Saikat Chowdhury

Clustering Assignment

Segmentation of countries for humanitarian ADD by an NGO using K-Means & Hierarchical Clustering.

Question 1: Assignment Summary

Answer:

Problem Statement: As a Data Scientist, we need to find the countries in direst need and help CEO of HELP International in using the fund money to reach right countries.

We used following technical approaches to get the list of countries which has really required/needed financial aid by NGO. Steps are below.

Solution Approach:

As we have the Data of countries like child mortality rate, GDP Per Capita, Income etc., we can use Clustering to segregate the countries into different groups. In the data provided, all the features are right-skewed which indicates us that it contains Outliers. As removing Outliers is not a feasible solution as we will lose data. We used Power Transformation to handle the skewness and also scaling.

1. Check for missing value, and treatment.

There was no missing values in the dataset so there was no needed to impute with any values. There are 167 rows and 10 columns in dataframe.

2. Check for outlier and treatment

There was outliers in the data in which I handled with upper and lower values (.05 and .95 – standard method). Dataset has no duplicate as well.

3. Perform the basic EDA to find the variability and distribution of the data, so as to identify if we need scaling the data.

Most of the data point are 'NOT Normally' distributed. Their variance are also differernt. Their range are also differnt All the above points indicates the need of standardising the data before we build the model. Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale is important here.

4. Data Scaling

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

5. Use Hopkins Method to check if the dataset is good enough for a cluster analysis

Before we apply any clustering algorithm to the given data, it's important to check whether the given data has some meaningful clusters or not? which in general means the given data is not random. The process to evaluate the data to check if the data is feasible for clustering or not is know as the clustering tendency. To check cluster tendency, we use Hopkins test. Hopkins test examines whether data points differ significantly from uniformly distributed data in the multidimensional space.

6. Using Hierarchical clustering to identify the optimal cluster value.

As mentioned in the 'Approach' section, we will use Hierarchical Clustering to identify appropriate cluster size with a good split of data (Max Intra-Cluster distance & Min Inter-Cluster Distance).

7. Use Silhouette and Elbow method to validate the optimal cluster values.

p is the mean distance to the points in the nearest cluster that the data point is not a part of q is the mean intra-cluster distance to all the points in its own cluster. The value of the silhouette score range lies between -1 to 1. A score closer to 1 indicates that the data point is very similar to other data points in the cluster, A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

8. Use K-Means Cluster method to build the final cluster model.

From the above 3 Iterations of K-Means, we could see that using 3 Clusters provided a better output in terms of a balanced cluster size. So we will consider the 'K-Means with 3 Clusters' as our FINAL MODEL.

9. Present the final report.

Recommended the name of top 5 countries in which they required financial aid by NGO.

10. Top 5 countries.

Top 10 Recommended countries which are in dire need of funds (Top 5 marked as bold):

- **Burundi**
- **Liberia**
- **Congo, Dem. Rep.**
- **Niger**
- **Sierra Leone**
- Madagascar
- Mozambique
- Central African Republic
- Malawi
- Eritrea

These countries have

- low GDPP
- low income and
- high child mortality

Question 2: Clustering

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

Answer:

K-Means	Hierarchical
a) **K- means clustering** a simply a division of the set of data objects into non- overlapping subsets (**clusters**) such that each data object is in exactly one subset).	a) A **hierarchical clustering** is a set of nested **clusters** that are arranged as a tree.
b) k-means is method of cluster analysis using a pre-specified no. of clusters. It requires advance knowledge of 'K'.	b) Hierarchical clustering also known as hierarchical cluster analysis (HCA) is also a method of cluster analysis which seeks to build a hierarchy of clusters without having fixed number of cluster.
c) 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.	c) Hierarchical methods can be either divisive or agglomerative.
d) One can use median or mean as a cluster centre to represent each cluster.	d) Agglomerative methods begin with 'n' clusters and sequentially combine similar clusters until only one cluster is obtained.
e) K-Means clustering can handle big data because the complexity is linear O(n) K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ.	e) Hierarchical cannot handle big data as the complexity is Quadratic i.e. O(n2) results are reproducible in Hierarchical clustering as it is up to us to decide number of clusters based on cutting the dendrogram.
f) Hierarchical clustering can't handle big data well.	f) but K Means clustering can.
g) While results are reproducible in Hierarchical clustering.	g) K Means is found to work well when the shape of the clusters is hyper spherical.
h) If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls.	h) KMeans produce tighter clusters than hierarchical clustering, especially if the clusters are globular

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

Answer:

We can understand the working of K-Means clustering algorithm with the help of following steps −

Step 1 − First, we need to specify the number of clusters, K, need to be generated by this algorithm.

Step 2 − Next, randomly select K data points and assign each data point to a cluster. In simple words, classify the data based on the number of data points.

Step 3 − Now it will compute the cluster centroids.

Step 4 − Next, keep iterating the following until we find optimal centroid which is the assignment of data points to the clusters that are not changing any more

- **4.1** − First, the sum of squared distance between data points and centroids would be computed.
- **4.2** − Now, we have to assign each data point to the cluster that is closer than other cluster (centroid).
- **4.3** − At last compute the centroids for the clusters by taking the average of all data points of that cluster.

K-means follows **Expectation-Maximization** approach to solve the problem. The Expectation-step is used for assigning the data points to the closest cluster and the Maximization-step is used for computing the centroid of each cluster.

While working with K-means algorithm we need to take care of the following things −

- While working with clustering algorithms including K-Means, it is recommended to standardize the data because such algorithms use distance-based measurement to determine the similarity between data points.
- Due to the iterative nature of K-Means and random initialization of centroids, K-Means may stick in a local optimum and may not converge to global optimum. That is why it is recommended to use different initializations of centroids.

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

Answer:

There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. 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. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.

We have another method called “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.
Both the above methods give us a Statistical aspect of selecting optimal number of clusters.

Sometimes, 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 segments.

The basic step of k-means clustering is simple. In the beginning we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids.

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

Answer:

In statistics, standardization (sometimes called data normalization or 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 (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0- 1000). The reason this importance is 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, (e.g., age and weight) 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 one field has a much greater range of value than another (because the field with the wider range of values likely has greater distances between values), 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. Standardization prevents variables with larger scales from dominating how clusters are defined. There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. 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 helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance. Although standardization is considered best practice for cluster analysis, there are circumstances where standardization may not be appropriate for your data (e.g., Latitude and Longitude). Whether to use Standardization or not, it depends on the data.

e) Explain the different linkages used in Hierarchical Clustering.

Answer**:**

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.

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. 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. The different types of linkages 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 distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together. Along with average linkage, it is one of the more 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-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

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 the new larger cluster.

THANK YOU