Deep Learning : Predicting Stock Prices Using LSTM

Ron Labe, **Eli Ruvinov,** **Tal Ben-Gozi.**

ronleib95@gmail.com **eli4056@gmail.com** **talbg9@gmail.com**

Git: https://github.com/Eliot100/DJIA-stock-project

Ariel University

Abstract

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).
Regression helps investment and financial managers to value assets and understand the relationships between variables, such as commodity prices and the stocks of businesses dealing in those commodities.

Classification is the attribution of labels to records according to a criterion automatically learned from a training set of labeled records. This task is needed in a huge number of practical applications, and consequently it has been studied intensively and several classification algorithms are available today. In finance, a stock market index is a measurement of value of a section of the stock market. Clearly, such a stochastic value is very **difficult to predict**. However, binary classification using high level technology like MLP, CNN, LSTM etc.. is a method to overcome the problem and forecast the direction of prices of stock markets. In our project we show you how we try to make it possible through top 25 historical news headlines of the same day. We took datasets containing daily historical news from 'Reddit' and use it to predict the value of the stock which is 1 if the price of our stock (DJIA) increase or stays the same and 0 if the price of the stock decrease.

Introduction

The aspiration of any investor is to forecast the market behavior with the aim of making the best decision when he comes to buying or selling shares of stocks seeking to maximize his profits. This is a difficult task because market behavior is volatile and influenced by many factors such as global economy, politics, investor expectation and others. However, advances in artificial intelligence and the growth of available data have made possible to try to forecast the stock price behavior. in some cases, investors tend to buy after positive news resulting in a stress of buying and higher stocks prices; and after negative news, they sell, resulting in a decrease of prices. This approach has a limitation since the market reacts to external information. Many people propose the use of text mining techniques and machine learning techniques to analyze textual data and take out information that can be relevant to the forecast process. Recently, with more computational capabilities and the availability to handle massive databases, it is possible to use more complex machine learning models, such as deep learning models, which presents a superior performance in traditional Natural Language Processing (NLP) tasks. The outstanding deep learning

models are: Convolutional Neural Network (CNN) Recurrent Neural Network (RNN), specifically the Long Short Term Memory architecture (LSTM) using Linear or Logistic regression or MLP.
Our project apply a deep neural network model that use as input events taken from historical news articles to forecast the direction of prices of DJIA (Dow Jones Industrial Average).

Related works

Stock Market prediction has been a trivial work for the Scientists and Mathematicians. At early stage Mathematics and statistics could predict the Stock Prices at lower efficiency using Regression models. With the change of time Artificial Neural Networks using historic values took its place. Deep Learning Algorithms has proven to be best in this field as we can not only train the values using historical data but also using data like Media, News , demand and supply.
The fundamental contribution of the paper is as follows:

1. we have used Python with TensorFlow giving Deep Learning aspect for prediction of the future values using historical values.

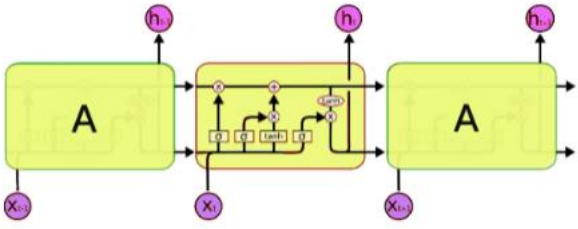
2. Multiple hidden layers are present in the system which makes the precision to be higher with better training support. Our layers are:
Input (187 values) -> LSTM (200 units) -> Dropout(0.2) -> LSTM (150 units) -> Activation('linear') -> Dense(15 units) -> Output (single value).

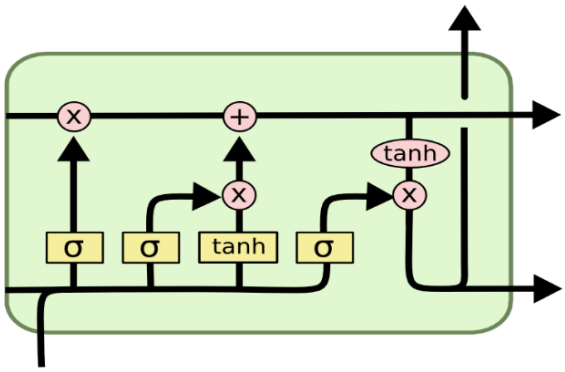
Background

MSE: Mean squared error (MSE) is the most commonly used loss function for regression. The loss is the mean overseen data of the squared differences between true and predicted values, or writing it as a formula:
Y_i: $true\ value$
$\hat{Y}_i$: $predicted\ value$

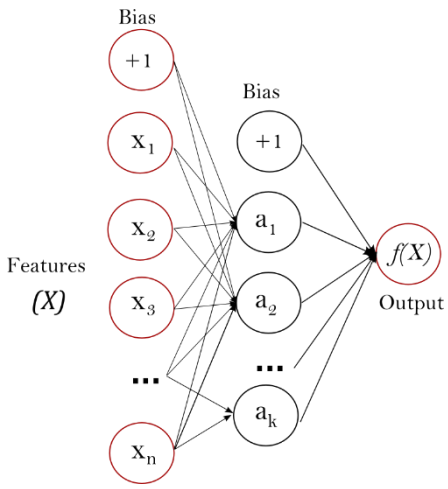
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Long short-term memory: Long short-term memory is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points ,but also entire sequences of data.
A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.





