UNIT –III

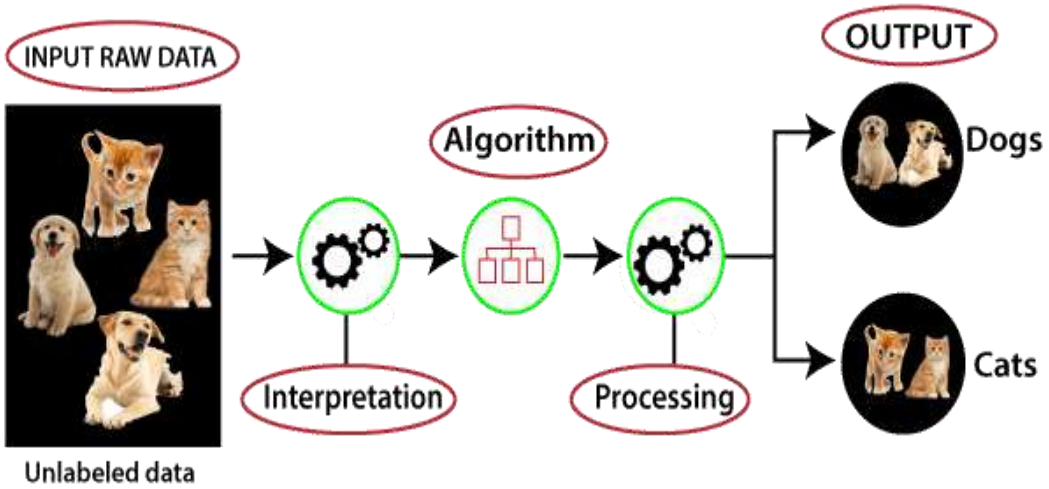
UNSUPERVISED LEARNING:

Unsupervised learning uses machine learning algorithms to analyse and cluster unlabelled data sets. These algorithms discover hidden patterns in data without the need for human intervention.

Unsupervised learning is when it can provide a set of unlabelled data, which it is required to analyse and find patterns inside. The examples are dimension reduction and clustering. The training is supported to the machine with the group of data that has not been labelled, classified, or categorized, and the algorithm required to facilitate on that data without some supervision. The objective of unsupervised learning is to restructure the input record into new features or a set of objects with same patterns.

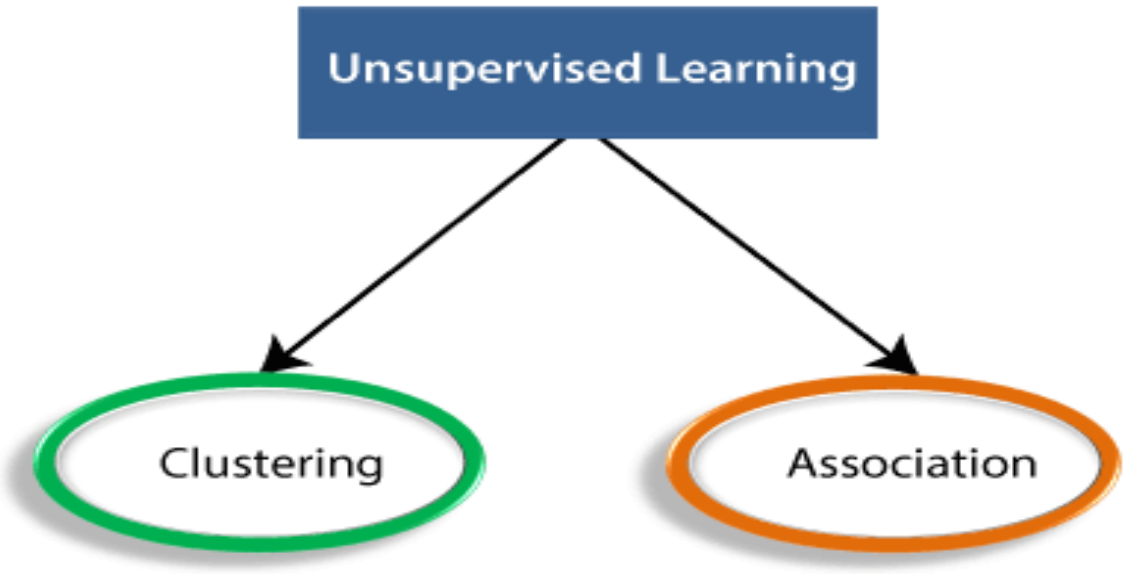
Cluster analysis is used to form groups or clusters of the same records depending on various measures made on these records. The key design is to define the clusters in ways that can be useful for the objective of the analysis. This data has been used in several areas, such as astronomy, archaeology, medicine, chemistry, education, psychology, linguistics, and sociology.

Google is an instances of clustering that needs unsupervised learning to group news items depends on their contents. Google has a set of millions of news items written on multiple topics and their clustering algorithm necessarily groups these news items into a small number that are same or associated to each other by using multiple attributes, including word frequency, sentence length, page count, etc.



