Introduction to Data Science Assignment 4

Due time: 11:59 pm, November 23, 2023

3 questions and 100 points in total

This assignment covers the sections of regularization, tree-based methods, ensemble methods and support vector machines.

Every student is expected to complete the assignment independently. Discussion between students is encouraged, but no plagiarism will be tolerated.

Please submit your answers in a PDF file. If you want to also submit your code separately, you can pack all files, including a readme file, into a compressed file.

Question 1 (60 points)

This question pertains to the bike-sharing dataset utilized in Assignment 2. You can find the dataset description at
<https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset>,
and you can download the dataset from
<https://archive.ics.uci.edu/ml/machine-learning-databases/00275/Bike-Sharing-Dataset.zip>. We will use only the data in file <day.csv>.

Now, randomly split this dataset in half – allocate 50% for training and the other 50% for testing. Train a model using the training set, making sure to find the optimal value for a tuning parameter through cross-validation. Test your trained model with the test set. Repeat this process five times – randomly splitting the dataset five times – and report the average performance over these five splits. Tuning parameters and performance metrics are detailed below. Any other tuning parameters not explicitly mentioned can be set to default values.

For each of the following questions, fit a corresponding model that predicts ride counts based on the weather variables (weathersit, temp, hum, windspeed), excluding the date/time variables.

1. (12 points) Fit a ridge regression with λ chosen by cross-validation. Report the test MSE obtained.
2. (12 points) Fit a LASSO model with λ chosen by cross-validation. Report the test MSE obtained, along with the number of non-zero coefficient estimates.

3. (12 points) Fit a regression tree. Plot the tree and interpret the results. What test MSE do you obtain? Use cross-validation to determine the optimal level of tree complexity, i.e., the optimal value for the tuning parameter of cost complexity pruning. Does pruning the tree improve the test MSE?
4. (12 points) Fit a random forest. Specify the number of trees in the forest, and use cross-validation to determine the optimal value for the number of variables considered at each split. Report the test MSE obtained.
5. (12 points) Reflect on and interpret the results obtained from the various models. Comment on the predictive accuracy of each model in predicting ride counts based on the weather variables. Consider aspects such as the overall performance, strengths, and limitations of each model. Analyze and compare the interpretability of the models, and discuss the practical implications of the findings. Justify your comments with reference to the specific characteristics and outcomes of each modeling approach.

Question 2 (20 points)

Suppose we estimate the regression coefficients in a regularized linear model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^{\,2}$$

for a particular value of λ. For each of the following four scenarios (1) through (4) below, indicate which of (a) through (e) is correct, and clearly provide an explanation to justify your answer.

1) (5 points) As we increase λ from 0, the training RSS will:
    a) Increase initially, and then eventually start decreasing in an inverted U shape
    b) Decrease initially, and then eventually start increasing in a U shape
    c) Steadily increase
    d) Steadily decrease
    e) Remain constant
2) (5 points) Repeat (1) for test RSS
3) (5 points) Repeat (1) for variance
4) (5 points) Repeat (1) for bias

Question 3 (20 points)

A linear decision boundary takes the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$ in two dimensions. Now, let's explore a non-linear decision boundary.

1. (5 points) Use a computer to draw the curve $(1 + X_1)^2 + (2 - X_2)^2 = 4$.
2. (5 points) On your sketch, identify the points satisfying $(1 + X_1)^2 + (2 - X_2)^2 > 4$ and $(1 + X_1)^2 + (2 - X_2)^2 \leq 4$. Clearly justify your answers.
3. (5 points) If a classifier categorizes an observation as the positive class when $(1 + X_1)^2 + (2 - X_2)^2 > 4$ holds true and as the negative class otherwise, what class does the observation (0, 0) belong to? How about (-1, 1), (2, 2), and (3, 8)? Clearly justify your answers.
4. (5 points) Prove that the decision boundary above (question 3.3) is not linear in terms of X_1 and X_2; however, it is linear in terms of X_1, X_1^2, X_2, and X_2^2.