Assignment: Part II

Question 1: Assignment Summary

I am working with HELP International NGO. After raising $ 10 million, the CEO of HELP needs to decide how to use this money strategically and effectively. The most significant issue is to which country do he need to extend humanitarian AID that are in dire need.

I have categorise the countries based on socio economic and health factors that determines the overall development of the country and then propose the list of countries to extend AID.

For this , first I have done data analysis and used 2 types of clustering namely K-Means and Hierarchical and compared the application and visualization of the two types of clustering.

From K-Means , I have found top 5 countries namely 'Burundi','Liberia','Congo Dem Rep','Niger' and 'Sierre Leone' which are in dire need of AID

From Hierarchical clustering, I have found only 1 country namely 'Luxembourg' which is in dire need of AID.

By comparing the two types of clustering, I have presented the CEO to go with the suggestion of 5 countries as mentioned in K-Means clustering as mentioned above

Question 2:Clustering

a) Compare and contrast K-Means Clustering and Hierarchical Clustering

Ans.

SNo	K-Means Clustering	Hierarchical Clustering
1	We need to find data point whose value are similar to each other. This method of finding something called as "Distance Measure" is known as Euclidean Distance	Here first we have to visually describe the similarity or dissimilarity between different points and then decide the appropriate number of clusters on the basis of these similarities or dissimilarities. Hierarchical methods can be either divisive or agglomerative
2	K-means divides the data in the first step itself. It subsequently refined the cluster to get most optimal grouping	Here a series of partition or merger takes place which may run from single cluster containing all objects to 'n' clusters that each contain a single object and vice-versa
3	One can use median or mean as a cluster center to represent each cluster	Agglomerative methods starts with 'n' clusters and subsequently one single cluster is formed by combining similar clusters

4	Here methods used are normally less computationally intensive and are suited with very large datasets	Hierarchical methods are especially useful when the target is to arrange the clusters into natural hierarchy
5	Since here one starts with random choice of clusters, the results produced by running the algorithm might differs many times	Here results are reproducible
6	It is set of data objects into non overlapping clusters such that each data object is in exactly one subset	It is a set of nested clusters that are arranged in a tree
7	Here convergence is guaranteed	It can handle any form of similarity or distance
8	K-value is difficult to predict and does not work well with global cluster	It requires computation and storage of nxn distance matrix which can be expensive for large dataset and can be slow also
9	There are 2 steps in K-means clustering a) Assignment Step b) Optimisation step	There are 2 steps in Hierarchical clustering a) Creating Dendogram b) Cutting the dendogram at an appropriate level

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

Ans.

Steps of K-means clustering algorithm are as follows:-

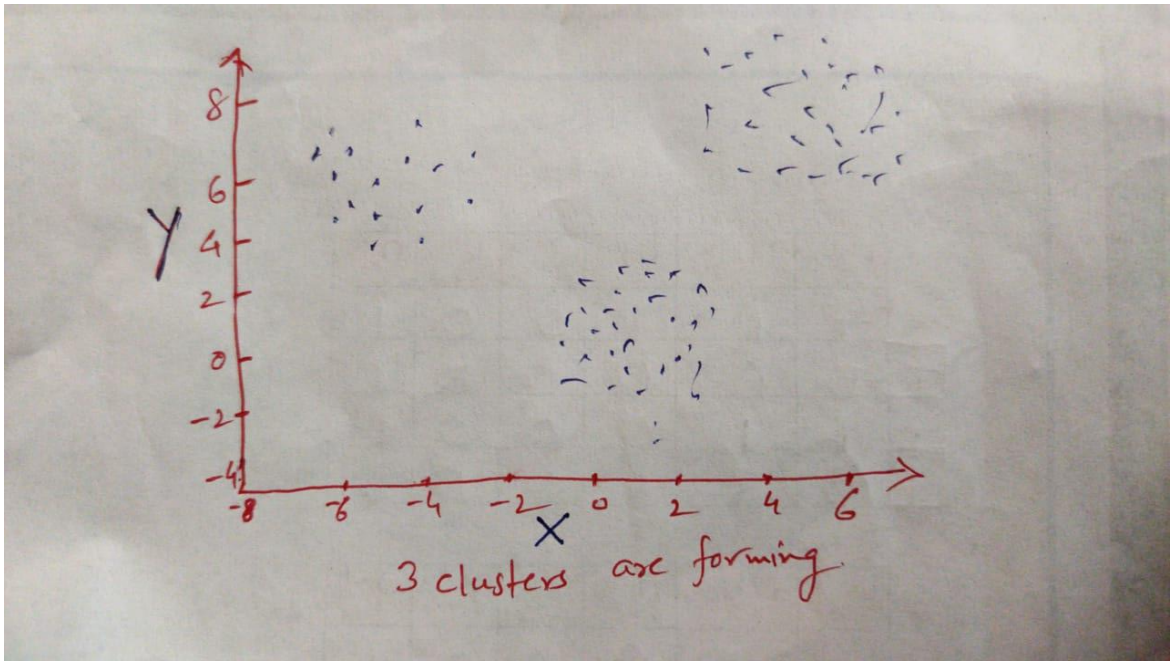
1. Select initial centroids. The input regarding number of centroids should be given by user
2. Assign the data points to the closest centroid
3. Recalculate the centroid for each cluster and assign the data objects again
4. Follow the same procedure until convergence. Convergence is achieved when there is no more assignment of data objects from one cluster to another or when there is no change in the centroid of 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.

