Arjun Singh Baghel

Roll Number - DDS1910090

Email ID: arjusingh89baghel@gmail.com

IIITB Email ID: arjunsinghbaghel.dds10@iiitb.net

Clustering & PCA Assignment Part-II

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)

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. 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.

And this is where you come in as a data analyst. Your job is to categorise 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.

Solution methodology:

I followed the below listed steps to complete the PCA & Clustering assignment
Step 1: Started with loading the country dataset in the python note book , read the data using inbuilt method
Step 2: Analyzed the data set by looking at the columns and rows
Step 3: Understanding of Data
Step 4: Exploring the data using EDA techniques
Step 5: Performed visualization of column data using univariate analysis methods
Step 6: Performed Outlier Analysis
Step 7: Performed Outlier Treatment, caped the values in the data set after observing their distributions across the mean with the help of plotting the box plot
Step 8: Visualized all the column variables after outlier treatment, still observed values are separated in gdpp column
Step 9. Scaled the data using scaling method available in python
Step 10: Performed PCA, found 4 Principal components
Step 11: Performed Clustering techniques:
a. Hopkins statistics: As the value is found to be approx. .731 deduced that the data has high tendency for clusters.
b. Used Silhotte method to calculate the K-value for K-means algorithm, which was found to be nearest to 2
c. Used Elbow method to further find the clusters using K-means, algorithm found 2 clusters having ID ,0 and 1.
d. Observed the clusters were not separated well, hence proceeded with Hierarchal clustering algorithm.
e. Using Hierarchal algorithm found 5 as the optimal values
f. Divided the PCA data set into the different clusters
Step 12: Analysis of Cluster: Calculated the mean for all the column, merged the dataset with the original based on the country names.
Step 13. Grouped the entire datasets based on the cluster id
Step 14: Visualized the data using bar plots
Step 15. Concluded the observation.
Observation: The countries grouped under cluster_ID=0 are the poorest countries.

Took that many numbers of principal components because to find the batter result and also that will be help for analysis.

Question 2: Clustering

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

Answer:

- 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).
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- 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 i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

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

Answer:

K-Means clustering is one of the frequently used clustering algorithms. It is non-hierarchical clustering method in which the number of clusters is decided a priori. The observations in the sample are assigned to one of the clusters. The following steps are used in K-means algorithm:

1. Choose K observations from the data set that are likely to be in different clusters. There are many ways of choosing these initial K values; easiest approach is to choose observations that are farthest. (in one of the parameters of the data)

2. The K observations chosen in step 1 are the centroids of those clusters.

3. For the remaining observations, find the cluster closest to centroid. Add the new observation closest to the centroid. Adjust the centroid after adding a new

observation to the cluster. The closest centroid is chosen based on appropriate distance measure.

4. Repeat step 3 till all observations are assigned to a cluster.

Note that centroids keep moving when new observations are added; also, observations may move to different clusters. An important aspect of K-means clustering is choosing the appropriate value of K. Initially the value of K is a guess; however, it can be decided, based on several measures such as CH(K) index, Silhoutte coefficient and elbow method.

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:

By visualization, we can say the number of clusters can be formed, but is limited only for

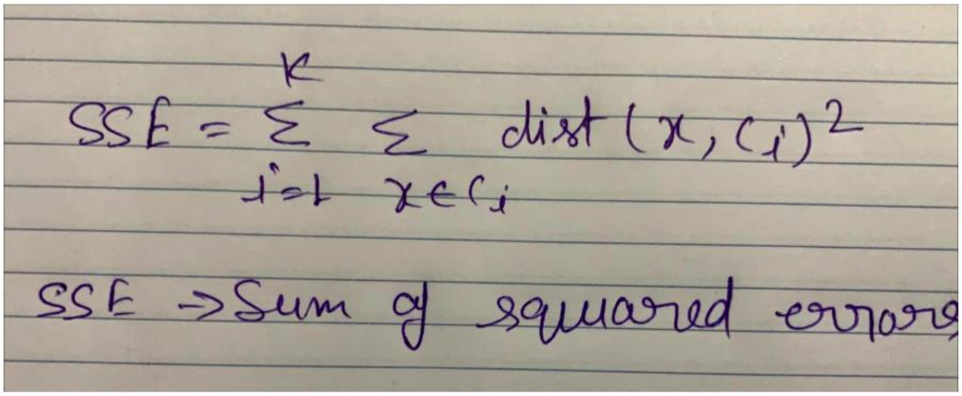
small data sets.

