ASSIGNMENT-II

Clustering Assignment

Question 1: Assignment Summary

Answer:

Problem Statement: Given a dataset of all the countries and some health and socio-economic data for each of the countries, we have to analyze and recommend some countries based on their condition so that they can be provided with some aid by the international organization HELP International.

Solution Approach: After I have read the data, I have checked for missing values and found none. Also I have performed univariate analysis of the fields and found that some of the fields contain some high values in it. These values may impact the final results after analysis and hence those fields were capped to 99 percentile.

After preparing the data for model building, using Elbow Curve and Silhauette's Curve, I decided the number of clusters for the K-Means method and the number of clusters was decided as 3. Using K-Means I created 3 clusters and analyzed them. After analyzing the clusters by plotting box plots and scatter plots, I found that countries of one of the clusters have very high per capita GDP and per capita income and very low child mortality rate. This cluster was for the highly developed countries. Another cluster's countries have moderate GDP, income and child mortality rate as well. These countries are developing countries. The other cluster's countries have very low per capita income and GDP but the child mortality rate is very high. These are the underdeveloped countries and these countries are in direst need of some aid.

Also, I have performed Hierarchical lustering method (both single and complete linkage). In complete linkage hierarchical method, I got 3 clusters and the clusters could be differentiated according to the characteristics of the countries' health and socio-economic features in a similar way as the clusters I got from K-Means method. But I choose K-Means method over Hierarchical method because the clusters I got from Hierarchical method, the number of countries were not evenly distributed in them and also the intra-cluster similarities were not very prominent.

I took all the countries from the underdeveloped countries' cluster from K-Means method and again sorted the countries based on per capita GDP, per capita income and child mortality rate and finally recommended the worst 10 countries that should be provided with financial and other aid.

Question 2: Clustering

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

Answer:

The most important methods Unsupervised Learning is Clustering which can be performed in two different ways – 1) K-Means and 2) Hierarchical.

Both of the above mentioned methods can be used for clustering data that are not labeled from beforehand. But there are some differences based on usage, efficiency etc.

If the number of data is less, Hierarchical Clustering can be used and it will provide good results. Whereas, if the number of data is high enough, we cannot use Hierarchical Clustering since it will take much longer time and memory. So, for large number of data, both the time complexity [O(n) for K-Means and O(n2) for Hierarchical] and space complexity will be much higher in case of Hierarchical Clustering.

A disadvantage for K-Means Clustering is that we have to pre-define the number of clusters before applying K-Means algorithm by using different techniques like Elbow Curve, Silhauette's Curve etc. However, in case of Hierarchical Clustering, we do not have to specify the number of clusters beforehand.

Since we start K-means algorithm with arbitrary centroids, the final result may differ every time after running the algorithm multiple times. However, the results from Hierarchical clustering are reproducible.

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

Answer:

In K-Means algorithm, we have to choose the number of clusters before we start the algorithm. Once the number of clusters, K, is decided, we randomly choose K number of centroids.

Assignment step: After randomly choosing the centroids, we calculate the distance between every points with the all the clusters and assign each points to the cluster from which the distance is lowest. This way we assign each points to a cluster.

Optimization step: After assigning all points to a cluster, we calculate the mean of all the points within a cluster and update the centroid to that mean point. This is how the centroids of each cluster is optimized.

These two steps are continued alternatively until the centroids are stabilized. And finally after the centroids are stable, we get the desired 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.

Answer:

The value of K can be chosen by following 2 different methods in K-Means algorithm:-

1. Elbow Curve
2. Silhauette's Curve

In Elbow curve, the sum of square is calculated varying the number of clusters. Number of clusters is varied from 2 to 10 and then for each number of clusters the within-cluster sum of square is calculated and plotted against the number of clusters. In the plot, where there is a considerable bend (elbow) and after which there is not much change in slope, that number of cluster is considered to be final value of K.

In Silhauette's Curve, the number of clusters is varied from 2 to 10 and for each number of clusters the average Silhauette's observation is calculated and plotted against the number of clusters. The number of cluster for which the drop in the average Silhauette's observation is highest, is considered to be the final number of clusters.

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

Answer:

Feature scaling is a technique to standardize the values that vary a lot with a very wide range. It is performed during preprocessing of data to handle the wide variance of data or difference in units. If scaling is not done, the model may treat higher value small and smaller value high i.e. without standardization, a model interprets 1 kilometer is less than 100 meters. This way, the model will provide faulty results.

So, feature is scaling is required for two different reasons:

1. Ease of interpretation
2. Faster convergence for gradient decent methods

e) Explain the different linkages used in Hierarchical Clustering.

Answer:

In Hierarchical Clustering there are two different ways that the algorithm can be applied:

1. Single Linkage
2. Complete Linkage
3. Average Linkage

In Single Linkage, the distance between two clusters is defined as the shortest distance between points in that two clusters. Here, we merge the two clusters whose two closest members have the smallest distance (the two clusters with the smallest minimum pairwise distance) in each step. The time complexity of Single Linkage clustering is O (n^2).

In Complete Linkage, the distance between two clusters is defined as the maximum distance between points in that two clusters. Here, we merge the two clusters with the smallest maximum pairwise distance in each step. The time complexity of Complete Linkage Clustering is O (n^2 log (n)).

In Average Linkage, the distance between two clusters is defined as the average distance between each point of one cluster to the points of the other cluster.