TYPES OF UNSUPERVISED LEARNING ALGORITHM:

The unsupervised learning algorithm can be further categorized into two types of problems:



CLUSTERING:

Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.

ASSOCIATION:

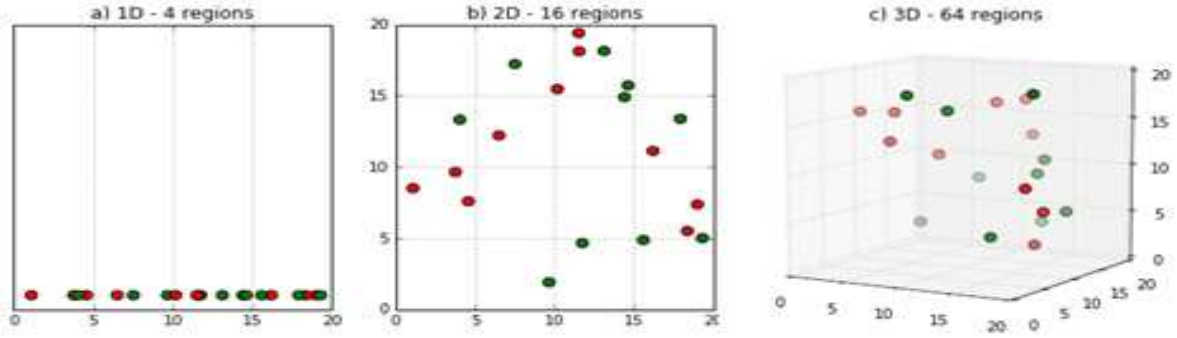
An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

CURSE OF DIMENSIONALITY:

The curse of dimensionality refers to the phenomena that occur when classifying, organizing, and analysing high dimensional data that does not occur in low dimensional spaces, specifically the issue of data sparsity and “closeness” of data.

ISSUES:

Sparsity of data occurs when moving to higher dimensions. The volume of the space represented grows so quickly that the data cannot keep up and thus becomes sparse, as seen below. The sparsity issue is a major one for anyone whose goal has some statistical significance.



As the data space seen above moves from one dimension to two dimensions and finally to three dimensions, the given data fills less and less of the data space. In order to maintain an accurate representation of the space, the data for analysis grows exponentially.

The second issue that arises is related to sorting or classifying the data. In low dimensional spaces, data may seem very similar but the higher the dimension the further these data points may seem to be. The two wind turbines below seem very close to each other in two dimensions but separate when viewed in a third dimension. This is the same effect the curse of dimensionality has on data.

INFINITE FEATURES REQUIRES INFINITE TRAINING:

When neural networks are created they are instantiated with a certain number of features (dimensions). Each datum has individual aspects, each aspect falling somewhere along each dimension. In our fruit example we may want one feature handling colour, one for weight,

one for shape, etc. Each feature adds information, and if we could handle every feature possible we could tell perfectly which fruit we are considering. However, an infinite number of features requires an infinite number of training examples, eliminating the real-world usefulness of our network.

Most disconcerting, the number of training data needed increases exponentially with each added feature. Even if we only had 15 features each being one 'yes' or 'no' question about the piece fruit we are identifying, this would require a training set on the order of 21532,000 training sample.

MITIGATING THE CURSE OF DIMENSIONALITY:

A careful choice of the number of dimensions (features) to be used is the prerogative of the data scientist training the network. In general the smaller the size of the training set, the fewer features she should use. She must keep in mind that each features increases the data set requirement exponentially.

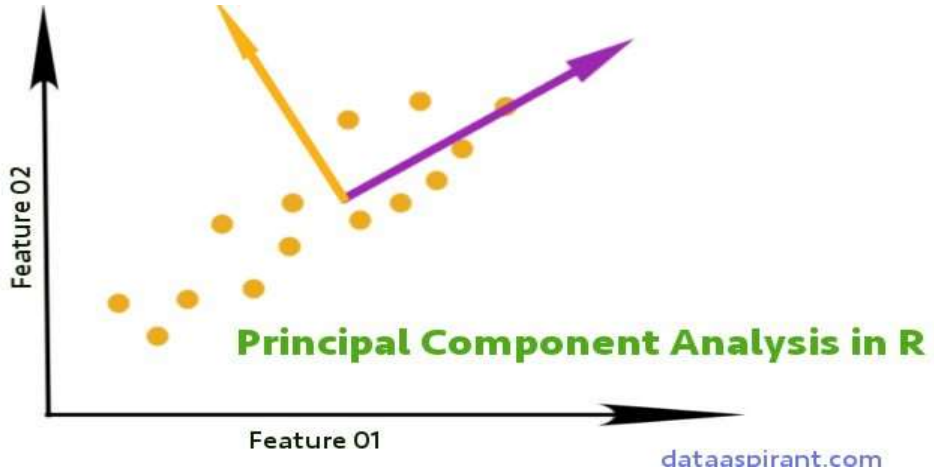
COMPONENT ANALYSIS:

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D.

