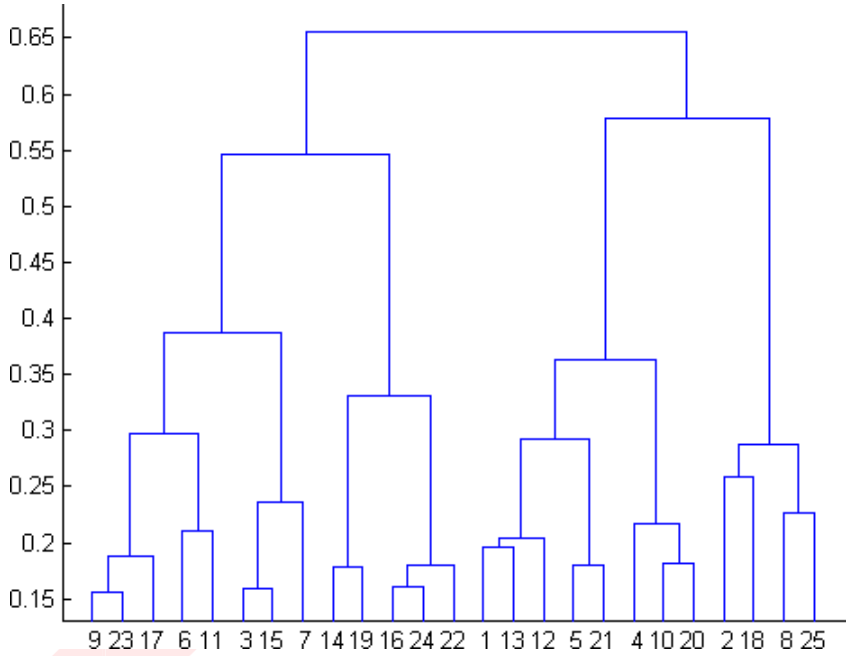
MACHINE LEARNING

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

Ans : 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

   Ans : 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

   Ans : formulating the clustering problem

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

   Ans : 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

   Ans : 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

   Ans : 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

   Ans : All of the above

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

   Ans : Supervised 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

   Ans : K- Means clustering

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

    Ans : 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

    Ans : All of the above

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

    Ans : Labeled data

MACHINE LEARNING

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

13. How is cluster analysis calculated?

Ans :

Clustering is the process of dividing uncategorized data into similar groups or clusters. This process ensures that similar data points are identified and grouped. Clustering algorithms is key in the processing of data and identification of groups.

The main types of clustering in unsupervised machine learning include K Means, hierarchical clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

For some clustering algorithms, such as K-means, one needs to know how many clusters there are beforehand. If the number of clusters is incorrectly specified, the results are not very informative. Certainly, domain knowledge of the data set may help determine the number of clusters. However, this assumes that you know the target and this is not true in unsupervised learning. We need a method that informs us about the number of clusters without relying on a target variable.

One possible solution in determining the correct number of clusters is a brute-force approach. We try applying a clustering algorithm with different numbers of clusters. We use popular metrics to assess cluster quality. They are

- **The elbow method** (used to decide how many clusters are we are going to create)
- The optimization of the **silhouette coefficient** (Used to evaluate how good it is)

The goal of clustering is to group data points in clusters so that,

1. points within a cluster are as similar as possible
2. points belonging to different clusters are as distinct as possible. This means that, in ideal clustering, the within-cluster variation is small whereas the between-cluster variation is large. Consequently a good clustering quality metric should be able to summarize it quantitatively.

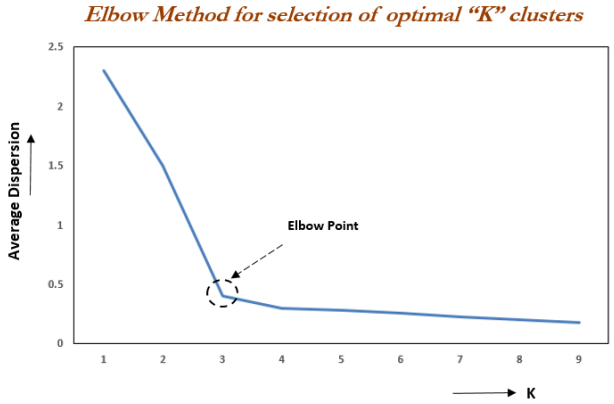
One such quality metric is inertia. **This is calculated as the sum of squared distances between data points and the centers of the clusters they belong to. Inertia quantifies the within-cluster variation**. The inertia is calculated by formula

