Question 1 – Clustering

Q. Assignment Summary (Unsupervised Learning – Clustering)

Answer: - We used following technical approaches to get the list of countries which has really required/needed financial aid by NGO. Steps are below.

- Check for missing value, and treatment.

There was no missing values in the dataset so there was no needed to impute with any values. There are 167 rows and 10 columns in dataframe.

- Check for outlier and treatment

There was outliers in the data in which I handled with upper and lower values (.05 and .95 – standard method). Dataset has no duplicate as well.

- Perform the basic EDA to find the variability and distribution of the data, so as to identify if we need scaling the data.

Most of the data point are 'NOT Normally' distributed. Their variance are also differernt. Their range are also differnt All the above points indicates the need of standardising the data before we build the model. 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 is important here.

- Data Scaling

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

- Use Hopkins Method to check if the dataset is good enough for a cluster analysis

Before we apply any clustering algorithm to the given data, it's important to check whether the given data has some meaningful clusters or not? which in general means the given data is not random. The process to evaluate the data to check if the data is feasible for clustering or not is know as the clustering tendency. To check cluster tendency, we use Hopkins test. Hopkins test examines whether data points differ significantly from uniformly distributed data in the multidimensional space.

- Using Hierarchical clustering to identify the optimal cluster value.

As mentioned in the 'Approach' section, we will use Hierarchical Clustering to identify appropriate cluster size with a good split of data (Max Intra-Cluster distance & Min Inter-Cluster Distance).

- Use Silhouette and Elbow method to validate the optimal cluster values.

p is the mean distance to the points in the nearest cluster that the data point is not a part of

q is the mean intra-cluster distance to all the points in its own cluster.

The value of the silhouette score range lies between -1 to 1. A score closer to 1 indicates that the data point is very similar to other data points in the cluster, A score closer to -1 indicates that the data point is not similar to the data points in its cluster.

.

- Use K-Means Cluster method to build the final cluster model.

From the above 3 Iterations of K-Means, we could see that using 3 Clusters provided a better output in terms of a balanced cluster size. So we will consider the 'K-Means with 3 Clusters' as our FINAL MODEL.

- Present the final report.

Recommended the name of top 5 countries in which they required financial aid by NGO.

Question 2 – Clustering

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

Answer: - Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Hierarchical clustering can't handle big data well but K Means clustering can. While results are reproducible in Hierarchical clustering. K Means is found to work well when the shape of the clusters is hyper spherical

In hierarchical k-means we pick some k to be the branching factor. This defines the number of clusters Figure 2: An example of k = 3 means hierarchical clustering. First sort the points into clusters and then recursively cluster each clustered set of points. at each level of the clustering hierarchy.

If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. 2) K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular. K-Means Disadvantages : 1) Difficult to predict K-Value

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

Answer: - Algorithmic steps for k-means clustering are follows.

1) Randomly select *‘c’* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:
5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

Advantages

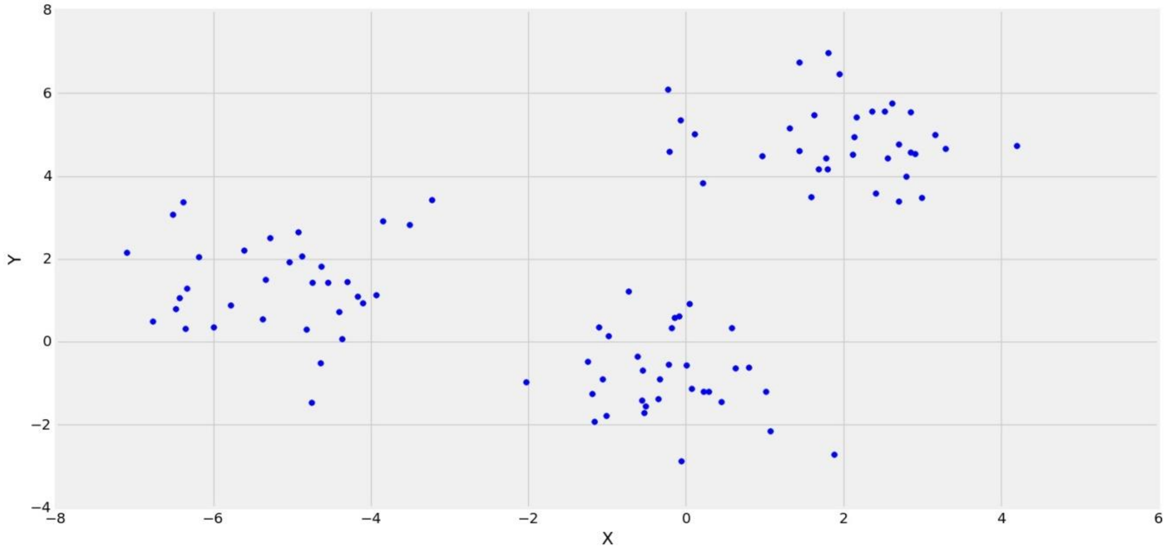
1) Fast, robust and easier to understand.

