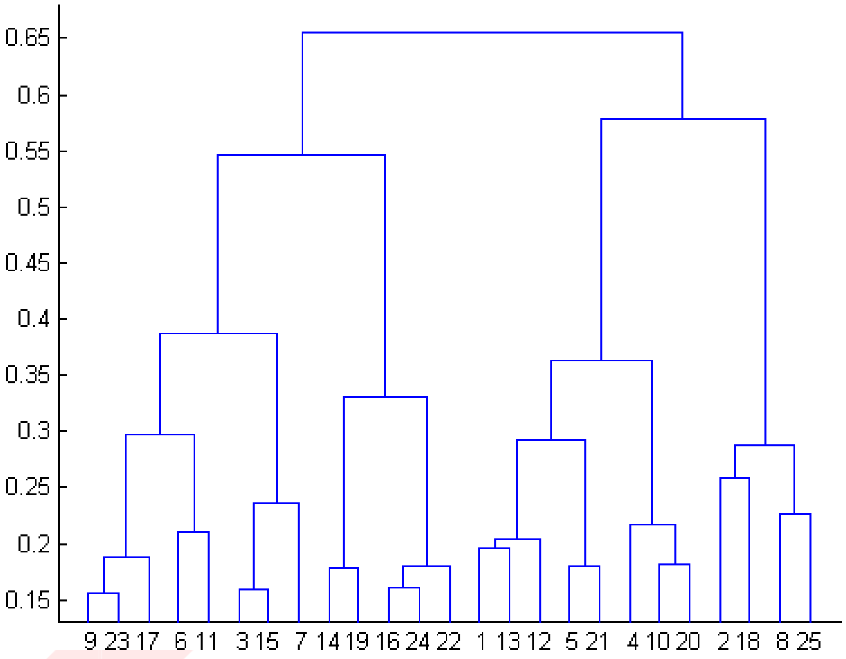
MACHINE LEARNING

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

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

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

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

MACHINE LEARNING

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

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

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

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

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

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

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

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

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

13. How is cluster analysis calculated?
14. How is cluster quality measured?
15. What is cluster analysis and its types?

MACHINE LEARNING

13. How is cluster analysis calculated?

Clustering or cluster analysis is the method of grouping the entities based on similarities. Defined as an unsupervised learning problem that aims to make training data with a given set of inputs but without any target values, It is the process of finding similar structures in a set of unlabeled data to make it more understandable and manipulative.

In Clustering, the machine learns the attributes and trends by itself without any provided input-output mapping. The clustering algorithms exact patterns and inferences from the type of data objects and then make discrete classes of clustering them suitably.

Types of Clustering Algorithms:

1. K-Means Clustering
2. Mini Batch K-Means clustering algorithm
3. Mean shift
4. Divisive Hierarchical Clustering
5. Hierarchical Agglomerative Clustering
6. Gaussian Mixture Model
7. DBSCAN
8. OPTICS
9. BIRCH Algorithm

16. How is cluster quality measured?

Clustering Approaches:

1. Partitioning approach: The partitioning approach constructs various partitions and then evaluates them by some criterion, e.g., minimizing the sum of square errors. It adopts exclusive cluster separation(each object belongs to exactly one group) and uses iterative relocation techniques to improve the partitioning by moving objects from one group to another. It uses a greedy approach and approach at a local optimum. It finds clusters with spherical shapes in small to medium size databases.

2. Connectivity Based:

- K-Means
- K-Medoids
- CLARINS

3. Density-Based Approach: This approach is based on Connectivity and density functions. It divides the set of objective into multiple exclusive clusters or a hierarchy of clusters. Density-based method:
   - DBSACN
   - OPTICS
4. Grid-based approach: This approach quantizes objects into a finite number of cells that form a grid structures. Fast processing time and independent of a number of data objects. Grid-based Clustering method is the efficient approach for spatial data mining problems.

Grid-based approach methods:

- STING
- WaveCluster
- CLIQUE

5. Hierarchical Approach: This creates a hierarchical decomposition of the data objects by using some measures. Hierarchical approach methods:

MACHINE LEARNING

- Diana
- Agnes
- BIRCH
- CAMELEON

Measures for Quality of Clustering:

1. Dissimilarity/Similarity metric: The similarity between the clusters can be expressed in terms of a distance function, which is represented by d(i, j). Distance functions are different for various data types and data variables. Distance function measure is different for continuous-valued variables, categorical variables, and vector variables. Distance function can be expressed as Euclidean distance, Mahalanobis distance, and Cosine distance for different types of data.

