Clustering and PCA 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:

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
- After the recent funding programs, 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.
- As a data analyst the job is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Then you need to suggest the countries which the CEO needs to focus on the most. The datasets containing those socio-economic factors

Solution Methodology:

1. Data Understanding

2. Performing PCA :

- Data Standardization
- Perform PCA and chose PCs that defines more than 85% variance
- Running PCA with the chosen number

3. Performing Clustering

Data Preparation for Clustering (Outliers and Hopkins Check)

K-Mean Clustering: Performing K-mean Clustering and choosing K with both elbow and silhouette curve, Run K-mean Clustering with chosen K, Visualize the Clusters and cluster Profiling

Hierarchical Clustering: Used both Single and Complete Linkage, chose one method based on the linkage, Visualized the Clusters

In the assignment I chose Principal components as 4 through Scree plot as 4 PCs explains variance in data more efficiently

Using Hopkins method decide whether the given data is good for Clustering and it showed 80% measure which is a good result

Using K-mean Clustering I decided to go with 4 clusters using Elbow and silhouette curve

Conducted Hierarchical Clustering and I find hierarchical clustering to be more efficient as it is both more flexible and has fewer hidden assumptions about the distribution of the underlying data.

In contrast, hierarchical clustering has fewer assumptions about the distribution of your data - the only requirement (which k-means also shares) is that a distance can be calculated each pair of data points

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:

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

K-means Clustering:

- The K-Means algorithm uses the concept of the centroid to create K clusters
- K- means is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. It is a division of objects into clusters such that each object is in exactly one cluster, not several.
- The k-means algorithm is parameterized by the value k, which is the number of clusters that you want to create. The algorithm begins by creating k centroids.
- It then iterates between an assign step (where each sample is assigned to its closest centroid) and an update step (where each centroid is updated to become the mean of all the samples that are assigned to it.
- This iteration continues until some stopping criteria is met; for example, if no sample is re-assigned to a different centroid Hierarchical Clustering:
- One of the major considerations in using the K-means algorithm is deciding the value of K beforehand. The hierarchical clustering algorithm does not have this restriction.
- In Hierarchical clustering, clusters have a tree like structure or a parent child relationship. Here, the two most similar clusters are combined and continue to Combine until all objects are in the same cluster.
- Agglomerative hierarchical clustering, instead, builds clusters incrementally, producing a dendrogram

 the algorithm begins by assigning each sample to its own cluster (top level).

- At each step, the two clusters that are the most similar are merged; the algorithm continues until all the clusters have been merged.
- Unlike k-means, you don't need to specify a k parameter: once the dendrogram has been produced, you can navigate the layers of the tree to see which number of clusters makes the most sense to your application.

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

K-Means algorithm is the process of dividing the N data points into K groups or clusters. Here the steps of the algorithm are:

- Start by choosing K random points the initial cluster centers.
- Assign each data point to their nearest cluster center. The most common way of measuring the distance between the points is the Euclidean distance.
- For each cluster, compute the new cluster center which will be the mean of all cluster members.
- Now re-assign all the data points to the different clusters by taking into account the new cluster centers.
- Keep iterating through the step 3 & 4 until there are no further changes possible. At this point, you arrive at the optimal clusters. Now having assigned each data point to a cluster, now we need to recompute the cluster centroids.

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

Value of K can be chosen by following either of the methods:

Choosing the number of clusters K in advance

There are several pointers that can help us decide the K for our K-means algorithm: -

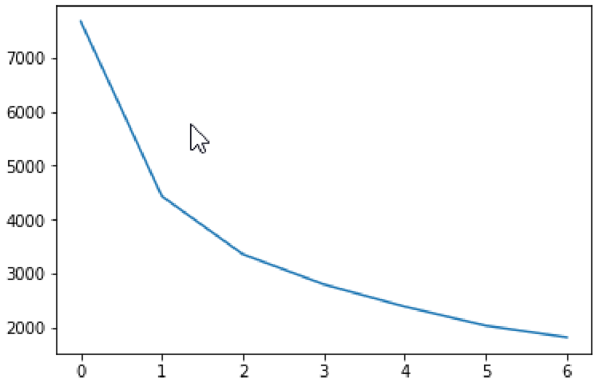
<u>1. Elbow method: -</u>

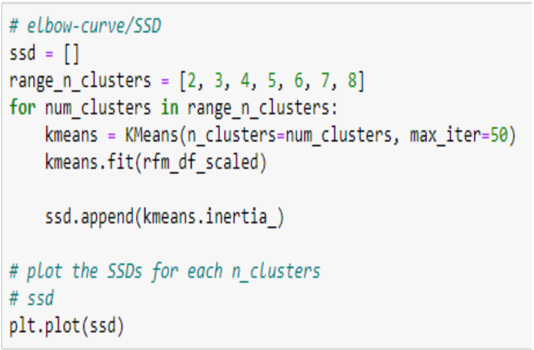
• Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.

• For each k, calculate the total within-cluster sum of square (wss).

• Plot the curve of wss according to the number of clusters k.

