Machine Learning Engineer Nanodegree

Financial Time Series Prediction via Machine Learning

Bryan Chik

16th July 2017

Background

Both academia and practitioners have been long searching for systematic ways in predicting asset prices. Predicting asset prices are difficult due to the fact that asset prices are determined by human agents in the market, and we are modeling human behavior.

Traditionally, predicting asset prices heavily utilize linear factor models as an attempt to quantify market expectations[1]. While in this project, I hope to apply machine learning techniques as a new class of methods to predict asset prices, with an ultimate goal to develop a profitable systematic long-only trading strategy based on the prediction using ETFs.

Problem Statement

Inspired by Google Cloud's recent use case, I hope to apply machine-learning techniques to predict Hong Kong's Hang Seng Index 1-day movement direction, given global index closing levels.

Since we are predicting movement direction (Up/Neutral/Down), this problem will be a classification problem.

Datasets and Inputs

The following data will serve as an input

- Daily index closing data consisting of 43 global exchange indices, spanning across major markets in Asia Pacific, Europe, US and Middle East, since 1st January 2010 to 30th June 2017. Full list of exchange indices are provided in the Appendix section.

The data is sourced from financial analytics provider FactSet Research Systems, via their Excel API.

Up/Neutral/Down labels are first generated using Hang Seng Index daily returns with the following threshold:

- Up – 1-Day return > 0.05%
- Neutral – 1-Day return within [-0.05% , 0.05%]
- Down – 1-Day return < 0.05%

[1] Roll, Richard; Ross, Stephen (1980). "An empirical investigation of the arbitrage pricing theory". *Journal of Finance*. **35** (5): 1073–1103. JSTOR 2327087. doi:10.2307/2327087

The features space will be using the rest of the global exchange indices closing daily return levels (i.e. 42 features).

The data will split into training, validation and test sets via 80/10/10 proportion. Instead of the usual practice where the training, validation and test sets are generated through randomizing the sample, we have to preserve the time structure of the data. We will be using the first 80% as training set, the following 10% as training set, and the final 10% as test set, in chronological order.

Solution Statement

Since we are estimating the index movement by up/neutral/down labels, we will apply classification algorithms and hope to search for the best classifier, based on the evaluation metrics defined in the "Evaluation metrics" section.

Pre-processing of the data is required, due to the different absolute levels of the indices, to preserve the structure of the data. Raw index levels cannot be used due to the non-stationary nature of the time-series data, so instead we will need to transform the index level to daily returns data.

After pre-processing of the data, we can then explore various supervised learning techniques as first attempts to build the model. Four Candidate algorithms include:

- Logistic Regression
- AdaBoost
- Feedforward Neural Networks
- LSTM

We will use cross-validation techniques to choose the best parameters for relevant models.

Benchmark Model

We can compare the classifier to 1 benchmark model:

- Gaussian Naïve Bayes

The reason behind is that empirical distribution for stocks/indices returns are noisy, but approximately Gaussian. We can see if our classification algorithm is better than computing likelihoods using the approximate empirical probabilities (using Gaussian Naïve Bayes).

Evaluation Metrics

The following metrics, which are relevant to classification problems, will be used to evaluate the performance of the model

- F-beta Score (with beta = 0.6)
- Precision

Precision and F-beta Score with emphasis on precision component is included due to the reason that we want to ultimately construct a long-only portfolio, which in turn consists of leveraged 2x ETFs that doubles your return when the index is up. Therefore, correctly predicting up label is more important.

Project Design

The outline of the project will be as follows:

1) *Raw Data Analysis*

   This section includes studying the raw data to identify whether preprocessing is required. Key areas to study will be:

   - Descriptive Statistics of the index data
   - Variance-Covariance structure of the data
   - Look-ahead bias due to different closing times

2) *Data Preprocessing*

   This section will be performing the identified preprocessing to the raw data and transform it into a usable format in the classification problem. Key tasks will be:

   - Turning the raw closing index level data to daily returns data to turn the series into a stationary series
   - Applying suitable lag to indices that closes after Hong Kong (e.g. US/Europe/Middle East)

3) *Learn the data!*

   This section will be performing the classification learning task, given the preprocessed data from above. We will try to fit the data with 4 specific classification algorithms

   - Logistic Regression
   - AdaBoost
   - Feed Forward Neural Networks
   - LSTM

   Evaluation metrics will be computed and compared across the benchmark models, to determine which algorithm works best.

4) *Improving the model*

   Finally, further improvement procedures will be explored to see if the performance of the best-chosen model from 3:

   - Applying PCA for reducing dimensionality to the features space, then apply the best-chosen classification algorithm from 3.
   - Applying Tree-based feature selection, then apply the best-chosen classification algorithm from 3.

Appendix

List of exchange indices within the raw data

Index Name	Country
Hang Seng Index	Hong Kong
SSE Composite Index	China
ASX All Ordinaries	Australia
India S&P BSE SENSEX	India
TOPIX	Japan
KOSPI Composite Index	South Korea
Taiwan TAIEX	Taiwan
FTSE Bursa Malaysia KLCI	Malaysia
FTSE Straits Times Index	Singapore
Philippines PSE PSEi	The Philippines
NEW ZEALAND NZX 50(CAP)	New Zealand
Thailand SET	Thailand
Euro STOXX	Eurozone
FTSE 100	UK
France CAC 40	France
Germany DAX (TR)	Germany
FTSE MIB	Italy
Belgium BEL-20	Belgium
Ireland ISEQ Overall	Ireland
Netherlands AEX	Netherlands
Norway OSE OBX TR	Norway
Spain IBEX 35	Spain
OMX Stockholm 30	Sweden
Switzerland SMI (PR)	Switzerland
Austria ATX	Austria
OMX Copenhagen 20	Denmark
OMX Helsinki 25	Finland
Portugal PSI 20	Portugal
Russia RTS	Russia
Czech Republic PX - 50	Czech
Hungary BUX	Hungary
Poland WIG	Poland
Turkey BIST 100	Turkey
S&P 500	US
DJ Industrial Average	US
Colombia IGBC	Colombia

Canada S&P/TSX Composite	Canada
Brazil Bovespa Index	Brazil
Mexico IPC	Mexico
Israel TA-125	Israel
Saudi Arabia All Share (TASI)	Saudi Arabia
FTSE JSE All Share	South Africa
Abu Dhabi Securities Exchange	UAE