Clustering & PCA Assignment -II

Problem Statement:

- HELP International 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. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.
- Recently they have raised $10 million and now, CEO of the NGO needs to decide how to use this money strategically and effectively.
- The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.
- We have received a dataset which includes various countries economic and social conditions such as GDP, Income, Fertility rate, Child mortality etc.
- As a data analyst, We need to categorize the countries using some socio-economic and health factors that determine the overall development of the country.
- Then we need to suggest the countries which the CEO needs to focus on the most.

***************Question 1: Assignment Summary**************

Solution Methodology:

We followed below steps for analysis and conclusion:

1. Data browsing and analysis:
   Country dataset in python evaluated. Data sanitised for null values.
2. Data formatting:
   Data formatted as Standard normal form for equal weightage on the PCA analysis
3. Data Visualization:
   Data was visualized using number of plots such as Heat-map, scatter plot to understand the correlation and data pattern

4. PCA analysis:
   4 variables identified based on Elbow curve and Silhouette Analysis
   On PCA analysis, I found that 94.4% of variance was explained by 4 variables.

5. Clustering:
   Elbow curve and silhouette_score built to identify number of K.
   Clustering performed based on PCA dataset and data was categorized into clusters using K-means and Hierarchical method
6. Data visualization on clusters:
   Data visualization is performed on columns -Income, GDPP and Child mortality for each cluster
7. Final conclusion:
   From 4 Cluster using K-Means, I found 38 countries worst affected in terms of economy, income and child mortality

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/standardisation before performing Clustering.
e) Explain the different linkages used in Hierarchical Clustering.

Q 2) Clustering

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

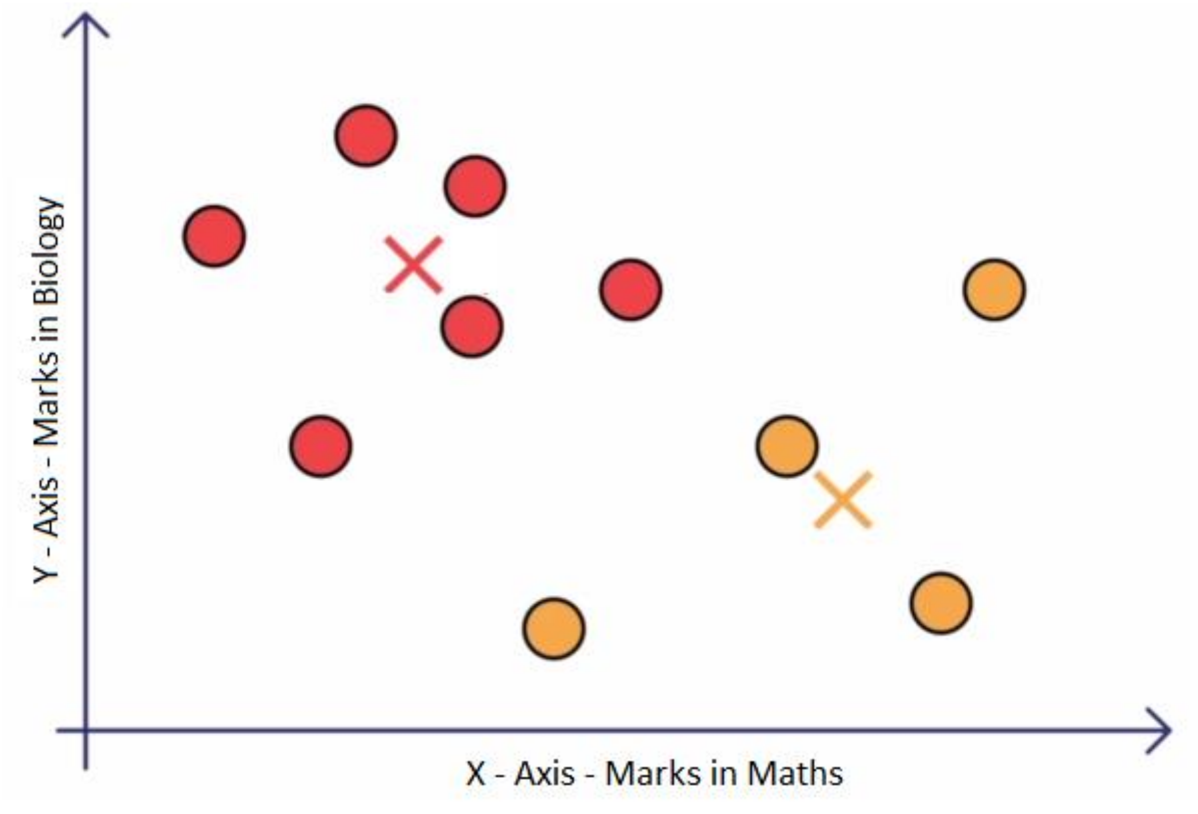
K-means is a iterative clustering algorithm which aims to find local maxima in each iteration and shift the mean value. This algorithm starts with all the data points assigned to a cluster of their own. two nearest clusters are merged into the same cluster.

