Subjective Questions

Assignment Summary:

In this assignment we had to find the 5 countries which were in direst need of aid from the NGO. The dataset contained numerical columns thus we did EDA by plotting distribution plot, pair plot and box plot. Here we decided to drop a few countries which were having high GDP values, since we were finding backward countries. Then we did a Hopkins test to check the clustering tendency. We chose the value of K to be 3 from elbow curve and silhouette scores. The model which we chose for clustering was K Means because it provided better results when compared to Hierarchical clustering and there was balance between the 3 clusters.

Compare and contrast K-means Clustering and Hierarchical Clustering:

- Hierarchical clustering can’t handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2).
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means are found to work well when the shape of the clusters is hyper spherical.
- K Means clustering requires prior knowledge of K, that is the number of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

Briefly explain the steps of the K-means clustering algorithm:

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers

1. Randomly select ‘c’ cluster centers
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers
4. Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, ‘Ci’ represents the number of data points in ‘ith’ cluster.

5. Recalculate the distance between each data point and new obtained cluster centers
6. If no data point was reassigned then stop, otherwise repeat from step 3

Finally, this algorithm aims at minimizing an objective function know as squared error function given by

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\| x_i - v_j \| \right)^2$$

Where, ‘||Xi-Vj||’ is the Euclidean distance between Xi and Vj.

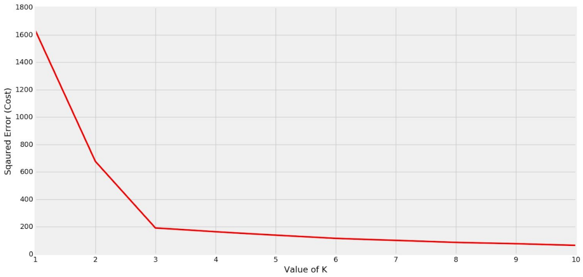
‘Ci’ is the number of data points in the ‘ith’ cluster.

‘C’ is the number of cluster centers.

How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it:

Elbow Method

The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the **elbow point**.



Silhouette Algorithm

Here, we assume that the data has already been clustered into k clusters typically by a K-Means clustering technique. Then for each data point, we define the following:-

C(i) -The cluster assigned to the ith data point

|C(i)| – The number of data points in the cluster assigned to the ith data point

a(i) – It gives a measure of how well assigned the ith data point is to it’s cluster

$$a(i) = \frac{1}{|C(i)|-1} \sum_{C(i),i\neq j} d(i,j)$$

b(i) – It is defined as the average dissimilarity to the closest cluster which is not it’s cluster

$$b(i) = min_{i\neq j}\left(\frac{1}{|C(j)|} \sum_{j\epsilon C(j)} d(i,j)\right)$$

The silhouette coefficient s(i) is given by:

$$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))}$$

We determine the average silhouette for each value of k and for the value of k which has the **maximum value of s(i)** is considered the optimal number of clusters for the unsupervised learning algorithm.

Explain the necessity for scaling/standardisation before performing Clustering:

Standardizing data is recommended because otherwise the range of values in each feature will act as a weight when determining how to cluster data, which is typically undesired. If one of your features has a range of values much larger than the others, clustering will be completely dominated by that one feature. However when the data is standardized this no longer becomes an issue and weights each feature as being equal when calculating the distance between each data point.

Explain the different linkages used in Hierarchical Clustering.

Single-Linkage

Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.

Complete-Linkage

Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average-linkage, it is one of the more popular distance metrics.

Average-Linkage

Average-linkage is where the distance between each pair of observations in each cluster are added up and divided by the number of pairs to get an average inter-cluster distance. Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

Centroid-Linkage

Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram. This problem doesn’t arise in the other linkage methods because the clusters being merged will always be more similar to themselves than to the new larger cluster.