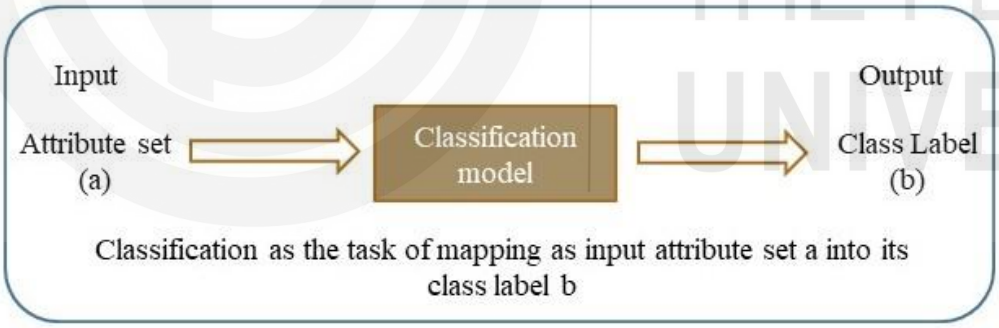
UNIT 10 CLASSIFICATION

10.0 INTRODUCTION

Data mining:
Classification is a single step in the data mining process. It is used for organizing objects based on some key features. Several approaches such as the nearest neighbor classification, Decision Tree Learning and Support Vector Machines are employed for data mining classification.

10.1 DEFINITION OF CLASSIFICATION

A classification problem requires that examples be classified into one of two or more classes. A classification can include reliable or discrete input variables. A two-class dilemma is also regarded as a challenge between two classes or binary classification. A multi-class grouping problem is often called a two-group problem.
Classification is the activity of learning a target function "c," which maps each attribute to one of the class labels "b".



Classification as the task of mapping as input attribute set a into its class label b

The attribute set 'a' can contain a number of attributes and can contain binary, categorical and continuous attributes. The class 'b' mark must have a distinct, i.e. binary or categorical attribute (nominal or ordinal).

Classification Models:

Descriptive Modeling:

It is a model of classification used to summarize data.

Predictive Modeling:

It is a classification model used to predict the class mark of unknown data.

Applications of Classification Models:

- Spam detection e-mails based on the header and content of the document.
- Classification of galaxies according to their types.
- Classification of students according to their qualifications.
- Patients are classified according to their medical history.
- Classification can be used for the approval of credit.

10.2 GENERAL APPROACHES TO SOLVING A CLASSIFICATION PROBLEM

- A classification technique is a systematic way to construct classification models based on data collection.
- Examples include decision tree classification machines, rule-based classification machines, neural networks, vector support machines and Bayes naive classifiers.
- Each technique uses the learning algorithm to find a template that best matches the relationship between the attribute set and the class mark of the data input.
- A training array consists of records that identify class marks. The training test is used for the creation of a classification model for the test set. The test collection consists of records with an unspecified class mark
- The estimation of the results of the classification model would be based on the number of test records correctly and incorrectly estimated by the model.
- These numbers are presented in a table called an uncertainty matrix.

10.3 EVALUATION OF CLASSIFIERS

It's probably wrong if it's easy.

You could have attacked a few data problems if you are new to a data science course or try to gather the bases yourself. You also played with various classifiers, feature sets, and perhaps different sets of parameters, etc. Finally, you can think about the evaluation: how well does your classification work?

I presume that you already know the basic evaluation methodology: you do not know how to assess a training set, maybe cross validation, and maybe even trust bars or even t-tests. It's Data Science 101, which you already know right now. Ok, that's not what I'm talking about. I don't talk about that. I don't talk about that. What I'm talking about is something more simple but harder: how do you calculate your classifier's performance?

The obvious response is to use accuracy: accurate classification of the number of instances. You have a classifier which takes test examples and takes classes for each class. His guess in any test example is either correct or wrong. You basically calculate how many right choices your classification makes, the difference between the number and accuracy of your classification is the total number of test examples. It's just so easy. Most findings of study are accurate and many practical ventures accurate. That's the default metric. What's the fuck with that? Precision is a simple measure that misleads several issues in the real world. Indeed, "but it succeeded in the laboratory, why does it not work in the field?" experience is best achieved by using accuracy as an evaluation metric.

What's wrong with accuracy?

Let us begin with a single two-class domain to illustrate the issues. Each classifier sees examples for this domain from the two classes and offers one of two possible rulings: Y or N. Every decision can be made with a test set and a particular classifier:

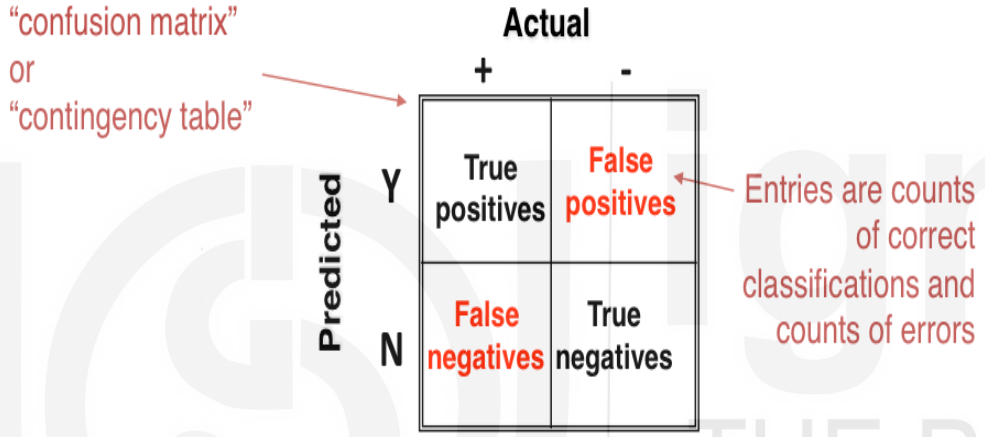
1. a positive example that has been graded as positive. This is a genuinely positive development.

2. A good example that is misclassified as negative. That's a wrong negative.

3. a negative case that was labelled as negative. That's a real negative.

4. A bad example is misclassified as good. That's a false positive thing.

A two-to-two matrix can be generated with the columns as true classes and the rows as hypothesized classes. It looks as follows:



Each case contributes exactly to the count in a single cell. This is your approach to your classification.
Precision is just the diagonal number divided by the total:

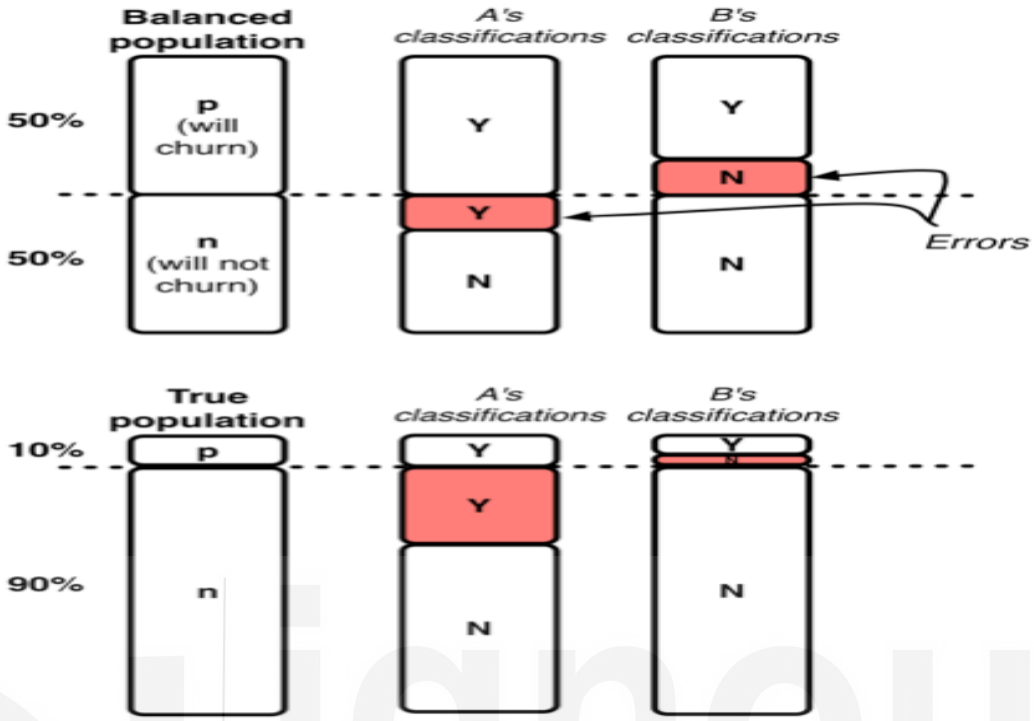
$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{\text{True positives} + \text{False positives} + \text{True negatives} + \text{False negatives}}$$

What is wrong with it?

The issue of class imbalance:

Accuracy only includes all examples and records a percentage of correct answers. Accuracy can be great when working with balanced (or roughly balanced) datasets. The higher you get from 50/50, the greater the precision. Consider a 99:1 divided negative to positive dataset. Only guessing that the majority class yields a 99% correct description!
Unbalanced realms are the law in the real world, not the exception! I have never handled a balanced dataset outside academia. The explanation is very simple. Machine learning is also used in the processing of populations consisting of a large number of negative (uninteresting) examples and a small number of positive examples (interesting, alarming).
For eg, fraud detection (1% is an immense problem; 0.1% is closer), screening for HIV (the incidence in the USA is around ~0.4%), prediction of disc drive failure (~1%

a year), click-through response rates (0.09%), development error rates (0.1%) - the list goes a long way. Unbalance is anywhere. Everywhere.



Why is this difficult assessment? An example is shown in the left diagram. We want to see which customers will probably churn (cancel their contracts) soon.
The top half reveals the trainers. The population has been balanced, so half is positive and half is negative. Two classifiers, A and B, have been qualified. They each make the same error percentage, so their accuracy in this particular data set is the same as seen in the top half by the red areas.
In the lower half the same two classifiers appear in a test setting, where the population is now 10% (will churn) to 90% (will not churn). That's far more practical. Notice that A is even worse now because its mistakes are false positives.
Two lessons are drawn from this example. One is that you need to know more than what accuracy tells you about your classifier. The other is that, regardless of the population in which you train, you still have to test the population you care for.

The question of error costs:

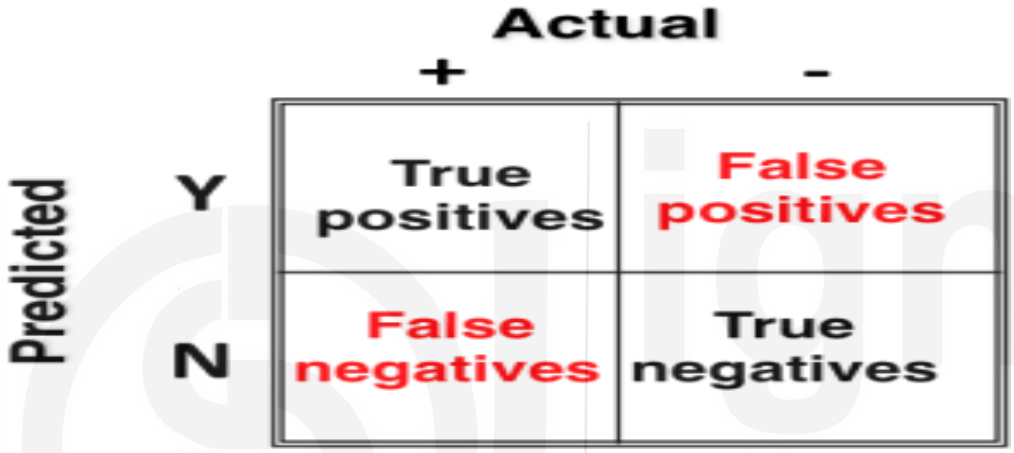
A second issue remains. Take the realm of click-through advertisement replies. The estimated click-through rate (previous class) is around 0.09 percent. That means that a classifier saying, "No, the consumer won't press," has a 99.91% accuracy. What is wrong with this? What is wrong with it? Nothing, strictly speaking! This is excellent performance—if you cared about precision.
The problem is that certain classes are far more important than others. In this scenario, you think much more for those who press 0.09% than about the 99.91% who don't. More specifically, it is not a big mistake to show an ad to certain people who don't click on it (a False Positive in the matrix above), but it isn't bad for people who want to click on an ad to not show it (a False Negative) is much worse. After all, aiming at the right people for the ad is the whole reason that you create a classifier first.
In several domains, different error costs are normal. In medical diagnoses, where there is no cancer, the expense is different from the lack of a true event. The cost of finding an insignificant document in document classification is different from the cost of losing an important document, and so on.

What about collections of public data?

You may have found that you've been working with databases from somewhere like the UCI Dataset Repository that do not seem to have these problems. The classes are always reasonably balanced, and it seems accurate and unproblematic when error costs are given. While these data sets may have originated with real issues, they have been artificially cleaned, balanced and provided with accurate error costs in many cases. Contest managers must be able to judge and announce a definite winner, so the judging criteria are squarely defined beforehand.

Other measurements

At this point, you are certain that the precision of plain classification is a poor measurement for real-world domains. Instead, what do you use? Many other assessment criteria were created. It is important to note that and way of synthesizing the confusion matrix is simply different.



