Machine Learning Assignment

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

L6. 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: (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

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

Q**13 to Q15 are subjective answers type questions, Answers them in their own words briefly.**
13. How is cluster analysis calculated?

Ans: Process of cluster analysis -

(a) Firstly, determine the number of cluster.

(b)Secondly, select a distance matric to measure the similarity between data points that can be Euclidean distance or any other.

(c) Choose any algorithm that can be K-means or any other and then run the algorithm on the given data.

(d) Finally, by examining the cluster centroid we can evaluate the result of cluster algorithm.

14. How is cluster quality measured?

Ans: Some metrics are given below to measured the quality of cluster -

(a) Silhouette score: By this metric, we can evaluate at which extinct the data points are fitted into the cluster and also what is the length of intracluster and intercluster. If the value of this Silhouette score is high, that means the data points or observations are well clustered and also the cluster is well seperated.

(b) Calinski-Harabasz Index: By help of this matric we measure the ratio of intercluster to intracluster. i.e ratio of distance between the cluster to distance between two data points within the cluster.

So, the value of Calinski-Harabasz Index is more that means data points between each cluster are tightly packed and the cluster are well seperated with each other.

15. What is cluster analysis and its types?

Ans: It is the technique by the help of which all the data points or observations are segregated into different groups or cluster according to their common characteristics. It doesn't require the labeled data for analysis. The algorithm runs and try to find similarity between the data points based on the details provided.

Types of Cluster Analysis -

(a) K-means Clustering

(b) Hierarchical Clustering