PCA and Clustering Q and A

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)

Ans: In this assignment “Clustering of countries” our main aim was to determine countries with direst need for an aid with respect to the features (child mortality, income and gdpp). We tried to understand the data in the beginning and inferred that there were 166 countries in the data set which had about 9 features each. Since we already knew about our main features to work with, it was a lot easier to start with the assignment. We started this assignment with data understanding and data cleaning and preparation. We tried to understand the structure of the dataframe after import data from the csv. We looked out for null values in the dataset and also for outliers.

We performed data standardization and performed PCA as it was told in the pre-assignment session. We carried out outlier analysis right after PCA and found out that it resulted into truncation of two countries from our dataset right away, which were Myanmar due to low import value and Nigeria due to inflation. We ran PCA and found out that 5 clusters were enough at PCA(0.94). We started with clustering (Hopkins Statistics) and found out that our data had high tendency to cluster. Later on we performed silhouette score analysis to choose optimum number of clusters and did also perform a sum of squared error test. We had to plot out dendograms for (single and complete) to perform hierarchical clustering. We inferred from the dendograms that clusters can either be 3 or 4 for this particular analysis so first we went with 4 clusters to be precise. When we performed K-means with 4 clusters we found out that cluster 1 was suffering badly in all three fronts (child mortality, income and gdpp). We also did a line plot analysis and a bar plot analysis (feature wise) to study the cluster closely.

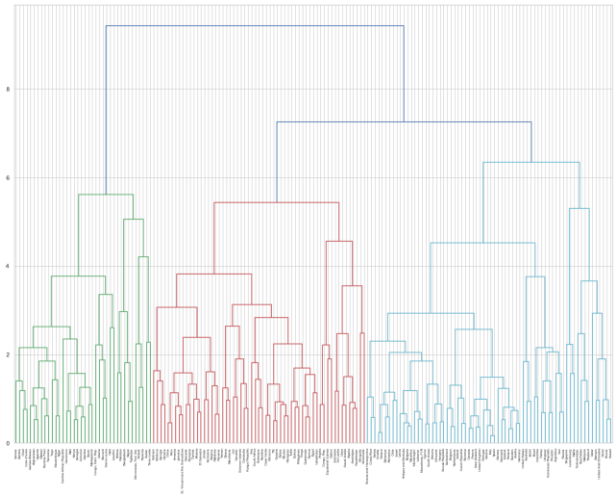
Since we weren’t convinced with 4 cluster analysis, we also went with 3 cluster analysis and found out that cluster 1 repeated its pattern and was same as how it was in 4-cluster analysis. We tried to list out all the top n countries suffering in (child mortality, income and gdpp) and recommended common countries from the 3 list.

Question 2: Clustering

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

Hierarchical clustering, as the name suggests is an algorithm that builds hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.

The results of hierarchical clustering can be shown using dendrogram. The dendrogram can be interpreted as:



K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*. The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the *K*-means clustering algorithm are:

1. The centroids of the *K* clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined.

Each centroid of a cluster is a collection of feature values which define the resulting groups. Examining the centroid feature weights can be used to qualitatively interpret what kind of group each cluster represents.

Difference between K-means clustering and hierarchical clustering

- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. $O(n^2)$.
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

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

The way K-means algorithm works is as follows:

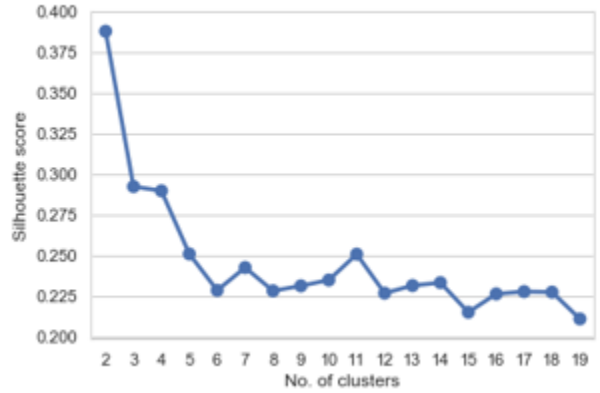
1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting Kdata points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids.