Principal component analysis (PCA) is a technique for reducing the dimensionality of such datasets, increasing interpretability but at the same time minimizing information loss. It does so by creating new uncorrelated variables that successively maximize variance.

PCA helps you interpret your data, but it will not always find the important patterns. Principal component analysis (PCA) simplifies the complexity in high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features.

PCA is more useful when dealing with 3 or higher-dimensional data. It is always performed on a symmetric correlation or covariance matrix. This means the matrix should be numeric and have standardized data



FACTOR ANALYSIS:

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors.

There are two types of factor analyses, exploratory and confirmatory. Exploratory factor analysis (EFA) is method to explore the underlying structure of a set of observed variables, and is a crucial step in the scale development process.

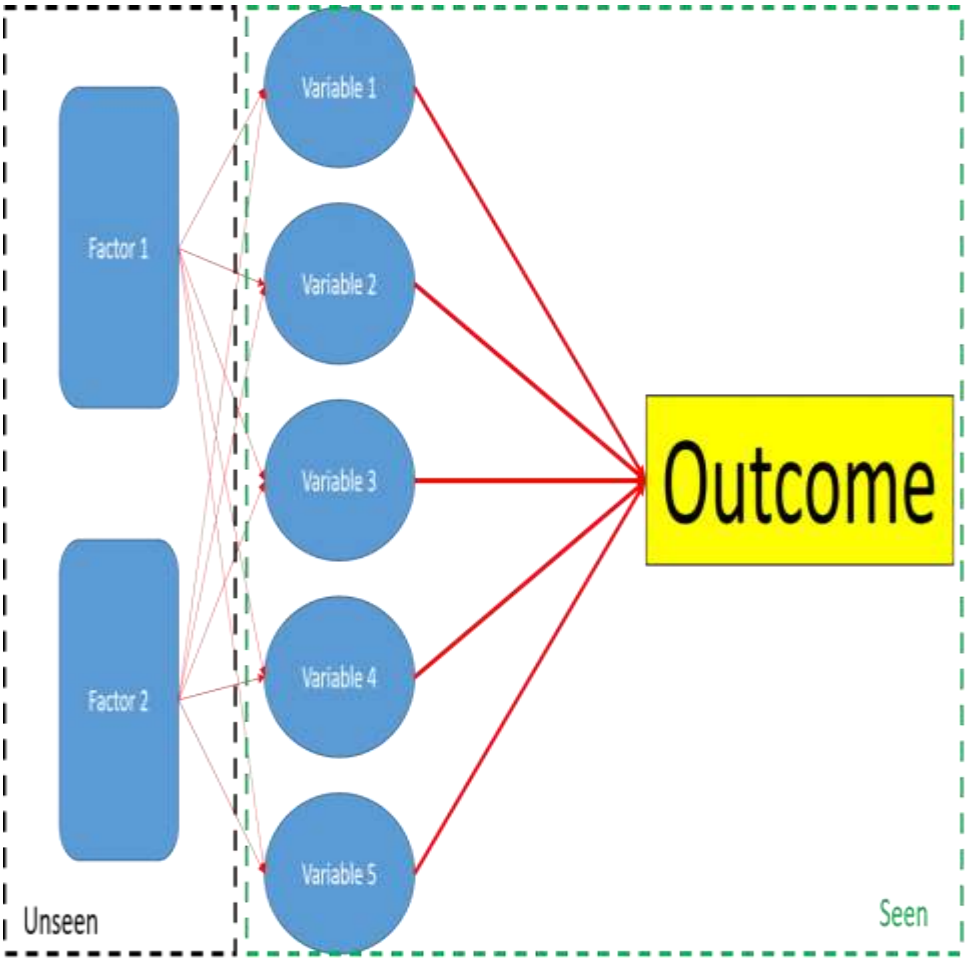
Factor analysis is used to identify "factors" that explain a variety of results on different tests. For example, intelligence research found that people who get a high score on a test of verbal ability are also good on other tests that require.

There are mainly three types of factor analysis that are used for different kinds of market research and analysis.

Exploratory factor analysis.

Confirmatory factor analysis.

Structural equation modelling.



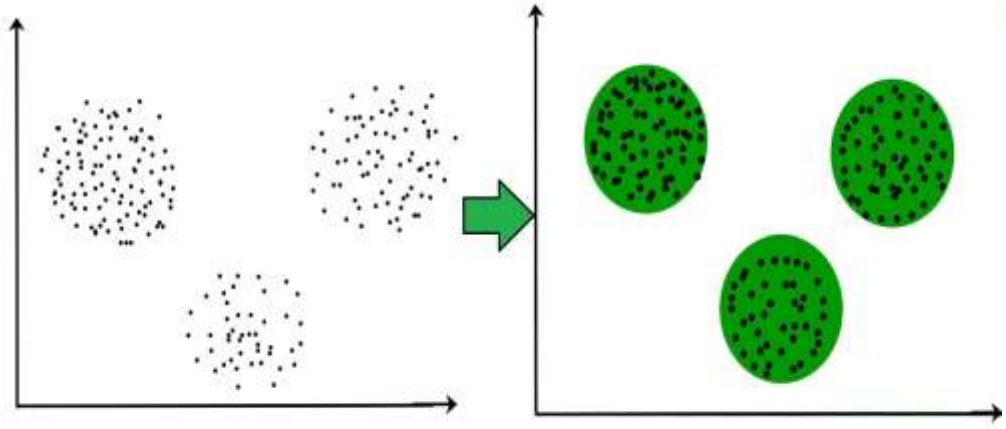
CLUSTERING:

