Machine Learning 1

Clustering & Principal Component Analysis

Questions

Word Limit: 200 - 300 words

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)

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.

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.

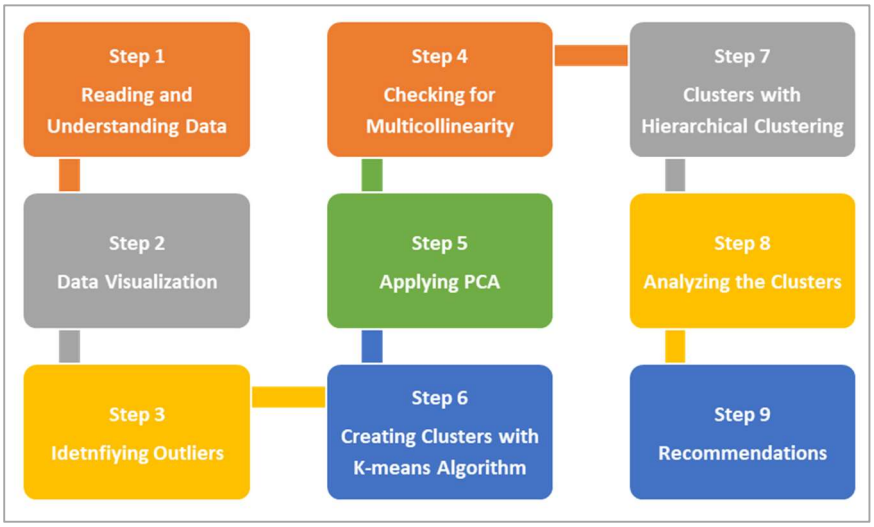
c. State at least three shortcomings of using Principal Component Analysis.

Assignment Summary

Client: HELP International (an international humanitarian NGO committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities)

Business Question: How to use the $ 10 million funds raised effectively by choosing countries that are in the direst need of aid.

Solution Methodology: The country data is provided with 9 different parameters: child mortality rate, net income per person, GDP per capita, inflation rate, life expectancy, fertility rate, import, export, spend on health, all of which are numeric variables.



There are 167 rows and 10 columns. There are no null values. All the attributes of the country are expressed as numeric value and there is no need to change the data type. No quality issues in the data provided. Since all attributes are numeric, we used distribution plot and box plot, to get insights on any skew in the data, presence of outliers etc. There are not many outliers, except for the variable gdpp (25 outliers). However, removing the outliers may result in some countries missing out the chance of getting the grant. So we have decided not to remove any outliers.

By doing a multivariate analysis with pair plot and heat map, we could see that: (a) income and gdpp has strong positive correlation; (b) child mortality and fertility rate has strong positive correlation; (c) both child mortality and fertility rate has strong negative correlation with life expectancy; (d) imports and exports has strong positive correlation. By doing a PCA, we can reduce the dimensionality with a fewer non correlated variables that are linear combination of the original variables and preserve maximum information. From the scree plot we could see that 94.5% of the variance is explained by 5 principal components. The model is built with 5 principal components.

With K-means Clustering Algorithm. To find the optimal number of clusters, we used the elbow curve and silhouette analysis; and both of them suggested 5 clusters. However, one of the clusters created had only one country. Probably this is because we did not remove the outliers. So we decided to create a model with 4 clusters. The result from K-means clustering was further used as it had clear groupings as follows: one of the clusters had 3 countries which were outliers, whereas the other 3 clusters correspond to the developed countries (30), developing countries (87) and under-developed countries (47). To identify the countries in dire need, I opted for countries in the cluster of under-developed countries with low gdpp and high child mortality rate.

Clustering

2.a. K-means Clustering & Hierarchical Clustering - Comparison

K-means Clustering	Hierarchical Clustering
Three steps: Initialization (randomly choose k points as centroids), Assignment (assign the different points to the nearest centroid) and Optimization (compute the centroid as an average of the points in the cluster). The assignment and optimization steps are repeated till the solution converges (i.e. the centroid do not change further)	In Hierarchical Clustering we can either choose a divisive approach (top-down: start with one single cluster of all data points and divide them in each step) or agglomerative approach (bottom up: start with n clusters by assigning each data point to its own cluster and start merging those which are close to each other).
Suitable when our data set is big.	Suitable when we have a smaller data set.
Non-linear Process.	Linear Process.
Time complexity: O(n)	Time complexity: $O(n^2)$
Not as computationally intensive as the hierarchical clustering algorithm.	Computationally Intensive. Every iteration processes the entire data and consumes a lot of computer's memory (RAM).
We need prior knowledge of the value of K i.e. we need to specify the number of clusters to be formed (and also the number of iterations)	No need to know the number of clusters. We arrive at the clusters by cutting the final dendrogram as per the requirement.
K-means may give a sub optimal solution if (a) you pick a K value too high or low (b) the data is not properly distributed or (c) the choices for initializing centroids goes wrong.	Produces better and more intuitive results compared to the k-means algorithm.
The algorithm is non-deterministic. The output varies each time we run the algorithm with different initial choices of cluster centers.	The results are reproducible i.e. we get the same result every time.

