Subjective Assignment

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)

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

Answer:

HELP International is an international humanitarian NGO that gained a funding of $ 10 million with their recent funding programmes and their CEO needs to decide which countries are in the direst need of aid, which is what he needs us to find out.

We are going to categorise these countries based on some social-economic factors. To categorise data we decide clustering is the way to go about it. First, we analyse the data and perform scaling to get them all in the same range. We decide to perform Principal component analysis on the data to get better transformed variables which help better describe the data. To select the number of principal components, we plot a scree graph with the cumulative explained variance ratios and see at which number more than 90% of variance is covered.

The numbers of components were selected as 4 and the transformation was done, after which we treated the outliers. Once outliers were removed, we checked whether the data can be clustered or not the Hopkins measure.

As the Hopkins value turned out to be greater than 0.5 we performed the first type of clustering i.e.; K- Means clustering. To decide the number of clusters, we ran iteration over a range of number of clusters and compared their silhouette scores. We also plotted the elbow curve. We then built two models using 2 different values of k which were 2 and 3. We performed mean analysis on the clusters formed in both the models.

Next we began the hierarchical clustering. We built dendograms using the single and complete linkage and decided the where the dendogram is to be cut. We obtained the cluster labels by cutting the dendogram at an optimal value resulting in 3 clusters. We performed mean analysis on these clusters.

By comparing the mean analysis of clusters in all the above created models, we concluded that the K- means clustering algorithm by taking number of clusters as 3 clusters resulted in the most 3 distinct clusters.

The cluster with worst socio – economic factors like gdpp, income and child_mort was selected as the one where the CEO should focus. Some of the countries in this cluster were - 'Afghanistan', 'Benin', 'Cameroon', 'Comoros', "Cote d'Ivoire", 'Gambia', 'Guinea', 'Lao', 'Malawi', 'Mozambique', 'Sudan', 'Togo', 'Zambia'.

Question 2: Clustering

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

Answer:

- K means clustering requires us to decide the number of clusters beforehand. In Hierarchical Clustering we get to decide the number of clusters after interpreting a dendrogram.
- K means can handle large data sets whereas hierarchical clustering cannot, as it takes a longer time than k means to execute.
- K means has a time complexity of O(n) as it is linear whereas hierarchical has O(n^2) as it is quadratic.

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

Answer:

K means clustering algorithm is an iterative algorithm that partitions the dataset with N data points into k distinct non-overlapping clusters. The steps are as follows:

- Start by choosing K random points the initial cluster centres.
- Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
- For each cluster, compute the new cluster centre which will be the mean of all cluster members.
- Now re-assign all the data points to the diffrent clusters by taking into account the new cluster centres.
- Keep iterating through the step 3 & 4 until there are no further changes possible and the solution converges.

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:

Statistical Aspect:

- One of the metrics that is commonly used to compare results across different values of k is the mean distance between data points to the cluster centroid are used as a function of k is to plot a Elbow Curve. The elbow point, where the rate of decrease sharply shifts, can be used to choose the optimal value of k.
- Then there is the average silhouette score method in which the global maximum is the point that will give us the most appropriate number of clusters

Business Aspect:

- Depending on the business requirements and limitations, we can decide on the number of clusters. Sometimes, the business can have specific kind of cluster requirements where they know how many clusters they want to build from the data they have.

- Another example, if a company has budget for only 2 categories of people that come on their platform, then we need to come up with 2 distinct clusters that best differentiate these 2 set of people from each other to come up with 2 different strategies for each set based on their common behaviour.

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

Answer:

- Since we need to compute the Euclidean distance between the data points, it is important to ensure that the attributes with a larger range of values do not out-weight the attributes with smaller range. Thus, scaling down of all attributes to the same normal scale helps in this process.
- Different attributes will have measures in different units and standardisation helps in making these attributes unit-free and uniform.

e) Explain the different linkages used in Hierarchical Clustering.

Answer:

- **Single Linkage**: The distance between 2 clusters is defined as the shortest distance between points in the two clusters
- **Complete Linkage**: The distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters
- **Average Linkage**: The distance between 2 clusters is defined as the average distance between every point of one cluster to every other point of the other cluster.

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

1. Facial Recognition
2. Image segmentation/compression
3. Recommender Systems
4. Computer vision

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

Answer:

Basis Transformation – A basis for a vector space of dimension n is a set of n vectors (α1, …, αn), called basis vectors, with the property that every vector in the space can be expressed as a unique linear combination of the basis vectors. We usually work with more than one basis for a vector space and we can easily transform coordinate-wise representations of vectors and operators taken with respect to one basis to their equivalent representations with respect to another basis. These transformations are termed Basis transformation.

Variance as information – PCA assumes that the variables which have minimum variance are least significant and the variables with maximum variance explain the data better. To get the variables that cover maximum variance, PCA is performed which is an orthogonal linear transformation that transforms the data to a new coordinate system(Basis transformation) to get new components which explain the data better.

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

Answer:

- PCA is limited to linearity, although non-linear techniques such as t-SNE also exist.
- PCA needs the components to be perpendicular, though in some cases, that may not be the best solution.
- PCA assumes that columns with low variance are not useful, which might not be true in prediction setups (especially classification problem with a high class imbalance).