Saarland University

The Elements of Stastical Learning

Assignement 4

Due Date: 13.12.2017

Thibault SCHOWING *Mat. 2571837*
Sarah MCLEOD *Mat. 2566398*
December 28, 2017

Problem 1

Prove that for linear and polynomial least squares regression, the LOOCV estimate for the test MSE can be calculated using the following formula:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2 \tag{1.1}$$

Where h_i is the leverage (3.37, ISLR p98)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_i' - \bar{x})^2} \tag{1.2}$$

We first have this equation, that can take long if n is big because it has to fit every model.

$$MSE_i = (y_i - \hat{y}_i)^2$$

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

So knowing that $\hat{y} = Hy$ and as it is a Leave One Out cross validation, we fit n times the model with one element out. So we have:

$$H = X(X^TX)^{-1}X^T$$
$$H^{-i} = X_{-i}(X_{-i}^TX_{-i})^{-1}X_{-i}^T$$

The hat matrix with all the data and with one out, respectively
Then we have

$$\hat{y}_i = x_i^T[X(X^TX)^{-1}X^T]y$$
$$\hat{y}_{-i} = x_i^T[X_{-i}(X_{-i}^TX_{-i})^{-1}X_{-i}^T]y_{-i}$$

The fitted values at x_i when using all the data points and when leaving one out. We can then do the following:

$$\hat{y}_{-i} = \sum_{i \neq j} H_{ij}y_j + H_{ii}\hat{y}_{-i}$$

$$\hat{y}_{-i} = \sum_{j}^{m} H_{ij}y_j - H_{ii}y_i + H_{ii}\hat{y}_{-i}$$

We have $\sum_j^m H_{ij}y_j = \hat{y}_i$ so:

$$\hat{y}_{-i} = \hat{y}_i - H_{ii}y_i + H_{ii}\hat{y}_{-i}$$

We substitute $\hat{y}_{-i}$ in the prediction error:

$$y_i - \hat{y}_{-i} = y_i - (\hat{y}_i - H_{ii}y_i + H_{ii}\hat{y}_{-i})$$

$$y_i - H_{ii}y_i - \hat{y}_{-i} - H_{ii}\hat{y}_{-i} = y_i - \hat{y}_i$$

$$y_i - \hat{y}_{-i} = \frac{y_i - \hat{y}_i}{1 - H_{ii}}$$

Taking the Mean Square Error leads to equation 1.1.

Problem 2

1. Ridge regression is done by minimizing the RSS with a quadratic penalty term:

   $$minimize(y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

   to show that the solutions take on the form:

   $$\hat{\beta}^{ridge} = (X^TX + \lambda yI)^{-1}X^Ty$$

   First we expand the equation to:

   $$y^Ty - y^TX\beta - yX^T\beta^T + X^T\beta^TX\beta + \lambda\beta^T\beta$$

   which simplifies to:

   $$y^Ty - yX^T\beta^T - yX^T\beta^T + X^T\beta^TX\beta + \lambda\beta^T\beta$$

   and:

   $$y^Ty - 2yX^T\beta^T + X^T\beta^TX\beta + \lambda\beta^T\beta$$

   We take the first derivative with respect to β, which gives us:

   $$0 - 2X^Ty + 2(XX^T)\beta + 2\lambda\beta$$

   setting this to zero this can be simplified to:

   $$2(XX^T)\beta + 2\lambda\beta = 2X^Ty$$

   and further simplified to:

   $$((XX^T) + \lambda I)\beta = X^Ty$$

   where I is the p x p identity matrix (added to the matrix math works out correctly). Solving for β gives:

   $$\hat{\beta}^{ridge} = (X^TX + \lambda yI)^{-1}X^Ty$$

2.

