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?
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:

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

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.

c) Agglomerative 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?

Ans. There are three steps that are necessary to calculate cluster analysis:

1) Copy your data into the table

2) Select more than one variable

3) Select the number of clusters you want to calculate

Clusters can be calculated using various grouping methods. These can be divided into

A) graph-theoretical

B) hierarchically

C) partitioning

D) optimizing

Data tab calculates you the k-means Cluster and hierarchical cluster.

14. How is cluster quality measured?

Ans. The Cluster quality can be measured as follows:

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric:
- There is a separate "quality" function that measures "goodness" of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, Boolean, categorical, and ordinal variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define "similar enough" or "good enough"

15. What is cluster analysis and its types?

Ans. Cluster analysis is an exploratory analysis that tries to identify structures within the data.

- Cluster: a collection of data objects

– Similar to one another within the same cluster – Dissimilar to the objects in other clusters

- Cluster analysis

– Grouping a set of data objects into clusters

Types of Cluster analysis

- Interval-scaled variables

.e.g., salary, height

- Binary variables
- e.g., gender (M/F), has cancer(T/F)
- Nominal (categorical) variables
- e.g., religion (Christian, Muslim, Buddhist, Hindu, etc.)
- Ordinal variables
- e.g., military rank (soldier, sergeant, captain, etc.)
- Ratio-scaled variables
- population growth (1,10,100,1000,...)
- 2 Variables of mixed types
- multiple attributes with various types