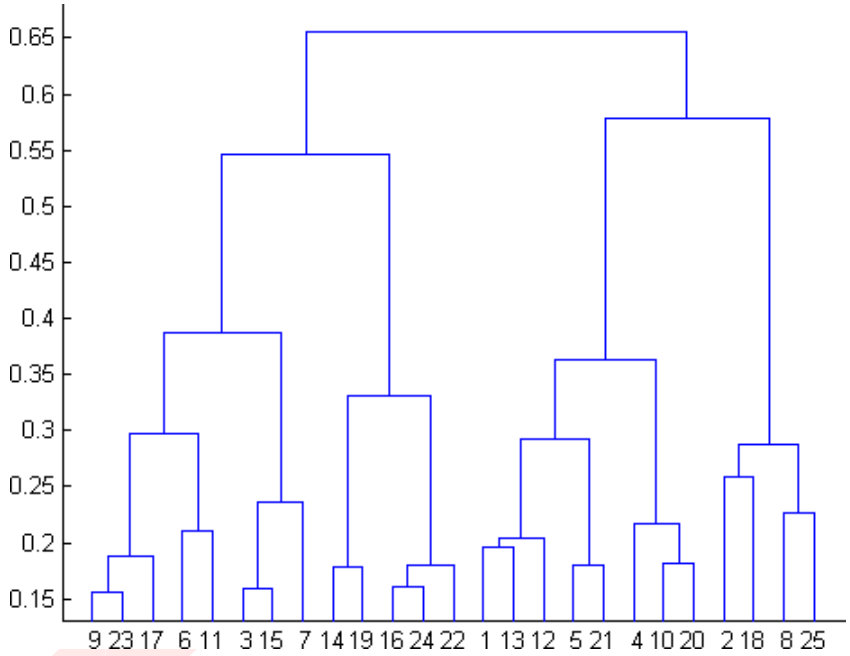
MACHINE LEARNING

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?

Answer: Cluster analysis is a process of grouping similar objects together, based on the similarity or distance between the objects. The specific method used to calculate this similarity or distance will depend on the type of data and the research question being addressed. Some commonly used methods for calculating similarity or distance include:

MACHINE LEARNING

- Euclidean distance: This method calculates the straight-line distance between two objects.
- Manhattan distance: This method calculates the distance between two objects by measuring the absolute differences of their coordinates and summing them.
- Cosine similarity: This method calculates the similarity between two objects based on the angle between their vectors.

Once the similarity or distance between objects is calculated, different clustering algorithms can be applied to group the objects into clusters. Some popular clustering algorithms include:

- k-means: This algorithm divides a dataset into k clusters, where each cluster is defined by the mean of the points within it.
- Hierarchical clustering: This algorithm builds a hierarchy of clusters, where each cluster is a subset of the previous one.
- Density-based clustering: This algorithm groups together objects that are closely packed together, and separates objects that are more sparsely located.

The choice of similarity or distance measure and the clustering algorithm used will depend on the characteristics of the data and the research question.

14. How is cluster quality measured?

Answer: Cluster quality can be measured using several metrics, such as:

- Silhouette score, which measures the similarity of each point to its own cluster compared to other clusters.
- Calinski-Harabasz index, which measures the ratio of between-cluster variance to within-cluster variance.
- Davies-Bouldin index, which measures the average similarity between each point's cluster and the clusters around it.
- CH index (Calinski and Harabasz)
- DB index (Davies and Bouldin)
- I-index (Xie and Beni)
- Rand index, which compares the pairs of points that are either in the same or different clusters in the predicted and true clusters.

It's important to note that no single metric is perfect and multiple evaluation metric should be used to get a comprehensive view of the quality of clustering.

15. What is cluster analysis and its types?

Answer: Cluster analysis, also known as clustering, is a method of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). Clustering is an unsupervised learning method, as it is used to find patterns or groupings in data without the use of predefined labels.
There are several types of cluster analysis, including:

1. Centroid-based clustering: In this method, clusters are formed around a central point, called a centroid. Examples of this type of clustering include k-means and k-medoids.
2. Hierarchical clustering: This method builds a hierarchy of clusters, where each cluster is divided into smaller clusters in a hierarchical tree structure. Examples of hierarchical clustering include agglomerative and divisive methods.
3. Density-based clustering: In this method, clusters are formed based on the density of objects in a given area. DBSCAN (Density-Based Spatial Clustering of Applications with

MACHINE LEARNING

Noise) is an example of density-based clustering.
4. Model-based clustering: In this method, clusters are modeled using a statistical distribution such as Gaussian mixture models.
5. Subspace Clustering: Clusters are formed by identifying subspaces of the feature space.
6. Spectral Clustering: Clusters are formed by using the eigenvectors of the similarity matrix of the data.

It is important to note that different types of clustering methods are more appropriate for different types of data and different clustering goals.