MACHINE LEARNING – WORKSHEET 1 (CLUSTERING)

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:

Ans=b) 4

1. In which of the following cases will K-Means clustering fail to give good results?

Ans=d) 1, 2 and 4

1. The most important part of \_\_\_\_\_ is selecting the variables on which clustering is based.

Ans=d) formulating the clustering problem

1. The most commonly used measure of similarity is the \_\_\_\_\_ or its square.

Ans=a) Euclidean distance

1. . \_\_\_\_\_ 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.

Ans =b) Divisive clustering

1. Which of the following is required by K-means clustering?

Ans=d) all answers are correct

1. The goal of clustering is to

Ans=a) Divide the data points into groups

1. Clustering is a\_

Ans=a) Supervised learning

1. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

Ans=a) K- Means clustering

1. Which version of the clustering algorithm is most sensitive to outliers?

Ans=a) K-means clustering algorithm

1. Which of the following is a bad characteristic of a dataset for clustering analysis\_

Ans=d) All of the above

1. For clustering, we do not require\_

Ans=b) Unlabeled data

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

1. How is cluster analysis calculated?

Ans=Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

The hierarchical cluster analysis follows three basic steps:

calculate the distances,

 link the clusters,

choose a solution by selecting the right number of clusters.

1. How is cluster quality measured?

Ans=To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

Cluster cohesion: Measures the closeness of the objects within the same cluster. A “lower within-cluster” variation indicates good compactness or good clustering. The separation method is implied to measure how well a cluster is separated from other clusters.

 A good clustering method will produce high quality clusters in which: – the intra-class (that is, intra intra-cluster) similarity is high. ... The quality of a clustering result also depends on both the similarity measure used by the method and its implementation.

1. What is cluster analysis and its types?

Ans=Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities.

Centroid Clustering

This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify. For example, if you’re a pet store owner you may choose to segment your customer list by people who bought dog and/or cat products.

Density Clustering

Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points...the more related they are. To determine this, the algorithm will select a random point then start measuring the distance between each point around it. For most density algorithms a predetermined distance between data points is selected to benchmark how closely points need to be to one another to be considered related.. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.

Distribution Clustering

Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid. The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data.

Connectivity Clustering

Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.