Statistics 151B Homework 2 Due: 2:00 PM Mon 5 Mar 2012

Theory

For this part of the homework, choose 5 out of the 6 problems. Each group should submit written solutions for those 5 problems at the beginning of section on Monday March 5th.

For extra credit, you may also submit a solution to the problem you did not choose to have graded. *You must clearly indicate which solution you are submitting for extra credit.* A correct solution gives you extra credit points, but an incorrect solution has no negative effect.

1. ESL 2.6.

   Note. A weighted least-squares problem has the form

   $$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i (y_i - x_i^\top \beta)^2$$

   for given non-negative weights $w_1, \ldots, w_n$. Informally, w_i conveys how important it is to fit y_i well, relative to the other observed y's.

2. ESL 3.5.

   Note. $\overline{x}_j$ is the mean value of the jth predictor variable in the dataset.

3. ESL 3.12.

4. ESL 3.16.

   Note. In the table, $\operatorname{rank}(|\hat{\beta}_j|) \leq M$ means $\hat{\beta}_j$ is one of the M largest coefficients in absolute value. Also, $(|\hat{\beta}_j| - \lambda)_+$ means $\max\{0, |\hat{\beta}_j| - \lambda\}$.

5. Suppose $p \gg n$ (many more predictor variables than observations), you have a design matrix X and a quantitative response vector y, and you plan to fit a linear regression model.

   (a) Explain why the ordinary least squares solution is not unique. What can you say about the residuals of any solution?

   (b) Is the ridge regression solution unique? Why or why not?

   (c) Suppose you compute a series of ridge solutions $\hat{\beta}(\lambda)$ for X and y, letting λ get monotonically smaller. What can you say about the limiting ridge solution as $\lambda \downarrow 0$?

6. ESL 4.2 (a), (c), (d). [Don't do (b), but use its result to do (c).]

   Hint. Part (d): any recoding of the response can be written in terms of the original coding as $y_i^* = a y_i + b$ for some a, b.

A regression analysis

Abalone (from the Spanish Abulón), genus *Haliotis*, are a species of shellfish (mollusks). The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope—a boring and time-consuming task. The dust created through the cutting of abalone shells is also dangerous; it can cause irritant bronchitis, other respiratory irritation responses, and allergic attacks.

It is thus of some interest to predict the age of abalone using covariates which are easier and safer to measure. The file `abalone.data` on the website contains observations of 4,177 abalone. In each case, there were eight measured covariates, and one response – the number of rings (age in years equals # rings + 1.5). The gender covariate has three levels: (I)nfant, (M)ale, and (F)emale. The other covariates are just what the names suggest.

You will study the prediction of abalone ring count from the predictors, using OLS, ridge regression, and the lasso.

What to turn in for this problem: email the following three files to the GSI and me by the stated due date and time.

1. A file containing the R commands you used. Call it `commands.txt`. This file should have only your input to R.

2. A file containing a transcript of your R session. Call it `transcript.txt`.

3. A file with the figures you generated, and a discussion of your analysis. Please remember to give specific attribution about who did what in your group. All group members are expected to contribute both to the problems and to the data analysis.

- Load `abalone.data` into R as a data frame `abalone`. Perform relevant summaries and inspections of the data.

- Split the rows of the data frame randomly into two sets: 80% of the rows in a training set, 20% in a test set. (Check the help page for the function `sample`.)

- Run a standard linear model fit on the training set. Perform relevant summaries and inspections, and generate any figures you feel are informative. Which covariates seem useful?

- Compute the mean squared error of the linear model on the test set. Use the function `predict` for this. It needs two arguments: a fitted linear model object, and the test set data frame, in that order. It returns the vector of predictions on the test set. You need to compute the average squared difference between these predicted values and the test-set values of `rings`.

- Download the file `hw2-funcs.R` from the website to your working directory. Load the R functions in that file into your session via `source("hw2-funcs.R")`.

- Run ridge regression, as follows.

- Load the MASS library with `library(MASS)`.
- Check the help page for the function `lm.ridge`, then use it to produce a sequence of ridge regression fits on the training set. Save the fits object in a variable. The vector of `lambda` values to use for the fits is `(1 / seq(1e-6, .05, 1e-4))`.
- Plot the ridge coefficient paths with the function `plot.ridge.coef` (from `hw2-funcs.R`). This function takes a fitted ridge object as its only argument.
- Suppose we know the best value of λ is 50 (i.e., $1/\lambda = .02$). Compute the mean squared error of ridge regression on the test set, using $\lambda = 50$. Use `predict.ridgelm` in `hw2-funcs.R` for this. It takes three arguments: a fitted ridge object, the test set data frame, and the desired value of λ, in that order. It returns the vector of predictions on the test set. You need to compute the average squared difference between these predicted values and the test-set values of `rings`.

- Run the lasso, as follows.
  - Install the `lars` package from the Internet, using `install.packages("lars")`.
  - Load the LARS library with `library(lars)`.
  - Check the help page for the function `lars`, then use it to produce a sequence of lasso fits on the training set. Save the fits object in a variable. The first argument to the `lars` function should be the output of the function call `model.matrix(rings ~ ., data = abalone.train)`.
  - Plot the lasso coefficient paths with the function `plot.lasso` in `hw2-funcs.R`. This function takes a fitted lasso object as its only argument.
  - Suppose we know the best value of $(|\beta|/\max|\beta|)$, the horizontal axis of the lasso coefficient plot, is 0.5. Compute the mean squared error of the lasso on the test set, using $(|\beta|/\max|\beta|) = 0.5$. Use `predict.lasso` in `hw2-funcs.R` for this. It takes three arguments: a fitted lasso object, the test set data frame, and the desired value of $|\beta|/\max|\beta|$, in that order. It returns the vector of predictions on the test set. You need to compute the average squared difference between these predicted values and the test-set values of `rings`.

- Which approach would you use to predict abalone age in the future? Why?