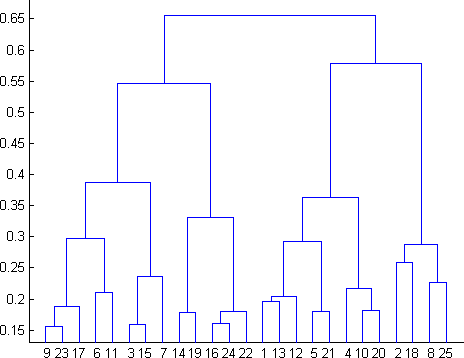
1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



Ans. 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:

Ans. 1, 2 and 4

3. The most important part of is selecting the variables on which clustering is based.

Ans. formulating the clustering problem

4. The most commonly used measure of similarity is the or its square.

Ans. 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.

Ans. K-means clustering

6. Which of the following is required by K-means clustering?

Ans. All answers are correct

7. The goal of clustering is to-

Ans. Divide the data points into groups

8. Clustering is a-

Ans. Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

Ans. K- Means clustering

10. Which version of the clustering algorithm is most sensitive to outliers?

Ans. All of above

11. Which of the following is a bad characteristic of a dataset for clustering analysis-

Ans. All of the above

12. For clustering, we do not require-

Ans. Labeled data

13. How is cluster analysis calculated?

Following are the methods of calculating cluster analysis :

Density-Based Methods: These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure), etc.

Hierarchical Based Methods: The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category

Agglomerative (bottom-up approach)

Divisive (top-down approach)

Partitioning Methods: These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example K-means, CLARANS (Clustering Large Applications based upon Randomized Search), etc.

Grid-based Methods: In this method, the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operations done on these grids are fast and independent of the number of data objects example STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest), etc.

14. How is cluster quality measured?

Ans. Cluster quality measured by two methods- extrinsic methods and intrinsic methods

If ground truth is available, it can be used by extrinsic methods, which compare the clustering against the group truth and measure. If the ground truth is unavailable, we can use intrinsic methods, which evaluate the goodness of a clustering by considering how well the clusters are separated. Ground truth can be considered as supervision in the form of “cluster labels.” Hence, extrinsic methods are also known as supervised methods, while intrinsic methods are unsupervised methods.

15. What is cluster analysis and its types?

Ans. The method of identifying similar instances and keeping them together is called clustering.

Clustering is an unsupervised approach which finds a structural pattern in a collection of objects which are “similar” amongst themselves and are “dissimilar” to the objects belonging to different cluster. Types of Cluster analysis are:

i. K-means clustering

ii. Hierarchical clustering

iii. DBSCAN (Density Based Spatial Clustering of Application with Noise)