Submitted By: Arjun B
Batch Number: DS0082

WORKSHEET 1 - MACHINE LEARNING

1. b) 4
2. d) 1, 2 and 4
3. d) Formulating the clustering problem
4. a) Euclidean Distance
5. b) Divisive clustering
6. d) All answers are correct
7. a) Divide the data points into groups
8. b) Unsupervised learning
9. d) All of the above
10. a) K-means clustering algorithm
11. d) All of the above
12. a) Labelled data

13. The objective of the clustering algorithm is to effectively split a data set S made up of n-tuples of real numbers into k clusters C1,..., Ck. A single element from each cluster Cj that has been chosen is referred to as a centroid.

    Step 1: Select the number of clusters, k.

    Step 2: Pick a starting set of k centroids.

    Step 3: The third step is to position the closest centroid for each data element (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: Choose a different centroid for each cluster.

    Step 5: Repeat Step 3 until the centroids remain the same.

14. Measures for assessing the quality of clustering

    1. **Ragbag-** Some categories' objects might not be able to be combined with those of other categories in some situations. The quality of the clusters is then evaluated using the Rag Bag method. We should classify the heterogeneous object into a rag bag category in accordance with the rag bag approach.
    2. **Small cluster preservation-**A small category of clustering can no longer be distinguished from the clustering if it is further broken up into tiny pieces, which adds noise to the overall clustering. A small category shouldn't be

divided into parts, per the small cluster preservation criterion, because doing so would further degrade the quality of the clusters because the fragments of the clusters are distinct.

3. **Dissimilarity/Similarity metric-** The distance function, denoted by d, can be used to express the similarity between the clusters (i, j). For various forms of data, the distance function can be expressed as Euclidean, Mahalanobis, or Cosine distance etc.
4. **Cluster completeness -** According to the ground truth, if any two data objects have similar attributes, they are assigned to the same category in the cluster. Effective clustering depends on the completeness of the clusters. The cluster completion rate is high if all of the objects are members of the same category.

15. Finding related groups of things to form clusters is done through the technique of cluster analysis. This method acts on unlabeled data and makes use of unsupervised machine learning. Every object contained in a cluster created by a set of data points would belong to the same group.

- Partitioning Method - Making partitions on the data in order to create clusters is the method of partitioning.If “n” partitions are done on “p” objects of the database then each partition is represented by a cluster and n < p. The two conditions which need to be satisfied with this Partitioning Clustering Method are:
    - One objective should only belong to only one group.
    - There should be no group without even a single purpose.

- Hierarchical Method - The given set of data objects is divided into hierarchical subsets using the hierarchical approach. On the basis of how the hierarchical decomposition is created, we may categorize hierarchical approaches and determine the classification's purpose. The two types of hierarchical methods are the agglomerative approach and the divisive approach.

- Density-based Method - This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold,

i.e. for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

- Grid-Based Method - In the Grid-Based technique, a grid is created by employing all objects, i.e., the object space is quantized into a set number of grid cells. The grid-based approach's quick processing time, which depends solely on the number of cells in each dimension of the quantized space, is one of its main advantages.

- **Model-Based Method** - In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects the spatial distribution of the data points.

- **Constraint Based Method** - With this approach, user- or application-oriented constraints are used to produce the clustering. The user expectation or the characteristics of the expected clustering results are examples of constraints. With the use of constraints, we may engage with the clustering process. The user or the requirements of the program might specify constraints.