For large data sets, we need to perform clustering algorithm.

A hint to find number of clusters is Elbow method:

i) Sum of squared error(SSE) for values of k is calculated.

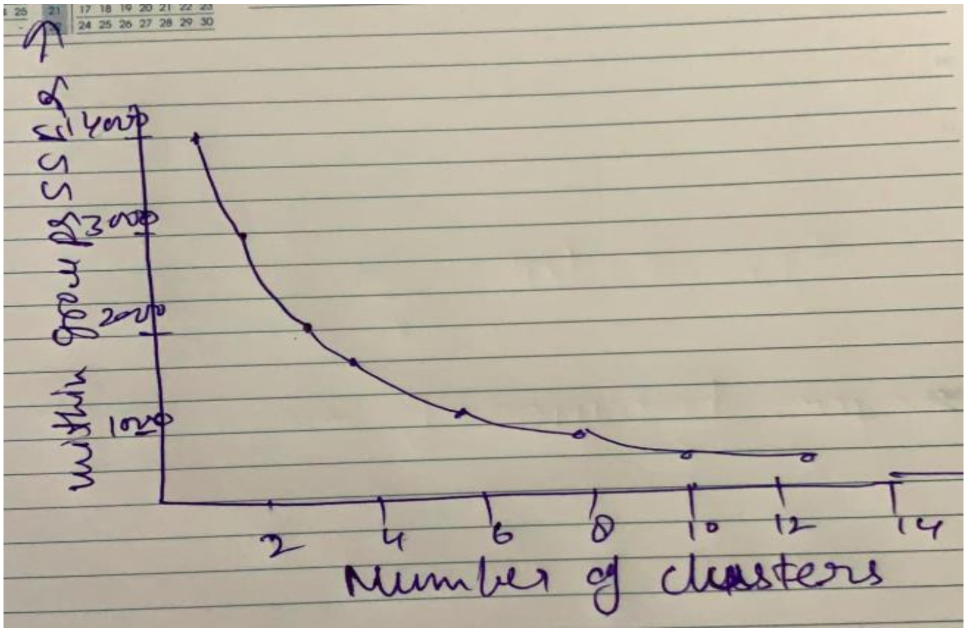
ii) SSE is the distance between each member of cluster and centroid.



$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(x, C_i)^2$$

SSE → Sum of squared errors

Plotting k against SSE, error decreases as k increases. It is because when the number of clusters increases, they are smaller, and distortion is also smaller. K is chosen at the point where SSE decreases abruptly, and this results in the elbow effect.



Clearly visible K = 6. At K = 6, SSE decreases abruptly.

Each centroid of a cluster is a collection of feature values which define the resulting groups.

Business Use in K- Means:

The data which are not explicitly labelled uses K-Means clustering. This assure business in finding

out, what type of group exists or to find out unknown groups in complex sets.

Some examples of used cases are:

Behavioral Segmentation:

-Purchase History by segmentation

-Activities on application, website, platform by segmentation.

-Personas based on interest.

-Profile based on activity monitoring.

Inventory Categorization:

-Types of activity in motor sensors

-Groups in health monitoring

-Group images

-Audio

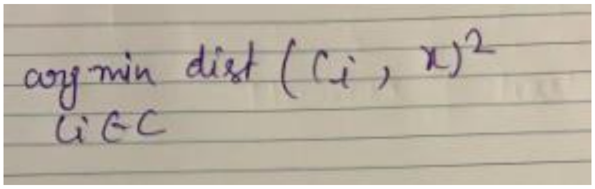
Detecting Anomalies:

-Separate activity group from bots.

-Outlier detection activity.

