Clustering Callers

Mitch Lowry

May 2016

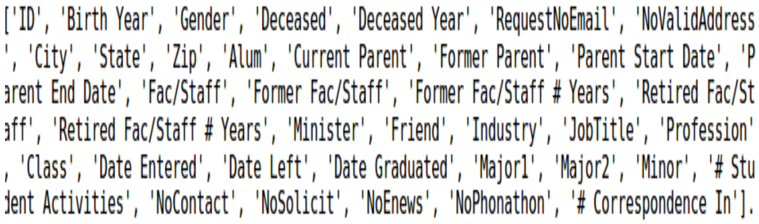
1 Introduction

The project undertaken for this class was provided by the Office of Advancement at Hendrix College. The Office of Advancement has been collecting data on potential donors and event attendees for a while, and the Office of Advancement gave us this data, hoping that we would be able to find some useful relationships. The objective of this data mining excursion was to tell the Office of Advancement how they can alter how they contact people in order to improve donation and event outcomes. This is important because contacting people in such a way that they donate more or attend more events implies that you are contacting these people in a way that they prefer, in a way that makes them want to be more affiliated with or donate more resources to the college. Additionally, improving event turnouts is a good indicator of the overall image of Hendrix. More people at events means more people want to be there, and it also means that there are more possibilities for the college to connect with people on a personal basis, which could improve future donation and event outcomes.

Clearly, our objective is a noble one, but how to go about accomplishing it is somewhat tricky. Our idea for accomplishing this consisted of two things. The first one was just looking around at the data a bit to see if we could find anything interesting. The second idea, and main focus of analysis in the paper, is that there are different groups of people who prefer to be contacted in different ways, and if we can try to put these people into different groups, we might be able to obtain some insights about how actions should be altered for each of the groups. To do this, we spent a decent amount of time looking into clustering algorithms.

2 Data

The Office of Advancement gave us four password protected Microsoft Excel files, Bio.csv, Events.csv, Actions.csv, and Gifts.csv. The college has been collecting biographical information on potential donors and event attendees. In Bio.csv, each row is contains a person's identification number along with that person's biographical information. The labels for this file are below:



```
['ID', 'Birth Year', 'Gender', 'Deceased', 'Deceased Year', 'RequestNoEmail', 'NoValidAddress
', 'City', 'State', 'Zip', 'Alum', 'Current Parent', 'Former Parent', 'Parent Start Date', 'P
arent End Date', 'Fac/Staff', 'Former Fac/Staff', 'Former Fac/Staff # Years', 'Retired Fac/St
aff', 'Retired Fac/Staff # Years', 'Minister', 'Friend', 'Industry', 'JobTitle', 'Profession'
, 'Class', 'Date Entered', 'Date Left', 'Date Graduated', 'Major1', 'Major2', 'Minor', '# Stu
dent Activities', 'NoContact', 'NoSolicit', 'NoEnews', 'NoPhonathon', '# Correspondence In'].
```

The identification number used in Bio.csv can be used to obtain information about that person's event attendance, donation history, and how that person has been contacted over time. Each row in Events.csv records an instance of someone attending some event on some date, and each row in Gifts.csv records an instance of someone or some organization donating some amount of money on same date. The most interesting file for this analysis is Actions.csv. The college can contact someone by email, phone, meeting, or mailing, and the college records each time it contacts someone as a row in Actions.csv, where the row contains the id of the person being contacted, how the person was contacted, and on what date the person was contacted. The file does not contain mass emails which are frequently sent out or the biannual Phonathon event; however, this file, which has been recorded since April 30, 2002, provides pretty useful information. Since the focus of this project was how actions could be altered, we only included data during the time frame during which data about actions was available. To do this, we sorted Events.csv, Gifts.csv, and Actions.csv by date and removed any rows in these files that had dates preceding the start of Actions.csv. These files were exported as csvs and data analysis was performed in an Ipython notebook, which is stored in this directory.

