HELP International – NGO Fighting Poverty

Subjective Question Answers
Author: Jithin Prakash K
DSC-June 2020

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

Note: You don't have to include any images, equations or graphs for this question. Just text should be enough.

Answer:

Problem Statement:

HELP International – NGO, has an objective to eliminate poverty by providing the people of backward countries with basic amenities and relief during the time of disasters and natural calamities.

A sum of $10 Million fund is raised recently. This amount needs to be strategically and effectively used across the globe. An analysis need performed to find the countries that are in direst need of money.

Solution Methodology:

Reading the data – First the data is loaded on to the data frame, and is checked for null values, types of datatype and the content of the data. It is observed from data dictionary that exports, health and imports are the percentage value of GDPP, so they are converted back to values using formula

$$\boldsymbol{Feature\_value = Feature\_Value * GDPP\_value/100}$$

Data Visualization and preparation – Most of the features in the data set are found to be numeric data, a box plot is plotted for all the features to analyze the outliers, there were outliers for all the features and hence the capping is done so as to cap the values for developed countries. Capping at 99th percentile, is carried out for 'exports', 'health', 'imports', 'income', GDPP' columns. Bar plots are plotted for developed and under developed countries by filtering and sorting the data feature wise. A heatmap is plotted to obtain the correlation and a pair plot is used to obtain the spread and to have overall view of all the features as scatter and distribution plots.

Data Preparation – The data is to be standardized as the values for each feature is in different scale. It is important to bring all the features to similar scale as the K-means algorithm is highly affected by the distance between the clusters. A standard scaler is used from sklearn's preprocessing module to standardize the data as per mean and standard deviation.

Data Check for cluster integrity – Hopkins analysis is used to find the Hopkin's score and to make sure the data can be properly clustered, a value around 0.90 was obtained, and hence the data can be clustered.

There are 2 types of clustering techniques used

1. K Means clustering

   K means clustering requires, number of clusters (k) as one of the inputs / parameters for the algorithm. Value for K is found using 2 methods, elbow curve based on sum of square distance and silhouette score. Elbow curve is plotted using the KMeans Inertia values and proper K is chosen, the same is found to be true using silhouette score as well. K value was found to be 3 (optimal between 5 and 3 is chosen). Using K value as 3, the clusters are found using KMeans algorithm and are assigned back to the cleaned dataframe. Visualization is plotted against each cluster to find the cluster which is assigned to countries that require aid. It was found that cluster0 belongs to such countries, further the dataframe is filtered for cluster-0 and then the country is sorted as per GDPP, child mortality and income, and the top 10 are considered to be the priority countries who require aid from the NGO.

2. Hierarchical clustering

   In case of hierarchical clustering, a dendrogram is drawn using single and complete linkages, and it was found that the dendrogram drawn/plotted using complete linkage made more clarity, hence the complete linkage is used to plot the agglomerative dendrogram. From the dendrogram, it was found that the number of clusters (k) as 4 is the best choice and the cut tree is used to cut at a distance of 8-10 to obtain 4 clusters. The visualization is plotted for the 4 clusters and it was found that the clusters 0 and 3 are the countries in dire need of the aid.

   Decision: Both hierarchical clustering and the KMeans clustering produced similar results and the countries match in both the list, so they top countries from both the list are considered, first five results of countries (under developed countries) who are in need of aid are, Burundi, Liberia, Congo, Dem. Rep., Niger, Sierra Leone

QUESTION 2: CLUSTERING

1. Compare and contrast k-means clustering and hierarchical clustering.

Answer:

- K-Means clustering requires parameter K (number of clusters) to process the algorithm, whereas hierarchical clustering is hierarchy-based (divisive or agglomerative) structure formed grouping nearest data points, and clusters and so on., once the dendrogram is constructed, it can be cut at different heights to obtain different number of clusters. The hierarchical method is either divisive or agglomerative.
- K-Means uses centroid, either average/mean or median based clustering and the process if carried out with new centroids until all the clusters are converged. In case of Hierarchical clustering, agglomerative clustering, the clustering starts from n data points and then clusters are grouped one by one to form n-1, n-2 etc. and finally to form a single cluster.
- In K-Means clustering the initial K data points are picked randomly and this affects the results of the final clustering. Each times the points picked makes a difference in the final results, hence the K-Means algorithm is not fool proof and the results keeps changing depending on the initial data points chosen. Hierarchical clustering can be visualized using dendrogram and remains same, i.e. results are reproducible.
- K-Means clustering clusters the similar data points as one cluster and there will be different clusters depending on the characteristic of data points, where as in hierarchical the clusters are combined to form a single cluster.
- K-Means clustering is computationally faster, i.e. less intensive whereas the hierarchical clustering is computationally intensive, i.e. when large datasets are used in hierarchical clustering, it takes a long processing time compared to K-Means clustering which is faster for large datasets.
- K-Means clustering is highly affected by the outliers. The clusters and centroids are affected by the outliers incase of K-Means clustering, where as the hierarchical clustering takes care of the outliers and cluster it separately.
- K-Means clustering is apt when the cluster structure is hyper spherical whereas the hierarchical clustering does not fetch better results for spherical shape-based clusters.

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

Answer

K-Means algorithm has following steps.

1. Decide the number of clusters (k), as per business needs, or using elbow curve or silhouette score (statistical methods).
2. Starts by selecting k (number of clusters) points randomly as centroid.
3. Calculates the distance from each data point to the centroid.
4. Group the data points to the nearest centroid, using the distance calculated, and assign as a cluster.
5. Find the new centroid by taking the mean of all the datapoints of the newly formed cluster.
6. Repeat steps 3, 4 and 5 until the cluster centroids are converged to form distinct clusters.

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

Answer

In K means clustering K is one of the parameters required for the KMeans algorithm to work. The most used statistical way of obtaining the value of K is using (i)Elbow Curve and (ii)Silhouette Score

Elbow Curve uses the inertia values (sum of squared distances) of KMeans algorithm which is plotted against multiple number of clusters, and the point from where the curve becomes flat is noted as the optimal k value.

Silhouette score formula,

$$silhouette\ score = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Where $b(i)$ is the mean/average distance to the points in the nearest cluster (inter cluster distance) and $a(i)$ is the mean/average distance to all the points in its own cluster (intra cluster distance).

The value of the silhouette score range lies between -1 (data point is not similar to the data points in its cluster) and 1 (data point is very similar to other data points in its cluster). The optimal k value / number of clusters is the number of clusters corresponding to the higher silhouette score.

Business aspect

More often, business or domain understanding helps in deciding the number of clusters as per the requirement, it can be such that the retail company has few categories of item that need to be clustered and in such cases the number of clusters can be directly taken from the domain subject matter experts. Another example can be a customer visit to a restaurant or pub where customers can be daily visitors, weekend visitors, or holiday visitors, so based on the requirement of the restaurant it can be clearly stated that the cluster would be 3 i.e. k=3.

4. Explain the necessity for scaling/standardization before performing clustering.

Answer

K-Means clustering and Hierarchical clustering are distance-based clustering technique, where the centroids or clusters are formed based on the distance between the datapoints. Most cases each feature would have its own unit and they can be of completed different scale. Let's suppose the age and the income as the parameters, where age will be in the range 0-100 where as income would be in a higher value spread, say 10000-50000. So, it is important to bring all the features to a standard scale before applying K-Means or hierarchical clustering (any distance-based measure). Usually standardization or normalization can be performed to bring all the features to same scale so that the units are not affecting the clusters formed/created.

If we have mixed numerical data in different units and scale, standardization is an important step in data preprocessing. Standardization or scaling helps to control and manage the variability of features in the dataset. This improves the efficiency and effectiveness/accuracy of clustering algorithms and results in good quality clusters.

5. Explain the different linkages used in hierarchical clustering.

Answer:

Hierarchical clustering in python is usually achieved in 2 ways.

