WORKSHEET 1 - MACHINE LEARNING

Submitted by Jwala R, DS0082 (17/01/2023)

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. **Clustering** algorithm's goal is to efficiently divide a data set S made up of n-tuples of real numbers into k clusters C1,..., Ck. A centroid is a single element from each cluster Cj that has been selected.

    Step 1: Select the number of clusters, k.
    Step 2: Pick a starting set of k centroids.
    Step 3: The third step is to place each data element's 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: Choose a different centroid for each cluster.
    Step 5: Go to Step 3 and repeat the procedure until the centroids stay the same (or some other convergence criterion is met)

14. **Measures for assessing the quality of clustering**:

    - **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.
    - **Cluster completeness -** If any two data objects have comparable qualities, they are placed in the same category of the cluster in accordance with the ground truth. Cluster completeness is a key factor in effective clustering. If the objects belong to the same category, the cluster completion rate is high.
    - **Small cluster preservation** - If a small category of clustering is further divided into little parts, those small bits of cluster create noise to the overall clustering and make it impossible to distinguish the small category from the clustering. According to the small cluster preservation criterion, it is not recommended to divide a small category into pieces since the pieces of clusters are distinctive, which further lowers the quality of clusters.
    - **Ragbag -** In certain circumstances, there may be a few categories whose objects cannot be mixed with those of other categories. The Rag Bag

technique is then used to measure the quality of those clusters. According to the rag bag method, we should put the heterogeneous object into a rag bag category.

15. Cluster Analysis is the process to find similar groups of objects in order to form clusters.This approach uses unsupervised machine learning and operates on unlabeled data. Each object in a cluster formed by a collection of data points would be a member of the same group.

Types:

- **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 categorise hierarchical approaches and determine the classification's purpose. The two types of hierarchical methods are agglomerative approach and 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 neighbourhood 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 of the 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 hypothesised 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 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 programme might specify constraints.