Machine Learning – Worksheet-1

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

Answer : 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) 1and2
   b) 2and3
   c) 2and4
   d) 1,2 and4

Answer : Option –d -1,2,4

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

Answer: Formulating the Clustering Problem

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 : 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

Answer : Division 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

Answer : 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

Answer : Divide the data points into groups.

8. **Clustering is a-**

    a) Supervised learning
    b) Unsupervised learning
    c) Reinforcement leaming
    d) None

Answer: Unsupervised 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

Answer : K-Means Clustering

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: 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

Answer: All of the above

12. For clustering, we do not require-

a) Labeled data
b) Unlabeled data
c) Numerical data
d) Categorical data

Answer: Labeled data

13. How is cluster analysis calculated?

Answer : The hierarchical cluster analysis follows three basic steps:

1) calculate the distances,

2) link the clusters, and

3) choose a solution by selecting the right number of clusters.

First, we have to select the variables upon which we base our clusters. In the dialog window we add the math, reading, and writing tests to the list of variables. Since we want to cluster cases we leave the rest of the tick marks on the default.

In the dialog box, Statistics we can specify whether we want to output the proximity matrix (these are the distances calculated in the first step of the analysis) and the predicted cluster membership of the cases in our observations. Again, we leave all settings on default.

In the dialog box, plots we should add the Dendogram. The Densogram will graphically show how the clusters are merged and allows us to identify what the appropriate number of clusters is.

The dialog box method allows us to specify the distance measure and the clustering method. First, we need to define the correct distance measure. SPSS offers three large blocks of distance measures for interval (scale), counts (ordinal), and binary (nominal) data.

For interval data, the most common is Square Euclidean Distance. It is based on the Euclidian Distance between two observations, which is the square root of the sum of squared distances. Since the Euclidian Distance is squared, it increases the importance of large distances, while weakening the importance of small distances.

If we have ordinal data (counts) we can select between Chi-Square or a standardized Chi-Square called Phi Square. For binary data, the Squared Euclidean Distance is commonly used.

Next, we have to choose the Cluster Method. Typically, choices are between-groups linkage (distance between clusters is the average distance of all data points within these clusters), nearest neighbor (single linkage: distance between clusters is the smallest distance between two data points), furthest neighbor (complete linkage: distance is the largest distance between two data points), and Ward's method (distance is the distance of all clusters to the grand average of the sample). Single linkage works best with long chains of clusters, while complete linkage works best with dense blobs of clusters. Between-groups linkage works with both cluster types. It is recommended is to use single linkage first. Although single linkage tends to create chains of clusters, it helps in identifying outliers. After excluding these outliers, we can

move onto Ward's method. Ward's method uses the F- value (like in ANOVA) to maximize the significance of differences between clusters.

14.How is cluster quality measured?

Answer : The quality of a clustering result **depends on both the similarity measure used by the method and its implementation**. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns. There is a separate "quality" function that measures the "goodness" of a cluster.

Elbow method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. We pick k at the spot where SSE starts to flatten out and forming an elbow.

There are two important elements in improving the quality of clustering: **improving the weights of the features in a document vector and creating a more appropriate distance measure**.

There are majorly two types of measures to assess the clustering performance. (i) Extrinsic Measures which require ground truth labels. Examples are **Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure**

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:

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.