# Module 5
# Machine Learning Algorithms for Big Data Analytics

## *5.1 Introduction*

***Artificial Intelligence* (AI)** refers to the science and engineering of making computers perform tasks, which normally require human intelligence. For example, tasks such as predicting future results, visual perception, speech recognition, decision making and natural language processing.

Two concepts in AI, 'machine learning' and 'deep learning' provide powerful tools for advanced analytics and predictions.

### Machine Learning

Machine Learning (ML) is a field of computer science based on AI which deals with learning from data in three phases, i.e. collect, analyze and predict. It does not rely on explicitly programmed instructions.

An ML program learns the behavior of a process. The program uses data generated from various sources for training. Learning from the outcomes from common inputs improves future performance from previous outcomes. Learning applies in many fields of research and industry. Learning from study of data enables efficient and logical decisions for future actions.

***Deep Learning* (DL)** refers to structured learning (DSL) or hierarchical learning. DL methods are advanced methods, such as artificial neural networks (ANN) such as artificial neural networks (ANN) or neural nets, deep neural networks, deep belief networks and recurrent neural networks. Learning can be unsupervised, semi-supervised or supervised. Applications of DL and ANN include computer vision, speech recognition, Natural Language Processing (NLP), audio recognition, social network filtering, machine translation, bioinformatics and drug design. DL methods give results comparable to and in some cases superior to human experts.





## 5.2 ESTIMATING THE RELATIONSHIPS, OUTLIERS, VARIANCES, PROBABILITY DISTRIBUTIONS AND *CORRELATIONS*

Independent variables represent directly measurable characteristics. For example, year of sales figure or semester of study. Dependent variables represent the characteristics. For example, profit during successive years or grades awarded in successive semesters. Values of a dependent variable depend on the value of the independent Variable.

***Predictor variable*** is an independent variable, which computes a dependent variable using some equation, function or graph, and does a prediction. For example, predicts sales growth of a car model after five years from given input datasets for the sales, or predicts sentiments about higher sales of particular category of toys next year.

*Outcome variable* represents the effect of manipulation(s) using a function, equation or experiment. For example, CGPA (Cumulative Grade Points Average) of the student or share of profit to each shareholder in a year using profit as the dependent variable. CGPA of a student computes from the grades awarded in the semesters for which student completes his/her studies. A company declares the share of profit to each shareholder in a year after subtracting requirements of money for future growth from the profit.

*Explanatory variable* is an independent variable, which explains the behavior of the dependent variable, such as linearity coefficient, non-linear parameters or probabilistic distribution of profit-growth as a function of additional investment in successive years.

*Response variable* is a dependent variable on which a study, experiment or computation focuses. For example, improvement in profits over the years from the investments made in successive years or improvement in class performance is measured from the extra teaching efforts on individual students of a class.

*Feature variable* is a variable representing a characteristic. For example, apple feature red, pink, maroon, yellowish, yellowish green and green. Feature variables are generally represented by text characters. Numbers can also represent features. For example, red with 1, orange with 2, yellow with 3, yellowish green 4 and green 5.

*Categorical variable* is a variable representing a category. For example, car, tractor and truck belong to the same category, i.e., a four-wheeler automobile. Categorical variables are generally represented by text characters.

## Relationships-Using Graphs, Scatter Plots and Charts

A relationship between two or more quantitative dependent variables with respect to an independent variable can be well-depicted using graph, scatter plot or chart with data points, shown in distinct shapes. Conventionally, independent variables are on the x-axis, whereas the dependent variables on the y-axis in a graph. A line graph uses a line on an x-y axis to plot a continuous function.

A scatter plot is a plot in which dots or distinct shapes represent values of the dependent variable at the multiple values of the independent variable Whether two variables are related to each other or not, can be derived from statistical analysis using scatter plots.

## Linear and Non-linear Relationships

A linear relationship exists between two variables, say x and y, when a straight line (y = a0 + a1.x) can fit on a graph, with at least some reasonable degree of accuracy. The a1 is the linearity coefficient. For example, a scatter chart can suggest a linear relationship, which means a straight line. Figure 6.1 shows a scatter plot, which fits a linear relationship between the number of students opting for computer courses in years between 2000 and 2017.

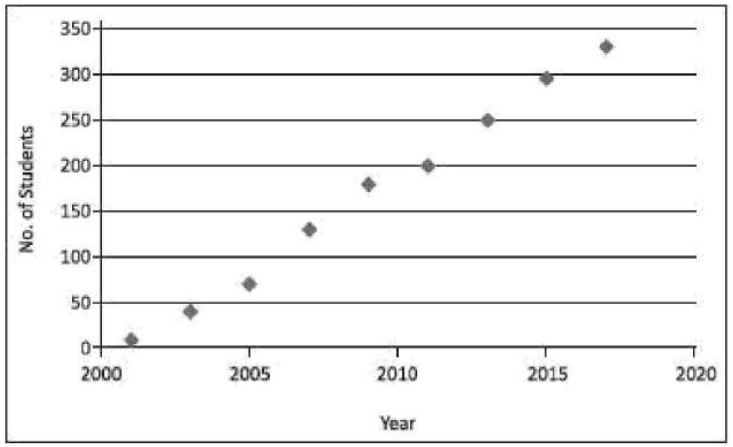



**Figure 5.1 Scatter plot for linear relationship between students opting for computer courses in years between 2000 and 2017**

A linear relationship can be positive or negative. A positive relationship implies if one





variable increases in value, the other also increases in value. A negative relationship, on the other hand, implies when one increases in value, the other decreases in value. Perfect, strong or weak linearship categories depend upon the bonding between the two variables.

A non-linear relationship is said to exist between two quantitative variables when a curve (y = a0 + a1.x + a2.x2 + •••) can be used to fit the data points. The fit should be with at least some reasonable degree of accuracy for the fitted parameters, a0, a1, a2 ••• Expression for y then generally predicts the values of one quantitative variable from the values of the other quantitative variable with considerably more accuracy than a straight line.

Consider an example of non-linear relationship: The side of a square and its area are not linear. In fact, they have quadratic relationship. If the side of a square doubles, then its area increases four times. The relationship predicts the area from the side.

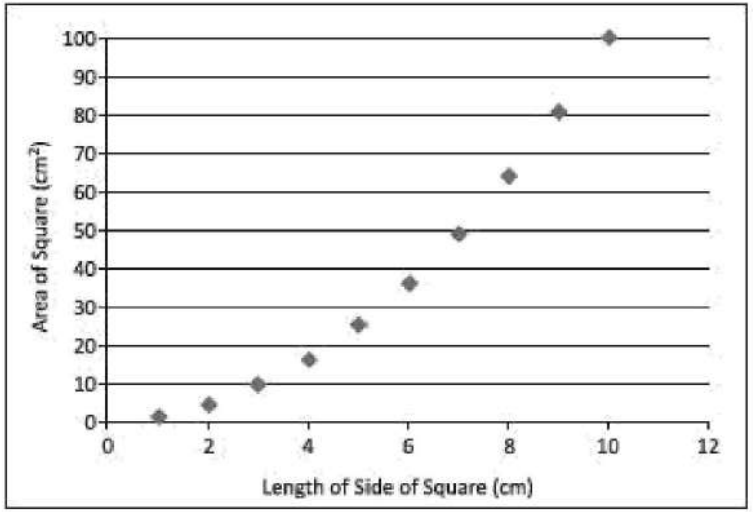



**Figure 5.2 scatter plot in case of a non-linear relationship between side of square and its area**

### Estimating Relationship

Estimating the relationships means finding a mathematical expression, which gives the value of the variable according to its relationship with other variables. For example, assume Ym = sales of a car model m in xth year of the start of manufacturing that model.

### Outliers

Outliers are data points that are numerically far distant from the rest of the points in a dataset, are termed as outliers. Outliers show significant variations from the rest of the points. Identification of outliers is important to improve data quality or to detect an anomaly

There are several reasons for the presence of outliers in relationships. Some of these are:

- Anomalous situation
- Presence of a previously unknown fact
- Human error (errors due to data entry or data collection)
- Participants intentionally reporting incorrect data (This is common in self-reported measures and measures that involve sensitive data which participant doesn't want to disclose)
- Sampling error (when an unfitted sample is collected from population).

Population means any group of data, which includes all the data of interest. For example, when analyzing 1000 students who gave an examination in a computer course, then the population is 1000. 100 games of chess will represent the population in analysis of 100 games of chess of a grandmaster.

Sample means a subset of the population. Sample represents the population for uses, such as analysis and consists of randomly selected data.

### Variance

Variance measures by the sum of squares of the difference in values of a variable with respect to the expected value. Variance can alternatively be a sum of squares of the difference with respect to value at an origin. Variance indicates how widely data points in a dataset vary. If data points vary greatly from the mean value in a dataset, the variance is large; otherwise, the variance is less. The variance is also a measure of dispersion with respect to the expected value.





A high variance indicates that the data in the dataset is very much spread out over a large area (random dataset), whereas a low variance indicates that the data is very similar in nature.

No variance is sometimes hard to understand in real datasets

### *Standard Deviation and Standard Error Estimates*

**Standard Deviation** With the help of variance, one can find out the standard deviation. Standard deviation, denoted by s, is the square root of the variance. The s says, "On an average how far do the data points fall from the mean or expected outcome?" Though the interpretation is the same as variance buts is squared rooted, therefore, less susceptible to the presence of outliers. The formulae for the population and the sample standard deviations are as follows:

The Population Standard Deviation: $\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2}$

The Sample Standard Deviation: $\sigma = \sqrt{\frac{1}{S-1}\sum_{i=1}^{S}(x_i-\bar{x})^2}$ ,

where N is number of data points in population, S is number in the sample, m is expected in the population or average value of x, and x is expected x in the sample.

**Standard Error** The standard error estimate is a measure of the accuracy of predictions from a relationship. Assume the linear relationship in a scatter plot of y (Figure 6.1). The scatter plot line, which fits, is defined as the line that minimizes the sum of squared deviations of prediction (also called the sum of squares error). The standard error of the estimate is closely related to this quantity and is defined below:

$$\sigma_{est} = \sqrt{\frac{\sum(y-y')^2}{N}},$$

where sest is the standard error in the estimate, y is an observed value, y¢ is a predicted value, and N is the number of values observed. The standard error estimate is a measure of the dispersion (or variability) in the predicted values from the expression for relationship.

### Probabilistic Distribution of Variables, Items or Entities

Probability is the chance of observing a dependent variable value with respect to some independent variable. Suppose a Grandmaster in chess has won 22 out of 100 games, drawn 78 times, and lost none. Then, probability P of winning Pw is 0.22, P of drawn game P0 is 0.78 and P of losing, PL = 0. The sum of the probabilities is normalized to 1, as only one of the three possibilities exist.

