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:

b) 4

2. In which of the following cases will K-Means clustering fail to give good results?
d) 1, 2 and 4

3. The most important part of is selecting the variables on which clustering is based.
d) formulating the clustering problem

4. The most commonly used measure of similarity is the or its square.
a) Euclidean distance

MACHINE LEARNING
ASSIGNMENT – 1

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.
b) Divisive clustering

6. Which of the following is required by K-means clustering?
d) All answers are correct

7. The goal of clustering is to-
a) Divide the data points into groups

8. Clustering is a-
b) Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
d) All of the above

10. Which version of the clustering algorithm is most sensitive to outliers?
a) K-means clustering algorithm

11. Which of the following is a bad characteristic of a dataset for clustering analysis-
d) All of the above

12. For clustering, we do not require-
a) Labeled data

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

13. How is cluster analysis calculated?

The hierarchical cluster analysis follows three basic steps:
1) calculate the distances,
2) link the clusters, and
3) choose a solution by selecting the right number of clusters.

14. How is cluster quality measured?

For measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available.
Here, *ground truth* is the ideal clustering that is often built using human experts.
If ground truth is available, it can be used by
extrinsic methods, which compare the clustering against the group truth and measure.
If the ground truth is unavailable, we can use
intrinsic methods, which evaluate the goodness of a clustering by considering how well the clusters are separated.

15. What is cluster analysis and its types?

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

TYPES:-

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.

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 centres.

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.

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.