3 Approach

Clustering algorithms are a good way to take large amounts of data and create subgroups based on specified characteristics. In this paper we implement the most common clustering algorithm, the k-means algorithm, which aims to partition n observations into k clusters where each observation is a d-dimensional vector that is put into the cluster that has a mean nearest to it's value. That is the goal is to minimize the sum over each of the clusters of the within cluster sum of squares value between the mean of the cluster and the points belonging to that cluster. This is an NP-hard problem, so there are a few locally optimized algorithms that accomplish this goal [Mac67]. The implementation used in this paper was Lloyd's algorithm, which initializes points, called centroids, in some say, and each observation belongs to the centroid nearest to it. The algorithm then repeats two steps until it terminates. It adjusts each centroid to be the mean of the observations that belong to it, and then, since the centroids have moved, the observations are reassigned to whatever centroid they are closest to. This continues until no observations change clusters, and it has been proven that the process monotonically decreases the sum of the within cluster sum of squares [Das13].

Note that this algorithm is finding a local optimum because it requires that we start by initializing centroids in some way. Analysis was done using the Python 3.0 kernel in a Jupyter notebook using and the scipy.cluster.vq library for clustering with the kmeans function, which initializes points simply by choosing centroids at random [com09]. There are certainly better methods for initializing these centroids, and this method is fairly poor in that the results of clustering are very sensitive to outliers. There are other methods of initialization, such as kmeans++, that cause clustering to be less sensitive to outliers [Das13].

Choosing the correct number clusters, k, for the clustering algorithm is a balance between compression and accuracy, and many times the correct value of k is ambiguous, but there are methods for estimating k. The most common method for doing this is called the elbow method. For this method the tradeoff between compression and accuracy is shown visually by making a connected scatterplot with a measure of the benefit of having more clusters on the y-axis and k on the x-axis is used. We are looking for a value of k where adding more values of k will not improve accuracy much, and this point tends to look like an elbow on the aforementioned connected scatterplot [CM12].

After performing the elbow method and clustering, principal component analysis (PCA) is used to get a visual representation of how good our clusters are. PCA is reducing the dimensionality of data by finding dimensions with the most explained variability [Shl01]. In this way, we can attempt to view d dimensional objects in two or three space.

4 Work Done

After loading the data into in Ipython notebook, I first removed data from impossible dates from Actions.csv, and then I checked all of the data for duplicate entries or other errors. Since we are interested in the amount of money donated by people, I figured that we should account for inflation and price changes, so I downloaded CPI data from the St. Louis Federal Reserve Bank and used it to add an additional number, the value of the gift in 1982-1984 dollars, to each recorded instance of a gift. I then created various hash tables with the given information. I created a hash table containing each person's given biographical information so demographics can be analyzed. I created the hash tables events history, gift history, and actions history where the key is an identification number of a person, and the value is the kind of history associated with that dictionary and person. These tables can be used to create time series data for individuals about their donations and event attendance; however, I used them simply to compute y_i the number of years each person i was in the data and the number of posthumous donors. We computed y_i by finding the difference between the date of the first form of contact, an action, event, or donation, and either the death date or the date the actions file stopped being recorded. This value is recorded as a floating point by uniformly at random choosing the death month of deceased people under the assumption that people die randomly throughout the year. The value is also computed in such a way that if person i is in the data for less than a year, we let $i = 1$ because these people would probably not donate much more anyway, and dividing by small fractions can really overstate how much people would have donated yearly and cause crazy outliers.

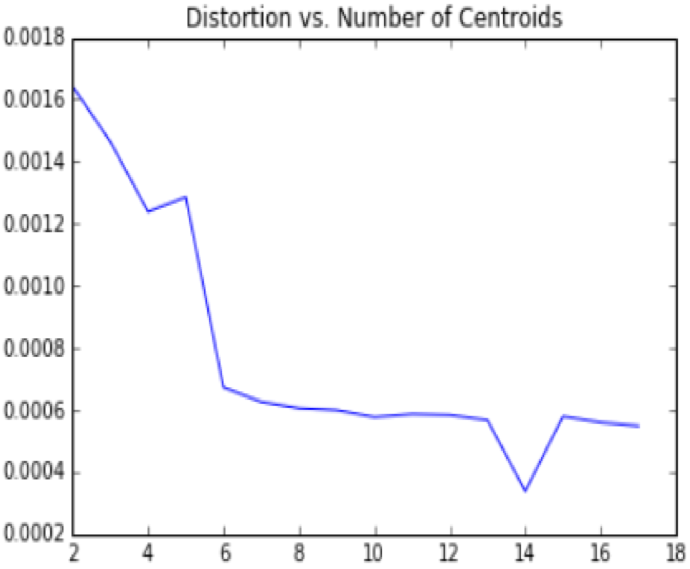
After having created these histories, I then created hash tables event count, action count, and gift count. The values of event count are the total number of events a person has attended; the value of actions count are $[a_i, b_i, c_i, d_i]$ where a_i is the number of time person i has been emailed, b_i is the number of time person i has been called, c_i is the number of times person i has had a meeting, and d_i is the number of times person i has been mailed; and the values of gift count are $[e_i, f_i, g_i]$ where e_i and f_i are the total nominal and real amounts donated by person i and g_i is the number of times person i donated.

