Project Midterm Report
Dior Xiong, Jee Su Byun, Shantanu Mandal

Stock Prices Forecast with Stacked Autoencoders and LSTM

1. Data Collection

We initially selected 11 stock indices to test the prediction ability of the proposed model. Those indices include:

- Korean: KOSPI (Korea Composite Stock Price Index)
- Vietnam: VN30 (VN30 Equal Weight Index)
- Bangladesh: DS30 (Dhaka Stock Exchange 30)
- Hong Kong: SHI (The Hang Seng Index)
- London: FTSE 100 Index (The Financial Times Stock Exchange 100 Index)
- USA: DJIA (Dow Jones Industrial Average)
- China: SSEC (Shanghai Stock Exchange Composite)
- India: BSESN (S&P Bombay Stock Exchange Sensitive Index)
- Switzerland: SMI (Swiss Market Index)
- Brazil: IBOVESPA

Beside, we will also choose technology indices for extra model testing:

- Korean: KOSDAQ (Korean Securities Dealers Automated Quotations)
- USA: NASDAQ (Nasdaq Composite)
- Indian: Nifty IT (NIFTYIT)
- China: CQQQ (Invesco China Technology ETF)
- London: FTSE AIM All-Share - Technology

From www.investing.com, we collected seven data attributes for each index:

- Date, Price (OHLC: Open, High, Low, Close), Trade Volume, % Change daily

We also considered collecting data from the following non-free-of-charge sources:

- Quantopian: https://www.quantopian.com
- Quandle: https://quandl.com
- NASDAQ: https://www.nasdaq.com/market-activity/quotes/historical
- Eoddata: http://www.eoddata.com/download.aspx
- TAMU May Business School: https://mays.tamu.edu/innovation-research-center/data-sets-in-mays-business-school/
- First Trade Data: http://firstratedata.com/

We wish to narrow down the industry sector such as oil, gold and technologies; we are still searching for the corresponding index data which reflect them individually. We also wish to find the hourly data for shorter term pattern analyses in the following weeks.

Below are the overview of techniques we will be applying to our model

2. Wavelet Transform

First of all, we apply wavelet transform for data denoising. Wavelet transform has the ability to decompose complex information and patterns into elementary forms. It is applied for data denoising in this project because of its ability to handle non-stationary financial time series data, which is useful in handling highly irregular financial time series. We apply the Haar function as the wavelet basis function because it can not only decompose the financial time series into time and frequency domain but also reduce the processing time significantly. The wavelet transform with the Haar function as a basis has a time complexity of O(n) with n being the size of the time series.

3. Stacked Autoencoder

After denoising, autoencoders will be applied for layer-wise training for the OHLC variables and technical indicators. Single layer AE is a three-layer neural network, the first layer being the input layer and the third layer being the reconstruction layer, respectively. The second layer is the hidden layer, designed to generate the deep feature for this single layer AE. The aim of training the single layer AE is to minimize the error between the input vector and the reconstruction vector. The first step of the forward propagation of single layer AE is mapping the input vector to the hidden layer, while the second step is to reconstruct the input vector by mapping the hidden vector to the reconstruction layer.

The activate function can have many alternatives such as sigmoid function, rectified linear unit (ReLU) and hyperbolic tangent. In this project, we set this to be a sigmoid function. The model learns a hidden feature from input by reconstructing it on the output layer. Stacked autoencoders is constructed by stacking a sequence of single-layer AEs layer by layer. The single-layer autoencoder maps the input daily variables into the first hidden vector. After training the first single-layer autoencoder, the hidden layer is reserved as the input layer of the second single-layer autoencoder. Therefore, the input layer of the subsequent AE is the hidden layer of the previous AE.

For training, the gradient descent algorithm will be used for solving the optimization problem in SAEs, and completing parameter optimization. Each layer is trained using the same gradient descent algorithm as a single-layer AE by solving the optimization function and feeds the hidden vector into the subsequent AE. The weights and bias of the reconstruction layer after finishing training each single-layer AE will be cast away. Depth plays an important role in SAE because it determines qualities like invariance and abstraction of the extracted feature. In this project, the depth of the SAE will be set to 5.

4. LSTM

After SAEs, we plan to use the time series data as sequence to train the LSTM. LSTM networks consists of an input layer, several hidden layers and an output layer. The number of input layers is the same as the number of features. We choose LSTM because it doesn't have the problem of vanishing gradients. LSTM is a gated version of recurrent neural network.

The main drawback of recurrent neural network is that it can not use previously seen data due to vanishing gradient problem. It is commonly known that in stock data pattern is the key factor to move the market in technical analysis. Hand-coded models such as clustering or feature extraction are sometimes difficult in time series data such as stock value. Therefore, the main purpose of the LSTM is to capture pattern throughout the time series data. We plan to show different performance benchmark comparison between LSTM and other deep neural networks. With our intuition we are confident that LSTM will work better that other feed forward network.

5. Future Plan:

The future objective of our projects can be divided into many folds.

1. We want to show different deep neural network model comparison on stock analysis and find out the best fitted model for our case.
2. Different benchmark comparison with market alpha.
3. Make a framework that is easily extendable with different prediction model.

6. Conclusion:

In this report, we discussed two different types of models, stacked auto encoder and LSTM. We extract features and build our prediction model via different neural network topologies. The main goal of this project is to find out an optimized model through different analysis and cross validation of the data. We'll show more detailed results on the different stock benchmarks.

Reference

1) Bao, W., Yue, J., & Rao, Y. (2017). A deep learning framework for financial time series using stacked autoencoders and long-short term memory. Plos One, 12(7). doi: 10.1371/journal.pone.0180944
2) Ramsey JB. The contribution of wavelets to the analysis of economic and financial data. Philosophical Transactions of the Royal Society B Biological Sciences. 1999; 357(357):2593–606.
3) Popoola A, Ahmad K, editors. Testing the Suitability of Wavelet Preprocessing for TSK Fuzzy Models. IEEE International Conference on Fuzzy Systems; 2006.
4) Abramovich F, Besbeas P, Sapatinas T. Empirical Bayes approach to block wavelet function estimation. Computational Statistics & Data Analysis. 2002; 39(4):435–51.
5) Hsieh TJ, Hsiao HF, Yeh WC. Forecasting stock markets using wavelet transforms and recurrent neural networks: An integrated system based on artificial bee colony algorithm. Applied Soft Computing. 2011; 11(2):2510–25.