PCA clustering Assignment Part: II

Question 1: Assignment Summary

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.

After the recent funding programmes, they have been able to raise around $ 10 million. Now the 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.

And this is where you come in as a data analyst. Your job is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most.

Solution:

We import the data into a data frame

1. **Cleaning & scaling:**

- Then we understand the pattern of data being present
- Check if the null values are being present and treat them
- Every variable in the data would be having its own scale of unit. Hence we will perform the Scaling of the data to make them standard

2. **Generation of PCA's:**

- After cleaning and standardizing the data, let us import the PCA function and implement the same on to the cleaned dataset.
- Once the PCA components are generated, we will check the cumulative variance explained by each Principal Component.

- Then we plotted the scree plot to find out the optimal number of principal components
- Create a PCA data frame and check the correlation between the principal component

3. K -means clustering:

- We will check the Hopkins statistics value before performing k- means clustering
- We will perform silhouette or elbow curve analysis for selecting the optimal number of clusters (K) value
- Once the K value is finalized we will import the K -means function and then perform clustering of the data
- We will merge this with the original data frame and analyse individual variables for all the clusters
- From the outcomes, we concluded that the countries falling in cluster_1 are in need to Aide from the NGO
- And we will then find the top 5 countries name which requires the Help of NGO

4. Hierarchical clustering:

- Firstly, we will again import the data in to separate variable
- Next we performed single linkage and complete linkage to identify the number of clusters to be formed from Hierarchical clustering
- From the Complete clustering it is evident to select the optimal number of clusters
- Later we assign the cluster ID's for the countries
- Once the Cluster ID's are assigned, we will join the original data to it
- Next we plotted all individual variables to understand which cluster is in need of Aide
- Then we found the top 5 countries which are in need of help of fud from NGO

5. Analysing the outliers:

- We generated the description table to check for the outliers
- Then we obtained the cluster_0 and understood that the outliers are being present in this specific clusters and they are treated without removing and clustering into a single cluster which represent the developed countries list which does not need the funding from NGO.

Question 2: Clustering

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

Hierarchical clustering:

Hierarchical clustering creates a hierarchy of clusters which may be represented in a tree structure called a dendrogram. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations.

K – means clustering:

The k-means algorithm assigns each point to the cluster who's Center (also called centroid) is nearest. The Center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

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

- We will select the optimal no of cluster (K) value based on business requirement or by performing silhouette or elbow curve analysis
- Then we group the data points into the clusters
- Then identifying the cluster centroids (mean point) of the current partition.
- Assigning each point to a specific cluster
- Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
- After re-allotting the points, find the centroid of the new cluster formed.

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

- The main intention of performing the clustering is to provide insights to improve the business
- Hence it is very important to consider from the business perspective while clustering
- Ideally any business, we would have to at least perform 2 to 3 clusters to understand, analyse, predict and provide insights for development.

From the statistical aspect we have two methods to select the K value.

- First one is the Silhouette plot analysis, which explains us how much of data is explained by how many o of components, we will choose the one which explains the highest init.
- Second one is the elbow curve analysis, in which from the plot we can find the K value where the elbow cut is found from it which would explain the maximum of data in it.

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

When we have a data consisting of multiple variable which are explained by their own scale of unit, one would vary from the other. But when we are performing the data this

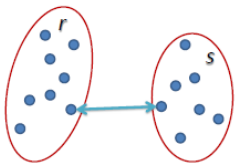
should be taken care. Because the values would vary highly in magnitude from each other and when clustering is performed without treating them, it would mislead us to different outcomes which would not be useful to implement.

Hence we should standardize the unit of all the variables within the data to a single one either by performing normal or standard scaling on to it. So that we can obtain outcomes without the influence of it.

e) Explain the different linkages used in Hierarchical Clustering.

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.

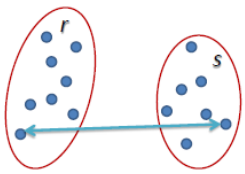


$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Single linkage

Complete Linkage:

In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest 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 furthest points.

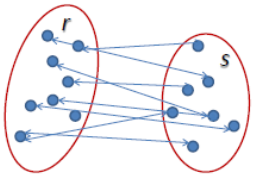


$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Complete linkage

Average Linkage:

In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters "r" and "s" to the left is equal to the average length each arrow between connecting the points of one cluster to the other.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Average linkage

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

- It is used mainly for the purpose of dimensionality reduction in various industrial applications
- Uncovering the latent variables/ themes/ concepts
- Data visualization and EDA: when very large numbers of variables are present in data set and it's very difficult to analyse, hence there PCA technique can be used
- Helps in creating uncorrelated feature to the dataset for analysis
- Noise reduction in the data as well

b) Briefly discuss the 2 important building blocks of PCA - Basis transformation and variance as information.

Basis Transformation:

- It is basically the same as "conversion of units"
- Basically, what you do in basis transformation is that you change the representation of the same point from one "unit" to another. Like in currency, the amount 1500 rupees can be represented as 22 US Dollars or 19 Euros or 2300 Yen and so on.
- So we convert and find a new set of basis vectors
- These basis vectors explain the information of dataset in best possible way, and thus allows us to perform dimensionality reduction, finding latent variables
- It allows us to represent data in a simpler way and makes easy to work on it

Variance as information:

- The data set does not have the columns in the order of variance that we want.
- In some scenarios, the variance might be equally distributed amongst the all columns.
- If we don't understand the variance contribution of columns it would be a tedious process and difficult to analyse
- The effective would be to understand the variance contribution of individual columns so that we can concentrate on the one which contribute and ignore the others.
- There by making it easier and effective to understand and find new basis vectors to work on the problem to effectively implement as a solution

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

- **Linearity:** PCA assumes that the principle components are a linear combination of the original features. If this is not true, PCA will not give you sensible results.
- **Large variance implies more structure:** PCA uses variance as the measure of how important a particular dimension is. So, high variance axes are treated as principle components, while low variance axes are treated as noise.
- **Orthogonality:** PCA assumes that the principle components are orthogonal.