Homework #5

STAT430 Spring 2019

Due date: April 14 (firm), 2019

(Please read the Homework Policy before you start)

Description: In this homework, you will practice using convolutional neural network to account for the cross-sectional and short-term temporal dependence among multiple time series. Meanwhile, you will practice how to construct features, labels, and data generators for handling relatively larger datasets.

Objective: Predict average price movement direction in the next minute for SPY (SP500 ETF), based on the close prices and volumes in the previous hour (60 minutes) of SPY and 20 other major stocks provided.

Dataset: This is one-minute time bar data of SPY and 20 largest component stocks of SPY. The data has been pre-processed from the tick data provided by IEX, and has been provided as is and for your homework only.

Practices:

1. There are 251 trading days in total from the dataset. Choose 1 to 150 as training set, 151 to 200 as validation set, and 201 to 251 as test set.
2. Prepare features and labels:
    - The first observation on each day takes all the close prices and volumes of all the 21 tickers during minutes 1~60 as features, and the price movement direction (0,1,2) on minute 61 as the label
    - The price movement direction on minute 61 is calculated based on the comparison between the following two values and a threshold $r = 0.08$, similar to the LOB example discussed in class:
        - average close price for minutes 1~60 vs average close price for minutes 61~120
    - The features and labels are created each day according to the above procedure. Therefore, for each day there are 271 observations with labels taking the price movements directions on minutes 61~331, respectively.
    - [Hint 1] Be careful about missing values of prices and volumes, and we should use appropriate methods to impute the missing values, respectively.
    - [Hint 2] After creating the labels, the following are distributions of the labels. Before moving on to the next step, you should make sure that your obtained numbers are roughly the same as follows:
        - the numbers of labels for 0, 1, 2 (decrease, stable, increase) are:
            * 10565, 14398, 15687 for the training set;
            * 3257, 5563, 4730 for the validation set;
            * 3545, 5173, 5103 for the test set.
3. Create appropriate data generator so that a large dataset like this one can be fed into your models.
4. Construct convolutional neural network appropriately to conduct classification of these 3 classes; do not use recurrent neural network for this homework, but you can try any techniques you have learned so far to increase the performance. The following components should be included in your analysis:
    - missing values are appropriately dealt with
    - variables are appropriately scaled
    - use data generators for model fitting and evaluation
    - use some callback functions based on your needs
    - the history plot containing at least acc and val_acc should be included
5. Try different techniques you have learned so far to tweak the model so that the following criteria are met:
    - accuracy based on the validation set $> 47\%$

- accuracy based on the test set $> 39\%$
- If you cannot achieve the above criteria, please use the `acc_lucky()` function discussed in class to show that the accuracy based on your model is better than the 3 types of guesses.
  - A blog post about classification accuracy

6. To be submitted:

   - The Rmd file of your work
   - A pdf print-out of your Rmd file

Your grade for this homework will be based on how much you follow the homework policy, the completeness of the practices, the appropriateness of the methods / models, and the particular criteria listed above.