Probability distribution is the distribution of P values as a function of all possible independent values, variables, situations, distances or variables. For example, if Pis given by a function P(x), then P varies as x changes. Variations in P(x) with x can be discrete or continuous. The values of P are normalized such that sum of all P values is 1. Assuming distribution is around the expected value x, the standard normal distribution formula is:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}$$

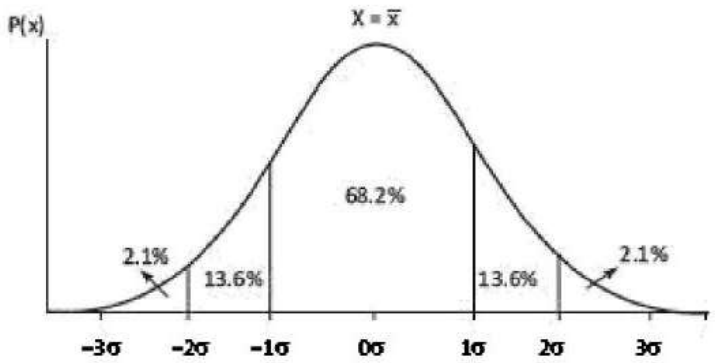



**Figure 5.3 Probability distribution function as a function of *x* assuming normal distribution around *x* = 'i, and standard deviation = s**

The figure also shows the percentages of areas in five regions with respect to the total area under the curve for P(x). The variance for probability distribution represents how individual data points relate to each other within a dataset.





variance is the average of the squared differences between each data value and the mean.

**Kernel Functions**

Kernel function is a function which is a central or key part of another function. For example, Gaussian kernel function is the key part of the probability distribution function. Figure 5.3 shows the probability normal distribution, which is a Gaussian function based on the Gaussian kernel function.

A kernel function[1], K* defines as

$$K^*(u) = \lambda . K(\lambda . u),$$

where $\lambda > 0$. Gaussian kernel function is

$$K^*(x) = \left[\frac{1}{(\sqrt{2\pi})}\right] e^{\left[-\frac{u^2}{2^2}\right]},$$

and when $u = \frac{\left\{\frac{x-\bar{x}}{2}\right\}}{\sigma}$, the distribution function is proportional to

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}.$$

$$\lambda = \left\{\frac{1}{\sigma\sqrt{2}}\right\}$$

Tricube kernel function is:

$$K^*(u) = (70/81)\,(1 - |u|^3)^3 \lambda . K(\lambda . u),$$

where $|u| \le 1$.

## Moments

*Moments* (0, 1, 2, ...) refer to expected values to the powers of (0, 1, 2 ...) of random variable variance. $0^{th}$ moment is 1, $1^{st}$ moment = $E(x) = \bar{x}$, (expected value), $2^{nd}$ moment is squared $V[(x_i - \bar{x})^2]$ = sum of product of $(x_i - \bar{x})^2$, and $P(x = x_i)$.

## Analysis of Variance

An ANOVA test is a method which finds whether the fitted results are significant or not. This means that the test finds out (infer) whether to reject or accept the null hypothesis. Null hypothesis is a statistical test that means the hypothesis that "no significant difference exists between the specified populations difference is just due to sampling or experimental error.

Consider two specified populations (datasets) consisting of yearly sales data of Tata Zest andJaguar Land Rover models. The statistical test is for proving that yearly sales of both the models, means increments and decrements of sales are related or not. Null hypothesis starts with the assumption that no significant relation exists in the two sets of data (population).

The analysis (ANOVA) is for disproving or accepting the null hypothesis. The test also finds whether to accept another alternate hypothesis. The test finds that whether testing groups have any difference between them or not.

F-test F-test requires two estimates of population variance- one based on variance between the samples and the other based on variance within the samples. These two estimates are then compared for F-test:

$$F = \frac{E1(V)}{E2(V)}$$

where El(V) is an estimate of population variance between the two samples and E2(V) is an estimate of population variance within the two samples. Several different F-tables exist. Each one has a different level of significance. Thus, look up the numerator degrees of freedom and the denominator degrees of freedom to find the critical value.





## Correlation

Correlation means analysis which lets us find the association or the absence of the relationship between two variables, x and y. Correlation gives the strength of the relationship between the model and the dependent variable on a convenient 0-100% scale.

R-Square Risa measure of correlation between the predicted values y and the observed values of x. R-squared (R2 ) is a goodness-of-fit measure in linear- regression model. It is also known as the coefficient of determination. R2 is the square of R, the coefficient of multiple correlations, and includes additional independent (explanatory) variables in regression equation.

Interpretation of R-squared The larger the R2 , the better the regression model fits the observations, i.e., the correlation is better. Theoretically, if a model shows 100% variance, then the fitted values are always equal to the observed values, and therefore, all the data points would fall on the fitted regression line

## Correlation Indicators of Linear Relationships

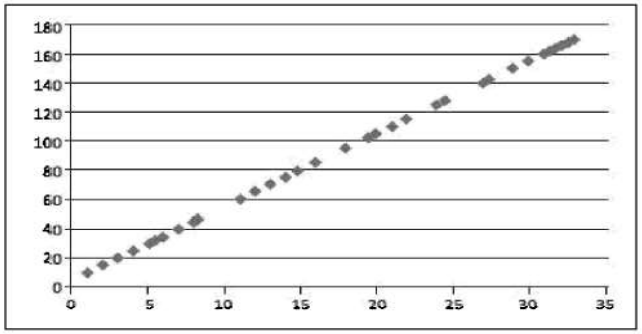
Correlation is a statistical technique that measures and describes the 'strength' and 'direction' of the relationship between two variables.

Relationships and correlations enable training model on sample data using statistical or ML algorithms. Statistical correlation is measured by the coefficient of correlation. The most common correlation coefficient, called the Pearson product-moment correlation coefficient. It measures the strength of the linear association between variables. The correlation r between the two variables x and y is:

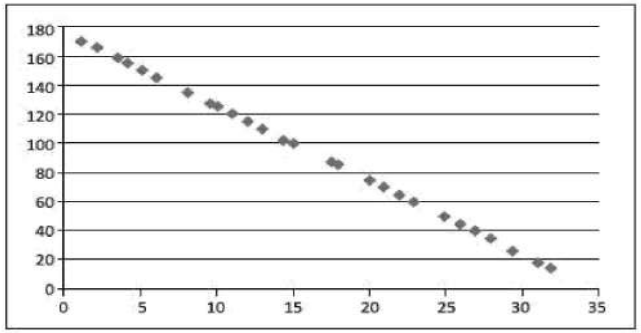
$$r = \left[\frac{1}{(n-1)}\right] \times \sum \left\{ \left[\frac{(x_i - \bar{x})}{\sigma_x}\right] \times \left[\frac{(y_i - \bar{y})}{\sigma_y}\right] \right\},$$

where n is the number of observations in the sample, xi is the x value for observation i, x- is the sample mean of x,Yi is the y value for observation i, y- is the sample mean of y, sx is the sample standard deviation of x, and Sy is the sample standard deviation of y.

| Value of r | Strength of relationship |
|---|---|
| -1.0 to -0.5 or 1.0 to 0.5 | Strong |
| -0.5 to -0.3 or 0.3 to 0.5 | Moderate |
| -0.3 to -0.1 or 0.1 to 0.3 | Weak |
| -0.1 to 0.1 | None or very weak |



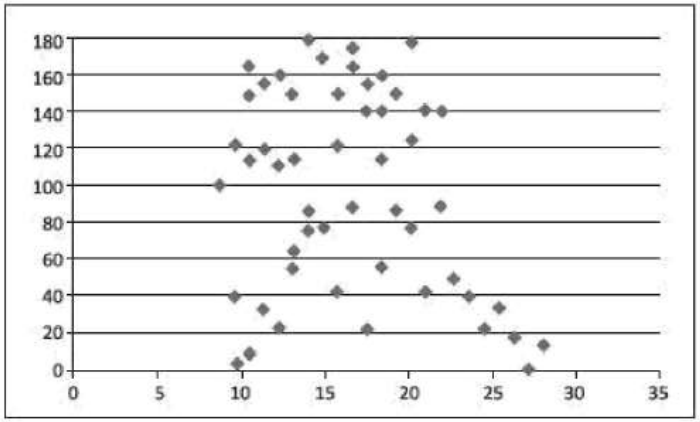

Perfect Positive Linear Relationship ($r = 1$)



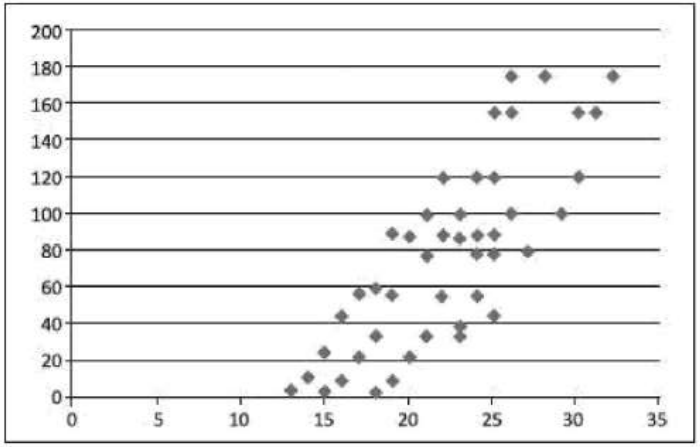

Perfect Negative Linear Relationship ($r = -1$)








No Relationship ($r \sim 0$)




Positive Linear Relationship ($r = 0.9$)

**Figure 5.4 Perfect and imperfect, linear positive and negative relationships, and the strength and direction of the relationship between variables**

### 5.3 Regression Analysis

Correlation and regression are two analyses based on multivariate distribution. A multivariate distribution means a distribution in multiple variables.

Suppose a company wishes to plan the manufacturing of Jaguar cars for coming years. The company looks at sales data regressively, i.e., data of

previous years' sales. Regressive analysis means estimating relationships between variables. Regression analysis is a set of statistical steps, which estimate the relationships among variables. Regression analysis may require many techniques for modeling and performing the analysis using multiple variables. The aim of the analysis is to find the relationships between a dependent variable and one or more independent, outcome, predictor or response variables. Regression analysis facilitates prediction of future values of dependent variables

Non-linear regression equation is as follows:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3,$$

where number of terms on the right-hand side are 3 or 4. Linear regression means only the first two terms are considered. The following subsections describe regression analysis in detail.

**Simple Linear Regression**

Linear regression is a simple and widely used algorithm. It is a supervised ML algorithm for predictive analysis. It models a relationship between the independent predictor or explanatory, and the dependent outcome or variable, y using a linearity equation.

$$y = f(a_0, a_1) = a_0 + a_1 x,$$

where $a_0$ is a constant and $a_1$ is the linearity coefficient.