2. Cluster completeness: Cluster completeness is the essential parameter for good clustering, if any two data objects are having similar characteristics then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category.

Let us consider the clustering C1, which contains the sub-clusters s1 and s2, where the members of the s1 and s2 cluster belong to the same category according to ground truth. Let us consider another clustering C2 which is identical to C1 but now s1 and s2 are merged into one cluster. Then, we define the clustering quality measure, Q, and according to cluster completeness C2, will have more cluster quality compared to the C1 that is, Q(C2, Cg) > Q(C1, Cg).

3. Ragbag: In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then the quality of those cluster categories is measured by the Rag Bag method. According to the rag bag method, we should put the heterogeneous object into a rag bag category.

4. Small cluster preservation: If a small category of clustering is further split into small pieces, then those small pieces of cluster become noise to the entire clustering and thus it becomes difficult to identify that small category from the clustering. The small cluster preservation criterion states that are splitting a small category into pieces is not advisable and it further decreases the quality of clusters as the pieces of clusters are distinctive. Suppose clustering C1 has split into three clusters, C11 = {d1, . . . , dn}, C12 = {dn+1}, and C13 = {dn+2}.

Let clustering C2 also split into three clusters, namely C1 = {d1, . . . , dn−1}, C2 = {dn}, and C3 = {dn+1,dn+2}. As C1 splits the small category of objects and C2 splits the big category which is preferred according to the rule mentioned above the clustering quality measure Q should give a higher score to C2, that is, Q(C2, Cg) > Q(C1, Cg).

17. What is cluster analysis and its types?

Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg., products, respondents, or other entities) based on a set of user selected characteristics or attributes

Types of Clustering Methods:

1. Connectivity-based clustering (Hierarchical Clustering)
2. Centroids-based clustering (Partitioning Methods)
3. Distribution-based clustering
4. Density based clustering (Model-based methods)

MACHINE LEARNING

5. Fuzzy Clustering
6. Constraint-based (Supervised Clustering)

1. Connectivity-based Clustering:

Hierarchical clustering also known as connectivity-based clustering, is based on the principal that every object is connected to its neighbors depending on their proximity distance. The clusters are represented in extensive hierarchical structures separated by a maximum distance required to connect the cluster parts.

2. Centroid – based or partition clustering:

Centroid-based clustering is the easiest of al the clustering types in data mining. It works on the closeness of the data points to the chosen central value. The datasets are divided into a given no of clusters, and a vector of values references every cluster. The input data variable is compared to the vector value and enters the cluster with minimal difference.

Pre-defining the no of clusters at the initial stage is the most crucial yet most complicated stage for the clustering approach. Despite the drawback, it is a vastly used clustering approach for surfacing and optimizing large datasets. The K-Means algorithm lies in this category.

3. Density based clustering (Model-Based Method):

This cluster is formed vary in arbitrary shapes and sizes and contain a maximum degree of homogeneity due to similar density. This clustering approach includes the noise and outliers in the dataset effectively.

When performing most of the clustering, we take two major assumptions, the data is devoid of any noise and the shape of the cluster so formed is purely geometrical.

4. Distribution-Based Clustering:

Distribution-based clustering has a vivid advantage over the proximity and centroid-based clustering methods in terms of flexibility, correctness, and shape of the clusters formed. The major problem however is that these clustering methods work well only with synthetic or simulated data or with data where most of the data points most certainly belong to a predefined distribution, if not, the results will overfit.

5. Fuzzy Clustering:

Fuzzy Clustering generalizes the partition-based clustering method by allowing a data object to be a part of more than one cluster. The process uses a weighted centroid based on the spatial probabilities.

The algorithm works by assigning membership values to all the data points linked to each cluster center. It is computed from a distance between the cluster center and the data point. If the membership value of the object is closer to the cluster center, it has a high probability of being in the specific cluster.

6. Constraint-Based:

The clustering process, in general is based on the approach that the data can be divided into an optimal number of “Unknown” groups. The underlying stages of all the clustering algorithms are to find those hidden patterns and similarities without intervention or predefined conditions. However in certain business scenarios, we might be required to partition the data based on certain constraints. Here is where a supervised version of clustering machine learning techniques come into play.