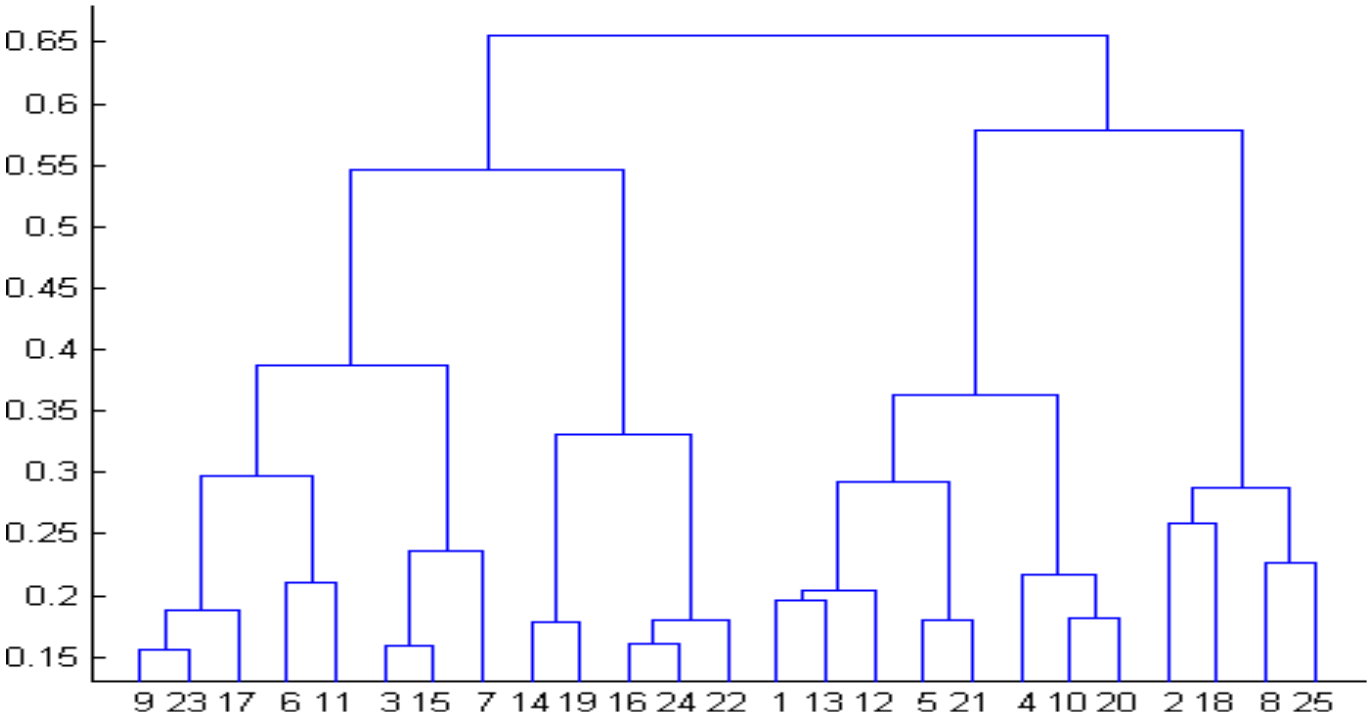
# MACHINE LEARNING

## ASSIGNMENT – 1

**Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.**

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:




a) 2
b) 4
c) 6
d) 8
Ans :- b) 4

2. In which of the following cases will K-Means clustering fail to give good results?
1. Data points with outliers
2. Data points with different densities
3. Data points with round shapes
4. Data points with non-convex shapes
Options:
a) 1 and 2
b) 2 and 3
c) 2 and 4
d) 1, 2 and 4
Ans :- d) 1, 2 and 4

3. The most important part of is selecting the variables on which clustering is based.
a) interpreting and profiling clusters
b) selecting a clustering procedure
c) assessing the validity of clustering
d) formulating the clustering problem
Ans :- d) formulating the clustering problem

4. The most commonly used measure of similarity is the or its square.
a) Euclidean distance
b) city-block distance
c) Chebyshev's distance

d) Manhattan distance
Ans :- a) Euclidean distance

5. is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.
a) Non-hierarchical clustering
b) Divisive clustering
c) Agglomerative clustering
d) K-means clustering
Ans :- b) Divisive clustering

6. Which of the following is required by K-means clustering?
a) Defined distance metric
b) Number of clusters
c) Initial guess as to cluster centroids
d) All answers are correct
Ans :- b) Number of clusters

7. The goal of clustering is to-
a) Divide the data points into groups
b) Classify the data point into different classes
c) Predict the output values of input data points
d) All of the above
Ans :- a) Divide the data points into groups

