Assignment: Part II – Rohit Keshari

Question 1: Assignment Summary

Problem Statement -

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.

After the recent funding programmes, they have been able to raise around $ 10 million. Now 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.

Answer(1):

The steps are below for this assignment:

Step 1: Reading and Understanding the Data

- Loading the data - Country-data
- Loading the data - Data-dictionary

Step 2: Data Preparation

- Shape of data
- Data information
- Data type information
- Duplicate country check:-check the spelling and unique country name
- Missing Value check

Step 3:EDA(Exploratory Data Analytics)

- Univariate analysis
- Heatmap
- Pairplot

Step 4: Data Preparation

- Outlier Analysis and Treatment
- Rescaling

Step 5: PCA(Principal Components Analysis)

- PCA on Scaled data
- Cumulative variance explained
- Principal Components Selection
- Apply the PCA on final data set with final number of principal components
- Visualize the PCA data

Step 6: Hopkins Statistics Test

- Hopkins Score Calculation

Step 7: Model Building - K-means Clustering

- K-means Clustering
- Silhouette Analysis
- Elbow Curves
- Decide the Cluster number
- Assign the cluster ID to country
- Visualize the data after assignment of cluster ID

Step 8: Model Building - Hierarchial Clustering

- Hierarchial Clustering – Single/Complete linkage
- Assign the cluster ID to country
- Visualize the data after assignment of ID
- Decide the clustering which you want to choose for final Analysis

Step 9: Final Analysis

- Final Country list Preparation on gdpp, child_mort and income

Step 10: Closing Statement and country list direst need of aid

Numbers of principal components are decided- : how many components are describing the more than 95 % of data.

Hierarchial Clustering Easy to implement, with a large number of variables, K--Means may be compute only faster than hierarchical clustering (if K is small).Kk--Means may produce better clusters than hierarchical clustering.

I would say hierarchical clustering is usually preferable, as it is both more flexible and has fewer hidden assumptions about the distribution of the underlying data.

Question 2: Clustering

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

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

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

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

e) Explain the different linkages used in Hierarchical Clustering.

Answer 2 (a)

- ***Time:***
  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).

- ***Shape of Clusters:***
  K-means works well when the shape of clusters is hyper-spherical (or circular in 2 dimensions). If the natural clusters occurring in the dataset are non-spherical then probably K-means is not a good choice.

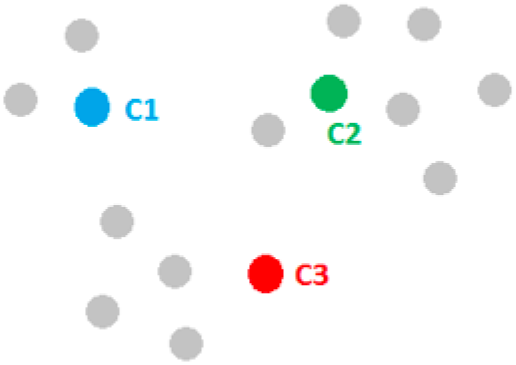
- ***Repeatability:***
  K-means starts with a random choice of cluster centers, therefore it may yield different clustering results on different runs of the algorithm. Thus, the results may not be repeatable and lack consistency. However, with hierarchical clustering, you will most definitely get the same clustering results.

Answer 2 (b)

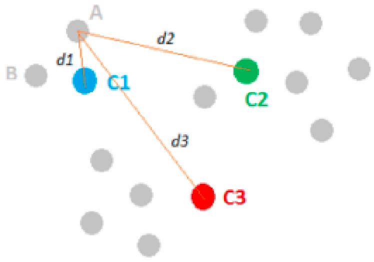
Clustering via K-means

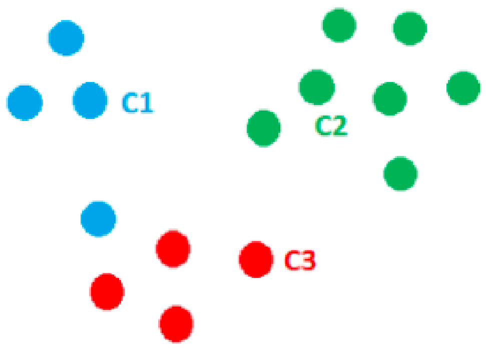
Among all the unsupervised learning algorithms, clustering via k-means might be one of the simplest and most widely used algorithms. Briefly speaking, k-means clustering aims to find the set of k clusters such that every data point is assigned to the closest center, and the sum of the distances of all such assignments is minimized. Below are steps for the same:

1: Initialize cluster centers

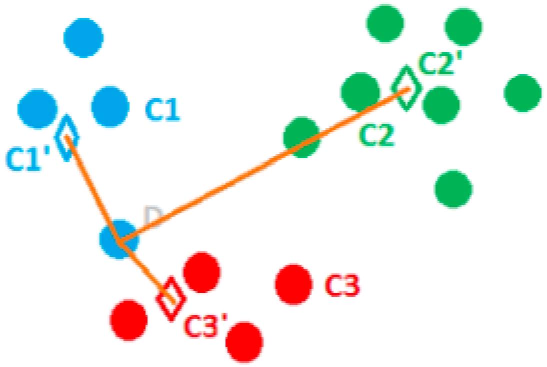


2: Assign observations to the closest cluster center.

