PCA and Clustering Assignment

Question 1: Assignment Summary

Ans:

In the assignment, we have a column called country which is the unique identifier in our data. We are supposed to find the countries which are struggling on various parameters. Following are the steps followed to find these countries.

Data understanding:

1. Get the data
2. Fix the data type
3. Fix the missing values

Perform PCA:

1. Data standardization
2. Perform PCA and choose the PCs that define more than 85% of the data variance
3. Run PCA with chosen number

Perform Clustering:

1. Data presentation for clustering
2. Outlier treatment
3. Hopkins check

Clustering:

1. K-means clustering
2. Hierarchical clustering

We perform all these steps to get the clusters of countries that need the most help.

Question 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.

Ans:

Difference between K-means and hierarchical clustering:

- Hierarchical clustering cannot 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

K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps:

1. Specify the desired number of clusters K: Let us choose k=2 for these 5 data points in 2-D space.

2. Randomly assign each data point to a cluster: Let's assign three points in cluster 1 shown using red color and two points in cluster 2 shown using grey color.
3. Compute cluster centroids: The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.
4. Re-assign each point to the closest cluster centroid: Note that only the data point at the bottom is assigned to the red cluster even though its closer to the centroid of grey cluster. Thus, we assign that data point into grey cluster
5. Re-compute cluster centroids: Now, re-computing the centroids for both the clusters.
6. Repeat steps 4 and 5 until no improvements are possible: Similarly, we'll repeat the 4th and 5th steps until we'll reach global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.

We choose the value of k-means in two ways:

1. Silhouette score method, where we select the value of k with maximum peak
2. Elbow curve method where we select the value of the elbow point
3. We can choose based on the business objective as well or randomly in some cases

Scaling/ standardization before performing clustering:

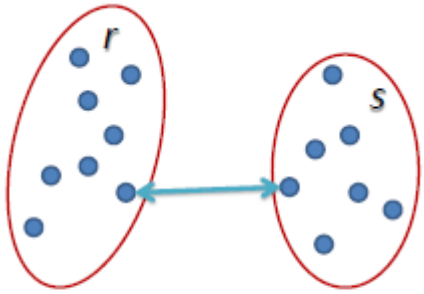
The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points in mathematical space. ... Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

Clustering algorithms such as K-means do need feature scaling before they are fed to the algo. Since, clustering techniques use Euclidean Distance to form the cohorts, it will be wise e.g to scale the variables having heights in meters and weights in KGs before calculating the distance.

Different linkages in Hierarchical Clustering:

Single Linkage

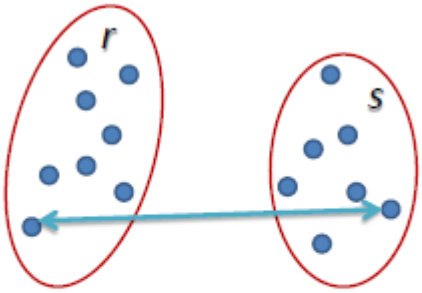
In single linkage hierarchical clustering, the distance between two clusters is defined as the *shortest* distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.



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

Complete Linkage

In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between tw 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.



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

Question 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.

Ans:
Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning. ... Models also become more efficient as the reduced feature set boosts learning rates and diminishes computation costs by removing redundant features.
Basis Transformation:
PCA simply takes points expressed in the standard basis and transforms them into points expressed in an eigenvector basis. In this process of transformation, some dimensions with low variance are discarded and hence the resulting dimensional reduction.
Variance as information:
In case of PCA, "variance" means summative variance or multivariate variability or overall variability or total variability. Below is the covariance matrix of some 3 variables. Their variances are on the diagonal, and the sum of the 3 values (3.448) is the overall variability.

Limitations of PCA:
the *limitations* of PCA, that is, when can you apply PCA on a dataset for dimensionality reduction. For this, we need to know about the assumptions that are used in PCA:

- Linearity: PCA assumes that the principle components are a linear combination of the original features. If this is not true, PCA will not give you sensible results.
- Large variance implies more structure: PCA uses variance as the measure of how important a particular dimension is. So, high variance axes are treated as principle components, while low variance axes are treated as noise.
- Orthogonality: PCA assumes that the principle components are orthogonal.