DS 5220: Supervised Machine Learning and Learning Theory Midterm

Due 2024/11/10 11:59pm

Instructions: The midterm is available from 2024/11/06 12:01am until 2024/11/10 11:59pm. You can choose any 24 hours in between to finish the exam. The exam solution must be submitted in one **PDF** file, including Jupyter Notebook printouts. To submit the exam, please upload your solution and click submit quiz on Canvas.

The exam is open book, open notes. You can refer to any materials posted on the course webpage or Canvas. However, you are **not allowed** to use ChatGPT or search the Internet to complete the questions. You must finish the exam individually. You are not allowed to talk to anyone about the exam until 2024/11/11.

You are expected to abide by the Northeastern University Honor Code. Any type of academic misconduct is not allowed, which includes 1) receiving or giving information about the content or conduct of an examination knowing that the release of such information is not allowed and 2) plagiarizing, whether intentionally or unintentionally. For the activities that are considered to be academically dishonest, refer to the Honor Code: https://catalog.northeastern.edu/handbook/policies-regulations/academic-integrity/

Problem 1 (Quiz questions) [30pts]

Write down the answer to the following questions. Construct the answer concisely. As a rule of thumb, for each part, the answer should not exceed one-half page in an A4 paper.

- [3pts] Explain the concept of generalization in machine learning. What's the relationship between training and generalization in machine learning? What's the relationship between regularization and generalization?

- [3pts] Can you write down the loss function of logistic regression for binary classification? Next, explain the relationship between logistic regression and maximum likelihood estimation.

- [3pts] Explain whether features need to be standardized before performing (a) subset selection methods (e.g., best subset selection/forward selection); (b) regularized methods (e.g., LASSO and ridge).

- [3pts] Can ridge regression be used for variable selection? Can LASSO be used for variable selection? If yes, how does the number of selected variables, bias, and variance vary with the regularization parameter λ? Explain why.

- [3pts] Describe the shape of the decision boundary of (i) logistic regression, (ii) LDA, (iii) QDA, (iv) k-nearest neighbors, and (v) Random forests.

- [3pts] Why is LASSO also called shrinkage? Can you provide a simple example of why LASSO can shrink small coefficient values to zero?

- [3pts] Explain the difference between leave-one-out cross-validation and k-fold cross-validation. In practice, what would you do to ensure the model does not overfit the training data?

- [3pts] Can you name a few performance metrics for classification? In the case of having imbalanced class labels, what metrics would you use? Explain why.

- [3pts] What is the main difference between random forests and gradient boosting? How do you decide which one to use in practice?

- [3pts] What is the difference between feature extraction and full fine-tuning for transfer learning? How do you decide which one to use? Explain why.

Currently, rental bikes are introduced in many urban cities to enhance mobility and comfort. It is important to make the rental bike available and accessible to the public at the right time as it lessens the waiting time. Eventually, providing the city with a stable supply of rental bikes becomes a major concern. The crucial part is predicting the bike count required at each hour for the stable supply of rental bikes. The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall) and date information.

For the midterm exam, we seek to predict the log of the count of bikes rented between 6pm-7pm using the input variables. There are 13 input variables in the data:

	Input variables
1	`Temperature`
2	`Humidity`
3	`Wind_speed`
4	`Visibility`
5	`Dew_point_temperature`
6	`Solar_Radiation`
7	`Rainfall`
8	`Snowfall`
9	`Holiday`
10	`Funtioning_Day`
11	`SeasonsAutumn`
12	`SeasonsSpring`
13	`SeasonsSummer`
	Output variable
15	`log_Rented_Bike_Count`

For every question in all problems,

1. Fit the model (and perform cross-validation if needed) using the training data (`SeoulBikeData_6pm_train.csv`).

2. Evaluate the model on the test set (`SeoulBikeData_6pm_test.csv`).

Problem 2 (Exploratory analysis and model selection) [28pts]

(a) [6pts] Explore the training data graphically to investigate the association between `log_Rented_Bike_Count` and input features. Which features seem useful in predicting `log_Rented_Bike_Count`? Scatter plots may be useful tools to answer this question. It is sufficient to list three features that you think are most useful. What is the sign of the association between these input features and `log_Rented_Bike_Count`?

(b) [6pts] Fit a linear model of `log_Rented_Bike_Count` on all input variables on the training data. Report the summary (e.g., coefficients and standard errors) of the fitted linear model. Which variables seem useful in predicting `log_Rented_Bike_Count` from the fitted linear model? Is the coefficient of `Temperature` aligned with your findings in (a)? Comment on your findings.

Next, we consider a linear model to predict `log_Rented_Bike_Count` using all input variables

(c) [6pts] Perform the best subset selection to select the best three-predictor model. Report the predictors selected. Report the training and test MSE obtained.

(d) [6pts] Perform forward stepwise selection to select the best three-predictor model. Report the predictors selected. Report the training and test MSE obtained.

(e) [4pts] Compare the training MSE from (c) and (d). Explain the ordering of training MSE. Next, compare the test MSE from (c) and (d). Comment on your findings.

Problem 3 (Regularization) [22pts]

In this problem, we consider a linear model to predict `log_Rented_Bike_Count` using all input variables. Define a list of λ values as

```
lambdas = 10**np.linspace(4, -6, 101)
```

Set the random state as 123 (i.e., `random_state=123`) when performing cross-validation.

(a) [6pts] Use cross-validation on the training data to find the optimal λ in the ridge regression model. Plot the coefficient of `Temperature`. Comment on how the coefficient varies with λ.

(b) [5pts] Refit a ridge regression model on the training data using the optimal λ selected by cross-validation. Report the training and test MSE obtained.

(c) [6pts] Use cross-validation on the training data to find the optimal λ in the lasso regression model. Report the coefficient of `Temperature`. Comment on how the coefficient varies with λ.

(d) [5pts] Refit a LASSO regression model on the training data using the optimal λ selected by cross-validation. Report the training and test MSE obtained.

Problem 4 (Regression trees and random forests) [20pts]

Finally, we use regression trees and random forests to predict `log_Rented_Bike_Count` using all input variables. Set the random state as 123 (i.e., `random_state=123`) when answering every part of this question.

(a) [5pts] Fit a regression tree to the training set. Plot the tree, and interpret the results. Report the training and test MSE obtained.

(b) [7pts] Explore whether pruning is useful to improve MSE. If so, determine the optimal level of tree complexity, prune the tree, and calculate the MSE of the pruned tree on training and test sets. If not, explain why.

(c) [8pts] Fit random forests to the training set. Try different numbers of features m, and report the corresponding training and test MSE. For each m, report the three most important variables. Does random forest improve upon regression trees? Explain why. (Hint: You may set `n_estimators` as 400. You may vary m from 3 to 13.)