Proposal: Numerai Trading Predictions

Zhiliang Gong
zhiliang.gong@gmail.com

Domain background

More and more hedge funds are using machine learning to predict the price of financial products, such as stocks, bonds, options, and futures. Top players in this field includes D. E. Shaw, Two Sigma, Renaissance Technologies, etc [1]. These quantitative trading firms are mostly playing a zero-sum game, as gains of one firm often come from losses of other firms. The pressure to stay on top of other firms is a paramount struggle for any quantitative hedge fund.

Problem statement

The difficulty to compete effectively against other players in quantitative trading stems from the them sharing the same set of information, such as company reports, prices over time, trading volumes, etc. This set of information could enable algorithms of one firm to "guess" the "thinking" of algorithms of other firms as the algorithms evolve with fresh inputs from the open market. Here comes Numerai, a startup in quantitative trading, working to crowdsource predictions from high performance individuals [2], to compete against the in-house designed algorithms of the big quantitative funds. Interestingly, they release a large pre-processed and encrypted financial dataset each week, which hides the meaning of each feature in the dataset, so that data scientists don't have to guess what the other data scientists might think about a given feature [3, 4]. In this project, I propose to prototype and train a machine learning model, improve the prototype model over fresh data, and then use the final model to make predictions for a validation set.

Datasets and inputs

The datasets will be one of the weekly datasets from Numerai. Each weekly dataset contains a training set with around 500k entries, a test and validation set around 200k entries, and a live set with around 100k entries without specifying the target. The target is binary, 0 or 1.

Solution statement

The solution to the problem is a multitude of machine learning methods in supervised and unsupervised learning. I anticipate to use clustering methods for dividing the datasets into feature sets, and try different combinations of regression and classification methods to refine predictions.

Benchmark model

The benchmark for financial predictors are chance. In the current situation, it's the ability to predict the target at an accuracy higher than 0.5 over extended period of time. The higher the better. Keep in mind that even a slight advantage over chance can bring enormous amounts of profits.

Evaluation metrics

The metric to evaluate the model is the accuracy score, i.e., the percentage of the correct predictions for data from the validation set.

Project design

As the ability to consistently predict data overtime is critical, I plan to slice the training set into 4 parts, with the first part for training and prototyping a model, and the remaining 3 parts for improving the model, without changing the structure of the model. The project will be carried out in five stages:

1. Prototype and train a model based on 1/4 of the training set available
2. Improve the model with the rest of the 3/4 added to the training set
3. Test the model with the test set
4. If testing result is bad, go back to prototyping the model
5. Validate the model using the validation set

References

1. http://www.streetofwalls.com/finance-training-courses/quantitative-hedge-fund-training/quant-firms/
2. https://techcrunch.com/2016/12/12/numer-ai-is-a-crowdsourced-hedge-fund-for-machine-learning-experts/
3. https://numer.ai/help
4. https://www.kaggle.com/numerai/encrypted-stock-market-data-from-numerai