Clustering Assignment

—Abhay Sahani

1. Assignment Summary

As per the problem statement, I was required to obtain a list of at least top 5 countries that were in dire need of aid. To aid this analysis, I was given a dataset of 167 countries which had data regarding various parameters such as GDPP, Income, Child Mortality, etc. First and foremost, data was checked for any duplicates, if it required any imputation; post which, univariate and bivariate analysis was performed from the point of view of a.) accounting for outliers in the features, b.) using bivariate analysis, looking at how variables such as GDPP; Child Mortality; and income interact with each other. What seemed to be a crucial step was to have an outlier treatment which could ensure that all data points were considered and not a single country has been left out from the analysis. To this end, values were capped and floored respectively at 99%ile and 1%ile. This was then followed by scaling of the data since clustering is based on Euclidean distance and features with a higher scale would be given more weightage by the algorithm. Hopkins statistic was also calculated to ensure that data could be clustered. Following scaling of data, both K-Means and Hierarchical clustering was undertaken to look at which is the worst performing cluster in terms of the chosen parameters by the business (GDPP, Income, Child Mortality). These clusters were subsequently visualised to aid our analysis of the data. In this particular case, both the types of clustering produced same list of countries that needed aid which were **Solomon Islands; Eritrea; Madagascar; Rwanda;** and **Kenya**.

2. Compare and contrast K-means clustering and Hierarchical Clustering.

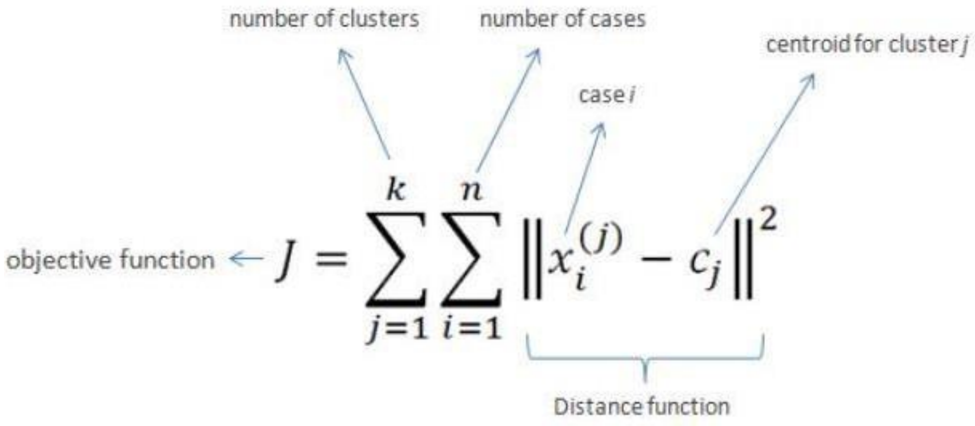
K-Means clustering algorithm uses a pre-specified number of clusters as a method of cluster analysis. On the contrary, Hierarchical Clustering seeks to build a hierarchy of clusters (either agglomerative or divisively) as a method for cluster analysis without having a fixed number of clusters.

K-means Clustering	Hierarchical Clustering
K-means using pre-specified number of clusters, the method assigns observation to each cluster to find the mutually exclusive cluster based on Euclidian/Manhattan distance.	Hierarchical methods can be either divisive or agglomerative.
To execute K-means, we need to provide the number of clusters we would want the dataset to be clustered into.	Basis the interpretation of dendrogram, one can stop at any number of clusters.
Cluster Centroid represents the cluster centres which is based on either median or mean.	Hierarchical clustering method can form clusters using either Agglomerative method (top-down) which begins with 'n' clusters and sequentially combine similar clusters until one cluster is obtained. Divisive method (bottom-up) works in the opposite direction.
The results produced by the algorithm might differ since one starts with selection of random clusters.	Hierarchical method of clustering yields reproducible results.
To ensure that each data object is exactly in just one subset, K-means clustering method divides data into non-overlapping subsets.	It produces dendrograms which is nothing but nested clusters which look like a tree.

K-means clustering works well when the structure of the clusters is hyper spherical (like circle in 2D, sphere in 3D).	Hierarchical cluster does not work well when the shape of the cluster is hyper spherical.
Advantages: Eventual convergence (Albeit setting max_iter at a large number).	Advantages: Ease of handling of any forms of similarity or distance.
Disadvantages: Ideal K-value can be difficult to ascertain right at the beginning of analysis.	Disadvantages: For large datasets, this can be expensive and slow since Hierarchical clustering requires computation and storage of a nxn distance matrix.

3. **Briefly explain the steps of the K-means clustering algorithm**

K-means clustering intends to subset n objects into k clusters in which each object belongs to cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinctions. The system does not have the knowledge of the best n clusters and hence it has to be provided by the user and the effectiveness of clustering has to be then further evaluated. The objective of the K-Means clustering is to maximise intra-cluster homogeneity and maximise inter-cluster heterogeneity.



