Machine Learning Engineer Nanodegree

Capstone Proposal

Guanyu Zhao, Feb 26, 2018

Proposal

Domain Background

Stock market prediction with machine learning techniques has a lot of applications. Lots of trading institutes predicts stock prices using deep learning and provides useful trade recommendations (buy/sell/hold) for the individual traders and asset management companies.

The accuracy of forecasts using time series models on economic data has received a great attention. The Box-Jenkins (Autoregressive Integrated Moving Average) ARIMA models have been widely used. These models give good forecasts for future observations but they are not accurate for nonlinear and non-stationary dataset [1].

With recent development of artificial neural network, It shows that recurrent neural network (RNN) can be a promising alternative to the traditional method, ARIMA, in forecasting especially in the case of nonlinear and non-stationary time series[2].

Problem statement

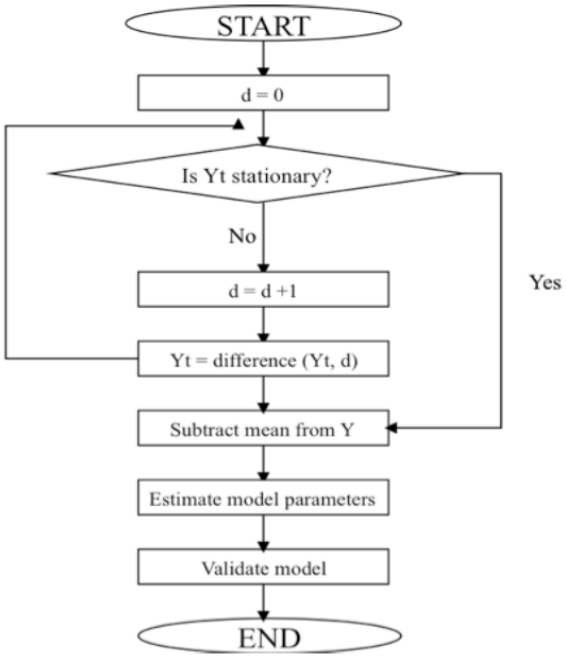
Time-Series often contain temporal dependencies that cause two otherwise identical points of time to belong to different classes or predict different behavior. This characteristic generally increases the difficulty of analyzing them. Existing techniques often depended on hand-crafted features that were expensive to create and required expert knowledge of the field[3]. The problem of this study is to use historical stock index data to predict the future. The aim of this study is to compare the performance of a classical (ARIMA) with an RNN (Long short term memory LSTM) forecasting techniques for financial time series of Dow Jones industrial index. Both techniques are very similar in that they attempt to discover the appropriate internal representation of time series data. In order to reach the goal of this study, a python package 'Statsmodel' is used for classical ARIMA model[6]. Tensorflow is used as a framework to establish the LSTM model.

Datasets and inputs

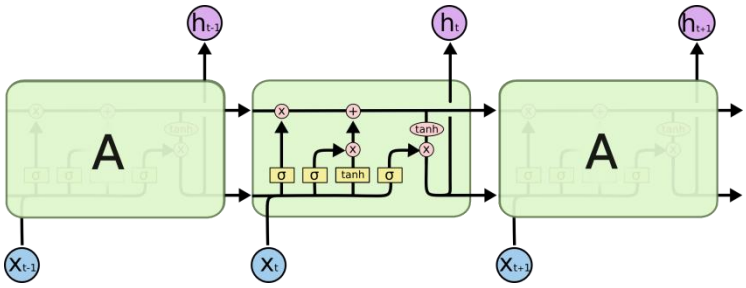
Dow Jones Industrial Average (DJIA) values from Jan 2013 to Dec 2017. The data was obtained using Google Finance API and includes the open, close, high and low values for a given day. The Dow Jones Industrial Average or simply the Dow, is a stock market index that shows how 30 large publicly owned companies based in the United States have traded during a standard trading session in the stock market. The value of the Dow is the sum of the price of one stock for each component company, and corrected by a factor which changes whenever one of the component stocks has a stock split or stock dividend, so as to generate a consistent value for the index.[5]

Solution statement

For the forecasting purposes the usage of ARIMA model can be summarized by the flowchart shown in the following figure:



The building block of LSTM model is shown in the figure below [4].



Benchmark model

The classical ARIMA model is selected as the benchmark model. The forecasting model based on LSTM is expected to outperform the classic model especially in dealing with nonlinear and non-stationary stock price index.

Evaluation Metrics

For selecting the best ARIMA model, Forecast Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) will be used to indicate the performance of different ARIMA models. And then root mean squared error (RMSE) will be used as metric to evaluate the forecasting performance between LSTM and ARIMA model.

Project design

- Data pre-processing
  The input data will be 5 years DJIA values from Jan 2013 to Dec 2017. There are in total 1260 observations. Only the Adj Close price will be used for training. The input data will be pre-processed using MinMaxScaler before fed into the model. The original data will be split into training(80%) and test(20%) dataset.

- Data exploratory data analysis (EDA)
  Some basic analysis will be performed such as candle plots, moving average plots with different rolling window sizes and envelope plots etc, in order to understand the trend and underlying triggering event of sudden change of stock market.

- ARIMA model selection
  Following the flowchart in solution statement, the best ARIMA model will be selected. First, log transformation and difference are used to transform the original data to stationary dataset. This can be checked with Dicky-Fuller test. Then, based on the ACF and PACF plots, the numbers of AR and/or MA terms can be tentatively identified. At last, grid search on some model parameters, and calculate BIC and AIC respectively and find the best ARIMA model.

- RNN network architecture
  Try different LSTM layers, dropout, activation function and epochs to find the optimal model.

- Result comparison and final discussion
  Compare prediction results for both training and testing set. Discuss both pro and cons for classical ARIMA model and RNN model with LSTM units.

Reference

1. Box, G.E.P. Jenkins, G.M. and Reinsel, G.C; Time Series Analysis: Forecasting and Control; 3rd edition; Prentice Hall: Englewood Cliffs, New Jersey
2. Enzo Busseti, Ian Osband, Scott Wong, Deep Learning for Time Series Modeling, cs229 Stanford University
3. John Cristian Borges Gamboa, Deep Learning for Time-Series Analysis, Cornel University
4. http://colah.github.io/posts/2015-08-Understanding-LSTMs/
5. https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average
6. http://www.statsmodels.org/stable/index.html