ISLR 6.6.3
6.6.5
8.4.5
9.7.3

① 6.6.3 Lasso as s increases

estimate the regression coefficients by minimizing
$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 \text{ subject to}$$
$$\sum_{j=1}^{p} |\beta_j| \le s$$
for a particular value of s.

As we increase s from 0

(a) Train RSS :- (iv) Steadily decrease

When s is sufficiently large, the β that minimizes RSS will be the least squares solution. Thus, training RSS will decrease monotonically.

(b) Test RSS :- (ii) Decrease initially, and then eventually start to increase in a U shape

When $s = 0$ $y = \bar{y}$
As s increases, more β are included, flexibility of model ↑, ∴ test RSS at some point will start increasing, ~~leading~~ to (overfitting happens).

(c) Variance :- (iii) steadily increase

As s goes from 0 to some high value, the flexibility of model increases, ∴ Variance increases.

(d) Squared Bias :- (iv) Steadily decreases

→ Because, model flexibility increases with s.

(e) Irreducible error :- (v) Remain constant

Irreducible error ε remains ~~des~~ regardless of model complexity. Its not dependent on X.

ISLR

6.6.5

Ridge regression tends to give similar coefficient values to correlated variables, whereas lasso may give different coefficient values to correlated variables.

Given

$n=2,\ p=2,\ x_{11}=x_{12},\ x_{21}=x_{22}.$

$y_1+y_2=0 \qquad x_{11}+x_{21}=0$

$x_{12}+x_{22}=0 \rightarrow \hat{\beta}_0=0$

In Ridge - $\hat{\beta}_\lambda^R$ are values the minimizes

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

choose $\hat{\beta}_\lambda^R = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ such that it minimizes:-

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

$$\Rightarrow \left(y_1 - \beta_0 - (\beta_1 x_{11} + \beta_2 x_{12}) \right)^2 + \left(y_2 - \beta_0 - (\beta_1 x_{21} + \beta_2 x_{22}) \right)^2 + \lambda (\beta_1^2 + \beta_2^2)$$

$$\Rightarrow (y_1 - \beta_0 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_0 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda (\beta_1^2 + \beta_2^2)$$

$\hat{\beta}_0 = 0$

(b) $$(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \cancel{\beta_0} - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda (\hat{\beta}_1^2 + \beta_2^2)$$

$$\Rightarrow (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_1 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda (\beta_1^2 + \beta_2^2)$$

$$\Rightarrow (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 \cancel{+} \left(\cancel{+ y_1} \cancel{+} \hat{\beta}_1 x_{11} \cancel{+} \hat{\beta}_2 x_{12} \right)^2 + \lambda (\beta_1^2 + \beta_2^2)$$

$\Rightarrow$

$$\Rightarrow (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{22})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

$$\Rightarrow (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{11})^2 + (-y_1 - \hat{\beta}_1 x_{21} - \hat{\beta}_2 x_{21})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

$$\Rightarrow (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{11})^2 + (-y_1 + \hat{\beta}_1 x_{11} + \hat{\beta}_2 x_{11})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

$$\Rightarrow (y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{11})^2 + (-1)^2(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{11})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

$$f(\hat{\beta}_1, \hat{\beta}_2) \quad 2(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{11})^2 + \lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2)$$

$f \rightarrow$

$\Rightarrow$

$$\frac{\partial(f)}{\partial \hat{\beta}_1} = 4(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{11})(-x_{11}) + 2\lambda\beta_1 = 0$$

$$2\lambda\hat{\beta}_1 = 4y_1 x_{11} - 4\hat{\beta}_1 x_{11}^2 - 4\hat{\beta}_2 x_{11}^2$$

$$\lambda\hat{\beta}_1 = 2y_1 x_{11} - 2\hat{\beta}_1 x_{11}^2 - 2\hat{\beta}_2 x_{11}^2$$

$$\hat{\beta}_1(\lambda + 2x_{11}^2) = 2y_1 x_{11} - 2\hat{\beta}_2 x_{11}^2$$

$$\hat{\beta}_1 = \frac{2y_1 x_{11} - 2\hat{\beta}_2 x_{11}^2}{(\lambda + 2x_{11}^2)}$$

$$\frac{\partial(f)}{\partial(\hat{\beta}_2)} = 4(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{11})(-x_{11}) = 0$$
$$+ 2\lambda\hat{\beta}_2$$

$$\Rightarrow 2y_1 x_{11} - 2\hat{\beta}_1 x_{11}^2 - 2\hat{\beta}_2 x_{11}^2 = \lambda\hat{\beta}_2$$

$$\hat{\beta}_2 = \frac{2y_1 x_{11} - 2\hat{\beta}_1 x_{11}^2}{(\lambda + 2x_{11})^2}$$

