Dhanish Ahamed

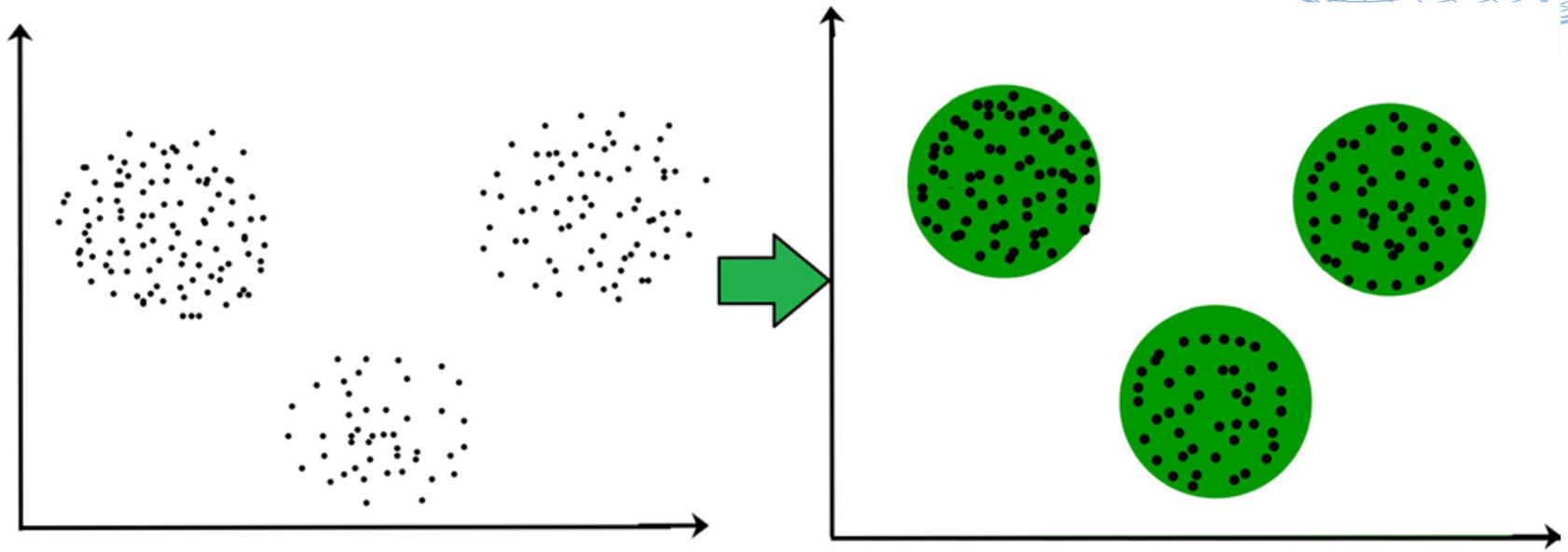
CLUSTERING

Machine learning Seminar

INDRODUCTION TO CLUSTERING

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

EXAMPLE



- The data points in the graph clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the picture.

WHY CLUSTERING

- Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present.
- There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need.

CLUSTERING METHODS

- Density-Based Methods
- Hierarchial Based Methods
- Partitioning Methods
- Grid-based Methods