I. Using SciPy library.
SciPy has modules for cluster which includes linkages, dendrograms and cut trees. These are used to obtain the dendrograms.

II. Using Sklearn library
Sklearn has cluster module which contain AgglomerativeClustering, which has attribute 'linkage'

Different linkage types/methods that are available in SciPy/sklearn are

1. **Single** Linkage: In case of Single linkage, distance between clusters is obtained as the minimum distance between all observations in the clusters.

   $$d(u, v) = min(dist(u[i], v[j]))$$

   - u and v are the clusters
   - i and j are the data points in cluster u and v respectively

   Single Linkage sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters. These clusters can appear spread-out.

2. **Complete** Linkage: In case of complete linkage, distance between clusters is obtained as the maximum distance between all observations in the clusters.

   $$d(u, v) = max(dist(u[i], v[j]))$$

   - u and v are the clusters
   - i and j are the data points in cluster u and v respectively

   This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together.

3. **Average** Linkage : In case of average linkage, it uses the average of the distances of each observation of the two clusters. The distance between clusters is taken as the average distance between each observation in one cluster to all the observations in other cluster.

$$d(u,v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| \ \times |v|)}$$

- u and v are the clusters
- i and j are the data points in cluster u and v respectively

This is also known as UPGMA (Unweighted Pair Group Method with Arithmetic Mean). Average-linkage and complete-linkage are the two most popular distance metrics in hierarchical clustering.

4. Weighted Linkage: This is also known as WPGMA (Weighted Pair Group Method with Arithmetic Mean) algorithm, which constructs dendrogram that reflects the structure present in a pairwise distance matrix. At each step, the nearest two clusters, say s and t are combined into a higher-level cluster $s \ \cup \ t$. Then, its distance to another cluster v is simply the arithmetic mean of the average distances between members of $v \ and \ s$ and $v \ and \ t$:

$$d(u,v) = \frac{dist(s,v) + dist(t,v)}{2}$$

where cluster u was formed with cluster s and t, and v is a remaining cluster in the forest

(Weighted linkage is not available in sklearn agglomerative clustering as per documentation)

5. Centroid Linkage: Centroid-linkage is the distance between the centroids of two clusters. As the centroids move with new observations, it is possible that the smaller clusters are more similar to the new larger cluster than to their individual clusters causing an inversion in the dendrogram.

$$dist(s,t) = ||C_s - C_t||_2$$

- where C_s and C_t are the centroids of clusters s and t respectively.
- When two clusters s and t are combined into a new cluster u, the new centroid is computed over all the original objects in clusters s and t. The distance then becomes the Euclidean distance between the centroid of u and the centroid of a remaining cluster v in the forest.

This is also known as the UPGMC (Unweighted Pair Group Method with Centroid) algorithm. *(Centroid linkage is not available in sklearn agglomerative clustering as per documentation)*

6. Median Linkage: In this linkage, it assigns $d(s,t)$ like the centroid method. When two clusters s and t are combined into a new cluster u, the average of centroids s and t give the new centroid u. This is also known as the WPGMC (Weighted Pair Group Method with Centroid) algorithm.

(Median linkage is not available in sklearn agglomerative clustering as per documentation)

7. **Ward** Linkage: In case of ward linkage, it minimizes the variance of the clusters being merged. The distance between clusters is calculated as the sum of squared differences within all the clusters observed. It uses the Ward variance minimization algorithm. The new entry $d(u,v)$ is computed as follows,

$$d(u,v) = \sqrt{\frac{|v|+|s|}{T}d(v,s)^2 + \frac{|v|+|t|}{T}d(v,t)^2 - \frac{|v|}{T}d(s,t)^2}$$

   - Where u is the newly joined cluster consisting of clusters s and t, v is an unused cluster in the forest
   - $T = |v| + |s| + |t|$ and $| * |$ is the cardinality of its argument. This is also known as the incremental algorithm.