Assignment: Part II (CLUSTERING AND PCA)

MADHAVAN RANGARAJAN

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:

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. 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. My job is to cluster the countries using some socio-economic and health factors that determine the overall development of the country. Then i need to suggest 5 countries (in dire need of financial aid) which the CEO needs to focus on the most.

Column Name	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services. Given as %age of the Total GDP
health	Total health spending as %age of Total GDP
imports	Imports of goods and services. Given as %age of the Total GDP
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	The average number of years a new born child would live if the current mortality patterns are to remain the same
total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

Solution Methodology:

1) EDA and Outlier Analysis :

   I started with reading the file in python , checking the head of the dataframe , checking the number of countries and number of features. After verifying that there are no missing values I used boxplots of the features to check the distribution and to detect outliers.

   As I found that gdpp and income had lots of outliers towards the right tail, high income and high gdpp being the traits of a 'DEVELOPED NATION' I choose to remove those countries. Also , avoiding data loss belonging to under developed economies being primary objective I did soft capping for features like child_mort , life_expec to bring the outliers to the 99th quantile value range

   After I removed the outliers, I subjectively verified the countries I removed are not under-developed.

2) Standardisation and PCA:

   I plotted out a heatmap displaying co-relation between variables. After there was clear indication of high multi co-linearity between the features and also as a fact that hierarchical clustering is computationally expensive I choose to use dimensionality reduction technique (principal component analysis) to reduce 9 features into 5 principal components. The choice of choosing 5 principal components using incremental PCA was made after plotting a scree-plot and observing that almost 95% of the variance in the data was captured by 5 principal components. As PCA depends on the standard deviation of the original data we are required to standardize the data before performing PCA to bring all variables under same scale and thereby giving same weight to every variable

3) Verifying Cluster Tendency:

   To check the cluster tendency we use Hopkins statistic. Hopkins statistic verifies if the data points differ from uniformly distributed data in multi dimensional space . It checks if the data can form meaningful clusters

4) K Means Clustering:

   I used elbow curve method and silhouette metric method to statistically choose optimal k for k-means . I also used business understanding along with statistical understanding to choose 3 clusters . I extracted the cluster labels

5) Hierarchical Clustering:

   I tried both types of linkage (single and complete) for agglomerative clustering . Complete linkage produced efficient result . I choose 4 clusters in case of hierarchical clustering. I extracted the cluster labels

6) K Means vs Hierarchical Clustering:

   I analysed the clusters by comparing how these three variables - [gdpp, child_mort and income] vary for each cluster of countries to recognise and differentiate the clusters of developed countries from the clusters of under-developed countries. I did these for both kmeans and hierarchical and compared the intra cluster heteroscedasticity.

   I visualized by choosing the first two Principal Components (on the X-Y axes) and plotting a scatter plot of all the countries and differentiating the clusters. I also did the same visualisation using two of the original variables ‘child_mor’ and ‘gdpp’.

After comparison I found out that both the methods are doing a effective job

7) Choosing Final Countries for Reporting:
The countries formed under the cluster ‘under-developed’ for k-means and hierarchical were 39 and 48 respectively. As we need to report only 5 countries to the CEO I first extracted the countries clustered under both methods as ‘under-developed’ , sorted those countries based on gdpp, income and child mortality then picked top 12 countries under each basis and used the intersection of these to finally arrive at 5 countries which are in dire need of financial aid.

Question 2: Clustering

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

Answer:
Both K-means and Hierarchical Clustering are unsupervised methods whose main motive is to cluster the data into groups such that it minimises intra cluster variance and maximises inter cluster variance (tightness within clusters and variation between clusters)

Description	K-means Clustering	Hierarchical Clustering
Choosing no of clusters	The choice no of clusters ‘k’ must be choosen beforehand	Hierarchical Clustering forms a tree like a structure and no of clusters can be choose after the linkage is created and fits the data
No of clusters (decision)	In statistical point of view k clusters by cluster validation methods like elbow-curve and silhouette metric	You can cut the dendrogram at any point according to business requirement but usually you can cut at the point in dendrogram where largest vertical distance without split is observed
Methodology	The concept of centroid and Euclidean distance is used by this algorithm	Linkage methods like complete, average and single are used to merge tree at each iteration
Time complexity and Handling big data	The time complexity of K-means is linear i.e O(n) and thus it can handle large data	The time complexity of Hierarchical Clustering is O(n^2) And thus cannot handle large data really well
Flexibility	K-means works effectively when the shape of clusters are hyper-spherical and if the data is not hyper-	Hierarchical Clustering doesn’t require the data to be hyper-spherical and it works efficiently on all shapes of data

	spherical it is not a good choice	
Repeatability	K-means runs with random choice of centroids and thus it may give different results on different runs of the algorithm. Thus, the results may lack consistency	Hierarchical Clustering gives the same results every time you run
Sensitivity To Outliers	K-means is highly sensitive to outliers. But , the presence of such outliers may sometimes form a cluster because of clumping	It is also sensitive to noise but the effect can be easily removed by cutting the dendrogram at a particular point

b) Explain the steps in k-means algorithm?

