MACHINE LEARNING ASSIGNMENT –

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

Ans 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

Ans 1, 2 and 4

3. The most important part of \_\_\_\_\_\_\_\_is selecting the variables on which clustering is based.

Ans formulating the clustering problem

4. The most commonly used measure of similarity is the\_\_\_ or its square.

Ans 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.

Ans 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 All of the above

7. The goal of clustering is to?

Ans Divide the data points into groups

8. Clustering is a?

Ans 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 all of the above

10. Which version of the clustering algorithm is most sensitive to outliers?

Ans 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 All of the above

12. For clustering, we do not require-

Ans Labeled data

13. How is cluster quality measured?

Ans To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set. The silhouette coefficient and other intrinsic measures can also be used in the elbow method to heuristically derive the number of cluster in a data set by replacing the sum of with-in cluster variances.

14. How is cluster analysed calculated ?

Ans it is calculated by measuring the distance between each data point and its centroid, squaring this distance, and summing these squares across one cluster. A good model is one with low inertia and a low number of cluster.

15. What is cluster analysis and its types?

Ans CLUSTER ANALYSIS- cluster analysis is a multivariate data mining technique whose goal is to group objects(e.g., 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.

TYPES OF CLUSTER ANALYSIS

* 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 method is known as agglomerative method.
* CENTROID BASED CLSUTING – in this type of clustering, clusters are represented by a central entity, which may or may not be a part if the given data set. K-Means method of clustering is used in this method, where k are the cluster centres objects are assigned to the nearest cluster centres.
* DISTRIBUTION BASED CLUSTERING – it is a type of clustering model closely related to statistics based on the models 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 the objects like correlation and dependence between attributes.
* DENSITY BASED CLUSTERING – in this type of clustering, clusters are defined by the area of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate cluster. The object in these sparse points are usually noise and border points in the graph. The most popular method in this type of clustering is DBSCAN.