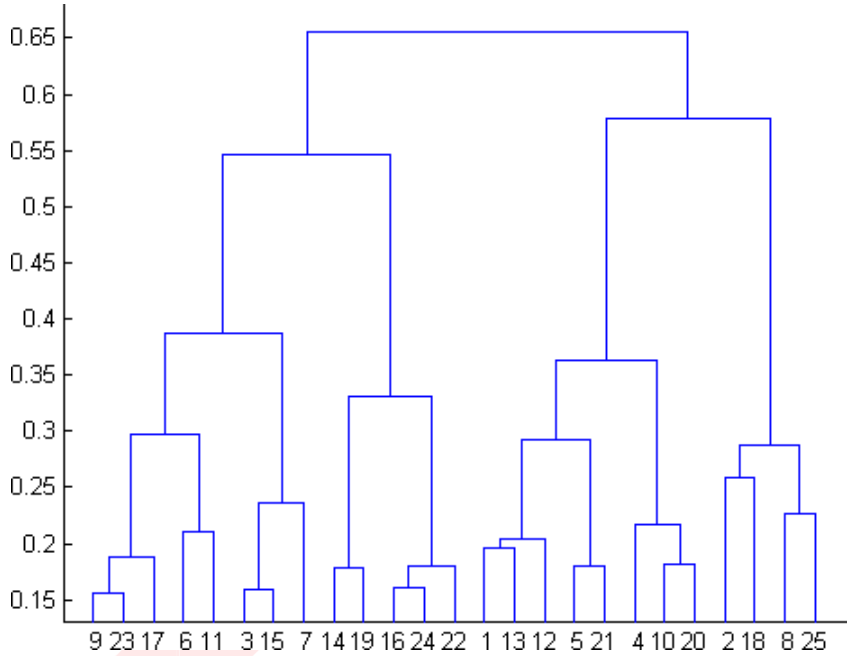
# MACHINE LEARNING

**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) 1,2 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

   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

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: (b) Divisive 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

# MACHINE LEARNING

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
    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: A rigorous cluster analysis can be conducted in 3 steps mentioned below:

a) Data preparation

b) Assessing clustering tendency (i.e., the clusterability of the data)

c) Defining the optimal number of clusters:

   We can find the optimal number of clusters by Elbow method, Average silhouette method, Gap Statistic method etc

d) Computing partitioning cluster analyses (e.g.: k-means algorithm, pam) or hierarchical clustering:

e) Validating clustering analyses: silhouette plot.

14. How is cluster quality measured?

Ans: Some of the internal measures we can deploy on clustering algorithms to measure the relative quality of different models are as follows:

a) Davies-Bouldin Index: The DB Index is calculated by the following formula:

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

where n is the number of clusters and σi is the average distance of all points in cluster i from the cluster centroid ci.

# MACHINE LEARNING

b) Dunn Index: The formula for the Dunn Index is as follows:

$$D = \frac{min_{1 \le i < j \le n} d(i,j)}{max_{1 \le k \le n} d'(k)}$$

where i, j and k are each indices for clusters, d measures the inter-cluster distance and d' measures the intra-cluster difference.
The Dunn Index captures the same idea as the DB Index: it gets better when clusters are well-spaced and dense. But the Dunn Index increases as performance improves.

c) Silhouette Coefficient: The Silhouette Coefficient is measured like so:

$$S(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

where a(i) is the average distance of point i from all other points in its cluster and b(i) is the smallest average distance of i to all points in any other cluster. To clarify, b(i) is found by measuring the average distance of i from every point in cluster A, the average distance of i from every point in cluster B, and taking the smallest resulting value.
The Silhouette Coefficient tells us how well-assigned each individual point is. If S(i) is close to 0, it is right at the inflection point between two clusters. If it is closer to -1, then we would have been better off assigning it to the other cluster. If S(i) is close to 1, then the point is well-assigned and can be interpreted as belonging to an 'appropriate' cluster.

15. What is cluster analysis and its types?

Ans: Clustering is the process by which we create groups in a data, like customers, products, employees, text documents, in such a way that objects falling into one group exhibit many similar properties with each other and are different from objects that fall in the other groups that got created during the process.

Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities.

The various types of clustering are:

- Connectivity-based Clustering (Hierarchical clustering)
- Centroids-based Clustering (Partitioning methods)
- Distribution-based Clustering
- Density-based Clustering (Model-based methods)
- Fuzzy Clustering
- Constraint-based (Supervised Clustering)