Formatting instructions for NIPS 2017

Forest Kobayashi
Department of Mathematics
Harvey Mudd College
Claremont, CA 91711
fkobayash@hmc.edu

Jacky Lee
Department of Mathematics
Department of Computer Science
Harvey Mudd College
Claremont, CA 91711
jaclee@hmc.edu

Abstract

Stock data is very difficult to analyze using classical methods, in large part due to heavy involvement of humans in stock pricing. In this paper, we examine a new approach to analyzing time-series stock data, particularly in the context of grouping correlated stocks.

1 Introduction

Here is a rough table of contents for our paper, so that the reader has a rough idea of where we'll be going:

1. Motivation (challenges to modelling stock data)
2. Long-term goal for the model / ideal model architecture
3. Classification
4. Findings
5. Conclusion

2 Background

There are many reasons why predicting stock prices is difficult, but for our model today, we will focus on just two:

1. The behavior of a given stock is influenced by hundreds of hidden variables — e.g., the current state of geopolitics, governmental fiscal policies (for instance, the US Treasury interest rate), the performance of stocks in various other markets, etc. Hence, the price of a stock at any given moment is an emergent phenomenon, influenced by many small movements of markets around the world. Further, these connections themselves are in a constant state of flux — as the market evolves, technology improves, and policies are rewritten, the weight each of these variables has in influencing pricing will wax and wane.
2. With the exception of some forms of algo trading, most trading strategies are being created and implemented by humans. That is, often humans are the ones who perform analysis of market information, and make decisions based off the findings. But humans are not naturally predisposed to rigorous quantitative analysis, and hence markets do not always behave rationally. As an example, consider `Bitcoin`.[1] Thus, an accurate financial model will need to have some method of predicting human psychology.

[1] Citation: Gary Evans

Using modern machine learning techniques, it is possible to control for some effects of (2) by using previous stock data. However, less work has been done to model the effects of (1). So, for today's model, we will focus on methods of preprocessing time series stock data so as to control for some of the effects of (1).

3 The Model

3.1 Goals

Our ultimate goal is to be able to use past stock data to train our model. Then given some stock trend data for some stock, use our model to predict the upcoming trends and return a vector of decisions. An example of the output is shown below.

$$\mathbf{y} = \left\langle P_{buy}(\mathbf{x}), P_{sell}(\mathbf{x}), P_{short}(\mathbf{x}), P_{nothing}(\mathbf{x}) \right\rangle$$

Here, $P_{buy}(\mathbf{x})$ represents how confident the model is in recommending the user to buy the stock. Similarly, $P_{sell}(\mathbf{x})$ represents selling the stock, $P_{short}(\mathbf{x})$ represents shorting the stock, and $P_{nothing}(\mathbf{x})$ represents doing nothing. All these probabilities should sum to 1.

3.2 Prototype Architecture

Our large-scale goal for the model is as follows:

