Machine Learning Engineer Nanodegree

Capstone Proposal

Prakhar Choudhary
May 23, 2017

Proposal

Domain Background

The stock market has always been a goldmine of adventures for mathematicians and statisticians.
They keep trying to find patterns such that the behaviour of stock market can be predicted, however the huge amount of data and buy/sell decisions carried out everyday makes it almost impossible to be analysed manually. This is where computers come in. Early research on stock market prediction was based on random walk theory and the Efficient Market Hypothesis (EMH)[1], which ironically is not so efficient. However in the recent few years the emergence of new organised data, high computational power and machine learning algorithms has given us the ability to use computer for predicting the behaviour of stock market.

Problem Statement

The goal of this project is to apply sentiment analysis and machine learning principles to find the correlation between ”public sentiment” and ”market sentiment”[2] and use it to predict stock market movement. Furthermore the project will learn from new data every day from the previous day’s data(using Yahoo Finance API). The project will be a web app that will suggest users to either Buy, Sell or keep stock depending on the user’s choice of stock out of the list of 30 companies present in Dow Jones Index. The project is inspired from a research paper by Anshul Mittal and Arpit Goel on “Stock Prediction Using Twitter Sentiment Analysis”.[2]

Datasets and Inputs

Two main datasets will be used in the project which are:

- Dow Jones Industrial Average (DJIA) values from June 2009 to December 2009. The data was obtained using Yahoo! Finance and includes the open, close, high and low values for a given day.
- Publicly available Twitter data containing more than 476 million tweets corresponding to more than 17 million users from June 2009 to December 2009. The data includes the timestamp, username and tweet text for every tweet during that period. Since we

perform our prediction and analysis on a daily basis, we split the tweets by days using the timestamp information.

Solution Statement

The solution to the proposed problem directly builds on the one used by Bollen et al[1] and Anshul Mittal[2]. The raw DJAI values are first fed into the preprocessor to obtain the preprocessed values. At the same time, the tweets are fed into sentiment analysis algorithm which output mood classes for each day. These moods and the processed DJIA values are then fed to our model learning framework which uses a regression algorithm(will try different algorithm and find which is the best to use) to learn a model to predict future DJIA values using them. The learnt model as well as the previous DJIA and mood values are used by the portfolio management system which runs the model to predict the future value and uses the predicted values to make appropriate buy/sell decisions.

Benchmark Model

The work is based on Bollen et al's strategy [1] and Anshul Mittal's Paper[2] on it using their own validation metric and sentiment analysis mechanism. They also attempted to predict the behavior of the stock market by measuring the mood of people on Twitter. While Bollen et al's work which classified public sentiment into 6 categories, namely, Calm, Alert, Sure, Vital, Kind and Happy, achieved a remarkable accuracy of about 87% in predicting the up and down changes in the closing values of Dow Jones Industrial Index (DJIA), Anshul Mittal's approach led to a 75.56% accuracy. Hence, the desired results will be to obtain accuracy close to the original Bollen et al's results.

Evaluation Metrics

In order to measure accuracy, we will use a novel validation technique proposed in the paper by Mittal and Goel[2], called the *k-fold sequential cross validation (k-SCV).* In this method, we train on all days up to a specific day and test for the next k days. The direct k-fold cross validation method is not applicable in this context as the stock data is actually a time series unlike other scenarios where the data is available as a set. Therefore, it is meaningless to analyze past stock data after training on future values. For the purpose of our analysis, we use k = 5.

Project Design

- *Data Pre-Processing*

Stock prices data collected is not complete understandably because of weekends and public holidays when the stock market does not function. The missing data is approximated using a simple technique by Goel [2]. Stock data usually follows a concave function. So, if the stock value on a day is x and the next value present is y with some missing in between. The first missing value is approximated to be (y+x)/2 and the same method is followed to fill all the gaps.

Tweets consists of many acronyms, emoticons and unnecessary data like pictures and URL's. So tweets are preprocessed to represent correct emotions of public. For preprocessing of tweets we employed three stages[3] of filtering: Tokenization, Stopwords removal and regex matching for removing special characters.
1) *Tokenization:* Tweets are split into individual words based on the space and irrelevant symbols like emoticons are removed. We form a list of individual words for each tweet.

2) *Stopword Removal:* Words that do not express any emotion are called Stopwords. After splitting a tweet, words like a,is, the, with etc. are removed from the list of words.

3) *Regex Matching for special character Removal:* Regex matching in Python is performed to match URLs and are replaced by the term URL. Often tweets consists of hashtags(#) and @ addressing other users. They are also replaced suitably. For example, #Microsoft is replaced with Microsoft and @Billgates is replaced with USER. Prolonged word showing intense emotions like coooooooool! is replaced with cool! After these stages the tweets are ready for sentiment
classification.

- *Sentiment Analysis*

In this project, we use four mood classes, namely, Calm, Happy, Alert, and Kind. The methodology is directly build upon the one mentioned in the paper by Goel[1]:-

1. WordList Generation
2. Tweet Filtering
3. Daily Score Computation
4. Score Mapping

Furthermore, for cross-validation we will use Granger Causality[4]. Granger Causality analysis finds how much predictive information one signal has about another over a given lag period. The p-value measures the statistical significance of our result i.e. how likely we could obtain the causality value by random chance; therefore, lower the p-value, higher the predictive ability.

- *Model Learning and Prediction*

We will then use the results of sentiment analysis algorithm to learn a model that can predict the stock index and it movement.
Implement Linear Regression, SVM and SOFNN and find out the results. We will then choose the one which gives most promising results.

- *Portfolio Management*

Having predicted the DJIA closing values one day in advance, we can use these predicted values to make intelligent sell/buy decisions. We develop a naive greedy strategy based on a simple assumption that we can hold at most one stock at any given time (or *s* stocks if all stocks are always bought and sold together) Following are the steps/features of our strategy-

➔ Pre-computation

We maintain a running average and standard deviation of actual adjusted stock values of previous k days

➔ Buy Decision

If the predicted stock value for the next day is n standard deviations less than the mean, we buy the stock else we wait.

➔ Sell Decision

If the predicted stock value is m standard deviations more than the actual adjusted value at buy time, we sell the stock else we hold.

- *Building an interactive web app*

A django based webapp will let the users select the company for which they want to know the predictions and investment suggestions.

Acknowledgement

[1] J. Bollen and H. Mao. “Twitter mood as a stock market predictor”. IEEE Computer, 44(10):91-94

[2] Mittal, Anshul, and Arpit Goel. ”Stock prediction using twitter sentiment analysis.” Stanford University, CS229(2011 http://cs229.stanford.edu/proj2011/GoelMittalStockMarketPrediction-UsingTwitterSentimentAnalysis.pdf) (2012).

[3] Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, Babita Majhi. “Sentiment Analysis of Twitter Data for Predicting Stock Market Movements”. arXiv:1610.09225v1 [cs.IR] 28 Oct, 2016

[4] http://www.scholarpedia.org/article/Granger_causality