2.b. K-means Clustering Algorithm

Initialization: The first step in K-means Clustering algorithm is to randomly choose some points as centroids. We have to be careful not to choose these points too close to each other or too far away from our data points. For smart initialization of centroids we use K-means++ algorithm.

- One of the data points is randomly chosen as a cluster center.
- Compute the distance between this data point and all other data points.
- Choose the next cluster center such that it is as far apart as possible from the already chosen cluster center(s). The probability of choosing a point as centroid is directly proportional to its distance from the nearest, previously chosen centroid.
- Repeat this step until we have K cluster centers.

Assignment: In this step, each data point is assigned to the closest cluster center. To do this we compute the Euclidean distance of the cluster centers from each point and choose the cluster center for which the distance measure is minimum. If there are 2 points X and Y with n dimensions each, i.e. X = (X_1, X_2, … X_n) and Y = (Y_1, Y_2, .. Y_n), the Euclidean Distance between them is:

$$D = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots . + (X_n - Y_n)^2}$$

Each point X_i is assigned to the closest cluster center (μ_k). The cost function to be optimized is

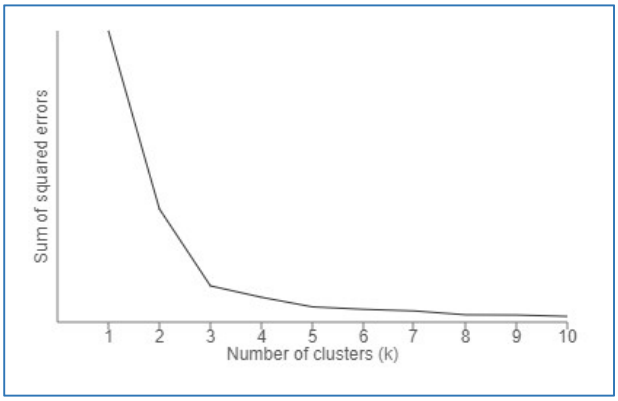
$$J = \sum_{i=1}^{n} \| X_i - \mu_{k(i)} \|^2$$

Optimization: The next step is to update each cluster center to the centroid i.e. the mean of the points assigned to it. μ_k = (X_1 + X_2 + X_3 .. + X_k) / n_k where k is the number of points assigned to the kth cluster. The assignment and optimization steps are repeated till the solution converges and the centroids do not further change. The number of iterations required for convergence depends on the choice of the initial cluster centers.

2.c. Choosing the value of K

In K-means clustering, we need to first choose the value of K i.e. the number of clusters need to be created. There are statistical methods to choose the value of K. However, if a statistical method suggests that we create 150 clusters then it will not make any business sense. We cannot handle that many segments. So what we do is to create different models with 2-15 segments and then profile the segments to see which one makes sense. This is very easy to do if we do a hierarchical clustering and observing the dendrogram to see where it makes sense to cut the dendrogram.

Statistically we use two techniques to find the optimal number of clusters:



Elbow curve: For different values of k (2, 3, 4 .... 10 and so on) we apply the algorithm and then compute the sum of the squared distances (SSD) of the points to their closest cluster center. We plot these SSD against the number of clusters, which gives us the elbow curve. We select the point at which the marginal decrease (or the rate of drop) in the SSD value starts to diminish, which means that adding a new cluster is not adding any value the model.

Silhouette score: The silhouette of a data point is a measure of how similar it is to data points within its cluster and how dissimilar it is matched to the data points in other clusters. Let q be the mean of the intra-cluster distance of a data point to all the points in its own cluster and p be the mean of the inter-cluster distance of the data point to the points in the nearest cluster that the data point is not part of. The silhouette score is computed as **(p - q) / max(p, q)**. The value of this score lies between 1 and -1. A score closer to 1 indicates that the point is very similar to other data points in its own cluster, whereas a score closer to -1 indicates intra-cluster dissimilarity.

2.d. Scaling / Standardization of Data

The K-means algorithm uses the Euclidean Distance Measure, which requires computing the distance between two data points or the data point and the respective cluster center. If the independent variables we use in the data are not in the same scale, then it will affect these distance measures. Suppose we use age and income as two parameters to form the cluster, and suppose the age is in the range (20, 60) whereas income is in the range of (20000, 80000), the magnitude of the income variable will severely affect the distance measure.

