Melody Zhao, Jessical Kwok

Math189R SP19

Midterm Proposal

Monday, Feb 10, 2019

Midterm Proposal Project

The data that we decided to investigate from is the historical stock prices for all S&P500 indices from 2013 to 2018. For each company, the data points include open, high, low close, and volume. There are more than 600,000 of data points for all of the 500 companies. First, we plan to visualize the data to see if there are any general trends between the companies. We can try dividing the companies by size, value, and stock prices. We can try visualizing if there are any trends among the larger companies vs. the smaller companies. Another way to divide the companies is by industries like technology or services. We are also considering comparing stock values to U.S. unemployment rate and GDP, then calculate the correlation between these factors.

Next, we plan on using our data and existing machine learning algorithms to see if we can create a prediction model. We plan on using supervised ML models such as linear regression, logistic regression and multilayer perceptron. Some ideas regarding prediction models include using 450 companies as the training set, and see if we can predict the stock prices of the last 50 companies based on their open, close, volume, and U.S. economy characteristics. Finally, we can verify the model by comparing to our existing data set. If we need to expand our model in the future, we can also utilize data from the other indices.

To visualize our data, we plan on using Python, including Numpy, Scipy libraries as well as neural packages for existing algorithms.