CS 260 Machine Learning Algorithms

Term Project Proposal

Author: Zhao Yang

UID: 904 169 203

Date: November 5th, 2015

**Motivation**

The motivation behind this project is to predict short-term stock price movement using information contained in a target company’s limit order book. A limit order book contains rich information about the market dynamics of a particular stock, which can generate numerous features that are potentially indicative of the movement of stock price in a short time frame. Such predictive model can be useful in the domain of high frequency trading (HFT). In fact, it is now common for investment banks to employ automated proprietary HFT systems in stock trading. This project aims to explore the predictive power of several models we learned in the class on the limit order data and search for the set of optimal features for the prediction of stock price movement.

**Background**

A limit order book lists all buying and selling offers at specific amount of shares and specific prices. Both ask (selling) and bid (buying) orders are sorted according to the price-time priority. Ask orders are first sorted in ascending order according to price and then orders with the same price are sorted according to the entry time. Similarly, bid orders are sorted in descending order with respect to price and then sorted by time. Once the top ask and the top bid orders cross, meaning the price of the two matches, an order is executed at the matching price.

Limit order books illustrate order flow imbalance, which illustrate the supply and demand for the stock. Because the short-term price change of the stock can be dominated by the supply and demand condition, it is expected that features which reflect the order flow imbalance can effectively predict price movement.

**Proposed Work**

In this project, I will use the limit order data of one stock across different time points during the day and across different trading days to predict the price movement during a short time interval (10 to 60 seconds) after the snapshot of the limit order data.

After acquiring the data, I will generate the following set of features for classification: Bid-Ask Spread, Bid-Ask Volume Difference, Mid-Price, Order Flow Imbalance, Execution Price, Average Price, and Average Volume. To have a fast start, I will first experiment with using only the Bid-Ask Spread and Bid-Ask Volume Difference features. After verifying the correctness of the training implementation and get feedback on using these two features, I will use cross validation to select the optimal set of features. New features may be added during the implementation of the project.

I plan to apply Logistic Regression with regularization, SVM, Bootstrap Aggregating, Decision Tree and K-Nearest Neighbor algorithms to this classification problem. The labels of the data will be the trend of the stock price after a specified time interval (between 10 - 60 seconds), {Up, Down}. Accuracy evaluation method will be the same as we applied in homework assignments.

**Timeline**

I have 5 weeks after the submission of the proposal. By the end of the first week, I should have gathered and processed data and generated some of the features from the data. By the end of the second week, I should have tested some models on the processed data and got feedback. Week 3 and 4 are for feature optimization, data reprocessing and model debugging. I anticipate that I will need to correct some bugs in producing features from the data and also generate some new features. A considerable amount of time will be spent on tuning the model. Each model has various hyper-parameters to try out. Graphs and charts will also be completed during week 4. Week 5 is for writing the report and preparing for the poster presentation.

**Deliverables / Evaluation**

The deliverables should include charts that report the accuracy rate for each model I applied, the program I used to produce the results in the charts, and a detailed explanation of the procedures and methods in the project. In the written report and the poster, I will elaborate on the research process and the techniques used in the project.

**Data**

The data consists of limit orders for a stock at various time points, with up to 30 price levels for each time point. The data will include the current stock price and the stock price shortly after the limit order data is captured. The limit orders contain all the information necessary to general the features, and the stock prices serve as labels.

Currently I have access to live level-2 data to NASDAQ stocks. The difficulty remains in automating the preservation of those live data. The NASDAQ software provides a manual channel to export limit order data to Excel, but does not provide an API. In fact, historical level-2 data is quite expensive to buy from the exchange or brokerage, usually costing thousands of dollars. The worst case scenario is that I manually export 100 data points per day for three days across the weekends.

Bloomberg Database contains comprehensive financial data of the market, and is accessible at the lab in Anderson Management School. Anderson students have access to the system, and the data is permitted for use by UCLA students for academic purposes. I am looking for a friend or professor to help collect data from that system.

**Libraries**

I plan to use the standard libraries in MATLAB as well as its math toolbox. After looking through its models, MATLAB should already contain all the models I want to apply. For data processing, I may use resources outside of MATLAB such as Python.

**Team**

I am working on this project by myself.

**Prior Discussion**

I discussed with Nikolaos Karianakis during discussion sessions for the past 3 weeks. I also stopped at his office Wednesday last week. I went to Professor Talwalkar’s office hour during the 3rd week to discuss with him my potential projects.