Exp(λ).
The mean and variance of X are

$$E(X) = \frac{1}{\lambda}$$

$$Var(X) = \frac{1}{\lambda^2}$$

respectively.
The CDF of exponential distribution is given as

$$F(x) = \begin{cases} 1 - exp(-\lambda x), & if\ 0 \leq x \leq \infty \\ 0 & otherwise \end{cases} \tag{10}$$

Joint Probability Distributions

If X and Y are discrete random variables, the joint probability distribution of X and Y is a description of the set of points (x, y) in the range of (X, Y) along with the probability of each point.
The joint probability distribution of two random variables is referred to as the bivariate probability distribution or bivariate distribution of the random variables.

Discrete Random Variables

The joint probability mass function of the discrete random variables X and Y, denoted as p_{XY}(x, y), satisfies

- p_{XY}(x, y) ≥ 0
- $\sum_x \sum_y p_{XY}$(x, y) = 1
- p(x,y) = P(X=x, Y=y)

Continuous Random Variables

The joint probability distribution of two continuous random variables X and Y can be specified by providing a method for calculating the probability that X and Y assume a value in any region R of two-dimensional space.
Analogous to the probability density function of a single continuous random variable, a joint probability density function can be defined over two-dimensional space.

Joint Cumulative Distribution Function

A bivariate random variable (X, Y) is continuous if there is a function f_{XY}(x, y) such that

$$F_{XY}(x, y) = P(X \leq x,\ Y \leq y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(x, y)\, dx dy$$

holds.
The function F_{XY}(x, y) is the joint cumulative distribution function of X and Y.

Covariance

The covariance between X and Y is defined as
Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X) E(Y).
This is based on Product and first moments.
The covariance is

- positive if, on average, larger values of X correspond to larger values of Y
- it is negative if, on average, greater values of X correspond to smaller values of Y.

If the random variables X_1 and X_2 are bivariate, the covariance matrix is defined as

$$Cov\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) \\ Cov(X_1, X_2) & Var(X_2) \end{pmatrix}$$

Also Note that Cov(X_1,X_2) = Cov(X_2,X_1)

Bivariate Normal Distribution

An extension of a normal distribution to two random variables is bivariate normal distribution.
An extension of a normal distribution to more than two random variables is multivariate normal distribution.
The probability density function of a bivariate normal distribution is

$$f_{XY}(x, y, \sigma_x, \sigma_y, \mu_x, \mu_y, \rho) =$$

$$\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-p^2}} exp\left\{ -\frac{1}{2(1-p^2)} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} \right] \right\}$$

$$-\infty < x < \infty, -\infty < y < \infty, \sigma_x > 0, \sigma_y > 0, -1 < \rho < 1.$$

Chi square(χ^2) Distribution

Let $Z_1, Z_2, ..., Z_n$ be n independent and identically N(0,1)-distributed random variables. The sum of their squares, $\sum_{i=1}^{n} Z_i^2$ is then χ^2 distributed with n degrees of freedom and is denoted

as χ_n^2

A random variable X has a χ_n- distribution if the PDF of X is given as

$$f(x) = \begin{cases} \frac{x^{\frac{n-2}{2}} exp(\frac{-x}{2})}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})}, & x > 0 \\ 0 \; otherwise \end{cases}$$

We write X χ_n^2.

The mean and variance of a random variable χ_n^2 distribution is

E(X) = n

Var(X) = 2n respectively.

The χ_n^2 distribution is not symmetric.

A χ_n^2 distributed random variable can only realize values greater than or equal to zero.

t - Distribution

Let X and Y be two independent random variables where X N(0,1) and Y χ_n^2. Then the ratio

$$\frac{X}{\sqrt{\frac{Y}{n}}} \simeq t_n$$

follows a t distribution (Student's tdistribution) with degree n of freedom. This is central tdistribution.

A random variable X has a t-distribution if the PDF of X is given as

$$f(x) = \frac{\Gamma\frac{n+1}{2}}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{x^2}{n}\right)^{\frac{-n-1}{2}} \; ; -\infty < x < \infty.$$

The mean and variance of a random variable X t_n distribution is

E(X) = 0, n > 1

Var(X) = $\frac{n}{n-2}$, n > 2 respectively.

The tdistribution is a symmetric distribution.

F - Distribution

Let X and Y be two independent random variables where X ‘ χ^2_m and Y χ^2_n . Then the ratio

$$\frac{\frac{X}{m}}{\frac{Y}{n}} \simeq F_{m,n}$$

follows the Fisher F-distribution with (m, n) degrees of freedom. We write X $F_{m,n}$.
The mean and variance of a random variable X $F_{m,n}$ distribution is

$$E(X) = \frac{n}{n-2},\ n > 2$$

$$Var(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)},\ n > 4$$

respectively.
The F random variable is nonnegative, and the distribution is skewed to the right.
The “degrees of freedom” specify the shape of the distribution.

Weak Law of Large Numbers

A positive integer n can be determined such that if a random sample of size n or larger is taken from a population with the density f(x) (with E(X) = μ), the probability can be made to be as close to 1 as desired that the sample mean $\bar{X}$ will deviate from μ by less than any arbitrarily specified small quantity. If n is any integer greater than $\frac{\sigma^2}{\epsilon\delta^2}$, then

$$P[|\bar{X_n} - \mu| < \epsilon] \geq 1 - \delta$$

where $\epsilon > 0$ and $0 < \delta < 1$.

