Introduction to Statistical Learning
Chapter 6 Exercise

Penghao Chen

April 17, 2019

1

(a) Best subset.

(b) Best subset.

(c)

i True.

ii True.

iii False.

iv False.

v False.

2

(a) iii is correct.

(b) iii is correct.

(c) ii is correct.

3

(a) iv is correct.

(b) ii is correct.

(c) iii is correct.

(d) iv is correct.

(e) v is correct.

4

(a) iv is correct.

(b) ii is correct.

(c) iii is correct.

(d) iv is correct.

(e) v is correct.

5

(a) The ridge optimization problem:

$$\min_{beta_1, beta_2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

(b) Let

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

Then

$$\frac{\partial F}{\partial \beta_1} = 2(\lambda \beta_1 - y_1 x_{11} + y_2 x_{21})$$
$$\frac{\partial F}{\partial \beta_2} = 2(\lambda \beta_2 - y_1 x_{12} + y_2 x_{22})$$

By first order condition, setting the partial derivatives to 0 and solve for β_1 and β_2, we get:

$$\beta_1 = \frac{y_1 x_{11} + y_2 x_{21}}{\lambda}$$
$$\beta_2 = \frac{y_1 x_{12} + y_2 x_{22}}{\lambda}$$

Remember we have $x_{11} = x_{12}$, $x_{21} = x_{22}$, hence $\beta_1 = \beta_2$.

(c) The ridge optimization problem:

$$\min_{beta_1, beta_2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

(d) Let's transform the optimization problem into:

$$\min_{beta_1, beta_2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

such that

$$\beta_1 + \beta_2 \leq s$$

Let

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

Plugging in $x_{11} = x_{12}$, $x_{21} = x_{22}$, $y_1 + y_2 = 0$, $x_{11} + x_{21} = 0$, $x_{12} + x_{22} = 0$, we get:

$$F = 2\left[y_1 - (\beta_1 + \beta_2)x_{11}\right]^2$$

By first order derivative, we get:

$$\begin{aligned} \frac{\partial F}{\partial \beta_1} &= \beta_1 + \beta_2 = \frac{y_1}{x_{11}} \\ \frac{\partial F}{\partial \beta_2} &= \beta_1 + \beta_2 = \frac{y_1}{x_{11}} \end{aligned}$$

Hence there is no unique solution for β_1 and β_2.

6

(a) For $p = 1$, we need to minimize the following function:

$$F = (y - \beta)^2 + \lambda \beta^2$$

By first order condition, we get:

$$\begin{aligned} \frac{\partial F}{\partial \beta} &\Rightarrow 2(\beta - y) + 2\lambda\beta = 0 \\ &\Rightarrow \beta = \frac{y}{1 + \lambda} \end{aligned}$$

This proves that (6.12) is solved by (6.14).

(b) For $p = 1$, we need to minimize the following function:

$$F = (y - \beta)^2 + \lambda|\beta|$$

By first order condition, we get:

$$\frac{\partial F}{\partial \beta} = \begin{cases} 2(\beta - y) + \lambda = 0 & \beta > 0 \\ 2(\beta - y) - \lambda = 0 & \beta < 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Solving the above equations, we get:

$$\beta = \begin{cases} y - \frac{\lambda}{2} & y > \frac{\lambda}{2} \\ y + \frac{\lambda}{2} & y < -\frac{\lambda}{2} \\ 0 & |y| \leq \frac{\lambda}{2} \end{cases} \tag{2}$$

This proves that (6.15) is solved by (6.13).

7

(a) Note that $\epsilon_i = y_i - \beta_0 - \sum\limits_{j=1}^{p} x_{ij}\beta_j$ follows $N(0, \sigma^2)$

The likelihood of the data is:

$$\begin{aligned} L &= \prod_{i=1}^{n} \phi(\epsilon_i) \\ &= \prod_{i=1}^{n} \phi(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j) \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} exp\left(-\frac{\sum\limits_{i=1}^{n}\left(y_i - \beta_0 - \sum\limits_{j=1}^{p} x_{ij}\beta_j\right)^2}{2\sigma^2} \right) \end{aligned}$$

This is $f(Y|X, \beta)$ in the Bayesian expression.