Grouping unlabelled examples is called clustering. As the examples are unlabelled, clustering relies on unsupervised machine learning. If the examples are labelled, then clustering becomes classification.

It is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

For ex– The data points in the graph below clustered together can be classified into one single group**.**



APPLICATIONS OF CLUSTERING IN DIFFERENT FIELDS:

MARKETING:

It can be used to characterize & discover customer segments for marketing purposes**.**

BIOLOGY:

It can be used for classification among different species of plants and animals.

LIBRARIES:

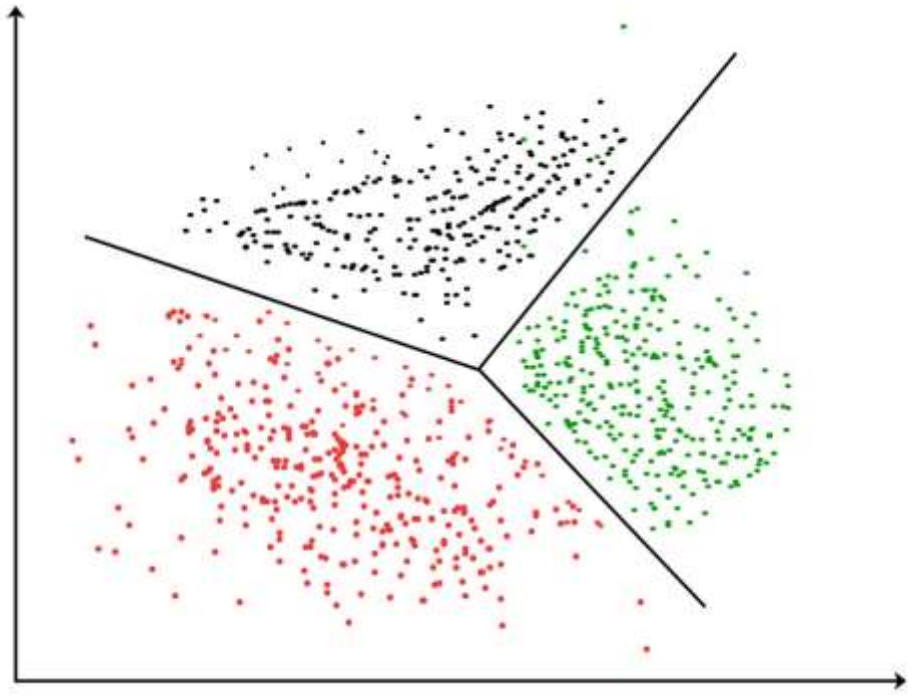
It is used in clustering different books on the basis of topics and information.

INSURANCE:

It is used to acknowledge the customers, their policies and identifying the frauds.

Clustering Algorithms:

K-means clustering algorithm – It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



REGRESSION:

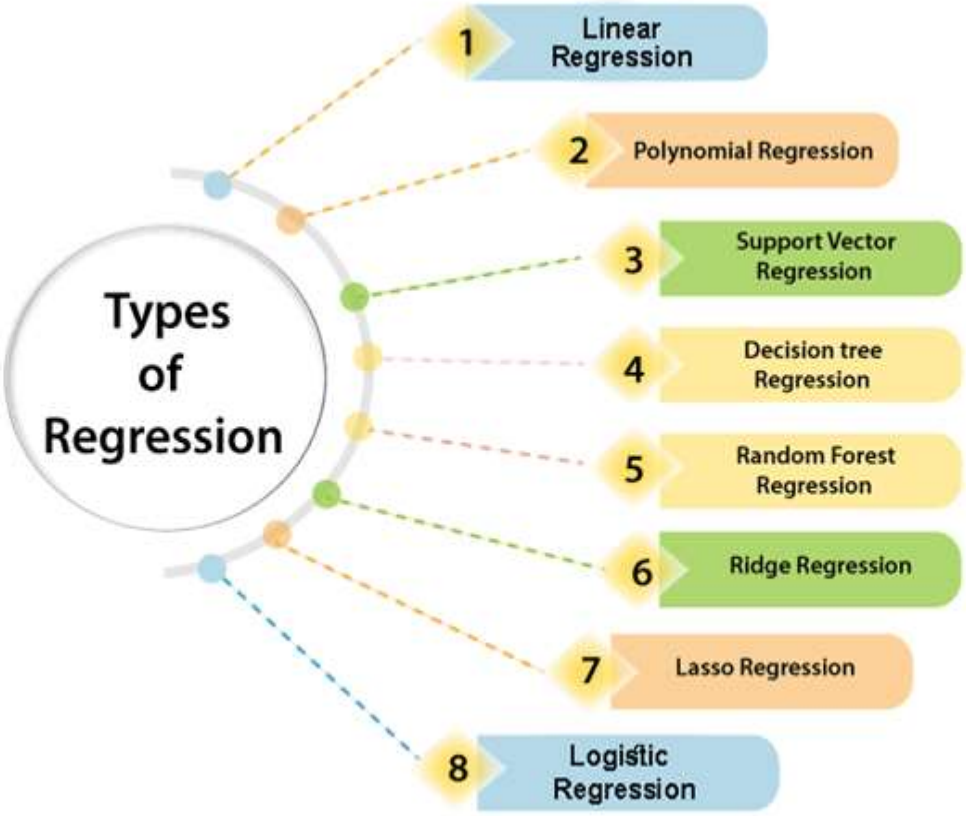
Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Also called simple regression or ordinary least squares (OLS), linear regression is the most common form of this technique. Linear regression establishes the linear relationship between two variables based on a line of best fit. Linear regression is thus graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of one variable when the value of the other is zero. Non-linear regression models also exist, but are far more complex.

