PG Diploma in Data Science Sept – 2020

Linear Regression Assignment – Subjective Questions

Name: Srihari R

Email: Sriharikrishna06@gmail.com

Assignment Summary

In this clustering of countries assignment, the main aim of this assignment is to find at least 5 countries which are in need of help from our client **HELP International** which 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.

The data is given from the client where it contains details of countries gdpp, income, child mortality rate, child fertility rate, imports, exports, life expectancy, health and much more. As a Data Analyst our role is to find the top countries which are in immediate need from our client. The imports, exports and health features are in % values of gdpp, which are first converted to absolute values

The data given is first checked for null values and outliers, there were no null values in the data given but we found many outliers in every feature of the data, therefore we removed data points which are upper quartile outliers in gdpp feature, which in turn removed most of the outliers from other features as well. Since, these are domain specific outliers, the remaining outliers are not treated since they might contain valuable information.

EDA is performed on the data where **univariate** analysis on life expectancy and health is made, and boxplots for all numerical columns to find outliers. **Bivariate** analysis is also performed on income vs. child mortality and income vs. total fertility

The data is then scaled using Standard scaling method, which makes it easier for the clustering algorithm.

To check for Cluster tendency we performed Hopkins test, which the score is from 0 to 1, where 0 means the data is not suitable for clustering and 1 means the data is perfect for clustering. To decide on the cluster number, we performed a series of tests using **silhouette score** and **elbow method,** which suggested that 3 is the good cluster number based on the data given.

In case of forming the clusters, we performed k-means and Hierarchical clustering (for both single and complete linkage) for cluster number 3 and 4 in order to choose the best model and cluster number. Kmeans for 3 clusters performed well in differentiating the data well for 3 clusters, when kmeans for 4 clusters where performed, only one single data point was clustered in the 4 cluster. In case of Hierarchical clustering, single linkage method performed very poorly and complete linkage formed clusters which did not differentiate clusters well from each other. Therefore, we choose K means as the final model and cluster number as 3. The clusters are checked by plotting line graph for gdpp, income and child mortality based on the clusters formed

The top 10 countries which require help are found by the countries from the cluster which has low gdpp and income and high child mortality, within this cluster we need 10 countries which need help; therefore, we sorted the countries where gdpp and income are sorted in ascending and child mortality is sorted in descending and retrieving only 10 countries from the sorted countries list

Subjective Questions

Question 1: Compare and contrast K-means Clustering and Hierarchical Clustering.

1. In K means clustering we need to decide on the cluster number before modelling but in hierarchical clustering, we can decide which cluster number based on the dendrogram formed.
2. We can use mean or median to know the cluster centroid for k means clustering whereas in hierarchical clustering, we use agglomerative methods which begins with n cluster (n = number of data points) and consecutively combine similar clusters until only one cluster remains
3. K means clustering is simply a division of set of data points based on the distance metric and cluster number used whereas hierarchical clustering is a set of nested clusters that are arranged as a tree

Question 2: Briefly explain the steps of the K-means clustering algorithm.

Step 1. Choose the K number based on a series of tests from elbow method and silhouette score and from client's requirement

Step 2. Select k random points or k points which are farthest from each other to avoid overlapping of clusters

Step 3. Assign the data points to the clusters based on distance metric, the least distance from the cluster is chosen for the cluster

Step 4. Recompute the centroid for the clusters using mean or median of the data point formed

Step 5. Repeat steps 3 and 4 until no new clusters are formed

Question 3: How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

The value of k is chosen based on the series of tests like elbow method and silhouette score, where elbow method is more of a decision rule and silhouette score is more of validation, therefore both are used in combination to choose the k number confidently before modelling. In case of business stand point, we can get help from the client and upon their requirement we can choose the k number, or we can get help from domain expert who has knowledge based on the problem statement.

Question 4: Explain the necessity for scaling/standardisation before performing Clustering

When the data is scaled from standardisation or normalisation before modelling which makes all the features under same scale as each other. If scaling is not performed then features with higher scale can easily dominate other features which has lower scales, forming entirely different clusters than required and the computational speed is also affected when there is no scaling. Entirely for this reason scaling is performed from standardisation or normalisation.

Question 5: Explain the different linkages used in Hierarchical Clustering

1. Single Linkage: Single-linkage 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.
2. Complete Linkage: Complete-linkage 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.

3. 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.
4. 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.