Assignment Part 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)

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

Answer 1

The assignment was around trying to figure out countries who are in the direst need for financial aid for HELP International where they can allocate their funds. The initial dataset provided had a list of health and socio-economic features based on which we had to cluster the countries.

The first step involved **identifying the outliers and treating them**- however, only the upper section of the outliers (wealthy countries) were considered and treated as the bottom outliers are our area of interest. The features were standardized and brought to the same scale. We went ahead with **PCA** to deal with multicollinearity present in the data and help in dimensionality reduction.

Using the **scree plot**, we identified the optimal number of principal components as 5 which were able to explain over 90% of the variance in the data. The next step involved transforming the original features into these components and perform clustering.

The **Hopkin’s statistic** (~0.85) tells us that the data is good to be clustered, and the **silhouette analysis** (along with the **elbow curve analysis**) told us that the optimal number of clusters should be 4. We went ahead with both K-means and Hierarchical clustering. The K-means clusters were visualized for all 5 principal components, and the visualization for the dendrogram for hierarchical clustering was done at n=4.

The clusters were analyzed by first merging the PCA data with the original set of features, then calculating mean per cluster for every feature. The clusters 1 and 2 consisted of countries that were not doing so well, and with the help of **binning** based on the cluster means, we identified following countries for the NGO: *Niger, Central African Republic, Comoros, Gambia, Benin, Mali, Senegal, Kenya, Cameroon*. Both types of clustering returned very identical sets of countries.

Question 2

State at least three shortcomings of using Principal Component Analysis.

Answer 2

1. Principal Component Analysis involves producing principal components which are linear combinations of the original variables. However, this may not be applicable in certain situations that require non-linear combinations, in which case a technique like t-SNE may be useful
2. Principal Component Analysis require the principal components to be perpendicular, orthogonal or uncorrelated. However, some situations may require us to include the correlated variables for modelling, in which case the alternative solution would be using Independent Components Analysis
3. Principal Component Analysis assumes that low variance data is not useful, however this may not be the case all the time. Particularly in the cases where we have class-imbalance, dropping one variable may significantly impact the performance of the model even if the variable may be having a low variance

----

Question 3

Compare and contrast K-Means Clustering and Hierarchical Clustering.

Answer 3

K-Means Clustering Algorithm is the process of dividing N-number of data points into K different groups or clusters.

The algorithm for the K-Means Clustering is as follows:

1. Start by choosing K random points, these will act as the initial cluster centers
2. Assign each data point to their nearest cluster center. This can be done using several distance metrics, however the most commonly used on is the Euclidean distance
3. For every cluster formed, compute the new cluster centers which will be the mean of all the respective cluster data points
4. Now, re-assign all the data points to the different clusters based on the distance metric by considering the new cluster centers
5. Iterate through the step 3 & 4 until there are no further changes

In the case of Hierarchical clustering, given N-number of datapoints, the steps in the hierarchical clustering are:

1. Calculate the NxN distance (similarity) matrix i.e. the distance from one datapoint to every other data point
2. Start by assigning each item to its own cluster, therefore, initially for N-datapoints there are N clusters
3. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster
4. Compute distances (similarities) between the new cluster and each of the old clusters
5. Repeat steps 3 and 4 until all items are clustered into a single cluster of size N
6. Then, based on the number of clusters that you want to create, say k, cut the dendrogram at the point which will return the k-clusters along with all the datapoints in the respective clusters

Now that we know how both the clustering algorithms work, let’s identify the **differences**:

- In the K-Means algorithm, the data was divided in the first step itself. In the subsequent steps, we refine our clusters to get the most optimal grouping. In hierarchical clustering, the data is not partitioned into a cluster in a single step. Instead, a series of partitions/merges take place, which may run from a single cluster containing all objects to n clusters that each contain a single object or vice-versa.
- The outputs of K-means is a set of data points that are assigned different clusters, whereas the immediate output of the hierarchical clustering is an inverted tree-shaped cluster, called a dendrogram
- For K-means, we need to decide the vale of K (number of clusters) beforehand, whereas this is not needed in the case of hierarchical clustering
- Hierarchical clustering is a linear process and hence it is computationally more expensive than K-Means, which is a non-linear method, therefore for larger sized data, K-means is preferred over hierarchical clustering