PCA and Clustering Assignment :

Question 1: Assignment Summary : Briefly describe the "Clustering of Countries" assignment that you just completed within 200-300 words.

Answer : HELP International is an international humanitarian NGO that is committed to fighting poverty and providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities. It runs a lot of operational projects from time to time along with advocacy drives to raise awareness as well as for funding purposes.

Using the dataset given need to categorise the countries using some socio-economic and health factors that determine the overall development of the country and then need to suggest CEO of HELP international NGO which are the countries that are in the direst need of aid.
Here Firstly we treated the data that was provided to us by doing soft treatment of outliers. Once done , We proceeded towards doing the PCA on the dataset through which I decided to use 3 PCA components as more than 90% of variance was covered with 3 components only.

Later on we did two types of clustering technique on data that is KMeans and hierarchical clustering. For Kmeans we used silhouette and elbow method. Based on the outcomes we got in graph … We choose K(cluster as):3. For Hierarchical clustering we used dendrogram through which we decided to choose the cluster as 3.

Later on We found the impact of child mortality , income and gdpp on three clusters. Cluster 0 had the highest child mortality , lowest income and lowest gdpp. We had 35 countries in cluster 0 using kmeans and 46 countries in cluster 0 using hierarchical clustering.

So in Cluster 0 using both methods we found the 5 countries in need of aid :

Guninew-Bissau , Burkina Faso , Guinea , Togo and Afghanistan.

Question 2: Clustering

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

Answer :) Difference between K Means 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. $O(n^2)$

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

Answer :

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

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

Answer : There are two methods that can be used to find K in K means that is :

1. The Elbow Method : *calculate the Within-Cluster-Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an elbow.*
2. The Silhouette Method : *The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).*

The Elbow Method is more of a decision rule, while the Silhouette is a metric used for validation while clustering. Thus, it can be used in combination with the Elbow Method.

Therefore, the Elbow Method and the Silhouette Method are not alternatives to each other for finding the optimal K. Rather they are tools to be used together for a more confident decision.

D) Explain the necessity for scaling/standardisation before performing Clustering.

Answer : The idea is that if different components of data (features) have different scales, then derivatives tend to align along directions with higher variance, which leads to poorer/slower convergence. Standardization is an important step of Data preprocessing. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

E) Explain the different linkages used in Hierarchical Clustering.

Answer : Complete-link clustering can also be described using the concept of clique. Let dn be the diameter of the cluster created in step n of complete-link clustering. Define graph G(n) as the graph that links all data points with a distance of at most dn. Then the clusters after step n are the cliques of G(n). This motivates the term complete-link clustering.

Single-link clustering can also be described in graph theoretical terms. If dn is the distance of the two clusters merged in step n, and G(n) is the graph that links all data points with a distance of at most dn, then the clusters after step n are the connected components of G(n). A single-link clustering also closely corresponds to a weighted graph's minimum spanning tree.

Average-link (or group average) clustering (defined below) is a compromise between the sensitivity of complete-link clustering to outliers and the tendency of single-link clustering to form long chains that do not correspond to the intuitive notion of clusters as compact, spherical objects.

Question : a) Give at least three applications of using PCA.

Answer : PCA is predominantly used as a dimensionality reduction technique in domains like facial recognition, computer vision and image compression. It is also used for finding patterns in data of high dimension in the field of finance, data mining, bioinformatics, psychology, etc.

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

Answer : So 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. The fraction of variance explained by a principal component is the ratio between the variance of that principal component and the total variance. For several principal components, add up their variances and divide by the total variance.

C) State at least three shortcomings of using PCA.

Answer : 1. Independent variables become less interpretable: After implementing PCA on the dataset, your original features will turn into Principal Components. Principal Components are the linear combination of your original features. Principal Components are not as readable and interpretable as original features.

2. Data standardization is must before PCA: You must standardize your data before implementing PCA, otherwise PCA will not be able to find the optimal Principal Components.

For instance, if a feature set has data expressed in units of Kilograms, Light years, or Millions, the variance scale is huge in the training set. If PCA is applied on such a feature set, the resultant loadings for features with high variance will also be large. Hence, principal components will be biased towards features with high variance, leading to false results.

Also, for standardization, all the categorical features are required to be converted into numerical features before PCA can be applied.

PCA is affected by scale, so you need to scale the features in your data before applying PCA. Use StandardScaler from Scikit Learn to standardize the dataset features onto unit scale (mean = 0 and standard deviation = 1) which is a requirement for the optimal performance of many Machine Learning algorithms.

3. Information Loss: Although Principal Components try to cover maximum variance among the features in a dataset, if we don't select the number of Principal Components with care, it may miss some information as compared to the original list of features.