Stock Price Prediction with Deep Learning and Feature Engineering

Yuhui Huang, Siyang Jing[*], Jiacheng Tian, Jiyu Xu

October 2, 2018

Problem and its relevance to investments Stock price predictability is one of the most important concerns for investors. Previous studies have mostly focused on stock price prediction at a low frequency. As machine learning techniques evolve in the area of finance, there has been growing interest in high frequency algorithmic trading [1]. With the increasing availability of high frequency trading data, better understanding of trading pattern will likely to lead to better prediction of stock price.

The data Quantopian(www.quantopian.com) has ongoing minute-level US equity pricing and volume data since January 1, 2002. We plan to use the data of 50 largest stocks as ranked by market capitalization. We conjecture that past stock returns are related with not only its own future returns but also the future returns of other stocks, and thus use 500 dimensional lagged stock returns (50 stocks and 10 lagged returns) as raw level input data. This large input makes deep learning a particularly suitable choice.

The model Following the practice in the research of Chong et al.[2], our model takes the form of $r_{t+1} = f \circ \phi(R_t) + \gamma$, $\gamma \sim \mathcal{N}(0, \beta)$. We assume the return of one hour prior to the current time point has influence on the return of the stock in the next minute, so t is in unit of minutes. As mentioned above, $R_t = [r_{1,t}, ..., r_{1,t-60}, ..., r_{50,t}, ..., r_{50,t-60}]^T$ is the 500 dimensional raw level input vector. ϕ transforms the data to features, or representations, and will be learned by principal component analysis, autoencoder, restricted Boltzmann machine, and other unsupervised learning techniques. f is the predictor function, and will be learned using recurrent neural network (RNN), which seems to be the state of the art deep learning method for financial time series data analysis [3].

Secret sauce We will try to combine other statistical techniques such as univariate autoregressive model (AR(10)) with our proposed model. In addition, online learning techniques, model retraining, and data assimilation will be added to incorporate the latest data to improve the performance.

How the model will be evaluated We evaluate the prediction model's performance on the test set with four measurements: normalized mean squared error, mean absolute error, root mean squared error, and mutual information. The training and test set are divided such that data from 2002 to 2017 is training set and data since 2017 is test set. We also employ bootstrap analysis on assessment of the accuracy of the estimator by random resampling with replacement from an original dataset.

Anticipated challenges We might have assumed incorrectly that the long term change of stock price does not significantly affect short term patterns. For example, the minute-level price fluctuation of equities during the Great Recession might be different from the pattern exhibited in normal years. This could contaminate the training data or lead to severe bias.

The promise With increasing amount of investors entering into the market, satisfactory stock price prediction will bring higher return for investors and further improve in capital scale.

[*] Corresponding author.

References

[1] M. Kearns and Y. Nevmyvaka, "Machine learning for market microstructure and high frequency trading," 2013.

[2] E. Chong, C. Han, and F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies," *Expert Systems with Applications*, vol. 83, pp. 187 – 205, 2017.

[3] M. Abe and H. Nakayama, "Deep Learning for Forecasting Stock Returns in the Cross-Section," *ArXiv e-prints*, Jan. 2018.