Subjective Questions for Clustering and PCA

Q:1 : Assignment Summery

Problem Statement:

Help CEO to make decision to choose the countries that are in the direst need of aid. Hence, My job as a Data analyst is to categorise the countries using some socio-economic and health factors that determine the overall development of the country. Then I need to suggest the countries which the CEO needs to focus on the most.

Solution Methodology :

Step 1 : Import libraries and load the data

Step 2 : Inspect the Dataset

Step 3 : Data Preparation

- 3.1 : Scaling the Data
- 3.2 : Perform PCA (We choose 5 components Because most of 98% information fall in 5 components
- 3.3 : Checking Outlier Based on PCA New Dataset

Step 4 : Perform K-means Clustering

- 4.1 : Perform the Hopekins Anakysis
- 4.2 : Finding the optimal number of Cluster using Two method
  - 4.2.1 : Using SSD
  - 4.2.2 ; Using Elow Curve (Based on Result we decide to choose k = 3)
- 4.3 : Perform Silhouette Analysis

Step 5 : Merge the Original Dats with Cluster Data

Step 6 : Analysis of K-Means

- 6.1 : Find the mean and concate the result
- 6.2 : Create Binning to find out Developed ,Developing,Poor Country
- 6.3 : Based on Binning We find the Poor country

Step 7 : Perform Hierarchical Clustering

7.1 : Perform Single & Complete Linkage

7.2 : Cut-Tree Based on Dendograms

7.3 Analysis the Hierarchical clustering

Step 8 : Conclusion :

- After comparing both K-means and Heirarchical clustering method, I am going with the K-means outcomes as the plots are clearly visible. As in both the methods, the top 8 under-developed countries are similar. I am considering the result of k-means outcome.

1. Burundi
2. Congo, Dem. Rep.
3. Niger
4. Sierra Leone
5. Central African Republic
6. Mozambique
7. Guinea-Bissau
8. Burkina Faso

Q:2 : Clustering

(A) Compare and contrast K-means Clustering and Hierarchical Clustering.

Hierarchical	K-means
Hierarchical clustering can't handle big data well	K Means clustering can Handel the Big Data
Hierarchical clustering is quadratic i.e. $O(n^2)$.	time complexity of K Means is linear i.e. O(n)
While results are reproducible in Hierarchical clustering.	In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ.

Hierarchical clustering by interpreting the dendrogram	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.

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

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

   3.1 Compute the sum of the squared distance between data points and all centroids.

   3.2 Assign each data point to the closest cluster (centroid).

   3.3 Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

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

The K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

Statistical Method :

1. Elbow method:-

1) Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance,by varying k from 1 to 10 clusters.

2) For each k, calculate the total within-cluster sum of square.

3) Plot the curve of wss according to the number of clusters k.

4) The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

2. Average silhouette Method :

1) Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance,by varying k from 1 to 10 clusters.

2) For each k, calculate the average silhouette of observations (avg.sil).

3) Plot the curve of avg.sil according to the number of clusters k.

4) The location of the maximum is considered as the appropriate number of clusters

Business Aspect :

1) Behavioral segmentation:

    - Behaviour of people

2) Attitudional Segmentation

    - Intension of customer

3) Demographic Segmentation

    - Gender

    - Age

- Location

- Income

- HouseType

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

In statistics, standardization (sometimes called data normalization or feature scaling) refers to the process of rescaling the values of the variables in your data set so they share a common scale.

Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit (e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000).

The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points.

When you are working with data where each variable means something different, (e.g., age and weight) so the fields are not directly comparable. One year is not equivalent to one pound, and may or may not have the same level of importance in sorting a group of records.

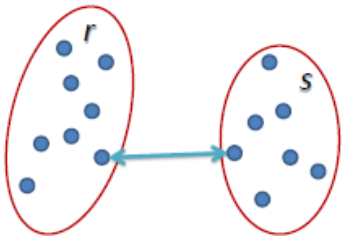
In a situation where one field has a much greater range of value than another it may end up being the primary of what defines clusters. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

(E) Explain the different linkages used in Hierarchical Clustering.

The **hierarchical clustering Technique** is one of the popular Clustering techniques in Machine Learning**.**

1) Single Linkage :

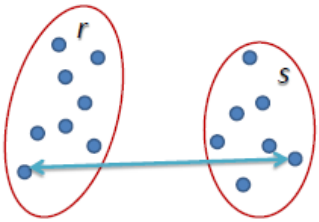
In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest Distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

2) Complete Linkage :

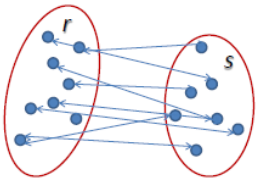
In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

3) Average Linkage :

In average linkage hierarchical clustering, the distance between two clusters is defined as theaverage distance between each point in one cluster to every point in the other cluster. For example,the distance between clusters “r” and “s” to the left is equal to the average length each arrow between connecting the points of one cluster to the other.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Q:3 Principal Component Analysis

(A) Give at least three applications of using PCA.

1) The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other

2) PCA is Widely used as a dimensionality reduction technique in domains like facial recognition, computer vision and image compression.

3) 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.

1) Variance :

The **Total Variance** is the sum of variances of all individual principal components.

The fraction of **variance explained** by a principal component is the ratio between the variance of that principal component and the total variance.

$$var(x) = \frac{\sum(x_i - \bar{x})^2}{N} \qquad cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

2) Basis Transformation :

In simple we can say that basis Trnasformation is to chage the basis of unit.

In statistics, Basis Transformation (sometimes called data normalization or feature scaling) refers to the process of rescaling the values of the variables in your data set so they share a common scale.

Often performed as a pre-processing step, particularly for cluster analysis, standardization may be important if you are working with data where each variable has a different unit

(e.g., inches, meters, tons and kilograms), or where the scales of each of your variables are very different from one another (e.g., 0-1 vs 0-1000).

The reason this importance is particularly high in cluster analysis is because groups are defined based on the distance between points.

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

1) PCA is limited to linearity, though we can use **non-linear techniques such as t-SNE** as well (you can read more about t-SNE in the optional reading material below).

2) PCA needs the components to be perpendicular, though in some cases, that may not be the best solution. The alternative technique is to use **Independent Components Analysis.**

3) 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).