Answer:

Step 1 – INITIALIZATION:

The first step is k-means randomly chooses 'k' cluster centroids. The choice of initial cluster centroids effects the final clusters. As k-means algorithm is not a convex function it may get stuck in local minima. This disadvantage can be slightly overcome by choosing the initial cluster centroids efficiently using 'k-means ++'

Step 2 – ASSIGNMENT:

The distance between datapoints and each cluster centroid is calculated and the point gets assigned to the cluster whose distance is minimum.

Step 3 – OPTIMIZATION:

The centroid of datapoints under every cluster is computed and it becomes the new cluster centroid

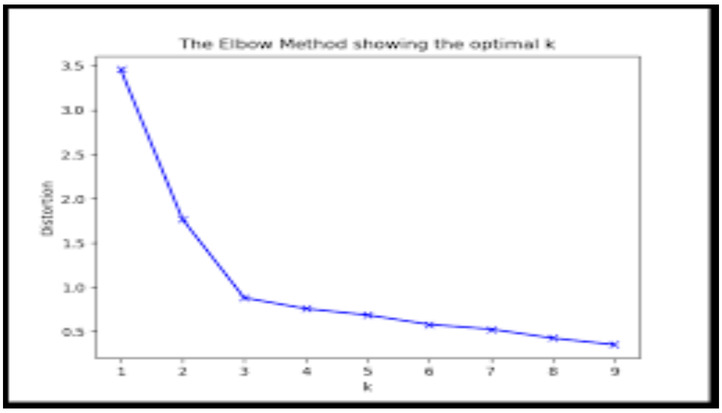
The steps 2 and 3 are computed for enough iterations until cluster centroids stop moving

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:

Statistical Approach to choose optimal 'k' :

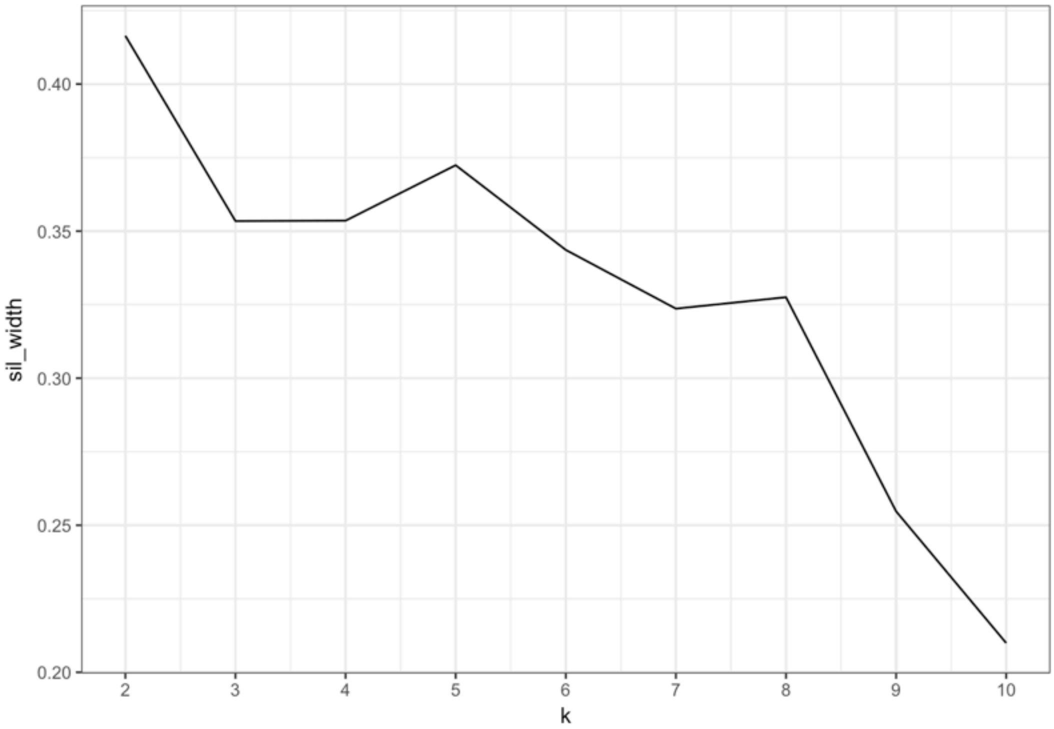
a) Elbow curve method:
Elbow curve is plotting K against variance allowing us to pick the optimal K value where the variance changes sharply.



