**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

**Ans-** 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:

a) 1 and 2

b) 2 and 3

c) 2 and 4

d) 1, 2 and 4

**Ans –** 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

**Ans –**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

**Ans –**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

**Ans –**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

**Ans--** 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

**Ans –**a) Divide the data points into groups

8. Clustering is a-

a) Supervised learning

b) Unsupervised learning

c) Reinforcement learning

d) None

**Ans –**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

**Ans –**a) K-Means clustering

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

**Ans –**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

**Ans --** d) All of the above

12. For clustering, we do not require-

a) Labeled data

b) Unlabeled data

c) Numerical data

d) Categorical data

**Ans—**a) Labeled data

13. How is cluster analysis calculated?

**Ans—**Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than those in other groups. There are different types of cluster analysis such as hierarchical clustering , k-means clustering ,distribution based clustering ,density based clustering. Etc.

K-Means clustering –i) it is used to partition the input dataset into k clusters.

ii) Select randomly k objects from datasets as initial cluster centers.

iii) Assign each observation to their closest centroid based on Euclidean distance between object and centroid.

iv) For each clusters update the centroid of clusters by calculating the mean values of all data points in that cluster

v) iterate steps iii and iv until cluster assignments stop changing

14. How is cluster quality measured?

**Ans—**Ground truth is the ideal clustering that are often built using human experts. Based on the availability of ground truth quality measurement methods can be divided into two. They are extrinsic or supervised methods if ground truth is available and intrinsic or unsupervised methods if ground truth is unavailable

Clustering quality is more if it satisfies the following criteria:

i)cluster homogeneity – more pure the clusters are,the better the clustering.

Q(q2,qg) >Q(q1,qg)

Q(q2,qg) 🡪cluster score of second cluster

Q(q1,qg) 🡪cluster score of first cluster

ii)cluster completeness –if any two objects belong to same category according to ground truth then they should be assigned to same cluster.

iii)Rag bag –They contain objects that cannot be merged with other objects. Ragbag criterion states that putting an heterogeneous object into a pure cluster should be penalized more than putting it into a ragbag.

iv)small cluster preservation-  If a small category is split into small pieces in a clustering, those small pieces may likely become noise and thus the small category cannot be discovered from the clustering.

15. What is cluster analysis and its types?

**Ans—**Cluster analysis is the task of grouping a set of datapoints in such a way that they can be characterised by their relevancy to one another.

There are 4 types of cluster analysis techniques. They are

i)centroid clustering –Here we choose the number of clusters that we want to classify. Algorithm starts by randomly selecting the centroids to group the data. The centroids and points ina cluster will adjust through all iterations resulting in optimized clusters.

ii)Density clustering –It groups the data points by how densely they are distributed. Algorithm says that the more dense the data is more related they are .

iii)Distribution clustering-It identifies the probability that a point belongs to a cluster. Around each possible centroid the algorithm defines the density distribution of each cluster.

iv)Connectivity clustering –It initially recognizes each datapoint as its own cluster. Primarily it is assumed that points closer to each other are more related.Then continually incorporate datapoints with other datapoints until all data points are engulfed into one big cluster.