8. Clustering is a-
a) Supervised learning
b) Unsupervised learning
c) Reinforcement learning
d) None
Ans :- b) Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
a) K- Means clustering
b) Hierarchical clustering
c) Diverse clustering
d) All of the above
Ans :- d) All of the above

10. Which version of the clustering algorithm is most sensitive to outliers?
a) K-means clustering algorithm
b) K-modes clustering algorithm
c) K-medians clustering algorithm
d) None
Ans :- a) K-means clustering algorithm

11. Which of the following is a bad characteristic of a dataset for clustering analysis-
a) Data points with outliers
b) Data points with different densities
c) Data points with non-convex shapes
d) All of the above
Ans :- d) All of the above

12. For clustering, we do not require-
a) Labeled data
b) Unlabeled data
c) Numerical data
d) Categorical data
Ans :- a) Labeled data

**Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.**

13. How is cluster analysis calculated?

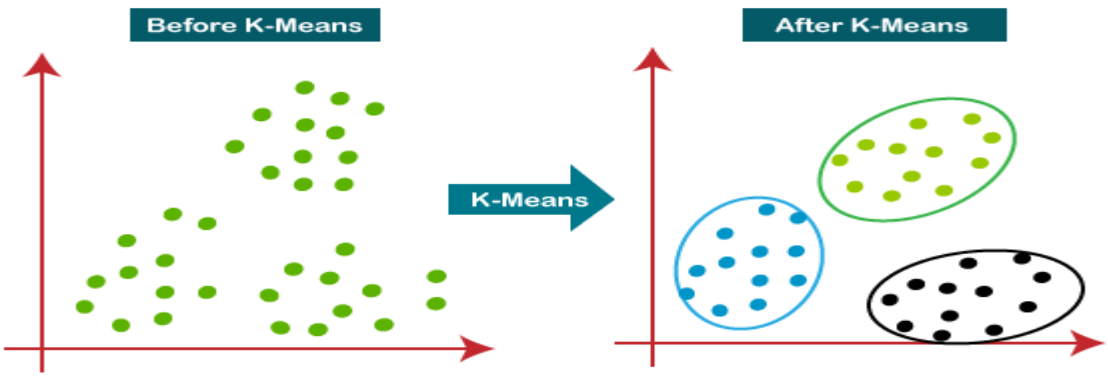
Ans :- The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

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 to the particular k-center, create a cluster.

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

The below diagram explains the working of the K-means Clustering 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.

14. How is cluster quality measured?

Ans :-

15. What is cluster analysis and its types?

Ans :- Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg., products, respondents, or other entities) based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.

There are a number of different methods to perform cluster analysis. Some of them are,

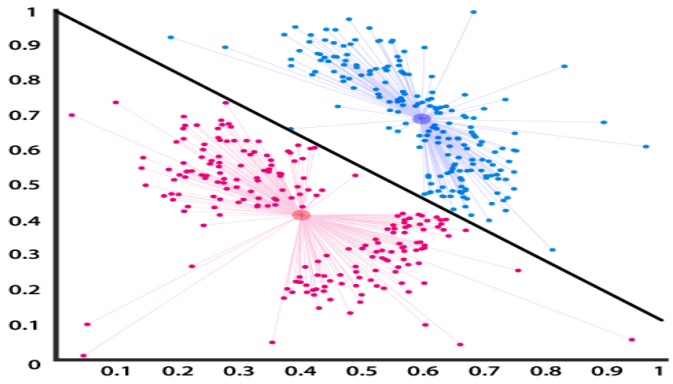
1. Hierarchical Cluster Analysis

In this method, first, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as Agglomerative method. Agglomerative clustering starts with single objects and starts grouping them into clusters.

The divisive method is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.

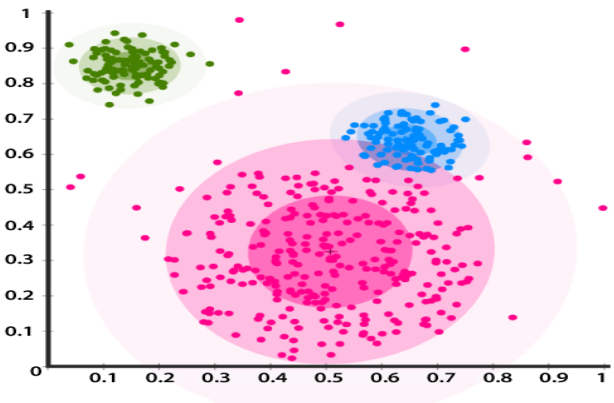
2. Centroid-based Clustering

In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centers and objects are assigned to the nearest cluster centers.




3. Distribution-based Clustering

It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster. This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.




4. Density-based Clustering

In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters. The objects in these sparse points are usually noise and border points in the graph. The most popular method in this type of clustering is DBSCAN.