MACHINE LEARNING – WORKSHEET I

(CLUSTERING) SOLUTIONS

1. b
2. e
3. d
4. a
5. b
6. d
7. a
8. b
9. a
10. a
11. d
12. a

Clustering is a process of segregating populations into a number of groups that are supposed to have similarity. Clustering Analysis is a concept in machine leaning which is used to group data points which are supposed to exhibit similar patterns. Clustering analysis is used on a unsupervised data, that means that data which do not have any labels. Clustering help to find natural groups in the feature space of the input data.

The following are the steps to perform clustering analysis.

* Set an optimal number of cluster centroids k. These will be the total number of clusters formed in your dataset
* For each value of k, calculate the total-within sum of square. This is done by taking the taking the mean vector of the points in that particular cluster until convergence.
* Perform this for every cluster.
* The optimal number of clusters k could be calculated by the Elbow method. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

Number of optimal clusters could be also found by other methods such as the Average silhouette method and the gap statistic method.

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. The ground truth could be generated the human expertise in the domain.

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.

Extrinsic Methods:

* Clustering Homogeneity: A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class.
* Cluster Completeness: A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster.
* Rag Bag: Elements with low relevance to the categories (e.g., noise) should be preferably assigned to the less homogeneous clusters
* Small Cluster Preservation: The small cluster preservation criterion states that splitting a small category into pieces is more harmful than splitting a large category into pieces.

Intrinsic Methods:

* Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters
* The Dunn Index (DI) is a metric for judging a clustering algorithm. A higher DI implies better clustering. It assumes that better clustering means that clusters are compact and well-separated from other clusters.

15:

Clustering is a process of segregating populations into a number of groups that are supposed to have similarity. Clustering Analysis is a concept in machine leaning which is used to group data points which are supposed to exhibit similar patterns. Clustering analysis is used on a unsupervised data, that means that data which do not have any labels. Clustering help to find natural groups in the feature space of the input data.

Types of Cluster Analysis are as follows:

* Centroid Clustering: In centroid cluster analysis you choose the number of clusters that you want to classify. the initial cluster centroids are selected randomly. After that, each gene is assigned to the closest cluster centroid. Then each cluster centroid is moved to the mean of the points assigned to it. This algorithm converges when the assignments no longer change. Eg : K-Means clustering
* 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 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 convergence. Eg, DBSCAN clustering
* 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.
* Hierarchical Clustering: The result of hierarchical clustering is a tree-based representation of the objects, which is also known as dendrogram. Observations can be subdivided into groups by cutting the dendrogram at a desired similarity level.