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

There is a popular method known as **elbow method** which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing *k*. As the value of *K* increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the **elbow point**.

Example

```
Text(0, 0.5, 'Silhouette score')
```



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

The issue is what represents a good measure of distance between cases.

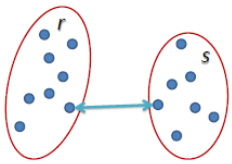
If you have two features, one where the differences between cases is large and the other small, are you prepared to have the former as almost the only driver of distance?

So for example if you clustered people on their weights in kilograms and heights in meters, is a 1kg difference as significant as a 1m difference in height? Does it matter that you would get different clustering's on weights in kilograms and heights in centimeters? If your answers are "no" and "yes" respectively then you should probably scale.

- k-nearest neighbors with a Euclidean distance measure is sensitive to magnitudes and hence should be scaled for all features to weigh in equally.
- Scaling is critical, while performing Principal Component Analysis (PCA). PCA tries to get the features with maximum variance and the variance is high for high magnitude features. This skews the PCA towards high magnitude features.

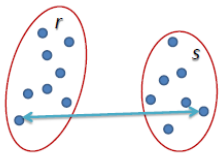
e) Explain the different linkages used in Hierarchical Clustering.

Single Linkages: In single linkage hierarchical clustering, the distance between two clusters is defined as the *shortest* distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.



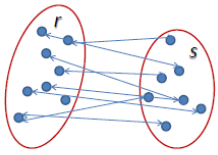
$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Complete Linkages: In complete linkage hierarchical clustering, the distance between two clusters is defined as the *longest* distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two furthest points.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Average Linkages: In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters "r" and "s" to the left is equal to the average length each arrow between connecting the points of one cluster to the other.



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

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

Data visualization is central to the Principal component analysis (PCA), it equally allows you to understand the data and reduce the dimension of data.

Applications of using PCA:

1. In MRI's and tensor imaging the data's dimension can be reduced using PCA.
2. For mapping atomic structures of elements we can use PCA to reduce dimensions.
3. For perfusion maps for blood flow distributions we can PCA to reduce dimensions.

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

PCA(Basis transformation) yields the directions (principal components) that maximize the variance of the data, whereas LDA also aims to find the directions that maximize the separation (or discrimination) between different classes, which can be useful in pattern classification problem (PCA "ignores" class labels).
In other words, PCA projects the entire dataset onto a different feature (sub) space, and LDA tries to determine a suitable feature (sub) space in order to distinguish between patterns that belong to different classes.

Variance :- PCA yields a feature subspace that maximizes the variance along the axes, it makes sense to standardize the data, especially, if it was measured on different scales. Although, all features in the Iris dataset were measured in centimeters, let us continue with the transformation of the data onto unit scale (mean=0 and variance=1), which is a requirement for the optimal performance of many machine learning algorithms.

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

PCA is focused on finding orthogonal projections of the dataset that contains the highest variance possible in order to 'find hidden LINEAR correlations' between variables of the dataset. This means that if you have some of the variables in your dataset that are linearly correlated, PCA can find directions that represents your data. Imagine two variables that represents the size of something in cm and inch respectively (the values of those variables are correlated by the formula 2.54 cm = 1 inch), if you add noise and plot the data you will get something similar to this picture:

but if the data is not linearly correlated (f.e. in spiral, where x=t*cos(t) and y =t*sin(t)), PCA is not enough.

- Relies on orthogonal transformations

Sometimes consider that principal components are orthogonal to the others it's a restriction to find projections with the highest variance:

- Large variance = low covariance = high importance

This assumption depends of what problem you want to solve:
* If you want to compress or remove noise from your dataset this assumption is an advantage
* for mostly any other problem (like Blind Source Separation) it is not useful. Based on Independent Component Analysis theory: uncorrelated is only partly independent.

- mean and covariance doesn't describe some distributions

There are many statistics distributions in which mean and covariance doesn't give relevant information of them. In fact, mean and covariance are used (or could be considered important) for Gaussians.

- scale variant

PCA, as you could've seen, is a rotation transformation of your dataset, which means that doesn't affect the scale of your data. That means that if you change the scale of just some of the variables in your data set, you will get different results by applying PCA