Substituting

$$C = \frac{2y_1 x_{11}}{(\lambda + 2x_{11}^2)} \quad \& \quad k = \frac{-2x_{11}^2}{(\lambda + 2x_{11})^2}$$

$$\hat{\beta}_1 = C + k\hat{\beta}_2 \quad -(1)$$
$$\hat{\beta}_2 = C + k\hat{\beta}_1 \quad -(2)$$

$$\hat{\beta}_1 = C + K(C + K\hat{\beta}_1)$$
$$\hat{\beta}_1 = C + KC + K^2\hat{\beta}_1$$
$$\hat{\beta}_1 = \frac{C(1+K)}{(1-K^2)} \quad (3)$$

(1) → (2)

$$\hat{\beta}_2 = C + K(C + K\hat{\beta}_2)$$
$$\hat{\beta}_2 = C + KC + K^2\hat{\beta}_1$$
$$\hat{\beta}_2 = \frac{C(1+K)}{(1-K^2)} \quad (4)$$

$$\therefore \hat{\beta}_1 = \hat{\beta}_2$$

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

$$\Rightarrow (y_1 - \beta_0 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_0 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda(|\beta_1| + |\beta_2|)$$

$$\Rightarrow 2(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{11})^2 + \lambda(|\beta_1| + |\beta_2|)$$

(d) Argue non-unique coefficient estimates

$\Rightarrow$

Lasso will select $\hat{\beta}^L$ that minimizes.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 \text{ give to constraint}$$

$$\sum_{j=1}^{p} |\beta_j| \leq s$$

$\therefore$ Lasso coefficients must minimize

$$2\left(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11}\right)^2 \geq 0 \text{ subject to}$$

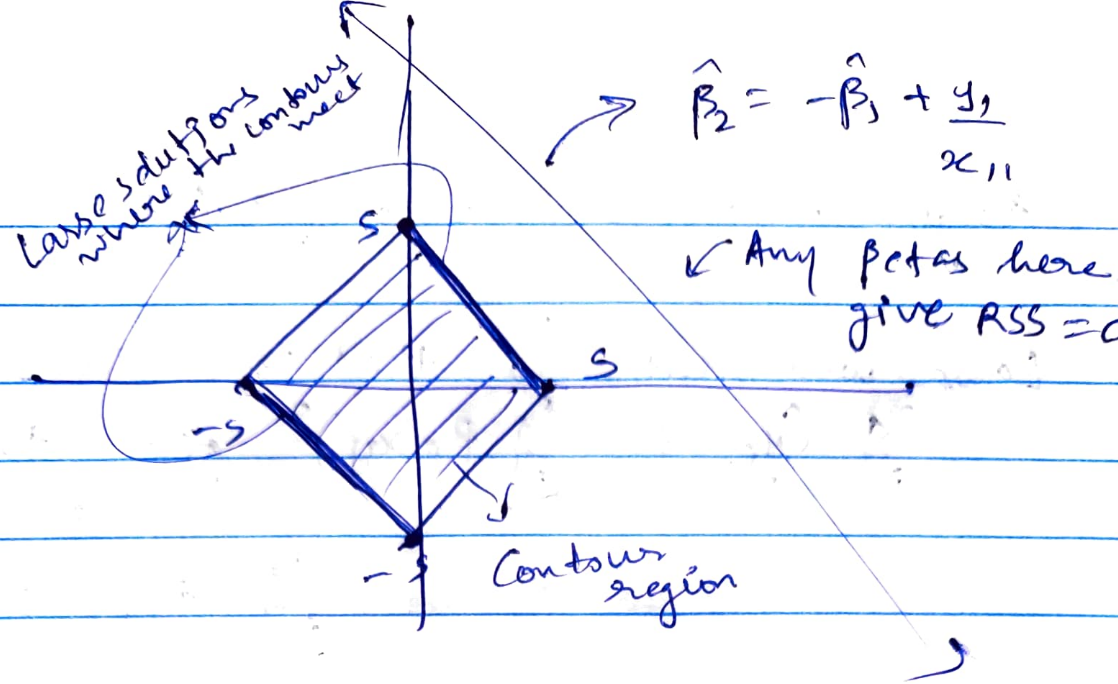
$$|\hat{\beta}_1| + |\hat{\beta}_2| \leq s$$

We can see that any $(\hat{\beta}_1, \hat{\beta}_2)$ that satisfy $\hat{\beta}_1 + \hat{\beta}_2 = \frac{y_1}{x_{11}}$ will have RSS of 0.