K-MEANS CLUSTERING

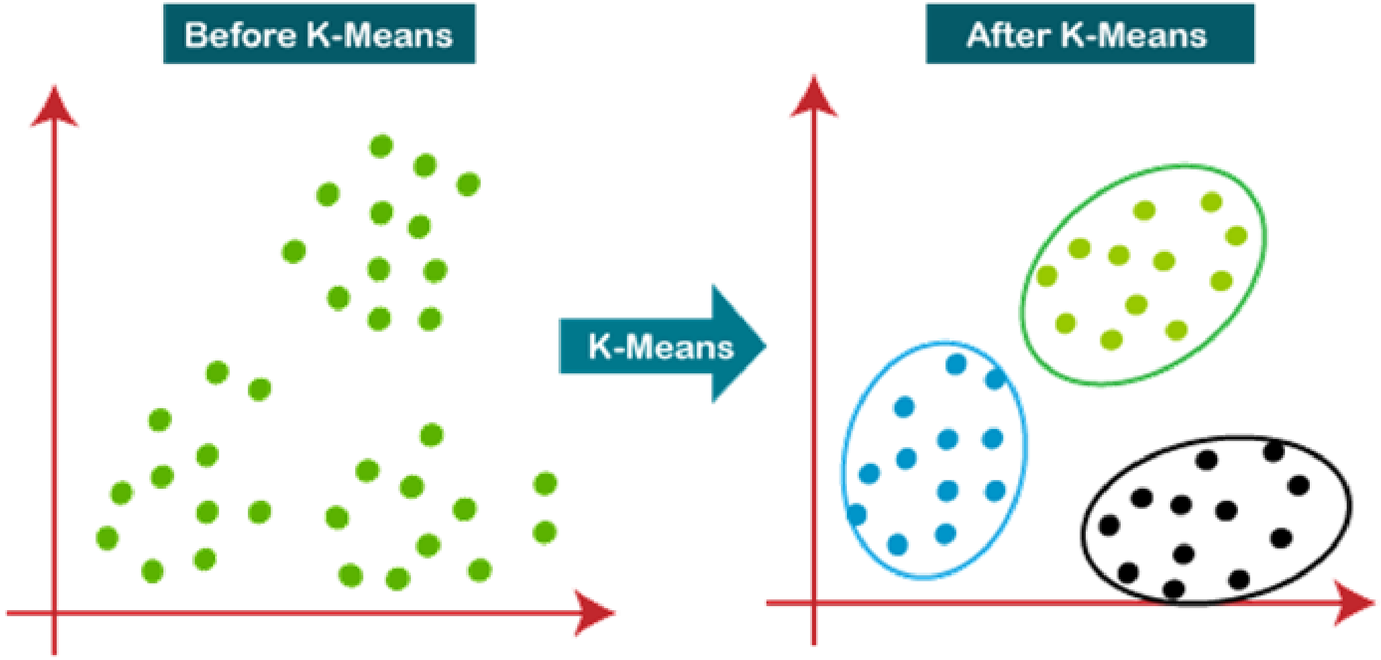
- K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- 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.
- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of the distance between the data point and their corresponding clusters.
- 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 k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near the particular k-center, create a cluster.

Hence each cluster has data points with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



ALGORITHM

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 reassign 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.

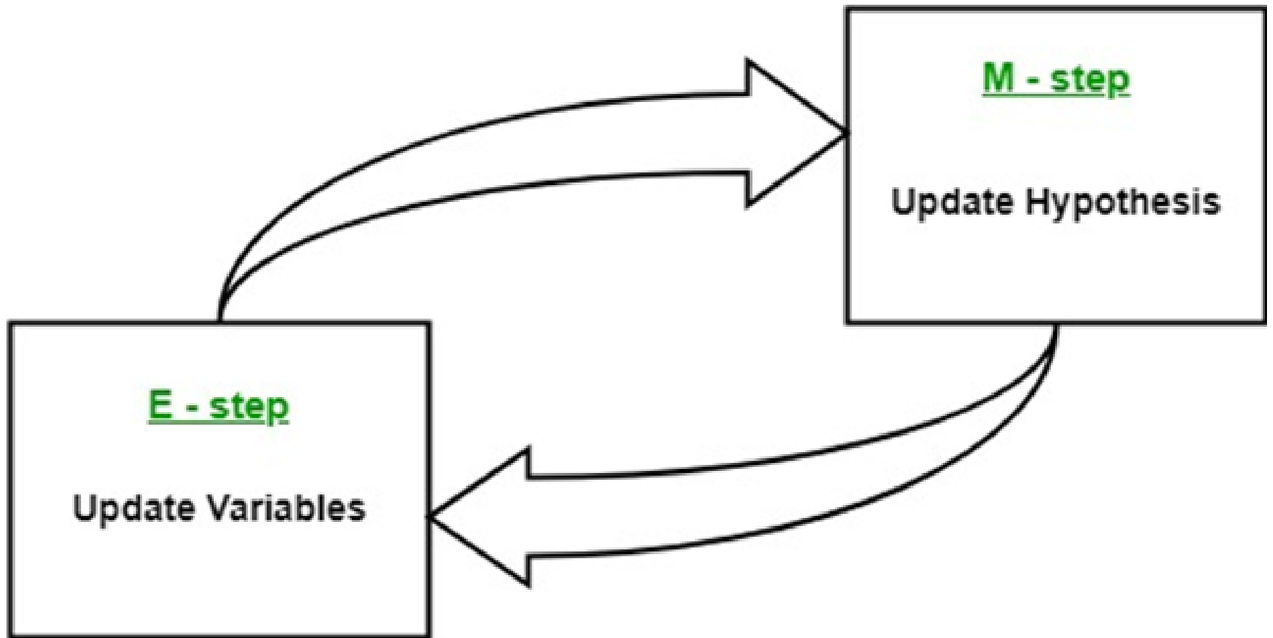
EXPECTATION-MAXIMIZATION ALGORITHM

ALGORITHM

- Expectation-Maximization algorithm can be used for the latent variables (variables that are not directly observable and are actually inferred from the values of the other observed variables) too in order to predict their values with the condition that the general form of probability distribution governing those latent variables is known to us.
- This algorithm is actually at the base of many unsupervised clustering algorithms in the field of machine learning.
- It was explained, proposed and given its name in a paper published in 1977 by Arthur Dempster, Nan Laird, and Donald Rubin.
- It is used to find the local maximum likelihood parameters of a statistical model in the cases where latent variables are involved and the data is missing or incomplete.