Simple linear regression is performed when the requirement is prediction of values of one variable, with given values of another variable.

The purpose of regression analysis is to come up with an equation of a line that fits through a cluster of points with minimal amount of deviation from the line. The best-fitting line is called the regression line. The deviation of the points from the line is called an 'error'.





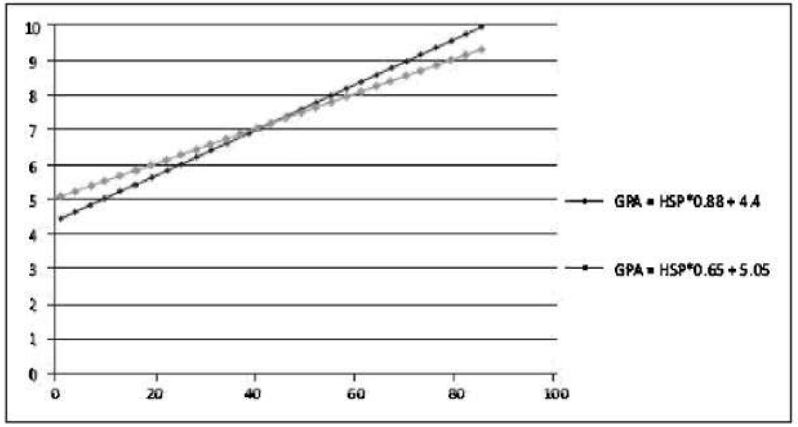



Figure shows a simple linear regression with two regression lines with different regression equations. Looking at the scatter plot, two lines can fit best to summarize the relation between GPA and high school percentage.

**Linear Square Estimation**

Assume n data-points, i = 1, 2, ..., n. A line out of two lines (Figure 6.6) that fits the data best will be one for which the sum of the squares of the n prediction errors (one for each observed data point) is as small as possible. This is the 'least squares criterion', which says that the best fit is one, which 'minimizes the sum of the squared prediction errors'. This implies that when the equation of the best fitting line is:

$$y_i' = b_0 + b_1 x_i,$$

where b0 and b1 are the coefficients which minimize the errors. The coefficients values make the sum of the squared prediction errors as small as possible

**Multiple Regression**

A criterion variable can be predicted from one predictor variable in simple linear regression. The criterion can be predicted by two or more variables in multiple regressions.

Multiple regressions are used when two or more independent factors are involved. These regressions are also widely used to make short- to mid-term predictions to assess which factors to include and which to exclude. Multiple regressions can be used to develop alternate models

with different factors. More than one variable can be used as a predictor with multiple regressions. However, it is always suggested to use a few variables as predictors necessarily, to get a reasonably accurate forecast. The prediction takes the form:

$$y = a + c_1 x_1 + c_2 x_2 + \ldots + c_n x_n.$$

More than one variable can be used as a predictor with multiple regressions. However, it is always suggested to use a few variables as predictors necessarily, to get a reasonably accurate forecast. The prediction takes the form:

Multiple regression analysis, often referred to simply as regression analysis, examines the effects of multiple independent variables on the value of a dependent variable or outcome.

**Modelling Possibilities using Regression.**

Regressions range from simple models to highly complex equations. Two primary uses for regression are forecasting and optimization. Consider the following examples:

1. Using linear analysis on sales data with monthly sales, a company could forecast sales for future months.
2. For the funds that a company has invested in marketing a particular brand, an analysis of whether the investment has given substantial returns or not can be made.
3. Suppose two promotion campaigns are running on TV and Radio in parallel. A linear regression can confine the individual as well as the combined impact of running these advertisements together.
4. An insurance company exploits a linear regression model to obtain a tentative premium table using predicted claims to Insured Declared Value ratio.
5. A financial company may be interested in minimizing its risk portfolio and hence want to understand the top five factors or reasons for default by a customer.
6. To predict the characteristics of child based on the characteristics of their





parents.

7. A company faces an employment discrimination matter in which a claim that women are being discriminated against in terms of salary is raised.
8. Predicting the prices of houses, considering the locality and builder characteristics in a locality of a particular city.
9. Finding relationships between the structure and the biological activity of compounds through their physical, chemical and physicochemical traits is most commonly performed with regression techniques.
10. To predict compounds with higher bioactivity within groups.

## 6.3.2 Predictions using Regression Analysis

Regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of another variable. Regression analysis is generally a statistical method to deal with the formulation of a mathematical model depicting the relationship amongst dependent and independent variables. The dependent variable is used for the purpose of prediction of the values. One or more variables whose values are hypothesized are called independent variables. The prediction for the dependent variable can be made by accurate selection of independent variables to estimate a dependent variable.

Two steps for predicting the dependent variable:

Estimation step: A function is hypothesized and the parameters of the function are estimated from the data collected on the dependent variable.

Prediction step: The independent variable values are then input to the parameterized function to generate predictions for the dependent variable.

K-Nearest-Neighbour Regression Analysis

Consider the saying, 'a person is known by the company he/she keeps.' Can a prediction be made using neighbouring data points? K-Nearest Neighbours (KNN) analysis is an ML based technique using the concept, which uses a subset of K = 1, 2 or 3 neighbours in place of a complete dataset. The subset is a training dataset.

Assume that population (all data points of interest) consist of k-data points. A data point independent variable is xi, where i = 1 to k. K-Nearest Neighbours (KNN) is an algorithm,

which is usually used for classifiers. However, it is useful for regression also. Predictions can use all k examples (global examples) or just K examples (K-neighbours with K = 1, 2 or 3). It predicts the unknown value Yp using predictor variable using the available values at the neighbours. The training dataset consists of available values of Yni at Xni with ni = 1 to K, where ni is the K-the neighbour, means just the local examples.

A subset of training dataset restricts k to K-neighbours, where K = 1, 2 or 3. This means using local values near the predictor variable. K = 1 means the nearest neighbour data points. K = 2 means the next nearest neighbour data points (xi,YJ K = 3 means the next to next nearest neighbour data points (Xj,yi).

First find all available neighbouring target (xi, Yi) cases and then predict the numerical value to be predicted based on a similarity measure. Prediction methods are as follows:

1. Simple interpolation, when predictor variable is outside the training subset
2. Extrapolation, when predictor variable is outside the training subset
3. Averaging, local linear regression or local-weighted regression.

**Euclidean Distance** The following equation computes the Euclidean distance $D_{Eu}$:

Sum of the squared Euclidean distance, $[D_{Eu}]^2 = \left[\sum_{i=1}^{v}(x_i - x_i')^2\right]$, and

$$\text{Euclidean distance } D_{Eu} = \left[\sum_{i=1}^{v}(x_i - x_i')^2\right]^{1/2}$$

$$\text{Euclidean distance } D_{Eu} = [(x_j - x_{j+1})^2 + (y_j - y_{j+1})^2]^{1/2}$$

**Manhattan Distance** The following equation computes the Manhattan distance $D_{Ma}$:

Manhattan distance $D_{Ma}$ = $D_{Ma} = \sum_{i=1}^{v}\left[|x_i - x_i'|\right]$ (6.20c)

Manhattan distance for three variables $v = 3$ (two independent variables and one dependent variable case) consists of three terms on the right-hand side in





Manhattan distance for three variables v = 3 (two independent variables and one dependent variable case) consists of three terms on the right-hand side in Equation.

**Comparison between Euclidean and Manhattan Distances**

Basically, Euclidean distance is the direct path distance between two data points in v-dimensional metric spaces. Manhattan distance is the staircase path distance between them. Staircase distance means to move to the next point, first move along one metric dimension (say, x axis) from the first point, and then move to the next along another dimension (say, y axis).

When v = 2, Euclidean distance is the diagonal distance between the points on an x-y graph. Manhattan distances are faster to calculate as compared to Euclidean distances. Manhattan distances are proportional to Euclidean distances in case of linear regression.

**Minkowski Distance** The following equation computes the Minkowski distance $D_{Mi}$:

$$\text{Minkowski distance } D_{Mi} = \left\{\sum_{i=1}^{v}\left[(x_i - x_i')^q\right]\right\}^{1/q}$$

**Hamming Distance** When predictions are on the basis of categorical variables, then use the Hamming distance. It is a measure of the number of instances in which corresponding values are found.

$$\text{Hamming Distance, } D_H = \sum_{i=1}^{v}|x_i - x_i'|.$$

when xi= x¢i then DH= O and when Xj not equal to x¢i then DH= 1. For example, Hamming distance DH= 1 between 10100111100 and 11100111100 because just one substitution is needed, change second bit from O to 1 at 10th place from the right to left positioned bits.

**Normalization Concept** Normalization factor in p-norm form in a v- dimensional space is

$$x_i = N^{-1} \cdot x_i \text{ where } N = \left(\sum_{i=1}^{v} |x_i|^p\right)^{1/p}$$

Here, xi is ith component of the vector X. The total number of components are Two-dimensional space v = 2, three-dimensional v = 3. The following example explains the meaning of distances, use of Euclidean and Manhattan distances, use distances for predictions, and the KNN regression analysis.

**5.4 FINDING SIMILAR ITEMS, SIMILARITY OF SETS ANDCOLLABORATIVE FILTERING**

The following subsections describe methods of finding similar items using similarities, application of near-neighbour search, Jaccard similarity of sets, similarity of documents, Collaborative Filtering (CF) as a similar-set problem, and the distance measures for finding similarities.

Finding Similar Items

An analysis requires many times to find similar items. For example, finding similar excellent performance of students in Python programming, similar showrooms of a specific car model which show high sales per month, recommending books on similar topic such as in Internet of Things by Raj Kamal from McGraw-Hill Higher Education, etc.

Application of Near Neighbour Search

Similar items can be found using Nearest Neighbour Search (NNS). The search finds that a point in a given set is most similar (closest) to a given point. A dissimilarity function having larger value means less similar. The dissimilarity function is used to find similar items.

NNS algorithm is as follows: Consider set S having points in a space M. Consider a queried point q EM, which means q is member of M. k-NNS algorithm finds the k-closet (1-NN) points to q in S.

Do not consider the number of items in which two users' preferences overlap. (e.g., 2 overlap items==> 1, more items may not be better.)

If two users overlap on only one item, no correlation can be computed.

The correlation is undefined if series of preference values are identical.





Greater distance means greater dissimilarity. Dissimilarity coefficient relates to a distance metric in metrics space in v-dimensional space. An algorithm computes Euclidean, Manhattan and Minkowski distances using Equations.

Distance metric is symmetric and follows triangular inequality. Meaning of triangular inequality can be understood by an example. Consider three vectors of lengths x, y, and z. Then, triangular inequality means

z < x + y. It is similar to the theorem of inequality that the third side of a triangle is less than the sum of two other sides, and never equal. The theorem applies to v-dimensional space also. Dissimilarity can be asymmetric, i.e., triangular inequality is not true (Bergman divergence).

