Clustering and PCA Assignment

Part – II

Student Name: Hafizullah Mahmudi

Student Id: APFE19M00734

IIIT Role Number: DDS1950112

Kabul, Afghanistan

23rd February, 2020

Institute: upGrad/IIIT-B

Question 1: Clustering of countries

HELP International is an international humanitarian NGOs that is committed to fighting poverty and providing fund for people of backward countries. HELP has got a fund of $10 million that needs to make a proper decision to use this money to the countries that needs the money most.

Provided with the dataset containing name of countries with key factors of each country's socio-economic and health profile, I worked to find list of countries which are in the direst need.

I used two type of clustering, K-Means and Hierarchical clustering to achieve this purpose. And Principle Component Analysis to prepare the data for K-Means and Hierarchical and then make the final clustering.

After applying the PCA on the data set, plotting the scree plot and extracting the cumulative result of component variance. 3 components were shown to be the most optimum that explains about 94.57% of the variance. While 4^{th} components add about 2% variance to the 3 components. So, we used 3 components to perform the clustering.

First, I used K-Means clustering to find categories of countries. Both Silhouette Analysis and Elbow Curve method shows 3 clusters as optimal choice to divide the countries into 3 subgroups.

Second, I used the hierarchical clustering to find the countries. The complete clustering with a cut of 4 clusters were most optimum which produced a good result of clustering countries into 4 categories/clusters.

By profiling the countries with the result of K-Means and Hierarchical clustering with gdpp, child_mort and income. K-Means resulted in 51 countries with lowest average gdpp, and income and highest average of child mortality but Hierarchical produced about 42 countries.

The Hierarchical produced a better result and narrowed down list of countries that are in most need of Aid. It is also possible to event narrow down to shrink the list of countries to 22.

Question 2: Clustering

Clustering is a method of unsupervised machine learning. The clustering helps to divide data points to a number of a number of subgroups that are more similar to the data of the same subgroup or cluster and is dissimilar to other clusters. Clustering is used in different fields like customer segmentation in marking, classification of plants and animals among different species, social media, human genetic clustering, earthquake studies, etc.

Two common types of clustering are K-Means clustering and Hierarchical clustering methods.

K-Means uses **distance measure** or **Euclidean Distance** to partition n observations in to k clusters that all observations within a k is similar to each other and **Hierarchical clustering** method constructs a hierarchy of clustering. K-Means clustering requires initial value of k to be chosen while hierarchical requires no value of k. The performance of K- means algorithm is better than Hierarchical Clustering Algorithm. K-Means is very sensitive to noise in the dataset while Hierarchical clustering is less sensitive

to noise in the dataset. K-Means algorithms Shows less quality while Hierarchical algorithm shows more quality and k -mean algorithm is good for large dataset while Hierarchical is good for small dataset.

In K-Means, following steps are necessary:

1. Choose number of k clusters
2. Assign each observation Xi to the closest cluster centroid μk
3. Compute and place new centroid for each cluster.
4. Repeat the above step until it results in no data point moving from one cluster to another.

There are two popular methods to determine the optimal value of k. Elbow Curve and silhouette score both helps in determining the optimal value of k in k-means clustering but the choosing proper value of k also depends on the business scenario that should be considered what value of the business domain is best fit.

The k-means clustering process is very sensitive to the presence of outliers in the data. Standardizing/scaling data helps to change the value of all numeric variables to specific range between 0 and 1 or similar range.

Type of Hierarchical linkages

- **Single Linkage**: The distance between 2 clusters is defined as the shortest distance between points in the two clusters
- **Complete Linkage**: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
- **Average Linkage**: Here, the distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

Question 3: Principal Component Analysis

PCA is a technique to reduce number of dimensions in a large data set and make smaller data set.

Basis is a unit in which we express the vectors of a matrix. For example, if we say a person with weight of 40 kg. Now 1 kg is the unit by which we express the weight of a person or thing. Change of basis or basis transformation is a way of representing the same observation into another basis or unit. Like we can express the weight of a person either in Kg or pounds. Looking the following image,

Patient ID	Height (cm)	Weight (kg)
p1	165	55
p2	155	71
p3	165	88
p4	160	105
p5	160	94

⟷

Patient ID	Height (ft)	Weight (lbs)
p1	5.4	121.3
p2	5.1	156.5
p3	5.4	194.0
p4	5.2	231.5
p5	5.2	207.2

Basis $\left\{ \begin{bmatrix} 1cm \\ 0\ kg \end{bmatrix}, \begin{bmatrix} 0cm \\ 1\ kg \end{bmatrix} \right\}$ $\left\{ \begin{bmatrix} 1ft \\ 0\ lbs \end{bmatrix}, \begin{bmatrix} 0ft \\ 1\ lbs \end{bmatrix} \right\}$

The basis vectors for the representation of the patient's information is given by $\left\{\begin{bmatrix}1\text{ft}\\0\text{ lbs}\end{bmatrix}, \begin{bmatrix}0\text{ft}\\1\text{ lbs}\end{bmatrix}\right\}$ to change from Kg to pound but the basis for the transform from Pound to Kg $\left\{\begin{bmatrix}1\text{cm}\\0\text{ kg}\end{bmatrix}, \begin{bmatrix}0\text{cm}\\1\text{ kg}\end{bmatrix}\right\}$

So $\begin{bmatrix}165\\55\end{bmatrix}$ is the same as $\begin{bmatrix}5.4\\121.3\end{bmatrix}$ when different basis vectors are being used.

Variance as information or variance is information is a measure of the importance of a column by checking its variance value. If a column has more variance of 0 then the column values are same but if a column has a value larger than 0 that mean the column contains more information.

Following are some of the shortcomings of PCA:

- PCA is limited to linearity.
- PCA needs the components to be perpendicular, though in some cases, that may not be the best solution.
- PCA assumes that columns with low variance are not useful, which might not be true in prediction setups

PCA is useful for Image segmentation, Recommender Systems, and patient statistics administered daily in a hospital.

The End