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.
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.

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.

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)

Solution Statement

I will implement three Machine learning algorithm. Polynomial regression, KNN and ensemble algorithms to see stock prediction performance.

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 polynomial regression.
Then, I will do performance compare with my algorithms using this evaluation metrics.

Evaluation Metrics

Accuracy = 1 - sum [absolute value(prediction[i] - test[i]) / test[i]] {i = 5 days, or 30 days}

This is percentage 0 ~ 100%. High accuracy shows better performance.
1 Week, 1 Month prediction will be calculated.

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.
This project CVS files are obtained from google/yahoo finance web site.
I need to focus on how long machine learning algorithm can show reasonable prediction out. For example, 1 day, 1 week, 1 month.
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.