**DWM QB**

**Applications of Web mining**

* Web mining helps to improve the power of web search engines
* used to predict user behavior.
* useful for a particular Website and e-service e.g., landing page optimization.
* useful to e-commerce websites and e-services

**What is web structure mining**

* Web structure mining is a tool that can recognize the relationship between web pages linked by data or direct link connection.
* This structured data is discoverable by the provision of web structure schema through database techniques for Web pages.
* The model is based on the topology of the hyperlinks.
* This can help in discovering similarity between sites or in discovering authority sites for a particular topic or discipline or in discovering overview or survey sites that point to many authority sites.

**Why is the naive Bayesian classification called naive?Explain Naive Bayesian classification algorithm.**

* The Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
* It is mainly used in text classification that includes a high-dimensional training dataset.
* The Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
* It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
* Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Why is it called Naïve Bayes?

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of other features. Such as if the fruit is identified on the basis of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.

**Explain Clustering**

* Clustering or cluster analysis is a technique, which groups the unlabelled dataset.
* It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."
* It does it by finding some similar patterns in the dataset such as shape, size, color, behavior, etc., and divides them as per those similar patterns.
* It is an [unsupervised learning](https://www.javatpoint.com/unsupervised-machine-learning) method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.
* It can be used to simplify the processing of large and complex datasets.

**Explain K means clustering with example-**

* K-Means Clustering is an [Unsupervised Learning algorithm](https://www.javatpoint.com/unsupervised-machine-learning), which groups the unlabeled dataset into different clusters.
* Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
* The algorithm divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters.
* It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

The working of the K-Means algorithm is explained in the below steps:

**Step-1:** Select the number K to decide the number of clusters.

**Step-2:** Select random K points or centroids. (It can be other from the input dataset).

**Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.

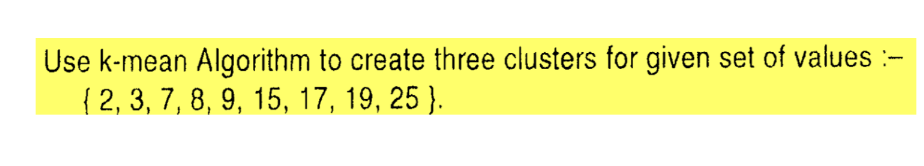
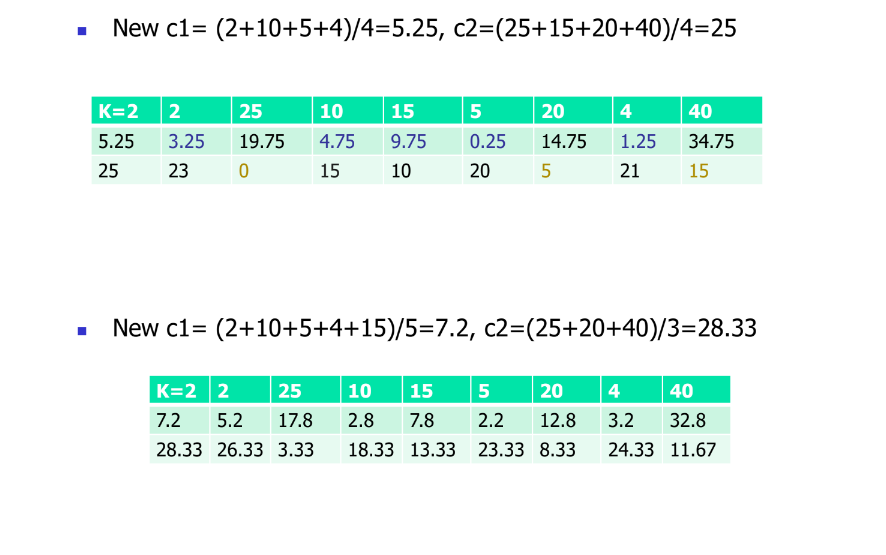
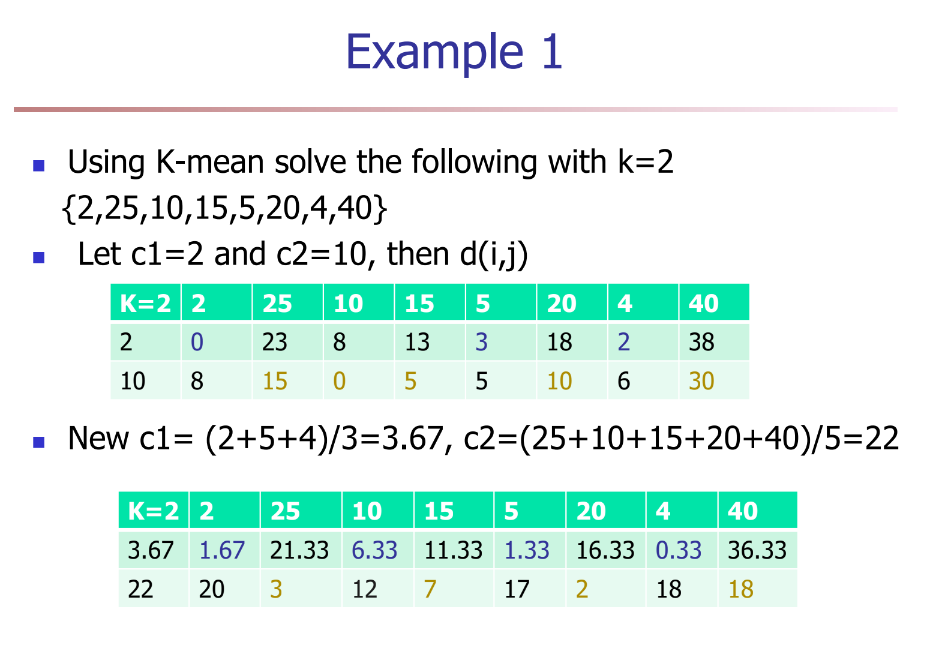
**Step-4:** Calculate the variance and place a new centroid of each cluster.