Here are some metrics that you can probably find:

- **True positive rate** = TP/(TP+FN) = 1 − false negative rate
- **False positive rate** = FP/(FP+TN) = 1 − true negative rate
- **sensitivity** = true positive rate
- **specificity** = true negative rate
- **positive predictive value** = TP/(TP+FP)
- **recall** = TP / (TP+FN) = true positive rate
- **precision** = TP / (TP+FP)

  F-score is the harmonic mean of precision and recall:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

  G-score is the geometric mean of precision and recall:

$$G = \sqrt{\text{precision} \cdot \text{recall}}$$

Two other steps are worth noting, but their precise meanings are too complex to clarify. Cohen's Kappa is an agreement metric that takes into account the amount of agreement that chance can expect. It does this by taking into consideration class imbalances and the propensity of the classifier to vote yes or no. The White Man U test is a general measure of the classifier's separation ability. It tests how well the classifier can classify positive instances over negative cases.

Expected cost:

Which is the right number to calculate, given all these numbers? We will avoid the usual statistical response ("It depends"), and reply: none of the above-mentioned.
The uncertainty matrix involves frequencies of the four results. The most accurate solution to the problem is to measure the estimated costs by using these four numbers (or equivalently, expected benefit). You should note that the normal form of an anticipated value is to take and calculate the likelihood of each circumstance by its value. We translate the numbers in the above matrix to cost and gain for each of the four. The number is the final calculation:
Planned costs = p(p) for cost(p) + p(n) for cost(s) (n)
Here p(p) and p(n) are each class's probabilities, also known as the class priors. Cost(p) is shorthand for the cost of a positive example. What is this? What is this? We need to divide it into two instances: if we get it right (a real good one), and if we go wrong (a false negative).
cost(p) = p(true positive) value(true positive) + p(false negative) value added cost (false negative)
Likewise, for the negative class: it is a true negative, if we have it correctly, otherwise it is a false positive. When we extend the entire cost equation, we get the messy (but hopefully comprehensible) equation:
Expected cost = p(**p**) × [*p*(true positive) × benefit(true positive)

+ *p*(false negative) × cost(false negative)]

+ p(**n**) × [*p*(true negative) × benefit(true negative)

+ *p*(false positive) × cost(false positive)]

This is much more accurate than precision, since it divides each of the four options for any choice and multiplies it by the costs or benefits associated with that decision. You can estimate the priors p(p) and p(n) directly from your results. The four p(true positive) expressions you obtain from your classifier; specifically, from the confusion matrix which comes from a test sample classifier. The value cost () and advantage () are extrinsic values which cannot be extracted from data. They quantify the actual effects of the decisions of the classifier. These must be calculated on the basis of expert domain information.
Ideally, you can test a classifier if you want its output to be reduced to a single number. If you have a set of classificatory and want to select the best, this method allows you to precisely measure the performance of each classifier.

10.4 CLASSIFICATION TECHNIQUES

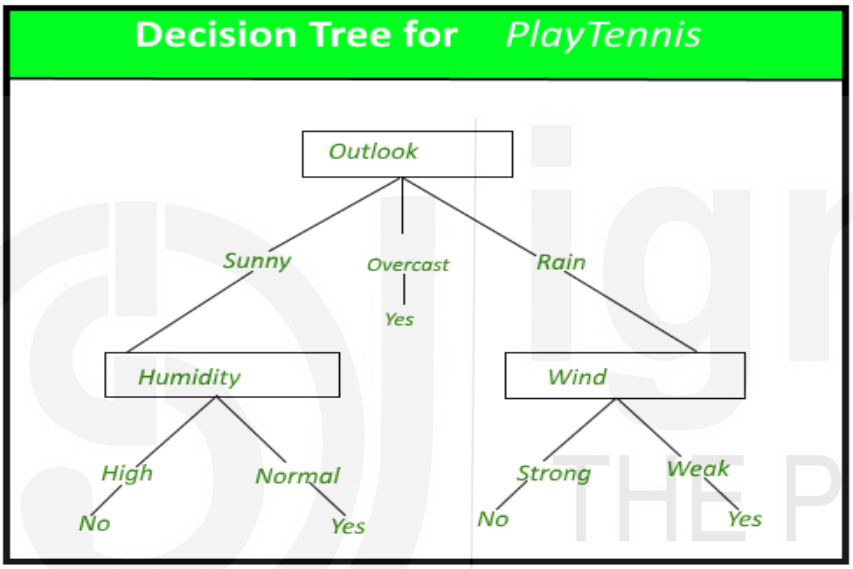
Classification is a technique used to predict group membership for data cases through data mining (machine learning). There are many methods of classification that can be used for classification purposes. We present the basic classification techniques in this chapter. Later on, we spoke about some key classification methods, including Bayesian networks, decision tree inference, neighbour classification, and SVMs, which have strengths, limitations, possible implementations and problems with their solution. The purpose of this study is to provide an exhaustive analysis of various classification techniques in machine learning. This work would allow both academia and new academics to further consolidate the foundation of the classification methods.
Classification is used to detect which category is connected to each data instance within a given dataset. It is used to classify data according to certain constraints into various groups. For classification a variety of significant classification algorithms are used, such as C4.5, ID3, k-nearest neighbour classifier, Naive Bayes, SVM and ANN. A classification methodology generally follows three classification approaches: statistical, machine learning, and neural network. This paper offers an inclusive

survey of various classification algorithms and their characteristics and weaknesses, thus examining these approaches.

10.5 DISCISION TREES – DECISION TREE CONSTRUCTION

Decision Tree:

Decision tree is the most effective and common prediction and classification method. A decision tree is a flux chart-like tree structure, where each internal node indicates an attribute test, each branch is the product of the test, and each leaf node has a class name.



Construction of Decision Tree:

A tree can be "learned" by dividing the source into subsets based on a value test attribute. This process is replicated in a recursive way, called recursive partitioning on each derived subset. The recursion is done when the subset at a node has all the same value as the target variable or when splitting doesn't add any more value to the forecasts. The construction of a decision tree classification involves no domain knowledge or parameter setting, so it is ideal for exploration of explorative knowledge. High dimensional data can be handled by decision trees. The decision tree classifier has excellent accuracy in general. Induction of decision tree is a standard inductive way of learning classification information.

Decision Tree Representation:

Decision trees identify instances by sorting them from the root to a leaf node that provides the instance classification. An instance is classified by starting from the tree root node, evaluating the specified attribute by this node, then moving down the tree branch corresponding to the attribute's value, as shown in the figure above. This process is repeated for the subtree that is rooted in the new node.

The decision tree in the figure above classifies the tennis court and the leaf classification as suitable for a certain morning. (In this situation, yes or no).
For example, the instance

(Outlook = Rain, Temperature = Hot, Humidity = High, Wind = Strong)

The left branch of this decision tree will be sorted down and therefore classified as a negative case.
In other words, we can assume that the decision tree represents a breakdown of constraints on instance attribute values.

(Outlook = Sunny ^ Humidity = Normal) v (Outllok = Overcast) v (Outlook = Rain ^ Wind = Weak)

Strengths and Weakness of Decision Tree approach:

The strengths of decision tree approaches are: • Decision trees can produce comprehensible rules.

- Classification of decision trees without much computation.
- Decision trees can accommodate both categorical and continuous variables.
- Decision trees clearly show which fields for prediction or classification are most important.
- The drawbacks of decision tree methods: • Decision trees are less suited to estimate tasks where the goal is to predict a constant attribute value.
- The decision trees are vulnerable to mistakes in many class problems and relatively limited numbers of training instances.
- Decision tree can be costly to train computationally. The decision-making method is computationally costly. Each splitting field must be sorted at each node before the best split can be identified. Combinations of fields are used in some algorithms and search must be made for optimal combined weights. Pruning algorithms can also be costly as many sub-trees of candidates have to be created and compared.

Measures of for selecting the Best Spilt:

There are several tests to decide the best way to split data.
Let P(i|t) indicate the fraction of Class I records at a t node. Measures to pick the best split also rely on the child's nodes' degree of impurity. The less impurity, the more the class distribution distorted. For eg, the class node (0.1) has no impurity, while the class node (0.5,0.5) has the highest impurity.

$$Entropy\ (t) = -\sum_{i=0}^{c-1} p(i|t) log_2 p(i|t)$$

$$\text{Gini}\ (t) = 1 - \sum_{i=0}^{c-1} [p(i|t)^2$$

$$\text{Classification error}\ (t) = 1 - \max[p(i|t)]$$

Splitting of Nominal Attributes:

For a nominal attribute, one-to-one mapping, you can have either binary or multi splits.
For equal distributions, the calculation of the Gini index is the same. It is preferable to have a smaller overall GINI index, but you cannot exclude single sections of the index. In our case, it's the multi-way split.

Splitting of Continuous Attributes:

To obtain a continuous property from a continuous attribute, we make a selection.
The values in our example have an increasing order of distinctness.
Between two positions of values that are adjacent to each other, is equal to the mean of the values on either side of that one.
The GINI index must be calculated for any attribute split split. The smaller GIN index can be used as a range if continuous is chosen.

Model Overfitting:

A classification model is made up of two groups of errors: misspellings and misclassifications.

- Training errors
- Generalization errors

Training Errors:

The percentage of training records that have errors due to mal-classification. or "To give you an example, a record in the training data has already existed in the test data, but the class prediction was incorrect." Errors of this kind are referred to as training errors.

Generalization Errors:

The model's forecasted variance. Let's assume for example the class label for the test data is under the last data point is identified, but it is the opposite of what is observed. It is difficult to estimate exactly how many errors of this kind are known as Extrapolation errors.
A good model should have a low number of faults while being easy to learn and validate.
It takes a great deal of trial and error when the tree is small. a suboptimal relationship between the model and the data.
The error rate increases as the tree grows, but the training error decreases. Using all available training data will lead to the phenomenon known as overfitting.
In this example, the tree with less nodes, there are more errors in training, and less in testing.

Methods for selecting the best split:

Unless a random state function is defined, train test split divides arrays into random subsets. The optimal split for training and research is said to be 80:20. You will have to change it according to the data set size and the complexity of the parameter.

Algorithm for Decision Tree Induction:

Bayesian Classification:

Baye's Theorem:

The theorem of Bayes named after British mathematician Thomas Bayes from the 18th century is a mathematical method to determine the probability of a condition. Conditional probability is the probability of an outcome, based on a prior result. The theorem of Bayes offers a way to revisit current predictions or hypotheses (update probabilities) with new or additional proof. The Bayes theorem can be used in finance to assess the danger of lending capital to potential lenders.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Naive-Bayesian Classification:

In statistics, Bayes naive classifiers are a family of simple "probabilistic classifiers" based on the theorem of Bayes, which has clear independence assumptions between characteristics. They are among the simplest Bayesian network models, but they achieve higher accuracy levels along with kernel density estimates. Classification systems from Naïve Bayes are very scalable and require a number of parameters in linear numbers in a learning problem.

Bayesian Belief Networks:

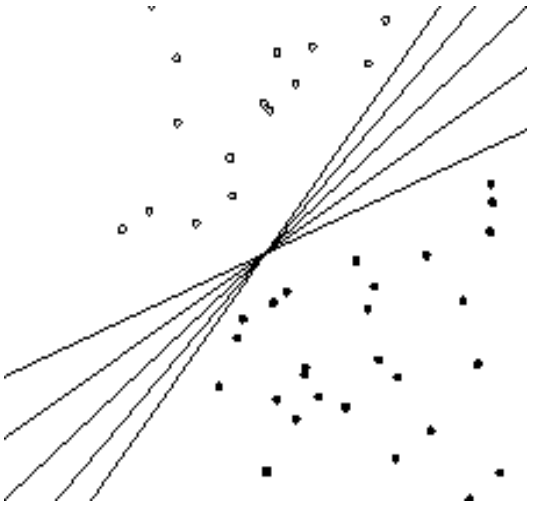
A Bayesian network (also known as a Bayes network, a belief network, or a decision-making network) is a probable graphical model representing a number of variables and their dependence through a directed acyclic graph (DAG). Bayesian networks are suitable for an incident that has happened and are likely to contribute to one of the potential established triggers.

Support Vector Machines:

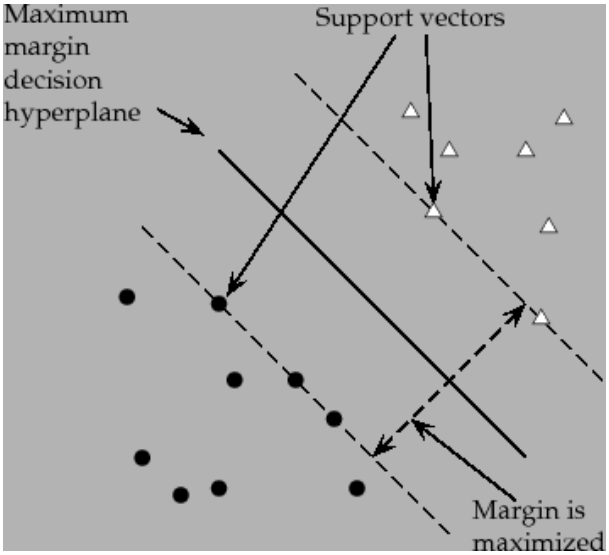
Support Vector Machine or SVM is a simple but effective algorithm for supervised machine learning that can be used to create both regression and classification models. SVM algorithm for both linear and non-linear separable datasets will work well.