**Jaccard Similarities set**

Let A and B be two sets. Jaccard similarity coefficient of two sets measures using notations in set theory as shown below:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

A n B means the number of elements or items that are same in sets A and B. A U B means the number of elements or items present in union of both the sets. Assume two set of students in two computer courses, Computer Applications CA, and Computer Science CS in a semester. Set CA 40 students opted for Java out of

60 students. Set CS 30 students opted for Java out of 50 students. Jaccard similarity coefficient Jjava (CA, CS) = 30/(60 + 50) x 100% = 27%. Two sets are sharing 27% of the members for Java course.

**Similarity of Document.**

An application ofJaccard similarity coefficient is in Natural Language Processing (NLP) and text processing. It quantifies the similarity in documents. Computational steps are as follows:

4. Find Bag of Words (Section 9.2.1.4) and remove words such as is, are, does, at, in, ....
5. Assign weighting factor is the Term frequency and Inverse Document Frequency (TF-IDF). Consider the frequency of words in the document.

6. Find k-shingles. A shingle is a word of fixed length. The k-shingles are the number of times the similar shingles extracted from a document or text. Examples of a shingle are Java, GP, 8.0, Python, 80%, Programming.
7. Find n-grams. A gram is a contiguous sequence of fixed length item (word
8. or set of characters, letters, words in pairs, triplets, quadruplets, ...) in a document or text. The n-grams are the number of times the similar items (1-grams, 2-grams, ..) extracted from a document or text. The 3-gram examples are lava GP 8.0, Python Programming 7.8, Big Data Analytics, 23A 240C 8LP, the numbers of which are extracted from the text.
9. Compute Jaccard similarity coefficient using Equation (6.22) between the documents.

**Collaborative Filtering as a Similar-Sets Finding Problem**

An analysis requires finding similar sets using collaborative filtering. Collaborative filtering refers to a filtering algorithm, which filters the items sets that have similarities with different items in a dataset.

CF finds the sets with items having the same or close similarity coefficients. Following are some examples of applications of CF:

Find those sets of students in computer application, and computer science who opt for the Java Programming subject in a semester.

Find sets of students in Java Programming subjects to whom same teacher taught and they showed excellent performance.

An algorithm finds the similarities between the sets for the CF. Applications of CF are in many ML methods, such as association rule mining, classifiers, and recommenders.

## Distance Measures for Finding Similar Items or Users

Distance can be defined in a number of ways. Distance is the measure of length of a line between two values in a two-dimensional map or graph. Set of Equations (6.20) measures distances.

For example, distance between (2014, 6%) and (2018, 8%) on a scatter plot when year is on the x axis and profit% on they axis is Distance = v [(2014 - 2018)2 + (6 - 8)2 ] = v (16 + 4) = 4.47, using Equation (6.20b). Distance can also be





similarly defined in v-dimensional space using Equation (6.20a).

Distances between all members in a set of points can be computed in metrics space using a mathematical equation. Metrics space means measurable or quantifiable space. For example, profit and year on a scatter plot are in metric space of two dimensions. Probability distribution function values are in metric space.

Consider student-performance measures 'very good' and 'excellent'. These parameters are in non-metric space. How are they made measurable? They become measurable when very good is specified as grade point average 8.5 which implies that a score between 8.0 to 9.0 is very good, and define 9.5 which implies that a score between 9.0 to 10.0 is excellent on a 10-point scale.

Euclidean distance $D_{Eu} = \left[\sum_{i=1}^{v} (x_i - x_i')^2\right]^{1/2}$

**Cosine distance**

Let **U** and **V** be two non-zero vectors, two documents in the vector space.

$$D_{Cos}(\mathbf{U}, \mathbf{V}) = \frac{\sum_i U_i V_i}{\sqrt{\sum_i U_i^2}\sqrt{\sum_i V_i^2}}$$

**Vector Cosine-Based Similarity Vector cosine similarity** in terms of angle between two vectors U and V is given by equation

$$\phi_{UV} = \cos^{-1}(\mathbf{U}, \mathbf{V}) = \frac{U \cdot V}{[\|U\| - \|V\|]}$$

**Concept of Sparse and Dense Vectors** Sparse vector uses a hash-map and consists of non-zero values. Hash-map is a collection, which stores data in (key- value) format (Section 3.3.1). Format is also called random access. Hashing means to convert a large value or string into shorter value or string so that indexing for searching is fast.

For example, assume a vector, which consists of array elements, (subject,

number of students opting, average GPA).

1. Dense vectors have elements (Hive, 40, 8.0), Oava, 30, 8.5), (FORTRAN, 0, 0), (Pascal, 0, 0). Dense vector consists of all elements, whether the element value is

O or not O.

2. Sparse vectors will be two only with elements (4, 40, 8.0) and (3, 30, 8.5). Random access Sparse vector means access to elements (key, value pairs) using key. Sparse vector consists of elements for which key is such that value is not O (Section 3.3.1).
3. Sparse vector has an associated hash-map in form of a hash-table. First row- Pascal, 1, second row- FORTRAN, 2, third row- Java, 3 and fourth row-Hive.
4. Hashing is a process of assigning a small number or small-sized string indexing, searching and memory saving purposes. Hash process uses a hash function, which results into not-colliding values. In case of two colliding numbers, the process assigns a new number. Sequential access sparse vectors mean two parallel accessing vectors, i.e., one to access keys and the other for values.

Edit distance DEd is a distance measure for dissimilarity between two set of strings or words. DEd equals the minimum number of inserts and deletes of characters needed to transform one set into another. Applications of edit distances are in text analytics and natural language processing, similarities in DNA sequences etc. DNA sequences are strings of characters.

**Hamming Distance**

If both U and V are vectors, Hamming distance DHa is equal to the number of

different elements between these two vectors. Recall Example 6.5 (iv) for Hamming distance between Jspi and Zspi. Hamming similarity-coefficient between car models Jaguar Land Rover and Zest is (1- 2/7) = 0.7. [70%]

Otta between two strings of equal length is the number of positions at which the corresponding characters differ. Otta is also equal to the minimum number of substitutions required to transform one string into the other. Otta is also equal to the minimum number of errors that need correction using transformation or substitution.

Hamming distance is therefore another distance measure for measuring the edit distance between two sets of strings, words or sequences.





**Frequent Item Set and Association rule Mining**

**Frequent Item Set**

Frequent itemset refers to a set of items that frequently appear together, for example, Python and Big Data Analytics. Students of computer science frequently choose these subjects for in-depth studies. Frequent itemset refers to a frequent itemset, which is a subset of items that appears frequently in a dataset.

Frequent Itemset Mining (FIM) refers to a data mining method which helps in discovering the itemsets that appear frequently in a dataset. For example, finding a set of students who frequently show poor performance in semester examinations. Frequent subsequence is a sequence of patterns that occurs frequently. For example, purchasing a football follows purchasing of sports kit. Frequent substructure refers to different structural forms, such as graphs, trees or lattices, which may be combined with itemsets or subsequences.

## Association Rule- Overview

An important method of data mining is association rule mining or association analysis. The method has been widely used in many application areas for discovering interesting relationships which are present in large datasets. The objective is to find uncovered relationships using some strong rules. The rules are termed as association rules for frequent itemsets. Mahout includes a 'parallel frequent pattern growth' algorithm. The method analyzes the items in a group and then identifies which items typically appear together (association)

## Apriori Algorithm

Apriori algorithm is used for frequent itemset mining and association rule mining. Apriori algorithm is considered as one of the most well-known association rule algorithms. The algorithm simply follows a basis that any subset of a large itemset must be a large itemset. This basis can be formally given as the Apriori principle. The Apriori principle can reduce the number of itemsets needed to be examined. Apriori principle suggests if an itemset is frequent, then all of its subsets must also be frequent. For example, if itemset {A, B, C} is a frequent itemset, then all of its subsets {A}, {B}, {C}, {A, B}, {B, C} and {A, C} must be frequent. On the contrary, if an itemset is not frequent, then none of its

supersets can be frequent. This results into a smaller list of potential frequent itemsets as the mining progresses.

Assume X and Y are two itemsets. Apriori principle holds due to the following property of support measure:

$$\forall\ X, Y: (X \subseteq Y) \rightarrow s(X) \geq s(Y)$$

Explanation: V means for all, and c means 'subset of' and can be 'equal to or included in'. Support of an itemset never exceeds the support of its subsets. This is known as the anti-monotone property of support.

The algorithm uses k-itemsets (An itemset which contains k items is known as a k-itemset) to explore (k+1)-itemsets in order to mine frequent itemsets from transactional database for the Boolean association rules (If Then rule is a Boolean association rule, as it checks if true or false).

The frequent itemset algorithm uses candidate generation process. The groups of candidates are then tested against the dataset. Apriori uses breadth-first search method and a hash tree structure to count candidate itemsets. Also, it is assumed that items within an itemset are kept in lexicographic order. The algorithm identifies the frequent individual items in the database and extends them to larger and larger itemsets as long as those itemsets are found in the database. The frequent itemsets provide the general trends in the database as well.

**Evaluation of Candidate rules**

Apriori algorithm evaluates candidates for association as follows:

$C_k$: Set of candidate-itemsets of size $k$

$F_l$: Set of frequent itemsets of size $k$

$F_1$ = {large items}

*for* (k=1; $F_k$ != 0; k++) *do* {

$C_{k+1}$ = New candidates generated from $F_k$

*for* each transaction t in the database *do*





Increment the count of all candidates in Ck+1 that are contained int Fk+i = Candidates in Ck+i with minimum support

}

Steps of the algorithm can be stated in the following manner:

Candidate itemsets are generated using only large itemsets of the previous iteration. The transactions in the database are not considered while generating candidate itemsets.

The large itemset of the previous iteration is joined with itself to generate all itemsets having size higher by 1.

Each generated itemset that does not have a large subset is discarded. The remaining itemsets are candidate itemsets.

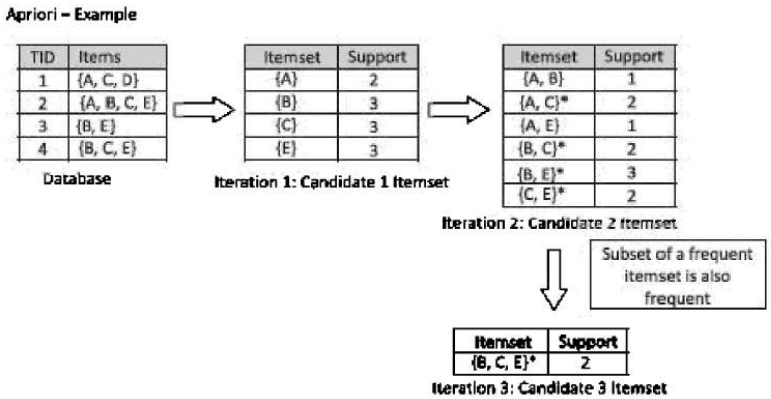



Figure shows Apriori algorithm process for adopting the subset of frequent itemsets as a frequent itemset.

