CLUSTERING ASSIGNMENT

Assignment based questions:

1. Assignment Summary

A NGO named HELP International worked to help the people of backward countries to fight poverty by providing them basic amenities and relief especially during extreme circumstances such as disasters and natural calamities. The CEO of this NGO was able to raise around $ 10 million via funding programmes. He has to decide how to spend this money strategically by considering those countries who are in the direst need of funds. Accordingly, the objective is to categorise the countries that are in the need of aid based on some social, economic and health factors that determine the overall development of the country. It clearly reflects the case of unsupervised machine learning. Therefore, in this analysis, we have used K Means and Hierarchical clustering to arrive at those countries that need help dealing with their economic and social problems. The frequency distribution plot shows that variables such as GDP per capita, income and chile mortality rate have same distribution and from policy perspective these variables plays an important role in determining the economic condition of the country. Therefore, countries were clustered based on these three variables. After performing the Hopkins test , we have scaled the variables to make the attributes unit-free and uniform. Based on elbow curve and Silhouette Analysis, we have initiated the K means analysis with 3 as k value. In case of Hierarchical clustering, looking at the natural groupings that is defined by long stems, we have obtained 3 clusters by cutting the dendrogram at the appropriate level. The clusters formed in both the cases were found to be identical. However, we have proceeded with the hierarchical clustering since there is no need to pre-determine the value of k as in K -means clustering. Thus, based on these 3 factors, we have arrived at the same set of countries which are in dire need of aid.

Clustering questions

1. Compare and contrast K-means Clustering and Hierarchical Clustering.

1. In K-means clustering, we have to initiate the analysis by specifying the desired number of clusters K and value of k (centroid). Thereafter the K-means algorithm assigns each observation to exactly one of the K clusters. However, the choice of initial cluster is important since it could impact the optimal clustering. Therefore, one has to be very cautious while selecting the initial K clusters. On the other hand, in hierarchical clustering, instead of pre-defining the number of clusters, we have to visualise the similarity or dissimilarity between the different data points and based on these similarities and dissimilarities, we can decide the number of cluster. The output of the hierarchical clustering algorithm resembles an inverted tree-shaped structure, called the dendrogram.

2. Briefly explain the steps of the K-means clustering algorithm

2. The following are the main steps of the K means algorithm:

(i) Initialisation: In this algorithm, we have to randomly picked up the K clusters and then initiate the k values based on elbow curve and Silhouette analysis.

(ii) Assignment: Each observation n is then assigned to the cluster whose cluster center i.e centroid is the closest to it. The squared Euclidean distance is used to find the centroid.

The equation of the assignment step is as follows:

$$C_k = Argmin \left\{ \sum_{k=1}^{K} \sum_{i=1}^{N} \left(x_i - \mu_k \right)^2 \right\}$$

Where : N is the number of observations, K is the total number of desired clusters and Ck denotes the set containing the observations in k th clusters.

(iii) Optimisation : For each of the K clusters, the centroid is computed such that the th cluster center is the vector of the p feature means for the observations in the k cluster.

The equation for the optimisation step is as follows.

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

(iv). Iteration: The process of assignment and optimisation is repeated until there is no change in the clusters or until the algorithm converges.

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

3. The intiation of analysis by specifying the number of clusters and choosing the value of k is of utmost importance in the algorithm since the having 2 clusters with lower values of K can cause the clusters not to be cohesive enough and higher values of k can lead to clusters not dissimilar enough.

The value of k is determined by the following two statistical measures:

a. Silhouette analysis: The Silhouette Analysis is a measure of how similar a data point is to its own cluster (cohesion) compared to other clusters (separation).

b. Elbow Curve Method: In this method the Sum of Squared Errors (SSE) is calculated for each value of K. Here, we always tried to minimize SSE, but as the SSE tends to decrease towards 0 with the increase in K, the number of clusters will increase and as the SSE becomes 0 then each data point will become its own cluster. Thus, the goal is to choose a small value of K which still has a low SSE, and the elbow usually represents the point from where the returns start diminishing with increasing values of K.

Besides the statistical approach, it is important to keep the business aspect in the mind. For example, let say NGO want to aid low developed countries based on GDP , income and child mortality rate. In such case, it can only support some countries with the limited funds. In such case, having more than 5 clusters will make no sense. Therefore, it is important to consider the business or policy objective before deciding the value of k.

4.Explain the necessity for scaling/standardisation before performing Clustering.

4. S**caling** is a method that is used to normalize the variables. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.

Scaling is required when the features in our dataset vary highly in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading.

There are following methods of scaling:

1. Standardization: In standardization, the values are replaced by z score. This redistributes the features with their mean $\mu = 0$ and standard deviation $\sigma = 1$

2. Normalization (Min-Max scaling) : This scaling brings the value between 0 and 1.The following formula is used:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

In clustering , since the distance metric used in the clustering process is the squared 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. Scaling helps in making the attributes unit-free and uniform.

5. Explain the different linkages used in Hierarchical Clustering.

5. In a hierarchical clustering algorithm, before any clustering is performed, it is required to determine the proximity matrix (displaying the distance between each cluster) using a distance function. This proximity or linkage between any two clusters can be measured by three distance functions

(i) Single: The distance between two clusters is defined as the shortest distance between two points in each cluster. However, it suffers from chaining issue. In such cases, clusters are not structured appropriately.

(ii) Complete: The distance between two clusters is defined as the longest distance between two points in each cluster. However, it suffers from crowding. Its score is based on the worst case dissimilarity between pairs, making a point to be closer to points in other clusters than to points in its own cluster. Thus, clusters are compact, but not far enough apart.

(iii) Average: The distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. In this case, clusters tend to be relatively compact as well as relatively far apart.

In general, complete or average linkages produce dendrograms which have a proper tree-like structure.