Central Limit Theorem

The central limit theorem tells that the sum of a large number of independent random variables has approximately a normal distribution.
It provides a simple method for computing approximate probabilities for sums of independent random variables

Theorem

Let $X_1, X_2, ..., X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2 .
Then for n large, the distribution of $X_1 + X_2 + , ..., X_n$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$.
It follows from the central limit theorem that

$$\frac{X_1 + X_2 + , ..., X_n - n\mu}{\sigma\sqrt{n}}$$

is approximately normal with mean 0 and variance 1, i.e. Standard normal distribution N(0, 1).

Continuity Correction

When we approximate the probabilities for discrete distributions, we incorporate the continuity correction also.
Let $X_1 + X_2 + , ..., X_n$ be a sequence of independent and identically distributed discrete random variables and let

$$Y = X_1 + X_2 + , ..., X_n$$

Suppose that we are interested in finding P(A) = P(l≤Y≤u) using the CLT, where l and u are integers. Since Y is an integer-valued random variable, we can write

$$P(A) = P\left(l - \frac{1}{2} \leq Y \leq u + \frac{1}{2}\right)$$

It turns out that the above expression sometimes provides a better approximation for P(A) when applying the CLT. This is called the continuity correction and it is particularly useful when Y is binomial.

Sample Mean Distribution

Let $X_1 + X_2 + , ..., X_n$ be a sample from a population having mean μ and variance σ^2. The central limit theorem can be used to approximate the distribution of the sample mean $\bar{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Since E($\bar{X_n}$) = μ, Var($\bar{X_n}$) = $\frac{\sigma^2}{n}$ and $\bar{X_n}$ is based on a linear combination of normally distributed

random variables, so when sample size n is large, then

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

has approx. a standard normal distribution N(0,1).

Statistical Inference

After completing the experiment, data is described and summarized with an aim to draw a statistical conclusion using the tools of inferential statistics.
A careful description and presentation of the data enable us to infer an appropriate probability model for a given data set which can be verified by using the additional data.
The tools of statistical inference lay the foundation of the formulation of a probability model to describe the data.

Unbiased Estimators

An estimator should be "close" in some sense to the true value of the unknown parameter. Formally, we say that $\hat{\theta}$ is an unbiased estimator of θ if the expected value of $\hat{\theta}$ is equal to θ. This is equivalent to saying that the mean of the probability distribution of (or the mean of the sampling distribution of) $\hat{\theta}$ is equal to θ.

Efficiency of Estimators

Let the parametric space of θ be Θ, i.e. $\theta \in \Theta$.
Let $\hat{\theta_1}$ and $\hat{\theta_2}$ be the unbiased estimators of θ.
Then $\hat{\theta_1}$ is said to be more efficient than $\hat{\theta_2}$ under the variance criterion for estimating θ when

$$Var(\hat{\theta_1}) \leq Var(\hat{\theta_2}) \; \forall \; \theta \in \Theta$$

and

$$Var(\hat{\theta_1}) < Var(\hat{\theta_2}) \; for \; at \; least \; one \theta \in \Theta$$

Cramer-Rao Lower Bound(CRLB)

Let $X_1, X_2, ..., X_n$ is a random sample from a distribution with f(x;θ) $\in \Theta$ and g(θ) is to be estimated.

CRLB provides a lower bound for the variance of any unbaised estimator of g(θ).

Consistency of Estimators

For a good estimator, as the sample size increases, the values of the estimator should get closer to the parameter being estimated.
This property of estimators is referred to as consistency.

Sufficiency of Estimators

Definition

Let $X_1, X_2, ..., X_n$ be a random sample from a probability density function (or probability mass function) $f_x(x,\theta)$.
A statistic T is said to be sufficient for θ if the conditional distribution of $X_1, X_2, ..., X_n$ given T = t is independent of θ.
Note: This method does not give a constructive way to find out a sufficient statistic. It can only verify if a statistic is sufficient or not.

Neyman-Fisher Factorization Theorem

Let $X_1, X_2, ..., X_n$ be a random sample from a probability density function (or probability mass function) $f_x(x,\theta)$.
A statistic T = $T(x_1, x_2, ..., x_n)$ is said to be sufficient for θ iff the joint density of $X_1, X_2, ..., X_n$ can be factorized as

$$f(x_1, x_2, ..., x_n; \theta) = g(t,\theta)h(x_1, x_2, ..., x_n)$$

where $h(x_1, x_2, ..., x_n)$ is nonnegative and does not involve θ and $g(t,\theta)$ is a nonnegative function of θ which depends on $x_1, x_2, ..., x_n$ only through t, which is a particular value of T.
This theorem holds for discrete random variables too.

Jointly sufficient

If $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_q)$' then T = $(T_1, T_2, ..., T_k)$ k not necessarily equal to q) is jointly sufficient for $\underline{\theta}$ if

$$f(x_1, x_2, ..., x_n; \underline{\theta}) = g(t, \underline{\theta})h(x_1, x_2, ..., x_n)$$

where $h(x_1, x_2, ..., x_n)$ is nonnegative and does not involve $\underline{\theta}$ and $g(t,\theta)$ is a nonnegative function of $\underline{\theta}$ which depends on $x_1, x_2, ..., x_n$ only through t, which is a particular value of T.

