Report

February 21, 2023

1 Part 1: missing data

(a) I've finished implementing CSDI in Tensorflow successfully, but I am still debugging the SSDI algorithm.

(b) The stock data is downloaded by a file 'download_data.py'.

(c)

(d) "If the holidays were suddenly declared by law not holidays anymore, the behavior of the investors would remain the same on the day before the holiday." For this assumption, my intuition is it does not hold in real life because I think people's investing behavior will be affected by the information whether the next day is a trading day. However, we do need to assume this otherwise there is no historical data for training to predict the price on the day before the holiday.

However, if the declaration was public for a long time, we don't need this assumption since we can remove the data before the declaration but still have enough data for training.

Other assumptions:

- If a day (for example Feb 2nd) is suddenly declared by law as a holiday, the behavior of the investors would remain the same on the day before the day(Feb 2nd).
- Every individual stock in the stock set should be viewed equally. Every interval of the same length should be viewed equally.
- Recent data matters more than distant ones.
- Historical trading data affects stocks' prices.

There are two methods to test these assumptions. One is a statistical method, the other is based on prediction models. The following description of the two methods is based on the holiday assumption. (The assumption appears in the question description.)

First, the statistical method uses data from the last k days. For a better description, let Feb 2nd denote the holiday which was suddenly declared by law not a holiday anymore. We take the average data of the last k days before Feb 1st, and then calculate the difference ratio between the average data and trading data on Feb 1st. We compare the difference ratio before and after the declaration and quantify the difference. If it's large, the assumption does not hold. Otherwise, it holds.

The second method is based on prediction models. A prediction model which predicts the trading data on Feb 1st is trained, and then we compare the prediction accuracy of data before and after the declaration. The assumption holds if the accuracy is similar.

(e) Objective function: the difference between the real added noise and the predicted noise.

(f) There are many missing data because of holidays, something happens to some companies, or some mistakes from the data providers. My idea to deal with missing data is to use interpolation. The way to test the validity is similar to the above-mentioned statistical method.

(g) Yes, we can use the missing data imputation method for the model, but we need to add a column for the price change and impute the predicted day. The answer is based on every input being a matrix of single stock historical data with the last row being masked. However, if we use the information of all stocks as input, we would like to make use of all the historical data for example to predict Europe market, the current-day data of the Hong Kong market is available, and we can use the Hong Kong market data and partially Europe market data of current day for USA stock prediction. We can also adopt Alcaraz(22) model for this task, but we need to be careful to make full use of past information and prevent using future information.