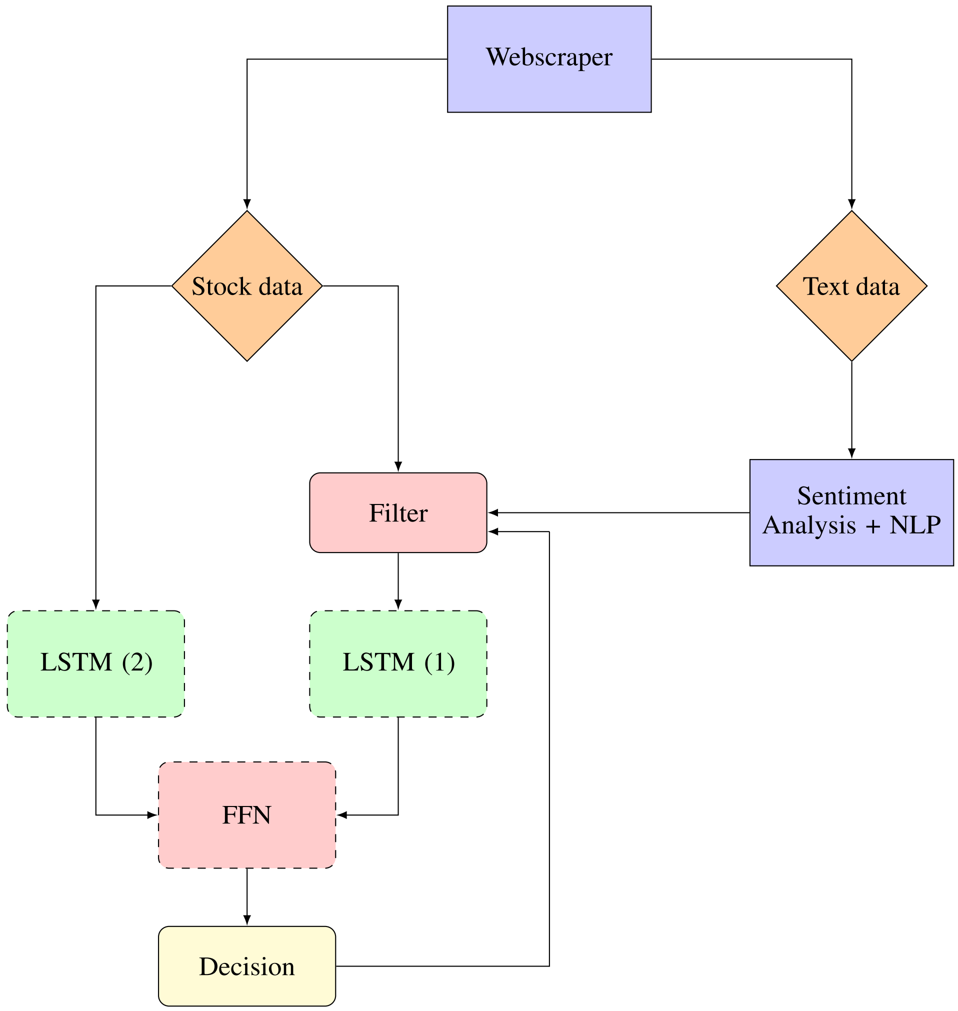

Figure 1: Model overview

4 Classification

Using LSTMs to predict stock pricing is, for the most part, a solved problem. Implementations and hyperparameters may differ, of course, but underneath, the structure of the model is largely the same. Hence, we decided to focus our work on augmenting the filtering stage shown above. We were particularly interested in finding new representations of market trends that we could feed into the network.

5 Methods

5.1 Data Scraping

We gathered intraday data for ten different stocks using the Alpha Vantage API (`https://github.com/RomelTorres/alpha_vantage`). The ten stocks were randomly selected from the stocks in the S&P500. A portion of the data for one stock is shown below.

```
                     1. open  2. high   3. low   4. close  5. volume
date
2018-05-08 09:30:00  1058.54  1058.54  1055.00  1055.1600    26407.0
2018-05-08 09:31:00  1054.99  1058.16  1054.99  1056.9248     6384.0
2018-05-08 09:32:00  1057.41  1058.95  1057.41  1058.5700     7160.0
```

```
2018-05-08 09:33:00  1058.30  1058.64  1056.93  1058.6400     6448.0
2018-05-08 09:34:00  1058.80  1060.00  1058.50  1059.9600     5548.0
```

For each stock, we computed the arithmetic mean of the high and low as a heuristic that summarized the price of the stock in that minute.

5.2 Polynomial Fitting on Windowed Data

We then analyzed this time series data by focusing in on windows of 20 minutes and fitting a cubic polynomial then extracting the constant, linear, and quadratic coefficients. Each consecutive window was only shifted by 2 minutes so consecutive windows would have significant overlap. Our reasoning was that this would make the coefficients be more similar and enable us to have smoother data.

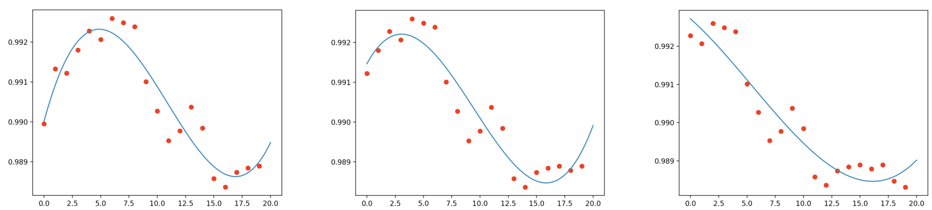

Figure 2: Sliding windows of data fitted with polynomials

5.3 Visualizing the Coefficients

We then plotted the coefficients of our polynomial fits in $\mathbb{R}^3$ to visualize our data.

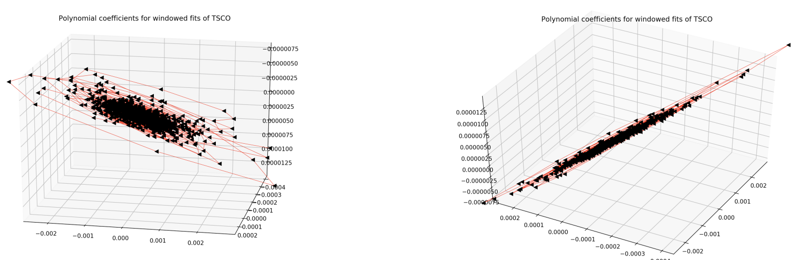

Figure 3: Visualization of the lower order coefficients in $\mathbb{R}^3$

The distribution of the coefficients in three dimensional space seems to be a multivariate Gaussian distribution. Furthermore, it seems to be very flat, appearing to be a line if viewed from the correct angle. The trajectory of the points also seems to be rotating around the center of the distribution.

5.4 Frequency Analysis