MACHINE LEARNING – WORKSHEET (CLUSTERING)

13. How is cluster analysis calculated?

Cluster analysis, or clustering, is an unsupervised machine learning task. It involves automatically discovering natural grouping in data. Unlike supervised learning , clustering algorithms only interpret the input data and find natural groups or clusters in feature space. Cluster analysis is also used to group variables into homogeneous and distinct groups. The Cluster analysis follows three basic steps:

* Calculate the distance
* Link the clusters
* Choose a solution by selecting the right number of clusters.

14. How is cluster quality measured?

To measure a cluster's quality within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set. We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether the Label is available.  If label is available, it can be used by **extrinsic methods**, which compare the clustering against the labels and measure. If the label is unavailable, we can use **intrinsic methods**, which evaluate the goodness of a clustering by considering how well the clusters are separated.

15. What is cluster analysis and its types?

There are four basic types of cluster analysis used in data science, they are Centroid Clustering, Density Clustering, Distribution Clustering, and Connectivity Clustering.

1. Centroid Clustering: The algorithm will start by randomly selecting centroids (cluster centers) to group the data points into the two pre-defined clusters. A line is then drawn separating the data points into the two clusters based on their proximity to the centroids. The algorithm will then reposition the centroid relative to all the points within each cluster. The centroids and points in a cluster will adjust through all iterations, resulting in optimized clusters. The result of this analysis is the segmentation of your data into the two clusters.
2. Density Clustering: This algorithm groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points, the more related they are. To determine this, the algorithm will select a random point then start measuring the distance between each point around it. For most density algorithms a predetermined distance between data points is selected to benchmark how closely points need to be to one another to be considered related. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.
3. Distribution Clustering: Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid The algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions The algorithm optimizes the characteristics of the distributions to best represent the data. Distribution clustering is a great technique to assign outliers to clusters, whereas density clustering will not assign an outlier to a cluster.
4. Connectivity Clustering: This algorithm initially recognizes each data point as its own cluster. The primary goal of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.