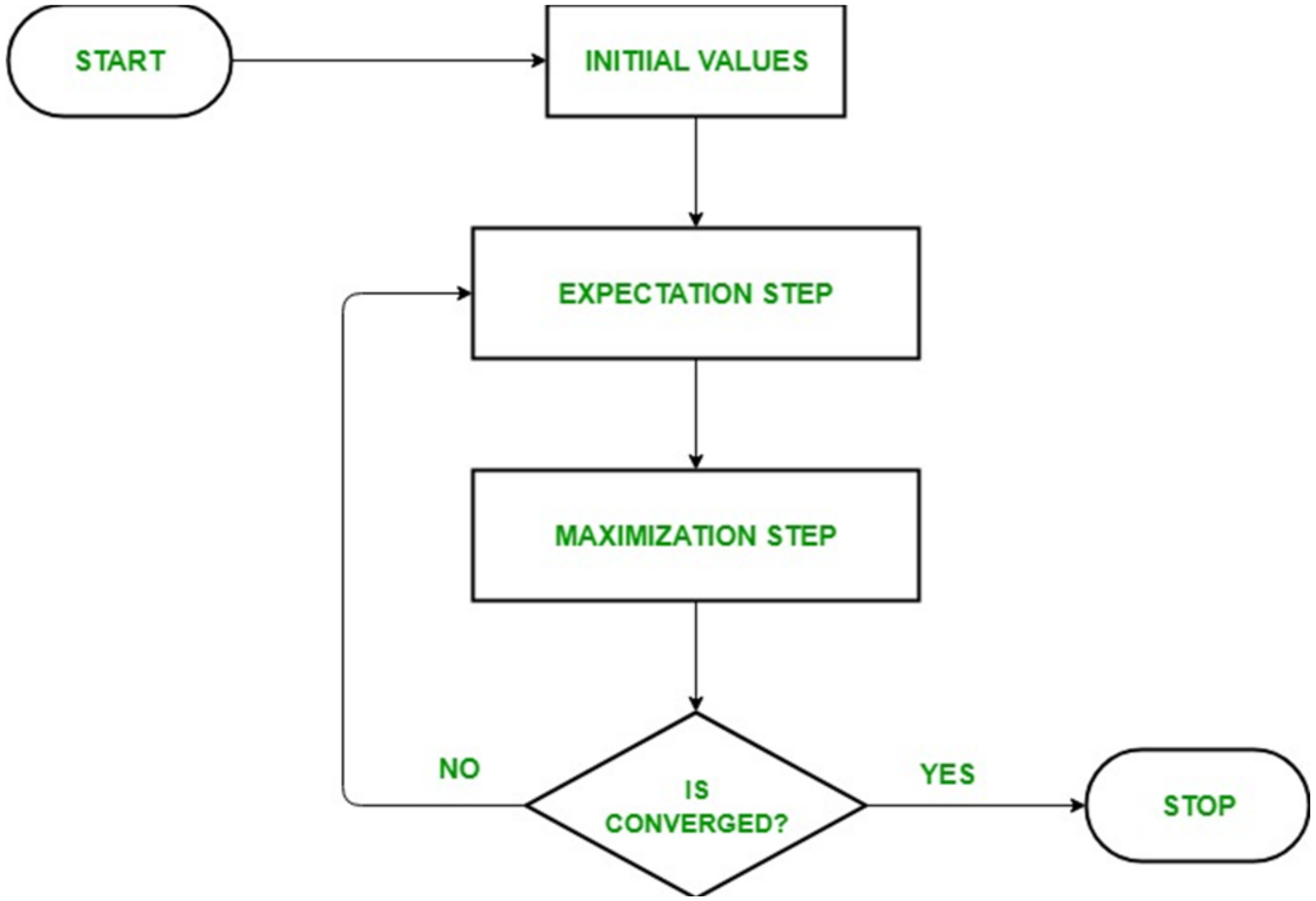
ALGORITHM

1. Given a set of incomplete data, consider a set of starting parameters.
2. Expectation step (E – step): Using the observed available data of the dataset, estimate (guess) the values of the missing data.
3. Maximization step (M – step): Complete data generated after the expectation (E) step is used in order to update the parameters.
4. Repeat step 2 and step 3 until convergence.



