Date- 01-03-2023

MACHINE LEARNING

ASSIGNMENT – 1

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

a) 2

2. In which of the following cases will K-Means clustering fail to give good results?
d) 1, 2 and 4

3. The most important part of is selecting the variables on which clustering is based.
d) formulating the clustering problem

4. The most commonly used measure of similarity is the or its square.
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.
b) Divisive clustering

6. Which of the following is required by K-means clustering?
d) All answers are correct

7. The goal of clustering is to-
d) All of the above

8. Clustering is a-
b) Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?
a) K- Means clustering

10. Which version of the clustering algorithm is most sensitive to outliers?
a) K-means clustering algorithm

11. Which of the following is a bad characteristic of a dataset for clustering analysis-
d) All of the above

12. For clustering, we do not require-
a) Labeled data

13. How is cluster analysis calculated?
Ans. KMeans is a centroid/distance-based algorithm where the distance between each point is calculated and then assigns it to a cluster. The distance can be calculated using

1. Euclidean Distance
2. City block/Manhattan Distance

14. How is cluster quality measured?
Ans. 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.

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.

15. What is cluster analysis and its types?

Ans. 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. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.

Hierarchical Clustering

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.