**Machine Learning Worksheet -1**

Q.1 (b)

Q.2 (d)

Q.3 (d)

Q.4 (a)

Q.5 (b)

Q.6 (d)

Q.7 (d)

Q.8 (b)

Q.9 (a)

Q.10 (a)

Q.11 (d)

Q.12 (a)

Q.13 First, an initial partition with k clusters (given number of clusters) is created.

* Then, starting with the first object in the first cluster, Euclidean distances of all objects to all cluster foci are calculated.
* If an object is detected whose distance to the center of gravity of the own cluster is greater than the distance to the center of gravity (centroid) of another cluster, this object is shifted to the other cluster.
* Finally, the centroids of the two changed clusters are calculated again, since the compositions have changed here.
* These steps are repeated until each object is located in a cluster with the smallest distance to its centroid (center of the cluster) (optimal solution).

Q.14 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 ground truth is available. Here, ground truth is the ideal clustering that is often built using human experts.

If ground truth is available, it can be used by extrinsic methods, which compare the clustering against the group truth and measure. If the ground truth is unavailable, we can use intrinsic methods, which evaluate the goodness of a clustering by considering how well the clusters are separated. Ground truth can be considered as supervision in the form of “cluster labels.” Hence, extrinsic methods are also known as supervised methods, while intrinsic methods are unsupervised methods.

Q.15 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.

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

## Centroid Clustering

This is one of the more common methodologies used in cluster analysis. In centroid cluster analysis you choose the number of clusters that you want to classify. For example, if you’re a pet store owner you may choose to segment your customer list by people who bought dog and/or cat products.

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 iteratations, resulting in optimized clusters. The result of this analysis is the segmentation of your data into the two clusters. In this example, the data set will be segmented into customers who are own dogs and cats.

## Density Clustering

Density clustering 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.

## 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.

These maps look a lot like targets at an archery range. In the event that a data point hits the bulls eye on the map, then the probability of that person/object belonging to that cluster is 100%. Each ring around the bulls eye represents lessening percentage or certainty.

Distribution clustering is a great technique to assign outliers to clusters, where as density clustering will not assign an outlier to acluster.

## Connectivity Clustering

Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise 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.