Regression analysis is a powerful tool for uncovering the associations between variables observed in data, but cannot easily indicate causation. It is used in several contexts in business, finance, and economics. For instance, it is used to help investment manager's value assets and understand the relationships between factors such as commodity prices and the stocks of businesses dealing in those commodities.

Regression as a statistical technique should not be confused with the concept of regression to the mean (mean reversion).

Formulating a regression analysis helps you predict the effects of the independent variable on the dependent one. Example: we can say that age and height can be described using a linear regression model. Since a person's height increases as age increases, they have a linear relationship.



For example, it can be used to predict the relationship between reckless driving and the total number of road accidents caused by a driver, or, to use a business example, the effect on sales and spending a certain amount of money on advertising. Regression is one of the most common models of machine learning.

Regression is a supervised machine learning technique which is used to predict continuous values. The ultimate goal of the regression algorithm is to plot a best-fit line or a curve between the data. The three main metrics that are used for evaluating the trained regression model are variance, bias and error.

LEAST SQUARES:

The least squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve. Least squares regression is used to predict the behaviour of dependent variables.

LEAST SQUARE METHOD FORMULA:

Suppose when we have to determine the equation of line of best fit for the given data, then we first use the following formula.

The equation of least square line is given by $Y = a + bX$.

Normal equation for 'a':

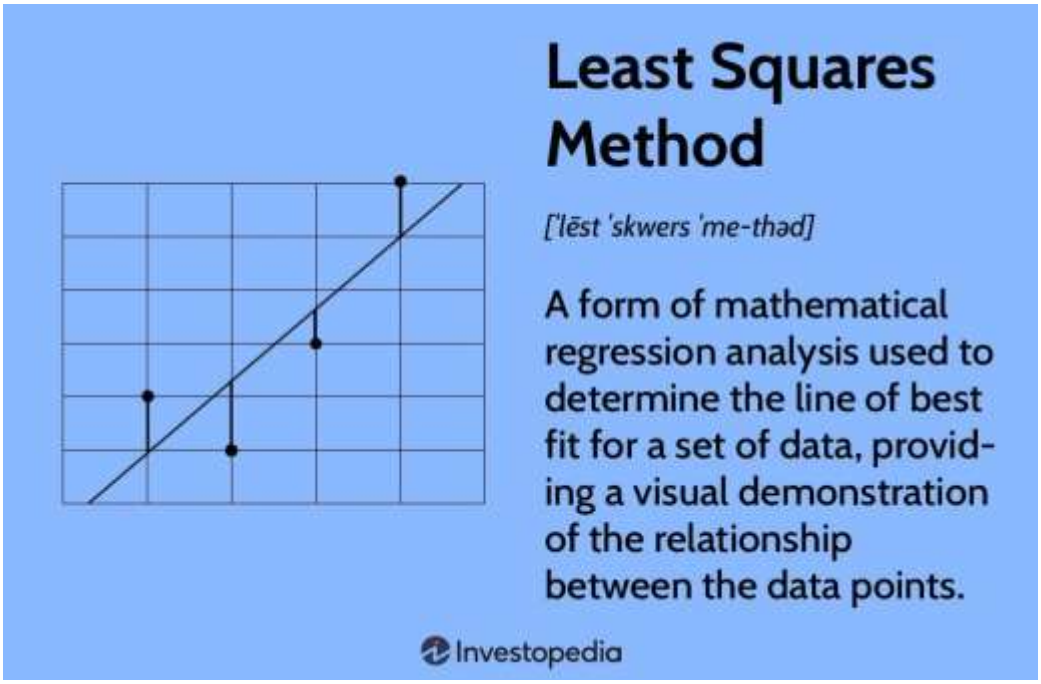
$\sum Y = na + b\sum X$.