Problem 3

Assume a scenario in which the number of observations equals the number of features ($n = p$) and X is the $n \times n$ identity matrix. Furthermore, assume that we perform regression without an intercept. In this setting, lasso simplifies to:

$$\underset{\beta}{\text{minimize}} \sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3.1}$$

Show that the lasso estimates take the form:

$$\hat{\beta}_j^{lasso} = \begin{cases} y_j - \frac{\lambda}{2}, & \text{if } y_j > \frac{\lambda}{2}; \\ y_j + \frac{\lambda}{2}, & \text{if } y_j < -\frac{\lambda}{2}; \\ 0, & \text{if } |y_j| \leq \frac{\lambda}{2}; \end{cases} \tag{3.2}$$

For equation 3.1 to be a minimum, we must derive it and equal it to zero:

$$\frac{d}{d\beta_j} \sum_{j=1}^{p} (y_j - \beta_j)^2 + \frac{d}{d\beta_j} \lambda \sum_{j=1}^{p} |\beta_j| = 0 \tag{3.3}$$

We can rewrite equation 3.1 with the matrix notation like following:

$$F(\beta) = (y - \beta)^T (y - \beta) + \lambda\beta = yy^T - y\beta^T - y^T\beta + \beta\beta^T + \lambda\beta$$

$$\frac{dF}{d\beta} = -2y + 2\beta + \lambda$$

$$\beta = y - \frac{\lambda}{2}$$

To make β negative y has to be smaller than $\frac{\lambda}{2}$
To make β positive y has to be greater than $\frac{\lambda}{2}$
To make β equals 0, $|y|$ has to be equal to $\frac{\lambda}{2}$

1. See the attached R code for the data normalization and splits.

2. We applied the best subset selection on the training set. Figure 1 shows the curves for R^2, adjusted R^2, C_p and BIC as a function of the number of predictors. As expected R^2 increases as the number of predictors increases. Adjusted R^2 and C_p both suggest that models with 7 predictors would be the best model, since these models have the highest adjusted R^2 and lowest C_p. These two also show that a model with 6 predictors will achieve about the same performance. BIC replaces the penalty term used by C_P ,$2d\hat{\sigma}^2$, with $log(n)d\hat{\sigma}^2$, where n is the number of observations. Since $log(n) > 2$ for any $n > 7$, BIC places a heavier penalty of models with more predictors. This can be observed in Figure 1, where the BIC statistic suggests a model with 2-4 predictors is best. We'll choose 7 feature model. Figure 2 shows the the selected features used for the best model for the range of predictors. In our selected model the features used are all the features except *gleason* . The **training error** for this model is 0.439. The **test error** is 0.516.

3. Figure 3 shows the values of the coefficients in relation to λ for the ridge regression fit on the training data. The results are as expected. On the left hand side λ is almost zero, therefore the ridge coefficient estimates are essentially the same as the least squares estimates. As $lambda$ increases the coefficient estimates shrink to zero. On the right hand side of the plot, when $lambda$ is large, all of the estimates are zero.

4. See code for 5-fold cross-validation for the ridge regression model. For this model the **training error** in MSE is 0.10586, and the **test error** is 0.44623.

5. Figure 4 shows the plot of the lasso coefficient estimates plotted against λ. In lasso the coefficients shrink to zero as λ gets sufficiently large, which can be seen in Figure 4. This is in contrast to ridge regression, where the coefficients approach but never shrink to zero.

6. See the code for 5-fold cross-validation for the lasso model. For this model the **training error** is 0.10342, and the **test error** is 0.4426. In ridge regression all off the coefficients are used, but the lasso coefficients are significantly smaller. In fact, many of the lasso coefficients, such as svi and lcp, are, for practical purposes, zero. Since lasso coefficients are generally a more sparse representation, since they go to zero, this is expected. The ridge regression coefficients and the lasso coefficients are:

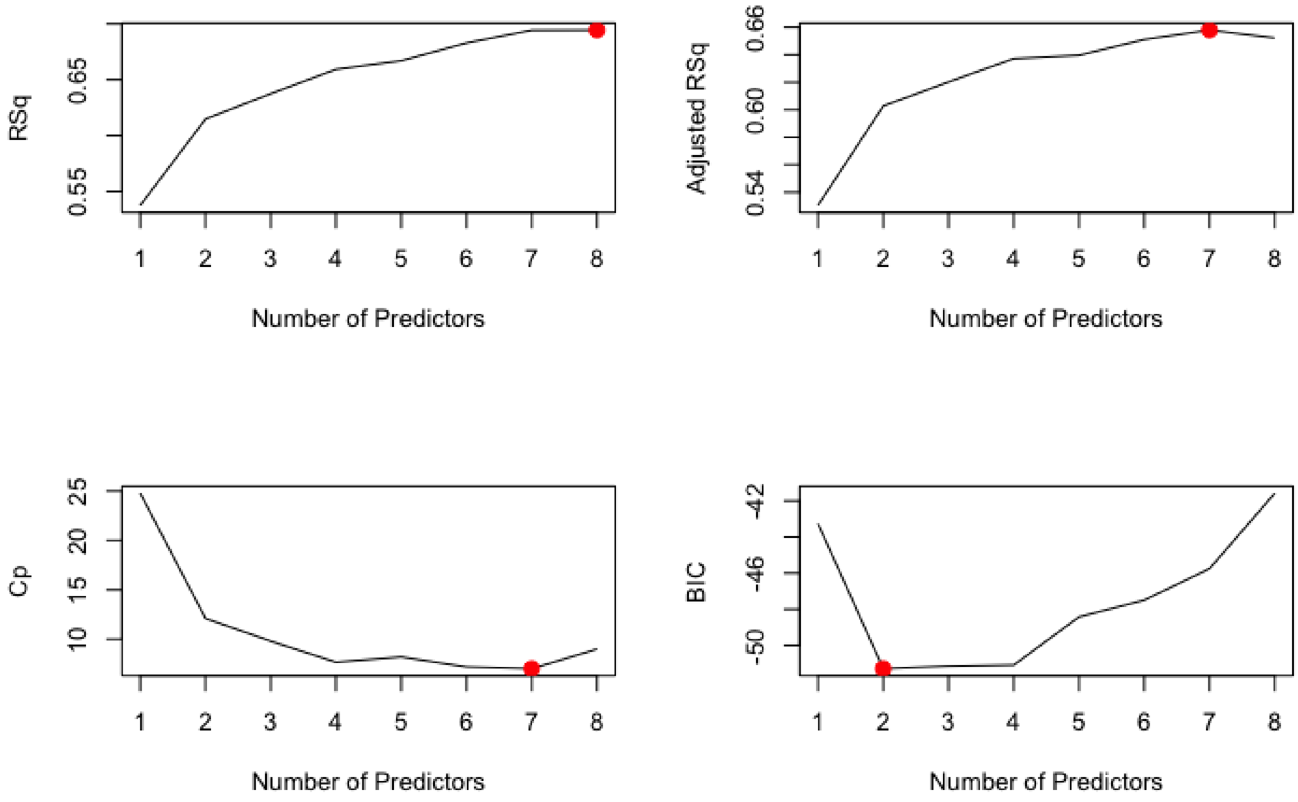

Figure 1: RSq, adjusted RSq, Cp and BIC values versus the number of predictors in best subset selection.

Features.png

Figure 2: Selected features for best model with predictors from BIC statistic.

Ridge.png

Figure 3: Ridge regression coefficients in relation to *lambda*.

Lasso.png

Figure 4: Lasso coefficients in relation to *lambda*.

	ridge	lasso
lcavol	-0.6361918	-1.157344e-04
lweight	-0.2638188	-7.258686e-05
age	0.1315373	-3.758792e-05
blph	-0.2072106	-4.337587e-05
svi	-0.3251784	-9.135083e-05
lcp	0.1615693	-8.125745e-05
gleason	-0.1387185	-5.804440e-05

7. The linear regression model trained on all the features has a higher MSE than the ridge regression and lasso models trained with cross-fold validation. The training error and test error is shown below, along with the error from the ridge regression and lasso cross-validation errors:

	train error	test error
ridge CV	0.10586	0.44623
lasso CV	0.10342	0.4436
linear	0.43919	0.5212