It is observed in the Apriori example that every subset of a frequent itemset is also frequent. Thus, a candidate itemset in Ck+1 can be pruned even if one of its subsets is not contained in Fk.

**Applications of Association rule mining**

**Market basket analysis** is a tool for knowledge discovery about co-occurrence of items. A co-occurrence means two or more things occur together. It can also be defined as a data mining technique to derive the strength of association between pairs of product items. If

people tend to buy two products (say A and B) together, then the buyer of product A is a potential customer for an advertisement of product B.

The concept is similar to the real market basket where we select an item (product) and put it in a basket (itemset). The basket symbolizes the transactions. The number of baskets is very high as compared to the items in a basket. A set of items that is present in many baskets is termed as a frequent itemset. Frequency is the proportion of baskets that contain the items of interest.

Market basket analysis can be applied to many areas. The following example explains the market basket model using application examples.

Market basket analysis generates If-Then scenario rules. For example, if X occurs then Y is likely to occur too. If item A is purchased, then item B is likely to be purchased too. The rules are derived from the experience. This may be the result of frequencies of co-occurrence of items in past transactions.

The rules can be used in several analytical strategies. The rules can be written in format If {A} Then {B}. The If part of the rule (A) is known as antecedent and the THEN part of the rule (B) is known as consequent. The condition is antecedent and the result is consequent.

The applications of market basket analysis in various domains other than retail are:

- Medical analytics: Market basket analysis can be used for conditions and symptom analysis. This helps in identifying a profile of illness in a better way. The analysis is also useful in genome analysis, molecular fragment mining, drug design and studying the role of biomarkers in medicine. The analysis can also help to reveal biologically relevant associations betweendifferent genes. Further, it can also help to find the effect of environment on gene expressions.
- Web usage analytics: FIM approaches can be used with viewing data on websites. The information contained in association rules can be exploited to learn about website browsing of visitor's behavior, developing website structure by making it more effective for visitors, or improving web marketing promotions. The results of this type of analysis can be used to inform website design (how items are grouped together) and to power recommendation engines (Section 6.8). Results are helpful in targeted marketing. For example, advertising content that people are probably interested in, based on past behavior of users.
- Fraud detection and technical dependence analysis: Extract knowledge so that





normal behavior patterns may be obtained in illegal transactions from a credit card database in order to detect and prevent fraud. Another example can be to find frequently occurring relationships or FIM rules

between the various parties involved in the handling of the financial claim. Some examples are:

- ♦ Financial institutions to analyze credit card purchases of customers to build profiles for fraud detection purposes and cross-selling opportunities.
- ♦ Insurance institution builds the profiles to detect insurance claim fraud. The profiles of claims help to determine if more than one claim belongs to a particular victim within a specified period of time.

- Click stream analysis or web link analysis: Click stream refers to a sequence of web pages viewed by a user. Analysis of clicks is the process of extracting knowledge from web logs. This helps to discover the unknown and potentially interesting patterns useful in the future. It facilitates an understanding of the behavior of website visitors. This knowledge can be used to enhance the way that web pages are interconnected or for increasing the sales of the commercial websites.
- Telecommunication services analysis: Market basket analysis can be used to determine the type of services being utilized and the packages customers are purchasing. This knowledge can be used to plan marketing strategies for customers who are interested in similar services. For example, telecommunication companies can offer TV Internet, and web- services by creating combined offers. The analysis might also be useful to determine capacity requirements.
- Plagiarism detection: It is the process of locating instances of similar content or idea within a work or a document. Plagiarism detection can findsimilarities among statements that may lead to similar paragraphs if all statements are similar and that possibly lead to similar documents. Formation of relevant word and sentence sequences for detection of plagiarism using association rule mining technique is also very popular technique.

**Finding Associations**

- Association rules intend to tell how items of a dataset are associated with each other. The concept of association rules was introduced in 1993 for discovering relations between

items in sales data of a large retailing company. Association analysis is applicable to several domains. Some of them are marketing, bioinformatics, web mining, scientific data analysis, and intrusion detection systems.

- The applications might be to find: products that are often purchased together, types of DNA sensitive to a new drug, the possibility of classifying web documents automatically, geophysical trends or patterns in seismicity to predict earthquakes and automate the malicious detecting characteristics.
- In medical diagnosis, for example, considering the co-morbid (co-occur) conditions can help in treating the patient in better way. This helps in improving patient care and medicine prescription.

**Finding Similarity**

Let A and B be two itemsets. Jaccard similarity index of two itemsets is measured in terms of set theory using the following equation:

$$\text{Jaccard itemsets similarity index} = 1 - \frac{|A \cap B|}{|A \cup B|} \times 100\%.$$

Explanation: n means intersection, number of those elements or items which are the same in set A and B. U means union, number of elements or items present in union of A and B.

**How will you define a similarity in a purchase of car model**

Assume two sets of car customers, youth Y and family F. Assume in set Y, 40 out of 100 youths and F 50 out of 200 families opted for the Tata Zest car model. Jaccard similarity index Jzest (Y, F) = 40/ (100 + 200).100% = 13%. Two sets are sharing 13% of the members who purchased a Zest.





# MODULE 5

# Chapter 1: Text Mining

Text mining is the art and science of discovering knowledge, insights and patterns from an organized collection of textual databases. Textual mining can help with frequency analysis of important terms, and their semantic relationships.

Text is an important part of the growing data in the world. Social media technologies have enabled users to become producers of text and images and other kinds of information. Text mining can be applied to large-scale social media data for gathering preferences, and measuring emotional sentiments. It can also be applied to societal, organizational and individual scales.

## 1.1 Text Mining Applications

Text mining is a useful tool in the hands of chief knowledge officers to extract knowledge relevant to an organization. Text mining can be used across industry sectors and application areas, including decision support, sentiment analysis, fraud detection, survey analysis, and many more.

1. *Marketing*: The voice of the customer can be captured in its native and raw format and then analyzed for customer preferences and complaints.
    1. Social personas are a clustering technique to develop customer segments of interest. Consumer input from social media sources, such as reviews, blogs, and tweets, contain numerous leading indicators that can be used towards anticipating and predicting consumer behavior.
    2. A 'listening platform' is a text mining application, that in real time, gathers social media, blogs, and other textual feedback, and filters out the chatter to extract true consumer sentiment. The insights can lead to more effective product marketing and better customer service.

2. The customer call center conversations and records can be analyzed for patterns of customer complaints. Decision trees can organize this data to create decision choices that could help with product management activities and to become proactive in avoiding those complaints.

3. *Business operations*: Many aspects of business functioning can be accurately gauged from analyzing text./
    1. Social network analysis and text mining can be applied to emails, blogs, social media and other data to measure the emotional states and the mood of employee populations. Sentiment analysis can reveal early signs of employee dissatisfaction which can then can be proactively managed.

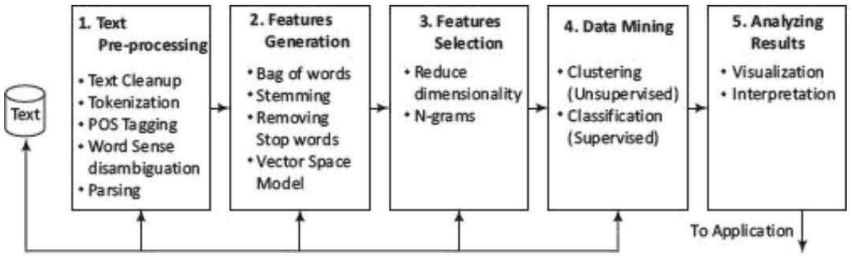
2. Studying people as emotional investors and using text analysis of the social Internet to measure mass psychology can help in obtaining superior investment returns.

3. *Legal*: In legal applications, lawyers and paralegals can more easily search case histories and laws for relevant documents in a particular case to improve their chances of winning.
   1. Text mining is also embedded in e-discovery platforms that help in minimizing risk in the process of sharing legally mandated documents.
   2. Case histories, testimonies, and client meeting notes can reveal additional information, such as morbidities in a healthcare situation that can help better predict high-cost injuries and preventcosts.

4. Governance and Politics: Governments can be overturned based on a tweet originating from a self-immolating fruit-vendor in Tunisia.
   1. Social network analysis and text mining of large-scale social media data can be used for measuring the emotional states and the mood of constituent populations. Micro-targeting constituents with specific messages gleaned from social media analysis can be a more efficient use of resources when fighting democratic elections.
   2. In geopolitical security, internet chatter can be processed for real- time information and to connect the dots on any emerging threats.
   3. In academic, research streams could be meta-analyzed for underlying research trends.

## 1.2 Text Mining Process

Text Mining is a rapidly evolving area of research. As the amount of social media and other text data grows, there is need for efficient abstraction and categorization of meaningful information from the text.




The five phases for processing text are as follows:

**Phase 1: Text pre-processing** enables Syntactic/Semantic text-analysis and does the followings:





1. Text *cleanup* is a process of removing unnecessary or unwanted information. Text cleanup converts the raw data by filling up the missing values, identifies and removes outliers, and resolves the inconsistencies. For example, removing comments, removing or escaping "%20" from URL for the web pages or cleanup the typing error, such as teh (the), do n't (do not) [%20 specifies space in a URL].
2. *Tokenization* is a process of splitting the cleanup text into tokens (words) using white spaces and punctuation marks as delimiters.
3. *Part of Speech (POS) tagging* is a method that attempts labeling of each token (word) with an appropriate POS. Tagging helps in recognizing names of people, places, organizations and titles. English language set includes the noun, verb, adverb, adjective,

   prepositions and conjunctions. Part of Speech encoded in the annotation system of the Penn Treebank Project has 36 POS tags.4
4. *Word sense disambiguation* is a method, which identifies the sense of a word used in a sentence; that gives meaning in case the word has multiple meanings. The methods, which resolve the ambiguity of words can be context or proximity based. Some examples of such words are bear, bank, cell and bass.
5. *Parsing* is a method, which generates a parse-tree for each sentence. Parsing attempts and infers the precise grammatical relationships between different words in a given sentence.

**Phase 2: Features Generation** is a process which first defines features (variables, predictors). Some of the ways of feature generations are:

1. *Bag of* words-Order of words is not that important for certain applications.

