**ML WORKSHEET-1**

**1**-(b)

**2**-(d)

**3**-(d)

**4**-(a)

**5**-(b)

**6**-(d)

7-(a)

**8**-(b)

**9**-(d)

**10**-(a)

**11**-(d)

**12**-(a)

**13**-

Cluster analysis is an exploratory analysis that tries to identify structures within the data.  Cluster analysis is also called segmentation analysis or taxonomy analysis.  More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known.  Because it is exploratory, it does not make any distinction between dependent and independent variables.  The different cluster analysis methods that SPSS offers can handle binary, nominal, ordinal, and scale (interval or ratio) data.

In SPSS Cluster Analyses can be found in Analyze/Classify….  SPSS offers three methods for the cluster analysis: K-Means Cluster, Hierarchical Cluster, and Two-Step Cluster.

K-means cluster is a method to quickly cluster large data sets.  The researcher define the number of clusters in advance.  This is useful to test different models with a different assumed number of clusters.

Hierarchical cluster is the most common method.  It generates a series of models with cluster solutions from 1 (all cases in one cluster) to n (each case is an individual cluster).  Hierarchical cluster also works with variables as opposed to cases; it can cluster variables together in a manner somewhat similar to factor analysis.  In addition, hierarchical cluster analysis can handle nominal, ordinal, and scale data; however it is not recommended to mix different levels of measurement.

Two-step cluster analysis identifies groupings by running pre-clustering first and then by running hierarchical methods.  Because it uses a quick cluster algorithm upfront, it can handle large data sets that would take a long time to compute with hierarchical cluster methods.  In this respect, it is a combination of the previous two approaches.  Two-step clustering can handle scale and ordinal data in the same model, and it automatically selects the number of clusters.

**14-**

# Clustering quality

Once clustering is done, how well the clustering has performed can be quantified by a number of metrics. Ideal clustering is characterised by minimal intra cluster distance and maximal inter cluster distance.

There are majorly two types of measures to assess the clustering performance.

(i) Extrinsic Measures which require ground truth labels. Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.

(ii) Intrinsic Measures that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.

**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.

# 4 Types of Cluster Analysis Techniques

# 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.