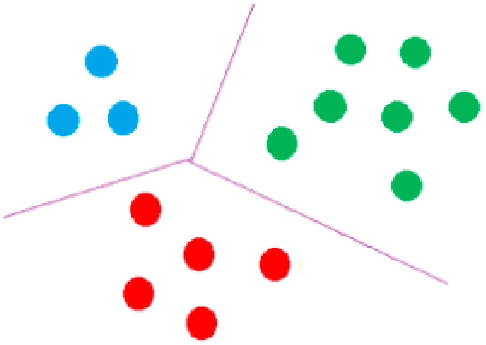




3: Revise cluster centers as mean of assigned observations



4: Repeat step 2 and step 3 until convergence



Answer 2 (C)

Below are two method deciding the K for K-means algorithm:-

- ✓ Silhouette Score
  - ➢ First, compute the score with respect to each K value and generally we start the K value from 2 to 10/15 cluster
  - ➢ Each K will have the average Silhouette score.
  - ➢ Now for visualization plot the score

    - ➢ Max. score will be chooser for clustering but sometime need to take business side, so we need to choose the score/cluster accordingly.

- ✓ Elbow Curve
    - ➢ First, compute the score with respect to each K value and generally we start the K value from 2 to 10/15 cluster
    - ➢ Each K value has been calculated SSD(sum of square distances)
    - ➢ Plot the SSD value with respect to number of K
    - ➢ Curve will look like elbow, so we need check the elbow and against that K number.
    - ➢ Choose the K number for clustering

Answer 2 (d)

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.

Cluster analysis is because groups are defined based on the distance between points in mathematical space. Standardization helps to make the relative weight of each variable equal by converting each variable to a unitless measure or relative distance.

The different attributes will have the measures in different units. Thus, standardisation helps in making the attributes unit-free and uniform.

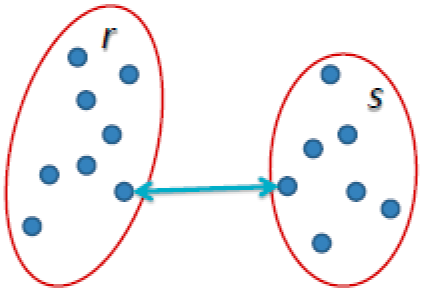
Answer 2 (e)

Hierarchical Clustering:

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. There are two types of hierarchical clustering.

Single Linkage*:*

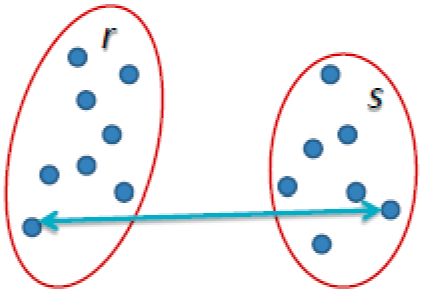
In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points.



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

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.



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

Question 3:

Principal Component Analysis

a) Give at least three applications of using PCA.

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

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

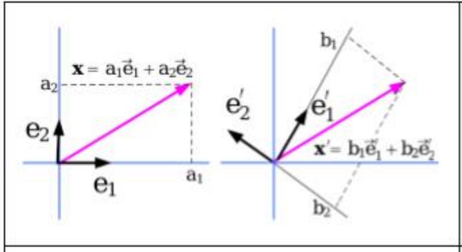
Answer 3 (a):

1. *Dimensionality Reduction*
2. *Data visualization: Digit Example*
3. *Image Compression*
4. *Noise reduction in the dataset*

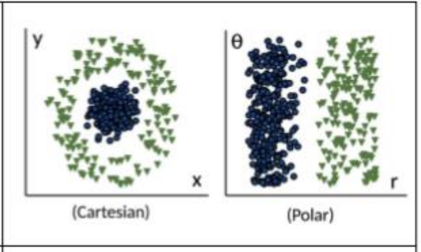
Answer 3 (b):

Basis transformation:

A basis for a vector space of dimension n is a set of n vectors, with the property every vector in the space can be expressed as a unique linear combination of the basis vectors. Since it is often desirable to work with more than one basis for a vector space, it is of fundamental importance to be able to easily transform coordinate-wise representations of vectors and operators taken with respect to one basis to their equivalent representations with respect to another basis. This process of converting the information from one set of basis to another is called basis transformation.



Transformation by rotation of basis

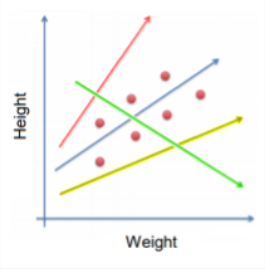


Transformation by change of basis

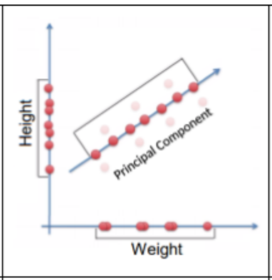
Variance as information:

Variance means - it measures how far a set of numbers are spread out from their average value.

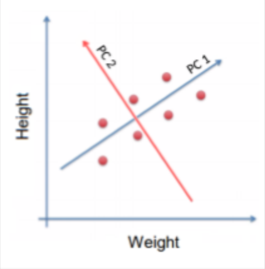
So, if we have higher variance means that we have more data or information. In PCA if we have more Variance as information. Its mean that we have good data for processing and like 4 principle components explaining the 95 percent but 15 components telling the 100% data.



Purely horizontal or vertical axis does not suffice the variation in both directions. Thus, an angled line is required.



The line closest to the data retains the maximum variation in the original data points. In fact, PCA finds lines/planes/surfaces closest to data points.



There is still some variance that is left in the perpendicular direction to the first PC. This forms the second PC.

Answer 3 (C):

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.*