Minimal sufficient

When we ask for sufficient statistics, we generally mean the minimal sufficient statistics. (Otherwise the whole sample itself can also be sufficient.)
By Neyman-Fisher Factorization Theorem, the statistics which we obtain is, in general, the minimal sufficient statistic

Method of Moments Estimators

Let $X_1, X_2, ...X_n$ be a random sample from either a probability mass function or a probability density function with p unknown parameters $\theta_1, \theta_2, ..., \theta_p$.
The moment estimators are found by

- equating the first p population moments to the first p sample moments and
- solving the resulting equations for the unknown parameters.

The resultant estimators are called as Method of Moments (MoM) estimators.

Method of Maximum Likelihood

The maximum likelihood estimator (MLE) of θ is the value of θ that maximizes the likelihood function $L(\theta; x_1, x_2, ..., x_n)$.
In the discrete case, the maximum likelihood estimator is an estimator that maximizes the probability of occurrence of the sample values
$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$
$\hat{theta}$ is the MLE of θ if $L(\hat{\theta}; x_1, x_2, ..., x_n). \geq L(\theta; x_1, x_2, ..., x_n). \forall \theta \in \Theta$.

Rao Blackwell Theorem

Rao Blackwell Theorem helps in obtaining a minimum variance unbiased estimator of a parameter.
Let $X_1, X_2, ...X_n$ is a random sample from a distribution with $f(x;\theta)$, $\theta \in \Theta$. Suppose $\delta(X)$ is an unbiased estimator of $g(\theta)$ and T is a sufficient statistic for θ. Define

$$\eta(T) = E(\delta(X)|T = t]$$

Since T is sufficient statistic, it is independent of θ.
So $\eta(T)$ is a statistic.

$\eta(T)$: Rao Blackwellised version of $\delta(X)$.

Confidence Intervals

An interval estimate for a population parameter is called a confidence interval.
The length of the interval reflects the uncertainty about μ.

Pivotal Quantity Method to Derive a Confidence Interval:

A general method for finding a confidence interval for an unknown parameter is as follows:

1. Let $X_1, X_2, ...X_n$ be a random sample of n observations.
2. Suppose we can find a statistic g$(X_1, X_2, ...X_n; \theta)$ such that
    - g$(X_1, X_2, ...X_n; \theta)$ depends on both the sample and θ but
    - the probability distribution of g$(X_1, X_2, ...X_n; \theta)$ does not depend on $\theta or any other unknown par$

Such g$(X_1, X_2, ...X_n; \theta)$ is called as pivotal quantity .
Now one can compute confidence intervals of the form $\hat{\theta_L}$ and $\hat{\theta_U}$ so that

$$P_\theta[\hat{\theta_L}(X) \leq \theta \leq \hat{\theta_U}(X)] \geq 1 - \alpha$$

SAMPLING THEORY

Sampling Unit

An element or a group of elements on which observations can be taken is called a sampling unit.
The objective of the survey helps in determining the definition of sampling unit.

Sampling Frame

List of all the units of the population to be surveyed constitutes the sampling frame.
All the sampling units in the sampling frame have identification particulars.
For example, all the students in a particular university listed along with their roll numbers constitutes the sampling frame.

Simple Random Sampling

Simple random sampling (SRS) is a method of selection of a sample comprising of n number of sampling units from the population having N number of units such that every sampling unit has an equal chance of being chosen.

Simple Random Sampling Without Replacement

SRSWOR

The sampling units are chosen without replacement in the sense that the units once chosen are not placed back in the population.
SRSWOR is a method of selection of n units out of the N units one by one such that at any stage of selection, any one of the remaining units have the same chance of being selected, i.e. $\frac{1}{N}$.

Simple Random Sampling With Replacement

SRSWR

The sampling units are chosen with replacement in the sense that the chosen units are placed back in the population.

SRSWR is a method of selection of n units out of the N units one by one such that at each stage of selection, each unit has an equal chance of being selected, i.e., $\frac{1}{N}$.

Sample Mean

Population mean is generally measured by arithmetic mean (or weighted arithmetic mean). Let us consider the sample arithmetic mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ as an estimator of population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$
Estimate population mean $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ by sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
$\bar{y}$ is an unbiased estimator of $\bar{Y}$ under SRSWR and SRSwOR cases.

$$E(\bar{y}) = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{Y}$$

Sample Varience

Variance of sample mean under SRSWOR

$$V(\bar{y_{WOR}}) = E(\bar{y} - \bar{Y})^2 = \frac{N-n}{Nn}S^2$$

Variance of sample mean under SRSWR

$$V(\bar{y_{WR}}) = E(\bar{y} - \bar{Y})^2 = \frac{N-1}{Nn}S^2$$

Standard Deviation of Sample Mean

$$\bar{y} = +\sqrt{Var(\bar{Y_{WOR}})} \; or \; +\sqrt{Var(\bar{Y_{WR}})}$$

Sampling for Proportions and Percentages

In many situations, the characteristic under study on which the observations are collected are qualitative in nature.

Sampling Procedure

Follow the same sampling procedures used in case of quantitative characteristics for drawing a sample for qualitative characteristic.

SRSWOR and SRSWR procedures for drawing the samples remains the same for qualitative and quantitative characteristics.

Mean

Estimate population proportion by sample mean

$$\bar{y} = p = \sum_{i=1}^{n} \frac{y_i}{n}$$

Variance

The variance of p under SRSWOR and SRSWR are

$$Var_{WOR}(p) = \frac{N-n}{N-1}\frac{PQ}{n}$$

$$Var_{WR}(p) = \frac{PQ}{n}$$

respectively.

