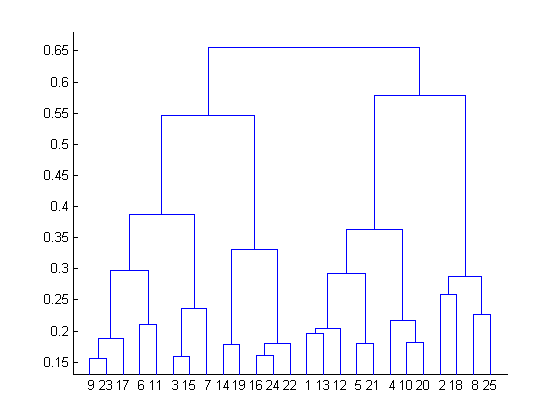
***MACHINE LEARNING***

1. In which of the following cases will K-Means clustering fail to give good results?



1. 2
2. 4
3. 6
4. 8

Answer – B). 4

The decision of the no. 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 no. of vertical lines in the dendrogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.

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). 1,2 & 4

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 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 – 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

Answer – 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

Answer – 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

Answer – D). All 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

Answer A). Divide the data points into groups

8. Clustering is a

a) Supervised learning

b) Unsupervised learning

c) Reinforcement learning

d) None

Answer 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

Answer – A). K-Means clustering

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 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

Answer D). all of the above

12. For clustering, we do not require

a) Labeled data

b) Unlabeled data

c) Numerical data

d) Categorical data

Answer A). Labelled data

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

13. How is cluster analysis calculated?

Answer - **Clustering** is an Unsupervised Learning algorithm that groups data samples into k clusters. The algorithm yields the k clusters based on k averages of points (i.e., centroids) that roam around the data set trying to center themselves — one in the middle of each cluster.

How is cluster analysis calculated:-

The hierarchical cluster analysis follows three basic steps:

* Calculate the distances.
* Link the clusters.
* Choose a solution by selecting the right number of clusters.

14. How is cluster quality measured?

Answer - Below measures can be used for cluster quality:

SSE (sum of the square error from the items of each cluster), Inter cluster distance, Intra cluster distance for each cluster, Maximum Radius, Average Radius. Roc curves can also solve our problem.

15. What is cluster analysis and its types?

Answer - Clustering or cluster analysis is a machine learning technique, which groups the unlabeled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points.

These types are Centroid Clustering, Density Clustering, Distribution Clustering and Connectivity Clustering.

**Centroid Clustering**

This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify. For example, if you’re a pet store owner you may choose to segment your customer list by people who bought dog and/or cat products.

**Density Clustering**

Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the denser the data points...the more related they are.

**Distribution Clustering**

Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid the algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions the algorithm optimizes the characteristics of the distributions to best represent the data.

**Connectivity Clustering**

Connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.