Clustering Assignment-Part II (K-Means + Hierarchical)

Submitted By : Snigdha Chakraborty

Question 1 – Assignment Summary

Q1. Briefly describe the "Clustering of Countries" assignment within 200-300 words

Problem Statement – As an analyst we need to categorize the countries using some socio economic and health factors and report back 5 countries, which are in dire need of aid, to the CEO of HELP International NGO

Technical Approach – Below are the list of steps performed to analyze the data and conclude the outcome:-

1. We started with business understanding and analyzing the problem statement followed by Data Understanding
2. During Data Cleaning we observed that there are no missing columns/rows and no duplicate data., hence moved to next step of EDA (Exploratory Data Analysis)
3. In EDA we analyzed the spread of data using various plot and statistics. We also did :-
    1. Univariate Analysis by plotting Bar graph of countries vs features. We concluded that most of the African Countries has lowest gdp, income, health, life expectancy and highest child mortality rate
    2. Bivariate Analysis by plotting pair plot and heat map. We found that imports – exports, child mortality rate – life expectance/total fertility are highly correlated variable.
4. Post that we moved ahead to Data preparation. Below steps were considered:
    1. Outlier Analysis and Treatment – Almost all the variables had outliers. Removing outlier is not chosen as it will shrink the size of the existing data and result into improper analysis. Hence, we did soft capping with 1-99 percentile.

Q1. Contd ..

1. Data Scaling – Scaled the date (except country) using standardized scaler
2. Validated Hopkins Statistics of the scaled data = 0.89, which indicates the data is good for cluster analysis

5. Model Building – As per elbow curve and silhouette score analysis we took 3 as optimal cluster and performed K-means clustering. We did validate the result by Hierarchical Clustering (Complete Linkage) and the results were in-line with previous outcome of k-means
   A. The time complexity of K Means is linear = O(n) while that of hierarchical clustering is quadratic = O(n2). So, considering the time complexity we opted to go with the clusters formed by K-means
6. Cluster Profiling – Based on statistics and economic factors the 3 clusters were identified as Under-developed, developed and developing countries
7. Under-developed countries were the one with low GDP, low Income and high child mortality rate and these made the cluster as best for our outcome
8. Conclusion – We filtered the list of countries from under-developed cluster as per above sequence and retrieved top 5 countries in need of aid - **Sierra Leone, Haiti, Chad, Central African Republic, Mali**

Question 2 – Clustering

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

SL.No	K-means Clustering	Hierarchical Clustering
1	K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. It is performed using pre-specified number of clusters	Hierarchical cluster analysis is an unsupervised clustering algorithm which seeks to build a hierarchy of clusters. It can be either divisive or agglomerative.
2	Optimal Number of clusters can be determine using Elbow Curve/SSD and Silhouette Score	Optimal number of cluster is determined by analyzing the dendrogram and cutting at appropriate stage
3	In K Means clustering, since one start with random choice of clusters, the results is produced by running the algorithm multiple times until convergence	In Hierarchical, Agglomerative is performed using Bottom-Up and Divisive is performed Top-Down approach.
4	Convergence depends on Choice of initial cluster centroids, number of repetition and impacts of outliers	The dendrogram depends on types of linkages single, complete, average. Mostly, complete and average linkage is widely used it produces clusters which have a proper tree-like structure
5	The time complexity of K Means is linear = O(n)	The time complexity of Hierarchical is quadratic = O(n2)
6	K-Value is difficult to predict and don't work well with global cluster	Hierarchical clustering requires the computation and storage of an n×n distance matrix. For very large datasets, this can be expensive and slow

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

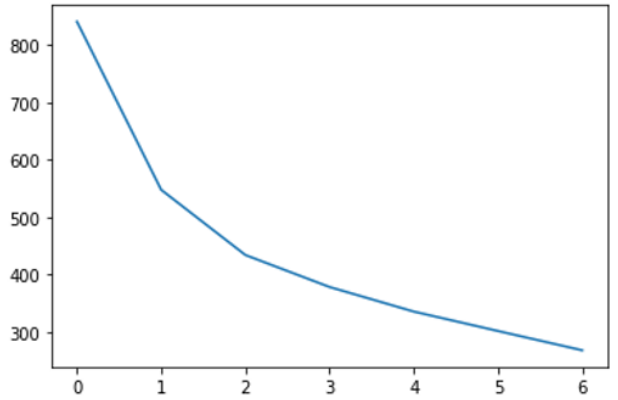
The steps of K-Means Clustering algorithm is mentioned below :-