Support vector machines: The linearly separable case

In this part, we show that Naive Bayes and Roccio are two study methods for linear classifiers, perhaps the largest text classifiers group. To simplify the discussion, this section only covers two classifiers and a linear classifier is defined when the linear combination of characteristics with one threshold is compared to a two classifier.



There are an infinite number of hyperplanes that separate two linearly separable classes.



The support vectors are the 5 points right up against the margin of the classifier.

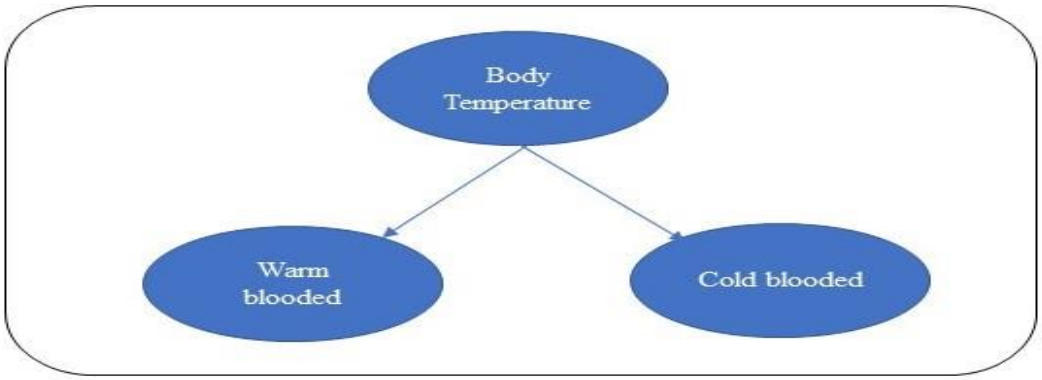
There are several possible linear separators in two-class, separable training data sets, like the one in Figure. An intuitively clear boundary between the data items of the two classes appears better than one that approaches examples of one or both classes very closely. Although some learning methods such as perceptron algorithms find just any linear separator, others, such as Naive Bayes, look for some criteria for the best linear separator. In particular, the SVM determines the requirement for a decision-making surface that is as far removed as possible from any data point. The distance between the decision surface and the nearest data point decides the classifier's margin. This construction method necessarily implies that the SVM decision function is defined entirely by a (usually small) data sub-set defining the separator location. These points are known as support vectors (in a vector space, a point can be thought of as a vector between the origin and that point). The figure indicates the sample problem margin and support vectors. Other data points do not play a role in deciding the chosen decision surface.

10.6 METHODS FOR EXPRESSING ATTRIBUTE TEST CONDITIONS

The methods used to express test conditions for attributes are as follows. You are You are:

Binary Attribute:

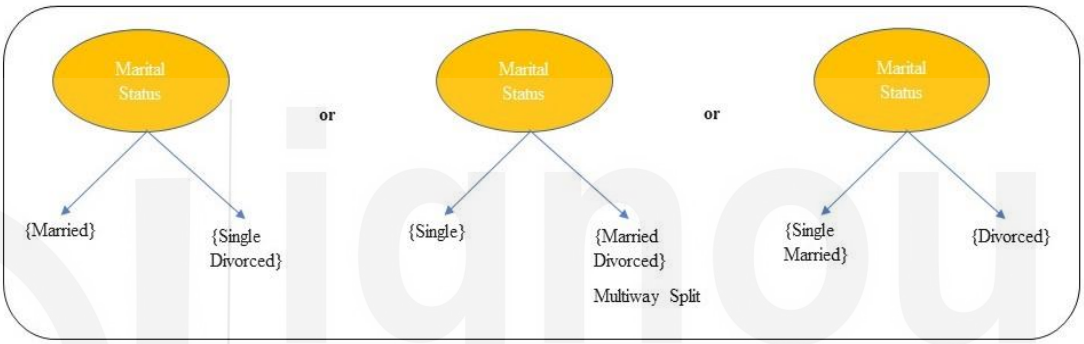
The binary attribute test condition generates two results as shown below.



Nominal Attributes:

As many values may exist for a nominal attribute, its test conditions can be expressed in two respects as shown below.

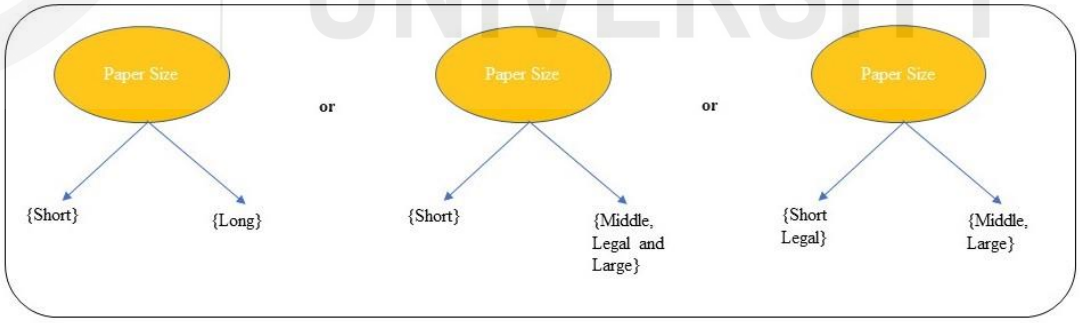
The number of outcomes for a multi-way division depends on the number of different values for the attribute.



Some algorithms, like CART, only support binary divisions. In this case, the k-attribute values can be split into 2k-1-1. For instance, the marital status is a value of three attributes, we can divide it in 22-1-1; that means 3 different ways.

Ordinal Attribute:

It can also create splits in binary or multiple ways. Ordinal attribute values may be grouped unless the grouping violates the attribute value's order property.



Different ways of grouping ordinal attribute values

In the example above, condition A and condition B fulfil order, but condition C violates the property of the order.

Continuous Attributes:

The test condition may be expressed as comparison (A<v) or (A>=v) test with binary results, or a query in the range of results with the results vi<=A<vi+1 for i=1,2,...,k.

UNIT 11 CLUSTERING

Clustering: Problem Definition:

A group of comparable data objects is classed as a cluster in clustering. A collection of data is referred to as a group. The cluster analysis divides data into groupings based on how closely they resemble one another. A label is assigned to a group once the data has been classified into several groups. By doing the classification, it aids in adaptation to shifts.

Clustering is an unsupervised Machine Learning-based Algorithm that divides a set of data points into clusters so that the objects are all part of the same group. Using clustering, data can be subdivided into smaller, more manageable groups. Data from each of these subgroups are grouped together into a single cluster. Because our client data has been segmented into groups, we're in a better position to determine who could benefit most from this offering.

- To put it simply, a cluster is a group of related things.
- It's a grouping of objects where the distance between any two members is less than the distance between any two people.
- A multidimensional space segment with a high density of objects that is related to the other segments.

Data mining Cluster Analysis: What Is It?

Data mining's Cluster Analysis refers to the process of identifying groupings of items that are similar in some respects but differ from one another in other respects.

A cluster is a grouping of comparable items that belongs to the same category.

A set of items in which the distance between any two things in the cluster is less than the distance between any two objects in the cluster and any object not placed inside the cluster.

High-density region related to other regions in multidimensional space.

Clustering is a technique used to turn a collection of unrelated abstract objects into a set of related classes.

It is a technique for dividing large collections of data or objects into smaller groups called clusters.

As a stand-alone tool, it provides insight into data distribution and can be used as a pre-processing step for other algorithms or as a pre-processing step in its own right.

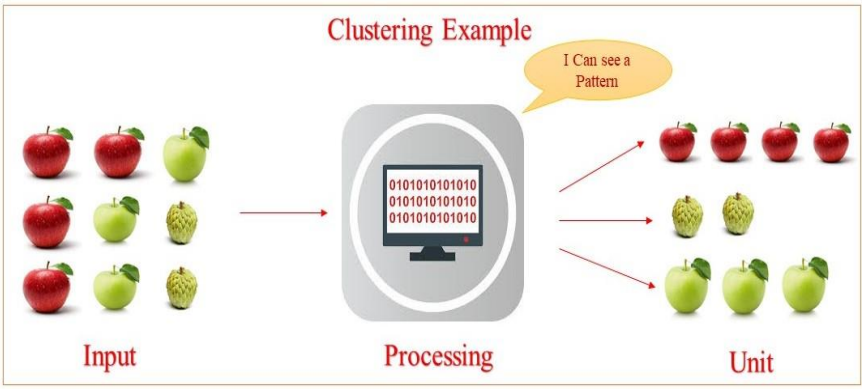


Fig: Example of Clustering

Uses for cluster analysis in data mining:

Data clustering analysis has a wide range of applications, including image processing, data analysis, pattern identification, and market research. With the use of Data clustering, firms can find new client groupings in their database. Buying patterns can also be used to classify data.

Data mining clustering aids in the classification of animals and plants by utilizing comparable functions or genes in biology. Finding out more about how species are structured becomes easier with this information. Data mining clustering identifies geographic areas. There are regions in the earth observation database that are similar to each other.

In the city, a particular type of dwelling is included in a specific neighborhoods. Information can be discovered more easily by classifying online files using clustering in data mining. Additionally, it's employed in detection software. If a credit card is being used fraudulently, the pattern of deceit can be identified using clustering in data mining.

To better understand each cluster, this information is useful. Using this tool, one can gain insight into the distribution of data and use it to do data mining tasks.

A cluster's data objects can be thought of as one big group.

While performing cluster analysis, we first divide the data set into groups. It allocates levels to groups based on data similarities.

An advantage of over-classification is that it is flexible and can assist identify crucial qualities that set differentiating groups apart.

Applications of cluster analysis in Data Mining:

- Clustering analysis is widely utilized in a variety of fields, including data analysis, market research, pattern identification, and image processing.
- It aids marketers in identifying distinct customer segments based on their customers' purchasing habits. They are able to classify their clients.
- It aids data discovery by assigning documents on the internet.
- For example, credit card fraud detection relies on clustering.
- Cluster analysis is a data mining function that provides insight into data distribution so that the properties of each cluster may be analysed.
- It can be used in biology to figure out plant and animal taxonomies, classify genes with similar functions, and have a better understanding of population structure.
- Earth observation databases use this data to identify similar land regions and to group houses in a city based on house type, value, and geographic position.

Clustering Requirements for Data Mining:

Its multiplicity of uses has made clustering analysis an emerging data mining topic. Due to the recent development of a variety of data clustering toolsets and their extensive use in a variety of fields, such as image processing and computational biology to mobile communication and medicine these methods are likely to become more prominent in the future. It's impossible to standardise data clustering algorithms, which is the fundamental problem. Advanced algorithms may work well with some data sets, but they fail or perform poorly when used with other sets. Despite numerous attempts to standardize algorithms that work effectively in all scenarios, little progress has been accomplished. So far, a slew of clustering utilities have been proposed. However, each method has pros and cons and isn't applicable in all real-world scenarios.

Interpretability:

Clustering should produce useable, intelligible, and interpretable results.

Clustering is a powerful tool.

Facilitates handling of stale or corrupted data:

Most of the time, the data is jumbled and unorganized. Because it can't be analyzed fast, data mining relies heavily on clustering. By organizing the data into groups of related data objects, grouping can provide some structure to the data. It becomes easier for the data specialist to process the data and also to uncover new things in the process of processing data.

High Dimensional:

Using data clustering, you'll be able to work with large and little data of any dimension.

Discovery of attribute clusters based on shape:

The technique for clustering detects groups of arbitrary shapes. There's also a little cluster here with a spherical form.

Usability of algorithms while dealing with various types of data:

Using clustering methods, a wide range of data types can be analyzed. Binary data, categorical data, and interval-based data are examples of possible data types.

Clustering Scalability:

Dealing with a large database is commonplace. The method must be scalable in order to deal with large databases.

Categorization of Major Clustering Methods:

The following categories describe several clustering techniques

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method:

Let's say "a" partitioning is done on the database's "b" objects in this method. Each partition and a<b will be used to represent a cluster. After the classification of items, C is the number of groups. There are a few requirements that must be met with this Partitioning Clustering Method, and they are as follows:

- Only one group should have a single objective at any given time.
- There should be no organization that does not have a single goal.

When using a Partitioning Clustering Method, keep in mind the following considerations:

- If we provide the number of partitions up front, there will be an initial partitioning (say m).
- Iterative relocation is a strategy that involves moving an object from one group to another in order to enhance partitioning.

K-Means Algorithm:

Clustering by K-Means is a machine learning or data science problem-solving technique that utilizes an unsupervised learning method. This article will teach you about the K-means clustering technique, how it works, and how to use it in Python.

What is K-Means Algorithm:

Using K-Means Clustering, an approach known as Unsupervised Learning, the unlabeled dataset can be divided into several clusters for analysis. For example, if K=2, then there will be two clusters, and if K=3, then there will be three clusters, and so on. K indicates the number of pre-defined clusters that need to be produced in the process.