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

K-Means clustering algorithm:

The first step of this algorithm is creating, c new observations, randomly located, called ‘centroids’. The number of centroids will be representative of the number of output classes (which is not known in advance). Now, an iterative process will start, made of two steps:

- ➢ First, for each centroid, the algorithm finds the nearest points (in terms of distance that is usually computed as Euclidean distance) to that centroid, and assigns them to its category;
- ➢ Second, for each category (represented by one centroid), the algorithm computes the average of all the points which has been attributed to that class. The output of this computation will be the new centroid for that class.



Every time the process is reiterated, some observations, initially classified together with one centroid, might be redirected to another one. Furthermore, after several reiterations, the change in centroids’ location should be less and less important since the initial random centroids are converging to the real ones. This process ends when there is no more change in centroids’ position.

c) How is the value of ‘k’ chosen in K-means clustering? Explain both the

statistical as well as the business aspect of it.

There are two methods to identify the Number of K.

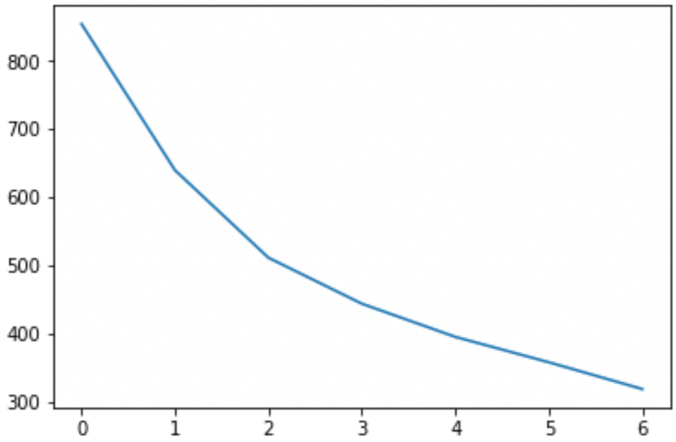
1. Elbow method and 2. Silhouette Score

Elbow method is the one which is mostly used to identify the number of K. The idea is that what we would like to observe within our clusters is a low level of variation, which is measured with the within-cluster sum of squares (WCSS):

And it is intuitive to understand that, the higher the number of centroids, the lower the WCSS. In particular, if we have as many centroids as the number of our observations, each WCSS will be equal to zero.

The idea is that, if the plot is an arm, the elbow of the arm is the optimal number of centroids.

