

Question 1: Assignment Summary

Answer:

Problem statement of the assignment was to cluster the given set of countries, and then compare them with these three variables - [gdpp, child_mort and income]. Basis on the analysis, then report back at least 5 countries which are in direst need of aid. There can be multiple ways of completing the assignment, but I followed below steps:

- 1) Read and understand the data
- 2) Clean the data, and make necessary modifications on the columns
- 3) Prepare the data for modelling (Perform PCA)
- 4) Modelling (KMeans, and Hierarchical clustering method)
- 5) Final analysis (Choose one method, and then analyzed the results)

Well, there were many choices that I had to made during the assignment. Couple of them are as follows:

- 1) 'exports', 'imports', and 'health' columns were given as percentages of GDP. Since GDP was not given, and instead GDPP was given, I converted these columns as values per person, since it would be ideal to convert all columns in one scale.

E.g.-

$\% \text{exports} * \text{GDP} = \text{Total exports}$

Since GDP is missing, and GDPP is given in question

$\% \text{exports} * \text{GDPP} * \text{population} = \text{Total exports}$

$\% \text{exports} * \text{GDPP} = \text{Total exports/population}$

$\% \text{exports} * \text{GDPP} = \text{Total export per person}$

- 2) To remove multicollinearity, PCA was used instead of RFE.
- 3) Hopkins statistics was calculated to see if data could be clustered, and indeed the result was positive, and data could be clustered.
- 4) I chose PCA with 5 components as it explained around 95% of the variance of the dataset.
- 5) 4 clusters were formed for the countries after seeing the elbow curve. Moreover, Silhouette score of 4 clusters was around 45 which is good. I gave more preference to elbow curve while choosing number of clusters as elbow curve was giving clear result.
- 6) Since KMeans clusters are more segregated compared to hierarchical clusters, we have chosen KMeans clustering for our analysis.

Question 2: Clustering

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

Answer:

1. The number of clusters in which the data points could be divided into, i.e. the value of the k must be pre-determined in K-means clustering, while it is the not the case with Hierarchical clustering.
2. The choice of the initial cluster centers can have an impact on the final cluster formation in K-means clustering, while results are reproduceable in Hierarchical clustering.

3. Time complexity of K-Means is linear i.e. $O(n)$ but time complexity of hierarchical clustering is quadratic i.e. $O(n^2)$.

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

Answer:

1. Start by choosing k initial centroids.
2. Assign each observation $X(i)$ of the dataset to the closest cluster centroid $u(k)$.
3. Update each centroid to the mean of the points assigned to it.
4. Again, perform step 2, and step 3 until there is no change in the clusters, or possibly until the algorithm converges.

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

Answer:

Multiple techniques can be used to choose the value of k in K-means clustering statistically:

1. **Elbow curve method:** The idea of the elbow method is to choose the k at which the sum of errors decreases abruptly. This produces an "elbow effect" in the graph, and then we choose that many clusters to start the clustering process.
2. **Silhouette coefficient:** It is a measure of how similar a data point is to its own cluster (cohesion), as compared to other clusters (separation).

While above are the statistical measures of choosing K , there are business aspects also which needs to be taken care of while choosing the value. In our current assignment also, we could have also chosen 2 number of clusters, but we chose 4. Because if we would have chosen 2, we would not have got countries which are in dire need of aid, but a general list of clusters which differentiates highly developed countries, from underdeveloped, and developing countries. We should always give importance to the business problem that we are trying to solve and accordingly choose the value of ' k '.

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

Answer:

Since the distance metric used in the clustering method is the Euclidean distance, we need to bring all the observations in the dataset to the same scale. This can be achieved either through either scaling or standardization. Clustering algorithm(s) works by taking into consideration the values of the columns in the dataset and try to optimize the distance between various observations of the dataset.

e) Explain the different linkages used in Hierarchical Clustering.

Answer:

Single Linkage: It considers the minimum of all the pairwise distances between the data points as the representative of the distance between 2 clusters. It produces dendrograms which are not structured properly.

Complete Linkage: Distance between any two (2) clusters is defined as the maximum distance between any two (2) data points in the cluster. It produces clusters which have proper tree like structure.

Average Linkage: Distance between any two (2) clusters is defined as the average distance between every point of one cluster to every point of the other cluster. It also produces clusters which have proper tree like structure.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

Answer:

There are lot of applications of the PCA, and some of them are described as following:

1. It helps in Dimensionality reduction and it also prevents loss of information due to dropping of variables.
2. It helps in Data visualization.
3. It helps in building predictive models by creating faster models, which have no multicollinearity.
4. It helps in finding latent themes.
5. It results in Noise reduction.

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

Answer:

Basis transformation:

Basis vectors are certain set of vectors whose linear combination can explain any other vector in that space. **For example:** For any dataset D1, we can find completely different set of vectors' B1 to represent it in the new basis system. So, if original dataset contains n columns, we try to get a smaller number of vectors, which explains all the original dataset variance.

Change of Basis in 1-Dimension

New Basis Representation (meter) = M * Old Basis Representation

$M(\text{inverse})$ * New Basis Representation (meter) = Old Basis Representation

When more than 1 dimension are involved, then M becomes a matrix rather than a simpler scaler.

M is a representation of old basis in new basis.

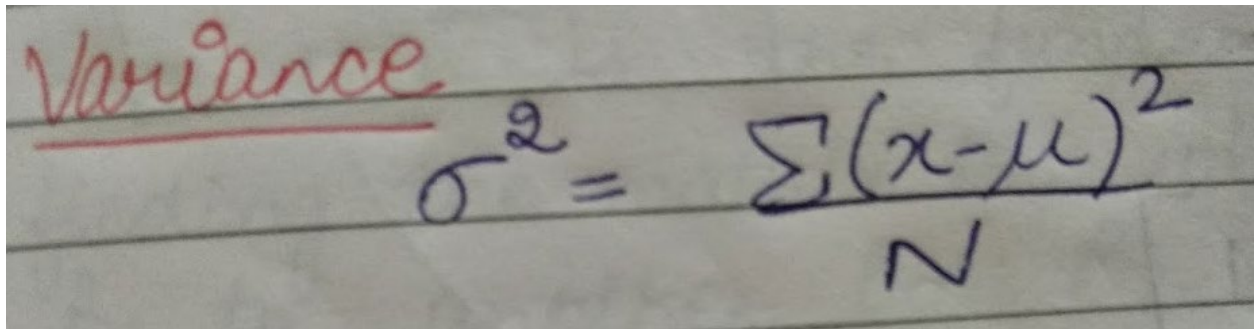
$M(\text{inverse})$ is a representation of new basis in old basis.

Change from one Basis to another

B1 is a set of Basis vectors, and B2 is another set of Basis vector's and moving from B1 to B2, would be done with following equation:

$$B1 = M * B2$$

Variance:



The image shows a handwritten formula for variance on lined paper. The word "Variance" is written in red ink and underlined. To its right, the formula $\sigma^2 = \frac{\sum (x - \mu)^2}{N}$ is written in blue ink.

- PCA measures the importance of the column by checking its variance. If it contains more variance, then it contains more information, and hence it is important.
- The very idea of the spread of the column data being equivalent to the variance is very good way to distinguish important directions from non-important directions.

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

Answer:

1. PCA automatically assumes that variables which have low variance are insignificant. But this assumption is not true in datasets which have high class imbalance.
2. PCA needs its components to be perpendicular. But in some cases, that might not bring the best solution.
3. PCA is very limited to linearity, though it can be used with some non-linear techniques too.