- Initially, a set of initial values of the parameters are considered. A set of incomplete observed data is given to the system with the assumption that the observed data comes from a specific model.
- The next step is known as the “Expectation” – step or E-step. In this step, we use the observed data in order to estimate or guess the values of the missing or incomplete data. It is basically used to update the variables.
- The next step is known as the “Maximization”-step or M-step. In this step, we use the complete data generated in the preceding “Expectation” – step in order to update the values of the parameters. It is basically used to update the hypothesis.
- Now, in the fourth step, it is checked whether the values are converging or not, if yes, then stop otherwise repeat step-2 and step-3 i.e. “Expectation” – step and “Maximization” – step until the convergence occurs.

FLOW CHART


START
INITIIAL VALUES
EXPECTATION STEP
MAXIMIZATION STEP
IS CONVERGED?
NO
YES
STOP

USAGE OF EM ALGORITHM :-

- It can be used to fill the missing data in a sample.
- It can be used as the basis of unsupervised learning of clusters.
- It can be used for the purpose of estimating the parameters of Hidden Markov Model (HMM).
- It can be used for discovering the values of latent variables.

ADVANTAGES OF EM ALGORITHM:-

- It is always guaranteed that likelihood will increase with each iteration.
- The E-step and M-step are often pretty easy for many problems in terms of implementation.
- Solutions to the M-steps often exist in the closed form.

DISADVANTAGES OF EM ALGORITHM:-

- It has slow convergence.
- It makes convergence to the local optima only.
- It requires both the probabilities, forward and backward (numerical optimization requires only forward probability).

HIERARCHIAL CLUSTERING

- Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and is also known as **hierarchical cluster analysis** or HCA.
- In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.

The hierarchical clustering technique has two approaches:

1. Agglomerative: Agglomerative is a **bottom-up** approach, in which the algorithm starts by taking all data points as single clusters and merging them until one cluster is left.

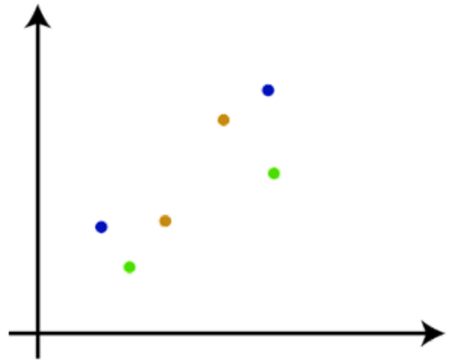
2. Divisive: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

AGGLOMERATIVE HIERARCHICAL CLUSTERING

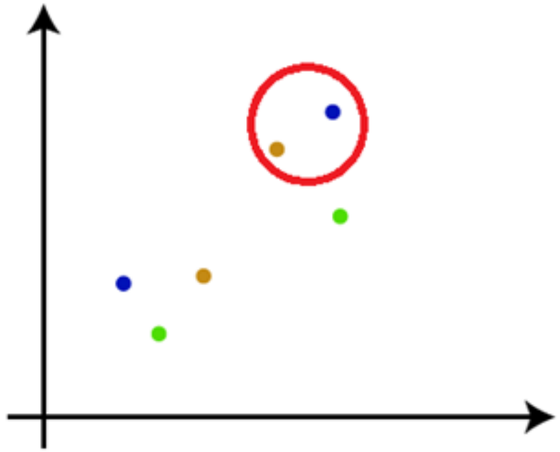
- The agglomerative hierarchical clustering algorithm is a popular example of HCA.
- To group the datasets into clusters, it follows the bottom-up approach. It means, this algorithm considers each dataset as a single cluster at the beginning, and then start combining the closest pair of clusters together. It does this until all the clusters are merged into a single cluster that contains all the datasets.
- This hierarchy of clusters is represented in the form of the dendrogram.

ALGORITHM

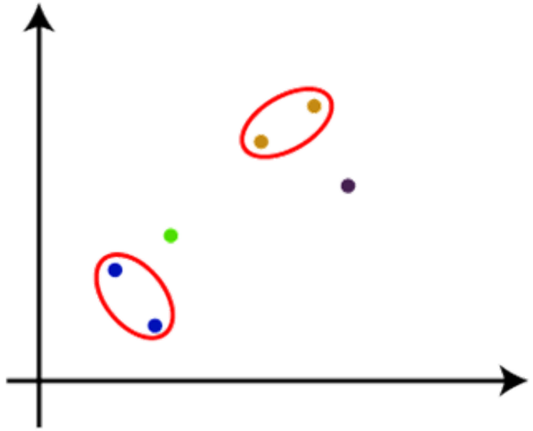
- Step-1: Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.