   Text document is represented by the words it contains (and their occurrences). Document classification methods commonly use the bag-of-words model. The pre-processing of a document first provides a document with a bag of words. Document classification methods then use the occurrence (frequency) of each word as a feature for training a classifier. Algorithms do not directly apply on the bag of words, but use the frequencies.
2. Stemming-identifies a word by its root.
   (i) Normalizes or unifies variations of the same concept, such as *speak* for three variations, i.e., speaking, speaks, speakers denoted by [speaking, speaks, speaker- + speak]
   (ii) Removes plurals, normalizes verb tenses and remove affixes.

   Stemming reduces the word to its most basic element. For example, impurification -+ pure.
3. *Removing stop words* from the feature space-they are the common words, unlikely to help text mining. The search program tries to ignore stop words. For example, ignores *a, at, for,* it, *in*
   and *are.*
4. *Vector Space Model* (VSM)-is an algebraic model for representing text documents as vector of identifiers, word frequencies or terms in the document index. VSM uses the method of term frequency-inverse document frequency (TF-IDF) and evaluates how important is a

word in a document.

When used in document classification, VSM also refers to the bag-of-words model. This bag of words is required to be converted into a term-vector in VSM. The term vector provides the numeric values corresponding to each term appearing in a document. The term vector is very helpful in feature generation and selection.

*Term frequency and inverse document frequency (IDF)* are important metrics in text analysis. TF-IDF weighting is most common- Instead of the simple TF, IDF is used to weight the importance of word in the document.

**Phase 3: Features Selection** is the process that selects a subset of features by rejecting irrelevant and/or redundant features (variables, predictors or dimension) according to defined criteria. Feature selection process does the following:

1. *Dimensionality* reduction-Feature selection is one of the methods of division and therefore, dimension reduction. The basic objective is to eliminate irrelevant and redundant data. Redundant features are those, which provide no extra information. Irrelevant features provide no useful or relevant information in any context.

   Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) are dimension reduction methods. Discrimination ability of a feature measures relevancy of features. Correlation helps in finding the redundancy of the feature. Two features are redundant to each other if their values correlate with each other.
2. *N-gram* evaluation-finding the number of consecutive words of interest and extract them. For example, 2-gram is a two words sequence, ["tasty food", "Good one"]. 3-gram is a three words sequence, ["Crime Investigation Department"].
3. *Noise detection and evaluation of outliers* methods do the identification of unusual or suspicious items, events or observations from the data set. This step helps in cleaning the data.

The feature selection algorithm reduces dimensionality that not only improves the performance of learning algorithm but also reduces the storage requirement for a dataset. The process enhances data understanding and its visualization.

**Phase 4: Data mining techniques** enable insights about the structured database that resulted from the previous phases. Examples of techniques are:

1. Unsupervised learning (for example, clustering)
   (i) The class labels (categories) of training data are unknown
   (ii) Establish the existence of groups or clusters in the data
   Good clustering methods use high intra-cluster similarity and low inter-cluster similarity. Examples of uses - biogs, pattern
   and trends.
2. *Supervised learning (for example, classification)*
   (i) The training data is labeled indicating the class
   (ii) New data is classified based on the training set
   Classification is correct when the known label of test sample is identical with the resulting class computed from the classification model.





Examples of uses are *news filtering application,* where it is required to automatically assign incoming documents to pre-defined categories; *email spam filtering,* where it is identified whether incoming email messages are spam or not.

Example of text classification methods are *Naive Bayes Classifier* and *SVMs.*

3. *Identifying evolutionary patterns* in temporal text streams-the method is useful in a wide range of applications, such as summarizing of events in news articles and extracting the research trends in the scientific literature.

**Phase 5: Analysing results**

(i) Evaluate the outcome of the complete process.
(ii) Interpretation of Result- If acceptable then results obtained can be used as an input for next set of sequences. Else, the result can be discarded, and try to understand what and why the process failed.
(iii) Visualization - Prepare visuals from data, and build a prototype.
(iv) Use the results for further improvement in activities at the enterprise, industry or institution.

## Text Mining Challenges

The challenges in the area of text mining can be classified on the basis of documents area-characteristics. Some of the classifications are as follows:

1. NLP issues:
   (i) POS Tagging
   (ii) Ambiguity
   (iii) Tokenization
   (iv) Parsing
   (v) Stemming
   (vi) Synonymy and polysemy
2. Mining techniques:
   (i) Identification of the suitable algorithm(s)
   (ii) Massive amount of data and annotated corpora
   (iii) Concepts and semantic relations extraction
   (iv) When no training data is available
3. Variety of data:
   (i) Different data sources require different approaches and different areas of expertise

(ii) Unstructured and language independency

4. Information visualization

5. Efficiency when processing real-time text stream

6. Scalability

## 1.3 Term Document Matrix

This is the heart of the structuring process. Free flowing text can be transformed into numeric data in a TDM, which can then be mined using regular data mining techniques.

1. There are several efficient techniques for identifying key terms from a text. There are less efficient techniques available for creating topics out of them. For the purpose of this discussion, one could call key words, phrases or topics as a term of interest. This approach measures the frequencies of select important terms occurring in each document. This creates a t x d Term–by–Document Matrix (TDM) where t is the number of terms and d is the number of documents (Table 11.1).
2. Creating a TDM requires making choices of which terms to include. The terms chosen should reflect the stated purpose of the text mining exercise. The list of terms should be as extensive as needed, but should not include unnecessary stuff that will serve to confuse the analysis, or slow the computation.

| | Term-Document Matrix | | | | |
|---|---|---|---|---|---|
| Document/ Terms | Investment | Profit | Happy | Success | ... |
| Doc 1 | 10 | 4 | 3 | 4 | |
| Doc 2 | 7 | 2 | 2 | | |
| Doc 3 | | | 2 | 6 | |
| Doc 4 | 1 | 5 | 3 | | |
| Doc 5 | | 6 | | 2 | |
| Doc 6 | 4 | | 2 | | |
| ... | | | | | |

**Table 1.1: Term-Document Matrix**





Here are some considerations in creating a TDM.

1. A large collection of documents mapped to a large bag of words will likely lead to a very sparse matrix if they have few common words. Reducing dimensionality of data will help improve the speed of analysis and meaningfulness of the results. Synonyms, or terms will similar meaning, should be combined and should be counted together, as a common term. This would help reduce the number of distinct terms of words or 'tokens'.
2. Data should be cleaned for spelling errors. Common spelling errors should be ignored and the terms should be combined. Uppercase- lowercase terms should also be combined.
3. When many variants of the same term are used, just the stem of theword would be used to reduce the number of terms. For instance, terms like customer order, ordering, order data, should be combined into a single token word, called 'Order'.
4. On the other side, homonyms (terms with the same spelling but different meanings) should be counted separately. This would enhance the quality of analysis. For example, the term order can mean a customer order, or the ranking of certain choices. These two should be treated separately. "The boss ordered that the customer orders data analysis be presented in chronological order'. This statement shows three different meanings for the word 'order'. Thus, there will be a need for a manual review of the TD matrix.
5. Terms with very few occurrences in very few documents should be eliminated from the matrix. This would help increase the density of the matrix and the quality of analysis.
6. The measures in each cell of the matrix could be one of several possibilities. It could be a simple count of the number of occurrences of each term in a document. It could also be the log of that number. It could be the fraction number computed by dividing the frequency count by the total number of words in the document. Or there may be binary values in the matrix to represent whether a term is mentioned or not. The choice of value in the cells will depend upon the purpose of the textanalysis.

At the end of this analysis and cleansing, a well-formed, densely populated, rectangular, TDM will be ready for analysis. The TDM could be mined using all the available data mining techniques.

## 1.4 Mining the TDM

The TDM can be mined to extract patterns/knowledge. A variety of techniques could be applied to the TDM to extract new knowledge.

Predictors of desirable terms could be discovered through predictive techniques, such as regression analysis. Suppose the word profit is a desirable word in a document. The number of occurrences of the word profit in a document could be regressed against many other terms in the TDM. The relative strengths of the coefficients of various predictor variables would

show the relative impact of those terms on creating a profitdiscussion.

Predicting the chances of a document being liked is another form of analysis. For example, important speeches made by the CEO or the CFO to investors could be evaluated for quality. If the classification of those documents (such as good or poor speeches) was available, then the terms of TDM could be used to predict the speech class. A decision tree could be constructed that makes a simple tree with a few decision points that predicts the success of a speech 80 percent of the time. This tree could be trained with more data to become better over time.

Clustering techniques can help categorize documents by common profile. For example, documents containing the words investment and profit more often could be bundled together. Similarly, documents containing the words, customer orders and marketing, more often could be bundled together. Thus, a few strongly demarcated bundles could capture the essence of the entire TDM. These bundles could thus help with further processing, such as handing over select documents to others for legal discovery.

Association rule analysis could show relationships of coexistence. Thus, one could say that the words, tasty and sweet, occur together often (say 5 percent of the time); and further, when these two words are present, 70 percent of the time, the word happy, is also present in the document.

## 1.5 Comparing Text Mining and Data Mining

Text Mining is a form of data mining. There are many common elements between Text and Data Mining. However, there are some key differences (Table 1.2). The key difference is that text mining requires conversion of text data into frequency data, before data mining techniques can beapplied.





| Dimension | **Text Mining** | **Data Mining** |
|---|---|---|
| Nature of data | Unstructured data: Words, phrases, sentences | Numbers; alphabetical and logical values |
| Language used | Many languages and dialects used in the world; many languages are extinct, new documents are discovered | Similar numerical systems across the world |
| Clarity and precision | Sentences can be ambiguous; sentiment may contradict the words | Numbers are precise. |
| Consistency | Different parts of the text can contradict each other | Different parts of data can be inconsistent, thus, requiring statistical significance analysis |
| Sentiment | Text may present a clear and consistent or mixed sentiment, across a continuum. Spokenwords adds further sentiment | Not applicable |
| Quality | Spelling errors. Differing values of proper nouns such as names. Varying quality of language translation | Issues with missing values, outliers, etc |
| Nature of Analysis | Keyword based search; co- existence of themes; Sentiment Mining | A full wide range of statistical and machine learning analysis for relationship and differences |

**Table 1.2: Comparing Text Mining and Data Mining**

## 1.6 Text Mining Best Practices

Many of the best practices that apply to the use of data mining techniques will also apply to text mining.

1. The first and most important practice is to ask the right question. A good question is one which gives an answer and would lead to large payoffs for the organization. The purpose and the key question will define how and at what levels of granularity the TDM would be made. For example, TDM defined for simpler searches would be different from those used for complex semantic analysis or network analysis.
2. A second important practice is to be creative and open in proposing imaginative hypotheses for the solution. Thinking outside the box is important, both in the quality of the proposed solution as well as in finding the high quality data sets required to test the hypothesized solution. For example, a TDM of consumer sentiment data should be combined with customer order data in order to develop a comprehensive view of customer behavior. It's important to assemble a team that has a healthy mix of technical and business skills.
3. Another important element is to pursue the problem iteratively. Too much data can overwhelm the infrastructure and also befuddle the mind. It is better to divide and conquer the problem with a simpler TDM, with fewer terms and fewer documents and data sources. Expand as needed, in an iterative sequence of steps. In the future, add new terms to help improve predictive accuracy.
4. A variety of data mining tools should be used to test the relationships in the TDM. Different decision tree algorithms could be run alongside cluster analysis and other techniques. Triangulating the findings with multiple techniques, and many what-if scenarios, helps build confidence in the solution. Test the solution in many ways before committing to deploy it.