Stratified Sampling

Important objective: Obtain an estimator of a population parameter which can take care of all salient features of the population.
If the population is heterogeneous with respect to the characteristic under study, then the sampling procedure used is stratified sampling.

Bootstrap Methodology

We are interested in finding the statistical properties of the estimators such as the expressions for measures of accuracy.
Deriving the variance or standard error of these statistics is difficult.
Asymptotic theory can be used to derive them but they are not available for small samples.
A modern alternative to the traditional approach is the bootstrapping method.
Bootstrap is a powerful, computer-based resampling method for statistical inference without relying on too many assumption.
Bootstrap is a resampling method.
Simple random sampling with replacement (SRSWR) is used.

Samples are drawn independently by SRSWR from an existing sample data of the same sample size n, and drawing inferences from these resampled data.

Simple linear regression model

Consider a simple linear regression model

$$y = \beta_0 + \beta_1 X + \epsilon$$

y : Dependent or study variable
X : Independent or explanatory variable.
β_0 : Intercept term
β_1 : Slope parameter.
ϵ : Unobservable error component.It accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of y.

OLSE

The ordinary least squares estimates (OLSE) of β_0 and β_1 are denoted as b_0 and b_1 respectively.
$b_0 = \bar{y} - b_1\bar{x}$
$b_1 = \frac{s_{xy}}{s_{xx}}$
where
$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

Maximum Likelihood Property

Assume ϵ_i's(i = 1,2,...n) are independent and identically distributed following a normal distribution N(0,σ^2).
The maximum likelihood estimates of the parameters β_0, β_1 and σ^2 of the linear regression model are

1. $\bar{b_0} = \bar{y} - \bar{b_1}\bar{x}$: same as OLSE
2. $\bar{b_1} = \frac{s_{xy}}{s_{xx}}$: same as OLSE
3. $\bar{s^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{b_0} - \bar{b_1}\bar{x}}{)}$: Different from OLSE

Multiple linear regression model

When we consider the regression modeling between the dependent and more than one independent variables, then the linear model is termed as the multiple linear regression model.
Let y denotes the dependent (or study) variable that is linearly related to k independent (or explanatory) variables $X_1, X_2, ..., X_k$ through the parameters $\beta_1, \beta_2, ..., \beta_k$ and we write

$$y = X_1\beta_1 + X_2\beta_2 + ,,, + X_k\beta_k + \epsilon$$

This is called as the multiple linear regression model.

- The parameters $\beta_1, \beta_2, ..., \beta_k$ are the regression coefficients associated with $X_1, X_2, ..., X_k$ respectively and
- ϵ is the random error component reflecting the difference between the observed and fitted linear relationship.

Principle of OLSE

The ordinary least squares (OLS) estimator of β is

$$b = (X'X)^{-1}X'y.$$

The estimator of σ^2 is

$$\hat{\sigma^2} = \frac{(y - Xb)'(y - Xb)}{n - k}$$

Maximum Likelihood Estimation

In the model $y = X\beta + \epsilon$, it is assumed that the errors are normally and independently distributed with constant variance i.e., $\epsilon \sim N(0, \sigma^2 I)$.
The likelihood function is the joint density of $\epsilon_1, \epsilon_2, ..., \epsilon_n$.
Maximizing the log likelihood
The maximun likelihood estimators (mle) of β and σ^2 are obtained by equating the first order derivatives of $ln\ L(\beta, \sigma^2)$ w.r.t β and σ^2 to zero.
Note: OLSE and mle of β are the same. So mle of β is also an unbaised estimator of β also mle of σ^2 is biased estimator of σ^2.

SQL

SQL stands for **Structured Query Language**. It is a database computer language designed for the retrieval and management of data in a relational database. SQL became a standard of the American National Standards Institute (ANSI) in 1986, and of the International Organization for Standardization (ISO) in 1987.

Why learn SQL?

SQL is the standard language for Relational Database System. All the Relational Database Management Systems (RDMS) like MySQL, MS Access, Oracle, Sybase, Informix, Postgres and SQL Server use SQL as their standard database language.
SQL is widely popular because it offers the following advantages -

- Allows users to access data in the relational database management systems
- Allows users to define the data in a database and manipulate that data
- Allows to embed within other languages using SQL modules, libraries & pre-compilers.

What is RDBMS?

RDBMS stands for Relational Database Management System.
RDBMS is the basis for SQL, and for all modern database systems such as MS SQL Server, IBM DB2, Oracle, MySQL, and Microsoft Access.
The data in RDBMS is stored in database objects called tables. A table is a collection of related data entries and it consists of columns and rows.

SQL Syntax

Most of the actions you need to perform on a database are done with SQL statements.
The following SQL statement selects all the records in a table:

```
SELECT * FROM table_name;
```

The most important point to be noted here is that SQL is *case insensitive*, which means SELECT and select have same meaning in SQL statements.
Some database systems require a semicolon at the end of each SQL statement.

Semicolon is the standard way to separate each SQL statement in database systems that allow more than one SQL statement to be executed in the same call to the server.