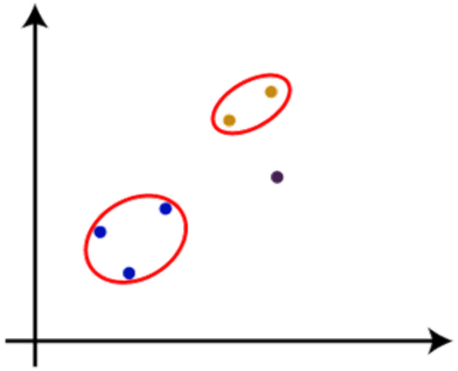
- Step-2: Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.

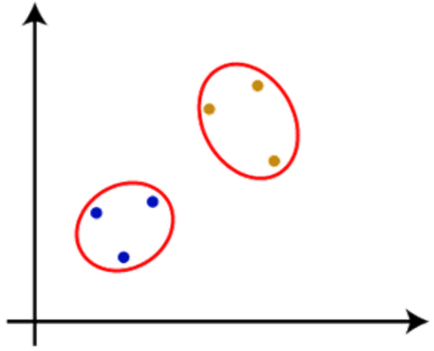


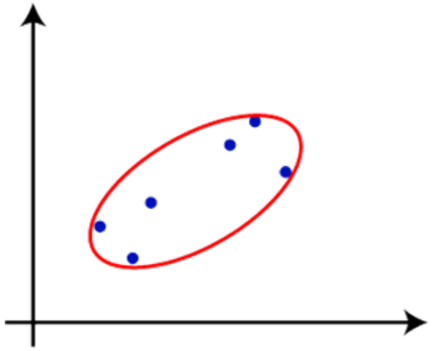
- Step-3: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.



- Step-4: Repeat Step 3 until only one cluster left. So, we will get the following clusters. Consider the below images:



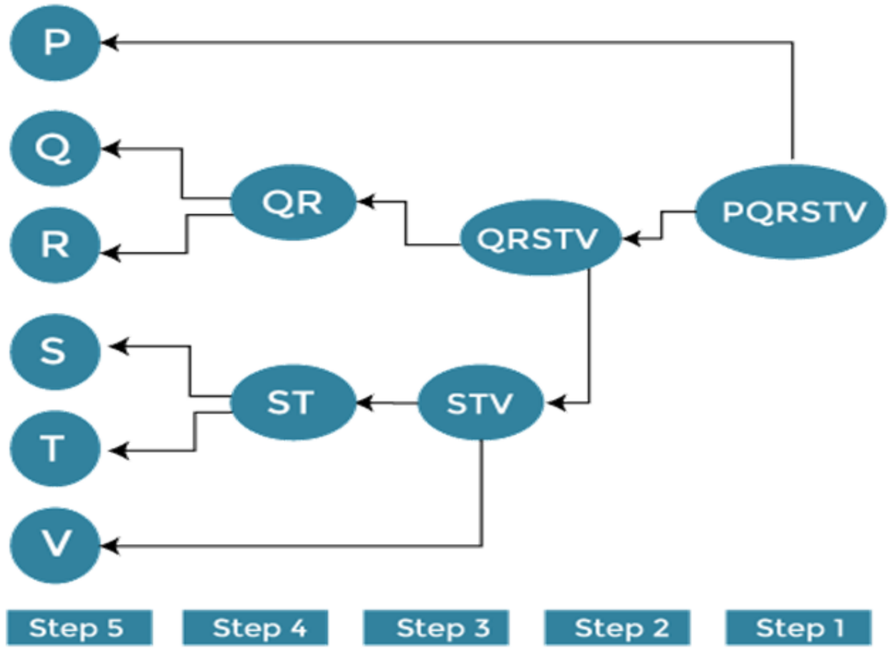




- Step-5: Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

DIVISIVE HIERARCHIAL CLUSTERING

- Divisive hierarchical clustering is exactly the opposite of 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 clusters.
- Finally, we are left with N clusters.



ADVANTAGES OF HIERARCHIAL CLUSTERING

- It is simple to implement and gives the best output in some cases.
- It is easy and results in a hierarchy, a structure that contains more information.
- It does not need us to pre-specify the number of clusters.

DISADVANTAGES OF HIERARCHIAL CLUSTERING

- It breaks the large clusters.
- It is Difficult to handle different sized clusters and convex shapes.
- It is sensitive to noise and outliers.
- The algorithm can never be changed or deleted once it was done previously.

THANK YOU!

Dhanish
Ahamed