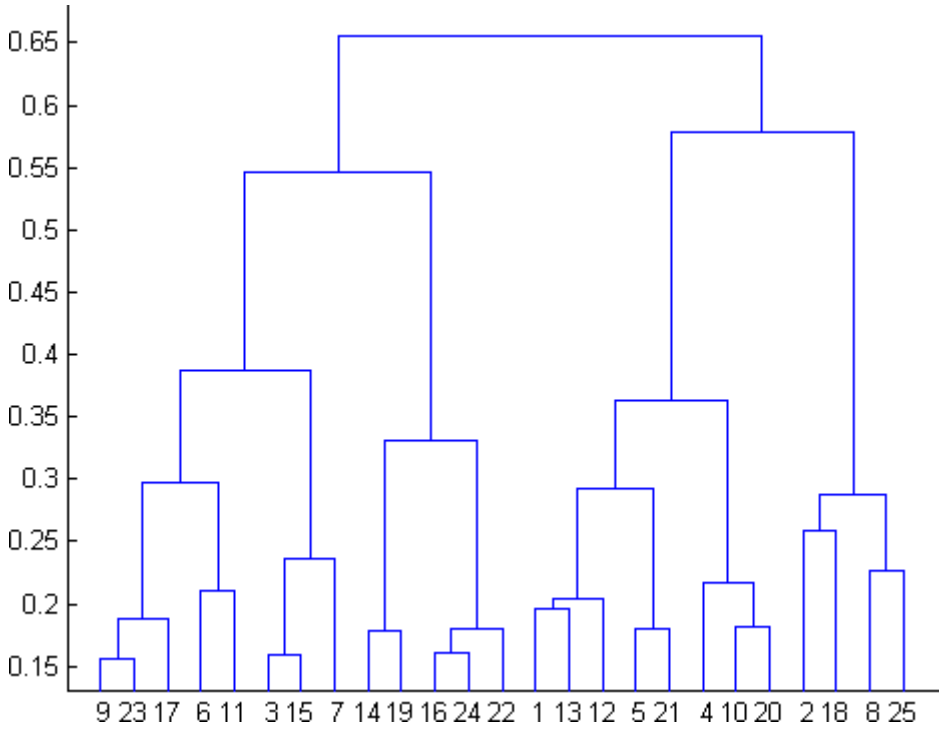
MACHINE LEARNING

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 : 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

Answer : 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

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 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

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 : 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

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(Labeled data)

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

13. How is cluster analysis calculated?

Clusters are those which makes a groups from entire data to classified or to describe the type of groups. Clusters contains a group of points which have some similarieties between them and aslo having dissimilaraties from rest of the clusters data points.
Calcualtion of clusters by using k means:
k represents a number of clusters to be made in data set.
Let’s take a uncluster data set from here we need to devide a data points into clusters so that to classified each clusters. If any new number comes into the picture that will goes into the respective cluster based on their similarities.
In step one took some data points from the data set and say them as centroids.
In step two all the data points are calculates a distance between all centroinds. By using Euclidean distance. Each centroids make a group of data points which is closer’s to them.
Then each centroids moves the position to the mean of respective group of data points.
Step two repeats until the centroids will not moves to next position.

Clustering can be done by using soo many methods some are :

- K-Means Clustering
- DBSCAN
- Hierarchical clustering
- Aglomerative clustering
- Devisive clustering

14. How is cluster quality measured?

We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available. Here, ground truth is the ideal clustering that is often built using human experts.
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.