Assignment PART – II

Question 1: Assignment Summary

Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words. Mention the problem statement and the solution methodology that you followed to arrive at the final list of countries. Explain your main choices briefly (why you took that many numbers of principal components, which type of Clustering produced a better result and so on) .

Answer:

The problem statement for the assignment was **to find out the top 5 such countries which are in dire need of aid as one of the NGO is looking to help such countries for a good global cause with the help of Clustering Algorithms. The factors provided on which the countries are to be chosen are CHILD MORTALITY , INCOME and GED Per Capita.**

Methodology Used: The assignment needs to be broken down and build the following roadmap to achieve the objective of finding such countries.

- The countries which are having **more number of children dying before age of 5 years,** having **less income than the average income** and having **low GDP per capita** than the average GDP per capita are the countries need to be aimed for.
- So in order to get there, we used the clustering algorithms such as **KMEANS** and **Hierarchical Clustering** algorithms to achieve the desired output.
- The original/raw data set cannot be used for clustering, hence we needed to first remove the outliers statistically, scale them in order to be on the same measurable scale and then apply the **technique of PCA.**
- The technique of PCA shrinks/reduces the dimensions of the scaled data frame and return us the **Principal Components** that explain an amount of Information variance from our data frame.
- The selection of PCA(4 in our assignment) is done based on the **Scree** plot. In our assignment we took **4 Principal Components** as together they were explaining around 95% of variance of information for us to cluster.
- We applied both the techniques of clustering i.e. **KMEANS** and **Hierarchical Clustering** both, but according to my analysis, **KMEANS** gives us the better result as in the final we got the top 5 countries from the Lower Class People cluster. Also Hierarchical clustering is not affective on a larger data set, due to which KMEANS got an advantage over Hierarchical clustering for this assignment.

Question 2: Clustering

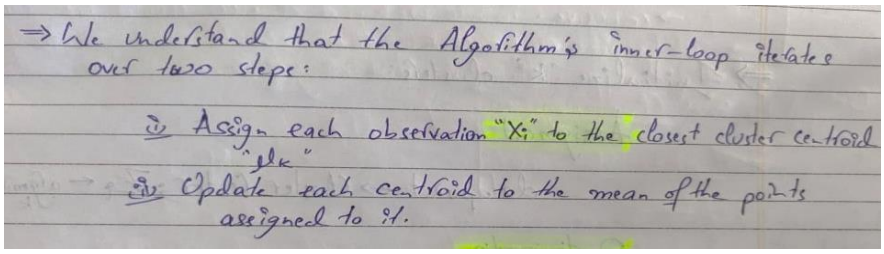
a) Compare and contrast K-means Clustering and Hierarchical Clustering.
b) Briefly explain the steps of the K-means clustering algorithm.
c) How is the value of 'k' chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.
d) Explain the necessity for scaling/standardization before performing Clustering.
e) Explain the different linkages used in Hierarchical Clustering.

Answer :

a.}

KMEANS Clustering	Hierarchical Clustering
• KMeans clustering is an iterative approach due to which linear amount of time (**O(n)**) is consumed while building a model.	• Hierarchical Clustering consumes quadratic amount of time (**$O(n^2)$**).
• Since KMeans start with random choice of clusters, hence it may produce different resulting clusters on different run.	• In Hierarchical Clustering, there is no need of prior knowledge of number of clusters, hence the results obtained on every run of the algorithm is same as the previous one.
• Since, KMEANS requires the prior knowledge of clusters, the algorithm finishes clustering once it has achieved the desired clusters.	• Whereas in Hierarchical clustering we can stop at whatever number of clusters we find to be fit for the problem statement by interpreting the dendogram.

b.} The steps involved in K-Means Clustering algorithm:



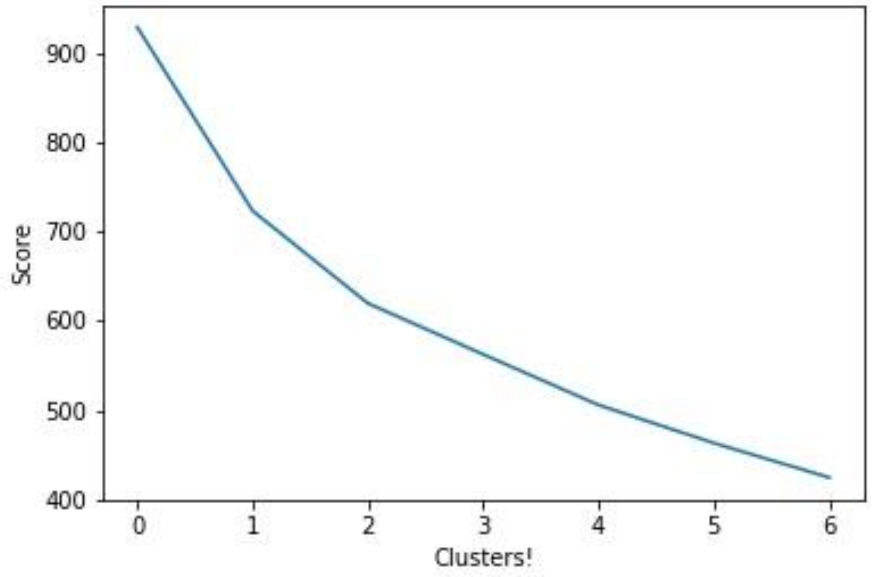
1. Start by choosing **'K'** initial centroids.
   a. Assignment Step : Randomly select a center and compute the Euclidean Distance to find the closest path and assign the points to the clusters respectively.
   b. Optimization Step : For every cluster we update the position of the cluster centers.
2. The grouping of clusters is done in such a way that:
   - Closesness/Tighteness of cluster is maximized.

- While, maximizing the distance between the clusters.

3. Each time the clusters are made, the centroids are updated. The updated centroid is the center of all the points which fall in the cluster associated with the centroid associated with the centroid.
4. Step 3 is iterated again and again till the centroid no longer changes, i.e. solution converges.

c.} The process of choosing the **Optimum Number of Clusters(K)**, statically is done by using :

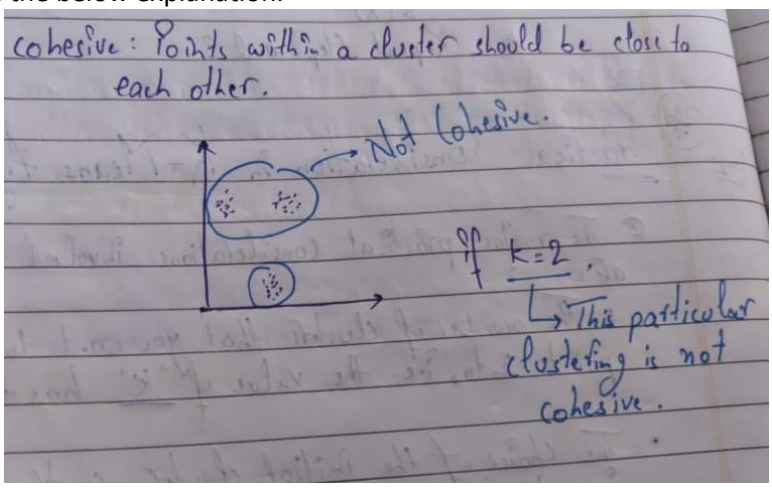
a) The Elbow Curve Method:

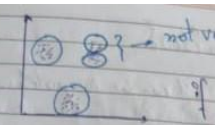


We can see that at index 1, which corresponds to **K=3** is the last significant dip, after which there is no more tight drop.

b) Silhouette Analysis: It is the measure of how similar a data point is to its own cluster(**cohesion**) compare to other clusters (**separation**).

The following images are prepared with respect to **K=2** for clarity purpose, but in our assignment we found **K=3**. Let's see the below explanation.





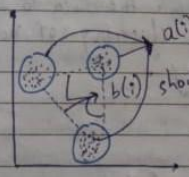
* So we have to capture a metric such that:

"Metric (Cohesion, Dissimilarity)"

⇒ So to compute "Silhouette Metric", we need to compare two measures i.e. a(i) and b(i) where,

- a(i) is the average distance from own cluster (cohesion) — As small as possible
- b(i) is the average distance from the nearest neighbor cluster (Separation) — As large as possible

For every i^{th} point



* $S(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$

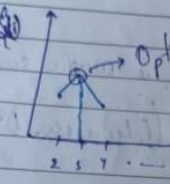
→ Ideal situation: $a(i) \ll b(i)$

→ S(i) can be at max 1; not more than this.

* Avg. S(i) = mean$_i$ {S(i)}

→ For every "k" it can be computed.

So plotting Avg. Si vs "k"



* → If we are worried about "K-Means" getting stuck in bad local optima, one way to solve this problem is if we try using multiple random initializations.