**Step-5:** Repeat the third steps, which means re-assign each datapoint to the new closest centroid of each cluster.

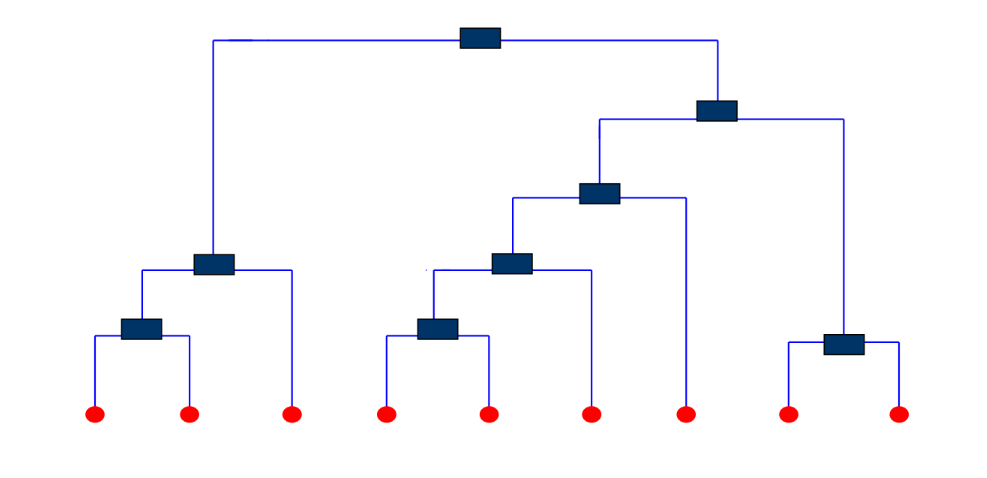
**Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.