SQL Expressions

An expression is a combination of one or more values, operators and SQL functions that evaluate to a value. These SQL EXPRESSIONs are like formulae and they are written in query language. You can also use them to query the database for a specific set of data.
There are different types of SQL expressions, which are -

- Boolean
- Numeric
- Date

Basic SQL Statements and Syntax

CREATE DATABASE

The SQL **CREATE DATABASE** statement is used to create a new SQL database.
Syntax:

```
1    CREATE DATABASE DatabaseName;
2
```

DROP DATABASE

The SQL **DROP DATABASE** statement is used to drop an existing database in SQL schema.
Syntax:

```
1    DROP DATABASE DatabaseName;
2
```

SELECT/USE DATABASE

When you have multiple databases in your SQL Schema, then before starting your operation, you would need to select a database where all the operations would be performed.
The SQL **USE** statement is used to select any existing database in the SQL schema.
Syntax:

```
USE DatabaseName;
```

CREATE TABLE

Creating a basic table involves naming the table and defining its columns and each column's data type.
The SQL **CREATE TABLE** statement is used to create a new table.
Basic syntax:

```
CREATE TABLE table_name(
CREATE TABLE table_name(
   column1 datatype,
   column2 datatype,
   column3 datatype,
   .....
   columnN datatype,
   PRIMARY KEY( one or more columns )
);
```

DROP TABLE

The SQL **DROP TABLE** statement is used to remove a table definition and all the data, indexes, triggers, constraints and permission specifications for that table.
NOTE : You should be very careful while using this command because once a table is deleted then all the information available in that table will also be lost forever.
Syntax:

```
DROP TABLE table_name;
```

INSERT Query

The SQL **INSERT INTO** Statement is used to add new rows of data to a table in the database.
Syntax:

```
INSERT INTO TABLE_NAME (column1, column2, column3,...columnN)
VALUES (value1, value2, value3,...valueN);
```

Here, column1, column2, column3,...columnN are the names of the columns in the table into which you want to insert the data.
You may not need to specify the column(s) name in the SQL query if you are adding values for all the columns of the table. But make sure the order of the values is in the same order as the columns in the table.

SELECT Query

The SQL **SELECT** statement is used to fetch the data from a database table which returns this data in the form of a result table. These result tables are called result-sets.
Syntax:

```
    SELECT column1, column2, columnN FROM table_name;

    /*If you want to fetch all the fields available in the field, then
   you can use the following syntax.*/

    SELECT * FROM table_name;

```

WHERE Clause

The SQL **WHERE** clause is used to specify a condition while fetching the data from a single table or by joining with multiple tables. If the given condition is satisfied, then only it returns a specific value from the table. You should use the WHERE clause to filter the records and fetching only the necessary records.
The WHERE clause is not only used in the SELECT statement, but it is also used in the UPDATE, DELETE statement, etc.
syntax:

```
    SELECT column1, column2, columnN
    FROM table_name
    WHERE [condition]

```

AND and OR Clause

The SQL **AND** & **OR** operators are used to combine multiple conditions to narrow data in an SQL statement. These two operators are called as the conjunctive operators.
These operators provide a means to make multiple comparisons with different operators in

the same SQL statement.
Syntax for **AND** clause:

```
    SELECT column1, column2, columnN
    FROM table_name
    WHERE [condition1] AND [condition2]...AND [conditionN];

```

Syntax for **OR** clause:

```
    SELECT column1, column2, columnN
    FROM table_name
    WHERE [condition1] OR [condition2]...OR [conditionN]

```

UPDATE Query

The SQL **UPDATE** Query is used to modify the existing records in a table.
You can use the WHERE clause with the UPDATE query to update the selected rows, otherwise all the rows would be affected.
Syntax:

```
    UPDATE table_name
    SET column1 = value1, column2 = value2...., columnN = valueN
    WHERE [condition];

```

DELETE Query

The SQL **DELETE** Query is used to delete the existing records from a table.
You can use the WHERE clause with a DELETE query to delete the selected rows, otherwise all the records would be deleted.
Syntax:

```
    DELETE FROM table_name
    WHERE [condition];

```

LIKE Clause

The SQL **LIKE** clause is used to compare a value to similar values using wildcard operators.
There are two wildcards used in conjunction with the LIKE operator.

The percent sign (%) The percent sign represents zero, one or multiple characters

The underscore (_) The underscore represents a single number or character

Syntax:

```
    SELECT FROM table_name
    WHERE column LIKE '%XXXX%'

    or

    SELECT FROM table_name
    WHERE column LIKE 'XXXX_'

    or

    SELECT FROM table_name
    WHERE column LIKE '_XXXX%'

```

TOP Clause

The SQL **TOP** clause is used to fetch a TOP N number or X percent records from a table.
Syntax:

```
    SELECT TOP number|percent column_name(s)
    FROM table_name
    WHERE [condition]

```

ORDER BY Clause

The SQL **ORDER BY** clause is used to sort the data in ascending or descending order, based on one or more columns. Some databases sort the query results in an ascending order by default.
Syntax:

```
    SELECT column-list
    FROM table_name
    [WHERE condition]
    [ORDER BY column1, column2, .. columnN] [ASC | DESC];

```

GROUP BY Clause

The SQL **GROUP BY** clause is used in collaboration with the SELECT statement to arrange identical data into groups. This GROUP BY clause follows the WHERE clause in a SELECT statement and precedes the ORDER BY clause.
Syntax:

```
    SELECT column1, column2
    FROM table_name
    WHERE [ conditions ]
    GROUP BY column1, column2
    ORDER BY column1, column2

```

DISTINCT Keyword

The SQL **DISTINCT** keyword is used in conjunction with the SELECT statement to eliminate all the duplicate records and fetching only unique records.
Syntax:

```
    SELECT DISTINCT column1, column2,.....columnN
    FROM table_name
    WHERE [condition]

```

PYTHON LIBRARIES USED IN DATA MINING

NUMPY

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

Why is Numpy so important?