# Chapter 2: Web Mining

Web mining is the art and science of discovering patterns and insights from the World-wide web so as to improve it. The world-wide web is at the heart of the digital revolution. More data is posted on the web every day than was there on the whole web just 20 years ago. Billions of users are using it every day for a variety of purposes. The web is used for electronic commerce, business communication, and many other applications. Web mining analyzes data from the web and helps find insights that could optimize the web content and improve the user experience. Data for web mining is collected via Web crawlers, web logs, and other means.

Here are some characteristics of optimized websites:

1. *Appearance*: Aesthetic design. Well-formatted content, easy to scan and navigate. Good color contrasts.
2. *Content*: Well planned information architecture with useful content. Fresh content. Search-engine optimized. Links to other good sites.
3. *Functionality*: Accessible to all authorized users. Fast loading times. Usable forms. Mobile enabled.

This type of content and its structure is of interest to ensure the web is easy to use. The analysis of web usage provides feedback on the web content, and also the consumer's browsing habits. This data can be of immense use for commercial advertising, and even for social engineering.

The web could be analyzed for its structure as well as content. The usage pattern of web pages could also be analyzed. Depending upon objectives, web mining can be divided into three different types: Web usage mining, Web content mining and Web structure mining (Figure 2.1).

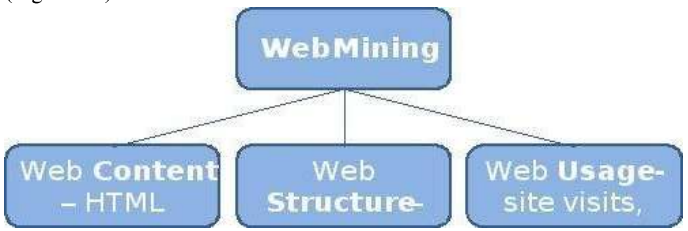



**Figure: 2.1 Web Mining structure**

## 2.1 Web content mining

A website is designed in the form of pages with a distinct URL (universal resource locator). A large website may contain thousands of pages. These pages and their content is managed using specialized software systems called Content Management Systems. Every page can have text, graphics, audio, video, forms, applications, and more kinds of content including user generated content.

The websites keep a record of all requests received for its page/URLs, including the requester information using 'cookies'. The log of these requests could be analyzed to gauge the popularity of those pages among different segments of the population. The text and application content on the pages could be analyzed for its usage by visit counts. The pages on a website themselves could be analyzed for quality of content that attracts most users. Thus the unwanted or unpopular pages could be weeded out, or they can be transformed with different content and style. Similarly, more resources could be assigned to keep the more popular pages more fresh and inviting.

## 2.2 Web structure mining

The Web works through a system of hyperlinks using the hypertext protocol (http). Any page can create a hyperlink to any other page, it can be linked to by another page. The intertwined or self-referral nature of web lends itself to some unique network analytical algorithms. The structure of Web pages could also be analyzed to examine the pattern of hyperlinks among pages. There are two basic strategic models for successful websites: Hubs and Authorities.

1. *Hubs*: These are pages with a large number of interesting links. They serve as a hub, or a gathering point, where people visit to access a variety of information. Media sites like Yahoo.com, or government sites would serve that purpose. More focused sites like Traveladvisor.com and yelp.com could aspire to becoming hubs for new emerging areas.
2. *Authorities*: Ultimately, people would gravitate towards pages that provide the most complete and authoritative information on a particular subject. This could be factual information, news, advice, user reviews etc. These websites would have the most number of inbound links from other websites. Thus Mayoclinic.com would serve as an authoritative page for expert medical opinion. NYtimes.com would serve as an authoritative page for daily news.

## 2.3 Web usage mining

As a user clicks anywhere on a webpage or application, the action is recorded by many entities in many locations. The browser at the client machine will record the click, and the web server providing the content would also make a record of the pages served and the user activity on those pages. The entities between the client and the server, such as the router, proxy server, or ad server, too would record that click.

The goal of web usage mining is to extract useful information and patterns from data generated through Web page visits and transactions. The activity data comes from data stored





in server access logs, referrer logs, agent logs, and client-side cookies. The user characteristics and usage profiles are also gathered directly, or indirectly, through syndicated data. Further, metadata, such as page attributes, content attributes, and usage data are also gathered.

The web content could be analyzed at multiple levels (Figure 2.2).

1. The *server side analysis* would show the relative popularity of the web pages accessed. Those websites could be hubs and authorities.
2. The *client side analysis* could focus on the usage pattern or the actual content consumed and created by users.
    1. Usage pattern could be analyzed using 'clickstream' analysis, i.e. analyzing web activity for patterns of sequence of clicks, and the location and duration of visits on websites. Clickstream analysis can be useful for web activity analysis, software testing, market research, and analyzing employee productivity.
    2. Textual information accessed on the pages retrieved by users could be analyzed using text mining techniques. The text would be gathered and structured using the bag-of-words technique to build a Term-document matrix. This matrix could then be mined using cluster analysis and association rules for patterns such as popular topics, user segmentation, and sentiment analysis.

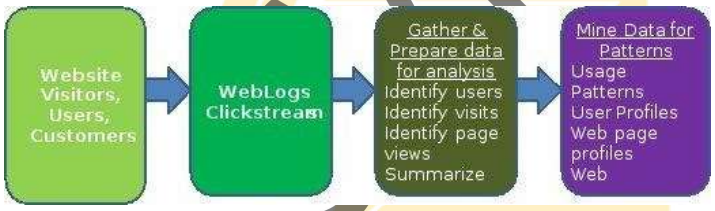



**Figure: 2.2 Web Usage Mining architecture**

Web usage mining has many business applications. It can help predict user behavior based on previously learned rules and users' profiles, and can help determine lifetime value of clients. It can also help design cross-marketing strategies across products, by observing association rules among the pages on the website. Web usage can help evaluate promotional campaigns and see if the users were attracted to the website and used the pages relevant to the campaign. Web usage mining could be used to present dynamic information to users based on their interests and profiles. This includes targeted online ads and coupons at user groups based on user access patterns.

## 2.4 Web Mining Algorithms

Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that rates web pages as being hubs or authorities. Many other HITS-based algorithms have also been published. The most famous and powerful of these algorithms is the PageRank algorithm. Invented by

Google co-founder Larry Page, this algorithm is used by Google to organize the results of its search function. This algorithm helps determine the relative importance of any particular web page by counting the number and quality of links to a page. The websites with more number of links, and/or more links from higher-quality websites, will be ranked higher. It works in a similar way as determining the status of a person in a society of people. Those with relations to more people and/or relations to people of higher status will be accorded a higher status.

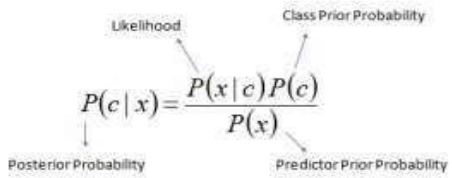
PageRank is the algorithm that helps determine the order of pages listed upon a Google Search query. The original PageRank algorithm formuation has been updated in many ways and the latest algorithm is kept a secret so other websites cannot take advantage of the algorithm and manipulate their website according to it. However, there are many standard elements that remain unchanged. These elements lead to the principles for a good website. This process is also called Search Engine Optimization (SEO).





# Chapter 3
# Naïve Bayes Analysis

Naïve Bayes technique is a is supervised machine learning technique that that uses probability theory based analysis.
It is machine learning technique that computes the probabilities of an instance of belonging to each of many target classes, given the prior probabilities of classification using individual factors.

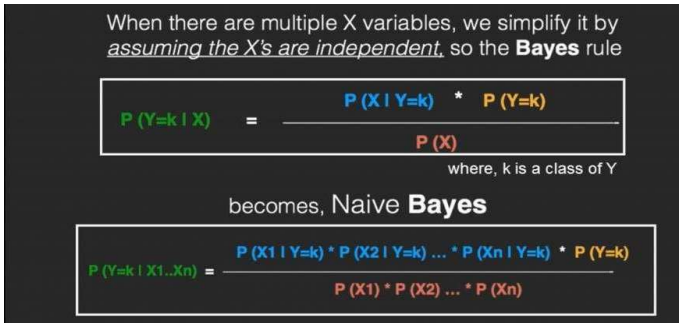


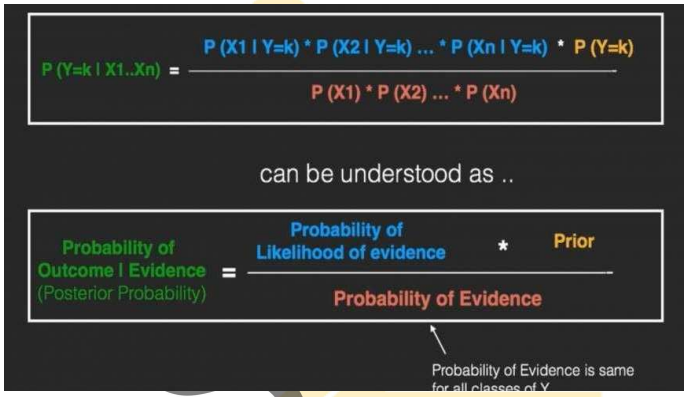

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).
- P(c) is the prior probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

## 3.1 Probability

- The Bayes Rule provides the formula for the probability of Y given X. But, in real-world problems, you typically have multiple X variables.
- When the features are independent, we can extend the Bayes Rule to what is called Naive Bayes.
- It is called 'Naive' because of the naive assumption that the X's are independent of each other. Regardless of its name, it's a powerful formula.




- In technical jargon, the left-hand-side (LHS) of the equation is understood as the posterior probability or simply the posterior.
- The RHS has 2 terms in the numerator.
- The first term is called the 'Likelihood of Evidence'. It is nothing but the conditional probability of each X's given Y is of particular class 'c'.
- Since all the X's are assumed to be independent of each other, you can just multiply the 'likelihoods' of all the X's and called it the 'Probability of likelihood of evidence'. This is known from the training dataset by filtering records where Y=c.
- The second term is called the prior which is the overall probability of Y=c, where c is a class of Y. In simpler terms, Prior = count(Y=c) / n_Records.