The silhouette Score received for clusters in our case are:

```
For n_clusters=2, the silhouette score is 0.28915373549301415
For n_clusters=3, the silhouette score is 0.28893459376584363
For n_clusters=4, the silhouette score is 0.29889857835637956
For n_clusters=5, the silhouette score is 0.30221787102857656
For n_clusters=6, the silhouette score is 0.23715857204944324
For n_clusters=7, the silhouette score is 0.2542062634731966
For n_clusters=8, the silhouette score is 0.22534567389476576
```

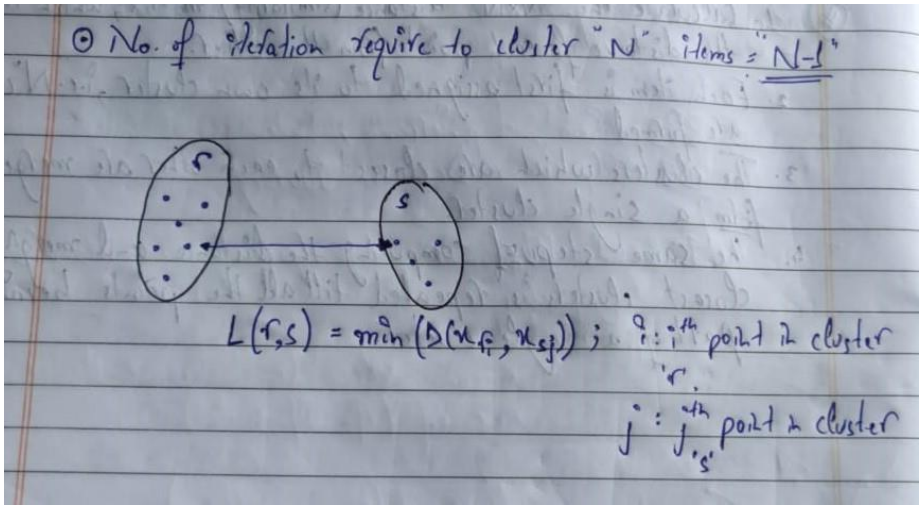
Hence the optimum cluster is **K=3 ,** although it's value is less than K=4, but we have to look for the business aspect also. The business aspect says, that we need to classify the countries where the data for individual population is provided. Hence we can classify the population into **Lower Class, Middle Class, Upper Class people.** Hence using this aspect, we finalize the cluster to be 3.

d.} **Importance of Scaling/Standardization**: Most of the times, our dataset will contain features highly varying in magnitudes, units and range. But machine learning algorithms such as KMEANS use Euclidean distance between two data points in their computations, this is a problem. If not scaled these algorithms will only take the magnitude of features neglecting their respective units. The results would vary greatly between different units, 10km and 1000mtrs.This will result in a biased model weighing towards the variables which have higher magnitude. In order to overcome this issue, the technique of **Standardization and Scaling** is used.

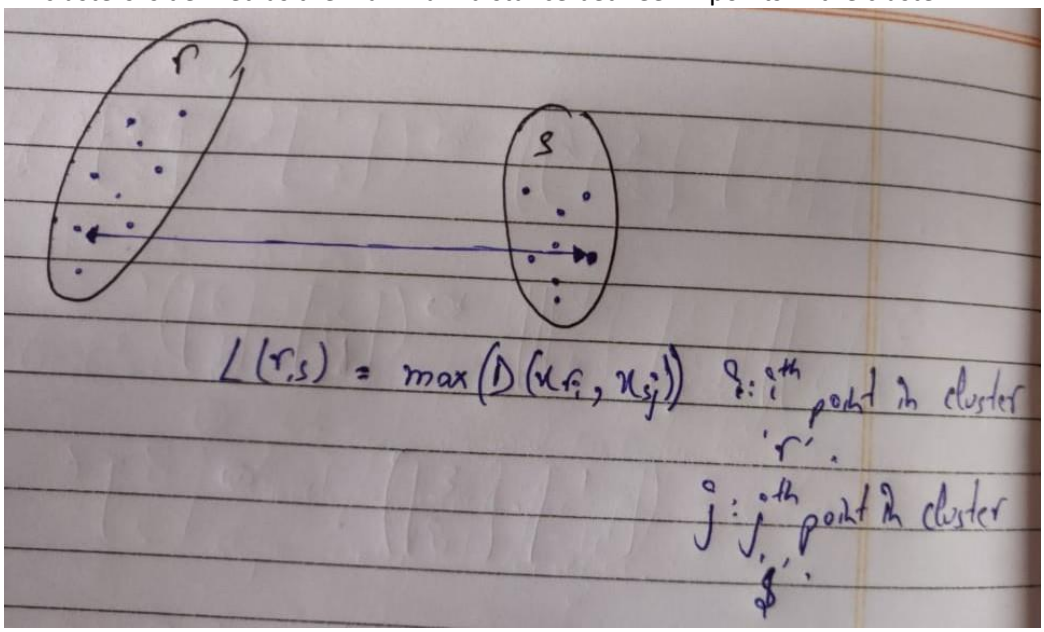
- Standardization is the process where the values are replaced by their respective **Z-Scores.**
- Scaling is the process where all the variables values are brought on similar scale of measurement.

