Worksheet-1

Machine learning

Question 1- What is the most appropriate no. of clusters for the data points represented by the following dendrogram:

Answer - b) 4

Question 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

Question 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

Question 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

Question 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

Question 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

Question 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

Question 8- Clustering is a

a) Supervised learning b) Unsupervised learning

c) Reinforcement learning d) None

Answer - b) Unsupervised learning

Question 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

Question 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

Question 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

Question 12- . For clustering, we do not require

a) Labeled data b) Unlabeled data

c) Numerical data d) Categorical data

Answer - a) Labeled data

Question 13- How is cluster analysis calculated?

Answer - Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. There are no criteria for good clustering. It depends on the user,

what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

Clustering Methods :

Density-Based Methods: These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure), etc.

Hierarchical Based Methods: The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category

Agglomerative (bottom-up approach)

Divisive (top-down approach)

Question 14 - How is cluster quality measured?

Answer - If all the data objects in the cluster are highly similar then the cluster has high quality. We can measure the quality of Clustering by using the Dissimilarity/Similarity metric in most situations. But there are some other methods to measure the Qualities of Good Clustering if the clusters are alike.

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.

Let us consider a clustering C1 and a cluster $C \in C1$ so that all objects in C belong to the same category of cluster C1 except the object o according to ground truth. Consider a clustering C2 which is identical to C1 except that o is assigned to a cluster D which holds the objects of different categories. According to the ground truth, this situation is noisy and the quality of clustering is measured using the rag bag criteria. we define the clustering quality measure, Q, and according to rag bag method criteria C2, will have more cluster quality compared to the C1 that is, $Q(C2, Cg) > Q(C1, Cg)$.

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, $C11 = \{d1, \ldots, dn\}$, $C12 = \{dn+1\}$, and $C13 = \{dn+2\}$.

Let clustering C2 also split into three clusters, namely $C1 = \{d1, \ldots, dn-1\}$, $C2 = \{dn\}$, 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, $Q(C2, Cg) > Q(C1, Cg)$.

Question 15- What is cluster analysis and its types?

Answer - 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,

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.

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.

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.

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.