Clustering and PCA Assignment

Assignment: Part II

Q 1:

Problem statement: A Well-Established NGO wants to provide basic amenities to the poor children and underdeveloped countries, which are need that aid. so, our task is to find list of countries which are desired aid from NGO.

Solution Methodology: we have reviewed available dataset. reviewed upon to apply unsupervised clustering algorithm to cluster data in similar behavior groups.

we assumed that basic factors of being under developing countries are less GDP, high child mortality, less income, less spends on health, less spend on education and less expend to export and imports.

Based on above points, we will select countries to provide funds to avail basic amenities to poor.

PCA Components: We took **4** PCA components on the base of **cumulative sum of variance ratio** diagram. First 4 components covering large percentage of area.

Clustering: In this assignment, we used two clustering method **KMeans** and **Hierarchica**l clustering. Both showing almost same countries name in same group but different cluster ids. Both methods have pros and cons.

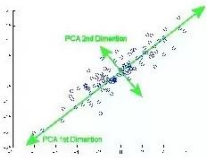
In KMeans, very under developing countries are in cluster id 0 while in Hierarchical same countries in cluster id 1.

KMeans, we chose three (k=3) clusters. On the base of **silhouette** score and **elbow curve**.

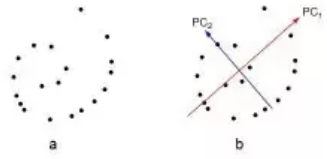
Q 2.

1.Relies on linear assumption:

If data is not linear correlated, PCA will fail to find hidden correlations between variables of the datasets.

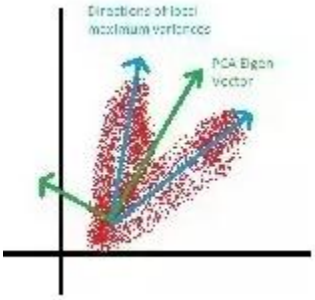


if data is spared as circle then there is lots of chance to loss information.



2**. Relies on orthogonal transformation:**

Its restriction to find projection with the Hight variance.



3. Scale variation:

if you change the scale of just some of the variables in your data set, you will get different results by applying PCA.

Q3.

K-Means: Advantages and disadvantages

Advantages:

- Ease to implement
- With many variables. K-Means may be computationally faster than hierarchical clustering (if k is small).
- k-Means may produce tither clusters than hierarchical clustering.
- An instance can change cluster (move to another cluster), when the centroids are recomputed.

Disadvantages:

- Difficult to predict the number of clusters (K-Values).
- Initial seeds have a strong impact on the results.
- Rescaling datasets will change results.

Hierarchical clustering: Advantages and Disadvantages

Advantages:

- Output will be a hierarchy. Means we can say structure is more informative than k-Means.
- This is also easy to implement.

Disadvantages:

- It's not possible to undo the previous step.
- Time complexity not suitable for large datasets
- Very sensitive to outliers.