ASSIGNMENT -1

MACHINE LEARNING

1. a) 2
2. d) 1,2 and 4
3. d) formulating the clustering problem
4. a) Euclidean distance
5. b) Divisive clustering
6. d) All answer are correct
7. a) divide the data points into groups
8. b) Unsupervised learning
9. d) All of above
10. a) K-means clustering
11. d) all of the above
12. a) labeled data

13. Clustering analysis is a form of exploratory data analysis in which observations are divided into different groups that share common characteristics. The purpose of cluster analysis (also known as classification) is to construct groups (or classes or *clusters*) while ensuring the following property: **within a group** the observations must be as **similar** as possible (intracluster similarity), while observations belonging to **different groups** must be as **different** as possible (intercluster similarity).

    Majorly three methods for the cluster analysis: *K-Means Cluster*, *Hierarchical Cluster*, and *Two-Step Cluster*.

    K-mean clusters- *K-means cluster* is a method to quickly cluster large data sets.

    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)

    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.

    Cluster analysis can be calculated as follows:
    1) calculate the distances,
    2) link the clusters, and
    3) choose a solution by selecting the right number of clusters.

14. A clustering-quality measure (CQM) is a function that, given a data set and its par- tition into clusters, returns a non-negative real number representing how strong or conclusive the clustering

is. A clustering-quality measure is a function that maps pairs of the form (dataset, clustering) to some ordered set (say, the set of non-negative real numbers), so that these values reflect how ‘good’ or ‘cogent’ that clustering is. Measures for the quality of a clusterings are of interest not only as a vehicle for axiomatizing clustering. The need to measure the quality of a given data clustering arises naturally in many clustering issues. The aim of clustering is to uncover meaningful groups in data. Clustering-quality measures may also be used to help in clustering model-selection by comparing different clusterings over the same data set.

15. Cluster Analysis is the process to find similar groups of objects in order to form clusters.

The clustering methods can be classified into the following categories:

- Partitioning Method.
- Hierarchical Method.
- Density-based Method.
- Grid-Based Method.
- Model-Based Method.
- Constraint-based Method.

a) **Partitioning method :** This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data. Its the data analysts to specify the number of clusters that has to be generated for the clustering methods. In the partitioning method when database(D) that contains multiple(N) objects then the partitioning method constructs user-specified(K) partitions of the data in which each partition represents a cluster and a particular region. There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Medoids), CLARA algorithm (Clustering Large Applications) etc.

Hierarchical Method: Hierarchical clustering is another unsupervised learning algorithm that is used to group together the unlabeled data points having similar characteristics. Hierarchical clustering algorithms falls into following two categories.

- **Agglomerative hierarchical algorithms** − In agglomerative hierarchical algorithms, each data point is treated as a single cluster and then successively merge or agglomerate (bottom-up approach) the pairs of clusters. The hierarchy of the clusters is represented as a dendrogram or tree structure.
- **Divisive hierarchical algorithms** − On the other hand, in divisive hierarchical algorithms, all the data points are treated as one big cluster and the process of clustering involves dividing (Top-down approach) the one big cluster into various small clusters.

Density-based Method:

To discover clusters with arbitrary shape, density-based clustering methods have been developed. – These typically regard clusters as dense regions of objects in the data space that are separated by regions of low density (representing noise).

Density-based clustering algorithms

- **DBSCAN:** grows clusters according to a density-based connectivity analysis. –
- **OPTICS:** extends DBSCAN to produce a cluster ordering obtained from a wide range of parameter settings. –
- **DENCLUE:** clusters objects based on a set of density Density-Based Methods distribution functions.

Grid-Based Method:

In Grid-Based Methods, **the space of instance is divided into a grid structure**. Clustering techniques are then applied using the Cells of the grid, instead of individual data points, as the base units. The biggest advantage of this method is to improve the processing time. The benefit of the method is its quick processing time, which is generally independent of the number of data objects, still dependent on only the multiple cells in each dimension in the quantized space.

An instance of the grid-based approach involves STING, which explores statistical data stored in the grid cells, WaveCluster, which clusters objects using a wavelet transform approach, and CLIQUE, which defines a grid-and density-based approach for clustering in high-dimensional data space.STING is a grid-based multiresolution clustering method in which the spatial area is divided into rectangular cells.

Model-Based Method: Model-based clustering is a statistical approach to data clustering. The observed (multivariate) data is assumed to have been generated from a finite mixture of component models. Each component model is a probability distribution, typically a parametric multivariate distribution. For example, in a multivariate Gaussian mixture model, each component is a multivariate Gaussian distribution. The component responsible for generating a particular observation determines the cluster to which the observation belongs. However, the component generating each observation as well as the parameters for each of the component distributions are unknown. The key learning task is to determine the component responsible for generating each observation, which in turn gives the clustering of the data. Ideally, observations generated from the same component are inferred to belong to the same cluster.

Constraint-based Method: constrained clustering is a class of semi-supervised learning algorithms. Typically, constrained clustering incorporates either a set of must-link constraints, cannot-link constraints, or both, with a data clustering algorithm.A cluster in which the members conform to all must-link and cannot-link constraints is called a chunklet.

Types of constraints:

Both a must-link and a cannot-link constraint define a relationship between two data instances. Together, the sets of these constraints act as a guide for which a constrained clustering algorithm will attempt to find chunklets (clusters in the dataset which satisfy the specified constraints).

- A **must-link constraint** is used to specify that the two instances in the must-link relation should be associated with the same cluster.

- A **cannot-link constraint** is used to specify that the two instances in the cannot-link relation should *not* be associated with the same cluster.

Some constrained clustering algorithms will abort if no such clustering exists which satisfies the specified constraints. Others will try to minimize the amount of constraint violation should it be impossible to find a clustering which satisfies the constraints. Constraints could also be used to guide the selection of a clustering model among several possible solutions.