Assignment 5: Social Network Profiles Clustering

Ryan Fischbach
Dr. Khuri

November 13th, 2020

1 The Dataset

The Social Network Profiles dataset contains 30,000 samples (hypothetical users of a social network) with 40 attributes. The first four attributes contain information about the sample (user):

Column Index	Attribute
0	Graduation Year
1	Gender
2	Age
3	Number of Friends

The rest of the 36 attributes are binary indicator variables suggesting that that user's account activity involves that item. This could suggest general interest in these activities or a mention of them in some way. These binary variables are: 'basketball', 'football','soccer', 'softball', 'volleyball', 'swimming', 'cheerleading', 'baseball', 'tennis', 'sports', 'cute', 'sex', 'sexy', 'hot', 'kissed', 'dance', 'band', 'marching', 'music', 'rock', 'god', 'church', 'jesus', 'bible', 'hair', 'dress', 'blonde', 'mall', 'shopping', 'clothes', 'hollister', 'abercrombie', 'die', 'death', 'drunk', 'drugs'.

2 Preprocessing the Data

To enable the best clustering of this social network dataset, some modifications were made.

1. Duplicates samples were removed to prevent that sample from having a larger weight than others. After removing duplicates, 29,350 samples remained.

2. Null/NA values were identified and mitigated. Null values were included in the "gender" and "age" attributes, likely because the user did not fully complete their profile on the social network. Imputing the samples with NA values with the mode for "gender" and mean for "age" was an option, but ultimately not chosen. Age and gender contain valuable information for clustering. After removing samples with NA values in gender or age, 23,968 samples remained.

3. 'gender' was binarized and converted to an integer (1 if Male, 0 if Female) to allow the clustering algorithm to use this attribute.

4. Binary indicator variables were removed because of their near 0 variances. The binary indicator attribute portion of the dataset was sparse and didn't add much value in terms of clustering. These attributes will be used after clustering to interpret clusters and determine astrategy for the marketing campaign.

5. Redundant (highly correlated) attributes were searched for, but none were identified in this dataset.

6. All attributes were rescaled via the SciKitLearn StandardScaler method to ensure that no attribute had a larger impact than another on clustering.

This data preprocessing left us with 23,968 samples with 4 attributes: 'gradyear', 'gender', 'age', 'friends'.

3 Data Exploration

After data preprocessing, data exploration was performed to identify any important patterns visually.

The correlation matrix of the remaining 4 attributes was created, but no interesting relationships emerged. The largest correlation relationship was 0.111 between 'age' and 'gradyear'. Additionally, the correlation matrix split between females and males, split between higher than average and lower than average friends, and split between higher than average and lower than average age was created.

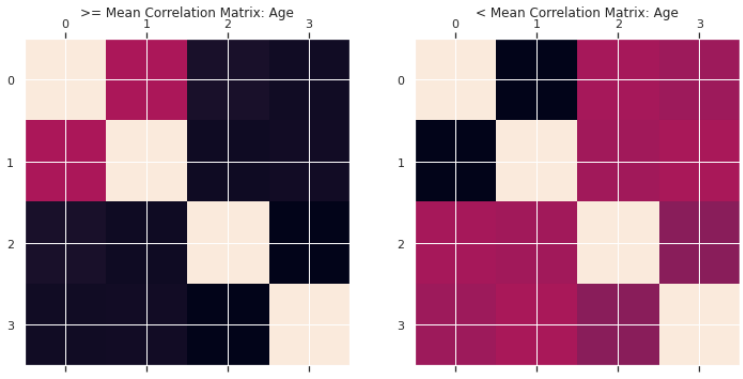

Figure 1: *Correlation matrix for two subsets of the data after preprocessing. On the left, the correlation matrix for every sample greater than or equal to the mean age for all samples. On the right, the correlation matrix for every sample less than the mean age for all samples. The right suggests that younger people tend to have a higher correlation between attributes than standard.*

4 KMeans Clustering

KMeans from the SciKitLearn library was used to cluster the data. Multiple values of k ranging from 2 to 10 were used and two metrics were computed to determine the best value of k for KMeans.

4.1 Elbow Plot

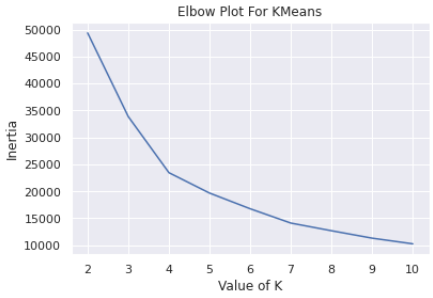

Figure 2: *An elbow plot was constructed using the inertia (a measure of cluster performance) on the value of k. The curve initially decreased in an exponential way, but reached an inflection point around $k = 5$ and began to start decreasing linearly after this point. For this reason, the elbow plot suggests the best number of clusters is 5.*

4.2 Average Silhouette Plot

Silhouette score is a metric to measure how similar a sample is to its cluster. The mean silhouette score for a value of K was computed via averaging the silhouette score for all samples in a cluster and average that score across clusters.

