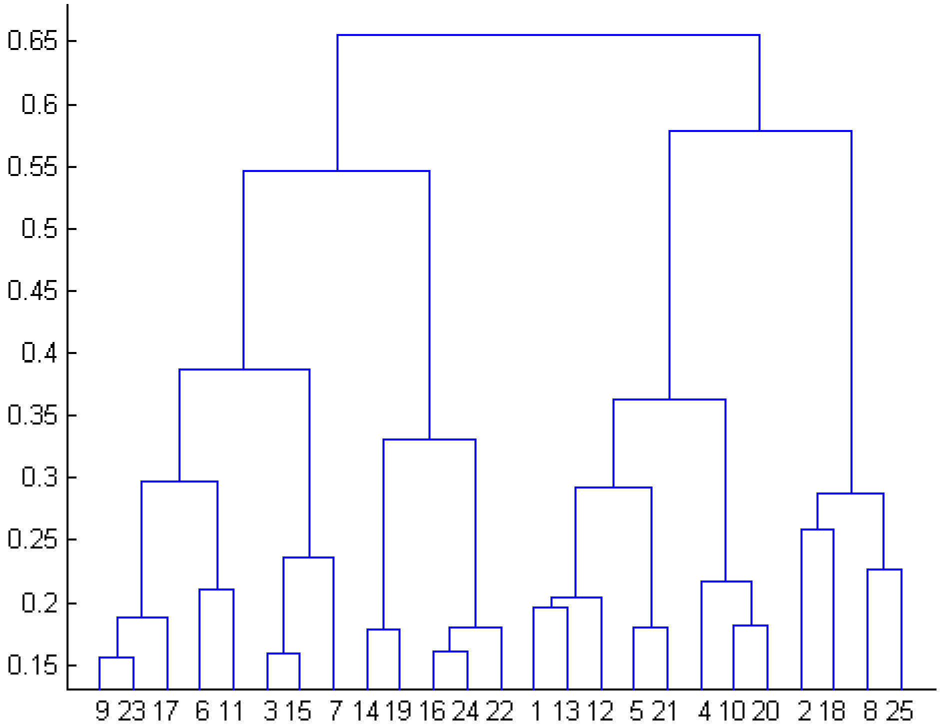
Machine Learning

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



A. 2

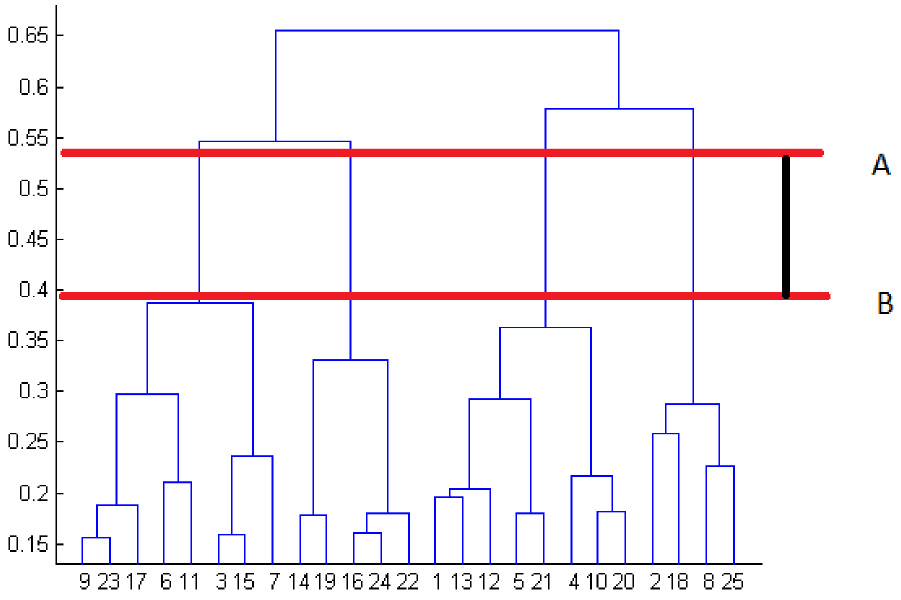
B. 4

C. 6

D. 8

Ans - 4 (B)

The decision of the number of clusters that can best depict different groups can be chosen by observing the dendrogram. The best choice of the no. of clusters is the number of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.



In the above example, the best choice of no. of clusters will be 4 as the red horizontal line in the dendrogram below covers maximum vertical distance AB.

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: D

K-Means clustering algorithm fails to give good results when the data contains outliers, the density spread of data points across the data space is different and the data points follow non-convex shapes.

3. The most important part 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

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

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

13. How is cluster analysis calculated?

The various types of calculating clustering analysis are:

1. Connectivity-based Clustering (Hierarchical clustering)
2. Centroids-based Clustering (Partitioning methods)
3. Distribution-based Clustering
4. Density-based Clustering (Model-based methods)
5. Fuzzy Clustering
6. Constraint-based (Supervised Clustering)

14. How is cluster quality measured?

Ans) To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.There are many ways to evaluate the performance of clustering models in machine learning. They are broadly divided into 3 categories-

1. Supervised techniques
2. Unsupervised techniques
3. Hybrid techniques

Supervised techniques are evaluated by comparing the value of evaluation metrics with some pre-defined ground rules and values.

For example- the Jaccard similarity index, Rand Index, Purity, etc.

Unsupervised techniques comprised of some evaluation metrics which cannot be compared with predefined values but they can be compared among different clustering models and thereby we can choose the best model.

For example - Silhouette measure, SSE

Hybrid techniques are nothing but a combination of supervised and unsupervised methods.

15. What is cluster analysis and its types?

Ans) Cluster analysis is a multivariate data mining technique whose goal is to group objects (eg., products, respondents, or other entities) 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.

There are a number of different methods to perform cluster analysis. Some of them are,

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