(b) Posterior distribution for β is:

$$\begin{aligned} p(\beta|X, Y) &\propto f(Y|X, \beta)p(\beta) \\ &= \frac{1}{2b(\sqrt{2\pi}\sigma)^n} exp\left(-\frac{\sum\limits_{i=1}^{n}\left(y_i - \beta_0 - \sum\limits_{j=1}^{p} x_{ij}\beta_j\right)^2}{2\sigma^2} - \frac{\sum\limits_{i} |\beta_i|}{b} \right) \end{aligned}$$

(c) If we maximize the likelihood function:

$$
\begin{aligned}
\max_\beta p(\beta|X,Y) &\Leftrightarrow \max_\beta log\left(p(\beta|X,Y)\right)\\
&\Leftrightarrow \max_\beta - \frac{\sum\limits_{i=1}^{n}\left(y_i-\beta_0-\sum\limits_{j=1}^{p}x_{ij}\beta_j\right)^2}{2\sigma^2} - \frac{\sum_i|\beta|}{b}\\
&\Leftrightarrow \min_\beta \frac{\sum\limits_{i=1}^{n}\left(y_i-\beta_0-\sum\limits_{j=1}^{p}x_{ij}\beta_j\right)^2}{2\sigma^2} + \frac{\sum_i|\beta|}{b}\\
&\Leftrightarrow \min_\beta \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \frac{2\sigma^2}{b}\sum_i|\beta|\right)\\
&\Leftrightarrow \min_\beta \sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \frac{2\sigma^2}{b}\sum_i|\beta|
\end{aligned}
$$

Letting $\lambda = \frac{2\sigma^2}{b}$, we get:

$$
\max_\beta p(\beta|X,Y) \Leftrightarrow \min_\beta \sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2 + \lambda\sum_i|\beta|
$$

Hence, solving the lasso optimization problem is equivalent to solving the maximum likelihood problem. The solution for lasso regression also maximizes the likelihood function for β.

By definition, values with the maximum likelihood is the mode.

(d) Now we are using a different prior. The posterior distribution for β is:

$$
\begin{aligned}
p(\beta) &= \prod_{i=1}^{p}\frac{1}{\sqrt{2\pi c}}exp\left(-\frac{\beta_i^2}{2c}\right)\\
&= \frac{1}{\sqrt{2\pi c}^p}exp\left(-\right)
\end{aligned}
$$

(e) With the new posterior, the likelihood function we are trying to maximize becomes:

$$
\begin{aligned}
p(\beta|X,Y) &\propto f(Y|X,\beta)p(\beta)\\
&= \frac{1}{(\sqrt{2\pi}\sigma)^n}\frac{1}{\sqrt{2\pi c}^p}exp\left(-\frac{\sum\limits_{i=1}^{n}\left(y_i-\beta_0-\sum\limits_{j=1}^{p}x_{ij}\beta_j\right)^2}{2\sigma^2} - \frac{\sum\limits_{i=1}^{p}\beta_i^2}{2c}\right)
\end{aligned}
$$

If we maximize the likelihood function:

$$\begin{aligned}
\max_{\beta} p(\beta|X,Y) &\Leftrightarrow \max_{\beta} log\left(p(\beta|X,Y)\right) \\
&\Leftrightarrow \min_{\beta} \frac{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2}{2\sigma^2} + \frac{\sum_{i=1}^{p}\beta_i^2}{2c} \\
&\Leftrightarrow \min_{\beta} \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \frac{\sigma^2}{c}\sum_{i=1}^{p}\beta_i^2\right)
\end{aligned}$$

Letting $\lambda = \frac{\sigma^2}{c}$, we get:

$$\max_{\beta} p(\beta|X,Y) \Leftrightarrow \min_{\beta} \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{i}^{p}\beta_i^2$$

Hence, solving the lasso optimization problem is equivalent to solving the maximum likelihood problem. The solution for ridge regression also maximizes the likelihood function for β.

By definition, values with the maximum likelihood is the mode. Observe that the posterior likelihood function for β is Gaussian as well. The mode and the mean overlaps in this situation. Therefore the ridge regression estimate is both the mode and the mean of the posterior distribution.