Shane Alvarez

Data Mining Project 1 Part 2: Clustering

Data Sets

For this part of the project, the data sets I choose describes patients of Thoracic Surgery and Credit Card defaults. I choose the surgery data set as it offered a lot of data about the patients' situations without many strong associations, and I figured that clustering may help to identify similarities between patients. The credit defaults data set was chosen for a similar reason. I wanted to see if it was possible to identify a group of individuals with common situations that possible promoted credit defaults.

R-Packages

I used the *foreign* and *dbscan* packages. The *foreign* package was needed to read the ARFF file that contained the surgery data. The *dbscan* package was used for plotting KNN distance data and, using that data, performing a *dbscan*. Additionally, I use *stats*, which provides AGNES clustering and distance calculations (I suspect this package was preinstalled, but include it nonetheless).

Data Modification

Because the values of the data for both datasets weren't very distributed or integer values, the data sets were modified inside of R to normalize the values because both sets included many T/F values. For the surgeries dataset, this included PRE5 (for which some values were much larger than others), and AGE. For credit cards, this include AGE and many all the bill amount and payment amounts.

Additionally, due to the large number of samples in the credit card defaults set, sample is taken for processing. This is primarily due to the plotting of results, which can overwhelm the R plotter.

Results

Running AGNES clustering on the surgeries data set revealed several possible clusters. From the dendrogram, there seems to be around 8 or fewer reasonable clusters to be identified. In the code, these groups are displayed. Printing some aggregate results from the different clusters reveal that there are certainly come trends, particularly in groups where some of the T/F values were all false. It also made it easy to identify the small but very similar group of people.

Running DBSCAN clustering on the surgeries data set revealed a PRE4 and AGE have some kind of correlation, but the abundance of T/F types in the results reduced the amount of information that could be gained from the data, as far as the plot is concerned. In addition to being harder to properly tune, it was also more difficult to gather granular data from groups. Indeed, less groups could be extracted. Also, the group of somewhat unique cases identified above were seemingly marked as outliers, and data they provided remain unseen. Finally, the plotting takes notably longer.

Running AGNES clustering on the credit defaults data set gave a dendrogram that seemed to best split the group into 4 clusters. It also seems that those clusters seem to be based around credit limits, showing a higher likelihood for defaults along the lower end and higher end of that spectrum. Additionally, the revealed a correlation between the higher limit balances and the bill amounts. In all, it seemed to reveal a lot of information, even on the sampled set.

Running DBSCAN clustering on the credit defaults data set quickly showed many more correlations than the AGNES clustering had. However, it should be noted that the plotting took several times longer to complete. Also, given the dimensionality of the data, it requires a very large plot to be usable. Additionally, the clustering gave only a single group and left the rest as outliers. Aggregating the data from this group provided no real insight.

Overall, ANGES seems very efficient for exposing larger trends in a data sets while still offering a lot of control. However, DBSCAN seems much more helpful in non-binary cases where trends can be better visualized and analyzed.