Homework 3 - Stat 495

YourNameGoesHere

Due Monday, Oct. 2 by midnight

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (anything besides our textbook(s), course materials in Moodle/Git repo, R help menu) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

-

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

-

PROBLEMS TO TURN IN: Additional 1-3

Additional 1 - Applications to College

Adapted from ISLR

```
data(College)
```

This data set contains information from 1995 on US Colleges from US News and World Report (see help file for details). Our goal is to predict the number of applications received (Apps) using the other variables as potential predictors.

> part a: Split the data into training and test data sets. (2/3 - 1/3 split is fine.)

SOLUTION:

> part b: Fit a "kitchen sink" linear regression model on the training set. Report the test error obtained, and comment on any issues with the model (such as conditions, etc.).

SOLUTION:

> part c: Fit a linear regression model using an automated variable selection method of your choice (from Stat 230) on the training set. Report the test error obtained, and comment on any issues with the model (such as conditions, etc.).

SOLUTION:

> part d: Fit a ridge regression model on the training set, with lambda chosen by cross-validation. Report the lambda chosen and the test error obtained.

SOLUTION:

> part e: Fit a lasso model on the training set, with lambda chosen by cross-validation. Report the lambda chosen and the test error obtained.

SOLUTION:

> part f: Comment on the results obtained across the four models. Address the following questions as part of your response. Do the test errors differ much? Do the coefficients differ greatly? In particular, if any variables were left out of the model in (c) or (e), is there any insight that they might have been removed based on the models in (b) or (d)? Which final model would you select here? Why?

SOLUTION:

Additional 2 - Soil predictions

The data comes from a Kaggle competition: https://www.kaggle.com/c/afsis-soil-properties. The original data set contained 3600 variables, 3599 possible predictors (really, 3578 and some other variables) and a response, Sand. The 3599 predictors were reduced to 106 (methods to be taught later this semester) that can "best" distinguish between the two levels of Depth (another variable in the data set). The resulting data set of 107 variables (106 predictors and the response variable, Sand) was saved in the data set "newsoil". The row numbers should be removed as demonstrated below.

```r
newsoil <- read.csv("https://awagaman.people.amherst.edu/stat495/newsoil.csv",
                    header = T)
newsoil <- select(newsoil, -X)
```

Our focus is on predicting the response variable Sand, using the selected variables from previous work.

> part a: Split the data set into a training and test set (75/25), with a seed of your choice. You may also wish to create appropriate x and y matrices for future function inputs at the same time.

SOLUTION:

> part b: Fit lasso models to predict Sand using all the possible predictors. Choose two lasso models - one that has a "best" lambda determined in some appropriate way, and another model with a different non-zero lambda of your choice. How many slope coefficients are set to 0 in each of your chosen lasso models?

SOLUTION:

> part c: Fit a ridge model to predict Sand using all the possible predictors. How many slope coefficients are set to 0 in your ridge model?

SOLUTION:

> part d: Compute test MSEs for both of your lasso models and your ridge model.

SOLUTION:

> part e: Write a few sentences to address the following questions. Does the test MSE from the model with the "best" lambda suggest it is in fact a better predictive model than your other lasso model? Is ridge better than the lasso models? Which final model would you choose here from these three models? Why?

SOLUTION:

> part f: What is the default setting for the normalize option in lars and the standardize option in glmnet? Why is this setting important to the model fit?

SOLUTION:

> part g: Explain what option you would change in order to fit an elastic net penalty (not OLS or ridge) using glmnet.

SOLUTION:

Additional 3

After reading through your portfolio reflections, I decided to try to adapt some class activities and assignments to better align with your goals. This is a little bit challenging because they are quite varied. However, the *College* data set that we used for Additional 1 does permit a host of different analyses, so we're going to try using it to help with this.

For this problem, your assignment is to tackle some aspect of your goals for class using the *College* data set. Include your work here, in the outline below. I'll give you feedback and work to correct any statistical issues, and note that while assessing this, I'm not looking for any one specific thing from any of you. I'm including some examples of what you might do below, based on some of the goals I read.

Examples, if you said . . .

- I want to demonstrate my understanding of method X. Can you apply method X to this data set? (You may have to do some work like create a variable for example to run an ANOVA; take one of the quantitative variables and cut it into 4 groups.) Try it. Write a summary of your findings.
- I want to work on my EDA. We know the (overall) goal here is to predict Apps. What EDA would you do (we skipped it above!)? Describe it, do some of it, etc.
- I want to practice commenting my code and making visuals extremely clear. Pick a simple analysis (predict Apps with like 2-3 other predictors), make your code awesome and visuals really good.
- I want to practice writing more precisely / shorter paragraphs for my results. Pick a simple analysis (predict Apps with like 2-3 other predictors), and write your results the way you typically would. Then, leave that there, and try a revision, working to improve the writing (this way, you see the original and the revision).

Other guidance:

- Spend at most 2 hours on this question. This is designed to let you practice something you wanted to work on, not take up all your time.
- If none of your goals seem to align with this, you can pick one based on the examples I listed above, or something similar that you want to work on.
- Use your best judgement for any models you need to fit here. If you just need a model to practice writing, it is okay if that's not the overall best model you might pick. On the other hand, if you want to practice model fitting, you should be spending time discussing that and showing your work for it. We will practice the entire data analysis process in future work.

part a: What aspect of your goals for class are you going to tackle for this assignment? How?

SOLUTION:

part b: Include your work here!

SOLUTION: