Assignment-II Clustering

Question 1 -> Summary of Clustering of Countries

Answer ->

Problem Statement: An International humanitarian NGO raised around $10 million in recent funding. The CEO needs to decide how to use this money and which countries to help. From the analysis we need to find out countries which are in direst need of financial aid.

The data we have, has details of 167 countries across the world. It has details like child mortality, value of total exports, imports, health expenditure, GDP, etc. I first converted the column values of exports imports and health from percentage of GDP per capita to unit values per capita by multiplying with value of GDP per capita and divide by 100. Plotted pair plot of all variables and saw that most of them are left skewed. I preprocessed it using PowerTransformer for better analysis which converted all values into logarithmic values. Plotting pair plot now gives clear understanding of variables dependency. In the next step scaled the data using standard scaler. Got the Hopkins score of around 0.85 which tells us that although data is not highly clustered but it is definitely not random and it supports clustering.

To begin with the modeling, plotted SSD curve which suggested k value of 2 or 3 is a good fit since SSD value after 3 does not decrease as large as from 2 to 3. I did silhouette analysis too just to be sure about the K value, got values of around 0.4 for K=2 and 3 and 0.3 after that.

Modelled data with KMeans algorithm choosing k=3, as per the pair plot between desired variables ('gdpp', 'child_mort', 'income') KMeans did a good job of separating the countries based on their parameter, tried KMeans with K=2, but that gives large number of countries in desired cluster (cluster with high child_mort, low income and low gdpp).

Tried Hierarchical Clustering too, single linkage failed to give any meaningful outcome but complete linkage separated out countries well too. Only difference was hierarchical clustering gave 67 countries in desired cluster and KMeans gave 50. As we have limited amount of raised fund and NGO wants to help only countries in direst need, I got the final list of countries from KMeans. From the 50 countries I got from cluster developed by KMeans, I choose top 10 based on their child_mort rate as it should be given importance over income and gdpp.

Question 2

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

Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 3 steps:

1. Specify the desired number of clusters K. Randomly assign each data point to a cluster.
2. Compute cluster centroids.
3. Re-assign each point to the closest cluster centroid.
4. Repeat step 2 & 3 until no improvements are possible.

Hierarchical clustering algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left. The results of hierarchical clustering can be shown using dendrogram. This algorithm also has top down approach starting with all data points assigned to same clusters.

Differences between both:

Time complexity of KMeans algorithm is linear whereas for hierarchical it is quadratic, therefor hierarchical clustering cannot be used for big data.

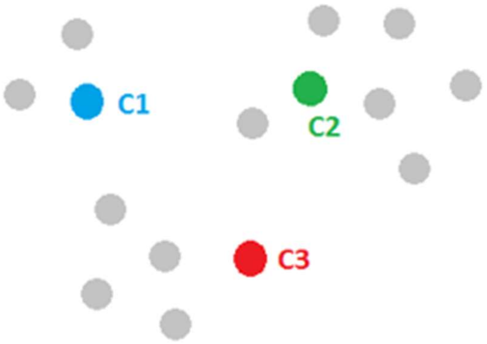
In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.

KMeans requires K to be known prior to running the algorithm whereas in hierarchical K can be decided by dendrogram.

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

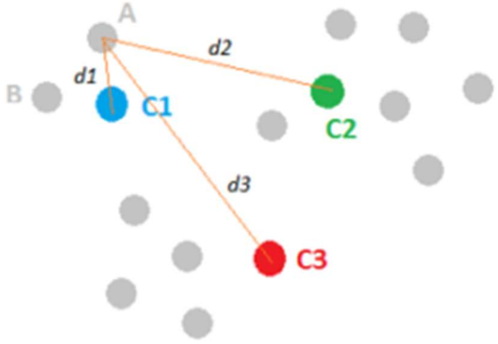
Below are the steps to be followed for KMeans Clustering:

1. We randomly pick three points C1, C2 and C3, and label them with blue, green and red colour separately to represent the cluster centres.