There are several reasons why NumPy stands out while evaluating complex mathematical equations on matrices:

- NumPy integrates C/C++ and Fortran codes in Python which execute faster as compared to Python.
- Python list is a heterogeneous collection of elements whereas a Numpy array is a homogeneous collection of elements stored in contiguous memory locations which results in faster access and execution.
- Performing simple arithmetic operations is way easier using Numpy array as compared to python lists.
- Parallel processing of sub-tasks of a huge task resulting in superfast execution with large data arrays.

Some Important Operations on Numpy arrays include:

- generating arrays with random numbers
- generating evenly spaced ndarrays
- reshaping of Numpy arrays
- flattening of a Numpy array
- transpose of a Numpy array
- expanding and squeezing of a numpy array
- slicing and negative slicing of Numpy arrays

- stacking and concatenating ndarrays
- broadcasting in Numpy arrays.

NumPy package also contains a Matrix library **numpy.matlib**. This module has functions that return matrices instead of ndarray objects.
It also contains **numpy.linalg** module that provides all the functionality required for linear algebra.

SciPy

SciPy, a scientific library for Python is an open source, BSD-licensed library for mathematics, science and engineering. The SciPy library depends on NumPy, which provides convenient and fast N-dimensional array manipulation. The main reason for building the SciPy library is that, it should work with NumPy arrays. It provides many user[-]friendly and efficient numerical practices such as routines for numerical integration and optimization.
By default, all the NumPy functions have been available through the SciPy namespace. There is no need to import the NumPy functions explicitly, when SciPy is imported.

Some SciPy Packages

SciPy-Cluster

K-means clustering is a method for finding clusters and cluster centers in a set of unlabelled data. Intuitively, we might think of a cluster as - comprising of a group of data points, whose inter-point distances are small compared with the distances to points outside of the cluster. Given an initial set of K centers, the K-means algorithm iterates the following two steps

- For each center, the subset of training points (its cluster) that is closer to it is identified than any other center.
- The mean of each feature for the data points in each cluster are computed, and this mean vector becomes the new center for that cluster.

These two steps are iterated until the centers no longer move or the assignments no longer change. Then, a new point x can be assigned to the cluster of the closest prototype. The SciPy library provides a good implementation of the K-Means algorithm through the cluster package.

SciPy-FFTpack

Fourier Transformation is computed on a time domain signal to check its behavior in the frequency domain. Fourier transformation finds its application in disciplines such as signal and noise processing, image processing, audio signal processing, etc. SciPy offers the fftpack module, which lets the user compute fast Fourier transforms.

SciPy-Ndimage

The SciPy ndimage submodule is dedicated to image processing. Here, ndimage means an n-dimensional image.
Some of the most common tasks in image processing are as follows

- Input/Output, displaying images
- Basic manipulations - Cropping, flipping, rotating, etc.
- Image filtering - De-noising, sharpening, etc.
- Image segmentation - Labeling pixels corresponding to different objects
- Classification.

SciPy-Optimize

The **scipy.optimize** package provides several commonly used optimization algorithms. This module contains the following aspects

- Unconstrained and constrained minimization of multivariate scalar functions (minimize()) using a variety of algorithms
- Global (brute-force) optimization routines (e.g., anneal(), basinhopping())
- Least-squares minimization (leastsq()) and curve fitting (curve_fit()) algorithms
- Scalar univariate functions minimizers (minimize_scalar()) and root finders (newton())

PANDAS

Pandas is a Python library used for working with data sets.
It has functions for analyzing, cleaning, exploring, and manipulating data.

Pandas allows us to analyze big data and make conclusions based on statistical theories.
Pandas can clean messy data sets, and make them readable and relevant.

Pandas Series

A Pandas Series is like a column in a table.
It is a one-dimensional array holding data of any type.

Pandas DataFrame

A Pandas DataFrame is a 2 dimensional data structure, like a 2 dimensional array, or a table with rows and columns.

MACHINE LEARNING ALGORITHMS

SUPERVISED MACHINE LEARNING

Supervised Machine Learning is the task of learning a function which maps an input to an output based on example input-output pairs. A supervised Machine learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels or predict the values for unseen instances.
It is further divided into 2 types:

