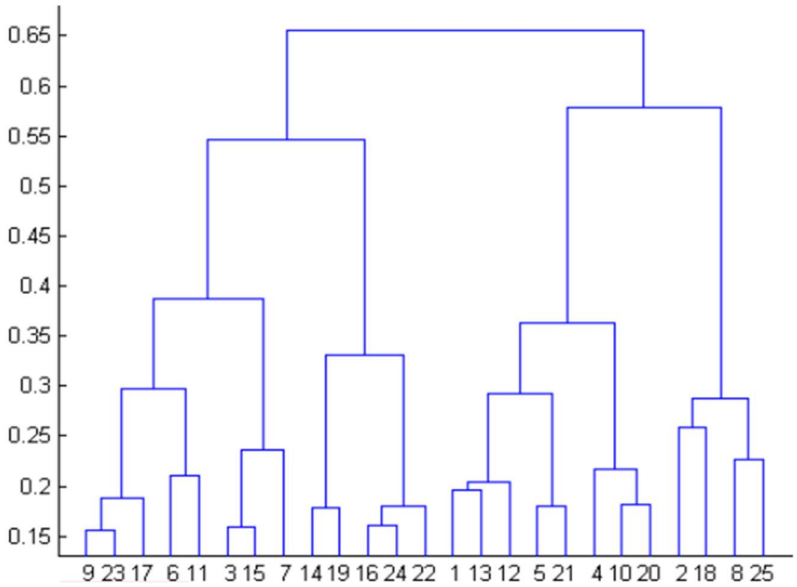
Assignment-1

MACHINE LEARNING

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:



a) 2
b) 4
c) 6
d) 8

Answer (b)

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

Assignment-1

b) 2 and 3
c) 2 and 4
d) 1, 2 and 4

Answer (d)

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)

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)

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)

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

Assignment-1

Answer (d)

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)

8. Clustering is a-

a) Supervised learning
b) Unsupervised learning
c) Reinforcement learning
d) None

Answer (b)

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)

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)

11. Which of the following is a bad characteristic of a dataset for clustering analysis-

Assignment-1

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)

12. For clustering, we do not require-

a) Labelled data
b) Unlabelled data
c) Numerical data
d) Categorical data

Answer (a)

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

13. How is cluster analysis calculated?

Answer-

Cluster analysis can be calculated by using different algorithm. K-means is a centroid-based clustering algorithm, where we calculate the distance between each data point and a centroid to assign it to a cluster. The goal is to identify the K number of groups in the dataset.

k-means is an iterative process of assigning each data point to the groups and slowly data points get clustered based on similar features. The objective is to minimize the sum of distances between the data points and the cluster centroid, to identify the correct group each data point should belong to.

Here, we divide a data space into K clusters and assign a mean value to each. The data points are placed in the clusters closest to the mean value of that cluster. There are several distance metrics available that can be used to calculate the distance.

- The first step is to define the K number of clusters in which we will group the data.

Assignment-1

- Centroid is the center of a cluster but initially, the exact center of data points will be unknown so, we select random data points and define them as centroids for each cluster.
- Now that centroids are initialized, the next step is to assign data points X_n to their closest cluster centroid C_k. In this step, we will first calculate the distance between data point X and centroid C using Euclidean Distance metric. And then choose the cluster for data points where the distance between the data point and the centroid is minimum.
- Next, we will re-initialize the centroids by calculating the average of all data points of that cluster.
- We will keep repeating steps 3 and 4 until we have optimal centroids and the assignments of data points to correct clusters are not changing anymore.

The number of clusters in the k-Means method must be determined before the start and is therefore not determined by the cluster method. The elbow method is a common way to determine the appropriate number of clusters.

14. How is cluster quality measured?

Answer-

Silhouette coefficient or silhouette score is a metric used to measure the quality of a clustering technique. Its value ranges from -1 to +1.

A coefficient close to -1 means clusters are assigned in the wrong way.

A coefficient close to 0 means cluster are indifferent, or we can say that the distance between clusters is not significant.

A coefficient close to +1 means clusters are well apart from each other and clearly distinguished.

Silhouette coefficient = (b - a)/max(a,b)

Where, a = the mean intra-cluster distance i.e., the mean distance to the other instances in the same cluster.

b = the mean nearest-cluster distance i.e., the mean distance to the instances of the next closest cluster.

15. What is cluster analysis and its types?

Answer-

Cluster analysis is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

Clustering analysis is an unsupervised Machine learning technique.

Types of clustering analysis-

- Centroid-based clustering (partitioning method)
- Density-based clustering
- Hierarchical clustering
- Distribution-based clustering
- Fuzzy clustering

The k-means clustering is the most common example of the centroid-based clustering.

In the hierarchical clustering, the dataset divided into cluster to create tree like structure called dendrogram.