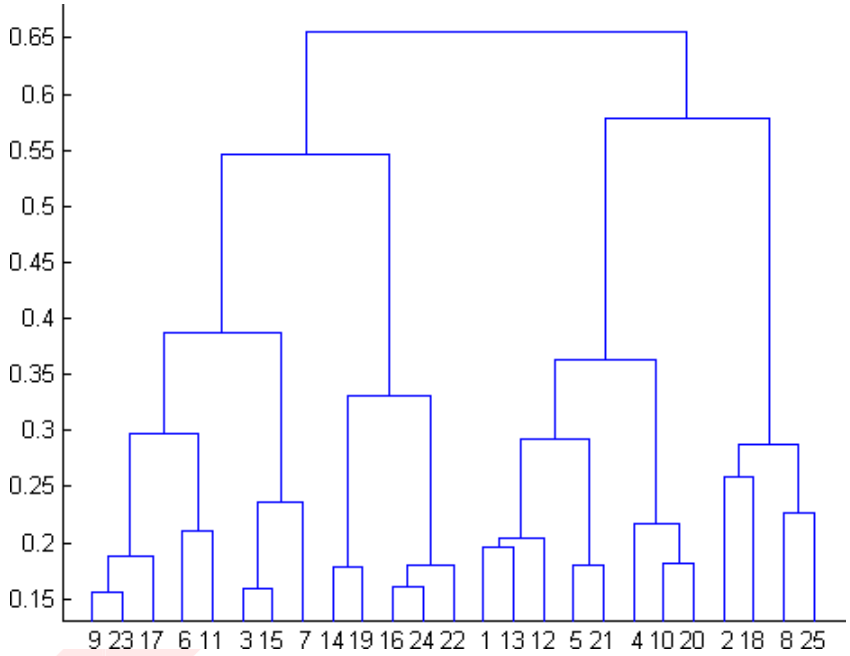
MACHINE LEARNING

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

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

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

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: - A

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

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

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

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

Answer: - 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: - D

MACHINE LEARNING

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

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

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

Answer: - A

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

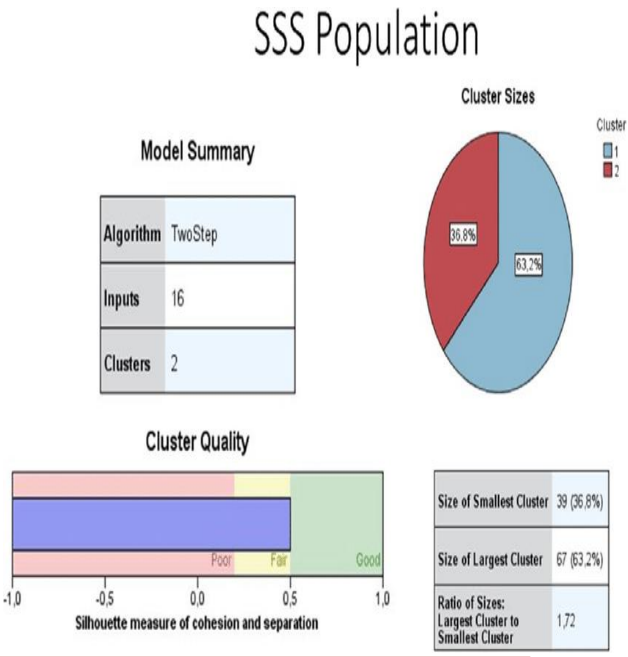
13. How is cluster analysis calculated?

Answer: - By varying k from 1 to 10 clusters. For each k, calculate the total within- cluster sum of square. Plot the curve of was according to the numbers of clusters k. The location of a bend in plot is generally considered as an indicator of the appropriate number of clusters.

14. How is cluster quality measured?

Answer: - 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.



MACHINE LEARNING

15. What is cluster analysis and its types?

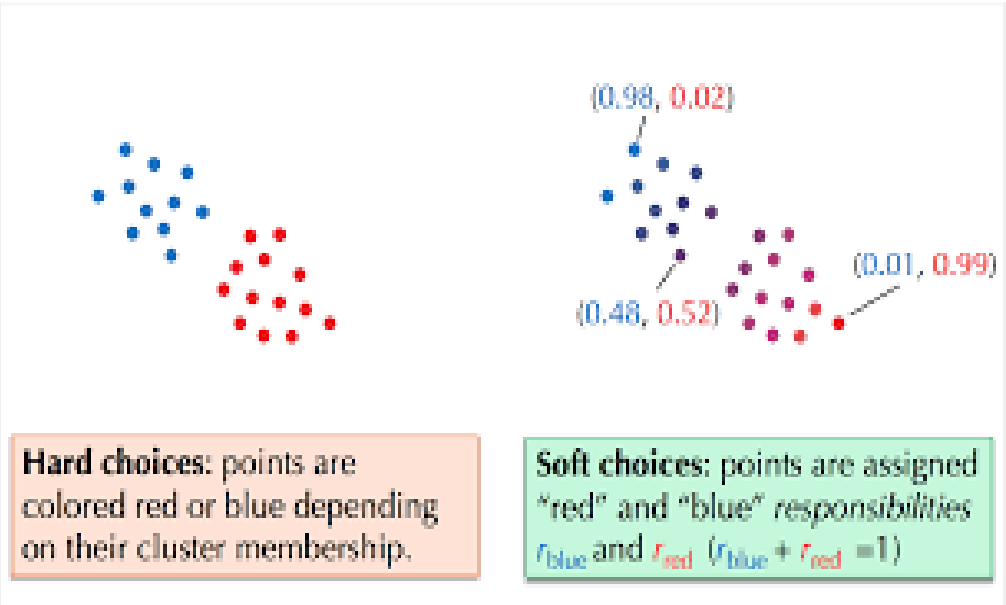
Answer: - Cluster is the method of identifying similar groups of data in a data set. Entities in each group are comparatively more similar to entities of that group than those of the other groups.

In other words, Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

TYPES OF CLUSTERING

Hard Clustering: - In Hard clustering, each data point either belongs to a cluster completely or not.

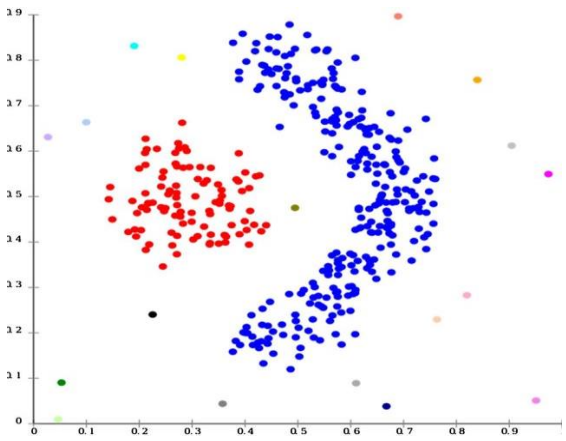
Soft Clustering: - In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned.



MACHINE LEARNING

Types of clustering Algorithum

- **Connectivity Models: -** These Models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away.



- **Centroid Models: -** These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category**.**

- **Distribution Models: -** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution. For example: Normal, Gaussian

- **Density Models: -** These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster.