STATS 3DS3
Homework assignment 6

Dr. Jeganathan

03/27/2021

Submit to Crowdmark using the link that was emailed to you.

Due before 10:00 PM on Wednesday, April 7th, 2021.

Assignments submitted after the due date will receive a zero grade.

Answer all the questions.

Grading scheme: $\{0, 1, 2\}$ points per question, including reference section, total of 34. We will convert this to 100%.

Your assignment must conform to the Assignment Standards listed below.

- Write your name and student number on the title page. Submit the title page for Q0. There is no grade for Q0. We will not grade assignments without Q0.
- You may discuss homework problems with other students, but you have to prepare the written assignments yourself. You can mention one helper's name in the helper's section only.
- Please combine all your answers, the computer code, and the figures into one PDF file, and submit a copy to Crowdmark.
- Please use **newpage** to write a solution for each part of a question.
- Please choose the pages for each part of a question in Crowdmark.
- No screenshots are accepted for any reason.
- The writing and referencing should be appropriate to the undergraduate level.
- Various tools, including publicly available internet tools, may be used by the instructor to check the originality of submitted work.

Question 1

The following question is from **ISLR** 6.8 Exercises.

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

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_{ij}x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \leq s$$

for a particular value of s. For parts (2.1) through (2.5), indicate whihc of i. through v. is correct. **Justify your answer**.

1.1) As we increase s from 0, the training RSS (residual sum of squares) will:

i. Increase initially, and then eventually start decreasing in an inverted U shape.

ii. Decrease initially, and then eventually start increasing in a U shape.

iii. Steadily increase.

iv. Steadily decrease.

v. Remain constant.

1.2) Repeat (1.1) for test RSS.

1.3) Repeat (1.1) for variance.

1.4) Repeat (1.1) for (squared) bias.

1.5) Repeat (1.1) for the irreducible error.

Question 2

The following question is from **ISLR** 6.8 Exercises.

In this exercise, we will predict the number of applications received using the other variables in the `College` data set in package `ISLR`. Read the description of the data in the help page.

```
library(ISLR)
data("College")
```

2.1) Split the data set into a training set and a test set (50%/50%). Set seed is 2021.

2.2) Fit a linear model using least squares on the training set, and report the test error obtained.

2.3) Use forward stepwise (BIC) to perform best subset selection, and report the test error obtained.

2.4) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

2.5) Fit a lasso model on the training set, with λ chosen by cross- validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

2.6) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these four (2.2, 2.3, 2.4, 2.5) approaches?

Question 3

The following question is from **ISLR** 4.7 Exercises.

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set (in the ISLR pacakge.)

```
library(ISLR)
data(Auto)
```

3.1) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. You can compute the median using the `median()` function. Note you may find it helpful to use the `dplyr::mutate()` and `ifelse()` functions to create the `mpg01`. After this step, you may drop `mpg` from the Auto data set.

3.2) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

3.3) Split the data into a training set and a test set (50%/50%). Set seed is 2021.

3.4) Perform logistic regression on the training data in order to predict `mpg01` using the variables that seemed most associated with mpg01 in (3.2). What is the test error of the model obtained?

3.5) Interpret the estimated odds ratio associated with one of the predictor variables.

Reference

Helper's name

You can write only one name.