Let us compute the distance measure for two points (25, 50000) and (30, 45000):

$$D = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2} = \sqrt{(25 - 30)^2 + (50000 - 45000)^2} = \sqrt{5^2 + 5000^2} = 5000.0025$$

The distance measure is influenced by the income variable with a greater magnitude. So it is required to apply scaling or standardization.

Standardization is a method which scales the data in such a way that the mean $\bar{x} = 0$ and the standard deviation $S = 1$. The x value is replaced with the Z-score computed as follows: $$x' = \frac{x - \bar{x}}{S}$$	Another method to scale the data is **min-max scaling**, which uses the following formula: $$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$ x is the original value and x' is the normalized value. The scaled value is compressed within the range [-1, 1].

2.e. Linkages in Hierarchical Clustering

In Hierarchical Clustering, we use the agglomerative approach to start with as many clusters as there are data points and then start merging the clusters that are closest to each other. There are three methods to compute the distance between two clusters.

Single Linkage: The distance is computed as the shortest distance between points in the two clusters. If we represent the points in cluster r as x_r and the points in cluster s as x_s, then in single linkage we compute the distance between the clusters as $\min(D(x_{ri}, x_{sj}))$. Dendrograms produced using single linkage may not be structured properly. So to get a proper tree-like structure we use complete linkage or average linkage.

Complete Linkage: The distance is computed as the longest distance between points in the two clusters i.e. $\max(D(x_{ri}, x_{sj}))$

Average Linkage: The distance is computed as the average distance between every point of one cluster to every other point of the other cluster.

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

---

3. Principal Component Analysis

Principal component analysis (PCA) is one of the most commonly used techniques for **dimensionality reduction** i.e. to convert large data sets into smaller ones containing fewer variables which are **not correlated** with each other and derived as **linear combinations of the original variables** there by **capturing maximum information** in the dataset.

3.a. Three Applications of using PCA

Using PCA for dimensionality reduction (i.e. converting large data sets into smaller ones containing fewer variables) has many applications:

1. To create uncorrelated features that can be input into a prediction model (linear regression or logistic regression). This makes the modelling process faster and more stable. It also solves the problem of multicollinearity without losing any information.
2. For data visualisation and EDA: PCA makes it easy to explore and visualize data.
3. Finding latent themes in the data: To find patterns in data of high dimensionality in various fields like finance, bioinformatics, psychology etc. For example, with the ratings given to different movies by Netflix users, PCA would be able to find latent themes like genre and, consequently, the ratings that users give to a particular genre.
4. For noise reduction in facial recognition, computer vision and image compression.

3.b. Building blocks of PCA

PCA is used for dimensionality reduction through feature extraction from the original variables; the new features (called principal components) are non-correlated. To obtain these principal components there are two mathematical concepts that we need to apply.

Basis transformation: If B_1 is a set of basis vectors and B_2 is another set of basis vectors. Assuming that B2 and B1 are written with respect to the standard basis vectors, to move from B_1 to B_2, we use the equation $B_1 = M\ B_2$ (the representation of B_1 in B_2 is equal to M) or $M = {B_2}^{-1}.B_1$

Each feature in the data set can be represented as vectors. In PCA, we transform the original data set so that the eigenvectors are the basis vectors and find the new coordinates of the data points with respect to this new basis. Principal components are simply the eigenvectors of the covariance matrix used as basis vectors. Each of the original data points is expressed as a linear combination of the principal components, giving rise to a new set of coordinates.

Variance as information. When the farthest points of the data are projected on the axis, the length of the projection becomes proportional to the variance. When the variances are unequally distributed among the original features or columns i.e. some columns have much less variance than others, it is easier to remove those columns and do dimensionality reduction. However this

may result in the loss of information. In PCA, we change the basis in such a way that the new basis captures the maximum variance.

3.c. Three shortcomings of using Principal Component Analysis.

PCA is limited to linearity i.e. the principal components are linear combination of the original variables; and hence, works well with linear models such as linear regression, logistic regression etc. However, it can be used with non-linear models for improving computational efficiency.

PCA needs the components to be perpendicular i.e. non correlated. PCA also requires the data to be highly correlated for it to create reasonable results. PCA produces components which are orthogonal and uncorrelated, but sometimes, correlated components can be the better choice. In such cases, we need to use Independent Component Analysis (ICA).

PCA assumes that columns with low variance are not useful. Hence it may lead to loss of valuable classes or variables in supervised learning procedures. Sometimes, when we have a set of non-correlated variables, even a column with low variance may be significant. For example in classification problems where the data has high imbalance (spam-ham classification, or fraud detection) certain columns of low variance may be significant to get the final output. So we should not use PCA to forcefully reduce the dimensionality.