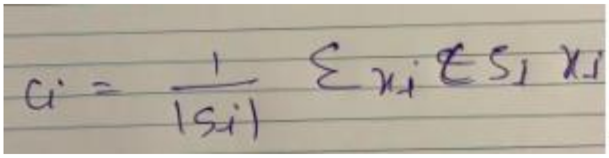
Statistical method of K- means Algorithm:

Data Assignment Step: Each centroid defines it presence in one of the clusters. Here, each data point is assigned to its nearest centroid based on Euclidean distance. In each data set C, if is the collection of centroids, then each data point x is assigned to a cluster based on



$$\underset{c_i \in C}{\arg\min}\ dist(c_i, x)^2$$

Where dist.(-) is the standard Euclidean distance. Let be the set of data points for each cluster. Centroid Update step: The centroids are recalculated, and this can be performed by taking the mean of all the data points assigned to the centroid’s cluster.



$$c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$$

Until a stopping criterion is met, step 1 and 2 iterates continuously. The result may be a local optimum which means assessing algorithm more than once with random starting centroids can give a better result.

Elbow method: k-means clustering, is used to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized. The total WSS measures the compactness of the clustering and we want it to be as small as possible. The Elbow method looks at the total WSS as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn’t improve much better the total WSS. The optimal number of clusters can be defined as follow:

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 (wss).
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.

Silhouette method: Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman and Rousseeuw 1990). The algorithm is similar to the elbow method and can be computed as follow:

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.

Gap statistics: The gap statistic compares the total within intra-cluster variation for different values of k with their expected values under null reference distribution of the data. The estimate of the optimal clusters will be value that maximize the gap statistic (i.e, that yields the largest gap statistic). This means that the clustering structure is far away from the random uniform distribution of points. The algorithm works as follow:

1. Cluster the observed data, varying the number of clusters from k = 1, …, *kmax*, and compute the corresponding total within intra-cluster variation *Wk*.
2. Generate B reference data sets with a random uniform distribution. Cluster each of these reference data sets with varying number of clusters k = 1, …, *kmax*, and compute the corresponding total within intra-cluster variation *Wkb*.
3. Compute the estimated gap statistic as the deviation of the observed *Wk* value from its expected value *Wkb* under the null hypothesis: Gap(k)=1B∑b=1Blog(W∗kb)−log(Wk)*Gap(k)=1B∑b=1Blog(Wkb∗)−log(Wk)*. Compute also the standard deviation of the statistics.
4. Choose the number of clusters as the smallest value of k such that the gap statistic is within one standard deviation of the gap at k+1: *Gap*(*k*)≥*Gap*(*k* + 1)−*sk* + 1.

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

Answer:

Scaling/Standardization comes into picture when features of input data set have large differences between their ranges, or simply when they are measured in different measurement units (e.g., Pounds, Meters, Miles … etc).

These differences in the ranges of initial features causes trouble to many machine learning models. For example, for the models that are based on distance computation, if one of the features has a broad range of values, the distance will be governed by this particular feature, and can create problem. So transforming the data to comparable scales can prevent this problem.

Standardization/scaling is referred as the process of rescaling the values of the variables in your data set so they share a common scale. Data clustering in one of the important techniques used in many applications in data mining, K-means is one of the most well-known methods of datamining that partitions a dataset into groups of patterns. Standardization is the central preprocessing step in data mining, to standardize values of features or attributes from different dynamic range into a specific range.

For example, in boundary detection, a variable that ranges between 0 and 100 will outweigh a variable that ranges between 0 and 1. Using variables without standardization can give variables with larger ranges greater importance in the analysis. Transforming the data to comparable scales can prevent this problem.

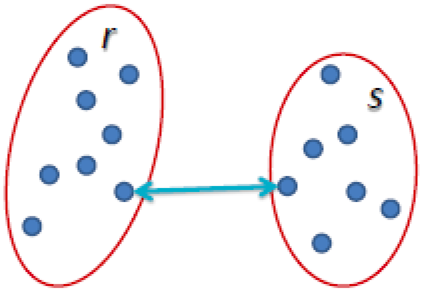
e) Explain the different linkages used in Hierarchical Clustering.

