Stat 222: Finance Project

1 Overview

A central goal for this project is to give you exposure to machine learning in the context of financial data. In the high-frequency era of trading, orders of stocks can be executed under a millsecond. The information about the thousands of orders is captured by the *limit order book* (LOB). In this project, we will explore the LOB data and gain insight on stock price movements at the millsecond scale. Your main task is to predict the stock price movements using LOB data via various machine learning techniques.

On the other hand, the project will provide you with the opportunity to practice reading and writing about applied statistical data analysis. For every data science project, it is crucial to understand the background of the problem and how the data is generated. This is often referred as *domain knowledge*. Domain knowledge gives rise to important intuition for further processing of the data, such as feature engineering, and more importantly, for asking the right questions. We will discuss more about stock orders and the LOB in the lecture, but I expect you to do some outside readings about finance or any relevant information.

To get started, please read the following paper, on which the project is based:

Alec N Kercheval and Yuan Zhang. Modelling high-frequency limit order book dynamics with support vector machines. Quantitative Finance, 15(8):13151329, 2015. [1]

You may also find this blog post [1] helpful.

2 Data Description

For this project, you will obtain two datasets:

1. LOB data: AAPL 05222012 0930 1300 LOB 2.csv

   This dataset contains the prices and volumes of AAPL from 9:30 to 12:28 on 2012/05/22 for 10 levels of bid and ask for each time-stamped event.

2. message book data: AAPL 05222012 0930 1300 message.csv

   This dataset contains information about each order from 9:30 to 13:00 on 2012/05/22.

[1] http://eugenezhulenev.com/blog/2014/11/14/stock-price-prediction-with-big-data-and-machine-learning/

If you are feeling ambitious, feel free to use the message book data, but you are not required to. In this project, we will be using only the data from 9:30 to 12:00.

3 Your Assignment

This is a team project. You should split up the tasks among your group members. You are required to do everything in Python. You might find that the package *scikit-learn* very useful in this project. While the idea of the project seems very straightforward, the details are quite involved, so start as early as possible.

1. Data preprocessing

   Your first task is to transform the dataset(s) into a data frame convenient for further analysis. You may want to look at Table 2 in Kercheval and Zhang's paper to get an idea what features might be useful for the prediction. You do not have to use all the features suggested in the paper. You are required to study both midprice movements and bid-ask spread crossings. For simplicity, time is measured in terms of the number of time-stamped trade events. You need to decide a suitable time horizon: if the time horizon is too small, the majority of the price movements will be stationary; if the time horizaon is too big, the prediction becomes harder.

   After the preprocessing, split the dataset into two parts: 9:30 to 11:00 and 11:00 to 12:00. For the rest of the project, except for the trading strategies implementation portion, use only the 9:30 to 11:00 data.

2. Model Fitting

   At this point, you should split your dataset into three parts: training set (50%), validation set (25%), and test set (25%). Your second task is to fit the various machine learning models to the training set. The minimum requirement is to fit support vector machines (SVM) with the radial basis function (RBF) kernels and random forests, but you are strongly encouraged to try other machine learning techniques such as gradient boosting. To tune the parameters of each machine learning algorithm, use the validation set; that is, choose the parameters that minimizes certain error criterion.

3. Model assessment

   Your third task is to test your models with the test set. What are the performances of each of your models? Be sure not to simply use the percentage accuracy as your sole performance measure. Use at least recall, precision, and F_1-measure to assess model performance[2].

[2]Refer to Kercheval and Zhang's paper and/or see Wikipedia for a description of these performance measures: https://en.wikipedia.org/wiki/Precision_and_recall

4. Interpretations

   One problem with more sophisticated machine learning algorithms is that interpretability is sacrificed in exchange for higher prediction accuracy. So far in this project, we have considered only machine learning algorithms with a strong black-box flavor. Your final task is to try coming up simple summary statistics of the LOB that are useful for predicting the price movements. These summary statistics can be taken from the feature sets in Kercheval and Zhang's paper, or new statistics you come up with. For example, consider the ratio of the total volume of top five ask-side orders and the total volume of top five bid-side orders. If this ratio is very big or very small, there is a strong indication of the imbalance of supply and demand, which might cause price movements. This final task is intended more open-ended and requires more intuition. After coming up with a couple summary statistics, try fitting a logistic regression and compare your results with those obtained from the machine learning approaches you tried earlier in this project. Can you interpret your new model?

5. Trading strategies implementation Come up with simple trading strategies based on your machine learning models. Implement the strategies on the 11:00-12:00 data and compute the total profit. To simplify the problem, make the following assumptions:

   (a) There is no transaction cost.

   (b) You can place only market orders, and assume that they can be executed immediately at the best bid/ask.

   (c) Your position can only be long/short at most one share.

   Since none of the assumptions are realistic, in your report, explain how you would modify your trading strategies if the assumptions are violated.

Timeline and logistics

Here is the tentative schedule:

Monday	Wednesday
	(2/24) Financial data
(2/29) Start Finance Project	(3/2) Machine learning
(3/7)	(3/9) Git
(3/14)	(3/16)
Spring break	*Spring break*
(3/28) Finance report	

4 Report Details

Your report should include at least six sections: an introduction, five body sections (each of which covers an aspect discussed in the "Your assignment" section), and a conclusion. Make sure to compare the predictions and model performances for midprice movements and bid-ask spread crossings. The report has a 9-page limit.

References

[1] Alec N Kercheval and Yuan Zhang. Modelling high-frequency limit order book dynamics with support vector machines. *Quantitative Finance*, 15(8):1315–1329, 2015. http://www.math.fsu.edu/~kercheva/papers/multi-svm.pdf.