With this approach, the unlabeled dataset is divided into as many clusters as possible, so that each dataset only belongs to one group.

There is no requirement for training because it automatically discovers the groups in the unlabeled dataset and clusters the data into them.

Each cluster has a corresponding centroid in this centroid-based approach. This algorithm's primary goal is to reduce the total distance between each data point and the clusters to which it belongs.

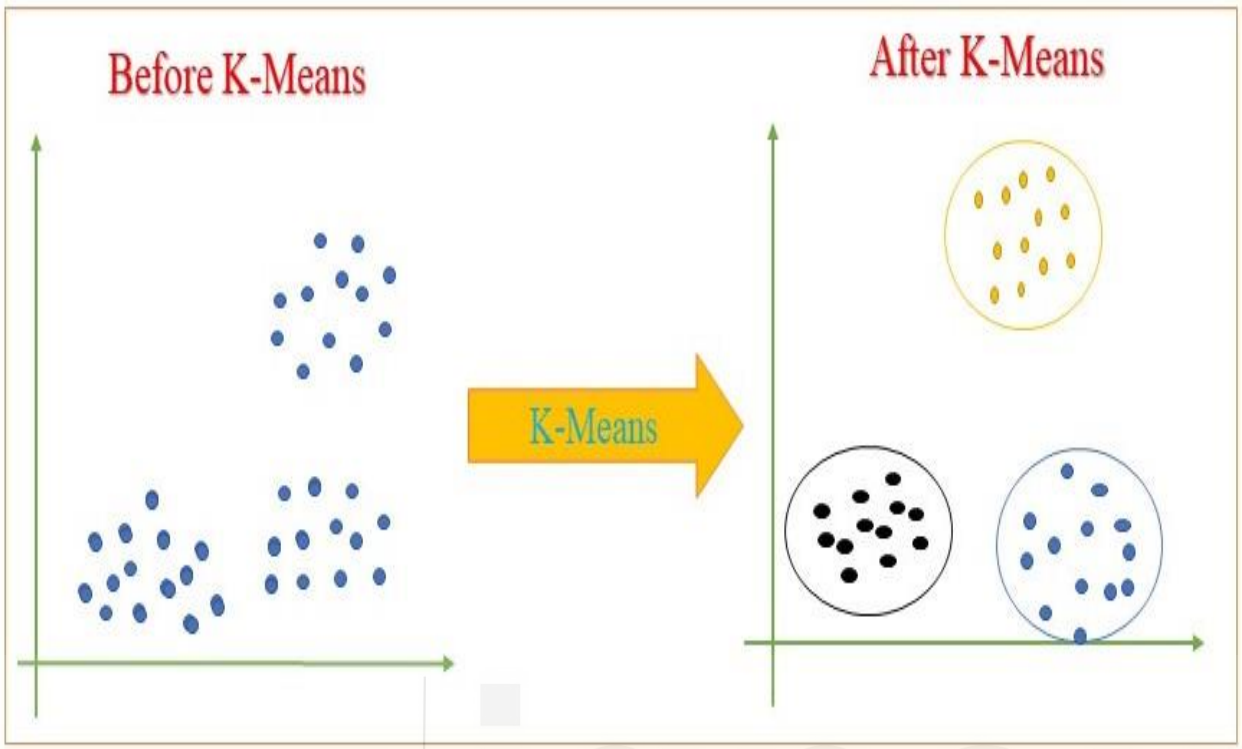
The algorithm starts with an unlabeled dataset, separates it into k clusters, and then continues the procedure until no better clusters can be found. To make this algorithm work, the value of k must be known beforehand.

k-means clustering has two primary functions.

- Iteratively arrive at the optimal value for K centre points or centroids.
- Each data point is assigned to the k-center that is the closest to it. A cluster is formed by data points that are close to a given k-center.

Consequently, data points with certain commonality are found in each cluster, and these clusters are separated from one another.

The K-means Clustering Algorithm's operation is illustrated in the diagram below:

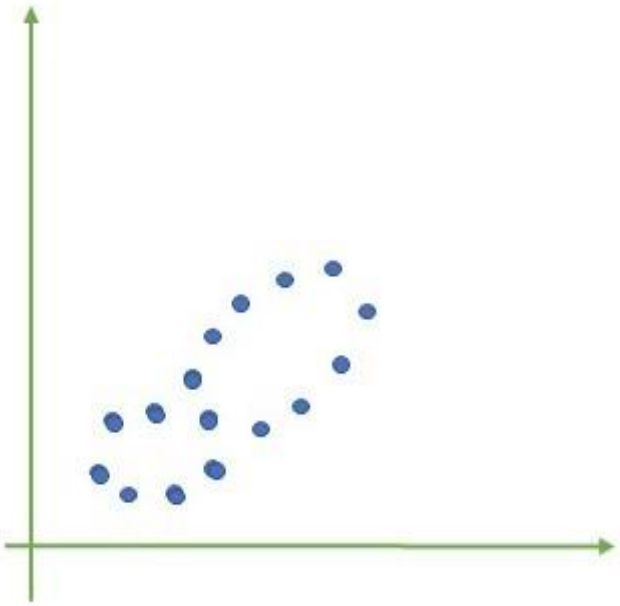


What is the K-Means Algorithm and How Does It Work?

- The number of clusters will be determined by selecting K.
- Precisely where you want to place K points or centroids is up to you. A dataset other than the one used as input can be used.
- Create the predefined K clusters by assigning each data point to its nearest centroid.
- Determine the standard deviation and re-locate the cluster centroid if necessary.
- Once again, reassign each data point to the cluster's new nearest centroid.
- FINISH is the next step if there are any reassignments. Otherwise, move to step 4.
- The prototype has been completed.

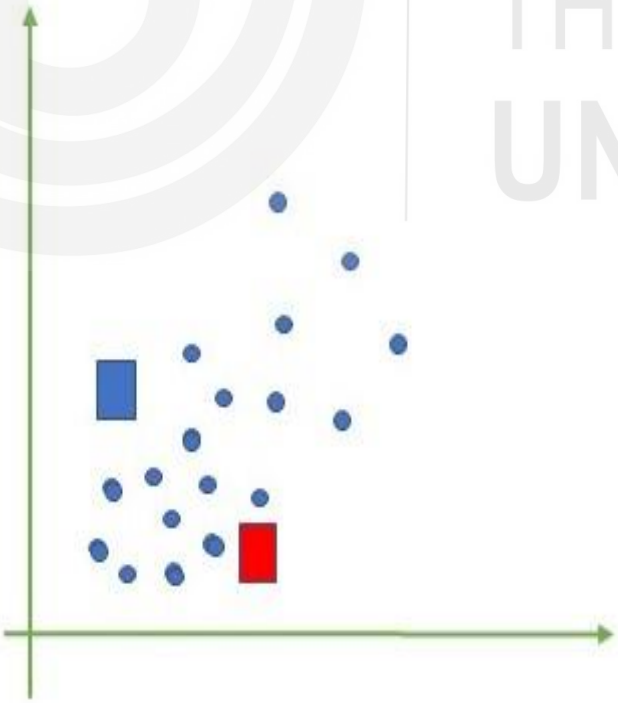
Let's take a closer look at the preceding phases by examining the accompanying diagrams:

Assume M1 and M2 are two variables. Below you'll find the x-y axis scatter plot for these two variables.



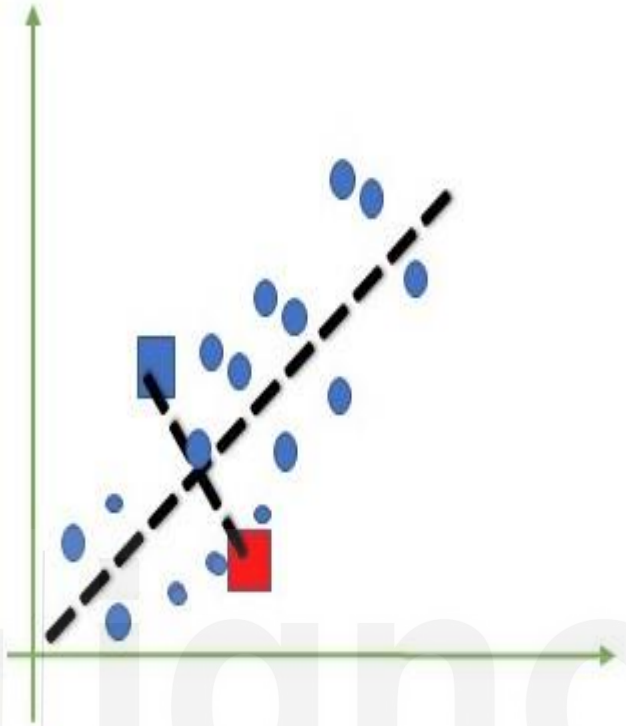
In order to identify the dataset and sort it into various groups, we'll use cluster number k, or K=2. This means that we'll make an effort to divide the datasets into two distinct groups.

To create the cluster, we'll need to pick k randomly distributed points (or centroid points). These points might be taken from the dataset or be taken from anywhere else. Here we are selecting k points from a dataset that does not include the two points below. Take a look at the image below:

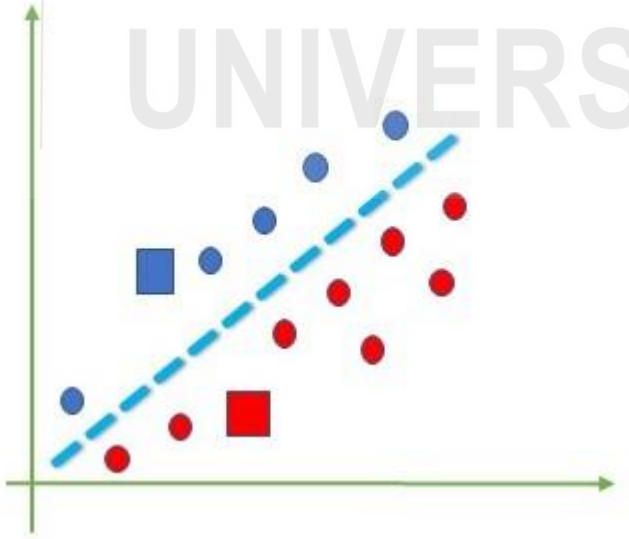


We will now give a K-point or centroid to each scatter plot data point. We'll figure it out by using some of the maths we've learned about measuring

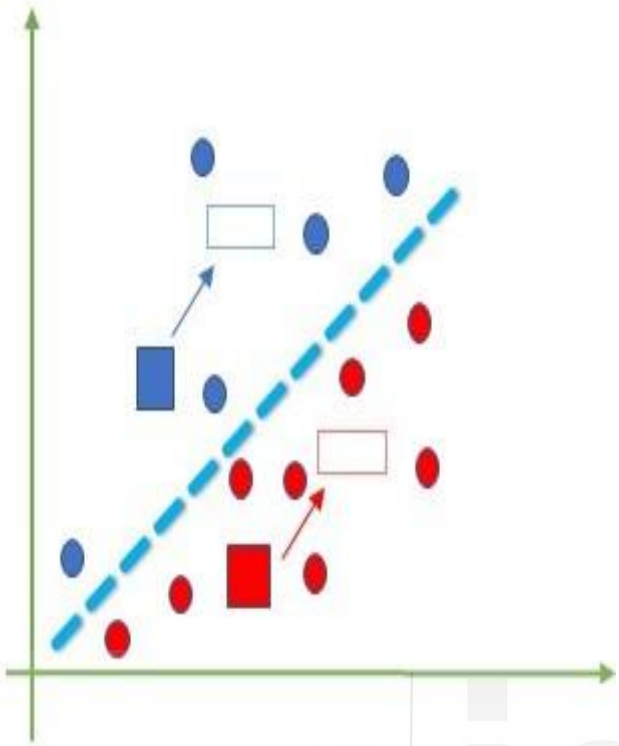
distances between points. As a result, we'll use the median of the two centroids as our point of departure. Take a look at the image below:



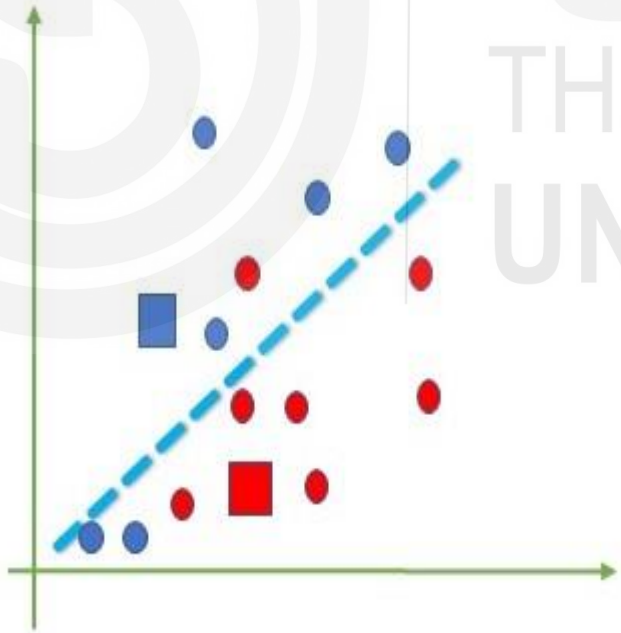
The image above shows that the points on the left side of the line are near the K1 centroid, which is blue, and the points on the right of the line are near the yellow centroid, which is yellow. To make things easier to understand, let's use blue and yellow to colour them.



