Machine Learning Engineer Nanodegree

Capstone Proposal

Takayoshi Nishida
4 November, 2017

Proposal

Domain Background

Predicting the stock price has been researched for long. Now many people try to predict with the machine learning algorithms, but there is not a single answer for this and it is still challenging problem.

It is also known that every country's stock market influences each other. So putting the another country's stock market price and predicting the stock price index worth challenging.

Problem Statement

The goal of this project is to predict the change rate of the close price of Nikkei 225 index compared to the previous day. Nikkei 225 is a stock market index for the Tokyo Stock Exchange in Japan.

My hypothesis is that the Nikkei 225 has a strong correlation with the close price of US stock prices (NASDAQ index) and JPY/USD foreign exchange rate of the previous day.

The target variable is the Nikkei 225's relative change rate from the previous day. For example, in case the Nikkei 225 index close price is "21450.04" and it was "21374.66" at the previous day, the relative change rate is ("21450.04" / "21374.66") ≒ 1.00352. Then, it is possible to know the error between the predicted rate and the actual rate.

Datasets and Inputs

I am going to prepare the input datasets from following 3 sources.

Those data sets has a long history data. Nikkei 225 has the data from 1950, but the NASDAQ starts from 2003 from this source. So I plan to use the data from 2003 to 2017.

1. Nikkei 225

The data starts from January 1950 to current date. This data can be obtained at Quandl.

- https://www.quandl.com/data/NIKKEI/INDEX-Nikkei-Index

The input feature data is the change from the previous day.

2. NASDAQ Index

The data starts from January 2003 to current date. This data can be obtained at Quandl.

- https://www.quandl.com/data/NASDAQOMX/COMP-NASDAQ-Composite-COMP

The input feature data is the change from the previous day of the NASDAQ index.

3. Currency Exchange - JPY/USD

The data starts from March 1991 to current date. This data can be obtained at Quandl.

- https://www.quandl.com/data/CURRFX/USDJPY-Currency-Exchange-Rates-USD-vs-JPY

The input feature data is the change from the previous day of the JPY/USD exchange rate.

Solution Statement

The solution to this problem is to apply LSTM (Long short-term memory) to predict the Nikkei 225 index of the next day. LSTM is a one kind of the RNN (Recurrent neural network).
I chose LSTM for this problem because:

- this is a type of time-series problem.
- LSTM has the advantages that can remember the past values better.

I will input the data of input_size: N, which means the data within N days will be input as a feature data. So inputting the N days data of the change rate for Nikkei 225, NASDAQ and JPY/USD exchange rate, I am going to predict the change rate of Nikkei 225.

At first, I am thinking to start with input_size N = 10 days. After that, I need to investigate what parameter would get the best score.

Benchmark Model

I'd like to use the MAPE (Mean absolute percentage error) to evaluate the prediction.
Benchmark model is MAPE calculated using using the [DummyClassifier (http://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html)](http://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html).
Comparing with DummyClassifier, we can figure out how the prediction model works well.

Evaluation Metrics

My prediction model calculates the change rate, so I think MAPE (Mean absolute percentage error) would be good choice to evaluate the model.

Project Design

Data preprocessing

The first step is collecting the data. As discussed above, I will obtain the data from Quandl.
The file is CSV and the data is numeric value.

Next, I am going to use the change rate for the input data, so it is needed to calculate the change rate. It is calculated by the specific day's value

Splitting the data

I will split the data. The data within 2003 to 2015 is for training dataset, the data within 2016 is for validation dataset and the data in 2017 (Until current date) for the test dataset.

Since this data is time series data, the data are not independent each other.
So It should be avoided to extract random 20% data for the test.
Instead, I am going to split the data by year. 2003 to 2015 year for training dataset, 2016 year for the validation and 2017 year for the test.

Model

I am going to build the model with LSTM (Long short-term memory).

I will start with the simple model first, which is input_size: 10 days, layer: 1. After seeing the result, I am going to adjust the parameters and try to get the best result.

Reference

- Understanding LSTM Networks (https://colah.github.io/posts/2015-08-Understanding-LSTMs/)
- 9 Mistakes Quants Make that Cause Backtests to Lie by Tucker Balch, Ph.D. (https://blog.quantopian.com/9-mistakes-quants-make-that-cause-backtests-to-lie-by-tucker-balch-ph-d/)
- Avoiding Look Ahead Bias in Time Series Modelling (https://www.datasciencecentral.com/profiles/blogs/avoiding-look-ahead-bias-in-time-series-modelling-1)