IST 565: Data Mining
Final Project Proposal
Max Gerstman, William D. Lombardi, James Lu, Adam Miller
March 27, 2018

Dataset

The dataset our group plans to use comes from Kaggle. It contains information that describes different crimes that occured in New York City from 2014-2015. There are 24 columns with over 1,000,000 records. These attributes range from nominal to ratio-like in nature. The link for the complete dataset can be found below:

https://www.kaggle.com/adamschroeder/crimes-new-york-city/version/1#

Preliminary Data Preparation

The first task with this particular set of data is going to be dimensionality reduction, because we can assume that training a model on over 1,000,000 examples will be prohibitively time-consuming. In order to do this, we first need to narrow the scope of our data. We plan to do this by focusing our efforts on a subset of the data which represents crimes that occured in bars/nightclubs in the borough of Queens. Hopefully, with reducing it to a single borough and a specific kind of location, the dataset will be a much more manageable size.

Also, using the 'complete.cases' function in R, we plan to find out if there are any incomplete or missing records that we can then omit or fill in using an appropriate average value.

Our Suggested Algorithms

Insofar as methods we'd like to investigate, we plan to try applying K-Nearest Neighbor, K-Means Clustering, and Support Vector Machine algorithms to the data. Based on our results, we will attempt to build a model that can predict the time and location at which any given bar/nightclub crime will take place. Another avenue of investigation we're interested in pursuing is using a Naive-Bayes algorithm to predict the level of crime that will be committed (i.e. felony, misdemeanor, or violation).