Normal equation for 'b':

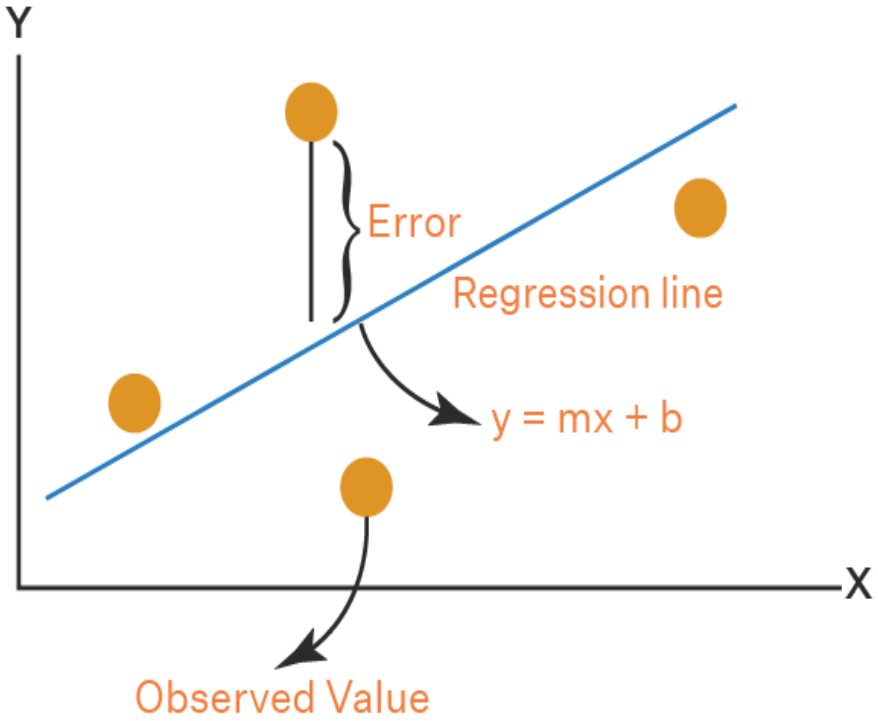
$\sum XY = a\sum X + b\sum X2$

The **least square method** is the process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum of the squares of the offsets (residual part) of the points from the curve. During the process of finding the relation between two variables, the trend of outcomes are estimated quantitatively. This process is termed as **regression analysis**. The method of curve fitting is an approach to regression analysis. This method of fitting equations which approximates the curves to given raw data is the least squares.

It is quite obvious that the fitting of curves for a particular data set are not always unique. Thus, it is required to find a curve having a minimal deviation from all the measured data points. This is known as the best-fitting curve and is found by using the least-squares method.



Least Square Method



COREELATION:

The word correlation is used in everyday life to denote some form of association. We might say that we have noticed a correlation between foggy days and attacks of wheeziness. However, in statistical terms we use correlation to denote association between two quantitative variables.

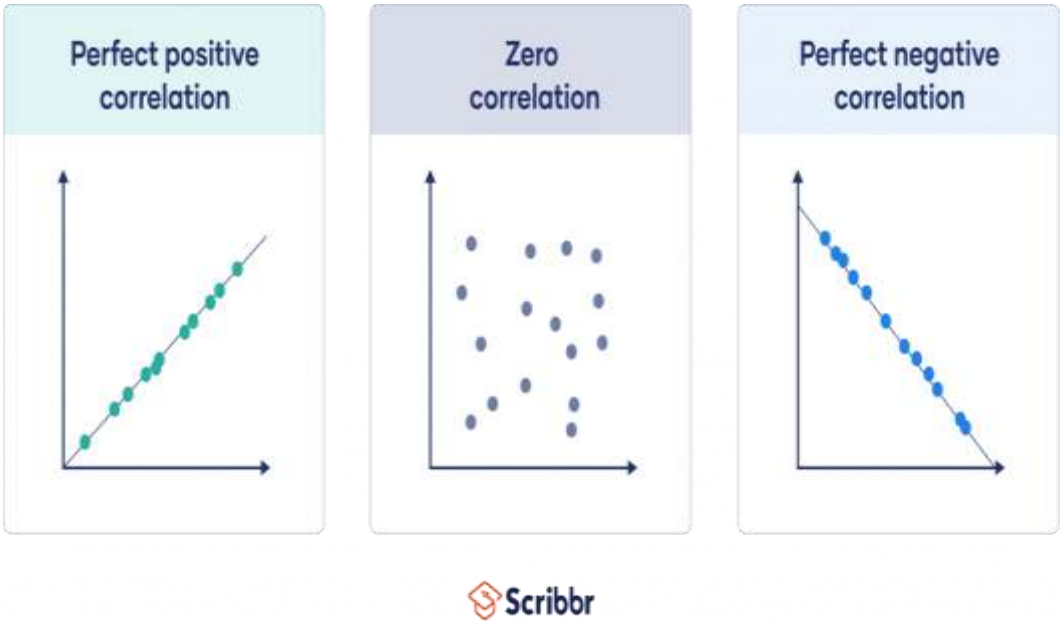
For example, there exists a correlation between two variables X and Y, which means the value of one variable is found to change in one direction, the value of the other variable is found to change either in the same direction (i.e. positive change) or in the opposite direction (i.e. negative change).

