Clustering Assignment-Part II

Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly (what EDA you performed, which type of Clustering produced a better result and so on)

ANS:

Problem Statement

HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

Solution

The following 5 steps were performed, to use clustering and arrive at 5 countries that would be in direst need.

1. EDA: In the EDA step we had analyzed the available data and figured that some columns have derived values, such as imports/exports/health which are mentioned as percentages of GDPP. Hence, we again derived the actual values for these fields and stored in *_total columns for further use.
2. Outliers Treatment: The second step was to verify presence of outliers. The GDPP related fields had outliers at the upper boundary corresponding to values for developed nations. We could have removed the developed nations from the dataset, but we kept them to have a better set for evaluation. Instead we capped the very high values to have the data points packed into a cluster.
3. K-Means Clustering: We then moved to run K-Means clustering on this dataset. Using Elbow method and Silhouette score, we determined an optimized number for K. The silhouette score provides a measure on how similar a point to its own cluster is and how dissimilar it is to neighboring clusters. A higher Silhouette Score means object is well matched to its own cluster. The K was determined to be 3.
4. Hierarchical Clustering: A hierarchical clustering was run on the dataset to determine optimal number of clusters available. The single linkage method did not yield a clearly defined cluster set. However Complete linkage did give us 3-5 clusters on the dendrogram.
5. Using both K-Mean and Hierarchical Clustering we found that the data points can be divided into 3 clusters, based on GDPP and Child Mortality. This gave us the insight to target the cluster with High Child Mortality and Low GDPP to find our set of countries.
6. Finally, we selected 5 countries, by further filtering the Cluster to include countries with Child Mortality rate of above 100, and GDPP and Income less than 1000 and 1500 respectively. A simple sort on this list gave us the top 5 candidate countries for the programs.

Question 2: Clustering

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

ANS:

K-Means Clustering	Hierarchical Clustering
Number of clusters are determined before a K-Means process is used	Number of clusters need not be pre-determined.
The output is always the number of clusters input.	The clusters are generated in hierarchical order, so that a single cluster is recursively broken down until each data point is in its own cluster.
Based on where the initial centroid is placed, the output can vary if run multiple times	This method if un multiple types will provide the same output.
The iterations to realign clusters can be capped at a certain number.	The hierarchical clustering always runs for the entire dataset.
The cost of running the method is low	The cost is generally high.

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

ANS:

The steps for run K-means cluster:

- Determine an optimal number of K (output clusters)
- Randomly place k centroids as far away from each other.
- Each point in the dataset is then assigned to the nearest centroid. This is done using Euclidean distance.
- Once all clusters are identified, each cluster is assigned a new centroid based on the Mean of the data points in a cluster, again using Euclidean distance.
- With the new centroids, all points are checked to see if any of the points should now be moved to a different cluster based on distance.
- If they are nearer to a different cluster centroid, the data points are assigned to the new cluster.
- This process is carried out recursively until, no data points are found that needs a reassignment of cluster or the maximum number of iterations are over.
- At that point, the clusters assigned to the datapoints are the clusters they belong to.

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

ANS:

The value of K in K-Means Clustering is chosen by using 2 methods.

Elbow Curve: For this K-Means clustering is run for 2-N number of clusters. Then the Sum of Separated Errors is plotted for each K. A higher K is preferred if the SSE is bettered considerably by moving to net K. If the reduction is not much (i.e. the curve straightens out) then the current K is selected.

Silhouette Score: The Silhouette Score gives a measure of the similarity of a data point with its own cluster and the dissimilarity of the data point from the neighboring cluster. The value ranges from -1 to 1, where a higher value means better similarity with its own cluster.

Elbow Curve in conjunction with Silhouette Score can be used to obtain an optimal K. A value of K that gives better reduction of SSE without losing on Silhouette Score is chosen as an optimal K.

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

ANS:

In a dataset, the features do not always have the same units/base. Some features are always measured in higher units say KMs, whereas some in higher values say Milliseconds. A high value in one feature for example Milliseconds may be mistaken for higher weightage to the said feature. This would make the model unstable. We already know this behavior creates an issue with Linear Regression model where higher values would end up with low beta. Since we use Euclidean distance in K-Means the problem remains here too.

Hence, it is required to Standardize or Normalize the data to put them in the same scale before using them in a model. This removes variability because of the unit size.

e) Explain the different linkages used in Hierarchical Clustering.

ANS:

There are 3 types of linkages used in Hierarchical Clustering.

- Single Linkage: The distance between two clusters is calculated as the shortest distance between any two points from each cluster.
- Average Linkage: The distance between two clusters is calculated as the average distance between each point of the cluster to each point of the other cluster.
- Complete Linkage: The distance between two clusters is calculated as the longest distance between any two points from each cluster.

Submitted By

Soumya Prakash Parida