2020 Fall EE5183 FinTech - Homework 4

Deep learning Model: Recurrent Neural Network

Due: Jan 7, 2021

INSTRUCTIONS

1. In this homework, the dataset is the daily historical data of the S&P 500, which is from Yahoo Finance. We will use this to build a regression model. The features are Date, Open, High, Low, Close, Adj Close, Volume which are common attributes for investors. We want to predict the 'Close' value of the next day based on historical data. RNN is usually used to process time series data because it can capture relationship between sequences.
2. Please install the following 2 packages for this homework, *mpl_finance* and *ta_lib*. Sometimes *ta_lib* may occur error if using pip install. You can go to this website to download whl file, and then install it as follows.
   Ex: `pip3 install TA_Lib0.4.17cp36cp36mwin_amd64.whl`
3. **Please only use TensorFlow, PyTorch, Keras or scikit-learn to build the model.**
4. **You should write your own codes independently. Plagiarism is strictly prohibited.**
5. **YOU MUST TURN IN hw4_STUDENT_ID.pdf FILE so that TA can score your homework.**
6. **ONLY *.py FILES ARE ALLOWED! NO *.ipynb. It is ok to wrap your py files in bash script.**
7. Report can be written in English or Chinese.

PROBLEMS

1. Regression:

In this exercise, you will implement a RNN model for regression using *S_P.csv*. The purpose of this exercise is to create and train a neural network to predict the 'Close' value of the next day. You need to split the data from 1994-2017 as training part and the data from 2018-2019 as validation part.

(i) (20%) Please use *S_P.csv* to plot
a.) Candlestick chart with 2 moving average line (10 days and 30 days).
b.) KD line chart.
c.) Volume bar chart.
Show your figures from 2019/1/1 to 2019/12/31. (Fig. 1 is an example of the year of 2018)
(Hint: You can use *mpl_finance* and *ta_lib* package to help you plot these stock charts.)

(ii) (15%) Please at least add 4 features from question (i) into your input which are 'Moving Average 10 days', 'Moving Average 30 days' and 'K,D from KD line chart'. And we want all features except Date to be normalized on a scale of 0 to 1 by the below equation. You can also add other features to help your model get better performance (e.g. If you think weekdays are important to stock price, you can add an one-hot attribute of weekday. Please discuss what did you do for data preprocessing.

$$z_i = \frac{x_i - min(x)}{max(x) - min(x)}$$

(iii) (10%) In RNN model, data dimension can be confused. Like the figure 2, RNN has 3 dimension which can be written as (*batch_size*, *time_step*, *input_dimension*). In this exercise, You can choose the *batch_size* on your design. The *time_step* should be 30 because we want to use the last 30 days to predict the 'Close' value of the next day. And *input_dimension* will be depend on your (ii) design.

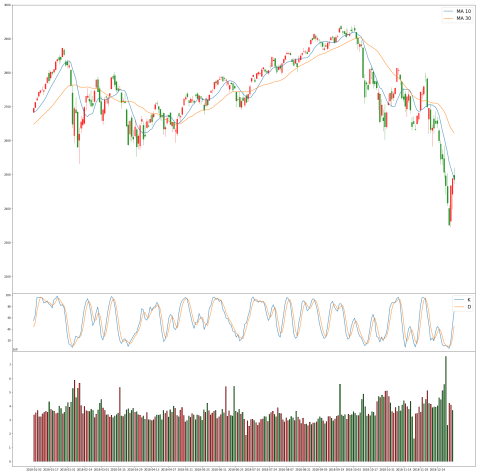

Figure 1: Example of Candlestick chart with moving average lines, KD line chart, volume bar chart. (The figure is just an example)

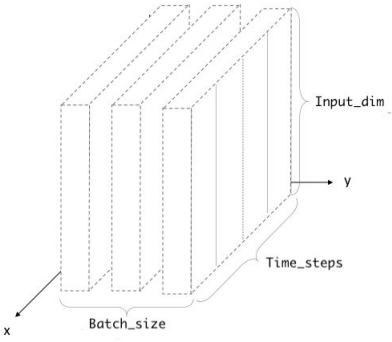

Figure 2: RNN input dimension.

(iv) (10%) Please construct a RNN model with *SimpleRNN* cell for predicting the 'Close' value of the next day according to mean square error

$$MSE = \frac{1}{n}\sum(y_i - \widetilde{y_i})^n$$

Please explain how do you design your model?

(v) (10%) Plot loss curve chart and the prediction of 'Close' value in validation part.

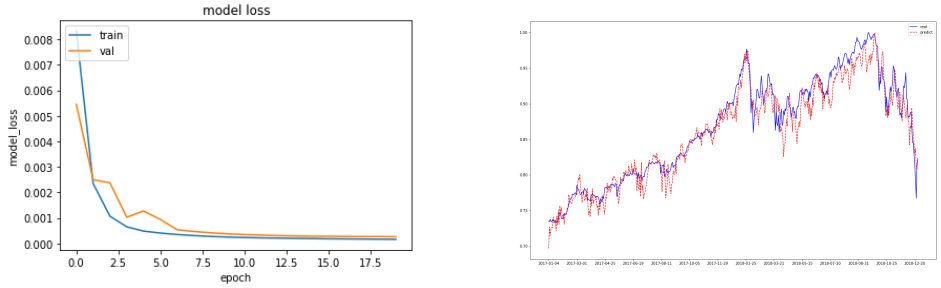

Figure 3: Example of loss curve and predict curve. (The figure is just an example)

(vi) (10%) Substitute *LSTM cell* for *SimpleRNN* and repeat (iv), (v).

(vii) (10%) Substitute *GRU cell* for *SimpleRNN* and repeat (iv), (v).

(viii) (5%) Discuss your findings from (iv) to (vii)?

(ix) (10%) Test the model on 2020 data. Does it perform well under COVID-19 conditions? (show results)

(x) (Bonus 10%) Follow (ix), how to improve the performance of the model? (you can explain your ideas or implement improved models, the latter one get more points.)