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data. Although in the broadest sense, "correlation" may indicate any type of association, in statistics it usually refers to the degree to which a pair of variables are linearly related. Familiar examples of dependent phenomena include the correlation between the height of parents and their offspring, and the correlation between the price of a good and the quantity the consumers are willing to purchase, as it is depicted in the so-called demand curve.

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. In this example, there is a causal relationship, because extreme weather causes people to use more electricity for heating

or cooling. However, in general, the presence of a correlation is not sufficient to infer the presence of a causal relationship (i.e., correlation does not imply causation).

Formally, random variables are dependent if they do not satisfy a mathematical property of probabilistic independence. In informal parlance, correlation is synonymous with dependence. However, when used in a technical sense, correlation refers to any of several specific types of mathematical operations between the tested variables and their respective expected values. Essentially, correlation is the measure of how two or more variables are related to one another.



LINEAR REGRESSION:

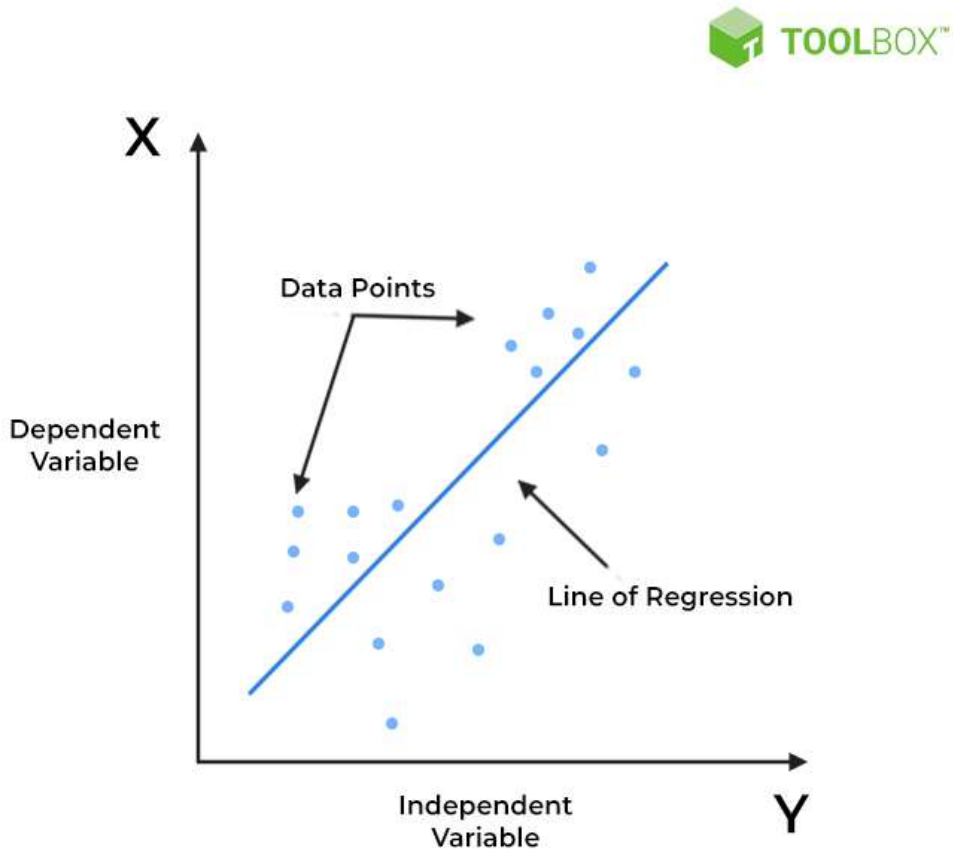
Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

The weight of the person is linearly related to their height. So, this shows a linear relationship between the height and weight of the person. According to this, as we increase the height, the weight of the person will also increase.

THERE ARE TWO KINDS OF LINEAR REGRESSION MODEL:-

Simple Linear Regression: A linear regression model with one independent and one dependent variable. Multiple Linear Regression: A linear regression model with more than one independent variable and one dependent variable.

Regression is a defense mechanism in which people seem to return to an earlier developmental stage. This tends to occur around periods of stress—for example, an overwhelmed child may revert to bedwetting or thumb-sucking. Regression may arise from a desire to reduce anxiety and feel psychologically safe.



