Clustering Assignment

Assignment Part II

Question 1: Assignment Summary

Problem Statement: HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people pf 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.

After the recent funding programmes, they have been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use the money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

And as a data analyst, my job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then make a list of suggestions of the countries which the CEO need to focus on the most.

Assignment Summary

Here we have a dataset of 167 countries with their corresponding socio-economic and health factors like child mortality rate, income, GDPP, etc. Unsupervised ML is used here. The countries have been segregated into different groups using Clustering. During the EDA, we found that the features are right-skewed which means that it contains outliers. But removing the outliers is not a feasible solution which causes a huge loss of data, so we used Power Transformer to handle the skewness and scaling. After pre-processing the data, we used the Elbow Curve and Hopkins Statistic to find the optimal number of clusters and to check whether the dataset is good enough for cluster analysis. Here we found 3 as the optimal clusters to be formed. Further used K-Means Clustering and Hierarchical Clustering (Both Single linkage and Complete linkage) to model and predict the countries. After plotting the scatter plots and box plots, we found that K-Means Clustering was a better approach to take for this particular dataset. And finally a list has been made of the top 20 countries in direst need. This list has been made considering high child mortality, low GDPP and low Income.

The Top 20 Recommended countries which are in direst need of aid:

1) Burundi
2) Liberia
3) Congo, Dem. Rep.

4) Niger
5) Sierra Leone
6) Madagascar
7) Mozambique
8) Central African Republic
9) Malawi
10) Eritrea
11) Togo
12) Guinea- Bissau
13) Afghanistan
14) Gambia
15) Rwanda
16) Burkina Faso
17) Nepal
18) Uganda
19) Guinea
20) Haiti

Q&A

1) Compare and contrast K-Means Clustering and Hierarchical Clustering.

K-Means	**Hierarchical**
Uses pre-defined number of clusters.	Does not require pre-defined or fixed number of clusters.
One can use median or mean as a cluster center to represent each cluster.	Agglomerative methods begin with ‘n’ clusters and sequentially combine similar clusters until one cluster is formed.
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.
Here, since one start random choice of clusters, the results produced by running the algorithm many times may differ.	Here, results are reproducible.

It is simply a division of the set of data objects into non-overlapping subsets (clusters).	It is a set of nested clusters that are arranged as a tree.
Here complexity is linear.	Here complexity is quadratic.

2) Briefly explain the steps of the K-Means clustering algorithm.

The K-Means clustering algorithm is a simple and easy clustering algorithm which is iterative and can be explained with the following steps:

a) First, we classify the number of clusters.
b) Next, we select randomly K data points and assign each data point to a cluster. Tis is called data classification.
c) Next is computing the cluster centroids. A centroid’s value is going to be the mean of all data in a cluster.

d) Next, keep iterating and following the step b and c. This will get the K-Means algorithm converged.

e) Finally consider the best clustering model from your derived iterative model.

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

There is a popular method known as elbow method which is used to determine the optimal value of ‘k’ to perform K-Means Clustering. The basic idea behind this is that it plots the various cost with changing k. As the value of k increases, number of elements get fewer in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where the distortion declines the most is the elbow point.

Next, we have another method called the ‘Silhouette Method’. The silhouette value measures how similar a point is to its own cluster compared to other clusters. 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.

Both the above methods give us a statistical aspect of selecting optimal number of clusters.

For the business aspect, it is up to the Business teams to decide the number of clusters. For example, in customer segmentation, marketing team may decide upon previous data to decide upon the number of customer segmentation.

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

In statistics, standardization (feature scaling) refers to the process of rescaling the values of the variables in your data set so they share a common scale. Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit, or where the scales of each of your particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space.

When you are working with data where each variable means something different, the fields are not directly comparable. One year is not equivalent to one pound, and may or may not have the same level of importance in sorting a group of records. In a situation where the field has a much greater range of value than another, it may end up being the primary driver of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

When we standardize the data prior to performing cluster analysis, the clusters change. We find that with more equal scales, the percent native American variable more significantly contributes to defining the clusters. Standardizing prevents variables with larger scales from dominating how clusters are defined. There is a popular method called elbow method which is used to determine the optimal value of K to perform K-Means clustering. 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.

Standardization also helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

Although it is considered best practice for cluster analysis, there are circumstances where standardization may not be appropriate for your data.

5) Explain the different linkages used in Hierarchical clustering.

The process of Hierarchical clustering involves either clustering sub-clusters 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.

It involves creating clusters that have predetermined ordering from top to bottom.

There are two types of hierarchical clustering:

Divisive and Agglomerative.

The different types of linkages describe the different approaches to measure the distance between two sub-clusters of data points.

They are:-

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 the distance is measured between farthest pair of observations in two clusters. This method produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with the average linkage, it is one of the most 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-linked and complete-linkage are the two most popular distance metrics.

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 new larger cluster.