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





```
# elbow-curve/SSD
ssd = []
range_n_clusters = [2, 3, 4, 5, 6, 7, 8]
for num_clusters in range_n_clusters:
    kmeans = KMeans(n_clusters=num_clusters, max_iter=50)
    kmeans.fit(rfm_df_scaled)

    ssd.append(kmeans.inertia_)

# plot the SSDs for each n_clusters
# ssd
plt.plot(ssd)
```

<u>Silhouette Analysis:</u>

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

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

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

• The location of the maximum is considered as the appropriate number of clusters.

<u>Business Aspect of K-mean clustering:</u>

The K-means clustering algorithm is used to find groups which have not been explicitly labeled in the data. This can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.

Thus, expertise and domain knowledge are necessary when choosing important variables and clusters as per the business need

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

Standardization of data, that is, converting them into z-scores with mean 0 and standard deviation 1, is important for 2 reasons in K-Means algorithm:

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

Thus, Standardizing data is recommended because otherwise the range of values in each feature will act as a weight when determining how to cluster data, which is typically undesired. However, when the data is standardized this no longer becomes an issue and weights each feature as being equal when calculating the distance between each data point.

e) Explain the different linkages used in Hierarchical Clustering.

The different types of linkages.

- Single Linkage: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters
- Complete Linkage: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the clusters

- Average Linkage: Here, 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.

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:

a) Give at least three applications of using PCA.

<u>The predictive model setup:</u> Having a lot of correlated features lead to the multicollinearity problem. Iteratively removing features is time-consuming and also leads to some information loss.

<u>Data visualization: It</u> is not possible to visualize more than two variables at the same time using any 2-D plot. Therefore, finding relationships between the observations in a data set having several variables through visualization is quite difficult.

- Fundamentally, PCA is a dimensionality reduction technique, i.e., it approximates the original data set to a smaller one containing fewer dimensions
- For creating uncorrelated features that can be input to a prediction model: With a smaller number of uncorrelated features, the modelling process is faster and more stable as well.
- Finding latent themes in the data: If you have a data set containing the ratings given to different movies by Netflix users, PCA would be able to find latent themes like genre and, consequently, the ratings that users give to a particular genre.
- Noise reduction

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

<u>Basis transformation</u>

- Basis is essentially the fundamental units in which you express your data. it is similar to how we use units for physical objects to measure things like height, weight, temperature, etc.
- In vectors and vector spaces, we use basis vectors to represent the points in space. every observation in the space can be represented by scaling and adding the scaled basis vectors. This process is also called a linear combination.
- one of the key ideas that helped you connect basis vectors and the idea of dimensionality reduction: using different basis vectors to represent the same points.

Here's a list of rules Basis transformation:

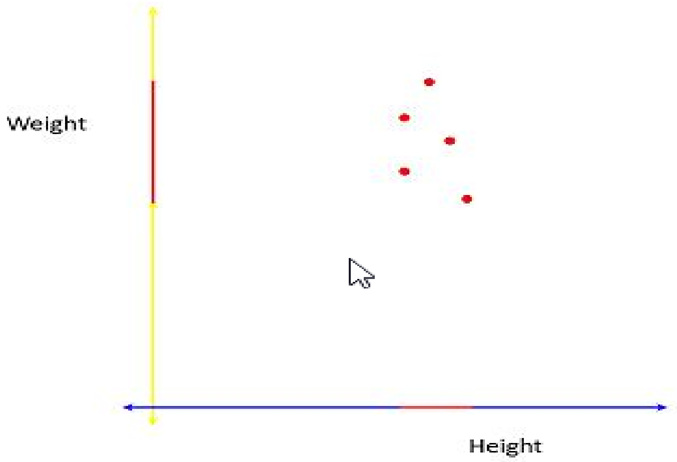
1. If you're moving from a basis space B to the standard basis, then the change of basis matrix M is the same as the basis vectors of B written as its column vectors. Therefore, if there is a vector v represented in B and you want to find its representation in the standard basis, then you'd have to perform Mv.

2. If you want to go the other way around, where you have v represented in the standard basis and want to find its representation in B you multiply it by its inverse - M−1v

3. Finally, if you want to find the change of basis matrix M where you move from two non-standard basis vectors - say from B1 to B2 then you can get that by calculating this value - B−12B1. Note that in all the above cases, the basis vectors should be represented in the same units.

<u>Variance as Information:</u>

- we didn't know as to how to find those "ideal basis vectors" and what exact properties they must satisfy and therefore the idea of variance as information enters
- PCA gauges the importance of a column by another metric called ‘variance’ or how varied a column’s values are.
- It is to measure the importance of a column by checking its variance values. If a column has more variance, then this column will contain more information.
- Look at the following image. Now, there is another elegant way of looking at variance geometrically. This give you an idea that Weight is a more important column than Height.



- The red line on the Height and Weight axes show the spread of the projections of the vectors on those axes. As you can see here, the spread of the line is quite good on the Weight axis as compared to the Height axis. Hence you can say that Weight has more variance than Height. This idea of the spread of the data being equivalent to the variance is quite an elegant way to distinguish the important directions from the non-important ones

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

shortcomings of using Principal Component Analysis:

- 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).
- 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.
- 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).