Lab 2a Report

Eric Sclafani

1 Data set background

The data set I chose to use this time is related to my research. I am on an IARPA project where our goal is to create a model that can do authorship attribution, which is the task of automatically detecting the author of a document. The first stage of this project involves creating long feature vectors indicative of authors stylistic tendencies.

I have been working on a system since last October that reads in text documents and converts them into vectors using feature extraction techniques. The data set I have been using in my research is called PAN 22 Authorship Analysis: Authorship Verification. This data set contains 1047 text documents complete with their author labels.

To get the specific feature data set for this lab, I processed the PAN22 data with my feature extractors and populated a csv file with them. Thus, every column is an individual feature extracted from text. Most of these features are count based. A.k.a, they count the relative frequencies of specific items and records these counts. The total amount of features is 407, which makes this a very high-dimensional data set.

One major drawback about this data set, which may even encourage me to switch text data sets, is that it only has 1047 data points. Also, each authors has a varying amount of documents, so a heavy class imbalance also exists

2 Implementation notes

Like in my previous lab, I have decided to keep all my code contained in one file since it's a small amount of lines. I have created section dividers that attempt to increase the readability of the code. Regarding the webpage

itself, I decided to display everything on one page since all of the components are technically related to each other, since most of them use the results from principal component analysis.

3 Observations

In the first scree plot, it is clear that after the third or fourth principal component, the explained variance doesn't increase much. The highest the cumulative explained variance gets is around 48%. I found it interesting that after the first components, the explained variance dips drastically.

Looking at the K-means error plot, I used the elbow method to find a good K to cluster my data. In my plot, the elbow lands at $k = 3$, with the sum of squares error being around 61,000.

Regarding the biplot and scatterplot matrix, because my data set has so many dimensions, it is difficult for me to interpret what these plots are telling me. In my biplot, two of the clusters are very close to each other (the blue and dark pink ones), while the other one is dispersed (yellow one). The same can be said for my scatterplot matrix.