COMP 3211 Final Project Report
Stock Market Forecasting using Machine Learning

Group Member: Mo Chun Yuen(20398415), Lam Man Yiu (20398116), Tang Kai Man(20352485)
23/11/2017

1. Introduction

1.1 Motivation

Forecasting is the process of predicting the future values based on historical data and analyzing the trend of current data. Processing powers of computers nowadays have become powerful enough to process large amount of data. By running simulations of future states based on present states, we can foresee the trend of stock market.

1.2 Approaches

Linear Regression is a powerful statistic methodology modeling the relationship between a scalar dependent variable y or more explanatory variables denoted X. However, the accuracy of prediction using Linear Regression is actually not satisfactory.

In this project, we attempted to evaluate Linear Regression using Q-learning and improve the recommendation on choices of actions the user shall take in stock trading.

2. Related work

2.1 ***Machine learning algorithm in Quantopian***

Quantopian[] is a public and open website where people and professionals can share their programs and exchange ideas in the machine learning in financial sector. In this website, there are lots of valuable resource, including various algorithms and models of machine learning in predicting stock price. Most of their algorithms and models are very complex which is out of our level of understanding now. However, the use of machine learning concept in finance provide us an insight and introduce a lots of useful module in Python that are very helpful to our project.

2.2 ***Financial Programming using Python***

Our group is completely new to the Python and data science. pythonprogramming.net[] is an excellent resource that help us to understand the python syntax and teach us to using different modules in python to manipulate the data and make prediction. There are also some similar project in this website that help us to understand the concept much quicker

3. Problem definition

3.1 Scope of Data

In this project, a selection of stock data in the Standard & Poor's 500(S&P 500) are used for the prediction of trend.

3.2 Definition of Prediction

Our Program is aimed to identify the trend of the price of the target stock. Prediction here refers to the general trend of the specific stock price.

3.3 Evaluation of the Accuracy of the Prediction

The accuracy of the prediction is evaluated and give a percentage of accurate result.

3.4 Recommendation for the user

Combining the accuracy and the prediction, recommendation can be given to the user to acknowledge them the trend of the target stock with known accuracy.

4. Challenges

4.1 Variable representation

The variables in machine learning equations are not easily to be applied in stock market.

We could not precisely represent their utilities and values. For example, concepts in stock trading like buy, sell and hold are extremely difficult to be implemented even though the actions and states of successor function could be defined:

Actions: Either buy, sell or hold
States: current stock price

However, it is not feasible to define the reward function. For example, you cannot define the reward of the action “hold”. If the stock price drops after you buy it, you cannot define it as a loss because you are still holding the stocks. The stock price is still possible to rebound. The future value of the stock is still an unknown. Defining the value to a loss would affect the data consistency as the cumulative amount of loss would be larger than the actual value of loss once you sold the stock. Therefore, the reward function of “hold” cannot be defined as a loss. However, it is also not suitable for you to define it as +0. As the stock value is actually decreasing, contributing all the loss to the corresponding “sell” action would greatly affect the data integrity. It is hard for the system to trace back the declination or the actual intermediate states values of the data.

4.2 Quality of data

Due to limited open source APIs on stock prices , we have encountered problems on finding optimal data sets for our Linear Regression algorithm. In this project, we have chose an alternative method to get the stock prices. We first get 500 companies in American Stock Market by using a website reader library called beautifulsoup4. And then, we get the stock values of those companies with source ‘Yahoo’. Although ‘Yahoo’ is a reliable source, it only contains limited data sets.

5. Infrastructure

5.1 Python

This program is written in python, one of the most used language in Machine Learning.

5.2 Python’s modules

In this project, various python’s modules are used to facilitate predictions, regression analysis, graph plotting, data manipulation and machine learning. These include sklearn[], pandas[], pandas-datareader and matplotlib[].

6. Solution

In general, this project is going to use linear regression analysis to predict the trend of the target stock by obtaining the slope of the the linear regression line. We will also provide the predicted price of the stock at corresponding time point.