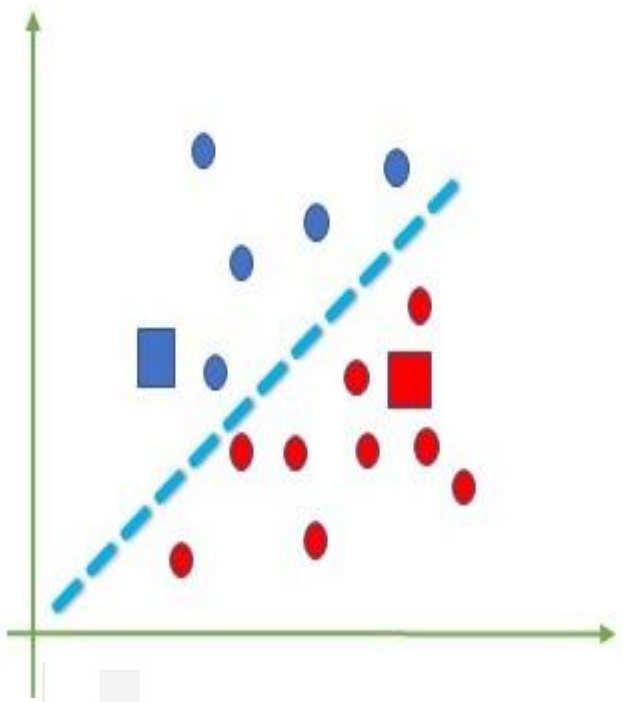
To discover the closest cluster, we'll choose a new centroid and go through the process again. New centroids will be found by computing the gravity centres of the existing centroids, and then selecting new centroids as follows:



Each data point will be assigned to the new centroid at this point. To do this, we'll use the same method we used to locate the median line. The median will look like the image below:

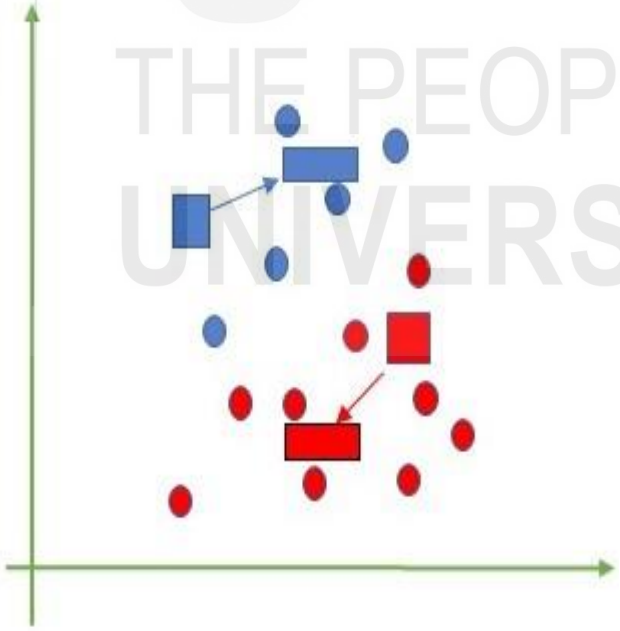


Two blue points and one yellow point may be seen above. The blue points are directly across from each other. This means that new centroids will be created using these three locations.

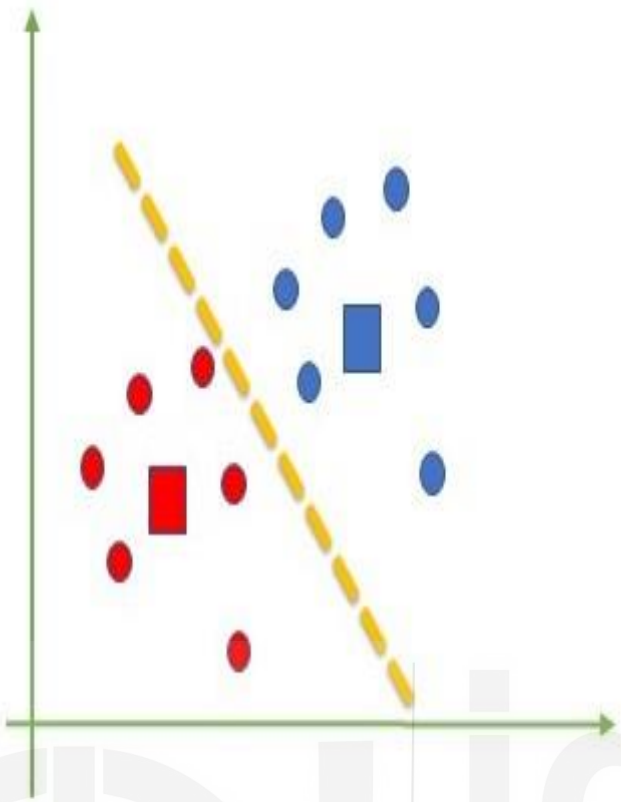


Because of the reassignment, we will return to step 4, which is the search for new centroids or K-points.

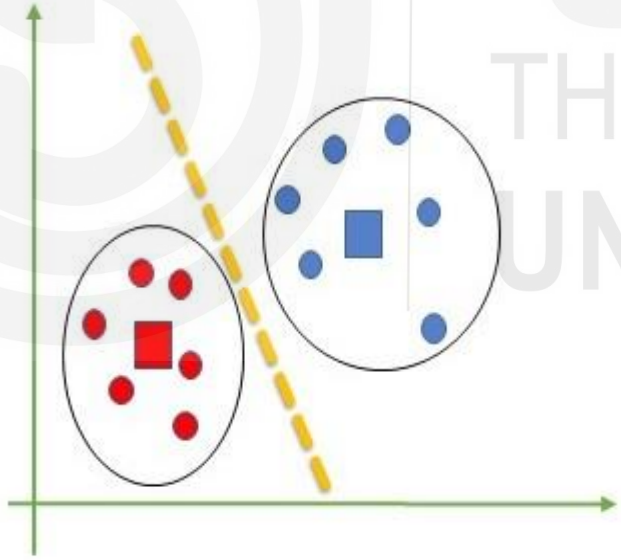
As indicated in the next image, we'll go through the process of identifying the centre of gravity for each of our centroids a second time.



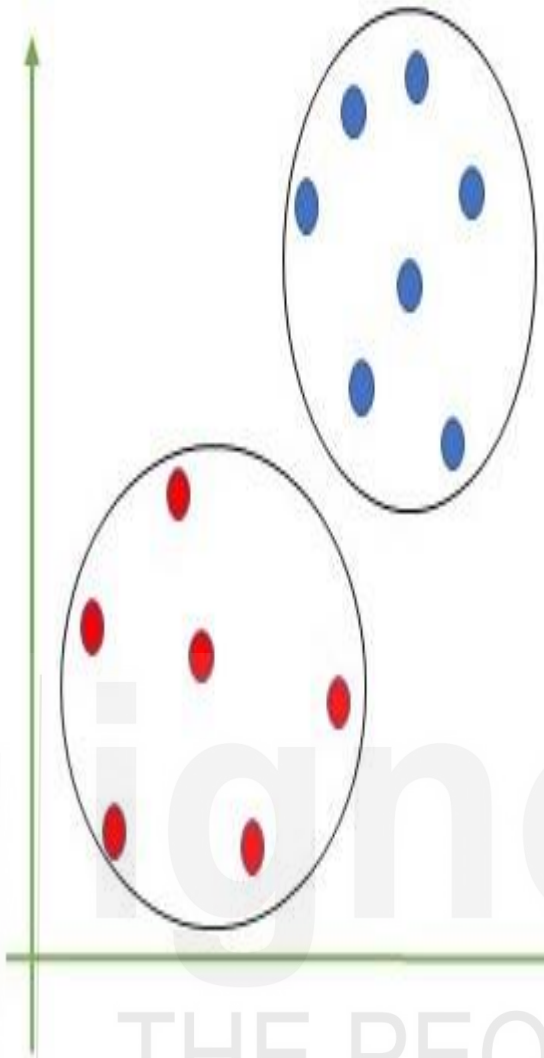
With the new centroids, we will redraw the median line and redistribute the data points. As a result, here is what the image will look like:



Since no dissimilar data points exist on either side of the line, our model has been built, as seen in the graphic above: Take a look at the image below:



Now that our model is complete, we can eliminate the centroids we previously thought existed, leaving us with the two final clusters displayed in the graphic below:



K-Medoids:

It's similar to the k-means clustering problem in that it uses k-medoids. The PAM algorithm was developed by Leonard Kaufman and Peter J. Rousseeuw, hence the name. K-means and K-medoids, two partitional algorithms, reduce the distance between cluster points and the central point of a cluster. Using this method, actual data points are selected as the cluster centers (medoids or examples), as opposed to using the k-means algorithm, which does not require the cluster centre to be an input data point. As a result, the cluster centers are easier to interpret (it is the average between the points in the cluster). As a result, unlike with k-means, the dissimilarity metrics used with k-medoids aren't restricted to Euclidean distance. Instead than minimizing the sum of squared Euclidean distances, k-medoids minimize the pairwise dissimilarities as opposed to the total.

The k-medoids approach is a classic clustering technique that separates an object data set into n clusters, with a fixed number of clusters k. (which implies that the programmer must specify k before the execution of a k-medoids algorithm). Some methods, such as the silhouette approach, can be used to assess a k value's "goodness.".

Cluster medoid objects are those that have an average dissimilarity to all other items in the cluster of one or less.

Hierarchical Method:

This method decomposes a set of data items into a hierarchy. Depending on how the hierarchical breakdown is generated, we can put hierarchical approaches into different categories. In this case, there are two methods to choose from.

- Agglomerative Approach
- Divisive Approach

Agglomerative Approach:

The bottom-up approach is another term for this strategy. At the outset, all of the teams are divided up. When all the groups have been merged, or the criterion for termination has been satisfied, the process continues merging them.

There are two methods for improving the quality of hierarchical clustering in data mining:

- At each level of hierarchical clustering, the object's links should be thoroughly examined.
- For the integration of hierarchical agglomeration, one can utilize a hierarchical agglomerative algorithm. As part of this method, micro-clusters are formed first. When data objects are micro clusters, they are subjected to macro clustering.

Divisive Approach:

With the Divisive technique, you work your way down, rather than starting at the top. All of the data items are initially maintained in a single cluster in this method. The continual iteration splits the group into smaller clusters. Iteration will continue until the condition of termination is met using the constant iteration method. After a group is divided or merged, it's impossible to go back and undo the changes.

Key Issues in Hierarchical Clustering:

Lack of a Global Objective Function:

The K-Means algorithm has a global objective function, but agglomerative hierarchical clustering approaches cluster on a local level. As a result, this technique has a distinct advantage over global functions in terms of cost and complexity.

Ability to Handle Different cluster Sizes:

To deal with merged clusters of varied sizes, we need a solution.

Merging Decisions Are Final:

One disadvantage of this method is that after two clusters have been merged, they can no longer be divided up for a better union in the future.

Density-based Method:

Density is the primary consideration in this Data Mining clustering approach. This clustering method is based on the concept of mass. This clustering technique causes the cluster to grow indefinitely. In the group's radius, each data point should have at least one number of points.

Grid-Based Method:

This Grid-Based Clustering Method makes use of a grid by grouping together the clustered objects. You can create a Grid Structure by breaking down the object space into discrete cells.

Advantage of Grid-based clustering method:

- The processing time is substantially faster with this method than with another, and so it can save time.
- The number of cells in the quantized space in each dimension determines how this method works.

Model-Based Method:

This approach hypothesizes a model for each cluster in order to locate the data that fits the hypothesis the best. This technique uses a clustering of the density function to find the clusters. It depicts the data's spatial distribution.

In addition, using this method, you may compute the number of clusters based on standard statistics, while also accounting for outliers and noise. As a result, the clustering methods it produces are solid.

Constraint-based Method:

This technique uses user- or application-oriented constraints to cluster data. When we talk about constraints, we're referring to things like user expectations or desirable clustering qualities. Constraints allow us to communicate with the

clustering process in an interactive manner. Both the user and the application have the ability to specify constraints.

Strengths and weakness of Hierarchical Clustering:

Heterogeneous clustering has the advantage of being both simple to understand and simple to implement. Among its flaws are the fact that it rarely offers the best solution, the large number of arbitrary decisions it requires, the inability to work with missing data, the inability to work well with mixed data types, the inability to handle very large data sets, and the common misinterpretation of the dendrogram are all problems. It's best to use latent class analysis as an alternative to other methods.

Outlier Analysis:

In statistics, an outlier is a data point that deviates significantly from the norm. Measuring and execution errors are two possible causes. Outlier analysis, also known as outlier mining, is the process of examining data that stands out for whatever reason. It's impossible to do data analysis without encountering outliers. An outlier is a data point that deviates significantly from the norm.

As a result, an outlier is critical since it indicates an experimentation flaw. There are numerous applications for outliers, such as identifying fraud and creating new market trends.