Multi-layer Perceptron (MLP):



MLP is a supervised learning algorithm that learn the function: $f(\cdot): R^m \rightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features $X = x_1, x_2, \ldots, x_m$ and a target y, it can learn a non-linear function approximator for either
classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers. The figure below shows a one hidden layer MLP with scalar output.

The leftmost layer, known as the input layer, consists of a set of neurons $\{x_i | x_1, x_2, \ldots, x_m\}$ representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation $w_1x_1 + w_2x_2 + \ldots + w_mx_m$, followed by a non-linear activation function $g(\cdot): R \rightarrow R$ - like the hyperbolic tan function. The output layer receives the values from the last hidden layer and transforms them into output values.

Project Description

This project is intended to predict the closing value of DJIA (Dow Jones Industrial Average) stock for each day between 2008-09-22 to 2016-07-01 using analysis of top headlines and historical stock data.
We pull the historical stock data and the headlines from kaggle website

We pull for each day data of 30 days backwards of: Open, High, Low, Close, Volume, Adj Close and the Stock Opening Price for the relevant day. the data taken from 'Yahoo Finance' website.
We analyze the headlines of 25 leading articles (25 headlines per day) from 'Reddit WorldNews Channel' (/r/worldnews).According to these articles, we want to know whether the weighting of all of them gives us a positive or negative index regarding the increase/decrease in the value of the stock.

The values we receive from the headlines analysis (with NLTK library) are : Subjectivity, Objectivity, Positive, Negative, Neutral.
The dataset has historical news headlines from 'Reddit WorldNews Channel'. They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date (Range: 2008-06-08 to 2016-07-01).

Results

First Model: Linear regression, we predict the values with 0.0101 loss.

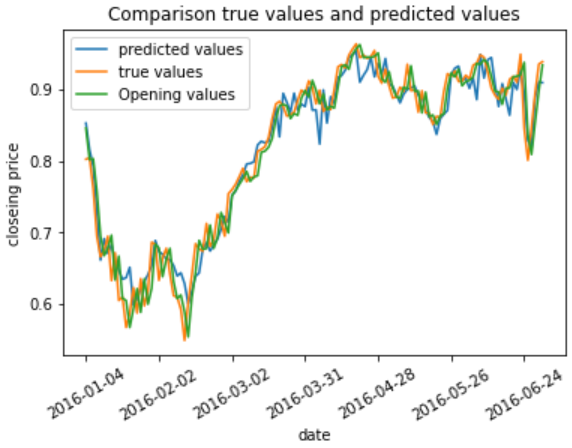
Second Model: Neural Network with single hidden layer (0.0052 loss).

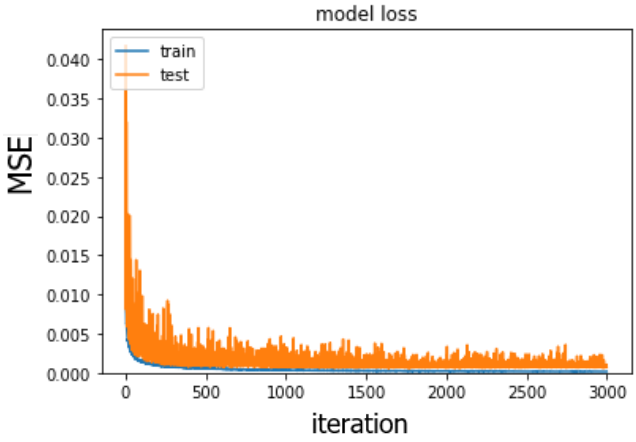
Third Model: Neural Network with three hidden layer (0.00134 loss).

Fourth Model: LSTM with MLP with hidden layers (Input -> LSTM -> Dropout -> LSTM -> Activation -> Dense -> Output) (0.000142 loss, after 2976 iterations).

```
Model: "LSTM with MLP"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_4 (LSTM)                (None, None, 200)         310400
_________________________________________________________________
dropout_2 (Dropout)          (None, None, 200)         0
_________________________________________________________________
lstm_5 (LSTM)                (None, 150)               210600
_________________________________________________________________
activation_2 (Activation)    (None, 150)               0
_________________________________________________________________
dense_4 (Dense)              (None, 15)                2265
_________________________________________________________________
dense_5 (Dense)              (None, 1)                 16
=================================================================
Total params: 523,281
```

(Fourth Model)





Conclusion

The system used for stock price prediction combining the long short-term memory for technical analysis and sentimental analysis for fundamental analysis successfully gave good accuracy.
Our significant conclusion is that it is possible to predict stock value (actual value price or increase/decrease), Through the analysis of headlines of (news) articles and historical data of the stock.
Our best way to predict the stock value (after 4 modules) is through LSTM prediction model with MLP.