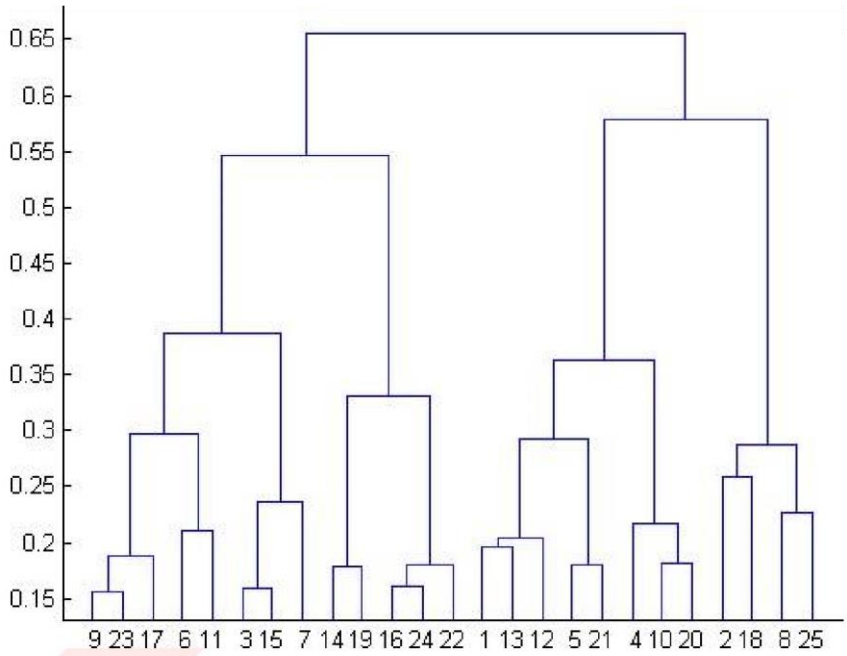
MACHINE LEARNING

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

FLIP ROBO

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 : a) K-Means clustering.

FLIP ROBO

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?

Most of the clustering algorithms use distances between the datapoints, density of clusters, centroids of clusters etc. The parameters like the distance function, cosine similarity, number of clusters depends on the dataset itself and the usage. The most common calculation method is to Calculate the distances between the datapoints within the cluster and between the clusters, then compare what the inter-cluster distances and the intra-cluster distances are indicating about the clustering.

14. cluster quality measured?

There are a few methods to measure the quality of a clustering technique. These can be categorized into 'Extrinsic' methods and 'Intrinsic' methods.

Intrinsic methods:

When there are no cluster labels, we can evaluate the goodness of the cluster by considering how well the clusters are separated. This method is considered as an Unsupervised method. There are many intrinsic methods. Below are a few of them:

1. **Davis-Bouldin Index:** The clustering is evaluated based on the intra-cluster and the inter-cluster distances. For a good cluster, the intra-cluster distance should be low and the inter-cluster distance should be high. For an algorithm that produces such a clustering the Davis-Bouldin index will be low because of the way it is calculated. Hence, we can say that a clustering algorithm that produces a collection of clusters with the smallest Davis-Bouldin index is the best algorithm based on these criteria.
2. **Dunn Index:** The Dunn index helps in identifying which algorithm is able to produce dense and well separated clusters. This method gives a value which is a ratio between the least inter-cluster distance and the highest intra-cluster distance. The value increases if the denominator (highest intra-cluster distance in this case) decreases. So, the algorithms that produce clusters with high Dunn index are more desirable.
3. **Silhouette coefficient:** Silhouette method helps in finding how similar a datapoint or an object is to its own cluster and how different it is to other clusters. The silhouette ranges from −1 to +1. A high Silhouette coefficient indicates that the object is very similar to its own cluster than to the other clusters. If most of the datapoints or objects have high values, then the algorithm has done a good job in clustering the data points and low values indicates a poor clustering. The similarity can be found using the distance between datapoints. Any distance metric such as Euclidean distance or Manhattan distance can be used to calculate the silhouette.

Extrinsic methods:

When the cluster labels are already available, those labels can be used to compare the clusters produced by the clustering algorithm. This method is also known as supervised method since the ground truth can be considered as a supervision in the form of "cluster labels". These types of evaluation methods measure how close the clustering is to the predetermined benchmark classes / cluster labels.

Some of the external evaluation/extrinsic methods are:

1. **Purity**: Purity can be seen as the opposite of entropy. We are trying to measure the extent to which the clusters contain only a single class. The number of datapoints that belongs to the most common class within each cluster is taken and are summed over all the clusters. Then it is divided by the total number of data points. This gives us a measure or how pure the clustering is. That is how well the datapoints are clustered such that each cluster have only one class. This measure doesn't penalize having many clusters, and more clusters will make it easier to produce a high purity. A purity score of 1 is always possible by putting each data point in its own cluster. Purity doesn't work well for imbalanced data. If a dataset of size 1000 consists of two classes, one containing 900 points and the other containing 100 points, then every possible partition will have a purity of at least 90 %.
2. **Rand Index:** The Rand index measures how similar the clusters produced by the clustering algorithm are to the predefined cluster labels. The Rand index basically computes the truth rate in the clustering by considering the number of True Positives, number of True Negatives, number of False Positives and the number of False Negatives. The True Positives and True negatives are decided based on the predetermined cluster labels. One issue with the Rand index is that false positives and false negatives are equally weighted. This may be an undesirable characteristic for some clustering applications. The F-measure addresses this concern, as does the chance-corrected adjusted Rand index.
3. **F-measure:** The F-measure is calculated from Precision and Recall. The Precision is the Number of True Positives out of all the predicted Positives. The Recall is the number of True Positives out of all the actual Positives. The F_1 score is the harmonic mean of the Precision and the Recall. The more generic F_β score applies additional weights, valuing one of precision or recall more than the other.
4. **Confusion Matrix:** The confusion matrix shows a visualization table of the True Positives, False Positives, True Negatives and False Negatives. This gives a measure of Type I and Type II errors produced by the clustering algorithm.

15. What is cluster analysis and its types?

    Cluster analysis is an exploratory analysis in which we try to group the datapoints to create clusters such that the datapoints within a cluster is similar to each other than to the datapoints in the neighbouring clusters. There are many applications for cluster analysis like diagnostic analysis in medicines, segment analysis of customers, grouping or taxonomy analysis etc.

    There are many types of clustering methods available. Some of the mostly used are the follows:

    1. Density based clustering – DBSCAN,OPTICS, HDBSCAN
    2. Hierarchy based clustering – Agglomerative, Divisive
    3. Fuzzy clustering – c-means
    4. Partitioning clustering – k-means, k-medoids
    5. Gird based clustering