Outliers are frequently mistaken for random fluctuations in the data. Outliers, on the other hand, are distinct from random noise in the following way:

- An outlier is a point of observation that differs from the rest because of its location.
- To better detect outliers, noise should be eliminated.

Data mining outliers can have a number of different causes.

Outliers in Data Mining can have a variety of causes. Following are a few of these contributing factors:

- Identifying financial fraud such as credit card hacking or other similar scams makes use of this technology.
- It's utilized to keep track of a customer's changing purchase habits.
- It's used to find and report human-made mistakes in typing.
- It's utilized for troubleshooting and identifying problems with machines and systems.

In Data Mining, why should outliers be handled?

Data Mining must deal with outliers for a variety of reasons. These are a few of the explanations:

The outcomes of databases are impacted by outliers.

- Outliers frequently produce good or useful discoveries and conclusions, allowing researchers to identify various patterns or trends.
- Even in the world of study, outliers can be useful. They can be a lifesaver when doing research.
- Data mining's most important subfield is outlier analysis.

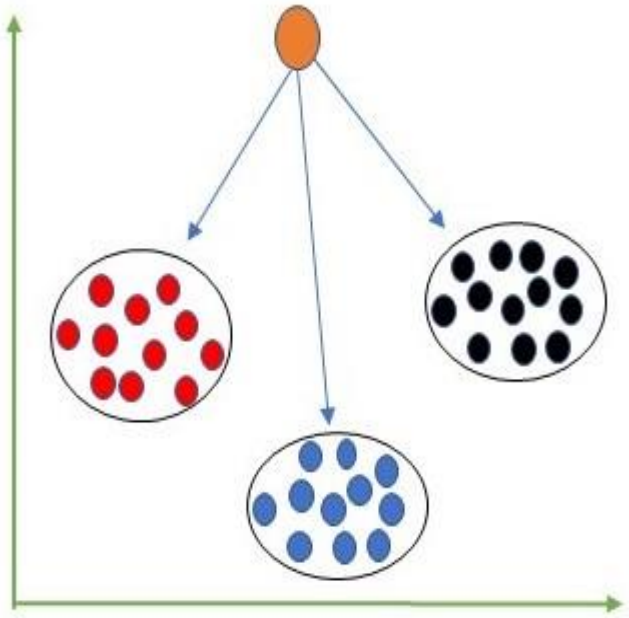
Uses for Detecting Outliers in Data Mining:

In Data Mining, it is common to utilize Outlier Detection to find anomalies. In data mining, it's utilized to find patterns or trends. The following are some examples of where outlier detection is put to use in data mining:

- Fraud Detection
- Telecom Fraud Detection
- Intrusion Detection in Cyber Security
- Medical Analysis
- Environment Monitoring such as Cyclone, Tsunami, Floods, Drought and so on
- Noticing unforeseen entries in Databases

Why Outlier Analysis:

For the most part, data mining algorithms ignore outliers like noise and exceptions. However, in certain applications like fraud detection, uncommon occurrences can be just as fascinating as the more common ones, therefore performing an outlier analysis becomes critical.



Detecting Outlier:

Each cluster has a mean value in the K-Means clustering method. Things are grouped together based on how near their arithmetic mean values are to one another. First, we must establish the threshold value so that any data point further away from its nearest cluster is considered an outlier for the purposes of this analysis. After that, we'll need to figure out how far the test data are from the mean of each cluster. The data will be labelled as an outlier if the distance between it and its nearest cluster exceeds a certain limit.

Detecting outliers using various methods:

Detecting outliers uses three main strategies. These are the strategies I've come up with:

The Statistical Approach

The Distance Based Approach

The Deviation Based Approach

Techniques for Detecting Outliers:

Detecting outliers involves using one of four methods:

- Numeric
- Z-Score
- DBSCAN
- Isolation forest

Numeric:

In a one-dimensional feature space, this is the simplest approach. An IQR (Inter Quartile Range) methodology is used to find outliers in this method. A

nonparametric property of this model is that it does not account for repeated measurements.

Z-Score:

Parametric outlier detection using Z-score is possible in a single or small feature space. To use Z-score, the data must have a normal distribution with standard deviation of one standard deviation. The following formula is used to determine Zi:

$Z_i = (x_i - \mu)/ ?$

DBSCAN:

The DBSCAN clustering technology is used in this Outlier Detection methodology. DBSCAN is a nonparametric outlier detection approach based on density in a one- or multi-dimensional feature space. All of the data points are categorised as either Core Points, Border Points, or Noise Points.

Isolation forest:

When dealing with large datasets in a one- or multi-dimensional feature space, Isolation Forest is a nonparametric method that is especially useful. Isolation numbers, which can be defined as the number of splits required to isolate a data point, are critical concepts in this strategy.

Unit 12: Text and Web Mining Page Nos.

12.0 TEXT MINING: INTRODUCTION

It is important to understand the language and its evolution. The text mining and Natural Language Processing (NLP) are connected now. NLP here stands for natural language processing and moving forward, the various application of NLP in the industry and its different components can be explored. The success of human race is because of the ability to communicate and share information now that is where the concept of language comes in however many such standards came up resulting in many such language with each language having its own set of basic shapes called alphabets and the combination of alphabets resulted in verse and the combination of these words arranged meaningfully resulted in the formation of a sentence now each language has a set of rules that is used while developing these sentences and these set of rules are also known as grammar not coming to today's world that is the 21st century according to the industry estimates only 21 percent of the available data is present in the structured format data is being generated.

12.1 OBJECTIVES

After going through this unit, you should be able to:

- understand the purpose of Text Mining.
- describe the general goals of Text Mining.
- discuss the evolution of Text Mining.
- describe various data analysis techniques performed by the Text Mining on the web.
- list, discuss an and
- describe various structures of operating system.

12.2 TEXT DATA ANALYSIS AND INFORMATION RETREIVAL

The tweets or messages on WhatsApp Facebook Instagram or through text messages and the majority of this data exists in the textual form which is highly unstructured in nature now in order to produce significant and actionable insights from the text data it is important to get acquainted with the techniques of text analysis so let's understand what is text analysis or text mining now it is the process of deriving meaningful information from natural language text and text mining usually involves the process of structuring the input text deriving patterns within the structured data and finally evaluating the interpreted output compared with the kind of data stored in database text is unstructured amorphous and difficult to deal with algorithmically nevertheless in the modern culture text is the most common vehicle for the formal exchange of information now as text mining refers to the process of arriving high-quality information from text the overall goal here is to turn the text into data for analysis.
Text mining has various areas to explore as shown in fig1

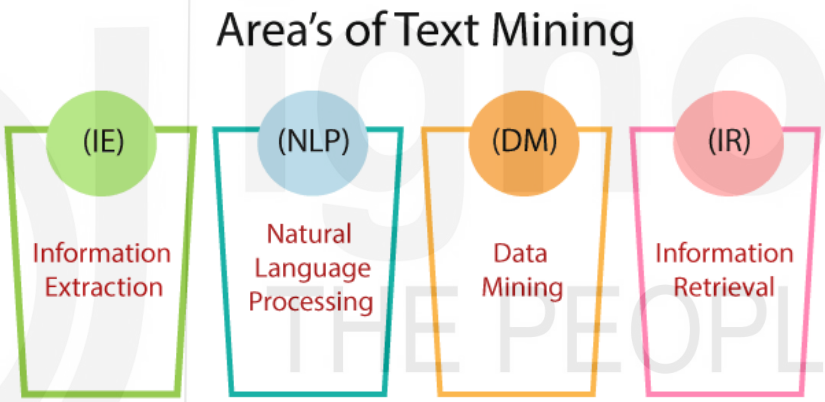


Figure1: Areas of Text Mining

The tweets or messages on WhatsApp Facebook Instagram or through text messages and most data exists in the textual form which is highly unstructured in nature now in order to produce significant and actionable insights from the text data it is important to get acquainted with the techniques of text analysis so let's understand what is text analysis or text mining now it is the process of deriving meaningful information from natural language text and text mining usually involves the process of structuring the input text deriving patterns within the structured data and finally evaluating the interpreted output compared with the kind of data stored in database text. Text mining refers to the process of arriving high-quality information from text the overall goal here is to turn the text into data for analysis.

Information Extraction is the techniques of taking out the information from the unstructured text data or semi-structured data contains in the electronic documents.
The process identifies the entities, then classify them and store in the databases from the unstructured text documents.

Natural Language Processing (NLP): The human language which can be found in WhatsApp chats, blogs, social media reviews or any reviews which are written in any

offline documents. This is done by the application of NLP or natural language processing so let's understand what is natural language processing so NLP refers to the artificial intelligence method of communicating with an intelligent system using natural language by utilizing NLP and its components one can organize the massive chunks of textual data perform numerous or automated tasks and solve a wide range of problems such as automatic summarization machine translation named entity recognition speech recognition and topic segmentation. So, let's understand the basic structure of an NLP application considering the chat pod here as an example we can see first we have the NLP layer which is connected with the knowledge base and the data storage now the knowledge base containing logs, a large history of all the chats which are used to train the particular algorithm and again we have the data storage where we have the interaction history and the analytics of that interaction which in turn helps the NLP layer to generate the meaningful output. So, now if we have a look at the various applications of NLP first, we have sentimental analysis now this is a field where NLP is used heavily, we have speech recognition now here we are also talking about the voice assistance like google assistant cortana and the siri.

Data Mining is the process of extracting significant data from standard language text. It uses various techniques of grouping the similar kind of data into clusters. It helps to predict the values or data after training the model from the previously recorded or historical data. Data mining technique is used to classify using variety of classification techniques or predictive modelling. It can predict the marketing trends and human marketing behavior, customer experience hence helping the business tycoons to build strategy for business and markets.

Text Data Analytics

First, let's define the term text mining and the term text for analytics. if you look at the two phrases literally mining emphasizes more on the process, so it gives us an algorithm a very good view of the problem analytics on the other hand emphasizes more on the result. Text mining and analytics implies to turn text data into high quality information or actionable knowledge so in both cases we have the problem of dealing with a lot of text data. The consumer is making decision to buy a product by learning from others' experience from social media reviews. The text mining supplies knowledge for optimal decision making.

We can utilize text mining in any industry, which means the data can be in any format, such as Word, PDF, XML, JSON, or anything else. Now we'll look at how text mining works. Text mining is a semi-automated procedure that necessitates the collection of data from a variety of sources. The textual data is in an unstructured format and is a free-flowing data, we cannot immediately apply data mining techniques to it, so we must first organize that thing. Once we develop a framework for the textual data, the next extremely crucial step is to structure it. After converting unstructured data into structured data, the next step is to cleanse the data. The next step is to clean the data and prepare it for analysis. Cleaning is the act of removing specific terms, such as stop words, and then removing any irrelevant words that may be included in the data. Once that is done, the data is available for analysis. Create a document matrix for a data analysis.

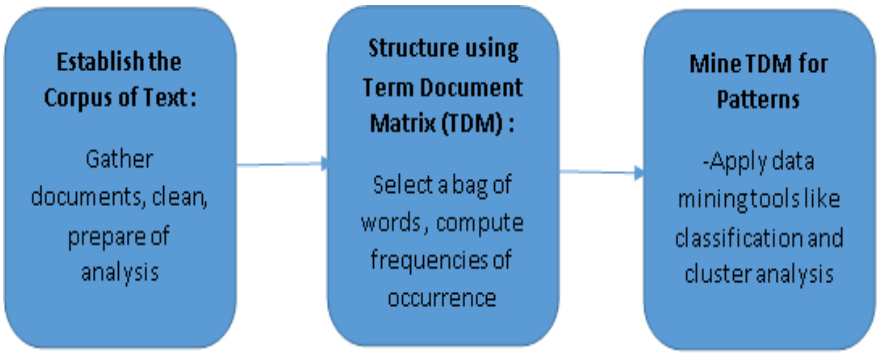


Figure 1: Text Data Analysis

Once the data matrix is developed from the input documents. The words found in those documents depending on the type of data various clustering techniques of machine learning can be applied.

Check Your Progress 1

1) Mention few applications of Text Mining?

……
……
……

2) What are the techniques used to analyze the web usage patterns?

……
……
……

12.3 DIMENSIONALITY REDUCTION OF TEXT

The conventional way of dealing with text is to collect information from various sources like text, pdf files, database files etc. After the preprocessing and cleaning of data, the analysis of data is done. The extraction of the summary of the whole content automatically is called Text Summarization. Another technique, Text Categorization existed to group the text predefined by users. The process of segmentation of text into relevant groups is called text clusters. Since, the text is quite huge and enormous to analyze so to reduce the size of the text, the dimensions of the text can be reduced. The text dimensions can be reduced by taking out the words order and their meaning. So, it can be analyzed into two ways.

a) Syntax Processing: The syntactical processing of the text can be achieved by tokenization and removal of stop words.

b) Semantic Processing: It means retaining the meaning of a text. The different techniques are applied for the same. A sentence has a main logical concept called predicate. The arguments of the predicate can be extracted from other parts of the sentence.
The identification of the predicate and the arguments for that predicate is known as semantic role labeling.

