An Introduction to Statistical Learning: Chapter 6

Due on January 15,2016 at 3:10pm

G. James et.al.

Tony Jiang

1. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p + 1$ models, containing 0, 1, 2, . . . , p predictors. Explain your answers:
(a) Which of the three models with k predictors has the smallest *training* RSS?
(b) Which of the three models with k predictors has the smallest *test* RSS?
(c) True or False:
i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by forward stepwise selection.
ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k + 1)$- variable model identified by backward stepwise selection.
iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k + 1)$- variable model identified by forward stepwise selection.
iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by backward stepwise selection.
v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k + 1)$-variable model identified by best subset selection.

Answers:
(a). Best subset has the smallest training RSS.
(b). Best subset approach is more likely to yield smaller test RSS as it searches broader range of possible models than other two approaches. Anyway, this is not guaranteed as there is no way to tell which model will do better than others.
(c) i. true: forward stepwise approach adds one extra variables at a time
ii. true : backward stepwise approach eliminates one variable at a time
iii. false: no guarantee between matching between backward and forward stepwise approaches.
iv. false :no guarantee between matching between backward and forward stepwise approaches
v. false: no guarantee between matching between best subset approaches with different number of variables included in the model

2. For parts (a) through (c), indicate which of i. through iv. is correct. 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.

Answers:
TSE=error+bias+variance
(a) iii
(b) iii.
(c). ii.

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

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right) \quad \text{subject to } \sum_{j=1}^{p}|\beta_j|\leq s$$

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

(a) As we increase s 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.

Answers:
(a). iv
(b). ii.
(c). iii
(d). iv
(e) v

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

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

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.

Answers:
note that now we are talking in terms of λ (this moves in the opposite direction of s
(a). iii
(b). ii.
(c). iv
(d). iii
(e) v

5. It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.
Suppose that $n = 2$, $p = 2$, $x_11 = x_12$, $x_21 = x_22$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_11+x_21 = 0$ and $x_12+x_22 = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.
(a) Write out the ridge regression optimization problem in this setting.
(b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$
(c) Write out the lasso optimization problem in this setting.
(d) Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique —in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions.

Answers: (a). let $x_{11} = x_{12} = x_1$ and $x_{21} = x_{22} = x_2$
we also know that $x_1 + x_2 = 0, y_1 + y_2 = 0$

Recall ridge regression target function

$$
\begin{aligned}
Q &= \sum \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\Big)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\
\beta_0 &= 0 \\
Q &= \Big(y_1 - \beta_1 x_1 - \beta_2 x_1\Big)^2 + \Big(y_2 - \beta_1 x_2 - \beta_2 x_2\Big)^2 + \lambda\beta_1^2 + \lambda\beta_2^2 \\
\frac{\partial Q}{\partial \beta_1} &= -2x_1\Big(y_1 - \beta_1 x_1 - \beta_2 x_1\Big) - 2x_1\Big(y_2 - \beta_1 x_2 - \beta_2 x_2\Big) + 2\lambda\beta_1 = 0 \\
\frac{\partial Q}{\partial \beta_2} &= -2x_1\Big(y_1 - \beta_1 x_1 - \beta_2 x_1\Big) - 2x_2\Big(y_2 - \beta_1 x_2 - \beta_2 x_2\Big) + 2\lambda\beta_2 = 0
\end{aligned}
$$

(b) we can easily see $\hat{\beta}_1 = \hat{\beta}_2$ satisfy the above two equations.
(c)similar to (a) , we have

$$
\begin{aligned}
Q &= \sum \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\Big)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \\
&= \Big(y_1 - \beta_1 x_1 - \beta_2 x_1\Big)^2 + \Big(y_2 - \beta_1 x_2 - \beta_2 x_2\Big)^2 + \lambda|\beta_1| + \lambda|\beta_2| \\
&= y_1^2 - 2x_1 y_1(\beta_1 + \beta_2) + x_1^2(\beta_1 + \beta_2)^2 + y_2^2 - 2x_2 y_2(\beta_1 + \beta_2) + x_2^2(\beta_1 + \beta_2)^2 + \lambda|\beta_1| + \lambda|\beta_2|
\end{aligned}
$$

rearrange and consider $y_1 + y_2 = 0$ and $x_1 + x_2 = 0$

obviously we have $x_1 y_1 = x_2 y_2$

this is equivalent to minimize Q'

$Q' =$

6. We will now explore (6.12) and (6.13) further.
(a) Consider (6.12) with $p = 1$. For some choice of y_1 and $\lambda > 0$,
plot (6.12) as a function of β_1. Your plot should confirm that (6.12) is solved by (6.14).
(b) Consider (6.13) with $p = 1$. For some choice of y_1 and $\lambda > 0$, plot (6.13) as a function of β_1. Your plot should confirm that (6.13) is solved by (6.15).

7. We will now derive the Bayesian connection to the lasso and ridge regression discussed in Section 6.2.2.
(a) Suppose that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$ where $\epsilon_1, ..., \epsilon_n$ are independent and identically distributed from a $N(0, \sigma^2)$ distribution. Write out the likelihood for the data.
(b) Assume the following prior for $\beta : \beta_1, ..., \beta_p$ are independent and identically distributed according to a double-exponential distribution with mean 0 and common scale parameter b: i.e. $p(\beta) = \frac{1}{2b} exp(-|\beta|/b)$. Write out the posterior for β in this setting.
(c) Argue that the lasso estimate is the mode for under this posterior distribution.

(d) Now assume the following prior for β : $\beta_1, ..., \beta_p$are independent and identically distributed according to a normal distribution
with mean zero and variance c. Write out the posterior for β in this setting.
(e) Argue that the ridge regression estimate is both the *mode* and the *mean* for β under this posterior distribution.