Capstone 1 Proposal for Springboard Data Science Career track
Version: 3.0
Author: Jiaqi Xu
Coach: Kevin Glynn
Date: 06/26/2020

Statistical Arbitrage on S&P 500 by Machine Learning Models

Background

S&P 500 is one of the most commonly followed equity indices which measures the stock performance of largest 500 companies listed on stock exchanges in US. The average annual total return of the index, including dividends, since inception in 1926 has been 9.8% (Wikipedia). From Berkshire Hathaway's annual meeting 2020, one of the most famous investors Warren Buffet indicated "In my view, for most people, the best thing is to do is owning the S&P 500 index fund". Besides directly investing on S&P 500 index, statistical arbitrage trading on composites of S&P 500 index is also a hot topic.

Literature Review

Krauss et al. (2017) developed a statistical arbitrage strategy based on deep neural networks (DNN), gradient-boosted trees (GBT), random forests (RAF), and their simple ensample and deployed it on the S&P 500 constituents from 1992 to 2015. The daily returns of 0.25 percent with annualized Sharpe ratio of 1.81 proves their methods performed better than general market. Using the same data as Krauss et al. (2017), Fischer and Krauss (2018) deployed long short-term memory (LSTM) networks for predicting out-of-sample directional movements for the constituent stocks of the S&P 500. Their daily returns of 0.46 percent and a Sharpe Ratio of 5.8 prior to transaction costs showed LSTM outperformed DNN, GBT and RAF. Fischer et al. (2019) performed a statistical arbitrage strategy based on logistic regression (LG) and RAF with 40 cryptocurrency coins on minute-binned data.

Goal

This project will compare different machine learning methods in S&P 500 trading with an extension data (start from year 2006 to the end of year 2019) different from Krauss et al. (2017). We also add more features by volumes in our feature creation part. The goal is to compare the recent results of machine learning methods to the previous results and check

the hypothesis that "profits are declining in recent years and there is a severe challenge to the semi-strong form of market efficiency".

Data

The dataset I will use for this project:

- Download historical S&P500 components on Wharton Research Data Services (WRDS) and Wikipedia, clean them and get a mapping table indicates which stock will be included in S&P500 at the end of the specific time.
- Based on the components from above resources, I will use Python yahoo finance module yahoo-finance-1.0.4 to download daily data of all components and S&P 500 index from 2005/01/01 to 2020/01/01 (15 years data). Since yahoo finance will not keep the delisted stock's information, several delisted stock's data in historical S&P 500 will not be available. This will not affect the prediction too much because most of the stock's information will be arrived.
- Other financial data such as currency rate, commodity prices, stock indices on other markets will also be considered as relative features in prediction process.

Approach

(1) Training and trading sets

The cleaned dataset will be split into training and trading parts. Like Krauss (2017), we set the length of the in-sample training window to 750 days (approximately three years) and the length of the subsequent out-of-sample trading window to 250 days (approximately one year). We move the training-trading set forward by 250 days in a sliding-window approach, resulting in 12 non-overlapping batches to loop over our entire data set from begin of year 2005 until end of 2019.

(2) Feature generation

For each training-trading set, the feature space (input) and the response variable (output) are created as follows:

Input:

Let $P^s = (P_t^s)_{t \in T}$ denote the price process of stock s or financial indicator, with $s \in \{1, \dots, n\}$.
Define the simple return $R_{t,m}^s$ for each stock or indicator over m periods as

$$R_{t,m}^s = \frac{P_t^s}{P_{t-m}^s} - 1 \quad (1)$$

Let $V^s = (V_t^s)_{t \in T}$ denote the volume process of stock s or financial indicator, with $s \in \{1, \dots, n\}$.
Define the simple volume change $V_{t,m}^s$ for each stock or indicator over m periods as

$$V_{t,m}^{s} = \frac{V_t^s}{V_{t-m}^s} - 1 \quad (2)$$

Where $m \in \{\{1, \dots, 20\} \cup \{40, 60, \dots, 240\}\}$

Output:

Construct a binary response variable $Y_{t+k,k}^{s} \in \{0,1\}$ for each stock s. The response $Y_{t+k,k}^{s}$ equals to one (class 1), if the k-period return $R_{t+k,k}^{s}$ of stock s is larger than the corresponding cross-sectional median return computed over all stocks and zero otherwise (class 0). Here we can try different k(s). For example, if k=1 it means one-period return prediction. We try to forecast probability $\mathcal{P}_{t+k}^{s}$ for each stock s to outperform the cross-sectional median in period t+k. Please note the binary response also can be extended to multinomial response if time is permit.

(3) Machine Learning models

All or parts of the following machine learning models will be used in classification process:

→ Deep neural networks (DNN)
→ Gradient-boosted trees (GBT)
→ Random forests (RAF)
→ Long short-term memory (LSTM)
→ Logistic regression (LG)
→ Simple ensemble of machine learning methods

(4) Trading

Sorting all stocks over the cross-section in descending order, separately above forecast methods, At the top, the most undervalued stocks according to the respective learning algorithm and at the bottom the most overvalued stocks with the lowest probability to outperform the cross-sectional median in period t + k. In consequence, go long the top q stocks of each ranking, and short the bottom q stocks, with $q \in \{1, \dots, \lfloor n/2 \rfloor\}$.

Deliverables

- Jupyter notebook with relative codes on GitHub
- Final Report includes data acquisition, data transformation, model implementation and analysis, results comparation and conclusions
- Presentation Slides
- PDF report

Reference

Krauss, C., Do, X. A., & Huck, N. (2017). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. European Journal of Operational Research, 259(2), 689-702.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. European Journal of Operational Research, 270(2), 654-669.

Fischer, T. G., Krauss, C., & Deinert, A. (2019). Statistical arbitrage in cryptocurrency markets. Journal of Risk and Financial Management, 12(1), 31.