LASSO REGRESSION:

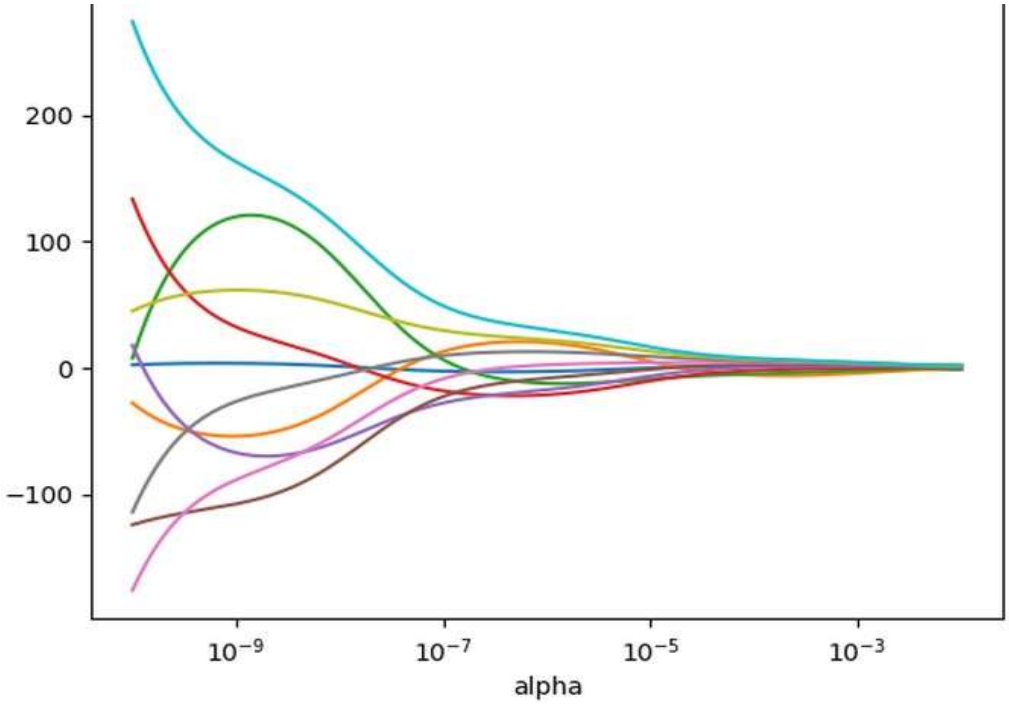
In Lasso regression, discarding a feature will make its coefficient equal to 0. So, the idea of using Lasso regression for feature selection purposes is very simple: we fit a Lasso regression on a scaled version of our dataset and we consider only those features that have a coefficient different from 0.

The LASSO method regularizes model parameters by shrinking the regression coefficients, reducing some of them to zero. The feature selection phase occurs after the shrinkage, where every non-zero value is selected to be used in the model.

Linear regression is a good model for testing feature selection methods as it can perform better if irrelevant features are removed from the model.

Overview. There are three types of feature selection: Wrapper methods (forward, backward, and stepwise selection), Filter methods (ANOVA, Pearson correlation, variance thresholding), and Embedded methods (Lasso, Ridge, Decision Tree).

LASSO involves a penalty factor that determines how many features are retained; using cross-validation to choose the penalty factor helps assure that the model will generalize well to future data samples.



Encoding is the process of using various patterns of voltage or current levels to represent 1s and 0s of the digital signals on the transmission link. The common types of line encoding are Unipolar, Polar, Bipolar, and Manchester.

ENCODING TECHNIQUES:

The data encoding technique is divided into the following types, depending upon the type of data conversion.

Analog data to Analog signals − the modulation techniques such as Amplitude Modulation, Frequency Modulation and Phase Modulation of analog signals, fall under this category.

Analog data to Digital signals − this process can be termed as digitization, which is done by Pulse Code Modulation PCM

. Hence, it is nothing but digital modulation. As we have already discussed, sampling and quantization are the important factors in this. Delta Modulation gives a better output than PCM.

Digital data to Analog signals − the modulation techniques such as Amplitude Shift Keying ASK

, Frequency Shift Keying FSK

, Phase Shift Keying PSK

, etc., fall under this category. These will be discussed in subsequent chapters.

Digital data to Digital signals − these are in this section. There are several ways to map digital data to digital signals.

FILE FORMATS:

A file format is a standard way that information is encoded for storage in a computer file. It specifies how bits are used to encode information in a digital storage medium. File formats may be either proprietary or free.

Some file formats are designed for very particular types of data: PNG files, for example, store bitmapped images using lossless data compression. Other file formats, however, are designed for storage of several different types of data: the Ogg format can act as a container for different types of multimedia including any combination of audio and video, with or without text (such as subtitles), and metadata. A text file can contain any stream of characters, including possible control characters, and is encoded in one of various character encoding schemes. Some file formats, such as HTML, scalable vector graphics, and the source code of computer software are text files with defined syntaxes that allow them to be used for specific purposes.

4 common file formats and how they're shared in Clinked

Word documents (. docx)

Web page images (. png and . jpg)

Adobe acrobat files (. pdf)

Multimedia files (. mp3 and . mp4)