Assignment 3: Discriminative Learning and Support Vector Machines

Due: May 13, 2022 at 11:59 pm

1 Instructions

The answers to the questions and the code should be submitted on GradeScope by 13 May 2022, 11:59 pm. You may submit the notebook downloaded as PDF, but please make sure that the questions are clearly segmented and labelled. You are encouraged to use the submission template provided. To secure full marks for a question both the answer and the code should be correct. Completely wrong (or missing) code with correct answer will result in zero marks.

2 Data and Preprocessing

1. (5 points) For this question you will need the "yfinance" and the "ta-lib" libraries.
   Using the yfinance library download daily Microsoft stock data with maximum history.
2. (10 points) Compute the following indicators using the pandas and ta-lib libraries:
   - Lagged High price
   - Lagged Close price
   - Lagged Low price
   - Bollinger bands using lagged close price with timeperiod=20
   - RSI using lagged close price with timeperiod=14
   - MACD using lagged close price with fastperiod=12, slowperiod=26, signalperiod=9
   - Momentum using lagged close price with timeperiod=12
   - OBV with lagged close price and lagged volume
   - ATR with lagged high, low, and close price
   - Continuously compounded returns of the Open price
   - CCI using lagged high, low, and close price.

Do not forget to drop the NA values.

3. (5 points) You want to long the stock if the return is greater or equal to 1 percent, short the stock when the stock return is less than or equal to -1 percent, and do nothing if the return is between -1 and 1 (exclusive). Create the label where

$$label_t = \begin{cases} 1 & \text{if} \quad R_t \geq 0.01 \\ 0 & \text{if} \quad -0.01 < R_t < 0.01 \\ -1 & \text{if} \quad R_t \leq -0.01 \end{cases} \tag{1}$$

4. (5 points) Create the feature space X and label vector y. Your feature space should exclude the following variables: 'Label','Returns','Open','Close','Volume','High','Low','Dividends', and 'Stock Splits'

5. (5 points) Create the train and test subsets. The training set will inlclude all data points except the last 30 days of the sample. The test sample will include the last 30 days of the sample.
You will need to fit a MinMaxScaler from the sklearn library to your train and test feature space *separately*. Therefore, you need to call the MinMaxScaler for both your training set and your test set.For the traning set the MinMaxScaler will use only the information contained in the training set (*hint:use fit_transform*). For the test set the MinMaxScaler will transform the test set using information in the training set *hint: use transform*.

3 Discriminative Learning

6. (30 points) Use Logistic Regression with solver=liblinear' to classify the label. Tune hyper-parameters 'penalty' and 'C' using GridSearchCV implementation and Time Series Split with n_split=5, test_size=1, gap=0. The grid search should search over: penalty: 'l1','l2' and C:[0.1, 0.5, 1, 2, 3, 4, 5, 10]

   Report the selected parameters, and accuracy on the training and test data.

4 Support Vector Machine

7. (30 points) Use Support Vector Machine to classify label. Tune hyper-parameters 'C' and 'kernel' using RandomizedSearchCV implementation and Time Series Split with n_split=5, test_size=1, gap=0.The randomized search should search over: 'C':[0.3,0.5,1,5,10,20,30,50,100] and 'kernel':['linear','rbf']. You will use the scoring method 'f1_macro' in the cross-validation.
   Report the selected parameters. Report the accuracy, precision, and recall on the test data.

8. (10 points) Show a graph that compares the cumulative sum returns of the logistics and SVM strategy returns to the market buy and hold returns.
   Which strategy is best? Is there a period one strategy is better than the other?

9. (10 points) Do you think these results are dependent on the cross-validation method and scoring? Do you think that if you change the scoring method you would get different results? *Hint: think of the mathematical formulas used for scoring in the Randomized Search CV and Grid Search CV. For SVM we used f1_macro.*

10. (10 points) **Bonus:** Provide supporting empirical evidence to your answers in question 9. **Hint:** Run additional models with your suggested CV and hyper-parameter selection. You can show a graph comparing your new method's performance relative to your answers in question 8.