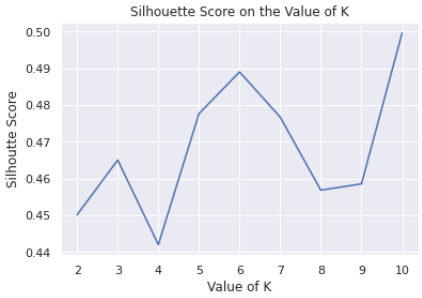

Figure 3: *The silhouette score was plotted on the value of k, with the higher k value being "better". The silhouette score increased as the value of k went up with high variability. k = 10 had the highest score of 0.5, with k = 5 and k = 6 having 0.48 and 0.49 respectively.*

The difference in silhouette scores between k = 5, 6, and 10 is negligible. The end goal of this clustering analysis is to provide different groups for marketing materials, thus a smaller number of clusters is better. k = 5 seems to be a good number of clusters based on this metric.

5 Hierarchical Clustering

Agglomerative Clustering from the SciKitLearn library was also used to cluster the data.

A dendrogram shows how samples clustered together based on their relative distance away from each other. This arrangement of clusters as well as the length of lines can show how well-separated samples or clusters are from each other. A dendrogram was produced (using Euclidean distance and Ward linkage) to determine how the social network profile data clustered and was used to find an ideal cut-off point for the number of clusters.

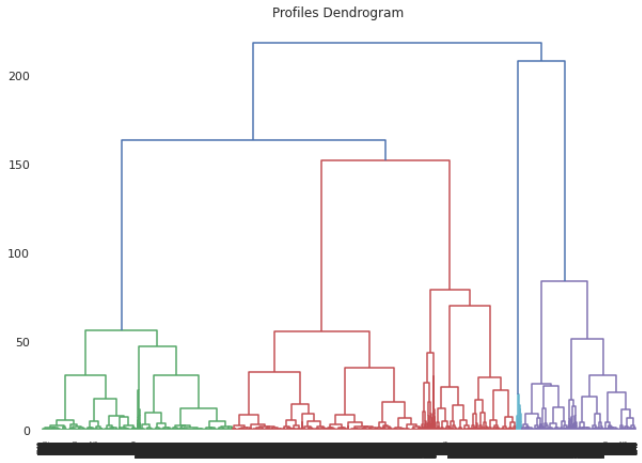

Figure 4: *The dendrogram has the distance on the y-axis and each sample on the x-axis. Many samples clustered together at low distances, suggesting that these samples are close together. As the distance increased, more well-defined clusters started to form around distance 100. There is a large distance (around 50) before clusters were merged, suggesting that these 5 clusters are well separated. Thus, $k = 5$ is validated via the dendrogram.*

6 Comparison Of Clusters Produced By KMeans And Agglomerative

To ensure that the resulting 5 clusters from both algorithms are similar and don't identify different patterns in the data, both results were compared.

Principle Component Analysis (PCA) was used to reduce the data's dimensionality to 2 so that the results could be visualized. A scatterplot was created to plot these two dimensions, with each color representing a different cluster.

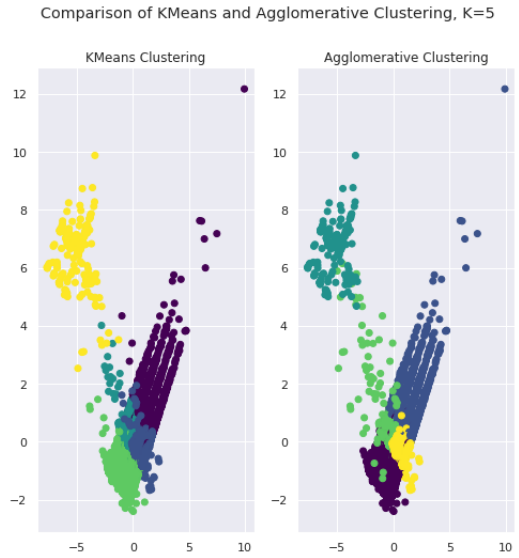

Figure 5: *Scatterplots created via PCA show the resulting clustering from both methods. KMeans Clustering is on the left and Agglomerative Clustering is on the right, both producing 5 clusters. The data seems to be clustered similarly, with minor variations towards the middle where there is a lot of overlap. This could be because PCA is not accurately representing the differences between these samples that exist in higher dimensional space or differences in how the algorithms cluster data.*

7 Analysis of Resulting Clusters and Recommendations for an Optimal Targeted Marketing Campaign

After verifying the similar clusters produced by the two algorithms, each cluster can now be analyzed to determine the optimal target advertising strategy for each cluster.