e.} Different Linkages in Hierarchical Clustering:

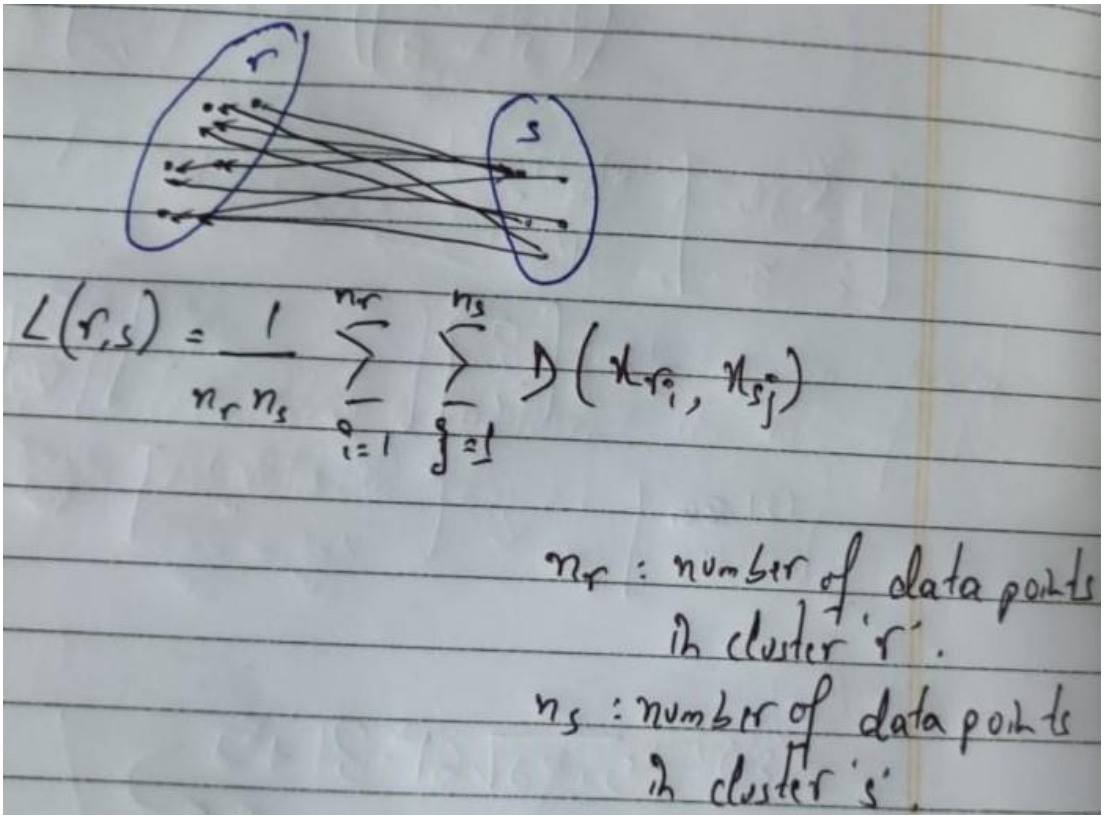
a. **Single Linkage:** In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster.
For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.



b. **Complete Linkage :** In Complete Linkage for Hierarchical Clustering, the distance between 2 clusters is defined as the maximum distance between 2 points in the cluster.



c. **Average Linkage:** Under average linkage of Hierarchical Clustering, the distance between 2 clusters is defined as the average distance between every point of one cluster to every point of the other cluster.



Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.
b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.
c) State at least three shortcomings of using Principal Component Analysis.

Answer:

a. The 3 Application of using PCA are:
    - Disease control/detection
    - Image Processing
    - Quantitative Finance
    - Energy pricing
b. The 2 important blocks of PCA:
    - Basis Transformation: The Basis Transformation is the technique used for converting the dimensions of the raw dataset into another dimension. The need of this transformation is because we need the raw data to be represented into some other dimension which is more useful and conveys a lot more information.
    - Variance as Information: Variance as information in terms of PCA stands for the variability explained by the converted principal components. The more the variance being explained by the components, the better the components are for explaining the data.
c. 3 Shortcomings of Using PCA:
    1. The PCA favors if the data needed to the algorithm is linear/ in 2D space. But if the non-linear data has to be handled, PCA fails to deliver the required objective.
    2. PCA always finds orthogonal principal components. It fails to achieve the objective if we need non-orthogonal principal components to be used for our model building.
    3. PCA always considered the low variance components in the data as noise and recommends to throw away those noisy components. But, sometimes those components play a major role in supervised learning task.