Assignment - 2

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)

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Solution:

Problem statement-HELP International a NGO has raised $10M and needs to decide on how to use this money effectively for countries in need of aid.

Approach: The below given steps were used to cluster the data into developed, developing and underdeveloped countries. Then based on the socio-economic and health factors that determined overall development of a country, 5 countries in dire need of aid were shortlisted

Step1 - Data Preparation:

- Checked for null values and duplicates.
- Outlier Treatment
- Feature Standardisation to scale the data.

Step 2 - Principal Component Analysis:

- Determined 5 principal components which captured 95% of variance.
- Correlation matrix confirmed that there were no correlation between the principal components.

Step 3 – Hopkins Statistics:

- Resulted in a score of greater than 70% indicating a good clustering tendency.

Step 4 – K-Means Clustering:

- Silhouette Analysis and Elbow curve method indicated approximately 2 and 3 values for K
- Visualization of 2 Principal Components indicated that a K value of 3 showed a better cluster segregation and made more sense to categorise the countries into developed, developing and underdeveloped.
- Results : Cluster0 : 47, Cluster1 : 29, Cluster2 : 79

Step 5 – Hierarchical Clustering:

- Complete Linkage method displayed the dendrograms clearly.
- Results : Cluster0 : 72, Cluster1 : 60, Cluster2 : 23

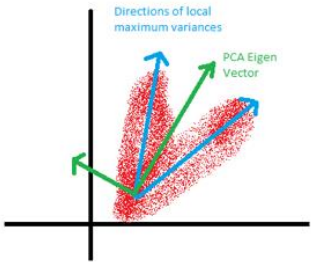
Step 6 – Results:

- Visualization with bar charts of all original variables indicated that the underdeveloped countries (cluster 0) had a high child mortality rate, very low income per person and low gdpp when compared to developed countries.
- The Hierarchical clustering did not seem to be as intuitive as K-Means and hence K-Means Clustering was preferred.
- Hence filtered the initial data (inclusive of outliers) by comparing each countries' child mortality, income and gdpp with the variables respective mean values and determined the 5 countries which are in dire need of aid

Question 2: State at least three shortcomings of using Principal Component Analysis.

Solution:

- PCA is limited to linearity. PCA has to be linear combinations of original features. There may be situations where data may need nonlinear combinations. PCA cannot be used in such situations.
- PCA requires principle components to be perpendicular or non-correlated or orthogonal. Sometimes the data that requires the correlated components to be present. PCA may not be useful in this situation. For e.g., in the diagram below, green colour vectors are principal components. But, the actual maximum variance directions are the blue colour vectors. PCA fails to find these vectors. However, Independent Component Analysis (ICA) works well for the above data and it gives the blue colour vectors as independent components



- PCA assumes that columns with low variance are not useful, which might not be true in prediction setups (especially classification problem with class imbalance like spam/ham classification or fraud detection)

Question 3: Compare and contrast K-means Clustering and Hierarchical Clustering.

Solution:

S.No	Description	K-Means Clustering	Hierarchical Clustering
1	Definition	K- Means is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. It is a division of objects into clusters such that each object is in exactly one cluster, not several.	Clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined together and continue to combine until all objects are in the same cluster.
2	Calculating number of Clusters	In K-Means, the desired number of cluster (K-Value) needs to be decided well ahead in the beginning.	In Hierarchical clustering, the dendrogram enables to determine the cluster number, however, the threshold value at which the dendrogram needs to be cut needs to be determined.
3	Repeatability	As K-means starts with a random choice of cluster centers, it may yield different clustering results on different runs of the algorithm.	Hierarchical clustering will mostly produce the same clustering results on each run.
4	Scalability & Flexibility	K means is scalable but cannot use for flexible data.	Hierarchical is Flexible but cannot be used on large data.

5	Run-Time Efficiency	K-Means clustering is a nonlinear process and it runs on an iteration value specified before. However, if it finds that all the centroids are converged before the specified iteration value is reached, it would stop further processing. Similarly in case where it has still not figured the convergence after execution of last specified iteration, it will inform that convergence point has not been reached and stop processing. Hence K Means clustering is relatively more efficient run-time wise and can be used for large data sets. It doesn't consume large amount of RAM.	Hierarchical clustering is a linear method and can be slow as it has to make several merge/split decisions at every iteration. Hence it requires huge RAM size. However, in normal systems with limited RAM size, hierarchical clustering can be used for small data size