$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

Algorithm:

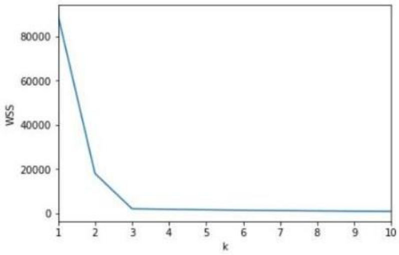
a.) Cluster the datapoints into k groups where k is predefined.
b.) Select k points at random as cluster centres.
c.) Assign objects to their closest cluster centre according to the Euclidean distance formula.
d.) Calculate the centroid or mean of the coordinates of all datapoints in each cluster.
e.) Repeat steps 2, 3, and 4 until the same points are assigned to each cluster in subsequent iterations.

There is a need to specify the number of clusters at the beginning of running K-Means clustering. To be able to figure out optimum value of K, it is prudent to run the model with different values of K since K-means clustering is sensitive to initialisation of clusters at the beginning.

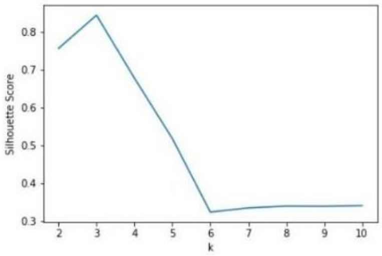
4. How are the value of ‘k’ chosen in K-means clustering? Explain both statistical and business aspect of it.

a.) Statistical Aspect: From the statistical aspect, we look at the primarily two methods of determining optimal number of clusters:

a. The Elbow Method: In this method, we calculate Within-Cluster Sum of Squared Errors for different values of k, and the choose the k for which WSS first starts to diminish. This phenomenon is visible as an elbow. Here, from this graph, 3 is the appropriate number of clusters.



b. Silhouette Method: This method measures how similar a point is to its own cluster (cohesion) compared to points in other neighbouring clusters (separation). Here, basically we look at the Global Maxima. The range of silhouette value varies between +1 and -1. Here, we see 3 clusters are optimum.



b.) Business Aspect: Here simply we are looking at the optimum number of clusters from the point of view of feasibility of time and resources to work upon the clusters. For example, a company that wants to execute a new marketing campaign has enough resources and time to make customised pitches for maybe three unique segments (i.e., 3 clusters) even though during our analysis, we found that optimum number of clusters to be at 5. But the business suggests to still go with 3 that is work on segments which might have overlapping elements since they do not have the bandwidth to work on 5 unique clusters.

5. Explain the necessity for scaling/standardisation before performing clustering.

Being a distance-based algorithm, K-means is affected by the scale of the variables. For example, assuming data has age variable which tells about the age of the person in years and an income variable which tells the monthly income of the person in INR:

ID	Age	Income (INR)
1	20	90,000
2	25	80,000
3	45	70,000
4	35	50,000

5	40	120,000

Here, the Age of the person ranges from 20 to 45 whereas the income variable ranges from 50,000 to 120,000. If we were to calculate the Euclidian distance between observation 1 and 2, i.e., Euclidean Distance = [((80000-90000)^2 + (25-20)^2)^0.5], it comes out to around 10,000.00124 . The high magnitude and overall scale of income affected the distance between the two points. The higher scale would force the algorithm to give more weightage to income which would create a bias. To keep the biasness at bay, we standardise the variables to ensure variables are at the same scale.

Normalization is the most commonly used method to scale the features which calculates the z score of all the observation. The other commonly used method is min-max scaling.

If we were to perform normalisation on the above dataset, it would something like this:

ID	Age	Income (INR)
1	-1.401826	0.345547
2	-0.862662	-0.086387
3	1.293993	-0.518321
4	0.215666	-1.382189
5	0.754829	1.641350

If we were to again calculate the Euclidian Distance between 1st observation and 2nd observation, the Euclidean Distance [((-0.086387-0.345547)^2 + (-0.862662+1.401826)^2)^0.5] would come out to be 0.69084.

Here we can observe that the formula is not biased towards income variable which is what we expect from K-means algorithm.

6. Explain different linkages used in Hierarchical Clustering.

a. Single Linkage: It is based on clustering points on a bottom-up fashion, i.e., at each step combining two clusters which contain the closest pair of clusters yet not belonging to the same clusters. These clusters can appear to spread out and hence this method is not extremely useful in dividing data into distinct classes.
b. Complete Linkage: Complete Linkage is where the distance measured is the longest distance measured between two points. This method usually produces tighter clusters than single linkage, but these tight clusters can end up close together.
c. Average Linkage: It is the average distance between each point in one cluster to every point in the other cluster.