PCA & Clustering Assignment

Name	**Rahul Kumar**
Email	**Rahul.cs068@gmail.com**
Phone	**+91-9538109454**
Course	**PGDDS**
Batch	**Dec 2018**

Contents

Comprehension

This part of the assignment is subjective and hence, you are required to write the answers and submit them in a PDF file. For writing normal text, you can use MS Word — or any other similar software which can convert word files to the .pdf format. For writing equations, drawing figures, etc. you can do so on a blank sheet of paper, photograph the images and upload them in the same word document.

Please limit your answers to 200-300 words per question. While calculating values, ensure you write all the necessary steps and formulae. Also, use the correct terminology to present the solution.

Question 1:

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 1:

Problem Statement:

HELP NGO 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 project that included a lot of awareness drives and 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. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

And here as data analyst, our job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then we need to suggest the countries which the CEO needs to focus on the most

To work for this problem, we would follow the below solution approach:

1. Read the data from the CSV file
2. Understand and clean the data
3. Scale the data using StandardScaler method
4. Do dimension reductionality using standard PCA method
5. Use Incremental process to see if there is any improvements in the dimension reduction
6. Identify outliers and clean them
7. Choose a random number for the K and do categorize countries
8. Use Silhouette Analysis to find the value of K
9. Use the value obtained from step 8 and do K-means clustering to categorize countries
10. Finally hierarchical clustering to categorize countries

As per above defined steps/approach, we loaded, cleaned and scaled the data using Scalar algorithm.

Before we do clustering, it is required to do PCA on the data to see whether the co-related factors can be removed and use only non-co-related data for the analysis.

Number of factors given in the problem are 9 and we could reduce this to 4 using PCA as these 4 PCA's explains 87% of the entire data.

- Child_mort
- Exports
- Health
- Imports

This PCA data is used in K means and Hierarchical Clustering Algorithm.

IQR method is used to remove outliers from the PCA dataset.

HopKins Analysis is used to check whether the clustering is required for the data are not and we found that HopKins value is greater than 0.7 and data require clustering

K-Means Clustering

We choose a random K to do clustering of the data. We use the libraries form sklearn to categorize the countries.

Randomly 4 is chosen for K and countries are clustered among these 4 categories.

- Cluster 1 - Countries which are economically poor and require large amount of help
- Cluster 0 - Countries which are economically above poor and require some amount of help
- Cluster 2 - Countries which are economically average and require very minimal amount of help
- Cluster 3 - Countries which are economically strong and may not require any help

To verify whether the chosen K is right or not we used Silhouette analysis which gave optimal k as 3.

We choose a K = 3 from Silhouette Analysis and do the clustering of the data.

Countries are categorized as 3 clusters

- Cluster 2 - Countries which are economically poor and require large amount of help
- Cluster 1 - Countries which are economically average and require some amount of help
- Cluster 0 - Countries which are economically strong and may not require any help

Dendrogram with method = ‘complete’ is used build the tree and we used Divisive clustering to cut the tree.

Using HC method also the countries are clustered.

Question 2:

State at least three shortcomings of using Principal Component Analysis.

Answer 2:

1. Independent variables become less interpretable: After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.

2. Data standardization is must before PCA: You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.

For instance, if a feature set has data expressed in units of Kilograms, Light years, or Millions, the variance scale is huge in the training set. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, principal components will be biased towards features with high variance, leading to false results.

Also, for standardization, all the categorical features are required to be converted into numerical features before PCA can be applied.

PCA is affected by scale, so you need to scale the features in your data before applying PCA. Use StandardScaler from Scikit Learn to standardize the dataset features onto unit scale (mean = 0 and standard deviation = 1) which is a requirement for the optimal performance of many Machine Learning algorithms.

3. Information Loss: Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features.

Question 3:

Compare and contrast K-means Clustering and Hierarchical Clustering

Answer 3:

Hierarchical clustering

It creates a hierarchy of clusters which may be represented in a tree structure called a dendrogram. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations.

Algorithms for hierarchical clustering are generally either agglomerative, in which one starts at the leaves and successively merges clusters together; or divisive, in which one starts at the root and recursively splits the clusters.

Any non-negative-valued function may be used as a measure of similarity between pairs of observations. The choice of which clusters to merge or split is determined by a linkage criterion, which is a function of the pairwise distances between observations.

Cutting the tree at a given height will give a clustering at a selected precision. In the following example, cutting after the second row will yield clusters {a} {b c} {d e} {f}. Cutting after the third row will yield clusters {a} {b c} {d e f}, which is a coarser clustering, with a smaller number of larger clusters.

Agglomerative hierarchical clustering

For example, suppose this data is to be clustered, and the Euclidean distance is the distance metric.

This method builds the hierarchy from the individual elements by progressively merging clusters. In our example, we have six elements {a} {b} {c} {d} {e} and {f}. The first step is to determine which elements to merge in a cluster. Usually, we want to take the two closest elements, according to the chosen distance.

Optionally, one can also construct a distance matrix at this stage, where the number in the i-th row j-th column is the distance between the i-th and j-th elements. Then, as clustering progresses, rows and columns are merged as the clusters are merged and the distances updated. This is a common way to implement this type of clustering, and has the benefit of caching distances between clusters. A simple agglomerative clustering algorithm is described in the single-linkage clustering page; it can easily be adapted to different types of linkage (see below).

Suppose we have merged the two closest elements b and c, we now have the following clusters {a}, {b, c}, {d}, {e} and {f}, and want to merge them further. To do that, we need to take the distance between {a} and {b c}, and therefore define the distance between two clusters. Usually the distance between two clusters and is one of the following:

- The maximum distance between elements of each cluster (also called complete-linkage clustering):
- The minimum distance between elements of each cluster (also called single-linkage clustering):
- The mean distance between elements of each cluster (also called average linkage clustering, used e.g. in UPGMA):
- The sum of all intra-cluster variance.
- The increase in variance for the cluster being merged (Ward's criterion).
- The probability that candidate clusters spawn from the same distribution function (V-linkage).

Each agglomeration occurs at a greater distance between clusters than the previous agglomeration, and one can decide to stop clustering either when the clusters are too far apart to be merged (distance criterion) or when there is a sufficiently small number of clusters (number criterion)

K-means clustering

The k-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

Example: The data set has three dimensions and the cluster has two points: X = (x1,x2,x3) and Y = (y1,y2,y3). Then the centroid Z becomes Z = (z1,z2,z3), where

The algorithm steps are:

- Choose the number of clusters, k.
- Randomly generate k clusters and determine the cluster centers, or directly generate k random points as cluster centers.
- Assign each point to the nearest cluster center, where "nearest" is defined with respect to one of the distance measures discussed above.
- Recompute the new cluster centers.
- Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed).

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments (the k-means++ algorithm addresses this problem by seeking to choose better starting clusters). It minimizes intra-cluster variance, but does not ensure that the result has a global minimum of variance. Another disadvantage is the requirement for the concept of a mean to be definable which is not always the case. For such datasets the k-medoids variants is appropriate. An alternative, using a different criterion for which points are best assigned to which centre is k-medians clustering.

----- End of File -----