Unsupervised Learning

Introduction:

Clustering is a division of data into a group of data points, similar data points are in one group called cluster and dissimilar data points are in another cluster. Categorical data is a collection of categories and each value represents some category. There are various algorithms introduced for clustering categorical data. These algorithms are available for clustering categorical data but no single algorithm can achieve the best result for all the data sets.

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/unknown 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 and job.

Optional: - In order to increase efficiency, we can reduce the data parameters by combining house loans and personal loans into one 'loan' category with answers 'yes', 'no' and 'unknown', and also, remove day of week of last known contact (as it isn't a sensible parameter for categorizing and clustering)

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 (or non-numerical) data. K-means requires numerical data.

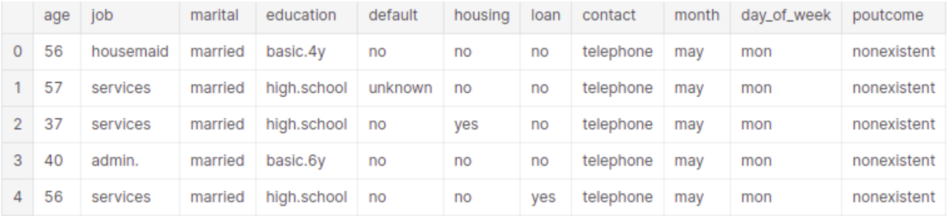
The K-Modes Clustering Algorithm extends the K-Means to cluster categorical data by using a simple matching dissimilarity measure, modes instead of means for cluster and a frequency-based method to update modes in K-Means Fashion to minimize cost function, hence being just as efficient as the K-Means Algorithm. It defines clusters based on the number of matching categories between data points. However, the clustering is sensitive to initial conditions.

K-Modes Clustering : -

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 at random, and assigned (as centers) to each of those clusters. Now, based on the categorical field values of these data points, those closest (with least dissimilarities) to these "centers" of the clusters are chosen and assigned to their respective clusters.

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

For example, in the given dataset,



	age	job	marital	education	default	housing	loan	contact	month	day_of_week	poutcome
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	nonexistent
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	nonexistent
2	37	services	married	high.school	no	yes	no	telephone	may	mon	nonexistent
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	nonexistent
4	56	services	married	high.school	no	no	yes	telephone	may	mon	nonexistent

The centre for cluster involving Entry 1,2,4 would be

Services,married,high.school,no,yes,telephone,may,mon

Hence any other entry matching (or having lowest dissimilarities) with this set of categorical values will be put into this Cluster.

Dissimilarities are counted in the following manner: -

1. Example:- Services,unmarried,graduate,no,no,telephone,jan,tuesday

Here the Dissimilarity count with above example is 5.

2. Example: - Services,married,high.school,yes,no,phone,may,mon

Here, the Dissimilarity count is 3.

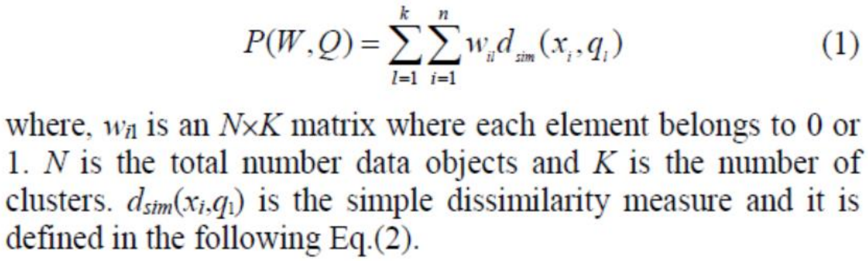
Hence Example 2 will be put into the cluster of entry 1,2,4.

Step-by-Step Algorithm:-

Step 1: Randomly select the K initial modes from the data objects such that Cj, j = 1,2,...,K

Step 2: Find the matching dissimilarity between each K initial cluster modes and each data objects using $d = \sum \delta \ where \ \delta = 1 \ if \ data \ is \ equal \ and \ 0 \ if \ unequal$

Step 3: Evaluate the fitness using



$$P(W,Q)=\sum_{l=1}^{k}\sum_{i=1}^{n} w_{il} d_{sim}(x_i,q_l) \qquad (1)$$

where, w_{il} is an $N\times K$ matrix where each element belongs to 0 or 1. N is the total number data objects and K is the number of clusters. $d_{sim}(x_i,q_l)$ is the simple dissimilarity measure and it is defined in the following Eq.(2).

Step 4: Find the minimum mode values in each data object i.e. finding the objects nearest to the initial cluster modes.

Step 5: Assign the data objects to the nearest cluster centroid modes.

Step 6: Update the modes by apply the frequency-based method on newly formed clusters.

Step 7: Recalculate the similarity between the data objects and the updated modes.

Step 8: Repeat the step 4 and step 5 until no changes in the cluster ship of data objects. Output: Clustered data objects

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: -$s(i) = \frac{(b(i)-a(i))}{\max\{a(i),b(i)\}}, \ if \ C(i) > 1$

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

$a(i)$ = similarity between point and its cluster's mode.

$b(i)$ = dissimilarity between point and other clusters' modes.

CODE:

https://colab.research.google.com/drive/1qdEJ7UXPBJQ2J7Cgp9H4Lje3pRh8Gyh5?usp=sharing