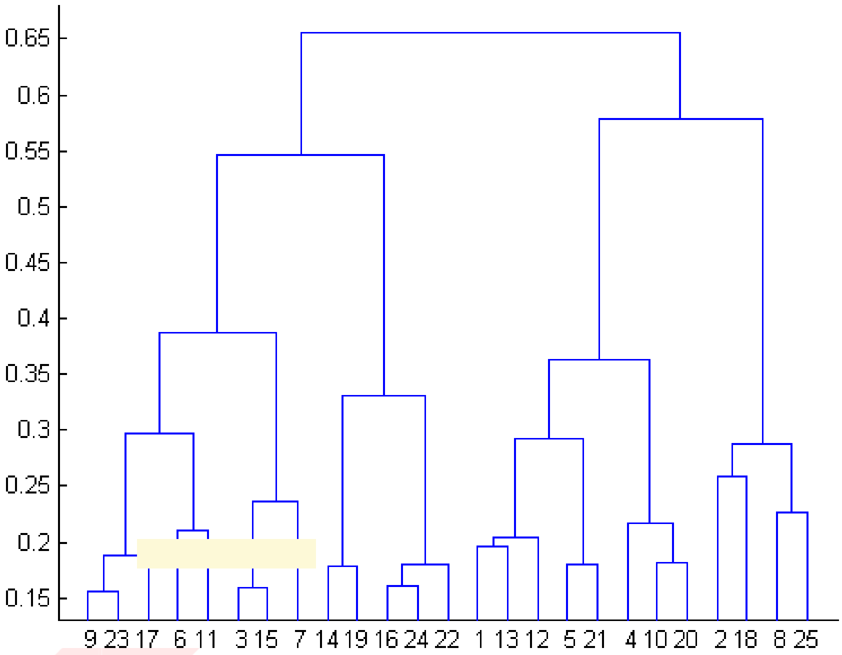
MACHINE LEARNING

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

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

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

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

MACHINE LEARNING

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

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

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

8. Clustering is a-
   a) Supervised learning
   b) Unsupervised learning
   c) Reinforcement learning
   d) None

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

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

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

12. For clustering, we do not require-
    a) Labeled data
    b) Unlabeled data
    c) Numerical data
    d) Categorical data

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

13. How is cluster analysis calculated?
14. How is cluster quality measured?
15. What is cluster analysis and its types?

MACHINE LEARNING

Q13:

In Clustering analysis data points are divided into different groups that share common characteristics.

There are 2 types of clustering analysis:

k-means clustering

Hierarchical clustering.

Clustering algorithms use the distance in order to separate observations into different groups.

Following steps used in k means clustering:

1)Distances are calculated between each data point by using eucledian distance.
2)Choose the number of clusters *k*
3) Make an initial selection of *k* centroids
4) Assign each data element to its nearest centroid (in this way *k* clusters are formed one for each centroid, where each cluster consists of all the data elements assigned to that centroid)
5) For each cluster make a new selection of its centroid
6) Go back to step 4, repeating the process until the centroids don't change.

Q14:

There are majorly two types of measures to assess the clustering Quality:

(i) Extrinsic Measures which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.

(ii) Intrinsic Measures that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.

Q15:

Cluster analysis is method of grouping a set of data points in such a way that they can be characterized by their relevancy to one another.

There are four basic types of cluster analysis they are

1)Centroid Clustering,

2)Density Clustering

3)Distribution Clustering,

4)Connectivity Clustering.

In centroid clustering ,algorithm will randomly selects centroids to group the data points into the pre-defined clusters.

In Density clustering groups data points by how densely populated they are. To group closely related data points, the algorithm partitions data based on density. It works on the principle that the more dense the data points the more related they are.

In Distribution clustering ,the probability that a point belongs to a cluster is calculated. Around each possible centroid The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions.

In connectivity clustering initially each data point is recognized as its own cluster. The primary technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are formed into one big cluster.

Answers:

Q1: B

Q2:D

Q3:D

Q4:A

Q5:B

Q6:D

Q7:A

Q8:B

Q9:A

Q10:A

Q11:D

Q12:B