After having created all of these connected hash tables, we could easily compute various metrics. The objective is to cluster by amount donated and how the person was contacted and see if there are any differences among clusters that provide some insight about how to alter actions of the Office of Advancement. Since people enter the data on different years, and leave the data on different years, we computed $\frac{e_i}{y_i}$, $\frac{f_i}{y_i}$, and $\frac{g_i}{y_i}$ for i in gift count as measures of donations and the average number of times a person is contacted per year as a measure of actions. In addition to this measurement, for each person we found what percentage of total types of contact fell into each category. For each person I created a seven dimensional vector and clustered over these all of these excluding

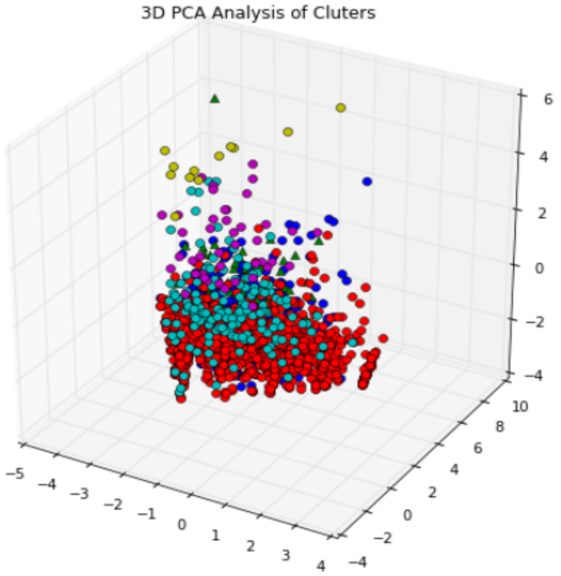
$\frac{e_i}{y_i}$.

Clustering was done many times and a bit of outlier filtering was done. There are 16686 individuals in these files since the start of Actions.csv, and because we were measuring percentages of contact, we decided to filter out people who have had 3 or less contacts because they cause a bunch of weird percentages that have a large influence on the clustering algorithm. After filtering these people out, we have 2231 people left, and then removing posthumous donators from that group leaves 2143 people. We initially clustered over these people, making sure to run the elbow graph analysis multiple times. When doing this we noticed that someones we would get a standard looking elbow graph with the elbow at four, but other times we would an elbow graph with kinks in it with the first kink at four, like the one below.

Fourth elbow graph including outliers:



Literature says that kinks in these graphs can be cause by extreme outliers [CM12]. We performed PCA analysis in two and three dimensions, and indeed there seemed to be a few extreme outliers. The most variation of the factors clustered over is present in the amount donated per year. There six individuals from this original clustering that donated absolutely staggering amounts of money, which was throwing our analysis off. We removed these people and reclustered. Our elbow graphs looked a lot less kinky. When choosing the number of clusters here, one should choose the number of clusters right before the first kink occurs, so long as a certain amount of variation is explained [CM12]. Based on this, we chose to create five clusters. Below is a three dimensional PCA analysis of the final clusterings, to give a visual idea of how good the clusters are.



5 Results

There are 1046 donors in any of these files since 2002 that have died, and of these dead donors 673 gave posthumous donations. This is a large percentage of donors, so developing good relationships with elderly donors is a good idea. For our analysis, we removed posthumous donors because the amount donated was adjusted by year based on date of death; however, we still have these people's identification numbers, so we could perform a separate analysis on these people, if desired.

Below is a table providing information on each of the final five clusters, which include people who have been contacted at least four times, are not posthumous donators, and donate less than an average of 48000 1982 dollars per year (more than 98000 current dollars).

