Stat 627/427 (Statistical Machine Learning)

Variable Selection and Shrinkage (Chap. 6)

1. (**Chap. 6, # 2, p.259)** Consider three methods of fitting a linear regression model - (a) lasso, (b) ridge regression, and (c) fitting nonlinear trends. For each method, choose the right answer, comparing it with the least squares regression:

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

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

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

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

2. (**Chap. 6, ≈# 6, p.261)** Ridge regression minimizes

$$\sum_{i=1}^{n}(Y_i - \beta_0 - X_{i1}\beta_1 - \ldots - X_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1}$$

   whereas lasso minimizes

$$\sum_{i=1}^{n}(Y_i - \beta_0 - X_{i1}\beta_1 - \ldots - X_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \,. \tag{2}$$

   Consider a "toy" example, where $n = p = 1$, $X = 1$, and the intercept is omitted from the model. Then RSS reduces to $RSS = (Y - \beta)^2$.

   (a) Choose some Y and λ, plot (1) and (2) as functions of β, and find their minima on these graphs. Verify that these minima are attained at

$$\hat{\beta}_{ridge} = \frac{Y}{1+\lambda} \quad \text{and} \quad \hat{\beta}_{lasso} = \begin{cases} Y - \lambda/2 & \text{if} \quad Y > \lambda/2 \\ Y + \lambda/2 & \text{if} \quad Y < -\lambda/2 \\ 0 & \text{if} \quad |Y| \leq \lambda/2 \end{cases} \tag{3}$$

   (b) Now choose some value of Y and plot ridge regression and lasso solutions (3) on the same axes, as functions of λ. Observe how ridge regression keeps a slope whereas lasso sends the slope to 0 when the penalty term is high.

Projects

3. (**Simulation project - Chap. 6, # 8, p.262)**

   In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

   (a) Use the **rnorm()** function to generate a predictor X of length as well as a noise vector ε of length $n = 100$ (you can refer to our first lab "First steps in R" for this command).

(b) Generate a response vector Y according to the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon,$$

where β_0, β_1, β_2, and β_3 are constants of your choice.

(c) Use the **regsubsets()** function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, \ldots, X^{10}$. What is the best model obtained according to C_p, BIC, and adjusted R^2 criteria? Show some plots to provide evidence for your answer and report the coefficients of the best model obtained.

(d) Repeat (c), using forward and backwards stepwise selection with **step**. How does your answer compare to the results in (c)?

(e) Now fit a lasso model with the same predictors. Use cross-validation to select the optimal value of λ. Create plots of the cross-validation error as a function of λ. Report the resulting coefficient estimates, and discuss the results obtained. Which predictors got eliminated by lasso?

(f) Now generate a response vector Y according to the model

$$Y = \beta_0 + \beta_7 X^7 + \varepsilon,$$

and perform best subset selection and the lasso. Discuss the results.

4. (**Real data analysis - Chap. 6, # 9, p.263**)

Predict the number of applications received based on the other variables in the **College** data set. Split the data set into a training set and a test set. Fit

(a) least squares regression
(b) ridge regression, with λ chosen by cross-validation
(c) lasso, with λ chosen by cross-validation
(d) PCR model, with M chosen by cross-validation
(e) PLS model, with M chosen by cross-validation

using the training set, then evaluate performance on the test set. For each method, report the cross-validation error.

Comment on the results obtained. How accurately can we predict the number of college applications? Is there much difference among the test errors resulting from these five approaches?