**Step-7**: The model is ready.

Example-



**What are Dendrograms**



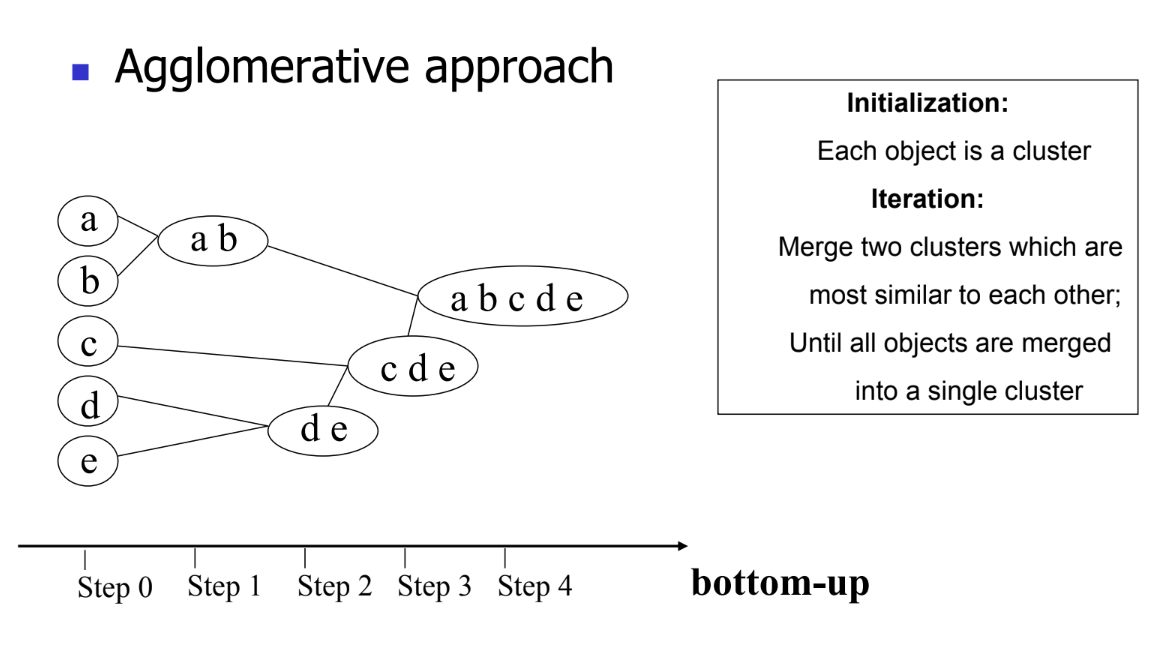
* Dendrogram is a binary tree that shows how clusters are merged/split hierarchically.
* It is most commonly created as an output from [hierarchical clustering](https://www.displayr.com/what-is-hierarchical-clustering/).
* The main use of a dendrogram is to work out the best way to allocate objects to clusters.

**Explain Hierarchical Clustering techniques**

A Hierarchical Clustering method works by grouping data objects into a tree of clusters. It refers to an unsupervised learning procedure that determines successive clusters based on previously defined clusters.It suffers from inability to perform adjustment once a merge or split decision has been executed.

