Clustering and PCA Assignment – Part II

Question 1: Assignment Summary

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)

Problem Statement: The CEO of the NGO needs to decide how to use this money strategically and effectively. The significant issues that come while making this decision are mostly related to choosing the countries that are in the direst need of aid.

Methodology:

1. Performed data conversion of columns export, health and imports from Percentage to number, as same as gdpp
2. Performed scaling for columns child_mort, exports, health, imports, income, inflation, life_expec, total_fer, gdpp and max scaling is 9.84
3. Performed PCA model SVD_solver to identify the number of dimensions reduce purpose
4. Generated Variance explained by PCA and cumulative for Scree Plot purpose
5. Performed dimension reduction by IncrementalPCA using K/Cluster 3
6. Performed Outlier
7. Calculated Hopkins statistic to ensure data is good (0.78)
8. Calculated Silhouette Analysis score to identify number of clusters is 2 (0.49)
9. Generated SSD/Elbow Curve plot to identify number of clusters is 3
10. Calculated KMeans clusters and added KclusterID column to the initial dataset for cluster profiling purpose
11. Performed Hierarchical cluster using complete Linkage
12. Generated dendrogram plot
13. Perform cut tree for clusters 3 and calculated cluster labels and added HClusterID column to the initial dataset for cluster profiling purpose
14. Performed mean for each variables/column for KclusterID and HClusterID and identified that least cluster for KclusterID is 2 and for HClusterID is 0

15. Performing inner join with the list of countries of KclusterID=2 and HClusterID=0 results 23 countries listed below
16. **Data Analysis** performed using KMeans and Hierarchical Clustering shows that the countries that re in direst need of aid are listed below, by taking into the consideration of Variables:
    a. child_mort: Death of children under 5 is more
    b. exports: Less % of goods and service
    c. health: Less % of spending
    d. imports: Less % of imports
    e. income: Less % of Income
    f. life_expec: Less average number of years a newborn would live
    g. gdpp: Less gdpp growth
17. **Country**: Afghanistan, Botswana, Comoros, Congo, Rep., Eritrea, Gabon, Gambia, Ghana, Iraq, Kenya, Lao, Liberia, Madagascar, Mauritania, Namibia, Pakistan, Rwanda, Solomon Islands, South Africa, Sudan, Tanzania, Uganda, Yeme

---

Question 2: Clustering

a) Compare and contrast K-means Clustering 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(n2).
- 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

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

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.

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

There are various techniques which are followed in order to get the exact value of k. The mean distance between the data point and the cluster is a most important factor which can determine the value of k and this method is common to compare.

There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.

d) Explain the necessity for scaling/standardization before performing Clustering.

Scaling/Standardization is the central preprocessing step in data mining, to standardize values of features or attributes from different dynamic range into a specific range.

Min-Max normalization is the process of taking data measured in its engineering units and transforming it to a value between 0.0 and 1.0. Whereby the lowest (min) value is set to 0.0 and the highest (max) value is set to 1.0. This provides an easy way to compare values that are measured using different scales or different units of measure.

e) Explain the different linkages used in Hierarchical Clustering.

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 3: Principal Component Analysis

a) Give at least three applications of using PCA.

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 transformation and variance as information.

Represents a change of basis and can be interpreted in several ways, such as the transformation of X into T by the application of P or by geometrically saying that P - which is the rotation and a stretch - transforms X into T or the rows of P, where {p1,…,pm} are a set of new basis vectors for expressing the columns of X.

The variances of components are sorted in decreasing order. By construction of PCA, the whole set of components keeps all of the original variance. The dimensions of the space are not then reduced but the change of axis allows a better representation of the data. Moreover, by retaining the q first principal components (with q<p), one is assured to retain the maximum of the variance contained in the original data for a q-dimensional space.

c) State at least three shortcomings of using Principal Component Analysis.

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.