CLUSTERING SUBJECTIVE QUESTIONS ASSIGNMENT

By- SATADHRITI CHAKRABARTY, DS C17 FEB 2020, GROUP 2

QUESTION 1: ASSIGNMENT SUMMARY

-HELP International is an NGO that has collected a fund of around $10 million. They want to use this fund to financially help countries with the worst socio-economic conditions. A list of 5-10 countries would help them prioritize the countries which are at the bottom of the socio -economic ladder.

Our job is to figure out this list of countries. Since we do not have any well-defined target variable here, we use clustering to segment the countries in various clusters and deriving which countries would be the among the first to receive financial aid.

For this purpose, we first capped the outliers of all the variables by looking at their boxplots and scaled the variables. We checked the Hopkins statistic to check whether the data is good enough for clustering. We then investigated the K-means algorithm first and from the combination of the Elbow curve and the Silhouette scores, concluded that the ideal number of clusters should be three. We then assigned each country to a cluster by considering the three primary drivers for evaluating a country's socio-economic condition i.e. child mortality, income and GDP per capita. We used pair plots amongst these three variables and a bar plot to identify the cluster with the worst performance on these three metrics. We finally got a list of 10 countries belonging to this worst performing cluster.

We used a similar approach in our investigation using the Hierarchical clustering and chose the number of clusters as 3 (though 4 would also have been a good choice, the results using 4 clusters was not as interpretable as using 3 clusters). We used similar visuals to get the list of 10 countries belonging to the worst performing cluster.

Incidentally, the results from the both the algorithms provided exactly same results, confirming our choice of clusters being the best possible with the data at hand.

QUESTION 2 a: Compare and contrast K-means Clustering and Hierarchical Clustering

- In K-means Clustering, the initial number of clusters need to be mentioned while in Hierarchical Clustering, we do not have such a pre-determined condition.

- In K-means, the clusters can change according to the initial choice of the centroids while in Hierarchical Clustering, there is no such inconvenience.

- In K-means, we have a cost function to be minimized given by:
  $$J = \sum_{i=1}^{n} ||X_i - \mu_{k(i)}||^2 = \sum_{k=1}^{K} \sum_{i\epsilon C_k} ||X_i - \mu_k||^2$$
  while in Hierarchical Clustering, there is no objective function to be minimized.

- Hierarchical Clustering is more sensitive to noise than K-means.

- K-means works well when the shape of the clusters is circular while Hierarchical clustering works well when the clusters are non-spherical.

- In general, K-means is good for large datasets and Hierarchical is good for smaller datasets as the Hierarchical algorithm must remember assignment of each cluster right from the beginning when we have ‘n’ clusters for ‘n’ data points.

- K-means is more time-efficient than Hierarchical clustering.

QUESTION 2 b: Briefly explain the steps of the K-means clustering algorithm

-The steps used in K-means algorithm are:

1) We choose the number of clusters initially based on our business understanding.

2) We select 'k' cluster center points for the 'k' clusters we form.

3) Assign each data point to its nearest cluster based on the Euclidean distance.

4) For each cluster, the centroid will be the mean of the distances of all the data points from the cluster center based on the Euclidean distance.

5) As we update the cluster center, reassign the points to the different clusters by considering the new cluster centers.

6) Keep iterating through steps 4 and 5 until there are no further changes.

QUESTION 2 c: How is the value of ‘k’ chosen in K-means clustering? Explain both the statistical as well as the business aspect of it.

- The value of ‘k’ in terms of statistics can be determined using the **Elbow curve** and the **Silhouette score**. The combination of both these methods can be the best approach to determine the number of clusters.

To understand the Elbow method, we first define the **sum of squared error (SSE)** which is the sum of the squared distance between each member of the cluster and its centroid. Mathematically,

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

where K is the number of clusters.

If we plot K against SSE, the error decreases as K gets larger. We choose the K at which the SSE falls abruptly.

The Elbow curve may not give the best results alone, hence we **combine it with the Silhouette score** to determine the optimal number of clusters.

Silhouette coefficient is a measure of how similar a data point is to its own cluster (cohesion) compared to other clusters (separation). Silhouette score, S(i) is given by:

S(i) = b(i) – a(i)/max{b(i), a(i)} where a(i) is the average distance from its own cluster and b(i) is the average distance from its nearest neighbor cluster.

The higher the value of the Silhouette coefficient across the values of ‘k’, better the result.

But we cannot always select the ‘k’ with the highest Silhouette score.

Along with all these, we need to **understand the business** as well. **Too many clusters may be difficult to interpret for the business**. We can try out our clustering with two/three values of ‘k’ and see which provides the best interpretability and ease as well as satisfy the business goal.

Combining all these methods, we arrive at an optimal number for ‘k’.

QUESTION 2 d: Explain the necessity for scaling/standardization before performing Clustering.

-Scaling of the continuous variables is extremely important before clustering. If we take the example of RFM analysis in case of an online/offline retail store, the monetary value will be naturally very high. The recency, which is in terms of number of days, is generally low. The frequency can even be lower. When we compute the distance between two data points, we can imagine each data point as a tuple of recency, frequency and monetary. The monetary, normally expressed in thousands, will result in the **sum of squared distance** between two data points with different monetary values to be very high and the **variable power of ‘monetary’ will overshadow the ‘frequency’ and ‘monetary’ variables** and the clusters will represent the variable ‘monetary’ and not the other two. Hence, standardization/scaling **is the most important data preprocessing** part in a K-means algorithm.

QUESTION 2 e: Explain the different linkages used in Hierarchical Clustering.

-The different linkages used in hierarchical clustering are:

1) **Single Linkage**: Here, the distance between 2 clusters is defined as the shortest distance between points in the two clusters.

2) **Complete Linkage**: Here, the distance between 2 clusters is defined as the maximum distance between any 2 points in the cluster.

3) **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.