MACHINE LEARNING - ASSIGNMENT – 1

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

Answer- 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 shape

Options: a) 1 and 2 b) 2 and 3 c) 2 and 4 d) 1, 2 and 4

Answer- 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

Answer- 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

Answer- 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

Answer- 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

Answer- 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

Answer- a) Divide the data points into groups

8. Clustering is a

a) Supervised learning

b) Unsupervised learning

c) Reinforcement learning

d) None

Answer- 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

Answer- 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

Answer- 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

Answer- d) All of the above

12. For clustering, we do not require

a) Labeled data

b) Unlabeled data

c) Numerical data

d) Categorical data

Answer- a) Labeled data

13. How is cluster analysis calculated?

Answer-Cluster Analysis is calculated by k-Mean clustering algorithm .It has following steps:

- Choose the number of cluster k
- Make an initial selection of k centroids
- Assign each data element to its nearest centroid (in this way k clusters are formed one for each centroid, where each cluster consist of all the data elements assigned to that centroid)
- For each cluster make a new selection of its centroid
- Go back to step 3 , repeating the process untill the centroids does not change (or some other convergence criterion is met)

14. How is cluster quality measured?

Answer- The cluster quality is measured by the average Silhouette coefficient value of all objects in the dataset .It is calculated using the mean intra-cluster distance and mean-nearest cluster distance. Silhouette Score ranges between [–1, 1], where higher the score the more well-defined and distinct cluster are.

15. What is cluster analysis and its types?

Answer-Clustering Analysis is a form of exploratory data analysis in which observations are divided into different groups that shares common characteristics. The purpose of cluster analysis is construct groups while ensuring following property, within group observation must be simmilar as possible, while observation belonging to different groups must be different as possible.

It is basically three types:

- K-Means Clustering
- Hierachical Clustering
- DB-SCAN