Answer:

Below are the different types of linkages used in hierarchical clustering-

Single Linkage – Commonly known as nearest neighbour clustering and is one of the oldest and famous technique to use. The distance between two groups are defined as distance between two closest data members. It results clusters where individuals are added sequentially to a single group.

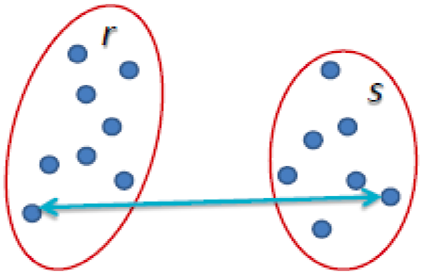
Let “r” and “s” on the left equals to the distance between the r and s point.



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

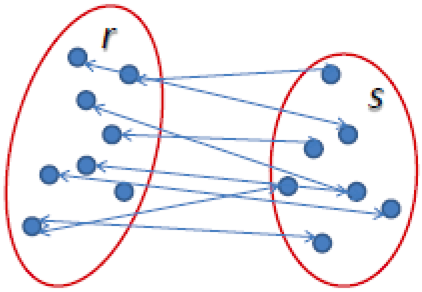
Complete linkage – It is also known as maximum method or furthest neighbour. The method defines as the farthest distance between two groups as distance between two farthest/ distant members. This method usually yields with the cluster which are quite apart and compact. One drawback is that outliers can cause merging of close groups later than it is optimal.

Let “r” and “s” on the left equals to the distance between the r and s the furthest point.



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

Average Linkage - In average linkage hierarchical clustering, the distance between two clusters is defined as the average 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.



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

Question 3: Principal Component Analysis

a) Give at least three applications of using PCA.

Answer:

Below are the applications of using PCA

1-Image compression- Principal component analysis (PCA) is one of the most popular and elaborate algorithms for image compression. Now a days Image compression broadly use in the industry to improve image quality. The image is represented as a small-dimensional vector (main component), which is then compared with benchmark vectors from the database. PCA allows us to represent high-dimensional data in a lower dimensional form. This means that we can use PCA to compress an image from the naive high-dimensional representation (RGB values for each pixel) to any size we choose.

2-Computer network attack - Network traffic data collected for intrusion analysis is typically high dimensional making it difficult to both analyze and visualize. Principal Component Analysis is used to reduce the dimensionality of the feature vectors extracted from the data to enable simpler analysis and visualization of the.

3-Bioinformatics- Bio scientist has to analyze the huge data sets of genes expression without losing the important features of it. PCA constructs linear combinations of gene expressions, called principal components. The principal components are orthogonal to each other, can effectively explain the variation of gene expressions, and may have a much lower dimensionality. PCA is computationally simple and can be realized using many existing software packages.

4- PCA can be used in neuroscience where it can identify that the specific properties of a stimulus that increases neuron's probability of generating any action.

5- In quantitative finance, PCA can be directly applied to risk credit or interest portfolios.

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

Answer:

Basis transformation: The process of change is basis is defined a Basis transformation. PCA takes the standard value points from the dataset and transform its points expressed in eigenvector basis. The hope is that this new basis will filter out the noise and reveal hidden dynamics. In other words, we can say that it is the process of identifying the principal components, a set of values of linearly uncorrelated variables.

Variance as Information: Variance (σ2) is a measurement of the spread between numbers in a data set. It measures how far each number in the set is from the mean and is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set. Using the concept of variance in terms of information related to data, is termed as variance as information where variance can clearly state that the information available in two different variables can be determined using only one variable.

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

Answer:

Shortcomings of using PCA are listed below:

1- PCA depends on the scaling of variables and hence can be fixed by scaling each individual feature by its standard deviation, so that it can end up to dimensionless features.

2- When linear correlation assumptions fail, PCA can no more capture it. Whereas in some cases coordinate transformation can restore linearity and PCA can be applied which is a rare case scenario.

3- Another restriction is mean removal process while constructing covariance matrix for PCA.

4- PCA uses variance as the measure of importance of variable. High variance variable are treated as principal components and low variance are treated as noise.

5- PCA assumes that principal components are orthogonal in nature.