Machine Learning Engineer Nanodegree

Capstone Proposal

Ronak Tanna
April 21st, 2017

Proposal

Domain Background

The **stock** of a corporation is constituted of the equity stock of its owners. A single share of the stock represents fractional ownership of the corporation in proportion to the total number of shares. A **share price** is the price of a single share of a number of saleable stocks of a company, derivative or other financial asset.

Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on an exchange. The successful prediction of a stock's future price could yield significant profit. The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable. Others disagree and those with this viewpoint possess myriad methods and technologies which purportedly allow them to gain future price information.

A lot of research has been done in this domain, few of which are:
- **Citation:** Saahil Madge."Predicting Stock Price Direction using Support Vector Machines". *Independent Work Report Spring*, 2015.
  This research paper is relevant to this section because it provides details on how to calculate price volatility and momentum for individual stocks and for the overall sector.
- **Citation:** Shunrong Shen, Haomiao Jiang and Tongda Zhang. "Stock Market Forecasting Using Machine Learning Algorithms".
  This project exploits the temporal correlation among global stock markets and various financial products to predict the next-day stock trend with the aid of SVM.

Problem Statement

Stock price prediction is one of the most widely studied and challenging problems, attracting researchers from many fields including economics, history, finance, mathematics, and computer science. The volatile nature of the stock market makes it difficult to apply simple time-series or regression techniques. Financial institutions and traders have created various proprietary models to try and beat the market for themselves or their clients, but rarely has anyone achieved consistently higher-than-average returns on investment. Nevertheless, the challenge of stock forecasting is so appealing because an improvement of just a few percentage points can increase profit by millions of dollars for these institutions. We will try to predict whether the price of a stock will increase or decrease in the following day which help us in identifying the 'winning' stocks.

Datasets and Inputs

We wanted to see how the SVM model, which has had such success in previous literature, would work in such an abnormally volatile market. We will also look into other Supervised

Learning algorithms and see how much difference they show when compared to SVM. We focus specifically on the technology sector. Focusing on a sector as opposed to the broad market allow us to test the model on companies that are similar to each other, making our results relatively standardized. We use the NASDAQ-100 Technology Sector Index (NDXT) as the general technology sector index. We look at 34 of the 39 stocks in the index. For each individual company we look at daily price data from the start of 2007 through the end of 2014. This allows us to analyze the fall of each company during the Recession as well as the recovery up to current times.

Solution Statement

Of course we cannot predict the exact closing price of the stock with hundred percent accuracy. However, our aim here is to be able to predict a trend or movement in the market which has an accuracy of anything better than just making a random decision. We will try using various machine learning algorithms to see which technique shows most potential. The idea is to be able to predict the change in state of a stock over either a short period of time or a over a long period of time taking into consideration the data points over a range of the number of days.

Benchmark Model

According to the Efficient Market Hypothesis it is impossible to predict the trend of the stock market. Hence, the benchmark Model for our algorithm to beat would be to make a prediction which better than random guessing. Hence a simple logistic regression model with an accuracy of 50% will be considered as the Benchmark Model.

Evaluation Metrics

In this project, our aim is to predict whether the NASDAQ 100 index companies will go 'Up' or 'Down'. As this a binary classification problem the accuracy metric we will use is accuracy. Accuracy is defined as the number of correctly classified points in comparison to the total number classifications made.

Project Design

In this final section, summarize a theoretical workflow for approaching a solution given the problem. Provide thorough discussion for what strategies you may consider employing, what analysis of the data might be required before being used, or which algorithms will be considered for your implementation. The workflow and discussion that you provide should align with the qualities of the previous sections. Additionally, you are encouraged to include small visualizations, pseudocode, or diagrams to aid in describing the project design, but it is not required. The discussion should clearly outline your intended workflow of the capstone project.

The basic idea behind the project is to gather data using the Quandl API for the companies we will be considering. Once the data has been gathered we will apply certain preprocessing techniques on the data in order to normalize it and get rid missing data points if any. We will then look at the statistical distribution of the data across various companies. Our next step would be to split our time-series data into training, testing as well as validation set. We then use various classification algorithms including SVM, Decision Tree classifier as well as Random Forest classifier in order to check the performance of these classifiers and see the determine the best of the lot.