$$\sum_{i=1}^{N} (x_i - C_k)^2$$

N is the number of samples within the data set, C is the center of a cluster. So the Inertia simply computes the squared distance of each sample in a cluster to its cluster center and sums them up. This process is done for each cluster and all samples within that data set. The smaller the Inertia value, the more coherent are the different clusters. When as many clusters are added as there are samples in the data set, then the Inertia value would be zero. So how to find the optimal number of clusters using the Inertia value? For this, the so called **Elbow-Method** can be used.

MACHINE LEARNING

The elbow method is used to determine the optimal number of clusters in k-means clustering. The elbow method plots the value of the cost function produced by different values of k. As you know, if k increases, average distortion will decrease, each cluster will have fewer constituent instances, and the instances will be closer to their respective centroids. However, the improvements in average distortion will decline as k increases. The value of k at which improvement in distortion declines the most is called the elbow, at which we should stop dividing the data into further clusters.



It consists in the interpretation of a line plot with an elbow shape. The number of clusters is were the elbow bends. The x axis of the plot is the number of clusters and the y axis is the Within Clusters Sum of Squares (WCSS) for each number of clusters.

When using K-means Clustering, you need to pre-determine the number of clusters. As we have seen when using a method to choose our k number of clusters, the result is only a suggestion and can be impacted by the amount of variance in data. It is important to conduct an in-depth analysis and generate more than one model with different k.

The distance of the data point from its nearest centroid can also be calculated to minimize the distances to arrive at the refined centroid. The Euclidean distance between two data points is measured and The measure of quality of clustering uses the SSE technique.

where distance calculates the Euclidean distance between the centroid of the cluster and the data points in the cluster. The summation of such distances over all the 'K' clusters gives the total sum of squared error. As we understand, the lower the SSE for a clustering solution, the better is the representative position of the centroid. After the centroids are repositioned, the data points nearest to the centroids are assigned to form the refined clusters. It is observed that the centroid that minimizes the SSE of the cluster is its mean. One limitation of the squared error method is that in the case of presence of outliers in the data set, the squared error can distort the mean value of the clusters.

on the basis of the proximity of the data points in this data set to the centroids, we assume the number of clusters in a dataset. next step is to calculate the SSE of this clustering and update the position of the centroids. Our aim is to minimize the homogeneity within the clusters and maximize the heterogeneity among the different clusters.

The *k*-means algorithm continues with the update of the centroid according to the new cluster and reassignment of the points, until no more data points are changed due to the centroid shift. *k*-means algorithm multiple times with different cluster centres to identify the optimal clusters.

MACHINE LEARNING

14. How is cluster quality measured?

Ans :

It can be said that a clustering algorithm is successful if the clusters identified using the algorithm is able to achieve the right results in the overall problem domain. cluster quality evaluation is done by below methods.

1. Internal evaluation

In this approach, the cluster is assessed based on the underlying data that was clustered. The internal evaluation methods generally measure cluster quality based on homogeneity of data belonging to the same cluster and heterogeneity of data belonging to different clusters. The homogeneity/heterogeneity is decided by some similarity measure.

For example, silhouette coefficient, which is one of the most popular internal evaluation methods, uses distance (Euclidean distances most commonly used) between data elements as a similarity measure. The value of silhouette width ranges between –1 and +1, with a high value indicating high intra-cluster homogeneity and inter-cluster heterogeneity.

For a data set clustered into K clusters, silhouette width is calculated as below

$$\text{Silhouette width} = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

a(i) is the average distance between the i th data instance and all other data instances belonging to the same cluster and b(i) is the lowest average distance between the i-the data instance and data instances of all other clusters.

1. External Evaluation

In this approach, class label is known for the data set subjected to clustering. However, quite obviously, the known class labels are not a part of the data used in clustering. The cluster algorithm is assessed based on how close the results are compared to those known class labels. For example, **purity** is one of the most popular measures of cluster algorithms – evaluates the extent to which clusters contain a single class.

