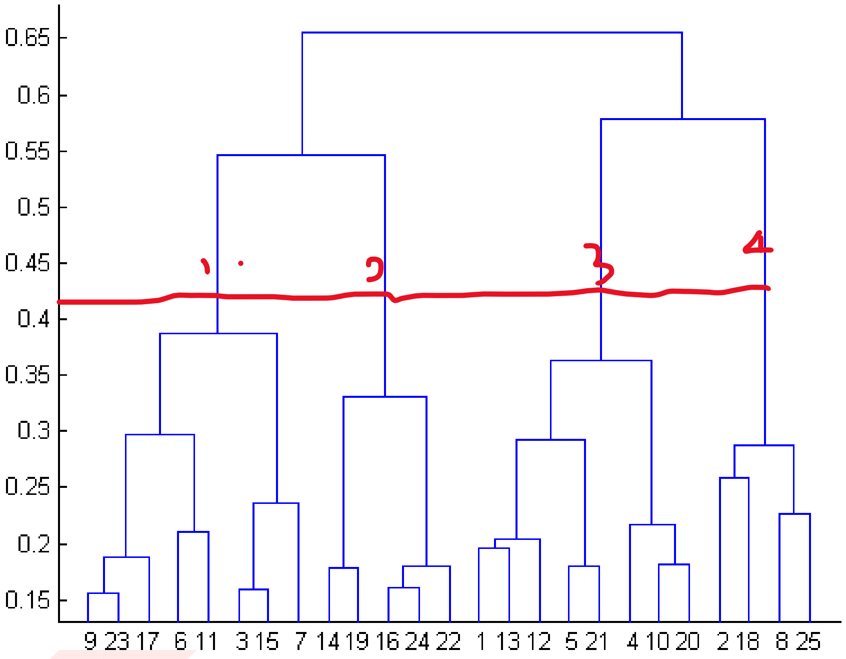
MACHINE LEARNING

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

Answer is (b).

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

   Answer is (d)

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

   Answer is (a)

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

Answer is (a).

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
   Answer is (b)

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
   Answer is (d)

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
   Answer is (a)

8. Clustering is a-
   a) Supervised learning
   b) Unsupervised learning
   c) Reinforcement learning
   d) None
   Answer is (b)

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
   Answer is (a)

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
   Answer is (a)

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
   Answer is (d)

12. For clustering, we do not require-
   a) Labeled data
   b) Unlabeled data
   c) Numerical data
   d) Categorical data
   Answer is (a)

MACHINE LEARNING

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

13. How is cluster analysis calculated?
14. How is cluster quality measured?
15. What is cluster analysis and its types?

Q. 13 Answer:

Depending on the nature of the dataset, cluster analysis can be calculated by employing clustering algorithms. The choice of algorithms is also dependent on whether the grouping of clusters is hierarchical or non-hierarchical.
For hierarchical clusters, calculation is based on the measurement of distance between datapoints and this mostly done by employing one of the following:

- Euclidean distance measure
- Squared Euclidean distance measure
- Manhattan distance measure
- Cosine distance measure

For non-hierarchical, K mean algorithm is mostly used to calculate clusters in the dataset.

Q. 14 Answer

By calculating the average silhouette coefficient value of all objects in the dataset.
The smaller the value the more qualitative the cluster is said to be, and the opposite is true.

Q 15 Answer

Cluster analysis is the methodical exploration of data in order to determine patten or groupings that form clusters within a given dataset. These patten can be used to predict users behaviors in given various stimulations.
Types of cluster analysis

a) Hierarchical cluster analysis
Similar and closest datapoints are identified and linked together to form a cluster. Then those clusters that are also similar and close by are linked together until the process cannot be expanded further. Agglomerative (bottom up) and divisive (top down) methods are the key features of hierarchical cluster analysis.

b) Centroid clustering
In this case clustering is built upon a key feature such as an estimated number of clusters in the dataset. K-Mean algorithm is used to determine clusters in the dataset.

c) Distribution clustering
This is based on statistical analysis of the dataset. Similarities and properties of objects such as correlation and dependence can be used to form distribution into clusters.

d) Density – based clustering.
Clusters are formed on basis of concentration. High density or concentrated distribution tend to determine the cluster of objects not withstanding the dissimilarity that may be apparent between them.

MACHINE LEARNING