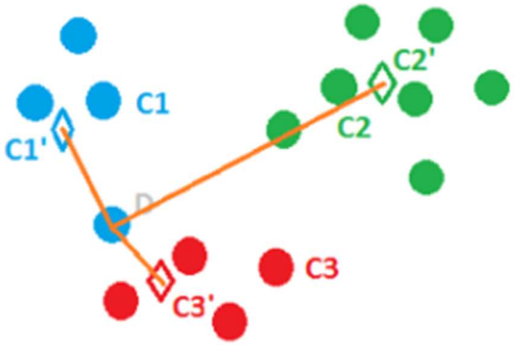


2. Assign all the data points to their closest cluster centres.

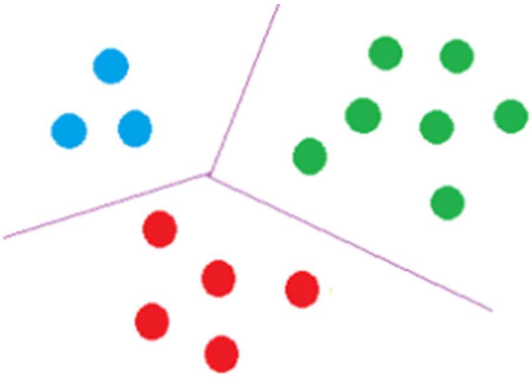
   For the gray point A, compute its distance to C1, C2 and C3, respectively. And after comparing the lengths of d1, d2 and d3, we figure out that d1 is the smallest, therefore, we assign point A to the blue cluster and label it with blue. We then move to point B and follow the same procedure. This process can assign all the points and leads to the following figure



3. Re-assign the cluster centres by calculating the centroid of all the points that are part of a single cluster, as shown in the figure below.



4. Repeat Step 2 and 3 until there is minor or no change in the position of cluster centroids.

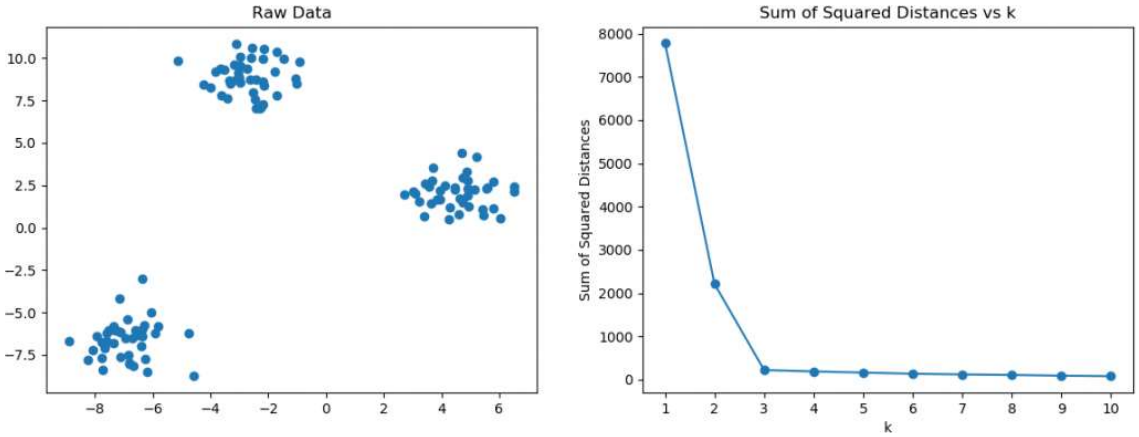


c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

From Statistical point of view, K is chosen using different measure which finds the optimal value of K. First measure is Sum of Squared Differences. We calculate sum of squared difference between data points and centroids using the below formula:

$$i{=}0\sum n(Xi{-}X)2$$

Where Xi represents a data point and overline{X}X represents the centroid that data point belongs to. This value is calculated for different values of K and a graph is plotted called SSD curve which looks like below, value ok K is chosen such that there is significant drop in SSD on reaching that value.



Second popular measure is The Silhouette Method. It measures how similar a point is to its own cluster compared to other clusters. The Silhouette value of each data point is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

Here a(i) is the measure of average distance of a data point from points in its own cluster and b(i) is measure of average distance from data points of next nearest cluster. More the Silhouette score better the value of K.

From business perspective, value of K is chosen such that it makes sense from the requirement point of view. For example, from the data set of customers of an e-commerce company we need to find out which group should be given 10% discount, 30% discount or no discount. Since we want to get 3-4 groups of customers value of K should be 3 or 4, any more clusters than that will not make sense even if SSD or silhouette suggest something else.

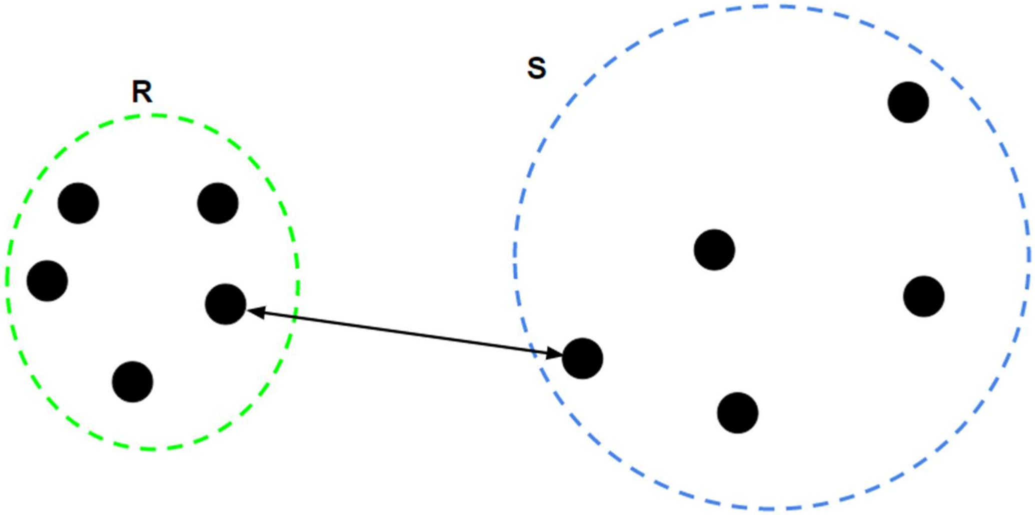
d) Explain the necessity for scaling/standardization before performing Clustering.