In the above figure the variance is sharply changed at k =3 thereby the plot forms a bent like structure (the point of change referred as elbow)

b) Silhouette Analysis

Silhouette metric calculates the ratio of average inter cluster distance to average intra cluster distance . It is a ratio of cohesion to seperation



The point where the silhouette score is high suggests better formation of clusters and optimal number of 'k' can be chosen

Business Understanding to Choose Optimal K:

Business Understanding , the use-case of clustering is much more important than statistical approach for choosing optimal 'k'. For example you want to cluster countries into groups and you want to identify countries which are under developed and fund them. In such case , even if statistical approach shows optimal k to be 2 according to the data ,clustering into 2 groups make a lot of countries in one group and there will be again a lot of subjective decision making involved to allocate funds to the country. In such case there is no use if the business objective is not made easy with the use of clustering. Thus business objective must be prioritized in choosing optimal k and statistical approach should be used for reference

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

Answer:
Consider a situation where you want to cluster the data using two features :

a) Price
b) Rating

Now price has value range from 0- 1 lakh whereas rating lies between the scale 0 to 5.
The distance metric used in k-means or hierarchical primarily is Euclidean distance. When both the features are not in the same units , higher weightage is given to the feature with larger numerical value and the distance is purely influenced by only price.

Standardization brings down all features in standard scale with same mean i.e zero and same standard deviation i.e 1. This allows all the features to become unit free and thus giving equal weight to price and rating.

e) Explain the different linkages used in Hierarchical Clustering.

Answer:
Single Linkage:
In single linkage the minimum of all pairwise distance between data points is representative of the distance between 2 clusters
Complete Linkage:
In complete linkage the maximum distance between data points is the maximum distance between 2 points in the cluster
Average Linkage:
It is nothing but average distance between every point of one cluster to every other point of the other cluster.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

Answer:

Dimensionality Reduction:
It is a kind of feature extraction technique which gets only k principal components out of m features which can be used for further analysis and building models. The purpose of PCA is such that it captures maximum variance of the data by reducing the features but still retaining most of the information present in the data

EDA:
Suppose you have a 'n' dimensional data which you want to visualize all together. The first 2 principal components captures maximum variance and thus we can visualize it

Finding Latent Themes and Noise Reduction:
PCA can be used to find the underlying themes in the data and it also reuces noise in the data

When multi co-linearity is present:
When the attributes of the data are highly co-related PCA generates principal components in such a way that data is orthogonal and thus multi co-linearity is removed. This helps model building easier , flexible and more stable

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

Answer:

Basis Transformation:

1) Basis is the fundamental units in which we express our data.
2) In vectors and vector spaces, we use basis vectors to represent the points in space.
3) If we are transforming basis in to the standard basis then the transformation matrix M = basis matrix B
4) Finding transformation matrix that moves between two nonstandard basis involves taking the inverse of matrix in the standard base and multiplying it by the matrix representing the basis of origin.
5) PCA transforms the information present in 2 vectors to represent in one vector (using the concept of direction of maximum variance -magnitude)
6) Projecting original data using orthogonal principal components results in reduction of dimensions

Variance as Information:

a) Variance is directly proportional to information.
b) The more the variance we will be able to capture our Principal component is more accurate
c) More important columns in a dataset are that have more variance
d) The ideal basis vectors are the directions which capture maximum variance
e) These ideal basis vectors capturing maximum variance are the principal components for the dataset

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

Answer:

1) **Data becomes less interpretable:**
After doing PCA on the data original variables will transform into Principal Components. Principal Components are nothing but linear combination of the original features. Principal components are less interpretable

2) **Information Loss:**
Although principal components try to capture maximum variance from the dataset , if optimum principal components are selected without looking at scree plot we may lose some information compared to original features

3) **PCA is a linear method:**
PCA needs the components to be perpendicular . If the data with the direction of maximum variance is not perpendicular In such cases it is not the ideal solution. Here, case we can use Independent Component analysis (ICA)

4)**Requires data to be highly co-related:**
When the features are not co-related pca should not be used forcefully to reduce dimensionality.