$$WCSS = \sum_{xi \in c} (x_i - \bar{x})^2$$

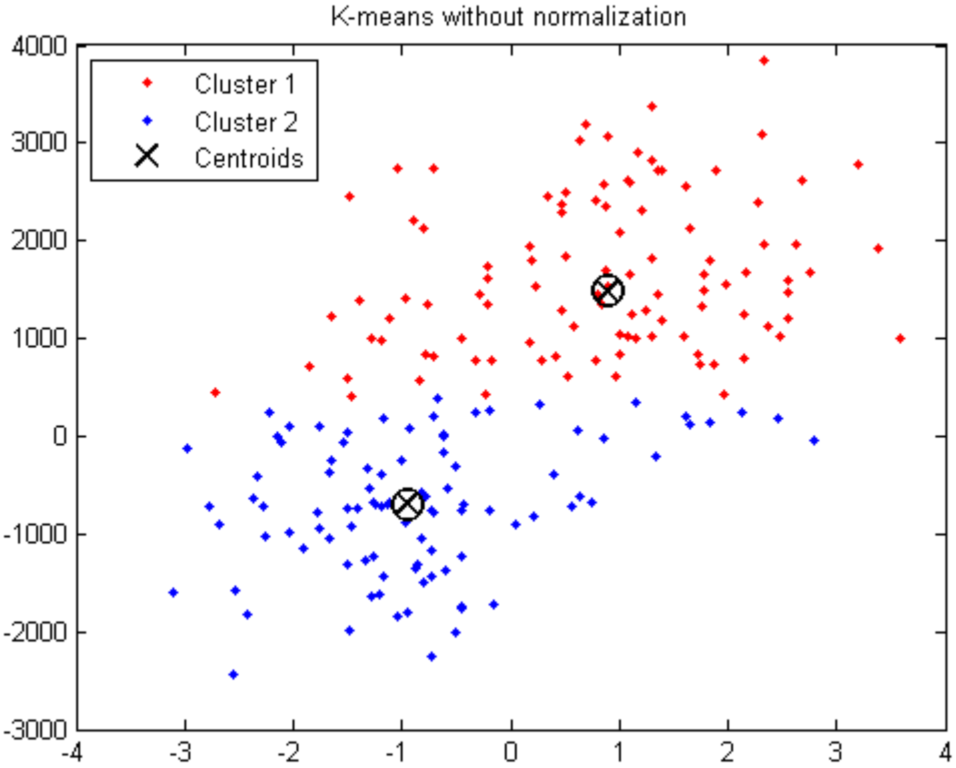


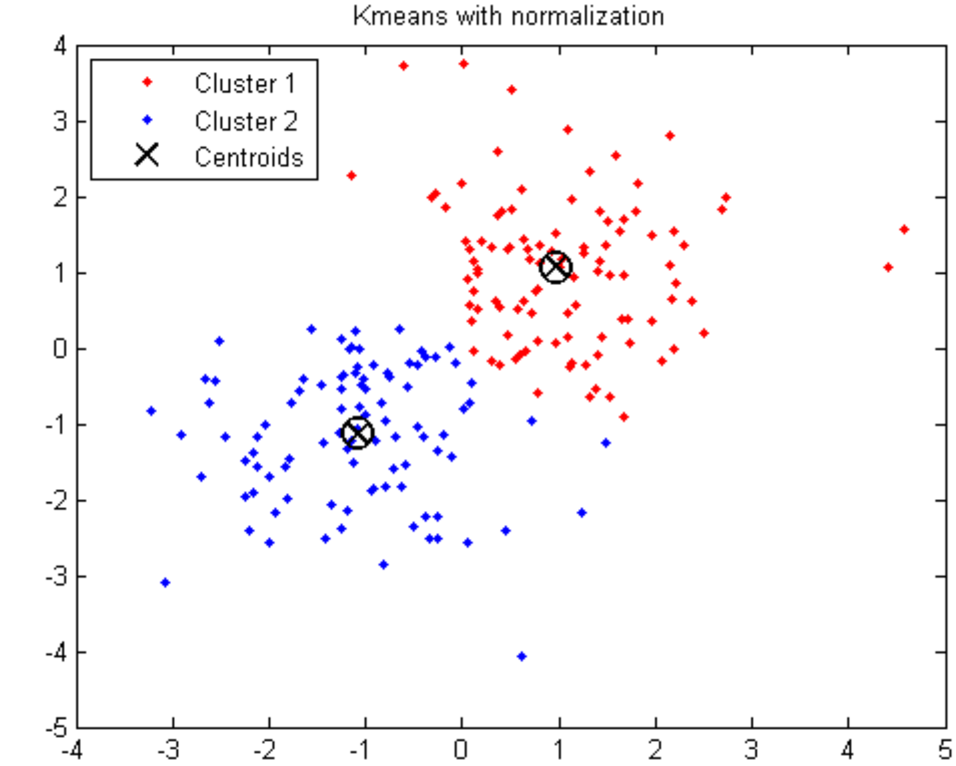
d) Explain the necessity for scaling/standardisation before performing

Clustering.