Considering the lasso constraint, the solution to the optimization problem will be where the contours of $2\left(y_1 - (\hat{\beta}_1 + \hat{\beta}_2)x_{11}\right)^2$ touch the lasso diamond.

As $(\hat{\beta}_1, \hat{\beta}_2)$ vary along $\hat{\beta}_2 = -\hat{\beta}_1 + \frac{y_1}{x_{11}}$ the contour will touch lasso diamond at many points instead of one.

$$\underbrace{\left(|\hat{\beta}_1| + |\hat{\beta}_2| \leq s\right)}_{\text{lasso diamond}} \qquad \underbrace{\begin{array}{l}\hat{\beta}_1 + \hat{\beta}_2 = s \\ \hat{\beta}_1 + \hat{\beta}_2 = -s\end{array}}_{\text{instead of one}}$$



This means there are ∞ solutions

$\hat{\beta}_\lambda^L$

$$\hat{\beta}_\lambda^L \in \{ (\hat{\beta}_1, \hat{\beta}_2) : \hat{\beta}_1 + \hat{\beta}_2 = s, \hat{\beta}_1 \in [0, s], \beta_2 \in [0, s] \}$$

$$\cup$$

$$\{ (\hat{\beta}_1, \hat{\beta}_2) : \hat{\beta}_1 + \hat{\beta}_2 = -s, \hat{\beta}_1 \in [-s, 0], \hat{\beta}_2 \in [-s, 0] \}$$

ISLR
8.4.5

① Majority vote :-

Accross 10 estimates

G, G, G, G, $\underbrace{R, R, R, R, R}$

O/P prediction → Red

② Average probability

Ave P (C = "Red" | X) = 0.45
Ave P (C = "Green" | X) = 0.55

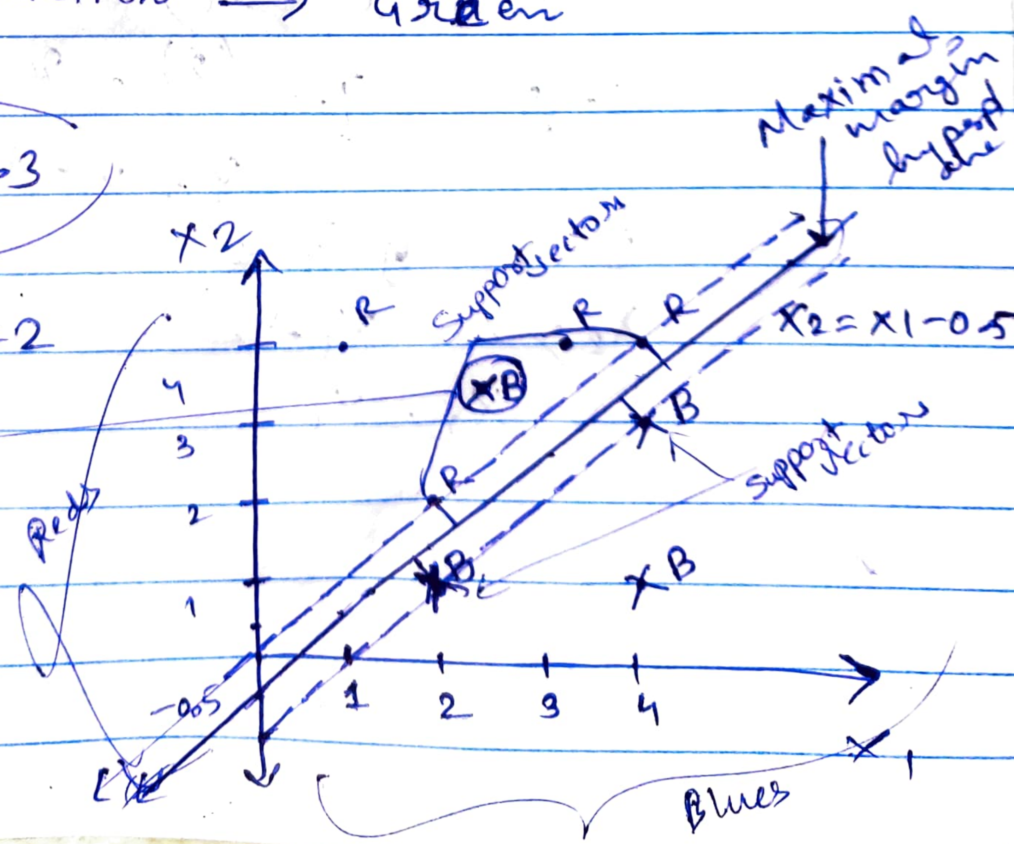
O/P prediction → Green

ISLR 9.7.3

(9) n = 7, p = 2

-- Additional point

Support vectors are on dashed lines.