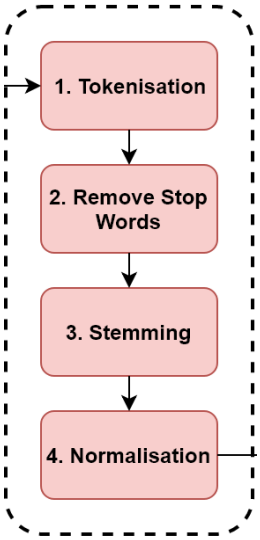


Figure 2: Text Preprocessing

Text Pre-processing: The Text mining is the process of cleaning up the data. It involves the steps of clean up. It means of removing of unnecessary text or data. The removal of ads from the web pages. It involves the conversion of cleaned text into binary formats.

- Tokenization: It converts the text into small words called tokens separated by white spaces. Transferring strings into a single textual token. Segmentation of running text into sentences and words.
- Cutting a text into pieces called *tokens*.
- Removing certain characters, such as punctuation.

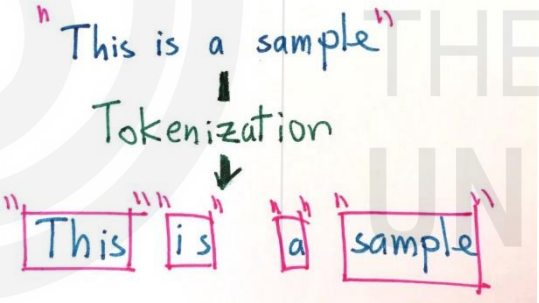


Figure 3: Tokenization

After the process of tokenization, the data are grouped into grams.

Bigrams: Bi means two so, grouping of two consecutive data together is called bigrams.
Trigrams: Tri means three, grouping of three consecutive data together is called trigrams
Similarly, it can be done,
N-grams: The groups of N consecutive data together form N-grams.

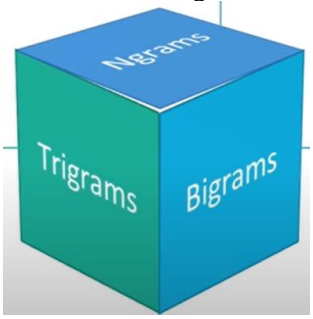


Figure 4: N-grams

POS Tagging (Part of Speech Tagging): It is the step of tagging a label or word to every token generated in the previous step. It is the identification of entity.

- According to the role of a word in a sentence, it can be tagged as a noun, verb, adjective, adverb, preposition, etc.
- Correct tags such as nouns, verbs, adjectives, etc. should be assigned.

Vector from text: Creating vectors from the text ensuring that it should not result in sparse matrix as it involves high computational costs. The important point to note here is that whatever technique we apply for dimension reduction from the text the meaning or the information of the text should not be lost. Once the text preprocessing is done then the next step is to transform the data in some structured form of a data base. The cleaned text is then transformed into generation of attributes. It can be achieved by two approaches as follows, first is

1) Bag of words:

2) Vector Space

For example, there are sample of reviews of a comedy movie so, the reviews of the viewer can be:

Review 1: This movie has a good sense of humor

Review 2: This movie is boring and long

Review 3: It is a one-time watch movie

You can easily observe three different opinions of three different viewers. You can see thousands of reviews about a movie on the internet. All these users generated text can help us out to takeout some interpretation in gauging that how a movie has performed. The above three reviews mentioned above cannot be given to the machine learning engine to analyze positive or negative reviews. So, we apply some text filtering techniques like Bag of words.

Bag of words (BOW): It is the kind of a model in which the text is written in the form of numbers. It can be represented as represent a sentence as a bag of words vector (a string of numbers).

First, the vocabulary of all the unique words in above three reviews. The vocabulary ‘This’, ‘movie’, ‘has’, ‘a’, ‘good’, ‘sense’, ‘of’, ’humor’, ‘the’, ‘is’, ‘long’, ‘boring’, ‘and’ ‘it’, ‘one-time’, ‘watch’.

	1 This	2 Movie	3 Has	4 a	5 good	6 sense	7 of	8 humor	9 is	10	11 Boring	12 and	13 it	14 One-time	15 watch	Length of the reviews (in words)
R1	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**8**
R2	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**6**
R3	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**1**	**4**

Vector of R1: [1 1 1 1 1 1 1 1 0 0 0 0 0 0 0]
Vector of R2: [0 1 0 0 0 0 0 0 1 1 1 1 0 0 0]
Vector of R3 : [0 0 0 0 0 0 0 0 1 0 0 0 1 1 1]

The basic idea of BoW model is to identify proper nouns, persons, locations, organizations, named identity.

Term Frequency – Inverse Document Frequency (TF-IDF): It is a measure to represent the count or the number of occurrences in the document for any term.
Let's say the term is represented as t, and document as d.

TF-IDF can be represented as the number of terms in the document d divided by the total number of terms in the document. So, every document will have its own TF score.

$$tf_{t,d} = \frac{n_{t,d}}{Number\ of\ terms\ in\ the\ document}$$

For example, look at the review of the movie again taken in the previous statement.

Let's calculate the TF for the review R1

Review 1: This movie has a good sense of humor

In this you will observe that the frequency of each word mentioned is as follows.
Vocabulary: 'This', 'movie', 'has', 'a', 'good', 'sense', 'of', 'humor'.
Number of words in Review 1 = 8.

TF for the word 'this' = (number of times 'this' appears in review 1) / (number of terms in review 1) = 1/8

Similarly, TF('movie') = 1/8 and so on.

Inverse Document Frequency (IDF): It is a measure to find out the occurrence of the words in number of documents. To compute the significant of the word we need to find out the TF and IDF both the measures.

$$idf_t = \log \frac{number\ of\ documents}{number\ of\ documents\ with\ term\ 't'}$$

For example, IDF for the Review 1 is:

IDF('this') = log (number of documents/number of documents containing the word 'this') = log (3/2) = 0.18

Similarly,

IDF ('movie') = log (3/3) = 0
IDF('is') = log (3/2) = 0.18

Hence, we can observe that the words with less score like 0 contains less importance than the high frequency words in occurring in many documents. We can now compute the TF-IDF score for each word in the corpus. Words with a higher score are more important, and those with a lower score are less important:

$$(tf\_idf)_{t,d} = tf_{t,d} * idf_t$$

The TF-IDF score for every word in Review 1:

TF-IDF ('this', Review 1) = TF ('this', Review 1) * IDF('this') = 1/8 * 0.18 = 0.0225
Similarly,

TF-IDF ('movie', Review 1) = 1/8 * 0 = 0

It has been observed that TF-IDF also gives larger values for less frequent words and is high when both IDF and TF values are high i.e., the word is rare in all the documents combined but frequent in a single document.

C. Feature Selection: It is also called attribute selection. It is the process of selecting a significant attribute. It is an important step in the creation of a model. The features selected helps to remove redundant features from the model.

D. Data Mining: The data mining techniques clustering and classification helps in prediction and generating the patterns.

E. Evaluate: The evaluation techniques help to evaluate the model. The model with better accuracy yields optimum results.

Check Your Progress 1

1) Calculate the TF-IDF('good' , Customer_Review 2) Review is : This product has good quality and good fit.

……
……
……

2) What is the difference between Stemming and Lemmatization in Text Mining?

……
……
……

12.4 TEXT MINING APPROACHES

The approaches of Text Mining are like a container of text documents and extraction of concepts is like a black box technique, from every document so much useful information can be taken out. The need to extract deep meaning from documents is important in terms of utilizing information to build up knowledge. The technology advancement in the field of text mining enables the task easier. There are various text mining techniques.

1) Numericizing Text
The challenges in performing text data analysis:
a) Large number of documents: There are large text documents containing huge amount of information. The objective is to find out the concept of the document, which can only be extracted from the randomly selected few documents. So, we cannot say how much would be the accuracy of the analysis as there would be large number of attributes.
b) Removal of few characters, rare words, numbers etc. The processing or indexing of the document needs preprocessing or cleaning of data. The words which appear in the document for very less time or infrequent can be removed.

c) *Removal of Stop words:* In text mining data analysis the number of words frequency is counted, then the number of times the word is appearing is counted. The documents

are classified based on the frequencies. In English language the stop words here referred to “the”, a”, “of”, “an”, “since”. These are words which are used very often to frame the sentence but provide less or unique meaning to the analysts.

- Adopting pre-defined stop words.

Sample text with Stop Words	**Without Stop Words**
TextMining – A technique of data mining for analysis of web data	TextMining, technique, datamining, analysis, web, data
The movie was awesome	Movie, awesome
The product quality is bad	Product, quality, bad

d) *Synonyms Identification:* Similarly, for the text mining analysis the synonyms of the words are identified. These words are then combined for indexing.
e) *Stemming Algorithms:* The stemming algorithms is another preprocessing step before indexing of documents in text mining enables to reduction of irrelevant words from the main word in the text.

- Slicing the end or the beginning of words.
- Intent of removing affixes.

Lemmatization

- Reduction of a word to its base form.
- Grouping together different forms of the same word.

f) *Support for different languages* is quite significant in text mining process.

Check Your Progress 2

1) What is the difference between Stemming and Lemmatization in Text Mining?

……
……
……

12.4 WEB MINING

Web mining as the name suggests that it involves the mining of web data. The extraction of information from websites uses data mining techniques. It is an application based on data mining techniques. The parameters generally to be mined in web pages are hyperlinks, text or content of web pages, linked user activity between web pages of the same website or among different websites. All user activities are stored in a web server log file.

Web Mining can be referred as discovering interesting and useful information from Web content and usage

12.4.1 Web Mining performs various tasks such as:

1) Generating patterns existing in some websites, like customer buying behavior or navigation of web sites.
2) The web mining helps to retrieve faster results of the queries or the search text posted on the search engines like Google, Yahoo etc.
3) The ability to classify web documents according to the search performed on the ecommerce websites helps to increase businesses and transactions.

12.4.2 Web Mining Features

- Web search, e.g. Google, Yahoo, MSN, Ask, Froogle (comparison shopping), job ads (Flipdog)
- The web mining is not like relation, it has text content and linkage structure.
- On the www the user generated data is increasing rapidly. So, Googles' usage logs are very huge in size. Data generated per day on google can be compared with the largest data warehouse unit.
- Web mining can react in real-time with dynamic patterns generated on the web. In this no direct human interaction is involved.
- Web Server: It maintains the entry of web log pages in the log file. This web log entries helps to identify the loyal or potential customers from ecommerce website or companies.
- Web page is considered as a graph like structure, where pages are considered as nodes, hyperlinks as edges.
  Pages = nodes, hyperlinks = edges
  - Ignore content
  - Directed graph
- High linkage
  - 8-10 links/page on average
  - Power-law degree distribution

Applications of Web Mining in eCommerce can be observed as

- Recommendations: e.g. Netflix, Amazon
- improving conversion rate: next best product to offer
- Advertising, e.g. Google AdSense
- Fraud detection: click fraud detection,
- Improving Web site design and performance

There are three types of web mining as shown in the following figure 5.

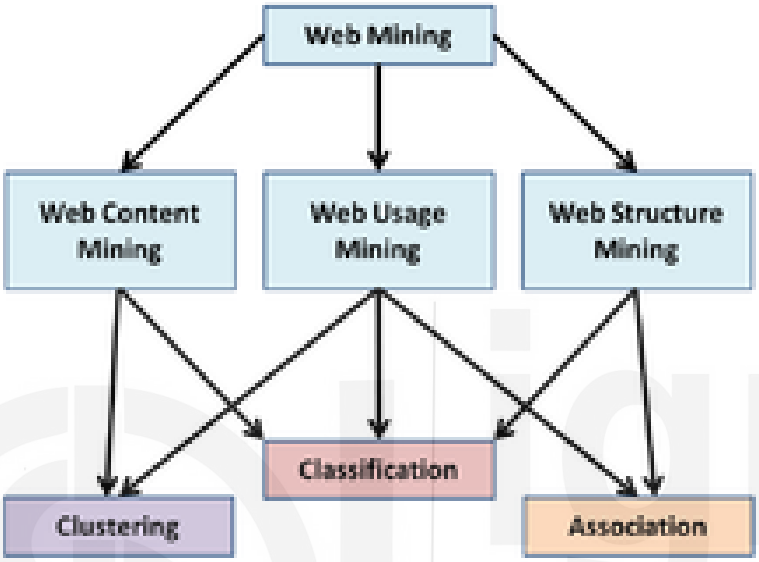


Figure 5: Three types of Web Mining (source: internet)

12.5 WEB CONTENT MINING

Web content mining is used for taking out information, data from web pages. It extracts the information and knowledge from the content of web pages. In web content mining each page is considered as an individual document. The Hypertext markup language (HTML) provides the logical layout and structure of the webpage from where the lot of useful information can be extracted. This extracted data provides easy level aggregation of web pages.

- The usefulness of web content mining is that it helps in distinguish topics or heading content of the web pages. You might have observed whenever you write a search text on the google search engine, you get a list of suggestions. It is internally performed by web content mining.
- Web content mining scans or reads the images or text and group of web pages. The input queries posted by the users and display the result of search engines. For example, if a user wants to search some information about some hotel in a particular region, the list will be displayed.

