2022 Fall EE5183 FinTech - Homework 2

Deep Learning Model: Recurrent Neural Network, BERT

Due: Nov 01, 2022

INSTRUCTIONS

1. Please only use PyTorch and scikit-learn to build the model.
2. You should write your codes independently. Plagiarism is strictly prohibited.
3. Report can only be written in English
4. All the figures are just examples. You do not need to be the same as the figures.
5. You must turn in hw2_student_ID.pdf and hw2_student_ID.ipynb and zip the files as hw2_student_ID.zip. TAs will grade your python code by Google Colab. Please ensure your code can be run on the Google Colab GPU environment. The wrong format will not be graded.
6. Please install the following two packages for this homework, *mpl_finance*, and *ta_lib*. Sometimes *ta_lib* may cause errors if using pip install. You can install it as follows.

```
url = 'https://launchpad.net/~mario-mariomedina/+archive/ubuntu/talib/+files'

ext = '0.4.0-oneiric1_amd64.deb -qO'

!wget $url/libta-lib0_$ext libta.deb

!wget $url/ta-lib0-dev_$ext ta.deb

!dpkg -i libta.deb ta.deb

!pip install ta-lib
```

PROBLEMS

In this homework, the dataset is the daily historical data of Apple Inc., which is from Yahoo Finance. We will use this to build a regression model. The features are Date, Open, High, Low, Close, Adj Close, and 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 relationships between sequences.

1. Regression: Stock Price Prediction

In this exercise, you will implement an RNN model for regression using APPLE.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 2011/01/01-2012/12/31 as the training part, the data from 2013/01/01-2013/06/30 as the validation part, and the data from 2013/07/01-2013/12/30 as the test part.

(i) (20%) Please use APPLE.csv to plot

a.) Candlestick chart with two moving average lines (10 days and 30 days).

b.) KD line chart.

c.) Volume bar chart.

Show your figures from 2011/01/01 to 2013/12/30. (Fig.1 is an example of the year 2011)(Hint: You can use *mpl_finance* and *ta_lib* package to help you plot these stock charts.)

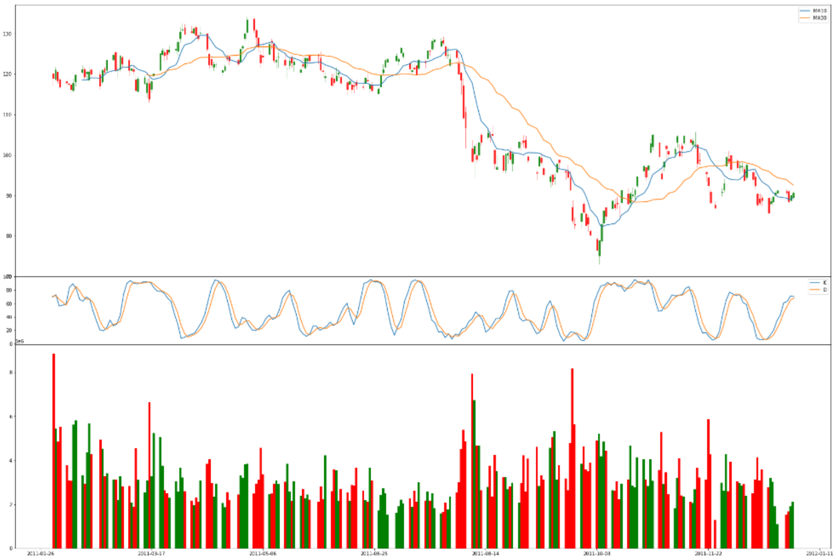

Figure 1: Example of Candlestick chart with moving average lines, KD line chart, volume bar chart. (The figure is just an example)

(ii) (15%) Please at least add four 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 weekdays.) Please discuss what you did for data preprocessing.

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

(iii) (10%) In the RNN model, data dimensions like figure 2, RNN has three dimensions which can be written as (batch size, time step, and input dimension). In this exercise, You can choose the batch size of 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 depend on your (ii) design.

