**MACHINE LEARNING – WORKSHEET (CLUSTERING**)

1. 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. 1 and 2

b. 2 and 3

c. 2 and 4

**d. 1, 2 and 4**

e. 1, 2, 3 and 4

**ANS- D**

1. 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**

**ANS- 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

**ANS- 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

**ANS- 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**

**ANS- 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

**ANS- A**

8- Clustering is a

a. Supervised learning

**b. Unsupervised learning**

c. Reinforcement learning

d. None

**ANS- 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**

**ANS- 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

**ANS- 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**

**ANS- D**

**12-** For clustering, we do not require

**a. Labeled data**

b. Unlabeled data

c. Numerical data

d. Categorical data

**ANS- A**

13- How is cluster analysis calculated?

ANS- Cluster Analyses can be found in Analyze/Classify…. SPSS offers three methods for the cluster analysis:

1. K-Means Cluster
2. Hierarchical Cluster
3. Two-Step Cluster.

K-means cluster is a method to quickly cluster large data sets. The researcher define the number of clusters in advance. This is useful to test different models with a different assumed number of clusters.

Hierarchical cluster is the most common method. It generates a series of models with cluster solutions from 1 (all cases in one cluster) to n (each case is an individual cluster). Hierarchical cluster also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis. In addition, hierarchical cluster analysis can handle nominal, ordinal, and scale data; however it is not recommended to mix different levels of measurement.

Two-step cluster analysis identifies groupings by running pre-clustering first and then by running hierarchical methods. Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods. In this respect, it is a combination of the previous two approaches. Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

14- How is cluster quality measured?

ANS- To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set. The silhouette coefficient and other intrinsic measures can also be used in the elbow method to heuristically derive the number of clusters in a data set by replacing the sum of within-cluster variances.

15-What is cluster analysis and its types?

ANS- Cluster analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters. Cluster analysis is also called classification analysis or numerical taxonomy. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects.

Clustering itself can be categorized into two types viz. Hard Clustering and Soft Clustering. In hard clustering, one data point can belong to one cluster only. But in soft clustering, the output provided is a probability likelihood of a data point belonging to each of the pre-defined numbers of clusters.