**Machine Learning Worksheet**

**Q1.** B

**Q2.** D

**Q3.** D

**Q4.** A

**Q5.** B

**Q6.** D

**Q7.** D

**Q8.** B

**Q9.** D

**Q10.** A

**Q11.** D

**Q12.** C

**Q13.** The hierarchical cluster analysis follows three basic steps:

1. calculate the distances,
2. link the clusters, and
3. choose a solution by selecting the right number of clusters.

**Q14**. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set.

**Q15.** Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. These techniques create clusters that allow us to understand how our data is related. The most common applications of cluster analysis in a business setting is to segment customers or activities. In this post we will explore four basic types of cluster analysis used in data science. These types are Centroid Clustering, Density Clustering Distribution Clustering, and Connectivity Clustering.

* **Connectivity models:** As the name suggests, these models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models can follow two approaches. In the first approach, they start with classifying all data points into separate clusters & then aggregating them as the distance decreases. In the second approach, all data points are classified as a single cluster and then partitioned as the distance increases. Also, the choice of distance function is subjective. These models are very easy to interpret but lacks scalability for handling big datasets. Examples of these models are hierarchical clustering algorithm and its variants.
* **Centroid models:** These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. K-Means clustering algorithm is a popular algorithm that falls into this category. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima.
* **Distribution models:** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from over fitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.
* **Density Models:**These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Popular examples of density models are DB scan and optics.