Ans.

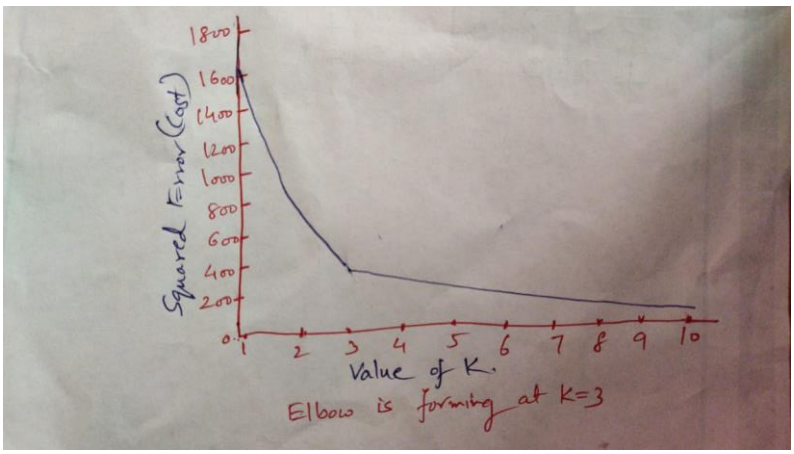
Elbow method is used to determine the value of 'k' to perform K-means clustering. It plots the various value of cost with changing k. As the k value increase, there will be fewer

element in the cluster. So the average distortion will decrease. The lesser of elements means closer to the centroid.



In the above figure, it is quite clear that distribution of points are forming 3 clusters.

Now let's see the plot for the squared error(cost) for different values of k



Clearly the elbow is forming at k=3. So the optimal value will be 3 for performing K-means.

Statistical Aspect :-

The K-means is one of the most popular and simple clustering algorithm that keeps data in the main memory. It is a well known algorithm for the efficiency in clustering large data sets and converges to acceptable results in different areas. The K-means algorithm with large number of variables is computationally faster than other algorithms.

Business Aspect :-

K-means clustering can be used to organize large set of retail data to generate competitive insights about the business which can help business to compete strategically in the retail market such as for Customer Segmentation, Delivery optimization, Document sorting and grouping, Customer retention, Discount analysis

d) Explain the necessity for scaling/standardization before performing clustering

Ans.

Scaling/Standardization refers to the process of rescaling the values of variables in the data set so that they share a common scale.

It is required as a pre processing step in clustering analysis where each variable has different unit(for e.g inches, meters, kilograms, etc) or where the scales of each of the variables are very different from one another (for e.g 0-1 or 0-1000). It is important in cluster analysis as groups are defined based on the distance between points.

When we are working with data where each variable means something different(for e.g age and weight, year and pound), the fields are not directly comparable. In such a situation where one field has much greater range of value than the other as they might have greater distance between values, it may end up being the primary driver of that defined clusters. So in such a situation standardization helps to make the relative weight of each variable to a unitless measure or relative distance.

When performing regression analysis, standardizing multi-scale variables reduces multicollinearity issues for models containing interaction terms

However as tree based analysis are not sensitive to outliers and do not require variable transformation, standardization of multiscale data is not necessary for Decision trees, Random Forest or Gradient boosting algorithm

e) Explain the different linkages used in Hierarchical Clustering?

Ans.

There are basically 3 types of different linkages in Hierarchical Clustering :-

1. Single Linkages – Here the distance between the two clusters is defined as the shortest distance between points in two clusters
2. Complete Linkages – Here the distance between the two clusters is defined as the maximum distance between points in two clusters
3. Average Linkages - Here the distance between the two clusters is defined as the average distance between every point of one cluster to every point of other cluster

Single linkage will produce Dendogram which are not structured properly where as complete or average linkage will produce clusters which have a proper tree like structure.