(iv) (10%) Please construct an RNN model with a Vanilla RNN cell for predicting the ’Close’ value of the next day according to the mean square error.

$$MSE = \frac{1}{n}\sum(y_i - \tilde{y}_i)^2$$

Please explain how you design your model.

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

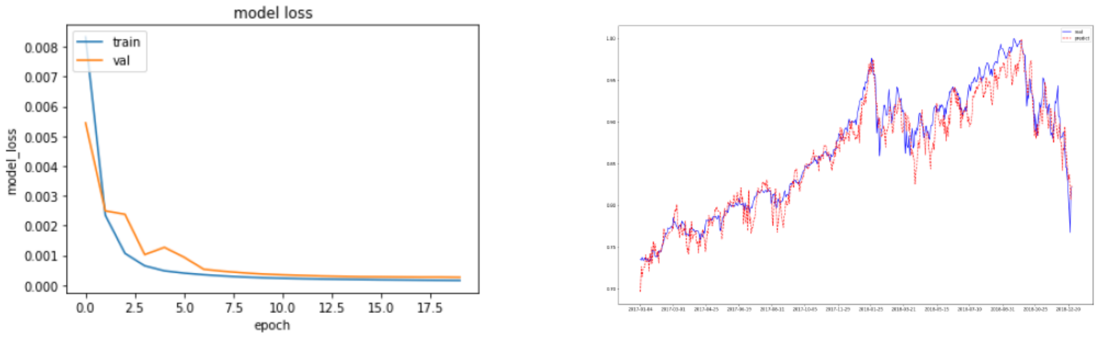

Figure 3: Example of loss curve and predict curve. (The figure is just an example)

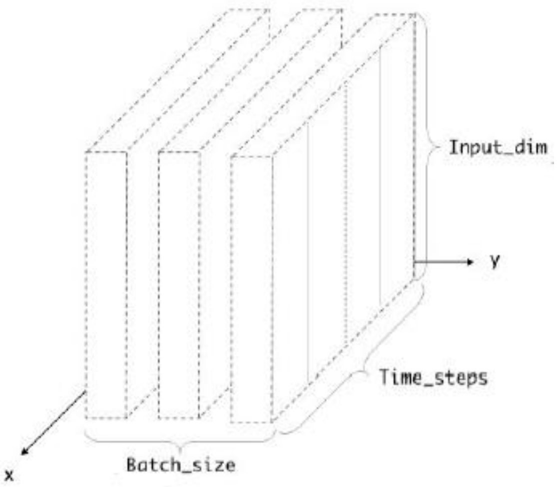

Figure 2: RNN input dimension.

(vi) (15%) Substitute LSTM cell for Vanilla RNN and repeat (iv), (v).

(vii) (15%) Substitute GRU cell for Vanilla RNN and repeat (iv), (v).

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

2. (Bonus 15 %) Daily News for Stock Price Prediction

The news could be one of the most important factors in stock market prediction. It is challenging to monitor news from many sources in real-time. We provide a dataset ‘APPLE_news.pkl ’, historical news of Apple Inc. from Bloomberg. The features are introduced as follows:

- Tickers: Stock symbol *Apple Inc. (AAPL)
- Date: Published date
- Title: Headline of the news
- Content: Content of the news

(i) (*Bidirectional Encoder Representations from Transformers*)

Please use the feature ‘title’ or ‘content’ in the ‘APPLE_news.pkl ’ dataset. You can concatenate daily title and content, then generate 768 dimensions embeddings using pre-train BERT first. Furthermore, you can also add the feature you used in 1-(ii) and news embedding to improve your model (You can choose one from Vanilla RNN, LSTM, or GRU). Please describe your idea, how it was implemented, and compare the results with previous models.

(Hint: You can use the Huggingface Transformers library with PyTorch.)

Reference document: https://huggingface.co/transformers/v3.0.2/