Statistical Methods for Bioinformatics [I0U31a] Assignment 02 - Chapter 6

Hamed Borhani

April 19, 2016

6.8.5

(a) Ridge optimization problem:

$$\underset{\beta}{\text{minimize}}\ \{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\} + \lambda\sum_{j=1}^{p}\beta_j^2\}$$
$$\underset{\beta}{\text{minimize}}\ \{(y_1 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda\beta_1^2 + \lambda\beta_2^2\}$$

(b)
$(y_1 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda\beta_1^2 + \lambda\beta_2^2 =$
$(y_1 - x_{11}(\beta_1 + \beta_2))^2 + (y_2 - x_{21}(\beta_1 + \beta_2))^2 + \lambda\beta_1^2 + \lambda\beta_2^2 =$
$y_1^2 + x_{11}^2(\beta_1 + \beta_2)^2 - 2y_1x_{11}(\beta_1 + \beta_2) + y_2^2 + x_{21}^2(\beta_1 + \beta_2)^2 - 2y_2x_{21}(\beta_1 + \beta_2)) +$
$\lambda\beta_1^2 + \lambda\beta_2^2 =$

$$\begin{aligned} &y_1^2 + x_{11}^2\beta_1^2 + x_{11}^2\beta_2^2 + 2x_{11}^2\beta_1\beta_2 - 2y_1x_{11}(\beta_1 + \beta_2) \\ &+y_2^2 + x_{21}^2\beta_1^2 + x_{21}^2\beta_2^2 + 2x_{21}^2\beta_1\beta_2 - 2y_2x_{21}(\beta_1 + \beta_2)) + \lambda\beta_1^2 + \lambda\beta_2^2 \end{aligned} \tag{1}$$

To find the minimum, we should take the first derivative with respect to β_1 and β_2 and set it equal to zero:

$$\frac{d}{d\beta_1}(1) = 2x_{11}^2\beta_1 + 2x_{11}^2\beta_2 - 2y_1x_{11} + 2x_{21}^2\beta 1 + 2x_{21}^2\beta_2 - 2y_2x_{21} + 2\lambda\beta_1 = 0$$
$$\Rightarrow \beta_1 = \frac{x_{11}^2\beta_2 + x_{21}^2\beta_2 - y_1x_{11} - y_2x_{21}}{x_{11}^2 + x_{21}^2 + \lambda}$$

$$\frac{d}{d\beta_2}(1) = 2x_{11}^2\beta_2 + 2x_{11}^2\beta_1 - 2y_1x_{11} + 2x_{21}^2\beta 2 + 2x_{21}^2\beta_1 - 2y_2x_{21} + 2\lambda\beta_2 = 0$$
$$\Rightarrow \beta_2 = \frac{x_{11}^2\beta_1 + x_{21}^2\beta_1 - y_1x_{11} - y_2x_{21}}{x_{11}^2 + x_{21}^2 + \lambda}$$

Since both β_1 and β_2 look similar, we can say that in this setting, the ridge coefficients are equal.

(c) Lasso optimization problem:

$$\underset{\beta}{\text{minimize}}\ \{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\} + \lambda\sum_{j=1}^{p}|\beta_j|\}$$
$$\underset{\beta}{\text{minimize}}\ \{(y_1 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda|\beta_1| + \lambda|\beta_2|\}$$

(d)

6.8.8

```
#a.
set.seed(21)
X = rnorm(100)
ep = rnorm(100)

#b.
Y = 1 + 3.7 * X + -0.25 * X^2 + 1.04 * X^3 + ep

#c.
library(leaps)
df <- data.frame(Y=Y,X=X)

regfit.full <- regsubsets(Y~poly(X, degree = 10,
   nvmax = 10), data = df)
reg.summary <- summary(regfit.full)

which.max(reg.summary$adjr2)
which.max(reg.summary$rsq)
which.min(reg.summary$cp)
which.min(reg.summary$bic)

par(mfrow = c(2,2))
plot(reg.summary$adjr2, xlab = "Number of Variables"
   , ylab = "Adjusted R2", type = "l")
plot(reg.summary$cp, xlab = "Number of Variables",
   ylab = "CP", type = "l")
plot(reg.summary$bic, xlab = "Number of Variables",
   ylab = "BIC", type = "l")
```

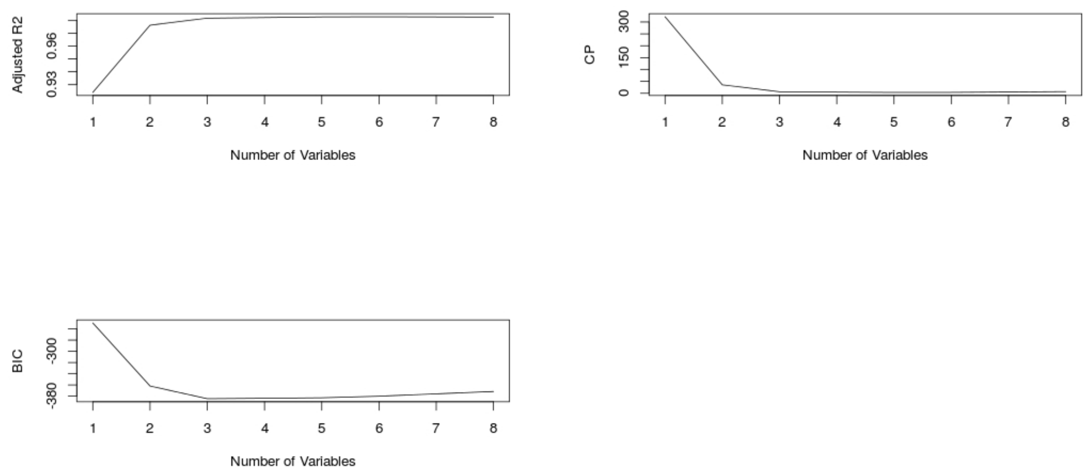

Figure 1: Adjusted R2, CP and BIC plots for all models

It seems that the best model is the model with 3rd degree polynomial:

(Intercept)	X^1	X^2	X^3
1.034283	65.716599	-5.104574	15.574039

6.8.10