Spring 2023, 958:588 Financial Data Mining Homework 3.

DUE: Monday, April 17th at midnight

INSTRUCTIONS: Submit all files on Canvas. For the free response parts, type your solution into an electronic document or take a picture of your solution and make sure that all hand-writing is legible. For the code parts, submit the completed files on Canvas. **Email and physical submissions will NOT be accepted.**

For code that imports other files such as csv files or R files, make sure that all the imported files are in the same directory as the code.

NOTE: It is required that you use R or Python in this homework.

NOTE: Your code must run without error. Points may be deducted if there is any error that causes the code to crash.

NOTE: You may work in groups with other students but you must write down the names of all the members of your group. You must write and submit your own solution.

Problem 1: Kernel Ridge Regression

Download either the R or Python version of `hw3_problem1` and `Boston.csv` and follow the instructions within. In this problem, we will implement Kernel Ridge Regression (KRR) with Gaussian kernel

$$K(X_i, X_{i'}) = \exp\left(-\frac{\|X_i - X_{i'}\|_2^2}{h^2}\right),$$

where $h > 0$ is the bandwidth that we will select via cross-validation.

Part a (code)

Complete the R or Python file `hw3_problem1` by filling in all parts labeled "FILL IN" in the code.

Part b (free response)

Examine the value of the matrix `mean_valid_errs`, whose (j, k)-th entry is the mean (across folds) validation error of bandwidth j and lambda k among the candidate sets `bandwidth_ls` and `lambda_ls`.

Use your observation to justify the importance of selecting both the bandwidth h and the regularization parameter λ in cross-validation.

Problem 2: Stock Prediction

Download the R or Python file `hw3_problem2` and `sp500_long.csv`. In this problem, we will predict the daily log-return of JPM stock using the *past* log-returns of 5 companies, including JPM.

Part a (code)

Complete the code by filling in all parts labeled "FILL IN". Report the final errors achieved by ridge regression as well as the baseline error.

Part b (free response)

Why can't we use cross-validation for this prediction problem?

Part c (free response)

Increase the value of the "lag" variable to 10, then to 15, 20, 25, and 30. What effect does this have on the covariates X_i? What effect does this have on both the in-sample and the test regression errors?

Problem 3: AdaBoost

In this problem, you will implement adaBoost using decision stumps as the weak learners. Download either the R or Python file `hw3_problem3`, `movies_hw3.csv`, and `sp500_long.csv`.

Part a (code)

Complete the definition of `decisionStumpClassifier` by filling in all parts labeled "FILL IN" in the function. You may use the small block of code right underneath the function definition to test the correctness of your implementation.

Part b (code)

First, complete the definition of `adaBoost` by filling in all parts labeled "FILL IN".

Part c (free response)

The code then applies adaBoost to predict whether a movie is rated as above average. A plot is produced by the code; what does the plot show? What does the plot say about how sensitive adaBoost is to the choice of the number of iterations M? Report also the final predictive error of adaBoost.

We next use adaBoost on the stock dataset in Problem 2 and use it to predict either a company's stock is going to go up or go down. Inspect the plot produced. What does the plot say about how sensitive adaBoost is to the choice of the number of iteration M?