- An example is better than an hour of theory. So let's see one

**Naive Bayes Example**

- Say you have 1000 fruits which could be either 'banana', 'orange' or 'other'.
- These are the 3 possible classes of the Y variable.
- We have data for the following X variables, all of which are binary (1 or 0).
- Long
- Sweet
- Yellow
- The first few rows of the training dataset look like this:

| Fruit | Long (x1) | Sweet (x2) | Yellow (x3) |
|---|---|---|---|
| Orange | 0 | 1 | 0 |
| Banana | 1 | 0 | 1 |
| Banana | 1 | 1 | 1 |
| Other | 1 | 1 | 0 |

- For the sake of computing the probabilities, let's aggregate the training data to form a counts table like this.

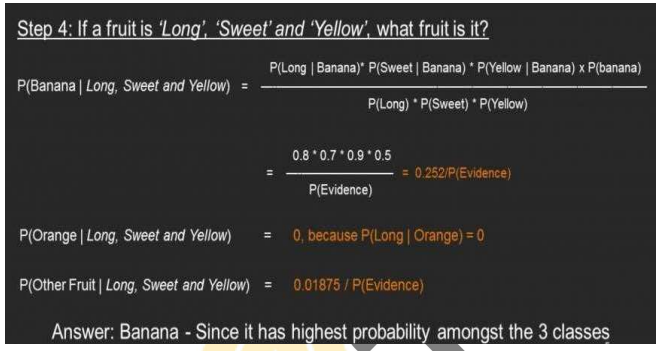
| Type | Long | Not Long | Sweet | Not Sweet | Yellow | Not Yellow | Total |
|---|---|---|---|---|---|---|---|
| Banana | 400 | 100 | 350 | 150 | 450 | 50 | 500 |
| Orange | 0 | 300 | 150 | 150 | 300 | 0 | 300 |
| Other | 100 | 100 | 150 | 50 | 50 | 150 | 200 |
| Total | 500 | 500 | 650 | 350 | 800 | 200 | 1000 |

- So the objective of the classifier is to predict if a given fruit is a 'Banana' or 'Orange' or 'Other' when only the 3 features (long, sweet and yellow) are known.
- Let's say you are given a fruit that is: Long, Sweet and Yellow, can you predict what fruit it is?
- This is the same of predicting the Y when only the X variables in testing data are known. Let's solve it by hand using Naive Bayes.
- The idea is to compute the 3 probabilities, that is the probability of the fruit being a banana, orange or other. Whichever fruit type gets the highest probability wins.
- All the information to calculate these probabilities is present in the above tabulation.
- Step 1: Compute the 'Prior' probabilities for each of the class of fruits.
  - P(Y=Banana) = 500 / 1000 = 0.50
  - P(Y=Orange) = 300 / 1000 = 0.30
  - P(Y=Other) = 200 / 1000 = 0.20
- Step 2: Compute the probability of evidence that goes in the denominator.
  - P(x1=Long) = 500 / 1000 = 0.50
  - P(x2=Sweet) = 650 / 1000 = 0.65
  - P(x3=Yellow) = 800 / 1000 = 0.80
- Step 3: Compute the probability of likelihood of evidences that goes in the numerator.
  - Here, I have done it for Banana alone.
  - Probability of Likelihood for Banana
  - P(x1=Long | Y=Banana) = 400 / 500 = 0.80
  - P(x2=Sweet | Y=Banana) = 350 / 500 = 0.70





  - P(x3=Yellow | Y=Banana) = 450 / 500 = 0.90
- Step 4: Substitute all the 3 equations into the Naive Bayes formula, to get the probability that it is a banana.




## Advantages

- When assumption of independent predictors holds true, a Naive Bayes classifier performs better as compared to other models
- Naive Bayes requires a small amount of training data to estimate the test data. So, the training period is less.
- Naive Bayes is also easy to implement.

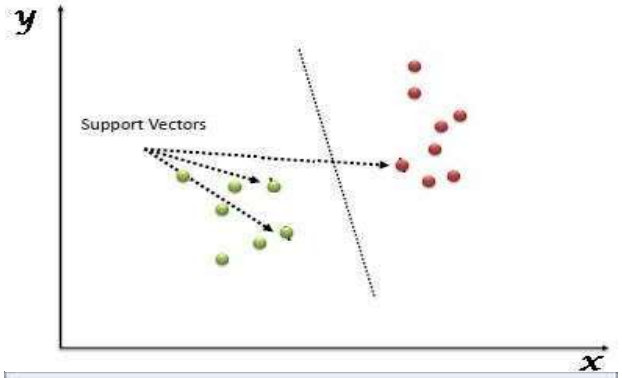
## Disadvantages

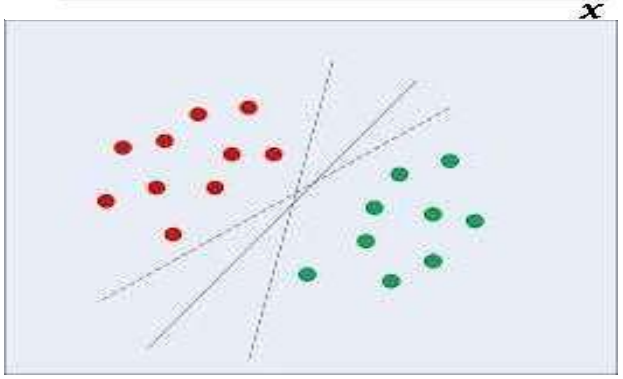
- Main imitation of Naive Bayes is the assumption of independent predictors. Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent.
- If categorical variable has a category in test data set, which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.

# Chapter 5

## Support Vector Machine

- "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 this 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 differentiate the two classes very well (look at the below snapshot).



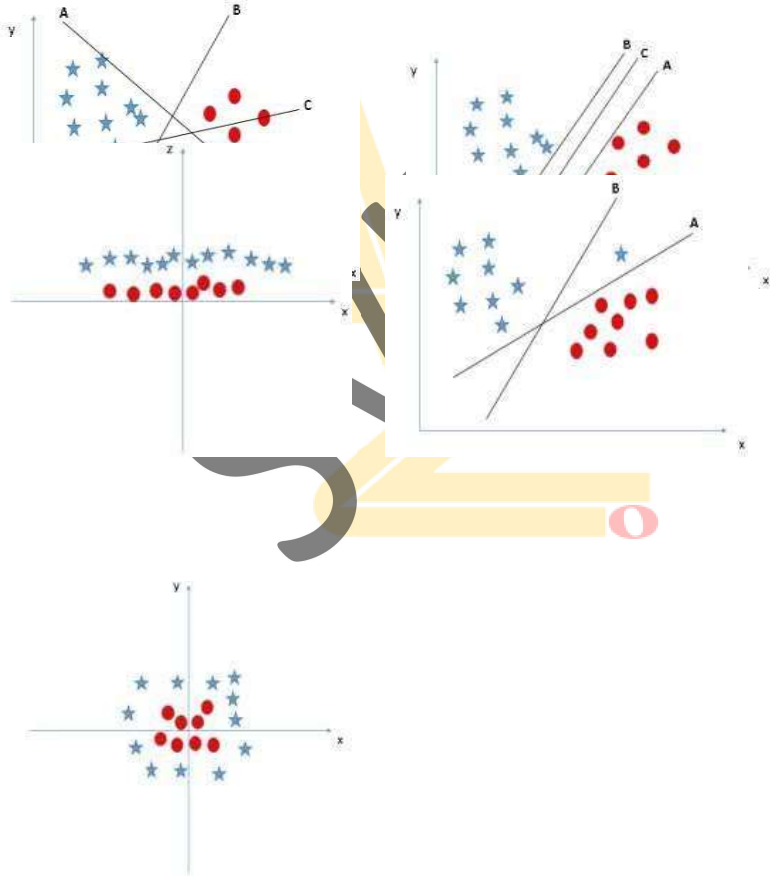





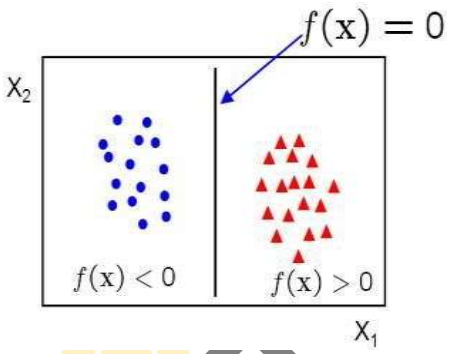


**How does it work?**

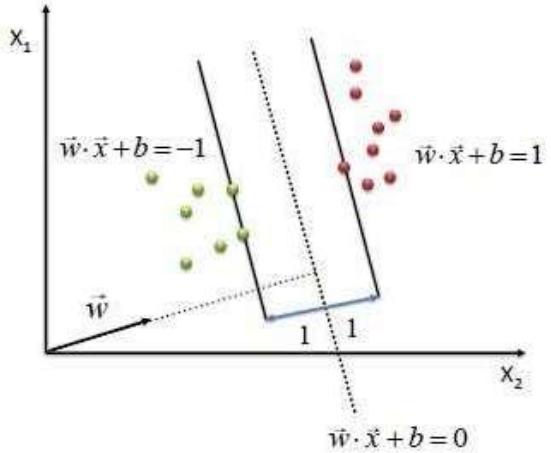
- Thumb rule to identify the right hyper-plane
- • Select the hyper-plane which segregates the two classes better
- • Maximizing the distances between nearest data point (either class) and hyper-plane. This distance is called as Margin.



## SVM Model




- f(x) = W.X + b
- W is the normal to the line, X is input vector and b the bias
- W is known as the weight vector




$$\max \frac{2}{\|w\|}$$

s.t.

$(w\cdot x+b)\geq 1, \forall x$ of class 1

$(w\cdot x+b)\leq -1, \forall x$ of class 2

## Advantages of SVM

- The main strength of SVM is that they work well even when the number of SVM features is much larger than the number of instances.
- It can work on datasets with huge feature space, such is the case in spam filtering, where a large number of words are the potential signifiers of a message being spam.





- Even when the optimal decision boundary is a nonlinear curve, the SVM transforms the variables to create new dimensions such that the representation of the classifier is a linear function of those transformed dimensions of the data.
- SVMs are conceptually easy to understand. They create an easy-to- understand linear classifier. By working on only a subset of relevant data,. they are computationally efficient. SVMs are now available with almost all data analytics toolsets.

## Disadvantages of SVM

The SVM technique has two major constraints

- It works well only with real numbers, i.e., all the data points in all the dimensions must be defined by numeric values only,
- It works only with binary classification problems. One can make a series of cascaded SVMs to get around this constraint.
- Training the SVMs is an inefficient and time consuming process, when the data is large.
- It does not work well when there is much noise in the data, and thus has to compute soft margins.
- The SVMs will also not provide a probability estimate of classification, i.e., the confidence level for classifying an instance.