Final Project Proposal

Qi Zhao, Yuandong Chen, Yunhang Zhu

1. Problem Statement

This project is a kaggle contest. Please reference to https://www.kaggle.com/c/two-sigma-financial-news.

The problem is mainly about how to use the current news data and stock market data to predict the trend of the stock price.

We will predict a signed confidence value, $\hat{y}_{ti} \in [-1,1]$, which is multiplied by the market-adjusted return of a given assetCode over a ten day window. If you expect a stock to have a large positive return--compared to the broad market--over the next ten days, you might assign it a large, positive confidenceValue (near 1.0). If you expect a stock to have a negative return, you might assign it a large, negative confidenceValue (near -1.0). If unsure, you might assign it a value near zero.

For each day in the evaluation time period, we calculate:

$$x_t = \sum_i \hat{y_{ti}} r_{ti} u_{ti},$$

where rti is the 10-day market-adjusted leading return for day t for instrument i, and uti is a 0/1 universe variable (see the data description for details) that controls whether a particular asset is included in scoring on a particular day.

The evaluation score is then calculated as the mean divided by the standard deviation of your daily x1 values:

$$\text{score} = \frac{\bar{x}_t}{\sigma(x_t)}.$$

If the standard deviation of predictions is 0, the score is defined as 0.

2. Related Work

1. The final project of NLP class(CS224N) of Stanford University.
   Author: Kari Lee and Ryan Timmons
   Title: Prediction the Stock Market with News Articles
   Content: Their trading system to act on these predictions outperformed a baseline strategy of simply holding on to equal amounts of the stocks in question for the test time period.

2. Gidofalvi trained a naive Bayesian text classifier by scoring news articles using a function combining the change in price with the β-value, which measures the volatility of the stock. The articles were given one of three labels based on the stock's movement compared to its expected movement. Using this method, the predictive power of the classifier was limited, but there was a strong correlation between the news article and the stock price behavior within a 20 minute window around the news article's release time.

3. Fung, et al designed a system which used a t-test based split-and-merge piecewise linear approximation algorithm to filter out the noise of the movement over the stock time series. Features were the words in a document weighted using term frequency inverse document frequency, and the optimization problem was solved using a Support Vector Machine. The resulting system was most stable and appropriate for predictions within 3-5 days.

3. Timeline

20 NOV ~ 21 NOV Clean data and data pre-processing
22 NOV ~ 23 NOV Plotting and analyze feature
24 NOV ~ 28 NOV Apply data to the model

29 NOV ~ 30 NOV Group Discussion, Tuning parameter and validation.
1 DEC ~ 2 DEC Start working on report
2 DEC ~ 3 DEC Preparing for the first draft.

4. Data collection and/or dataset/corpus

The training and test data are from kaggle.com.

We will be predicting future stock price returns based on two sources of data:

1. Market data (2007 to present) provided by Intrinio - contains financial market information such as opening price, closing price, trading volume, calculated returns, etc.
2. News data (2007 to present) Source: Thomson Reuters - contains information about news articles/alerts published about assets, such as article details, sentiment, and other commentary.

5. Data Pre-processing:

1. Find missing value percentages and use only valid columns that are not nullable.
2. Find all entries whose assetName attribute is unknown and group these entries by their assetCode.
3. Use heat map to have an intuition of relationship between columns.
4. Use pie chart to evaluate trading volumn of different assetCode.
5. Find outliers on returnsOpenNextMktres10 vs volumn(open/close) diagram.

6. Approach

We are working on LSTM-RNN network which is suitable for sequential inputs, Linear Regression, Logistic Regression, SVM, Adaboost, Recommendation System and try to pick one or two to compare the performance of these models. Finally, we may or may not use the bagging to optimize the trained model.

7. Evaluation:

1. The evaluation criteria is the score calculated above.The higher the score is, the model is better.
2. Considering the recall, precision, f1-score as criteria for the evaluation.
3. Using different ratio of train and test data and compare the results.
4. Analyze the bias and variance of the model.
5. Calculate AUC to evaluate our model.

8. What are you trying to understand better to gain insights on this problem?

We observe the data first and find some features like "sentiment Class" contribute much to the trend of the stock market performance. For example, if the value of "sentiment class" is 1, means positive news, the stock price of corresponding company may have a better performance. But if the value is -1, means negative news, the stock price may have a worse performance.

We also go through some old events in which the news influence the stock market a lot and try to find some intuition to help us understand how the news data work in this problem.

9. Reference

1. Gidofalvi, Gyozo. *Using News Articles to Predict Stock Price Movements*. Department of Computer Science and Engineering, University of California, San Diego. 2001.
2. Fung, Gabriel, et. al. *The Predicting Power of Textual Information on Financial Markets*. IEEE Intelligent Informatics Bulletin. Vol. 5. No. 1. June 2005.
3. The final project of NLP class(CS224N) of Stanford University. https://nlp.stanford.edu/courses/cs224n/2007/fp/timmonsr-kylee84.pdf