Data 698 - Project Proposal
Tze Fung Lung, Jim
February 15, 2019

Topic: Portfolio optimization and Machine learning with visualization analysis for S&P 500

10 Keyword: S&P500 Stocks Return Risk Strategic Linear k-Nearest Moving-Average ARIMA LSTM

Abstracts

Predicting how the stock market will perform is one of the most difficult things to do. There are so many factors involved in the prediction – physical factors vs. physiological, rational and irrational behavior, etc. All these aspects combine to make share prices volatile and very difficult to predict with a high degree of accuracy.

The S&P 500 is widely regarded as the best single gauge of large-cap U.S. equities. The index includes 500 leading companies and captures approximately 80% coverage of available market capitalization.

1. Description of the Problem

We'll look at the S&P 500, an index of the largest US companies. The S&P 500 is an American stock market index based on the market capitalization of 500 large companies having common stock listed on the NYSE, NASDAQ Exchange. I will load all 500 dataset in S&P 500 for analysis by using portfolio optimization to get the possible several stocks with higher return and lower risk. And using the machine learning predict the investment trend for S&P 500 index.

- What are the top 20 higher monthly return among all 500 number of stocks in S&P500 by Mathematical programming? The target is to find out the top valuable, higher return with lower risk of stocks.

- Could I invest these top 20 stocks now by analysis for the trend of S&P500 index by Machine learning? It is to determine if I could invest these stocks by choosing the most accuracy model with the trend.

2. Why the problem is interesting

Automatic trading without anyone involved will be the trend of stock market near future. I would like to use the data science methods to make a strategic for investment.
I will study which method of machine learning would be more accurate, suitable for prediction by using root-mean-squared error, that the prediction will be more meaningful in use.

3. What other approaches have been tried

First of all, I will construct the portfolio optimization in order to achieve a maximum expected return given their risk preferences due to the fact that the returns of a portfolio are greatly affected by nature of the relationship between assets and their weights in the portfolio.
The top 20 monthly return of stocks will be get into the portfolio optimization. Then in order to get the higher return and lower risk of stocks, the portfolio optimization will be conducted to find out which are the best choose of investment and generate the visualization for returns and volatility.

For the next part, I will work with historical data about the S&P500 price index to understand if I can invest in market this moment. I will implement a mix of machine learning algorithms to predict the future stock price of this company, starting with simple algorithms like averaging and linear regression, and then moving on to advanced techniques like Auto ARIMA and LSTM.

And I will compare the models by using root-mean-squared error (RMSE) to measure of how model performed and measure difference between predicted values and the actual values.

4. Discussion on your hypothesis is and how you specific solution will improve

Stock market analysis is divided into two parts – Fundamental Analysis and Technical Analysis.
Fundamental Analysis involves analyzing the company’s future profitability on the basis of its current business environment and financial performance. Technical Analysis, on the other hand, includes reading the charts and using statistical figures to identify the trends in the stock market.

We’ll scrape all S&P 500 tickers from Wiki and load all 500 dataset to be in cleaning and appending the adjusted closing price from 2008 to 2018.

Moving Average - The predicted closing price for each day will be the average of a set of previously observed values. Instead of using the simple average, we will be using the moving average technique which uses the latest set of values for each prediction.

Linear Regression - The most basic machine learning algorithm that can be implemented on this data is linear regression. The linear regression model returns an equation that determines the relationship between the independent variables and the dependent variable.

K-Nearest - Neighbors Another interesting ML algorithm that one can use here is kNN (k nearest neighbours). Based on the independent variables, kNN finds the similarity between new data points and old data points.

ARIMA - ARIMA is a very popular statistical method for time series forecasting. ARIMA models take into account the past values to predict the future values.

Long Short Term Memory (LSTM) - LSTMs are widely used for sequence prediction problems and have proven to be extremely effective.