For a data set having 'n' data instances and 'c' known class labels which generates 'k' clusters, purity is measured as

$$\text{Purity} = \frac{1}{n}\sum_{k}\max(k \cap c)$$

Purity is quite simple to calculate. We assign a label to each cluster based on the most frequent class in it. Then the purity becomes the number of correctly matched class and cluster labels divided by the number of total data points.

MACHINE LEARNING

15. What is cluster analysis and its types?

Ans :

Cluster analysis is a multivariate data mining technique whose goal is to groups objects. This process ensures that similar data points are identified and grouped. Clustering algorithms is key in the processing of data and identification of groups based on a set of user selected characteristics or attributes.

Types of Cluster Analysis

The clustering algorithm needs to be chosen experimentally unless there is a mathematical reason to choose one cluster method over another. It should be noted that an algorithm that works on a particular set of data will not work on another set of data. There are a number of different methods to perform cluster analysis. Some of them are,

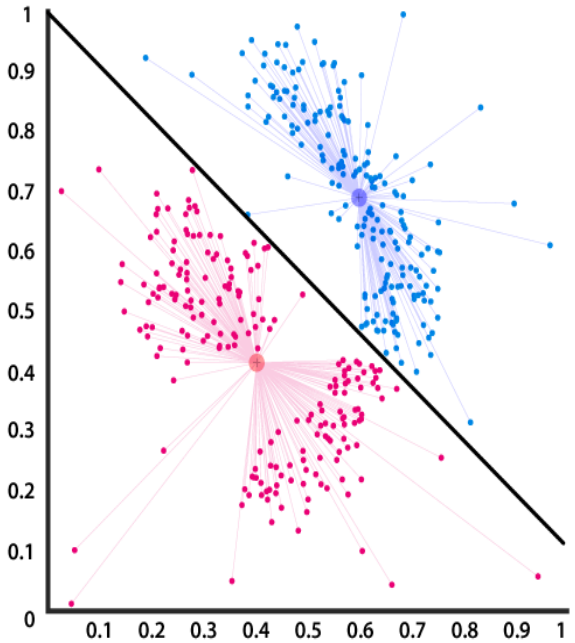
1. Hierarchical Cluster Analysis

In this method, first, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as Agglomerative method. Agglomerative clustering starts with single objects and starts grouping them into clusters.

The divisive method is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.

2. K-mean or Centroid-based Clustering

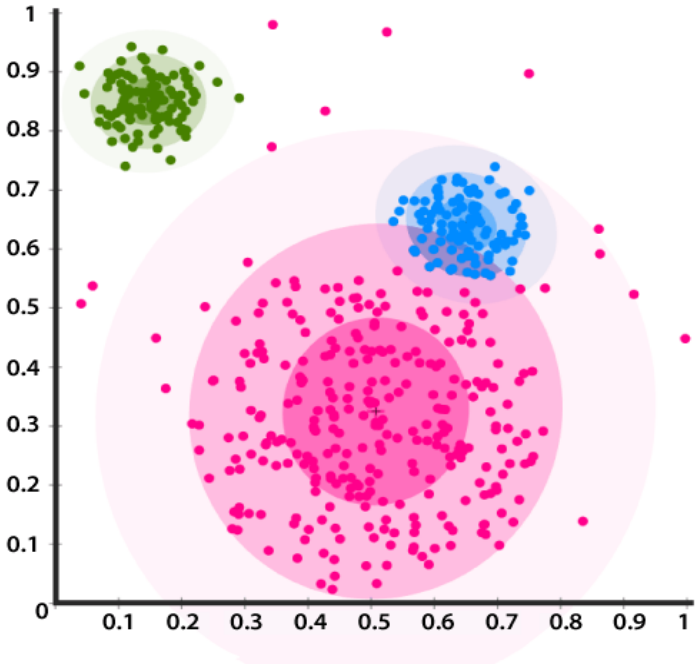
In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centers and objects are assigned to the nearest cluster centres.



MACHINE LEARNING

3. Distribution-based Clustering

It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster. This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.



4. Density-based Clustering

In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters. The objects in these sparse points are usually noise and border points in the graph. The most popular method in this type of clustering is DBSCAN.