Clustering is built on PCA. If some variables have large variance and some have very small variance then in this case large variance data will have larger impact than smaller variances.

K-means clustering is "isotropic" in all directions of space and therefore tends to produce more or less round (rather than elongated) clusters. Here, I will give one example where clustering performed on Non-normalized data and how it changed when clustering is performed on normalized data.





The comparative analysis shows that the distributed clustering results depend on the type of normalization procedure.

Standardizing either input or target variables tends to make the training process better behaved by improving the numerical condition of the optimization problem and ensuring that various default values involved in initialization and termination are appropriate. Standardizing targets can also affect the objective function.

Standardization of cases should be approached with caution because it discards information. If that information is irrelevant, then standardizing cases can be quite helpful. If that information is important, then standardizing cases can be disastrous.

e) Explain the different linkages used in Hierarchical Clustering

Hierarchical Clustering:

Clustering tries to find structure in data by creating groupings of data with similar characteristics. The most famous clustering algorithm is likely K- means, but there are a large number of ways to cluster observations. Hierarchical clustering is an alternative class of clustering algorithms that produce 1 to n clusters, where n is the number of observations in the data set. As you go down the hierarchy from 1 cluster (contains all the data) to n clusters (each observation is its own cluster), the clusters become more and more similar (almost always). There are two types of hierarchical clustering: divisive (top-down) and agglomerative (bottom-up).

Divisive:

Divisive hierarchical clustering works by starting with 1 cluster containing the entire data set. The observation with the highest average dissimilarity (farthest from the cluster by some metric) is reassigned to its own cluster. Any observations in the old cluster closer to the new cluster are assigned to the

new cluster. This process repeats with the largest cluster until each observation is its own cluster.

Agglomerative:

Agglomerative clustering starts with each observation as its own cluster. The two closest clusters are joined into one cluster. The next closest clusters are grouped together and this process continues until there is only one cluster containing the entire data set.

There are different linkages are used as metrics in Hierarchical clustering.

Single-Linkage

Single-linkage (nearest neighbor) 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.

Complete-Linkage

Complete-linkage (farthest neighbor) 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.

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.

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.

Q3)

a) Give at least three applications of using PCA.

PCA is predominantly used as a dimensionality reduction technique in domains like facial recognition, computer vision and image compression. It is also used for finding patterns in data of high dimension in the field of finance, data mining, bioinformatics, psychology, etc

1. Image recognition: One of the major use of PCA is in image recognition. Any square image of size NxN pixels can be represented as a NxN matrix where each element is the intensity value of the image. (The image is formed placing the rows of pixels one after the other to form one single image.) So for any set of images, we can form a matrix out of these matrices, considering a row of pixels as a vector.
2. Another example of using PCA is in financial sector. In financial sector it is used to reduce the financial ratios found from different statements. From given a large number of variables available in this sector, the requirement is to reduce the number of variables and group/cluster the data accordingly.
3. Third and most fitted example of PCA is for doing socio and economic survey on countries or continents. One of the example is given as below:

Genetic variation in a sample of 3,000 European individuals genotyped at over half a million variable DNA sites in the human genome. Despite low average levels of genetic differentiation among Europeans, we find a close correspondence between genetic and geographic distances; indeed, a geographical map of Europe arises naturally as an efficient two-dimensional summary of genetic variation in Europeans. The results emphasize that when mapping the genetic basis of a disease phenotype, spurious associations can arise if genetic structure is not properly accounted for.

PCA has two important building blocks Basic transformation and variance as information.

Basic transformation:

The first step of PCA is to perform transformation. In transformation step, given dataset is converted into standard normal. Change of basis is performed for dimensionality reduction.

when we have one dimension, the calculations for the change of basis are pretty straightforward. All we need to do here is to multiply the factor M which gives us the method of transforming from one basis to another.

When there are more than one dimension involved, M becomes matrix rather than simple scalar.
M using the following equation

$M=B^{-1}_{2}B_{1}$

c) State at least three shortcomings of using Principal Component Analysis.

1. Independent variables become less interpretable: After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.

2. Data standardization is must before PCA: You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.

3. Information Loss: Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features.