Work sheet 1 MACHINE learning

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 :- (a)

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

Answer :- (d) 1,2&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 :- (c) Agglomerative 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 :- (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

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 :- clustering is an unsupervised learning , it deals with finding a structure or patterns in a collection of unsupervised data. It is calculated using various algorithms that group similar data points together based on similarity measures.

The process of cluster analysis typically involves the following steps

>data preprocessing

>distance/similarity calculations

>selecting the number of clusters

> Visualization

14. How is cluster quality measured?

Answer: Cluster quality can be measured using several evaluation metrics, such as:External evaluation metrics, which use external information, such as class labels, to evaluate the quality of the clusters. Examples include adjusted Rand index, Fowlkes-Mallows index, and Jaccard similarity coefficient. Internal evaluation metrics, which use only the information within the clusters to evaluate the quality of the clusters. Examples include silhouette coefficient, Calinski-Harabasz index, and Davies-Bouldin index. Other metrics such as purity, entropy, and mutual information.

15. What is cluster analysis and its types?

Answer: Cluster analysis is a method used to classify a set of objects into groups (clusters) based on their similarity. The goal of cluster analysis is to divide a set of data points into clusters such that the points within a cluster are as similar as possible to each other, and as dissimilar as possible to the points in other clusters.

There are two main types of cluster analysis:

Hierarchical clustering: This type of clustering builds a hierarchy of clusters, where each cluster is a subset of the previous one. There are two types of hierarchical clustering: agglomerative and divisive. In agglomerative clustering, the process starts with each data point being a separate cluster, and then the closest clusters are merged until there is only one cluster remaining. In divisive clustering, the process starts with all the data points in one cluster, and then the cluster is split until each data point is in its own cluster.

Partitioning clustering: This type of clustering divides the data points into a fixed number of clusters. K-means is the most commonly used partitioning method, which partitions the data points into k clusters by iteratively moving the centroid of each cluster to the mean of the data points assigned to the cluster.