After obtaining the trend, we will evaluate the accuracy of the prediction by using Q-learning. The Q-value obtained will reflect the accuracy of the model with a heavier weight of present state.
Lastly, combine the prediction and the q-value by using a simple weighted sum to give a recommendation to the user.

6.1 Default Setting and Asumption

In the program, we will predict the trend of the target company by using the price data of the previous month, the actual number of days in the previous month will vary as stock exchange will close at certain days. We assume there are 30 days in a month for simplicity.

The stock price on the 7th day since the date the user inputted will be predicted by default. It is assume that the 7th day since the user inputted will be the working day of the stock exchange where actual stock price on that day would be available.

6.2 Regression Analysis and Visualization

With the aid of the sklearn module and matplotlib, we can do various regression analysis on the data and visualize it on a graph. In our case, we visualize the three

regression model, namely Radial Basis Function model, Linear model and Polynomial model. The are visualized on a graph to give user insight in the analysis.
See fig.1 as example(figure in session 8.1).

6.3 Linear Regression and Prediction

For simplicity, we use linear regression model for our prediction as it is easy to obtain the regression line slope, which can indicate the trend of the stock price. We use about 30 days of data to predict the trend of the upcoming week and output the predict stock on the 7th day since the date user inputted. The prediction follows the assumptions mentioned above.

See fig.2 as example(figure in session 8.2).

6.4 Q-learning

The Q-learning is not directly applied on the regression function, but it analyzes the quality of the function by considering whether the prediction result is reliable or not, reliability depends on the coefficient(slope of the linear regression line) of regression.
The following are the procedure and equation for the Q-learning.

i. Obtain the predicted slope of the linear regression and the actual trend of stock price on a large scale:

- The actual trend is defined as increasing if the actual price of stock from the beginning is less than that the actual price of the stock on the 7th day since the beginning.

- The actual trend is defined as increasing if the actual price of stock from the beginning is greater than that the actual price of the stock on the 7th day since the beginning.

ii.Define action,states, transition function and assigning future reward, discount, instantaneous reward and alpha(learning rate) value

- There is only 1 action, which is to use the linear regression model.
- There are two state: Prediction is correct(s_0). Prediction is wrong(s_1)
  s_0: actual trend is increasing and the coefficient (slope of the linear regression line) is positive or actual trend is decreasing and the coefficient (slope of the linear regression line) is negative
  s_1: Other than s_0

- Future reward($V(s')$) is 1 if reaching s_0, -1 if reaching s_1
- instantenous reward/transition reward($R(s,a,s')$) is always 0
- alpha(α) is 0.01
- discount(γ) is 1
- Transition function: unknown

iii. Sample based Q-value iteration

- From we will learn in the lecture note,
  Running average: $Q(s,\alpha) \leftarrow (1-\alpha)Q(s,\alpha) + \alpha[sample]$
  sample:$R(s,a,s') + \gamma(max)Q(s',a')$
  $V(s') = maxQ(s',a')$
  Thus, $Q(s,\alpha) \leftarrow (1-\alpha)Q(s,\alpha) + \alpha(R(s,a,s') + \gamma(V(s')))$
- By substituting the aforementioned value into the new running average. The resultant running average for our project:

$$Q(s,\alpha) \leftarrow 0.99Q(s,\alpha) + 0.01(V(s'))$$

$$V(s') = 1/-1$$

iv. Obtaining the Q-value using the resultant running average

6.5 Combining Q-value and prediction to give recomendation

- define accuracy(a) as number of correct prediction/total number of episode recommendation(r) is calculated as:
  $r = a(prediction) - (1-a)(Q(s,a))$

7. Limitiation

7.1 Limitation of quantitative analysis

The future values of stock prices are not only based on its historical data, they are also affected by some external factors, for example industry performance, investor sentiment and also economic factors. Due to the uncertainty in the stock market, there is no universal mathematical principle for predicting the future trend accurately. Quantitative analysis can only be applied to problems of computing mathematical principles. Therefore, the decision making based only on the quantitative analysis can lead to severe loss in investment. Quantitative analysis only provide insight from a mathematical perspective.