(b) Visually it looks like the optimal separating hyperplane is

$$x_2 - x_1 + 0.5 = 0$$

Hyperplanes $\Rightarrow$

$$\underset{(2,1)}{d_1} \rightarrow \frac{|2-1+0.5|}{\sqrt{2}} = \frac{1.5}{\sqrt{2}}$$

$$\underset{(2,2)}{d_2} \rightarrow \frac{|0.5|}{\sqrt{2}} = \frac{0.5}{\sqrt{2}}$$

① $y - y_1 = 1(x - x_1)$

(2,2) $y - 2 = 1(x - 2)$

$\boxed{y = x}$ — ① $\boxed{x_2 = x_1}$

② $y - y_1 = 1(x - x_1)$

(2,1) $y - \cancel{y}\, 1 = x - 2$

$\boxed{x_2 = x_1 - 1}$

$$\text{Margin} = \frac{\sqrt{2}}{4}$$

(c) Classification Rule for Maximal margin classifier.

"Red" if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$ else "Blue"

$\rightarrow (0.5, -1, 1)$

i.e. "Red" when $X_2 > X_1 - 0.5$

& "Blue" when $X_2 \leq X_1 - 0.5$

The maximal marginal classifier thus becomes

$f(x) = X_2 - X_1 + 0.5$

here if $f(x') > 0 \rightarrow$ "Blue"

else "Red"

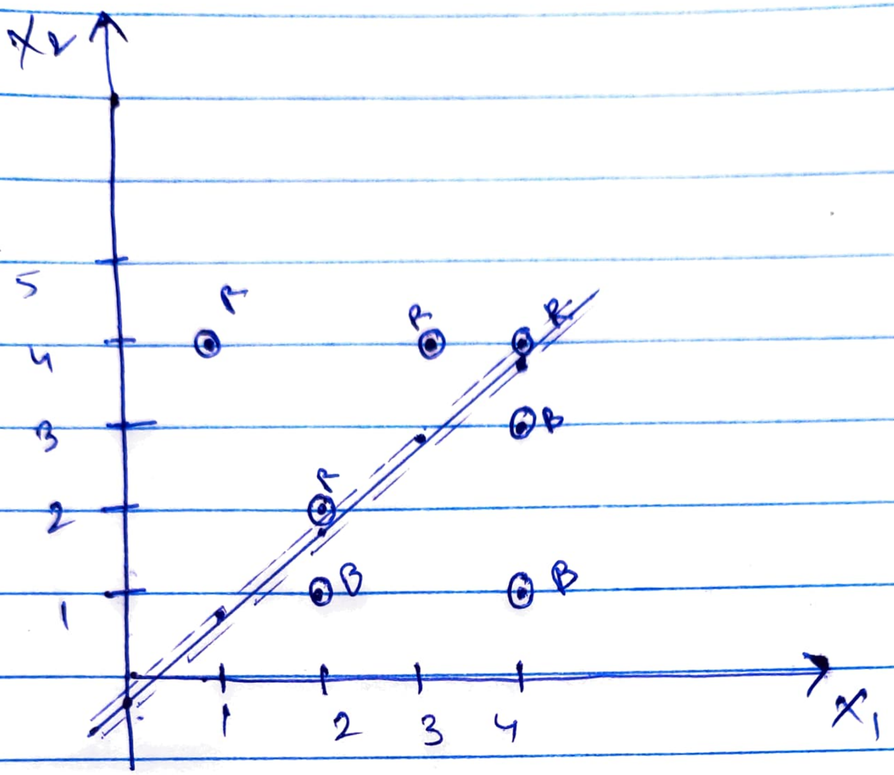
(d)

(e) Check answer (a)

(f) Observation ⑦ is not a support vector. Any small movement of this point won't change the fact that $X_2 - X_1 + 0.5 = 0$.

If observation 7 moves inside of the margin then it will start influencing the position of maximal hyperplane.

(g)



$x_2 = X_1 - 0.25$

The Margin M in this case is extremely small.

Not an optimal hyperplane

(h) ~~Check answer~~ (a)