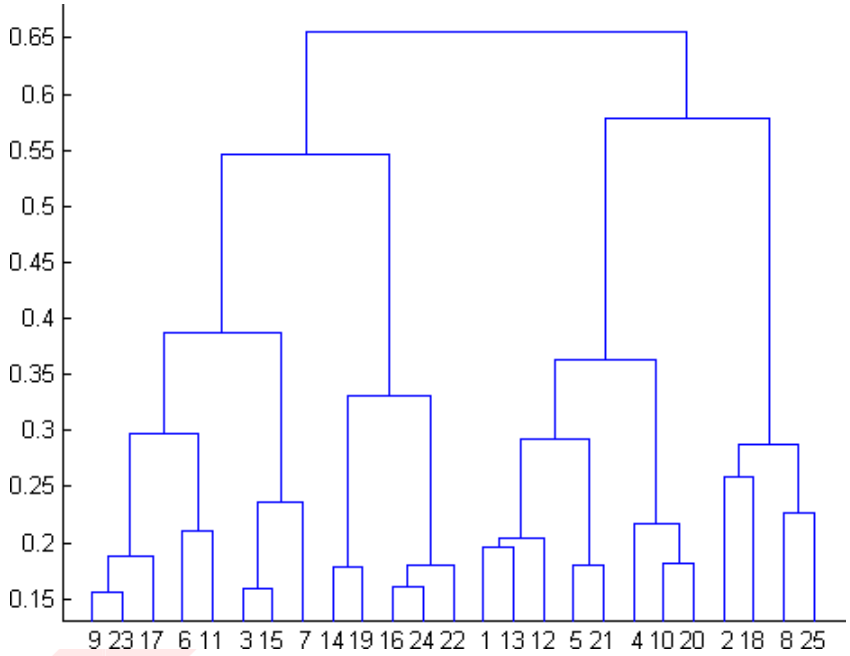
MACHINE LEARNING

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
ANS: b) 4

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
   ANS: d) 1,2and 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
   ANS : d) formulating the clustering problem

MACHINE LEARNING

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
   ANS : a) Euclidean 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
   ANS : C) Agglomerative 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
   ANS : 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
   ANS : a) Divide the data points into groups

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

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
   ANS: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
    ANS: a)K-means clustering algorithm

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
    ANS : d) All of the above

12. For clustering, we do not require-
    a) Labeled data

MACHINE LEARNING

b) Unlabeled data
c) Numerical data
d) Categorical data

ANS: a) Labeled data

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

13. How is cluster analysis calculated?

ANS: Cluster analysis differs from many other statistical methods since it's mostly used when researchers do not have an assumed principle or fact that they are using as the foundation of their research.This analysis technique is typically performed during the exploratory phase of research, since unlike techniques such as factor analysis, it doesn't make any distinction between dependent and independent variables. Instead, cluster analysis is leveraged mostly to discover structures in data without providing an explanation or interpretation.

14. How is cluster quality measured?

ANS: There are three categories to measures cluster quality

1) External measures,
2) Internal measures
3) Relative measures.

External measures - we can consider they are supervised, employ criteria not inherent to the datasets itself. That means we may have some prior or expert knowledge. For example, some ground truth. Then we can comparing the clustering results against the prior or expert specified knowledge, using certain clustering quality measure.

Internal measure - Then the second kinds of measure are called internal measure, which is unsupervised. That means the criteria derived from the data itself. In that case, we will evaluate the goodness of clustering by considering how well the clusters are separated and how compact the clusters are. For example, we can use silhouette coefficient.

Relative measure - The third one is a relative measure. That means we can directly compare different class rings using those obtained via different parameter setting for the same algorithm. For example, For the same algorithm, we use different number of clusters. We may generate different clustering results.

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. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.

MACHINE LEARNING

Types of Cluster Analysis

1) Hierarchical Cluster Analysis
2) Centroid-based Clustering
3) Distribution-based Clustering
4) Density-based Clustering