There are 2 types of hierarchical clustering-

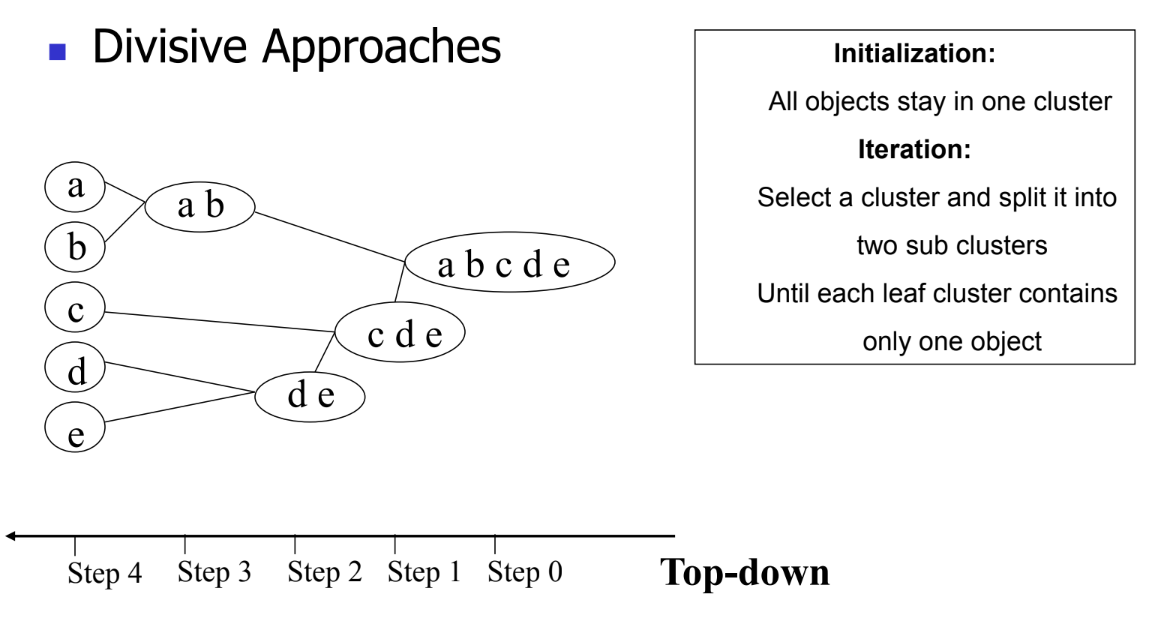
1. Agglomerative (bottom-up/ merging)



It is used to group similar objects in clusters. In agglomerative clustering, each data point acts as an individual cluster and at each step data objects are grouped in a bottom-up method. At each iteration, the clusters are combined with different clusters until one cluster is formed.

1. Determine the similarity between individuals and all other clusters.
2. Consider each data point as an individual cluster.
3. Combine similar clusters.
4. Recalculate the proximity matrix for each cluster.
5. Repeat step 3 and step 4 until you get a single cluster.

1. Divisive( top- down/ splitting)



Divisive hierarchical clustering is exactly the opposite Agglomerative Hierarchical clustering. In Divisive Hierarchical clustering, all the data points are considered an individual cluster, and in every iteration, the data points that are not similar are separated from the cluster. The separated data points are treated as an individual cluster. Finally, we are left with N clusters.

**Explain BIRCH , DBSCAN**

BIRCH -

* BIRCH (balanced iterative reducing and clustering using hierarchies) is an unsupervised data mining algorithm that performs hierarchical clustering over large data sets.
* With modifications, it can also be used to accelerate k-means clustering and Gaussian mixture modeling with the expectation-maximization algorithm.
* An advantage of BIRCH is its ability to incrementally and dynamically cluster incoming, multi-dimensional metric data points to produce the best quality clustering for a given set of resources (memory and time constraints).
* In most cases, BIRCH only requires a single scan of the database.

DBSCAN-

* Density-Based Clustering refers to one of the most popular unsupervised learning methodologies used in model building and machine learning algorithms.
* The data points in the region separated by two clusters of low point density are considered as noise.
* The surroundings with a radius ε of a given object are known as the ε neighborhood of the object

**Explain FP tree**

* The frequent-pattern tree (FP-tree) is a compact data structure that stores information about frequent patterns in a database. Each transaction is read and then mapped onto a path in the FP-tree.
* This is done until all transactions have been read. Different transactions with common subsets allow the tree to remain compact because their paths overlap.
* A frequent Pattern Tree is made with the initial item sets of the database. The purpose of the FP tree is to mine the most frequent pattern. Each node of the FP tree represents an item of the item set.
* The root node represents null, while the lower nodes represent the item sets. The associations of the nodes with the lower nodes, that is, the item sets with the other item sets, are maintained while forming the tree

**Explain Page rank algorithm.**

* PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results.
* PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is.
* The underlying assumption is that more important websites are likely to receive more links from other websites.
* PageRank is a way of measuring the importance of website pages.
* The PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page.

