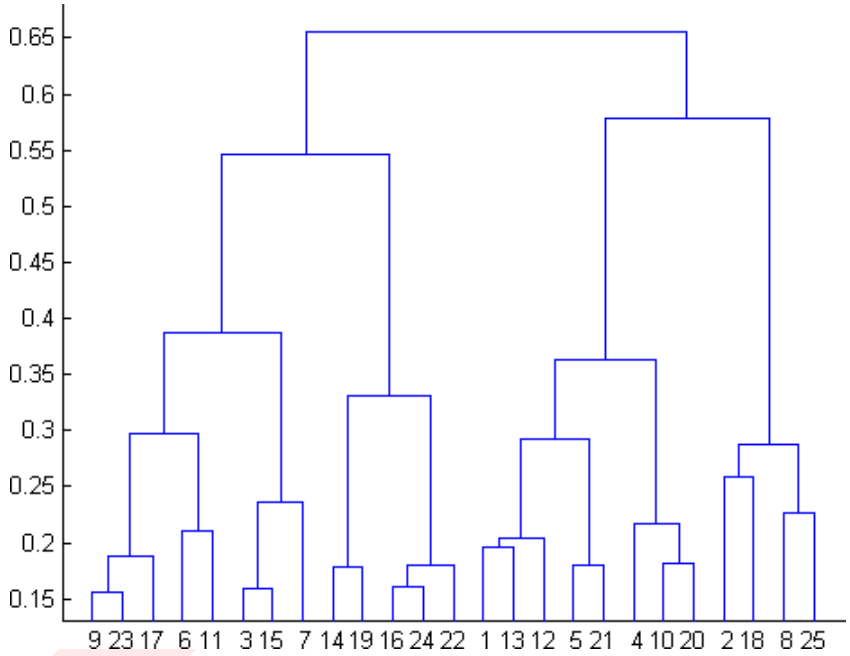
MACHINE LEARNING

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?

Ans :- To calculate cluster analysis, only three steps are necessary:

- Copy your data into the table
- Select more than one variable
- Select the number of clusters you want to calculate

MACHINE LEARNING

Clusters can be calculated using various grouping methods. These can be divided into

- graph-theoretical
- hierarchically
- partitioning
- optimizing

14. How is cluster quality measured?

Ans:- A cluster is the collection of data objects which are similar to each other within the same group.If all the data objects in the cluster are highly similar then the cluster has high quality. We can measure the quality of Clustering by using the Dissimilarity/Similarity metric in most situations.But there are some other methods to measure the Qualities of Good Clustering if the clusters are alike.

1. **Dissimilarity/Similarity metric**: The similarity between the clusters can be expressed in terms of a distance function, which is represented by d (i, j). different functions are there for various data types and data variables. Distance function measure is different for continuous-valued variables, categorical variables, and vector variables.
2. **Cluster completeness:** Cluster completeness is the essential parameter for good clustering, if any two data objects are having similar characteristics, then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category.
3. **Ragbag:** In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then the quality of those cluster categories is measured by the Rag Bag method.

15. What is cluster analysis and its types?

Ans :- Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data.

Types of Cluster Analysis

Broadly, there are 2 types of cluster analysis methods. On the basis of the categorization of data sets into a particular cluster, cluster analysis can be divided into 2 types - hard and soft clustering.

1.**Hard Clustering**

In a given dataset, it is possible for a data researcher to organize clusters in a manner that a single dataset is placed in only one of the total number of given clusters.
This implies that a hard-core classification of datasets is required in order to organize and classify data accordingly.

2.**Soft Clustering**

The second class of cluster analysis is Soft Clustering. Unlike hard clustering that requires a given data point to belong to only a cluster at a time, soft clustering follows a different rule.
In the case of soft clustering, a given data point can belong to more than one cluster at a time. This means that a fuzzy classification of datasets characterizes soft clustering.