CLUSTERING ASSIGNMENT PART 2

Question 1: Assignment Summary

Help international is an international NGO that wants to fight poverty and help the countries in need by donating a part of 10-million-dollar fund available to them. To do so, they first had to decide which countries need the donation the most. For this purpose, we had to cluster the countries and find the cluster that has countries in the at most need of the funding. To do so, first I proceeded with EDA to find out the bottom countries for each of the factors given and I stored the countries that appeared in more than one factor. Next, I removed the outliers, by capping only the top countries from GDPP, Income, Import and Export and not the bottom ones as they are the subject of interest for us. It did solve most of the outlier problem. Another issue was that the variables were highly correlated to each other and hence I found PCA to be an appropriate solution. By screeplot, I decided that 5 components explain 99% of the variation in the data and hence are appropriate number of clusters. Then after standardising the data, I calculated the Hopkins statistics to see whether or not there is a tendency of clustering in the data. I did that 10 times to make sure the results are reliable and got none of the value less than 0.7 and hence concluded that the data has high tendency to cluster. After then, it was a task to decide how many clusters to choose, by performing Silhouette score and SSD analysis and chose 3 clusters to be appropriate for the for the analysis and hence performed K-Means analysis on the data to get 43 countries. I then performed hierarchical clustering (both single and complete linkage, but the dendogram for single linkage was inconclusive) with four cluster to get 47 countries. 39 of the countries were common in both. I then deleted the countries that had more than median value for GDPP and Income and less than median value for child mortality to get the final 5 countries: -

- 1. Congo, Dem. Rep
- 2. Niger
- 3. Sierra Leone
- 4. Mozambique
- 5. Central African Republic

Question 2: Clustering

a) Compare and contrast K-means Clustering and Hierarchical Clustering.

k-means Clustering	**Hierarchical Clustering**
k-means, using a pre-specified number of clusters, the method assigns records to each cluster to find the mutually exclusive cluster of spherical shape based on distance.	Hierarchical methods can be either divisive or agglomerative.
K Means clustering needed advance knowledge of K i.e. no. of clusters one want to divide your data.	In hierarchical clustering one can stop at any number of clusters, one find appropriate by interpreting the dendrogram.
One can use median or mean as a cluster centre to represent each cluster.	Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
Methods used are normally less computationally intensive and are suited with very large datasets.	Divisive methods work in the opposite direction, beginning with one cluster that includes all the records and Hierarchical methods are especially useful when the target is to arrange the clusters into a natural hierarchy.
In K Means clustering, since one start with random choice of clusters, the results produced by running the algorithm many times may differ.	In Hierarchical Clustering, results are reproducible in Hierarchical clustering

b) Briefly explain the steps of the K-means clustering algorithm.

Algorithmic steps for k-means clustering

Let $X = \{x_1, x_2, x_3, \ldots\ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots\ldots, v_c\}$ be the set of centres.

1) Randomly select *'c'* cluster centres.

2) Calculate the distance between each data point and cluster centres.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centres..

4) Recalculate the new cluster center using:
where, *'c_i'* represents the number of data points in *i^{th}* cluster.

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

5) Recalculate the distance between each data point and new obtained cluster centres.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

The statistical methods are:

1. The Elbow Method
2. The Silhouette Method

Elbow Method

This is probably the most well-known method for determining the optimal number of clusters. *It is also a bit naive in its approach.*

*Calculate the **Within-Cluster-Sum of Squared** Errors (WSS) for **different values of k**, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an **elbow.***

Within-Cluster-Sum of Squared Errors is:

- The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster centre.
- The WSS score is the sum of these Squared Errors for all the points.
- Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.

The Silhouette Method

The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). The range of the Silhouette value is between +1 and -1. A **high value is desirable** and indicates that the point is placed in the correct

cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.

Business aspect

It is also important to not entirely depend on the statistical methods but also combine the business requirements to make the decision about the number of clusters.

If your using your data for splitting things into categories, try to imagine how many categories you want first. If it's for data visualization, make it configurable, so people can see both the large clusters and the smaller ones.

If you need to automate it, you might want to add a penalty to increasing k, and calculate the optimal cluster that way. And then you just weight k depending on whether you want a ton of clusters or you want very few.

d) Explain the necessity for scaling/standardisation before performing Clustering.

Standardization is an important step of Data pre-processing. it controls the variability of the dataset, it converts data into specific range using a linear transformation which generate good quality clusters and improve the accuracy of clustering algorithms.

K-means clustering is "isotropic" in all directions of space and therefore tends to produce more or less round (rather than elongated) clusters. In this situation leaving variances unequal is equivalent to putting more weight on variables with smaller variance.

e) Explain the different linkages used in Hierarchical Clustering.

The process of Hierarchical Clustering involves either clustering sub-clusters (data points in the first iteration) into larger clusters in a bottom-up manner or dividing a larger cluster into smaller sub-clusters in a top-down manner. During both the types of hierarchical clustering, the distance between two sub-clusters needs to be computed. The different types of linkages describe the different approaches to measure the distance between two sub-clusters of data points. The different types of linkages are: -

1. **Single Linkage:** For two clusters R and S, the single linkage returns the minimum distance between two points i and j such that i belongs to R and j belongs to S.

2. **Complete Linkage:** For two clusters R and S, the single linkage returns the maximum distance between two points i and j such that i belongs to R and j belongs to S.

3. **Average Linkage:** For two clusters R and S, first for the distance between any data-point i in R and any data-point j in S and then the arithmetic mean of these distances are calculated. Average Linkage returns this value of the arithmetic mean.