Stock Prediction with Price Trend

Project 1 for Deep Learning and Finance, Spring 2024
Deadline: 2024.04.29 22:00 (Not 23:59)

1 Data Description

In this project, we provide you data of A-share stock price trend in two formats, one in image format and the other in numerical format.

1.1 Image Dataset

The image dataset includes about 90,000 randomly sampled images over stock trade records in A-share market. Specifically, we include data from 2010-2012 in training set and data from 2013 in test set. All images have sizes of 180 * 96.

A typical price trend image looks like Figure 1. A stock's daily open, close, high and low price with trading volume are included in one graph. Such images were preprocessed according to method from (1), and the corresponding processed data is shown in Figure 2. Each image includes price and volume data of 64 days, and the starting price is set to 1 for all the graphs.

Note: OHLC chart for Tesla stock with 20-day moving average price line and daily volume bars. Daily data from January 1, 2020 to August 18, 2020.

Figure 1: A Tesla stock price trend image

The structure for image dataset is organized as follow:

- train: Folder of training set images, ranging from 2010 to 2012. The folder included two subdirectories: 0 and 1, which correspond to images with negative 5-day ahead return and positive 5-day ahead return respectively.
- test: Folder of test set images, including data from 2013. The folder included two subdirectories: 0 and 1, which correspond to images with negative 5-day ahead return and positive 5-day ahead return respectively.

Figure 2: An A-Share stock's processed price trend image

The path to image dataset in the server is "./data/image".

1.2 Numerical Dataset

Numerical dataset provides the same A-share stock data from 2010 to 2013, except this dataset include all trade records during the period. All data are included in file "data.csv", here is a description for each column in the file:

- code: code for stock.
- date: the trading date.
- preclose: adjusted close price on previous trading day.
- open: adjusted open price.
- high: adjusted high price.
- low: adjusted low price.
- close: adjusted close price.
- tradestatus: the trading status, usually indicates whether the stock is being traded or closed.
- volume: trading volume
- amount: trading amount

- ret: current trading day's return = (close - preclose) / preclose. Note the return is shown as a percentage number.
- 5_ret: 5-day ahead return. This is the label of task 2. Note that the task for this project is a classification problem, so you need to process the return into category variable.

The path to numerical dataset in the server is "./data/data.csv".

2 Task 1

Given the image dataset, please create a CNN-based network for classifying the 5-day ahead return of each image into positive class or negative class. The result should be the classification accuracy of your model over test dataset. Please also report the structure and hyperparameters used in your network, and analyze how they affect the accuracy.

Hint: You can use "ImagerFolder" from PyTorch to directly load the image dataset, a tutorial can be found in this link: Image Folder Tutorial. The model structures used in original paper (1) are provided in Figure 3. You can inherit their CNN network, or create your own CNN network. There is no requirement for creating a validation set, it is up to you to decide how to split your training set.

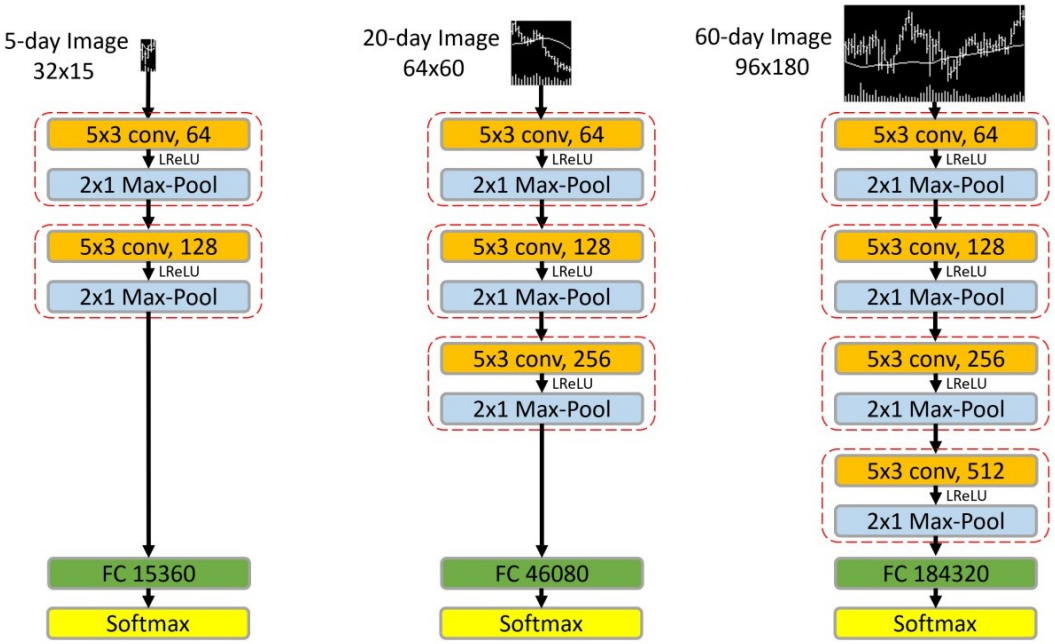

Figure 3: Networks from original paper

3 Task 2

Given the numerical dataset, please create a RNN-based network for classifying the 5-day ahead return of each trading day. The result should be the classification accuracy of your model over test dataset. Please also report the structure and hyperparameters used in your network, and analyze how they affect the accuracy.

Hint: You need to split the dataset by yourself in this task, do make sure to use data from 2013 as your test set. You need to preprocess the data into features similar to images from image dataset. Unlike in images, price and volume data have very different scales when displayed in numerical format, so you have to figure out how to process them into proper features. There is no requirement for the structure of your RNN, you can choose the popular LSTM, GRU or any other network you want.

Task 1 and 2 use stock data from same period with different methods. Although data from task 1 are randomly sampled, we want you to compare your experiments between two tasks and analyze pros and cons of each method. Include your analysis in the report under Task 2.

4 Requirement

You need to submit all codes and a short report (pdf format) with the following requirements:

- Completing and successfully running baseline networks for task 1 and 2.
- Including out-of-sample accuracy for both tasks in your report. You can also support it with tables like confusion matrix and do performance analysis with metrics like precision and recall rate. Do analysis on how different model structures, hyperparmeters and data split affect the result.
- Exploring the pros and cons from two types of models and data.

5 Attention

- You need to submit all codes and a report (at least two pages in PDF format).
- The report should show all improvement attempts (whether successful or not) and corresponding results in detail. Provide an analysis of the results.
- Plagiarism of is not permitted. If open-source code is borrowed, be sure to include a description.

References

[1] JIANG, J., KELLY, B., AND XIU, D. (re-)imag(in)ing price trends. *The Journal of Finance 78*, 6 (2023), 3193–3249.