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)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Answer:

Our main objective for this assignment is to find the countries that are in direst need of aid. Our job is to find those countries using socio-economic and heath factors which will show overall development of the country.

So, in order to do so first I analyze the dataset – It has total 167 countries with no missing values, I transformed some of the variables like health, imports and exports from % of GDP to 5 of GDPP as it makes more sense to the dataset. Then I visualized the data by finding the correlation between variables to check the multicollinearity and found that most of the variables are having multicollinearity. To remove this issue, I used PCA – principal component analysis, this concept will not only solve our multicollinearity problem but also it will help us reduce the dimensions of the dataset. So, before applying PCA on the dataset I standardized the dataset by using standard scaler and then I applied PCA on the scaled dataset. After this I got 9 features or components as PCA will create components equal to or less than the variables present in the dataset.

Now I have plotted the first two principal components on x-y axes to understand the feature sense. And I find out by scatter plot that some of the features like gdpp, life expectancy etc. are in the direction of first principal component which means they shows strong variance as first component always have high variance explained than other components.

Now to choose number of components I decided it by scree plot. I have concluded from the plot that the ideal number of components to select is 5 as it explains approx. 97% of the variance. So, going forward I will use number of principal component equal to 5.

Now I checked the outliers of the dataset and I found there are outliers on all the variables. I checked this by describing the dataset in percentiles and by plotting boxplot. So, after this I made two approaches – one with outliers and another without outliers. By this way I will get an idea about the dataset and will choose the final countries from any one of them. And when I have used incremental PCA the transformed data on principal components have removed the multicollinearity as all the values are having correlation near to 0. So, first I K-means clustering with outliers and found the list of countries which need aid. Again, same thing I did with hierarchical clustering and found different results than k-means. So, from this I have to choose which method gave us correct outcome and I have choose k-means output because even though for less records hierarchical clustering have much advantage than k-means, I have to check the business requirements and more precise information is giving from k-means.

Now, same analysis I did without considering the outliers. This time I got promising output from hierarchical clustering but one point I found that the countries which are excluded because of outliers are having low income and low gdpp, which means if I am doing outliers then I am missing most important countries and it is certainly not meeting our business needs.

So, my final data was from 1st method which was by K-means and I got total 47 countries which are in direst need of aid.

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.

Answer:

a)

K-Means Clustering	**Hierarchical Clustering**
We need to have desired number of clusters ahead of time.	We can decide the number of clusters after completion of plotting dendrogram by cutting the dendrogram at different heights
It is a collection of data points in one cluster which are similar between them and not similar data points belongs to another cluster.	Clusters have tree like structures and most similar clusters are first combine which continues until we reach a single branch.
Works very good in large dataset	Works well in small dataset and not good with large dataset
The main drawback of k-Means is it doesn’t evaluate properly outliers.	Outliers are properly explained in hierarchical clustering
K-means only used for numerical.	Hierarchical clustering is used when we have variety of data as it doesn’t require to calculate any distance.

b)

Step 1: Randomly select K points as initial centroids.

Step 2: All the data points closet to the centroid will create cluster center according to Euclidean distance function.

Step 3: Once we assign all the points to each of k clusters, we need to update the cluster centers or centroid of that cluster created.

Step 4: Repeat 2,3 steps until cluster centers reach convergence.

c)

‘K’ value is chosen randomly in K-Means clustering based on statistical aspect. From business aspect, we need to first understand the dataset and based on that we decide number of ‘k’. for example, we have a dataset of variables like ‘pen’, ‘pencil’, ‘books’, ‘notebooks’, ‘mobiles’, ‘charger’, ‘laptop’. Now if we want to have k values based on statistical aspect, we can use silhouette score to determine that but based on business aspect, after viewing the dataset we can easily make cluster = 2, one in electronics category and another non-electronics.

d)

It is definitely a good idea to do scaling/standardisation because our variables may have units at different scale and as our method stresses more on calculation of direction of space or distance, so if we have one variable with high scale units then while calculating for k-Means or hierarchical it will create a big difference as the clusters will tend to move with the variables having greater values or variances. By applying standardisation/scaling will increase the performance of our model.

e)

Linkage is a technique used in Agglomerative Clustering.

Linkage helps us to merge two data points into one using below linkage technique.

Single linkage: The distance between two clusters is calculated by the minimum distance between two points from each cluster.

Complete linkage: The distance between two clusters is calculated by the maximum distance between two points from each cluster.

Average linkage: The distance between two clusters is the average distance between every point of one cluster to the another every point of other cluster.

Ward linkage: The distance between clusters is calculated by the sum of squared differences with all clusters.

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)

PCA used in application like image compression, facial recognition, Finance (Trading/stock), Health (Neuroscience)

b)

Building blocks of PCA –

1. Basis Transformation:
    a. Basically, it is just like a conversion, we just change the representation of the same point from one unit to another. It converts a set of variables (correlated variables) into another variables (non-correlated variables) which is finally called principal components.
    Example: let’s take two basis vectors – [2,4] and [4,6]. Now let’s use these vectors and represent the same point differently.
    Take a new point [6,10] which can be written as [6,10] =1*[2,4] + 1*[4,6]
    Hence, we can see that the new basis system [1,1] used to represent [6,10]
    b. We find new basis transformation where most of the variance in the dataset is along the axes. We find the new components of the dataset in a best possible way and helps us to do operations like finding latent variables, dimension reduction etc.

2. Variance:

a. More variance of a column meant more information which means it is more important for creating new models. Therefore, the low variance column can be removed from our dataset without affecting much to the output. This is what dimensionality reduction does. It removes unnecessary columns and gives our data more informative.

c)

Three shortcomings of PCA:

1. PCA is a linear transformation method hence it is limited only to linearity.
2. PCA needs all the components perpendicular to each other, hence for some cases it is not the best way.
3. PCA only keeps high variance components and removes low variances as it finds as it is not useful, which is not true for some cases. Ex: Health domain - for prediction setups(Classification with imbalance variables)