Assignment 8 - Ridge Regression & Lasso

your name goes here

Due: Wednesday, March 28, 2018

Summary

We move now to computational methods for model building: Ridge Regression and the Lasso. Using ideas other than asymptotic distribution theory (e.g., F distributions), we can think about which variables are most important in a particular linear model. Note that there are open research questions pertaining to building inference for ridge regression and the Lasso.

The homework in this assignment comes primarily from the alternative textbook, *An Introduction to Statistical Learning*, `http://www-bcf.usc.edu/~gareth/ISL/`.

Assignment

1. For parts (a) through (c), indicate which of i. through iv. is correct (look up the term flexible in your book). Justify your answer.

   (a) The lasso, relative to least squares, is:

   i. More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   ii. More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

   iii. Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   iv. Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

   (b) Repeat (a) for ridge regression relative to least squares.

   (c) Repeat (a) for non-linear methods relative to least squares.

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

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

   for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

   (a) As we increase λ from 0, the training RSS 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.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

(e) Repeat (a) for the irreducible error (look up the term irreducible error in your book).

3. Consider the `NCI60` data set in the `ISLR` package. I messed with the response variable (so the example isn't very realistic). Pretend the goal is to predict the severity of disease with the 6830 gene expression vectors.

```
require(ISLR); require(glmnet)
data("NCI60")
NCI.genes <- NCI60$data
set.seed(47)
NCI.severity <- rnorm(length(NCI60$labs), mean=as.numeric(as.factor(NCI60$labs)),sd=2)
```

(a) Try running OLS on the data. What happens? Look at the lm output (including the coefficients) and explain. Try plotting or numerically summarizing your coefficients to say something meaningful. **Please do not print out lists of 6830 things!!! Make your output meaningful!**

(b) Can you use forward stepwise methods? Which gene do you think would be the first gene into the model? Note: you don't have to run forward (it's kind of a pain to write the code), just think about how you might assess which gene is most significant in a simple bivariate way.

(c) Can you go backwards? Explain.

(d) Run ridge regression on the data.

i. Find the λ value that minimizes the cross validated error.

ii. Provide plots for both the cross validated error and the coefficients as functions of λ.

iii. Did the variable(s) that you found in (b) have high coefficients? Note that the coefficient vector is 6831 long, because there is an intercept coefficient!

(e) Run lasso on the data.

i. Find the λ value that minimizes the cross validated error.

ii. Provide plots for both the cross validated error and the coefficients as functions of λ.

iii. Did the variable(s) that you found in (b) have high coefficients? Note that the coefficient vector is 6831 long, because there is an intercept coefficient!

(f) Make a pairs plot of the three sets of coefficients (all, RR, Lasso). Give some interpretation of the plot. (n.b. You should remove the intercept and then cbind the non-intercept coefficients. Also, you may need as.numeric on the ridge regression and lasso coefficient vectors.)