Answers for MACHINE LEARNING (CLUSTERING) Worksheet-1

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. d. All of the above
8. b. Unsupervised learning
9. a. K- Means clustering
10. a. K-means clustering algorithm
11. d. All of the above
12. a. Labeled data
13. Cluster analysis (for example K-means) is calculated by using following steps:
    - Deciding the number of clusters.
    - Randomly initializing the cluster centroids that is equal to the number of clusters.
    - Calculating the distance (example: euledian distance) of each data points to the centroids, and assigning the data point to that centroid having the least distance.

- Re-computation of the centroids. This is done by taking the mean of all data points assigned to that centroid's cluster. In this step all the cluster centroids are updated.
- Now, step 2 and 3 re-iterates until the cluster distribution does not change, in which the algorithm gets converged forming the decided number of clusters.

Here, the above steps are used for the K-means algorithm which is a centroid clustering technique, it is to be noted that other clustering algorithms uses different approach and may be more robust to data distribution, however, the idea for calculation of cluster analysis is to find the method (mathematically) to group the data points based on their similarities.

14. The quality of clusters is measured by two important factors:
    - Inter-cluster distance
    - Intra-cluster distance

Inter-cluster distance: It is the distance between two data points belonging to two different cluster.

Intra-cluster distance: It is the distance between two data points belonging to the same cluster.

For any clustering algorithm to improve its quality of clusters formed, the Inter-cluster distance should increase and Intra-cluster distance should decrease.

One of the well-known techniques that uses the above concept is called **Silhouette.** Here, with the help of Silhouette co-efficient score, we can quantify the quality and validate the clusters formed.

Silhouette co-efficient is calculated using the mean intra-cluster distance (a) and the mean inter-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is given by:

(b-a)/max(b, a)

Its value ranges from -1 to 1, where, -1 indicates worst score, 0 indicates overlapping clusters and 1 indicates best score.

15. Cluster analysis is the task of grouping the data points based on their similarities or relevancy to one another. It is done so to understand the hidden patterns or structure within the data.

    The 4 basic clustering analysis used in data analytics are:

    - **Centroid Clustering**: In this clustering technique, the centroid for each pre-defined cluster needs to be initialized randomly, and based on the similarities between each point and the centroids, the data points get assigned to the respective centroids forming a cluster. This process is iterated until the algorithm gets converged.

    - **Density Clustering:** This technique works on the bases of how dense the data points are in order to form a cluster. The algorithm starts by selecting random point to find the distance between each point around it, so it needs this pre-defined distance between data points so to understand how closely they are to consider them related. Also, the algorithm will find other data points as well as that are closely related to one another. This process iterates continuously by selecting different random data points to start with until the optimum clusters are formed.

    - **Distribution Clustering:** This clustering technique is based on the probability distribution of the data points for it to belong to a particular cluster. Here, around each centroid the algorithm defines density distribution for each cluster, which gives the probability of a data point belonging to a given cluster.

    - **Connectivity Clustering:** In this clustering technique, each data point is initially identified as a single cluster, but as the algorithm progresses, the data points or small clusters within the proximity gets engulfed with each bigger cluster in a hierarchical fashion.

Therefore, these type of clustering techniques are also referred to as **hierarchical clustering.**