ASSIGNMENT-1

MACHINE LEARNING

1. a) 2.
2. c) 2 and 4.
3. d) Formulating the clustering problem.
4. a) Euclidian 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) Labeled data.
13. Calculation of Cluster analysis:

    1. It starts by putting every point in its own cluster, so each cluster is a singleton
    2. It then merges the 2 points that are closest to each other based on the distances from the distance matrix. The consequence is that there is one less cluster
    3. It then recalculates the distances between the new and old clusters and save them in a new distance matrix which will be used in the next step
    4. Finally, steps 1 and 2 are repeated until all clusters are merged into one single cluster including all points.

There are 5 main methods to measure the distance between clusters, referred as linkage methods:

1. Single linkage: computes the minimum distance between clusters before merging them.
2. Complete linkage: computes the maximum distance between clusters before merging them.
3. Average linkage: computes the average distance between clusters before merging them.
4. Centroid linkage: calculates centroids for both clusters, then computes the distance between the two before merging them.
5. Ward's (minimum variance) criterion: minimizes the total within-cluster variance and find the pair of clusters that leads to minimum increase in total within-cluster variance after merging.

14. We can measure the quality of Clustering by using the dissimilarity/similarity metric in most situations. But there are some other methods to measure the qualities of good clustering.

1. Dissimilarity/Similarity metric:

   The similarity between the clusters can be expressed in terms of a distance function. Which is represented by d(i, j). Distance functions are different for various data types and data variables. Distance function measure is different for continuous-valued, categorical and vector variables. Distance function can be expressed as Euclidean distance, Mahalanobis distance, and Cosine distance for different types of data.

2. Cluster completeness:

   It is the essential parameter for good clustering, if any two data objects are having similar charcteristics then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category.

3. Ragbag:

   In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then the quality of those cluster categories is measured by the Ragbag method. According to the Ragbag method, we should put the heterogeneous object into a ragbag category.

4. Small cluster preservation:

   If a small category of clustering is further split into small pieces, then those small pieces of cluster become noise to the entire clustering and thus it becomes difficult to identify that small category from the clustering. The small cluster preservation criterion states that are splitting a small category into pieces is not advisable and it further decreases the quality of clusters are distinctive.

15. Cluster analysis is a data analysis/mining technique that explores the naturally occurring groups within a dataset known as “clusters”. Cluster analysis doesn’t need to group data points into any predefined groups, it is unsupervised learning method.

    In unsupervised learning, insights are derived from the data without any predefined labels or classes.

Types of cluster analysis:

- Hierarchical cluster analysis.
- Centroid-based clustering.
- Distribution-based clustering.
- Density-based clustering.