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

2nd Course in Mathematics of Big Data (MATH189 2018), Claremont, CA, USA.

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.

In order to reach this sort of conclusion, we want our model to compensate for the effects of (1) in a clever way. In traditional stock trading practices, there are a few guidelines that are used for similar determinations:

1. Someimtes industries follow trends together. For instance, if we see a price drop in one oil stock, we might expect to see drops in other oil stocks.
2. When interest rates go up, yield-bearing financial assests increase in value, while the value of equity stocks might decrease.
3. If an earnings report reveals that a stock overestimated its expected profits, the value of the stock will decrease.
4. In times of high uncertainty, money might be pulled out of equity assets and moved to yield-bearing financial assets. Hence, stock prices might go down.
5. Bad PR can result in a temporary dip in stock pricing.

These are external factors that cannot be determined purely from the stock data itself. Often, most of this information comes from news outlets, press releases, and other media-sharing forums. So, we need our model to perform the following:

1. Scrape real-time text data from news websites, financial websites, and (possibly) academic journals in the case that we're looking at something like a technology stock.
2. Perform some sort of Natural Language Processing and Sentiment Analysis to determine how this information might affect the market if / when it becomes widely publicized.
3. Somehow combine this information with time-series stock data to make a prediction of how the market might behave in the short and long term.
4. Make a trading decision based on this information.

We propose the following prototype model.

3.2 Prototype Architecture

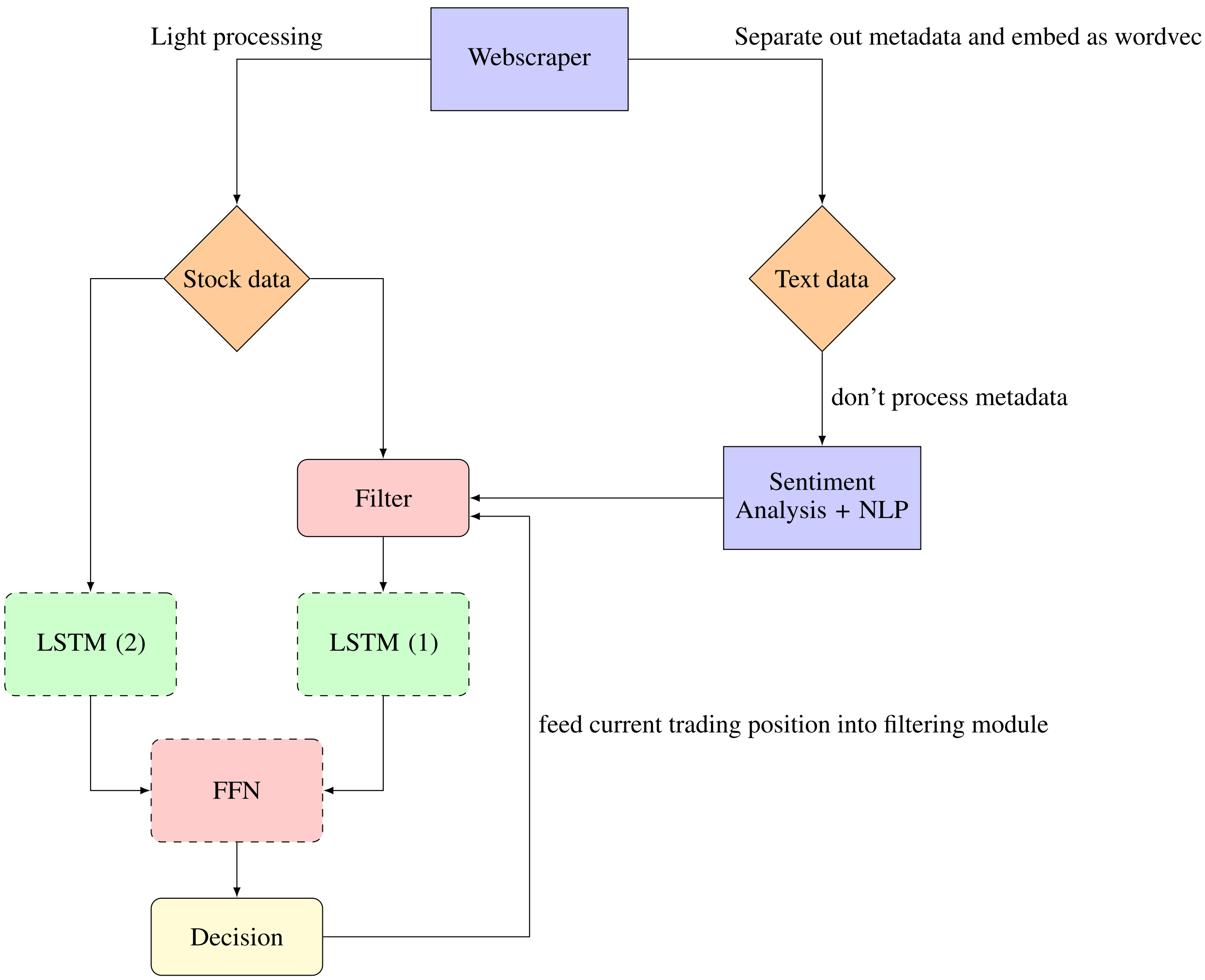

Figure 1: Model overview

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.

4.1 Polynomial Fitting on Windowed Data

4.2 Frequency Analysis

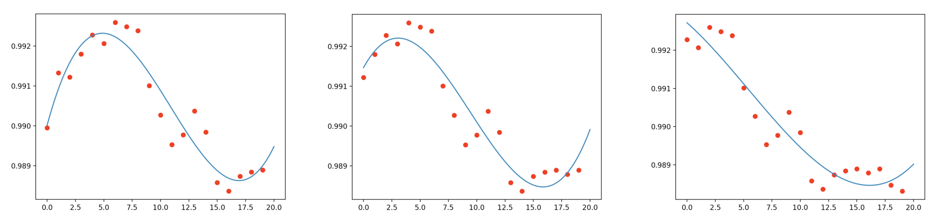

Figure 2: Sliding windows of data fitted with polynomials