1. Select initial centroids. The input regarding the number of centroids is given by the user at random as cluster centers
2. Calculate the Euclidean distance between each data point and cluster centers
3. Assign the data points to the closest centroid according to the Euclidean distance function
4. Recalculate the centroid for each cluster as mean of assigned observations
5. Recalculate the distance between each data point and new obtained cluster centers and assign the data objects again
6. Follow the same procedure until convergence. Convergence is achieved when there is no more assignment of data objects from one cluster to another, or when there is no change in the centroid of clusters

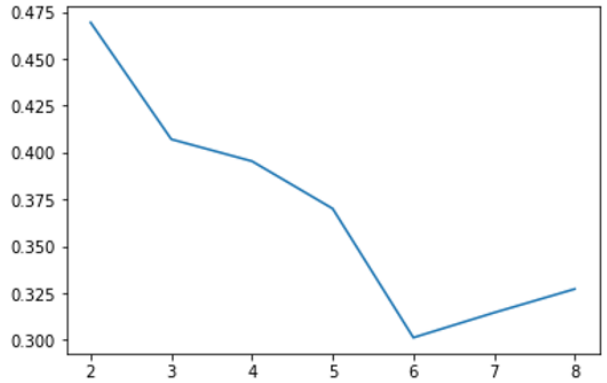
c) How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. It is performed using pre-specified number of clusters i.e. K. Domain and business understanding also helps to determine logical value of K based on various factors. Such as in countries assignment global standard of dividing country also helped us to come to a value in-line with statistical result. Statistically there are two ways to determine the optimal number of clusters:-

1. Elbow Curve/SSD : Calculate the Within-Cluster-Sum of Squared Errors (WSS) by varying k from 1 to 10 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. The point where this distortion declines the most is the elbow point.



2. 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.



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

The process of standardization is used to normalize the data by doing feature scaling. In this process we rescale the values of the data variables so that they share a common scale. Mainly, we perform data scaling when each variable has a different unit or the scales of each of variables are very different from one another.

The reason scaling is important in cluster analysis is because groups are defined based on the distance between points, so large variation may end up being the primary driver of clusters definition. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance. There are two ways we can perform standardization:-

1. **Min-Max Scaling** – It is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]

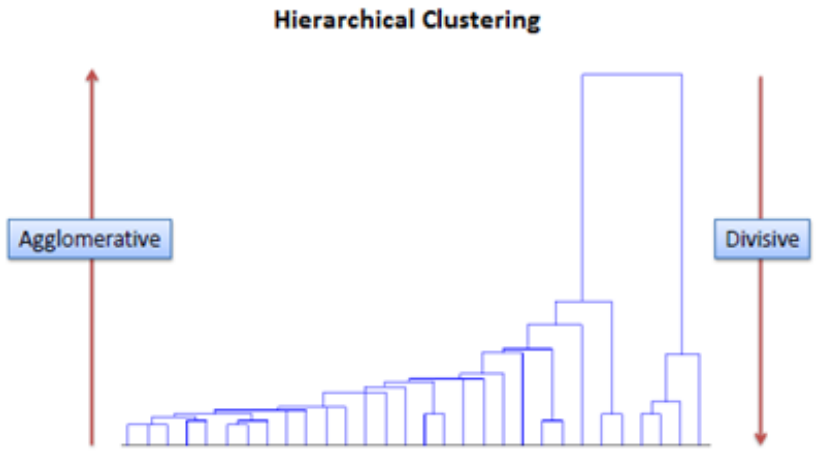
$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

2. **Standardization** – Also known as Z-Score normalization. It makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance.

$$z = \frac{x_i - \mu}{\sigma}$$

e) Explain the different linkages used in Hierarchical Clustering.

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. There are two types of hierarchical clustering, Divisive (Top-Down) and Agglomerative (Bottom-Up). It results in an inverted tree-shaped structure, called the dendrogram. Linkage can be defined as the pairwise distances between the data points. There are mainly 3 types of linkages:-



1. Single Linkage - Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters
2. Complete Linkage - Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
3. Average Linkage - Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster

Usually single linkage produces non structured dendrogram where as complete/average linkage produces clusters with proper tree like structures and hence is more convenient in clustering