Machine Learning Engineer Nanodegree

Capstone Proposal

Hongqiao Li
December 2017

Proposal

Domain Background

In recent years, the machine learning becomes a very important technique which can be used in many areas, such as driving, searching, and economic area. This project will focus on how to use the machine learning techniques to predict the adjusted close price for a stock.

The stock price will be affected by many factors, such as tax rate, interests, profitability, new acquisition, etc. The stock is a time series feature. It will be affected highly by previous stock states. We need to also analyze the previous stock price to predict today's stock price. It is really hard for the human to gather and process some much information. So it is very important to use the machine learning way to predict the stock.

Problem Statement

There are more than seven thousand stocks in the stock market. I will choose the VGT stock to predict. The VGT is an index fund which contains the stocks of large, mid-size and small U.S. companies within the information technology sector.[1] It can represent the growth trend of the entire technology area.

Since the stock's price is affected by macroeconomics and microeconomic, there are two kinds of features I would use to predict the price. Macro-features and Micro features. The macro-features will be the interest rate, the unemployment rate, reserve ratio.[2] The Micro features will be the indicators which represent the trend, momentum, and volume of the stock.

For the model itself, I try both statistical machine learning and deep learning models.

Datasets and Inputs

In this project, I will use the dataset downloaded from Yahoo.[1] This dataset contains the historic market data from 1/30/2004 to 12/13/2017.

The data contains the following seven fields: Date, the date information. Open, the stock price at the opening of the stock market High, the highest stock price on a given day. Low, the lowest stock price on a given day. Close, the stock price at the end of a given day. Adj Close, the closing price adjusted for stock splits and dividends. Volume, how many stocks were traded.

There are 3494 rows data in this dataset.

Solution Statement

This project will predict the adjusted close price for a given day. This project will have steps:

1.Preprocess data. macro-features
a. Normalize the unemployment data and interest rate.

Micro features
a. Normalize historical stock price data.
b. Generate the indicators by using stock data

Combine these two kinds of data.

2.Split train and test data. The project will split the data into 8:1:1 pieces, which are training data, validation data and testing data. It means, the data ranges from 1/30/2004 to 3/9/2015 will be the training data. The data ranges from 3/10/2015 to 7/26/2016 will be the validation data. The data ranges from 7/26/2016 to 12/13/2017 will be the test data.

3.Training the model. This project will use both the machine learning models and the deep learning models to train. It will choose the best model as a final model.

Benchmark Model

The benchmark model will use the linear regression model with the same features, training, validation and testing data as the other models. The benchmark model will be compared by using the R^2 value.

Evaluation Metrics

The project will use R^2 value as the evaluation metric and the mean square error as the loss function, since the stock problem is a regression problem.

Project Design

This project will predict the adjusted close price for VGT stock.

The dataset will be the historical stock price for VGT downloaded by Yahoo.

The features are the macro-features and the micro-features.
macro-features: the interest rate and the unemployment rate
micro-features: everyday stock price and indicators

The preprocess will normalize the stock price. It also uses stock price to generate the indicator[3]. The micro-features (eg: size 20) and the macro-features (eg: size 30) will be combined to have a new feature (eg: size 20 + 30 = 50).

The model will be chosen from both deep learning models and machine learning models (linear regression, decision forest regression and boosted decision tree regression, etc.) by evaluation.

The loss function will be the mean square error. The evaluation function will be R^2.

---

[1]https://finance.yahoo.com/quote/VGT?p=VGT
[2]https://fred.stlouisfed.org/
[3]https://pypi.python.org/pypi/stockstats