To get the identifying characteristics for each cluster, an average for all attributes in that cluster was created.

Cluster	gradyear	gender	age	friends
0.0	2006.490828	0.000000	18.219017	21.606847
1.0	2008.504748	0.000000	16.293284	22.956671
2.0	2007.808511	0.244681	101.088707	27.808511
3.0	2007.700943	0.031881	17.076433	112.372699
4.0	2007.349453	1.000000	17.516645	22.885996

From the average value of each attribute for each cluster, the clusters can start to be distinguished.

Cluster 0: Young women (average gender = 0, average age = 18.21) who graduated in 2006 and have an expected number of friends in the social network (average friends = 21.6). These women have below-average mean for almost all binary variables, suggesting that they do not interact with the social network or do not have any association with these activities.

I would advertise "young adult" products to this group such as clothes, technology, and other items that are bought by young people with their sources of income. The average age of this group is 18, suggesting that this group is nearly close to being done with schooling and could potentially have their source of income outside of their parents. Thus, this cluster has some purchasing power that isn't available with younger groups.

Cluster 1: Young women (average gender = 0, average age = 16) who graduated in 2008 and have an expected number of friends in the social network (average friends = 22.95). These women graduated on average two years later so are two years younger relative to Cluster 0. This cluster had higher means almost all binary indicator variables, suggesting that they are more interactive with the social network than Cluster 0. This group had the highest mean 'soccer', 'volleyball', and 'swimming' values, suggesting that they could play these sports in school or outside of school.

I recommend advertising women's sporting items such as balls, cleats, or clothing to this group. This cluster has a high interest in several sports, but only has an average age of 16. Thus, they likely don't have much money to spend beyond their allowance or what their parents will buy for them. Parents are more likely to spend money on activities the kids enjoy, and all indications point to this cluster enjoying these specific sports.

Cluster 2: Mostly old women (average gender = 0.24, average age = 101) who graduated in 2007 and have an expected number of friends in the social network (average friends = 27.8). Suspicion tells us that a decent amount of these users likely lied about their age to gain access to the social network, providing false data. Some samples are likely legitimate as well. This cluster has

the highest average in 'music', 'tennis', 'rock', and 'die'. This tells us that music, specifically rock, could be an interest to this group.

I suggest advertising a combination of music and grown-up products to this cluster. The music seems to appeal to this cluster, and would likely also appeal to the younger users who faked their age. Additionally, I would suggest adding on products that have a higher price tag because these individuals likely have sources of income or savings, but avoiding technology.

Cluster 3: Young women (average gender = 0.03, average age = 17) who graduated in 2007 and have the highest number of friends out of the social network (average friends = 112). These are the highest use individuals out of all users on the platform. They have the largest mean value for 'mall' and 'shopping', and 'abercrombie'. This could be because of their increased use on the platform or because they have an interest in these items.

To advertise to this group, I would suggest using products such as eCommerce websites, clothes, or other social media networks that appeal to their current activities. This group already uses the platform and has high engagement, so they will be likely to see and click on an ad for another technology or one of their interests.

Cluster 4: Young men (average gender = 1.0, average age = 17.5) who graduated in 2007 and have a reasonable number of friends (average friends = 22.88). They have the highest mean value for 'football', 'basketball', 'baseball', and 'sports'.

I would advertise men's sports products, sports drinks, or sports streaming services to this group. They still aren't old enough to have a stream of income, so no high price tag items will likely appeal to them and generate sales. Items that their parents can buy for them or that they can buy in a convenience store will likely produce the most sales and activity from advertisements.

8 Citations

[1] Social Network Profiles.csv
[2] Tan et al. 2005. Introduction to Data Mining, 88 pages.
[3] Numpy. 2020. https://numpy.org
[4] Pandas. 2020. https://pandas.pydata.org
[5] Matplotlib. 2020. https://matplotlib.org
[6] Seaborn. 2020. https://seaborn.pydata.org
[7] Scikit learn. 2020. https://scikit-learn.org/stable/
[8] Natalia Khuri. 2020. lecture-7, 80 pages.
[9] Natalia Khuri. 2020. lecture-8.