Part 2 Dissertation Progress Report

Learning the Stock Market: Deep Learning and Sentiment Analysis-Based Stock Price Prediction

Ian Tai
ihbt2@cam.ac.uk

Supervisors: Dr. Sean Holden, Prof. Stephen Satchell
Director of Studies: Dr. Sean Holden
Overseers: Prof. Anuj Dawar, Dr. Timothy Griffin

Summary of Work Completed

Outline

1. Finished data collection for financial history data
2. Built LSTM system for price prediction
3. Conducted large-scale data collection for news headline data via web-scraping
4. Built Naive Bayes classifier for sentiment classification
5. Implemented semi-supervised learning for sentiment classification using K-Nearest-Neighbours and Support Vector Machine with radial basis function kernel
6. Built data processing system for end-to-end data collection to model training integration

Financial Data: Collected historical stock data of companies in the NASDAQ 100 index with a time interval of 5 minutes from various Bloomberg terminals around the University. This data included opening price, volume traded, bid price, bid volume, ask price, and ask volume.

LSTM: Built deep LSTM system which supports multi-layer configurations. The model currently correctly predicts the rise and fall of a single stock at around 52% accuracy after repeated training with normally-randomized initial weights. Since the system currently only utilizes historical prices from the same single stock, this accuracy is acceptable and should increase when our features become more sophisticated.

News Data: Experimented with different Twitter API libraries before finally settling for a web-scraper instead because of numerous advantages such as speed and usability. Collected around 1.4 million tweets from the 35 biggest news headlines in the world, spanning the past year. I randomly sampled a small subset (300 tweets) to manually classify sentiment.

Naive Bayes: Implemented a Naive Bayes classifier as the first steps towards a viable sentiment classification model. Tested on a small subset of the large twitter dataset, using feature generation techniques such as Bag of Words and Doc2Vec. The results were unsatisfactory, which led to the development of more sophisticated models.

Semi-Supervised Learning: Used popular Machine Learning library scikit-learn to implement semi-supervised learning for a larger subset of the twitter data. Results were also unsatisfactory, which led me to believe that dataset may be too noisy, or the classification approach incorrect. I will spend more time on this portion in the near future.

Data Processing: Built a simple data processing system that wrote inputs and outputs to temporary files to convert to the corresponding format for different libraries and self-built modules.

Difficulties

There has only been one main unexpected difficulty:

None of the aforementioned sentiment classifiers performed to a sufficient standard, with the best result (RBF kernel of SVM) having only a 57% accuracy on the validation set, after much experimentation with different classifiers and knowledge representations.

I will need to fix this by either changing my dataset, or use a specific subset that doesn't include noisy data, or change my representation of the input data within the model.

Schedule

Because of the difficulty outlined above, the sentiment analysis portion of my project, which was supposed to be completed before Christmas, is yet to be fully completed. However, the rest of the project is on track. This setback puts me around 2 weeks behind.

I will try to fix this by allocating time to the sentiment analysis problem in the next stage of work, which initially included only dissertation writing.