**MACHINE LEARNING – WORKSHEET (CLUSTERING)**

1. . Which of the following is an application of clustering

a. Biological network analysis

b. Market trend prediction

c. Topic modeling

**d. All of the above**

**ANS- D**

1. On which data type, we cannot perform cluster analysis?

a. Time series data

b. Text data

c. Multimedia data

**d. None**

**ANS-D**

1. . Netflix’s movie recommendation system uses

a. Supervised learning

b. Unsupervised learning

**c. Reinforcement learning**

d. All of the above

**ANS- C**

1. The final output of Hierarchical clustering is

a. The number of cluster centroids

**b. The tree representing how close the data points are to each other**

c. A map defining the similar data points into individual groups

d. All of the above

**ANS- B**

1. Which of the step is not required for K-means clustering?

a. a distance metric

b. initial number of clusters

c. initial guess as to cluster centroids

**d. None**

**ANS- D**

1. Which is the following is wrong?

a. k-means clustering is a vector quantization method

b. k-means clustering tries to group n observations into k clusters

**c. k-nearest neighbor is same as k-means**

d. None

**ANS-C**

1. . Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?

1. Single-link

2. Complete-link

3. Average-link

Options:

a. 1 and 2

b. 1 and 3

c. 2 and 3

**d. 1, 2 and 3**

**ANS- D**

1. Which of the following are true?

1. Clustering analysis is negatively affected by multicollinearity of features

2. Clustering analysis is negatively affected by heteroscedasticity

Options:

**a. 1 only**

b. 2 only

c. 1 and 2

d. None of them

**ANS-A**

1. **ANS- A**
2. For which of the following tasks might clustering be a suitable approach?

a. Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

**b. Given a database of information about your users, automatically group them into different market segments.**

c. Predicting whether stock price of a company will increase tomorrow.

d. Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

ANS- B

1. **ANS- A**
2. **ANS-B**
3. What is the importance of clustering?

**ANS-**Clustering is used to segment the data. Unlike classification, clustering models segment data into groups that were not previously defined. Classification models segment data by assigning it to previously-defined classes, which are specified in a target. Clustering models do not use a target.

Clustering is also useful for exploring data. If there are many cases and no obvious groupings, clustering algorithms can be used to find natural groupings. Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build supervised models.

Clustering can also be used for anomaly detection. Once the data has been segmented into clusters, you might find that some cases do not fit well into any clusters. These cases are anomalies or outliers

1. How do you cluster a profile?

**ANS-** To do this, you can profile your clusters after you have run a cluster analysis. Profiling involves generating descriptions of the clusters with reference to the input variables you used for the cluster analysis. Profiling acts as a class descriptor for the clusters and will help you to ‘tell a story’ so that you can understand this information and use it across your business.

This approach is most useful where a multivariate cluster analysis has been conducted so that the clusters can be described from multiple angles and provide a more holistic description of your shopper’s behaviour. Once you have described and profiled each cluster, you can make inferences from the information provided.

For example, if you only have access to POS data, you would be able to assume the income group of the customers in the cluster based on the average spend per shopping trip, products purchased and brands purchased.

1. How can I improve my clustering performance?

ANS- Clustering algorithm can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm.

When the data has overlapping clusters, k-means can improve the results of the initialization technique.

When the data has well separated clusters , the performance of k-means depends completely on the goodness of the initialization.

Initialization using simple furthest point heuristic (Maxmin) reduces the clustering error of k-means from 15% to 6%, on average.