2) Relatively efficient: O(tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n.

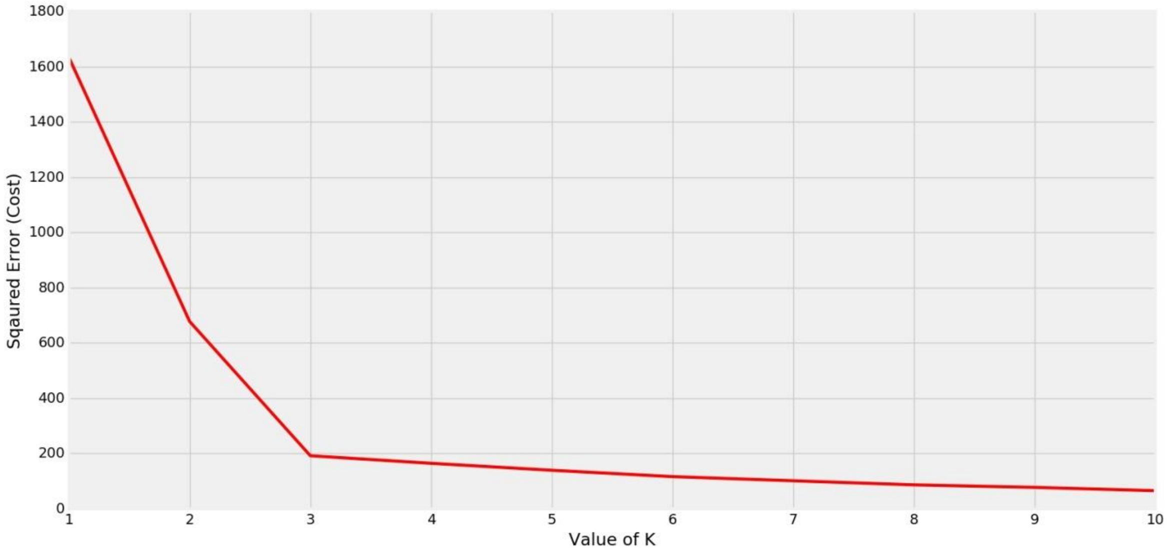
3) Gives best result when data set are distinct or well separated from each other.

Q. (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: - There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing *k*. As the value of *K* increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid. So, the point where this distortion declines the most is the elbow point.



In the above figure, its clearly observed that the distribution of points are forming 3 clusters. Now, let’s see the plot for the squared error(Cost) for different values of K.



Clearly the elbow is forming at K=3. So the optimal value will be 3 for performing K-Means.

K Means Numerical Example. The basic step of k-means clustering is simple. In the beginning we determine number of cluster K and we assume the centroid or center of these clusters. We can take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids.

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

Answer: - When we standardize the data prior to performing cluster analysis, the clusters change. We find that with more equal scales, the Percent Native American variable more significantly contributes to defining the clusters. Standardization prevents variables with larger scales from dominating how clusters are defined.

in average it should improve the results, and it probably will not worsen them. The reason is that normalization gives the same importance to all the variables. The standard example is considering age (in year) and height (in cm).

K-means needs to compute means, and the mean value is not meaningful on this kind of data. As explained in this paper, the k-means minimizes the error function using the Newton algorithm, i.e. a gradient-based optimization algorithm. Normalizing the data improves convergence of such algorithms.

When we standardize the data prior to performing cluster analysis, the clusters change. We find that with more equal scales, the Percent Native American variable more significantly contributes to defining the clusters. Standardization prevents variables with larger scales from dominating how clusters are defined. There is a popular method known as elbow method which is used to determine the optimal value of K to perform the K-Means Clustering Algorithm. The basic idea behind this method is that it plots the various values of cost with changing k. As the value of K increases, there will be fewer elements in the cluster.

Q. (e) Explain the different linkages used in Hierarchical Clustering.

Answer: - 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, *Divisive* and *Agglomerative*.

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.

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.