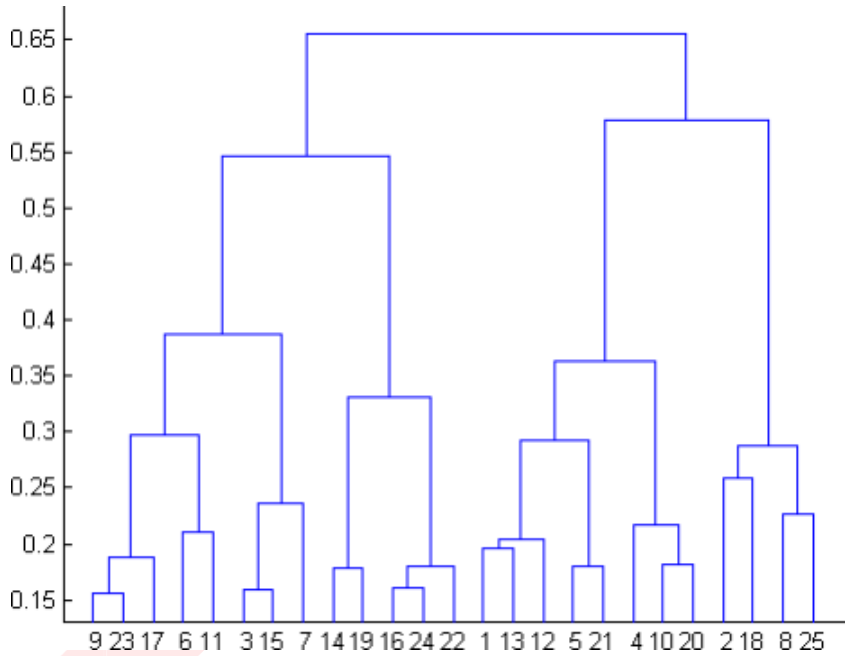
MACHINE LEARNING

#please note answers are marked in red

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 :It is calculated on the Euclidian distance between two observations ,which is the square root of the sum of squared distances. There are also three types of methods to calculate cluster analysis they are K-Means Cluster, Hierarchical Cluster, and Two-Step Cluster.

14. How is cluster quality measured?
Ans : if the data in the cluster is with high similarity then the cluster quality is said to be with high quality considering most of the hidden parts are discovered

15. What is cluster analysis and its types?

Ans : Cluster analysis is a multivariate data mining technique whose goal is to groups objects based on a set of user selected characteristics or attributes

The types are centroid-based, density-based, distribution-based, hierarchical, constraint-based, and fuzzy clustering.