ML Stocks Progress Report
CS 4478/5599 Machine Learning

Tyler Hedgepeth, Nickolas Nathan Taylor

October 22, 2020

1 The Problem

The problem we are attempting to solve is the ever present problem of predicting future stock prices. This task is generally considered impossible for a human to do, given the sheer amount and complexity of the data to consider. Even professional traders end up with results no better than random. We hypothesize that a machine learning algorithm will yield some improvement to human, or random, prediction. Given the age of technology and stock trading, there are ample data and tools to work with. Using this information, we will be predicting 'breakout' stocks–stocks which will see a specified percentage of growth within a specified amount of time.

Using 'candle' datasets that contain 'end of day' or EOD information we can train a regression model on this time-series data and predict the features of the candle for the next day. A candle contains the most pertinent information: Opening price, closing price, daily high, daily low, and volume sold.

Another consideration is to train different models for different groups or industries within our dataset. There are many websites that group various stocks into industries and those can be references to build different datasets. Because we are not considering any outside factors directly into our model, we could compensate for those by only training on groups that are affected by the same external stimuli.

2 Initial Models

We have been looking into different model possibilities. The most promising so far is a time-series regression model. The time-series requirement is really a 'no-brainer' in that anyone could tell you the stock market changes over time.

There are two main cross-validation methods we are considering: sliding window and forward chaining. At this point we don't know enough to commit to one or the other. Sliding window validation looks a set span of time and considers all datapoints therein to predict an equal span of target future datapoints. Forward chaining begins the same way, by looking at a span of datapoints at the beginning of the set to predict the next span, but instead of repositioning the window we consider all the datapoints up to a new point. It seems this will give more weight to data earlier in the dataset which may not be ideal.

3 Data Gathering

Stock market data is readily archived and available. We will be training our models on the Huge Stock Market Dataset from Kaggle. This dataset includes both Stocks and ETFs (Exchange Traded Fund, ie groups of stocks), but ETFs will not be used.

There were several factors in deciding to use this dataset.

- Amount of data. The dataset as a whole contains over 770MB of stock data in text file format.

- Single stocks. Many other datasets exist which are focused on stock averages like the S&P 500. Since our goal is to predict a specific breakout stock, this approach makes the most sense.

- Modern data. This dataset include data through 2017. While the last few years are still unavailable, stock market tendencies have not changed much.

- Relevant data. Many datasets exist to compare stock market data between countries. Ideally, we would be able to use this model to our own advantage, limiting us to US stocks.

Overall, compared to other datasets which were available, the Huge Stock Market Dataset appears to be the best solution for our goal. Once a working model is achieved, daily stock data can be obtained from Quandl for real-time training.

4 Dataset Description

The dataset is separated first by text files. Each text file represents a stock, where the file's name represents the stock's label. Within each text file, each line represents the stock's attributes for a specific day. These attributes include

- Date. The date for which the other attributes are calculated in yyyy-mm-dd format.

- Open. The stocks value at the day's open.

- High. The highest value reached in the day.

- Low. The lowest value reached in the day.

- Close. The stocks value at the day's close.
- Volume. The number of stocks which were traded that day.
- OpenInt. 'Open interest'. It is a complicated attribute involving contracts bought and sold. Most values are 0, and this value is not relevant to our goal.

An example of a line of data is as follows:
2006-03-03,24.349,24.706,24.33,24.449,2250685,0

5 Schedule to Finish

We have little under a month to build this model and write a report. Below is our calendar of dates and goals:

Date	**Goal**
Oct 26	Have datasets and environment ready to start experimenting
Nov 2	Pseudocode mapped out and some basic helpers written (Decide on model)
Nov 9	Model defined and ready for testing and peer review
Nov 16	Begin writing report
Nov 19	Report Due