12.5 WEB STRUCTURE MINING

Web structure mining helps the web pages to extract knowledge. This task is not easy as the challenge in accessing the unstructured web data link is huge requires lot of computational costs. Since, the web data is unstructured automatic generation of knowledge or pattern is difficult.
It is used to find out the link structural parameters of hyperlink. It helps to find out the network link or if the hyperlink is directly associated with the data. The social network analysis in web structure mining treats web as a directed graph, in which the webpages are as vertices associated with the hyperlinks.

12.6 MINING MULTIMEDIA DATA ON THE WEB

The websites are flooded with the multimedia data like, video, audio, images, and graphs. This multimedia data has different characteristics. The videos, images, audio, and pictures have different methods of archiving and retrieving the information. The multimedia data on the web has different properties this is the reason the typical multimedia data mining techniques cannot be applied. This web-based multimedia has texts and links. The text and links are the important features of the multimedia data to organize web pages. The better organization of web pages help in effective search operation. The web page layout mining can be applied to segregate the web pages into the set of multimedia semantic blocks from non-multimedia web pages. There are few web-based mining terminologies and algorithms to understand.

PageRank: This measure is used to count the number of pages the webpage is connected to other websites. It gives the importance of the webpage. The Google search engine uses the algorithm PageRank and rank the web page very significant if is frequently connected with the other webpages on the social network. It works on the concept of probability distribution representing the likelihood that a person on random click would reach to any page. It is assumed the equal distribution in the beginning of the computational process. This measure works on iterations. Iterating or repetition of page ranking process would help rank the web page closely reflecting to its true value.

HITS: This measure is used to rate the webpage. It is developed by Jon Kleinberg. It uses hubs and authorities to be determined from a web page. Hubs and Authorities defines a recursive relationship between web pages.

- This algorithm helps in web link structure and speeds up the search operation of a web page. Given a query to a Search Engine, the set of highly relevant web pages are called Roots. They are potential Authorities.
- Pages that are not very relevant but point to pages in the Root are called Hubs. Thus, an Authority is a page that many hubs link to whereas a Hub is a page that links to many authorities.

Page Layout analysis: It extracts and maintains the page-to-block, block-to-page relationships from link structure of web pages.

Vision page segmentation (VIPS) algorithm: It first extracts all the suitable blocks from the HTML Document Object Model (DOM) tree, and then it finds the separators between these blocks. Here separators denote the horizontal or vertical lines in a Web page that visually cross with no blocks. Based on these separators, the semantic tree of the Web page is constructed. A Web page can be represented as a set of blocks (leaf nodes of the semantic tree). Compared with DOM-based methods, the segments obtained by VIPS are more semantically aggregated. Noisy information, such as navigation, advertisement, and decoration can be easily removed because these elements are often placed in certain positions on a page. Contents with different topics are distinguished as separate blocks.

You can understand simply by considering following points,

- The web page contains links and links contained in different semantic blocks point to pages of different topics.
- Calculate the significance of web page using algorithms PageRank or HITS. Split pages into semantic blocks

Apply link analysis on semantic block level. For example, in the below figure 1, it is clearly shown. From Figure 1, we can see the links in different blocks point to the pages with different topics. In this example, one link points to a page about entertainment and another link points to a page about sports.

Figure 6: example of a sample web page (new.yahoo.com), showing web page with different semantic blocks (red, green, and brown rectangular boxes. Every block has different importance in the web page. The links in different blocks points to the pages with different topics.

To analyze the web page containing multimedia data there is a technique known as Link analysis. It uses two most significant algorithms PageRank and HITS to analyze the significance of web pages. This technique uses each page as a single node in the web graph. But since, web page with multimedia has lot of data and links. So, cannot be considered as a single node in the graph. So, in this case the web page is partitioned into blocks using vision page segmentation also called VIPS algorithm. So, now after

extracting all the required information the semantic graph can be developed over world wide web in which each node represents a semantic topic or semantic structure of the web page.

VIPS algorithm helps in determining the text for web pages. This is the closely related text that provides content or text description of web pages and used to build image index. The web image search can then be performed using any traditional search technique. Google, Yahoo still uses this approach to search web image page.

Block-level Link analysis: The block-to-block model is quite useful for web image retrieval and web page categorization. It uses kinds of relationships, i.e., block-to-page and page-to-block. Let's see some definitions. Let P denote the set of all the web pages,

P = {p1, p2,.., pk}, where k is the number of web pages.

Let B denote the set of all the blocks,

B = {b1, b2, …, bn}, where n is the number of blocks.

It is important to note that, for each block there is only one page that contains that block. $b_i \in p_j$ means the block i is contained in the page j.

Block-Based Link Structure Analysis: This can be explained using matrix notations. Consider Z is the block-to-page matrix with dimension n × k. Z can be formally defined as follows:

$$Z_{ij} = \begin{cases} 1/s_i & \text{if there is a link from block } i \text{ to page } j \\ 0 & \text{otherwise} \end{cases}$$

where s_i is the number of pages that block i links to. Zij can also be viewed as a probability of jumping from block i to page j.

The block-to-page relationship gives a more accurate and robust representation of the link structures of the web unlikely, HITS as at times it deviates from the web text information. It is used to organize the web image pages. The image graph deduced can be used to achieve high-quality web image clustering results. The web page graph for web image can be constructed by considering measuring which tells the relationship between blocks and images, block-to-image, and image-to-block. Likewise, page-to-block and block-to-pages.

12.7 AUTOMATIC CLASSIFICATION OF WEB DOCUMENTS

The categorization of web pages into the respective subjects or domains is called classification of web documents. For example, in the following figure, it has shown various categories like, books, electronics etc. let's say you are doing online shopping on the amazon website and there are so many webpages so when you search for electronics the respective web page containing the information of electronics is displayed. This is the classification of products which is done on the textual and image contents.

Figure 7: Types of web documents containing different types of data.

The problem with the classification of web documents is that every time the model is to be constructed by applying some algorithms to classify the document is mammoth task. The large number of unorganized web pages may have redundant documents.

The automated document classification of web pages is based on the textual content. The model requires initial training phase of document classifiers for each category based on training examples.
In the figure below it is shown that the documents can be collected from different sources. After the collection of documents data cleansing is performed using extraction transformation and loading techniques. The documents can be grouped according to the similarity measure (grouping of the documents according to the similarity between the documents) and TF-IDF. The machine learning model is created and executed, and different clusters are generated.

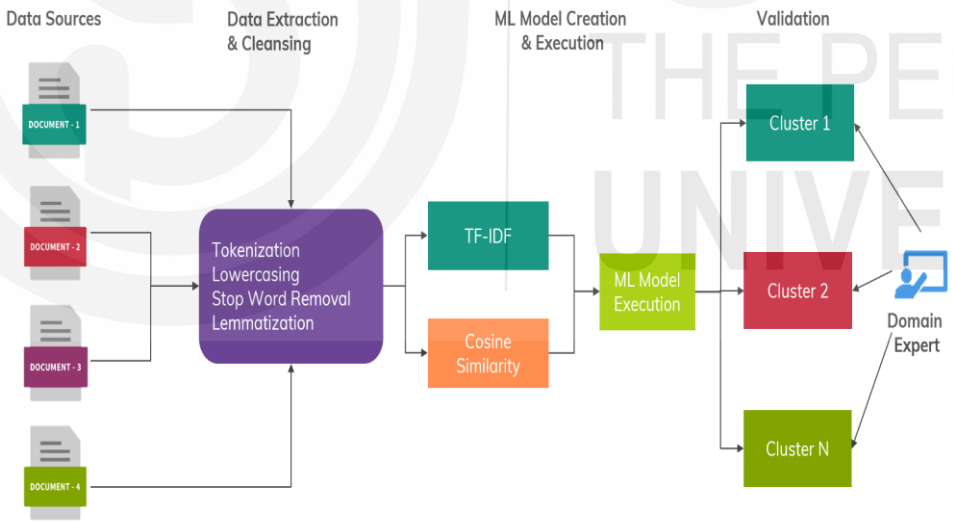


Figure 8: Automatic Classification of Web documents

Automated document classification identifies the documents and groups the relevant documents without any external efforts. There are various tools available in the market like RapidMiner, Azure, Machine Learning Studio, Amazon Sage maker, KNIME and Python. The trained model automatically reads the data from documents (PDF, DOC, PPT) and classifies the data according to the category of the document. This trained model is already trained with the Machine Learning and Natural Language Processing techniques. There are domain experts who perform this task efficiently.

Benefits of Automatic Document Classification System

1) It is more efficient system of classification as produces improved accuracy of results and speed up the process of classification.
2) The system incurs in less operational costs
3) Easy data store and retrieval.
4) It organizes the files and documents in a better streamlined way.

12.8 WEB USAGE MINING

Web Usage Mining: This term web usage mining is used to mine the data of web log records. It helps to find out the navigation and access patterns of web pages. This mining helps to study the user access pattern of web pages. It uses webserver log files to mine or track the behavior of the user. It helps to know the interests or the perception of the user, while the web structure mining and content mining focusses more on the intention of the user. After analyzing and exploring the regularities in web blogs, can identify potential customers for ecommerce. It enhances the quality and delivery of internet connection and the improves web server system performance.

The web log file contains all the information and data pertaining to web sites and user activities. The URL has a lot of information related with IP address of the request generated and at what time the request initiated. Like Amazon which is an ecommerce website where millions of transactions take place. So, huge number of web log entries are recorded gigabytes every day. So, it is quite important to have efficient web mining techniques to generate relevant information in less time.

The web mining technique involves taking out knowledge from the raw data stored in web log files. So, these raw data needs preprocessing cleaning, transformation and loading of data to retrieve and analyze important information. The web log data is ready after preprocessing of data.

The second source of information is URL (Uniform Resource Locator), it contains IP address, time, and web page content. Using the information of web log, a multidimensional cube is constructed. so that the knowledge can be extracted from various dimensions. The analysis of OLAP cube helped to identify top number of users and top users accessing the web page. Now the data mining techniques can be applied to find out the necessary information about the user accessing behavior of web pages and using association rule mining a pattern can be generated. The web log data can be used majorly for tracking and analyzing the user behavior on the social network.

The weblog files helped used to analyze system performance, web caching, identifying web page, understanding the nature of web traffic, understanding the nature of web traffic, purpose of that network behavior. For example, some studies have proposed adaptive sites: websites that improve themselves by learning from user access patterns. Weblog analysis may also help build customized Web services for individual users.

Check Your Progress 3

1) What web log file contains?

……
……

2) Show the diagram for Image mining process with brief explanation?

…….

……

1.12 SOLUTIONS/ANSWERS

Check Your Progress 1

Answer 1: Challenges in Web Mining:

- The web page link structure is quite complex to analyze as the web page is linked with many more other web pages. There exists lot of documents in the digital library of the web. The data in this library is not organized.
- The web data is uploaded on the web pages dynamically on regular basis.
- Diversity of client networks having different interests, backgrounds, and usage purposes. The network is growing rapidly.
- Another challenge is to extract the relevant data for subject or domain or user.

Answer 2: Techniques used to analyze the web usage patterns are as follows:

1) Session and web page visitor analysis: The web log file contains the record of users visiting web pages, frequency of visit, days, and the duration for how long the user stays on the web page.
2) OLAP (Online Analytical Processing): OLAP can be performed on different parts of log related data in a certain interval of time.
3) Web Structure Mining: It produces the structural summary of the web pages. It identifies the web page and indirect or direct link of that page with others. It helps the companies to identify the commercial link of business websites.

Answer 3: Application of Web Mining:

- Digital Marketing
- Data analysis on website and application accomplishment.
- User behavior analysis
- Advertising and campaign accomplishment analysis.

Check Your Progress 2

Answer 1: The main difference between BLHITS (Block HITS) and HITS

BLHITS	HITS
Links are from blocks to pages	Links from pages to pages
Root is top ranked blocks	Root is top ranked pages
Analyses only top ranked block links	Analyses all the links of all the pages
Content analysis at block level	Content analysis at page level

Check Your Progress 3

Answer 1: Weblog data can be used with Web content and Web linkage structure mining to aid Web page ranking, Web document classification, and the creation of a multilayered Web information base, because it provides information about what kind of users will visit what types of Web pages. The mining of a user's interactions with the web is a particularly intriguing use of Web usage mining.

Answer 2:

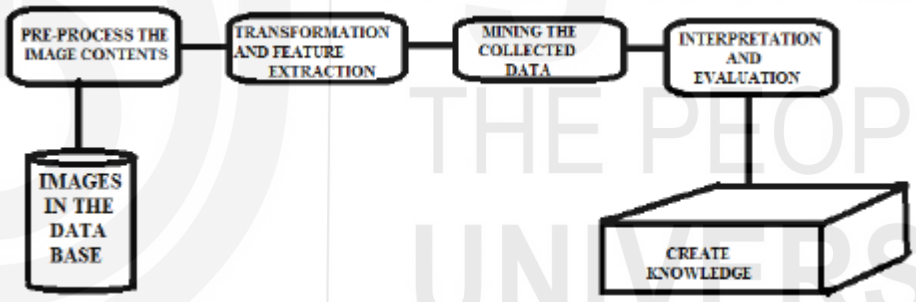


Figure : Image Web Mining