Measure	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
number	68	12	43	230	1778
adj amt	5877	40689	16912	1421	145
dev amt	1909	4885	4389	656	157
ave freq	3.36	1.43	2.52	3.41	1.56
dev freq	5.71	0.85	2.94	4.92	2.51
ave email	0.257	0.235	0.235	0.238	0.356
dev email	0.207	0.245	0.176	0.186	0.250
ave mail	0.178	0.094	0.158	0.241	0.244
dev mail	0.151	0.110	0.172	0.170	0.188
ave phone	0.266	0.344	0.299	0.261	0.220
dev phone	0.154	0.200	0.182	0.173	0.193
ave meeting	0.300	0.328	0.308	0.260	0.181
dev meeting	0.147	0.198	0.159	0.180	0.177
act freq	1.79	1.99	1.99	1.38	0.918
dev act	0.89	1.15	1.22	0.89	0.734
male percent	0.69	0.75	0.56	0.60	0.55
female percent	0.31	0.25	0.44	0.40	0.45
deceased percent	0.103	0.333	0.163	0.065	0.014
ave age	72.9	79.0	74.6	68.3	57.1
dev age	12.3	11.60	11.6	13.6	15.3
alum percent	85.3	58.3	51.2	72.2	74.9
former parent	19.1	33.3	34.5	36.5	25.8
no solicit	4.4	0.0	4.7	1.3	1.6

In the above table, adj amt is the average adjusted yearly amount donated, ave freq is the average yearly number of donations, the next four averages are average percentages of contact, act freq is the average number of times per year contacted, and all statistics starting with dev are standard deviations. There are many trends that one could spend a lot of time looking at.

One thing to notice is the particularly large number of females in cluster 3 even though the average age is around 68, during a time when there were proportionally less women in college. It is likely that there are a lot of married couples in this group or people in a high-paying gender neutral field, and this can be easily checked using the ids of married people that were emailed to us. There is also a large number of women in cluster four, and both cluster three and cluster four have a large percentage of former parents, so it is likely that there are a large number of married couples in these groups, even ones with children. One idea is to try to get Hendrix students and alumni dating by promoting dating culture on campus, hosting events for single alumni, or other methods.

Based on the similar ages in groups one through four, and the fact that people in cluster four are contacted the least and donate the least, it might be a good idea to contact people in this cluster a little more. They receive phone calls less frequently than the first three clusters, and more mail. I think I good plan here would be to tone the mail down slightly and call them more often. Even better would be to look at them individually and try to figure out a way to coax them into coming to a meeting or two.

6 Conclusion and Future Work

This short exercise in data mining turned out to be mostly unfruitful; however, much of the ground work for future clustering has been done here. I have processed the data and stored it in connected hash tables and wrote a function you can pass keys, outliers, and a maximum number of clusters, and it will perform the elbow method for you using whatever you are clustering over. I have also made a pretty easy process for extracting the biographical information of people clustered, but I could work on making it more automated. These tools make it easy to cluster over different subsets of the data, perhaps to look for kinds of alumni, parents, etc. I have stored the ids of different subgroups in lists (parents, former parents, alumni, staff, etc.), and I could simply pass these into the function that I wrote. Perhaps more interesting relationships can be found by pulling out all of the biographical information for clusters and performing an analysis somehow. It would be interesting to cluster people based on their donation history and then see if there are any interesting trends in the biographical information about these cluster. If the Office of Advancement is interested lists of ids for anyone in these clusters, posthumous donors, or other categories, just shoot me an email.

References

[CM12] Mahima Chandra and Kevin Mabe. *An Algorithm to Find a Business-Appropriate 'Optimum' on an Elbow Curve.* 2012. URL: www.kevinmabe.com/files/Finding_Elbow2.doc.

[com09] The Scipy community. "K-means Clustering and Vector Quantization". In: (2009). URL: https://docs.scipy.org/doc/scipy-0.7.x/reference/cluster.vq.html.

[Das13] Sanjoy Dasgupta. "Lecture 3 - Algorithms for k-means Clustering". In: (2013). URL: http://cseweb.ucsd.edu/~dasgupta/291-geom/kmeans.pdf.

[Mac67] J. B. MacQueen. "Some Methods for classification and Analysis of Multivariate Observations". In: *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (1967), pp. 281–297. URL: http://projecteuclid.org/euclid.bsmsp/1200512992.

[Shl01] John Shlens. *A Tutorial On Principal Component Analysis.* Princeton University, 2001. URL: https://www.cs.princeton.edu/picasso/mats/PCA-Tutorial-Intuition_jp.pdf.