Clustering & KMeans Assignment - II

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:

The Business problem of the assignment was around finding a group of countries (based on certain socio-economic factors) from the available data set that are in direst need of money. This list would help the NGO HELP International channel their funds appropriately so that countries in need get the benefit.

The key strategy followed here was to reduce the dimensionality of the features (10) using PCA (Principal Component Analysis) that would only consider the variables that explain the maximum variance in the dataset so that the results are not biased. The Scree plot was visualized plotting the cumulative variance captured by variables and as per the chart, 5 Principal components were picked which explained ~95% of the overall variance in the dataset. By plotting the first 2 Principal Components against the 10 available variables in the dataset, we found that variables child_mort, total_fer, gdpp, income and life_expec had highest magnitude along the first Principal Component, hence these variables were chosen to be used for comparision when the clusters have been formed.

Next, the Hopkins score on the data was tested to make sure that the data is good for clustering and there is no absolute randomness.

The Clustering was approached using KMeans and Hierarchical both the techniques. With the help of Silhouette score & Elbow curve, the appropriate number of clusters was chosen as 4.

So the entire data was divided into 4 groups with each group or the cluster exhibiting some features of the same group. Comparing the clusters with the mean values of the of the 5 variables found during PCA, we could find the countries that need the funds the most.

The same analysis was done using Hierarchical Clustering as well with 4 clusters and we got almost similar results for the list of countries. Upon contrasting the 2 menthods – KMeans & Hierarchical, I believe KMeans is more efficient as we iteratively optimize the clusters assigned and keep improving the clustering we do, whereas this is not the case with Hierarchical.

Question 2:

State at least 3 shortcomings of using Principal Component Analysis

Answer:

There are 3 major shortcomings of PCA which are as follows (Alternatives do exist with some pitfalls)

a) The Principal Components in PCA have to be linear combinations of the original columns

- ➔ But why do we have to limit ourselves to linearity when we can go non-linear? Going linear might not be possible always
- ➔ t-SNE is an alternative to PCA when we choose to go non-linear, but it is computationally very expensive

b) PCA requires the Principal Components to be uncorrelated/orthogonal/perpendicular to each other
- ➔ But sometimes the data demands that correlated components to represent the data to represent useful insights
- ➔ ICA (Independent Component Analysis) overcomes this drawback, and can be used as an alternative, but it is several times slower than PCA

c) PCA assumes that the low variance components are not very useful –
- ➔ In supervised learning situations, this can lead to loss of valuable information. This is especially true for highly imbalanced classes/variables. For eg., to classify email as spam or ham, we may have only a very few variables that contribute to describing an email as ham and may not exhibit high variance. With PCA we might tend to drop those variables accounting to low variance but this would lead to loss of information and hence would be incorrect

Despite some drawbacks, PCA is a very efficient and powerful technique to reduce complexity in data and discover patterns.

Question 3:

Compare and contrast K-Means Clustering & Hierarchical Clustering

Answer:

Difference between K-means & Hierarchical

K Means Clustering	Hierarchical Clustering
➔ K Means compares the distance of each point from the cluster centroid and not with other points. So this consumes comparatively less time	➔ Hierarchical compares distance of every point in one cluster to every point in the other cluster. So hierarchical clustering is time consuming
➔ K means needs the number of clusters to be pre-determined. The algorithm would generate those many clusters, but number of clusters may not be known in advance in many cases	➔ Hierachical does not need to pre determine the number of clusters. We can decide where to cut the dendogram(based on height or average distance between clusters). The point where we cut the tree would give us the number of clusters.
➔ In K means, iteratively the point moves to the most optimal cluster based on the minimum distance from the centroid	➔ Hierarchical is Linear. i.e, a point once allocated to a cluster remains with it, it cannot be moved to another cluster
➔ K-means calculates the distance of a point from the cluster Centroid and keeps moving the points in different clusters depending on this distance and hence does not consume	➔ Hierarchical would consume more memory. It calculates distances of every point from every other point in the dataset. Thus the entire distance

much memory and suitable for large volume of data.	information from the start of n clusters till everything converges to one cluster has to be maintained. Hence it is not suitable for large volume of data
➔ The output of K Means depends on the initial choice of clusters, hence we need to run the algorithm multiple times to arrive at the most optimal clusters. Choosing different cluster centroids will result in different clusters.	➔ This does not vary the output depending on what clusters we pick as each observation in the dataset is an independent cluster in the beginning.