Classification: takes real value input and produces a discrete output

Regression takes real value as input and produces a continuous output

Linear Regression

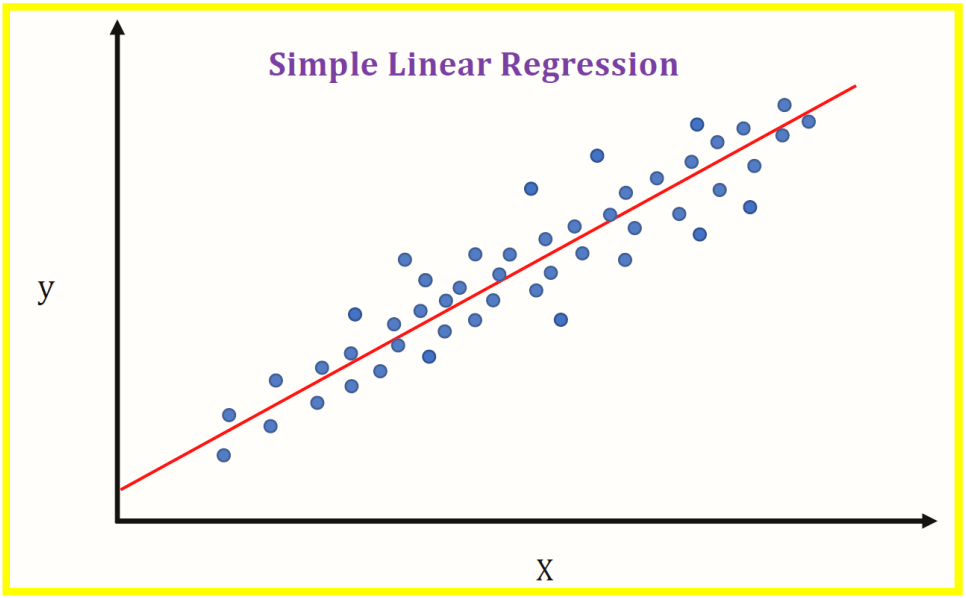

Figure 1: Simple Linear Regression

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called simple linear regression; for more than one, the process is called multiple linear regression.

Linear Regression is the first type of regression algorithm to be studied rigorously, and to be used extensively in practical applications as the models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters.

Logistic Regression

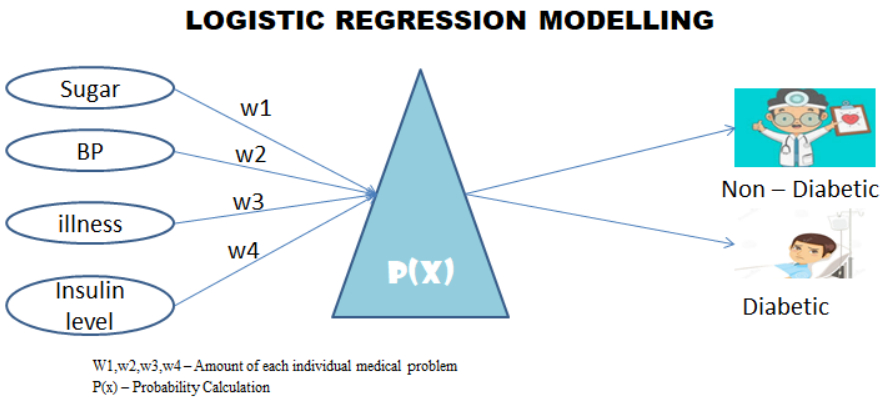

Figure 2: Logistic Regression Modelling

In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

It is named after the logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

Decision Tree

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal, its also

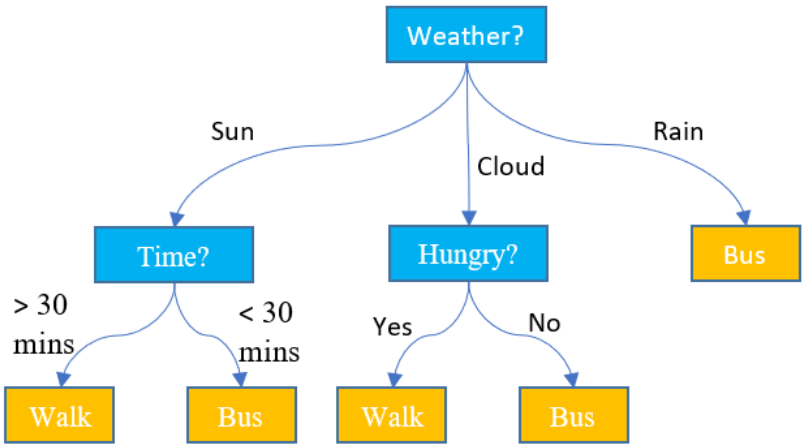

Figure 3: Decision Tree example

widely used in machine learning.

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute. This process is then repeated for the subtree rooted at the new node.
Some Strengths of Decision tree approach are:

- they are able to generate understandable rules.
- perform classification without requiring much computation.
- they are able to handle both continuous and categorical variables.
- they provide a clear indication of which fields are most important for prediction or classification.

Some of its weaknesses include:

- they are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- they are prone to errors in classification problems with many class and relatively small number of training examples.

- they can be computationally expensive to train.

Random Forest

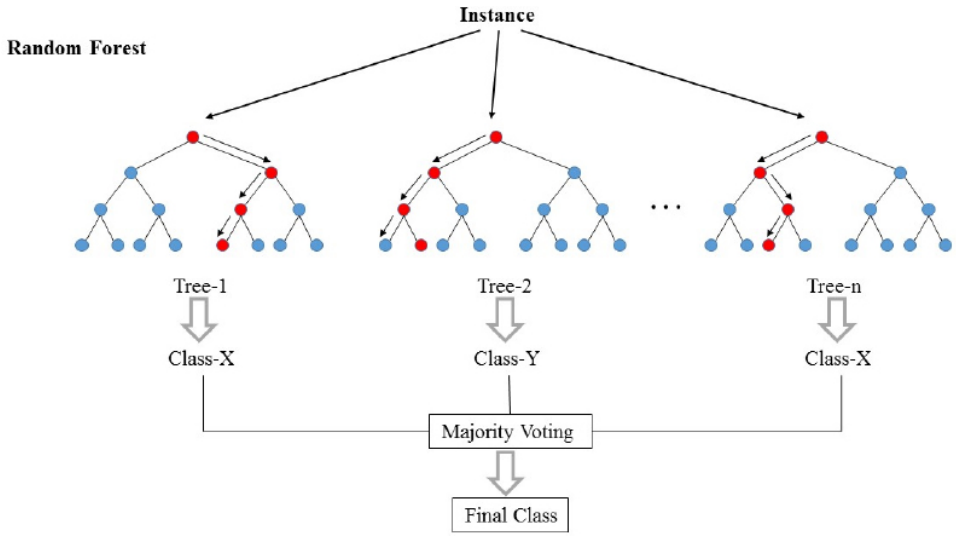

Figure 4: Random Forest Classification

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned.

