Predicting Stock Market Movements Using Public Sentiment Data and Sequential Deep Learning Models

Team: Junge Zhang, Yichao Shen, Lyuang Fu, Zhewen Song
Mentor: Josua Krause, PhD

Background and Motivation

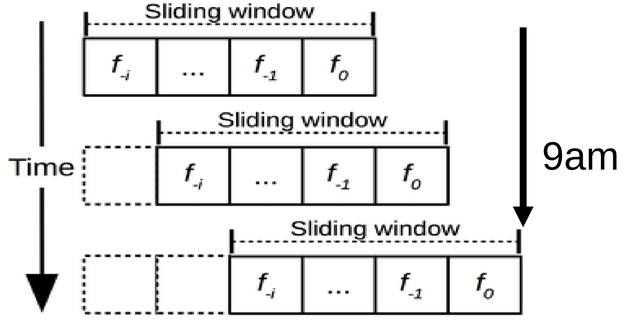
- It is a general interest to predict stock market movement which changes erratically every day.
- This project aims to use news analytics to forecast the stock market with time series data from news and social network (including sentiment, impact of news stories, etc.).

Data Setup

- Data cover 19 stocks from 2014 to 2018.
- 141 numerical features from news or social network and 6 numerical features from the stock market.
- Target is the open to open price (9 am – 9 am) change (in percentage) between the next two consecutive days.

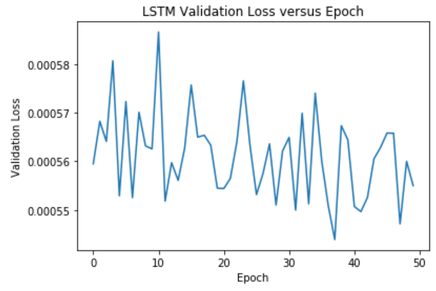
Data Preprocessing

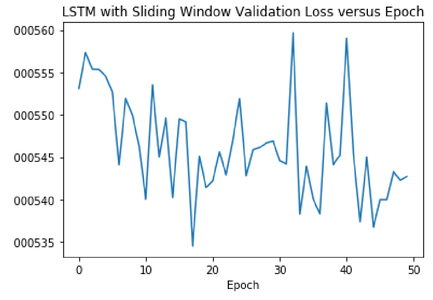
- **Aggregation:** Aggregate hourly data into daily data for baseline model (only used in baseline model).
- **Sliding window:** Use a sliding window of 24 hours to crop previous day's data and add more data in the neural network.
- **Clipping:** Clip top and bottom data that are out of range of 3 standard deviation to reduce the impact of outliers.
- **Noise:** Add Gaussian noise in the data to prevent the model from overfitting.

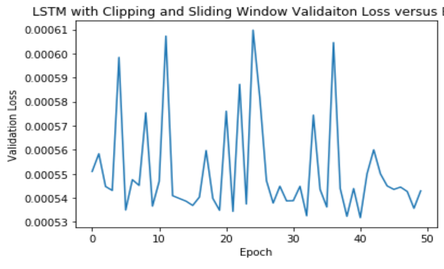


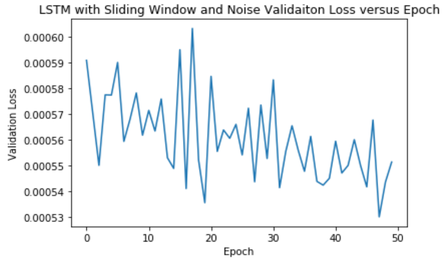
Model

- **Baseline:** Lasso Regression.
- **LSTM:** Long short-term memory based recurrent neural network which generally has good performance on time series data.
- **Customized piecewise loss function:** The training loss for LSTM is **mean squared error** when the prediction and the target has the **same sign** while it is **four times mean squared error** for **different signs.** The piecewise loss aims to penalize more when the prediction is wrong in stock price increasing or decreasing.







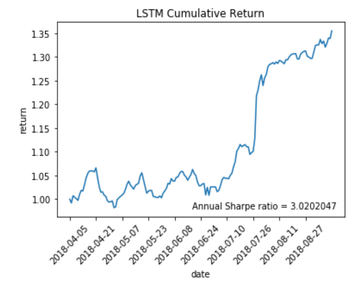


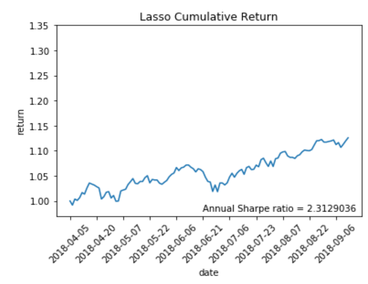
Evaluation Metrics

- **MAE**: Mean Absolute Error.
- **Direction accuracy:** Computed by the accuracy on whether the prediction and the target values have the same signs.
- **Sharpe ratio:** Financial measurement that indicates the return of investment with respect of risk. It is computed by the ratio between the mean and the standard deviation of the daily return. Normally, a Sharpe ratio over 2 is considered as good.

Results

	MAE	Direction Accuracy	Sharpe ratio
LSTM	0.0094206	0.5309211	0.9847754
LSTM Clipping + Window	**0.0093911**	0.5654605	1.8502489
LSTM Noise + Window	0.0094107	0.5093421	0.1014509
LSTM Window	0.0093949	0.5404605	**3.0202048**
Lasso	0.0096155	**0.5716247**	2.3129036





Conclusion

- Lasso hit a good Sharpe ratio. However, Lasso benchmark can only predict same positive values. It works out coincidentally in this time frame but that strategy would not generalize very well. The time frame of the data is in a bull-market so just holding stocks already gives easy returns. However, this does not always hold true.
- Currently with limited amount of data, the LSTM model shows its potential with reduced MAE and good Sharpe ratio. However, the direction accuracy needs to be further optimized.

Further Study

- Train with more data.
- Find better way to solve data skewness problem.
- Discover a model or a training metric that can better optimize the direction accuracy and predict stocks in the context of inherent time series data.