MLND Capstone Proposal:

LSTM Predictor for S&P 500

Yichao Zhang

Oct, 2018

Domain Background

Inverstment companies use financial model to predict stocks. Some useful feature informations are hidden behind the historical data and other relative performance. However, stock price is not explicitly depend on these features, the prediction problem has never been well solved by any current math model.

Machine learning has the potential to create a complex statistical model for stock price prediction.

Problem Statement

Standard & Poor's 500 (S&P 500 (https://en.wikipedia.org/wiki/S%26P_500_Index)) is a very important index in stock trading market. Here we are going to build a machine learning model to predict the close price of S&P 500 in the next day.

Datasets and Inputs

The data is from Yahoo Finance (https://finance.yahoo.com/quote/%5EGSPC/), we download the data of recent 20 years (from 1998/08 to 2018/07) for training and testing.

Each data instance of one day includes: Date, Open, High, Low, Close, Adj Close and Volume. Adj Close (https://www.investopedia.com/terms/a/adjusted_closing_price.asp) is Adjusted Closing Price, and will be eliminated if it is all the same with Closing Price. For each instance:

Input:

Open, High, Low, Close, Adj Close and Volume in the recent 50 days.

Output:

The Close in the next day, that means the Close shifted by -1 day.

Solution Statement

In this projest, we will use two regression models in supervised learning algorithms: Linear Regression and LSTM (https://en.wikipedia.org/wiki/Long_short-term_memory). We will compare the performance of these two models and choose the better one.

Benchmark Model

We will use the Linear Regression model as a benchmark model. And LSTM will be our primary model.

Evaluation Metrics

The error of each model will be represented by Root-mean-square deviation (RMSE (https://en.wikipedia.org/wiki/Root-mean-square_deviation)), which is the square root of the mean of squared errors.

Project Design

1. Data Preprocessing
    - Down load the original data of S&P 500 from yahoo
    - Select data columns
    - Scale the data
    - Convert data type.
2. Develop Linear Regression Model
    - Create model
    - Fit the model
    - Prediction
    - Check Performance
3. Develop LSTM Model
    - Create model
    - Fit the model
    - Prediction
    - Check Performance, and compare with Linear Regression model.
4. Plot graphs and write the report