MACHINE LEARNING – ASSIGNMENT 1 (SATENDER SINGH) DS2302

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

Answer : b

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

Answer : d

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

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

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

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

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

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

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 : d

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

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

12) For clustering, we do not require-
a) Labelled data
b) Unlabelled data
c) Numerical data
d) Categorical data
Answer : a

13) How is cluster analysis calculated?
Cluster analysis groups the data which was not labeled
Step 1: Choose the number of clusters k
Step 2: Make an initial selection of k centroids
Step 3: Assign each data element to its nearest centroid (in this way k clusters are formed one for each centroid, where each cluster consists of all the data elements assigned to that centroid)
Step 4: For each cluster make a new selection of its centroid
Step 5: Go back to step 3, repeating the process until the centroids don't change (or some other convergence criterion is met)

14) How is cluster quality measured?
Answer: To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

15) What is cluster analysis and its types?

Answer: 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.

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,

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.