**What are web crawlers?**

* A Web crawler is an Internet bot that systematically browses the World Wide Web and that is typically operated by search engines for the purpose of Web indexing.
* A Web crawler starts with a list of URLs to visit.
* Web search engines and some other websites use Web crawling or spidering software to update their web content or indices of other sites web content.
* Web crawlers copy pages for processing by a search engine, which indexes the downloaded pages so that users can search more efficiently.

**Explain Web content mining.**

* Web content mining is referred to as text mining. Content mining is the browsing and mining of text, images, and graphs of a Web page to decide the relevance of the content to the search query.
* This browsing is done after the clustering of web pages through structure mining and supports the results depending upon the method of relevance to the suggested query.
* With a large amount of data that is available on the World Wide Web, content mining supports the results lists to search engines in order of largest applicability to the keywords in the query.
* It can be defined as the phase of extracting essential data from standard language text. Some data that it can generate via text messages, files, emails, documents are written in common language text. Text mining can draw beneficial insights or patterns from such data.

**What are outliers**

* Outlier is a data object that deviates significantly from the rest of the data objects and behaves in a different  manner.
* They can be caused by measurement or execution errors.
* The analysis of outlier data is referred to as outlier mining.
* An outlier cannot be termed as a noise or error. Instead, they are suspected of not being generated by the same method as the rest of the data objects.

**Explain Market Basket Analysis**

* Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns.
* It involves analyzing large data sets, such as purchase history, to reveal product groupings and products that are likely to be purchased together.
* Basically, Market basket analysis in data mining is to analyze the combination of products which have been bought together.
* Implementation of market basket analysis requires a background in statistics and data science and some algorithmic computer programming skills. For those without the needed technical skills, commercial, off-the-shelf tools exist.

**What are different ways to improve the efficiency of apriori algorithm**

Following are the different ways to improve efficiency of apriori algorithm-

1] Hash-based itemset counting

In hash-based itemset counting, you need to exclude the k-itemset whose equivalent hashing bucket count is less than the threshold is an infrequent itemset.

2] Transaction Reduction

In transaction reduction, a transaction not involving any frequent itemset becomes not valuable in subsequent scans.

3] Partitioning −

A partitioning technique can be used that needs two database scans to mine the frequent itemsets. It includes two phases involving In which the algorithm subdivides the transactions of D into n non-overlapping partitions.

4] Sampling −

The fundamental idea of the sampling approach is to select a random sample S of the given data D, and then search for frequent itemsets in S rather than D. In this method, it can trade off some degree of accuracy against efficiency.

**Explain Apriori algorithm with example**

* Apriori algorithm refers to the algorithm which is used to calculate the association rules between objects.
* It means how two or more objects are related to one another.
* In other words, we can say that the apriori algorithm is an association rule learning that analyzes that people who bought product A also bought product B.
* Apriori algorithm is also called frequent pattern mining.
* Generally, you operate the Apriori algorithm on a database that consists of a huge number of transactions.

Example-

**What are different attribute selection methods in the Decision tree induction algorithm?**

Attribute selection method is a procedure to determine the splitting criterion that “best” partitions the data tuples into individual classes.

This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.

Attribute Selection Measures-

Information gain − Information gain is used for deciding the best attributes that render maximum data about a class. It follows the method of entropy while aiming at reducing the level of entropy, starting from the root node to the leaf nodes.

Gain ratio − The information gain measure is biased approaching tests with several results. It can select attributes having a high number of values. For instance, consider an attribute that facilitates a unique identifier, including product ID.

Gini index − The Gini index can be used in CART. The Gini index calculates the impurity a data partition or collection of training tuples

**Explain Frequent pattern Growth mining frequent item sets with an example.**

**Explain decision tree based classification algorithm with example**