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

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

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

b) selecting a clustering procedure
c) assessing the validity of clustering

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

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
b) Classify the data point into different classes
c) Predict the output values of input data points
d) All of the above

8. Clustering is a

b) Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
a) K- Means clustering

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?

1. Calculate the distance
2. Link the clusters.
3. Choose a solution by selecting the right no of clusters.

14. How is cluster quality measured?

To measure the quality of a clustering use the average silhouette coefficient value of all the objects in the data set.

The Silhouette Coeffcient is calculated using the mean intra –cluster distance (a) and the mean nearest –cluster distance (b) for each sample. The Sihouette Coefficient for a sample is (b-a)/max(a,b). To clarify ,b is the distance between a sample and the nearest cluster that the sample is not a part of it.

15. What is cluster analysis and its types?

Cluster analysis is a multivariate data mining technique whose goal is to group objects based on a set of user selected characteristics or attributes.

Type of Clustering

1. Centroid –based Clustering.
2. Density – based Clustering
3. Distribution- based Clustering.
4. Hierarchical Clustering.