7.2 Limitation of Linear Regression

The accuracy of the prediction by Linear Regression is actually not high enough to make a good decision on stock trading. Linear Regression is limited to linear relationships. The algorithmn already assume the system is a straight-line. However, for stock trading, the values of the system could be either a raise, a drop or remain constant. The data values are scattered and fluctuated.

Apart from that, Linear Regression is not a complete description of relationships among variable. It only provides the functionality to investigate on the mean of the dependent variable and the independent variable. However, it is not applicable for the situation we encountered in stock market. And hence, the prediction is actually suppressed by this constraint.

8. Result and Analysis

We are going to predict the trend of AAPL(Apple inc.)

8.1 Visualization of Regression Model:

Visualization of Regression Model of AAPL(Apple inc.) from 22/10/2017 to 22/11/2017

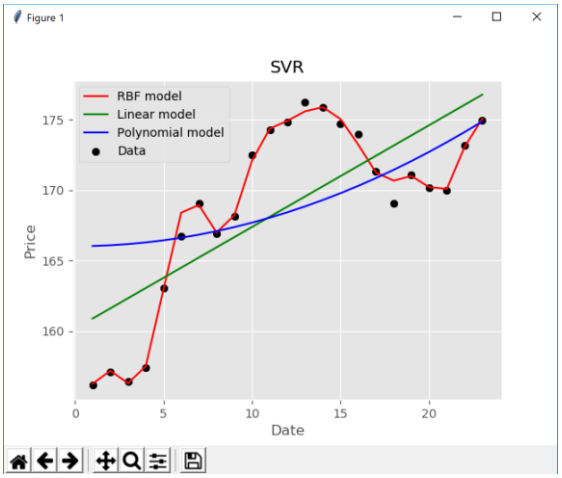

fig.1

8.2 Prediction:

Prediction output for AAPL(Apple inc.)

```
Warning (from warnings module):
  File "C:\Users\Harry\AppData\Local\Programs\Python\Python36\lib\site-packages\sklearn\cross_validation.py", line 41
    "This module will be removed in 0.20.", DeprecationWarning)
DeprecationWarning: This module was deprecated in version 0.18 in favor of the model_selection module into which all the refactored classes
and functions are moved. Also note that the interface of the new CV iterators are different from that of this module. This module will be re
moved in 0.20.

Training already done

type in ticker (-1 to exit): AAPL

Warning (from warnings module):
  File "C:\Users\Harry\AppData\Local\Programs\Python\Python36\lib\site-packages\sklearn\utils\validation.py", line 578
    y = column_or_1d(y, warn=True)
DataConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example
using ravel().

Predicted price:  186.855908131  Coefficient:  0.721244993083
Accuracy of prediction:  0.539610271903  Q value:  0.0208230855974
Recommendation:  0.543821945104928 6
```

fig.2

Explanation:
the recommendation (r) is given by the following equation:

$$r = (a)(prediction) + (1-a)(Q(s,a))$$

where a = accuracy:
(number of correction prediction)/(total number of episode)

prediction = 1, if $s_0$is achieve
prediction = -1, if $s_1$is achieve

Meaning of recommendation:
if r = 0, it predicts no trend.
if r>0, it predicts a increasing trend.
if r<0, it predicts a decreasing trend.

9. Conclusion

In this project, a Q-learning algorithm was implemented to give a recommendation to the user on the dependability of the result from Linear Regression. User can refer to the recommendation rating and then make decisions on stock trading. Linear Regression is a statistic methodology that is being criticized for its accuracy. Only depending on the result of Linear Regression cannot make a good decision on stock trading. Therefore, we have introduced a novel way of using machine learning to evaluate the rating of trust on the Linear Regression. Combining Linear Regression with Q-learning, we could produce a more accurate prediction for the user whether the stock price would follow the predicted trend of Linear Regression.