Machine Learning Engineer Nanodegree

Capstone Proposal

Junghoon Lee
December 11, 2017

Proposal

Stock Prediction Machine Learning

Domain Background

Machine Learning is using everywhere. There are few things we cannot do with Machine Learning in these days. Stock price prediction is the most interesting issue with this technology. Many hedge fund companies are using machine learning for stock prediction and keeping the best portfolio.
Prediction methodologies fall into two broad categories. They are fundamental analysis and technical analysis.
Fundamental Analysts are concerned with P/E ratio, validity of stock, evaluation of company past performance. Warren Buffett is the most famous fundamental investor. He did not trade stock frequently, but keep long time. (Murphy, John J. (1999). *Technical analysis of the financial markets*.)
However, technical analyst are not concerned with any of the company's fundamental, but rely on chart information. For example, exponential moving average, candlestick pattern.
This, on the other hand, is solely based on the study of historical price fluctuations. Practitioners of technical analysis study price charts for price patterns and use price data in different calculations to forecast future price movements (Turner, 2007).
Since I have been investing money on stock market, this project result will be my best information for future investing. I want to see how much machine learning algorithm can predict real stock price.

Problem Statement

There are a lot of input data to predict future stock price. It will be quite difficult to make a special set or a equation for all stocks in stock market(Nasdaq). Some stock (S&P 500 ETF) is quite converged market signal(interest rate, oil price and so on), but some small stocks(special market) are not follow market signal. So, this domain is really fit to machine learning algorithm.
So , in this project, I will predict “adj close” stock price for a day, week, month time frame. Volume, 20-days mean stock value, 10-days mean stock value are used to input feature.

Datasets and Inputs

In order to implement/verify machine learning algorithm, stock adj close price, volume, date, dividends, split information will be used. These data can be gathered from yahoo/google finance. Nice python API set is available already. (https://pypi.python.org/pypi/yahoo-finance) In this project, S&P 500 ETF stock(IVV) will be used for machine learning algorithm. Since this has 500 companies stock in it, it will be easier to predict.
I will also check how much input data, for example, 1 years stock price, 2 years stock price, can provide better prediction result.
In addition, how many days Machine learning algorithm can predict. 1 day, 1 week, 1 month.

Solution Statement

I will implement three Machine learning algorithm. Polynomial regression, KNN and ensemble algorithms to see stock prediction performance.
In this project, S&P 500 ETF stock(IVV) will be used for machine learning algorithm. Since this has 500 companies stock in it, it will be easier to predict(less volatile).
“Adj close” stock price of S&P 500 ETF(IVV) will be predicted.
I will check range of input data, for example, 1 years stock price, 2 years stock price, can provide better prediction result.

I will investigate on how many days Machine learning algorithm can predict. 1 day, 1 week, 1 month.

Benchmark Model

Although many hedge fund companies use machine learning algorithm, that is not available on public. So, I will create benchmark prediction model using simple feature set(stock price, volume) and linear regression.

Evaluation Metrics

R2 score will be used as evaluation metrics.

Project Design

Input feature is the most importance on this project. So, I need to care about a specific stock(company)'s target market and need to find out which index value can give some relation. S&P 500 ETF (IVV) stock will be used in this project. This project CVS files are obtained from google/yahoo finance web site. Volume, 20-days average stock price, 10-days average stock price are used to input feature.
I need to focus on how long machine learning algorithm can show reasonable prediction out. For example, 1 day, 1 week, 1 month.
Since this is stock forecasting, I always took training data from old days and took test data from the later than training date. This is look ahead bias in time series machine learning.

I will start this with very simple(2 or 3 input feature) polynomial regression algorithm for benchmark.
Then, more input features are added on Polynomial regression algorithm.
Next, KNN will be tried with same input features.
Finally, ensemble algorithm with Polynomial regression and KNN.
All performance score will be calculated by evaluation metrics on above.