Standardizing your data prior to cluster analysis is also extremely critical. Clustering is an unsupervised learning technique that classifies observations into similar groups or clusters. A commonly used measure of similarity is Euclidean distance. The Euclidean distance is calculated by taking the square root of the sum of the squared differences between observations. This distance can be greatly affected by differences in scale among the variables. Generally, variables with large variances have a larger effect on this measure than variables with small variances. For this reason, standardizing multi-scaled variables is advised prior to performing clustering.

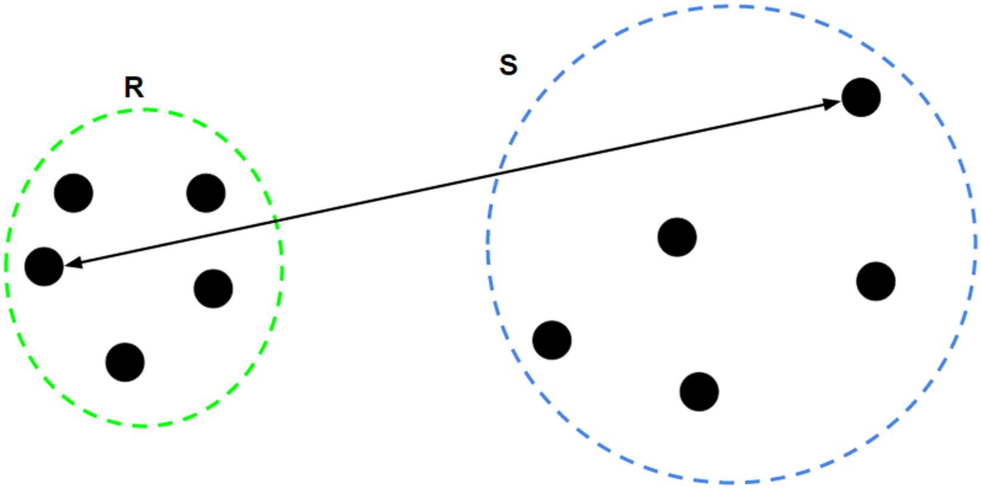
e) Explain the different linkages used in Hierarchical Clustering

To compute the distance between two clusters in hierarchical clustering, there are different approaches called single linkage, complete linkage and average linkage.

Single Linkage: Distance between two clusters is considered as minimum distance between two points, one from each cluster.



Complete Linkage: Distance between two clusters is considered as maximum distance between two points, one from each cluster.



Average Linkage: Distance between two clusters is equal to arithmetic average of distances between all pairs of datapoints taken one from each cluster.

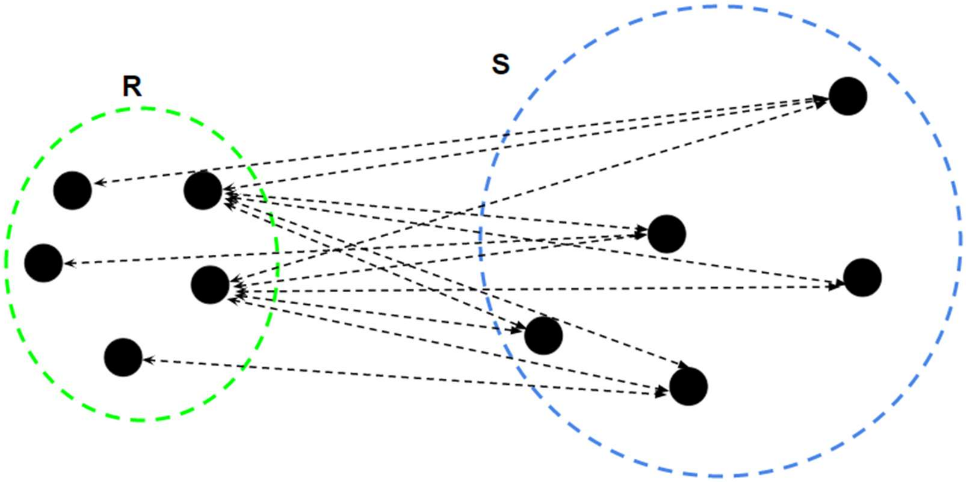


Image Source: GeeksForGeeks