Unsupervised Learning

Preprocessing Data:

Since the data is being collected from a banking institute, we need to make sure the data we use for clustering is relevant to the banking sector as much as possible.

First, we remove entries with null value in their ages and jobs (as they are the most difficult to “guess”). Other fields can be guessed by implying another clustering algorithm on those fields relating them to the age.

Clustering Algorithm:

Since categorical data is involved, we can use **K-modes clustering algorithm** which uses **Gower Distance**. Gower Distance measures distance between two data points by comparing their similarities- 0 for identical points and 1 for maximally dissimilar. K-modes clustering algorithm uses modes to measure differences in clusters.

Some other algorithms that came in my way were Hierarchical Clustering which was rejected as the final output is a single cluster containing multiple clusters which makes it more computationally demanding when dealing with large datasets. K-means Clustering was rejected as it works on Euclidean Distance which fails on categorical data. K-means requires numerical data.

K-Modes Clustering Algorithm: -

In this, we choose ‘K’ number of clusters to be formed. This number is determined according to certain requirements/conditions mentioned in the next section. Then, ‘K’ number of data points are chosen and assigned to each of those clusters. Now, based on the categorical field values of these data points, those closest to these “centers” of the clusters are chosen and assigned to their respective clusters.

Now, mode of each categorical field value is taken and assigned as the centers of these clusters and the process is repeated till convergence is achieved. By convergence, it means each data point is already in closest mode cluster.

Determination of ‘K’:

In order to determine ‘K’, we can use the **Elbow Method.** In this, we assume a range of ‘K’, suppose 1 to 6. We run a K-modes method for each of this value of K and plot the graph of its cost (sum of dissimilarities between clusters) vs ‘K’. This graph will have a comparably sharp bend. The value of ‘K’ at that bend is used to get the correct number of cluster formation. This method does have a bit of ambiguity which can be fixed by using **Silhouette Method** but **Silhouette Method**, generally, uses Euclidean Distance, making it unsuitable for our use case.

The formula for **Silhouette Method** is as follows: -

For C(i)=1, s(i)=0

= similarity between point and its cluster’s mode.

= dissimilarity between point and other clusters’ modes.