Big Data in Finance (PD3-16)

LIU, Zhaoyu  
Professor Daniel Palomar

Introduction
Sentiment analysis has been applied to financial news in recent years. However, these sentiment analyses only classify news as positive or negative sentiment scores and didn't predict the stock trend. This project aims to use financial news as the only data source for the stock trend prediction to find the relationship between a piece of financial news and the stock trend.

The differences between this project and the previous works are the following:
- Instead of sentiment analysis, this project uses news data to predict stock trend.
- This project collects data from 100 companies and builds a data set of 968 pieces of financial news.

Methodology
This project used Multi-layer neural network model and the Long Short Term Memory (LSTM) model for training and testing. Due to the specialties of the two models, their input data are of different formats. Thus, we used different modeling methods of news for these two neural network models.

One-hot and Multi-layer neural network model
For multi-layer neural network model, the modeling of the test of news is the one-hot model.

Evaluation & Result
In real world, we want to know when investors should invest in stock market and when to sell their stock shares. Thus, given a period of time, we are interested in the time when the maximum value occurs. We could invest before the time when maximum occurs and sell shares on or after that. Another situation is that we are interested whether the stock price will go up or down.

Therefore, we built our evaluation method to calculate the accuracy that our model predicts the time when maximum value occurs. There is a simple function in Tensorflow that could do this job, the argmax() function.

We performed 3 tests on both of the two neural network models, to predict the maximum value on a 30-day basis, 5-day basis and the direction of stock movement for one day. The results for the multi-layer neural network model and the LSTM model are showing below:

<table>
<thead>
<tr>
<th>Model</th>
<th>30-DAY Basis (maximum)</th>
<th>5-DAY Basis (maximum)</th>
<th>One day direction (pass / neg)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-layer</td>
<td>0.04</td>
<td>0.215</td>
<td>0.55</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.05</td>
<td>0.333</td>
<td>0.55</td>
</tr>
<tr>
<td>Random guess</td>
<td>1/30 = 0.033</td>
<td>1/5 = 0.2</td>
<td>1/2 = 0.5</td>
</tr>
</tbody>
</table>

Testing results
In conclusion, the multi-layer has similar accuracy with random guess. Which means the pair of one-hot model and multi-layer neural network model is not very efficient at predicting stock trend. While the pair of Glove Word2Vec model and LSTM model shows a significant increase in accuracy on a 5-day basis test compared to random guess. And the one day prediction is also better than the random guess. This result shows