Random decision forests correct for decision trees habit of overfitting to their training set. They generally outperform decision trees, however data characteristics can affect their performance. It adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

Bayes Classifier

In statistics, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong independence assumptions between the features. They are among the simplest Bayesian network models.

They are among the simplest Bayesian network models. Despite their naive design and apparently oversimplified assumptions, naive Bayes classifiers have worked quite well in many complex real-world situations.
An advantage of naive Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.

K-Nearest Neighbors(KNN)

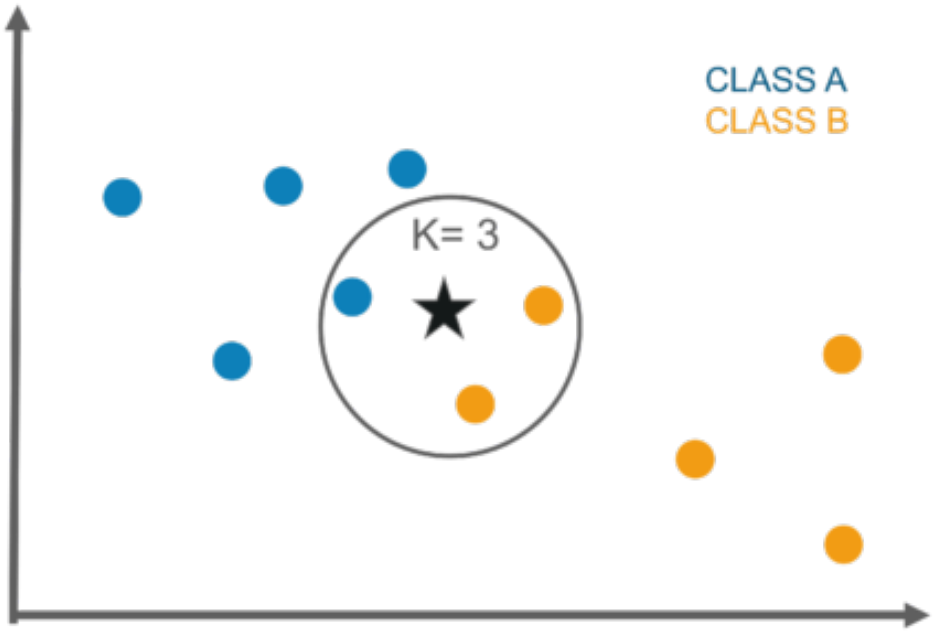

Figure 5: Classifying unknown example with K=3

The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

KNN works by finding the distances between a query and all the examples in the data,

selecting the specified number examples (K) closest to the query, then votes for the most frequent label (in the case of classification) or averages the labels (in the case of regression).

Support Vector Machines(SVM)

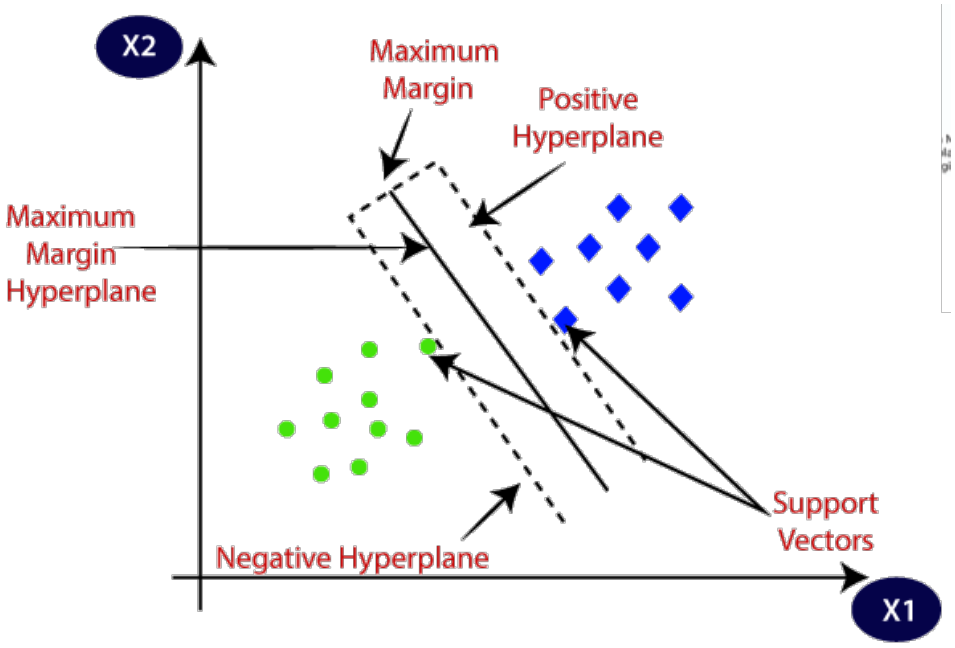

Figure 6: Support Vector Machine example

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes.

Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. If the number of input features has dimension n, then the hyperplane will be a (n-1) dimensional affine plane.

Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line).