Predicting Stock Price Changes Using Past Prices and News Articles

Ang Li (al171@stanford.edu)

Objective

The goal of the project is to predict price changes in the future for a given stock. Information that might be leveraged to make these predictions more accurate include prices from previous days, and financial news articles related to the company of interest.

To further reduce volatility in the results, predictions are only made on articles related to the company of interest. The model outputs classification (1 or 0) of the sign of the next day’s price change. The model achieved 59% accuracy and an expected gain over 100 days.

Naive Bayes for News Articles

• Naive Bayes using the multinomial event model with Laplace smoothing was trained on 400 days, and tested on 200 days of data. Output of the model is the probability of a positive relative price change.

• Numbers, percentages, money amounts were substituted with fixed tokens that consolidate these words together.

Features for NN Model

• 41 features in total to predict today’s relative price change
  - Relative price changes for the same stock from last 20 trading days.
  - News predictors (between 0 and 1) for the last 20 trading days.
  - News predictor (between 0 and 1) for today.

Neural Network

• The neural network was structured to have 1 hidden layer with 20 nodes.

• The input layer has 41 nodes, and the output layer has 1 node.

• The activation functions used in the hidden and output layer are sigmoids.

• Mini-batch gradient descend (with batch size 20) with forward and backward propagation was applied to learn parameters in the neural network.

• L2 Regularization with lambda = 0.0001 was also applied to reduce overfitting to training data.

Results

• One of the most important evaluation metric in this project is percentage gain if trading according to the predictions. E.g., a value of 0.01% daily gain means the expected gain over 100 days is 1%.

• Naive Bayes news article positiveness predictor achieved an accuracy of 55.5% on the test set.

• Naive Bayes predictions could be improved to increase accuracy, which would in turn increase with the prediction accuracy of subsequent models. Ways to improve Naive Bayes predictions include removing insignificant words, reducing stemmed words into base forms, and so on.

• Neural networks as structured in previous section with all features (including news) produced the best accuracy (59%) and daily gain (0.0423%).

Discussion & Future Work

• Overall, there was a lot of overfitting to training data despite the regularization techniques used, due to the fact that the stock market is way too complex to be captured in merely price changes and news.

• Higher quality news data would have helped with prediction accuracy. The news data pulled from xignite.com appear to have mediocre quality.

• Neural Bayes predictions could be improved to increase accuracy, which would in turn increase with the prediction accuracy of subsequent models. Ways to improve Naive Bayes predictions include removing insignificant words, reducing stemmed words into base forms, and so on.

• Neural networks with other activation functions and different architecture (e.g. using ReLU or a different number of hidden units) could be explored.

• Overall, there was a lot of overfitting to training data despite the regularization techniques used, due to the fact that the stock market is way too complex to be captured in merely price changes and news.

• Higher quality news data would have helped with prediction accuracy. The news data pulled from xignite.com appear to have mediocre quality.

• Neural Bayes predictions could be improved to increase accuracy, which would in turn increase with the prediction accuracy of subsequent models. Ways to improve Naive Bayes predictions include removing insignificant words, reducing stemmed words into base forms, and so on.

• Neural networks with other activation functions and different architecture (e.g. using ReLU or a different number of hidden units) could be explored.

• Augment the neural network by adding more input features, such as news in other industries, political influences, competitor trends, and many more, to capture the full scale of complexities of a stock market.