By S.Anil

Assignment-based Subjective Questions Dataset – Country Data Model – Clustering – Unsupervised Machine Learning.

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 (what EDA you performed, which type of Clustering produced a better result and so on)

Solution:

The steps ascertained for the Assignment are hereunder:

Problem statement *– To identify atleast 5 countries which are in direst need of aid from the analysis work:*

1. EDA – Visualizing the data
   - A) Using distplot - Identifying the distribution of data in each numerical columns
   - B) Boxplots to view the outliers if any.
   - C) Pair plots and heatmap to view the relation between individual features with each other.

2. Outlier Treatment-
   - The clustering process is very sensitive to the presence of outliers in the data.
   - In this case, specifically the outliers need not to be dropped at all. As all the countries are to be considered for evaluation. The extremities should be considered, if any, and may form a characteristic of cluster. Thus soft capping was implemented

3. Scaling - the distance metric used in the clustering process is the Euclidean distance therefore, standardized scaling was attempted as it is important for clustering and all the attributes of the countries data be on the same scale.

4. Checking the tendency of the data: Hopkins Test

5. Checking the best value for K: ssd, silhouette method
   - Optimum value considered by SSD and silhouette method is 3.
   - Logically as well the consideration is 3 as countries considered based on their health and econometrics in three categories.

6. Perform KMeans with the final value of k.

7. Visualize the clusters using scatter plot – On visualizing the cluster (Cluster Ids the dataset is divided into 3 clusters i.e. Developed, Under- Developing and Un-Developed Countries.)
8. Perform Cluster Profiling: GDPP, CHILD_MORT, INCOME.

9. Hierarchical Clustering - cluster analysis which seeks to build a hierarchy of clusters without having fixed number of cluster.

10. Single Linkage: Dendogram: single linkage doesn't produce a good enough result.
11. Complete Linkage: Dendogram (FARTHEST NEIGHBOR CLUSTERING) shows :

At y-axis point 6 , we can see that 3 clusters are being formed clearly and considering the problem statement 3 seems an ideal choice for cluster numbers.

12. Visualize the cluster on the basis of performing Cluster Profiling: **GDPP, CHILD_MORT, INCOME**

13. Using both the results, reporting the countries that are in need of the AID: The top 10 Countries which require the aid (Based on the above Cluster Profiling) are as follows on the basis of Hierarchal Clustering(Reasons for choosing Hierarchal Clustering):

- Eliminating the k Means limitation of predefined consideration of number of clusters.
- Data set is small.
- Slightly more data points considered in Un-developed Countries Segment (Cluster-labels = 0
)

- Central African Republic
- Sierra Leone
- Haiti
- Chad
- Mali
- Nigeria
- Niger
- Angola
- Congo, Dem. Rep.
- Burkina Faso

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.

---

a) Solution :

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, instead, builds clusters incrementally, producing a dendogram. As the picture below shows, 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 of the clusters have been merged. Unlike k-means, you don't need to specify a k parameter: once the dendogram has been produced, you can navigate the layers of the tree to see which number of clusters makes the most sense to your particular application.

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:

where, 'ci' represents the number of data points in ith cluster.

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

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).

c) Solution:

The basic idea behind k-means consists of defining k clusters such that total **within-cluster variation (or error) is minimum**.

1. **The Elbow Method:** Calculate the **Within-Cluster-Sum of Squared** Errors (WSS) for **different values of k**, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an **elbow.**
   - The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.
   - The WSS score is the sum of these Squared Errors for all the points.
   - Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.
2. **The Silhouette Method**
   - The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.

The Silhouette Value s(i) for each data point i is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

Source: Wikipedia

Note: s(i) is defined to be equal to zero if i is the only point in the cluster. This is to prevent the number of clusters from increasing significantly with many single-point clusters.

Here, a(i) is the measure of similarity of the point i to its own cluster. It is measured as the average distance of i from other points in the cluster.

For each data point $i \in C_i$ (data point i in the cluster C_i), let

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

Source: Wikipedia

Similarly, b(i) is the measure of dissimilarity of i from points in other clusters.

For each data point $i \in C_i$, we now define

$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j)$$

Source: Wikipedia

d(i, j) is the distance between points i and j. Generally, Euclidean Distance is used as the distance metric.

3. Methodology of identifying K, on the basis of business acumen :

The clustering can be considered on the basis of business call or expertise, despite of the validations and scores it would be the call for the expert or the purpose of making this decision. Not ideally higher scores would be leading to desired clusters; it would help to have a better statistical approach with high scores however it would still lead to complex details.

d) Solution:

Standardization is thecentral preprocessing step in data mining, to standardize values of features or attributes from different dynamic range into a specific range. Otherwise the range of values in each feature will act as a weight when determining how to cluster data, which is typically undesired.

e) Solution :

1. - Single linkage:

Single linkage returns minimum distance between two point , where each points belong to two different clusters.

This measure defines the distance between two clusters as the minimum distance found between one case from the first cluster and one case

2. Complete linkage:

It returns the maximum distance between each data point.

This measure is similar to the single linkage measure described above, but instead of searching for the minimum distance between pairs of cases, it considers the furthest distance between pairs of cases

3. Average linkage:

It returns the average of distances between all pairs of data point .

This method is supposed to represent a natural compromise between the linkage measures to provide a more accurate evaluation of the distance between clusters. For average linkage, the distances between each case in the first cluster and every case in the second cluster are calculated and then averaged