/ /

Hw3-problem2 - conceptual questions - ISL book - chap3

1] Sales = 2.939 + 0.046×TV + 0.189×radio − 0.001×$\underbrace{\text{newspaper}}$

isn't associated with sales in the presence of TV & radio

3] $Y = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_4 - 10X_5$

a) i and iii can be true.

b) $50 + 20\times4 + 0.07\times110 + 35\times\overset{\text{female}}{\overset{\uparrow}{1}} + 0.01\times4\times110 - 10\times4\times1 = 137$

c) small coefficient for interaction GPA×IQ (X_4) means it has a bit association with sallary in the presence of other feartures.

5] $\hat{y}_i = \hat{\beta}x_i \Rightarrow \hat{y}_i = \dfrac{\sum_{i=1}^{n} x_i y_i}{\sum x_i^2} \times x_i$

مخرج مشترک دارند می توان جدا کرد

$$= \sum \underbrace{\left(\frac{x_{i'} x_i}{\sum x_{i'}^2}\right)}_{a_{i'}} y_{i'}$$

$$\Downarrow$$

$$\boxed{\hat{y}_i = \sum a_{i'} y_{i'}}$$

ISL_book_chap6

1] a) between different models, a model which has better fit to training data, has smallest training RSS.

b) if the data have large P(features), may overfit and in this situation forward stepwise selection has smaller test RSS.

c) i) True   ii) True   iii) false
iv) false   v) True

2] a) iii → lasso regression can remove some features by zero coefficient and decrease flexibility (complexity)

b) iii → ridge is also act like lasso.

c) ii → non-linear methods can lead to increase in flexibility.

5] a) $(y_1-\hat{\beta}_1x_{11}-\hat{\beta}_2x_{12})^2+(y_2-\hat{\beta}_1x_{21}-\hat{\beta}_2x_{22})^2+\lambda(\hat{\beta}_1^2+\hat{\beta}_2^2)$ → ridge regression

b) $\hat{\beta}_1=\hat{\beta}_2$ $(y_1^2+\hat{\beta}_1^2x_{11}^2+\hat{\beta}_2^2x_{12}^2-2\hat{\beta}_1x_{11}y_1-2\hat{\beta}_2x_{12}y_1+2\hat{\beta}_1\hat{\beta}_2x_{11}x_{12})$

$+(y_2^2+\hat{\beta}_1^2x_{21}^2+\hat{\beta}_2^2x_{22}^2-2\hat{\beta}_1x_{21}y_2-2\hat{\beta}_2x_{22}y_2+2\hat{\beta}_1\hat{\beta}_2x_{21}x_{22})$

$+\lambda\hat{\beta}_1^2+\lambda\hat{\beta}_2^2$

/ /

$\rightarrow$ minimization $\quad \frac{\partial L(\beta)}{\partial \beta_1} = 0 \rightarrow (2\hat{\beta}_1 x_{11}^2 - 2x_{11}y_1 + 2\hat{\beta}_2 x_{11}x_{12}) +$

$(2\hat{\beta}_1 x_{21}^2 - 2x_{21}y_2 + 2\hat{\beta}_2 x_{21}x_{22}) + 2\lambda\hat{\beta}_1 = 0$

$$\left.\begin{array}{l} x_{11} = x_{12} = x_0 \\ x_{21} = x_{22} = x_1 \end{array}\right\} \rightarrow (\hat{\beta}_1 x_0^2 - x_0 y_1 + \hat{\beta}_2 x_0^2) + (\hat{\beta}_1 x_1^2 - x_1 y_2 + \hat{\beta}_2 x_1^2) + \lambda\hat{\beta}_1 = 0$$

$$\Rightarrow \hat{\beta}_1(x_0^2 + x_1^2) - x_0 y_1 - x_1 y_2 + \hat{\beta}_2(x_0^2 + x_1^2) + \lambda\hat{\beta}_1 = 0$$

$$\lambda\hat{\beta}_1 + \hat{\beta}_1(x_0^2 + x_1^2) + \hat{\beta}_2(x_0^2 + x_1^2) = x_0 y_1 + x_1 y_2$$

$x_0 \quad x_1$

$x_{11} + x_{21} = 0$

$$\beta_1(\overbrace{x_0^2 + x_1^2 + 2x_0x_1}^{\cancel{(x_0+x_1)^2}\ 0} - 2x_0x_1) + \hat{\beta}_2(\overbrace{x_0^2 + x_1^2 + 2x_0x_1}^{0} - 2x_0x_1) + \lambda\hat{\beta}_1$$

$$\lambda\hat{\beta}_1 - 2\beta_1 x_0 x_1 - 2\hat{\beta}_2 x_0 x_1 = x_0 y_1 + x_1 y_2$$

$$\Rightarrow \lambda\hat{\beta}_1 = 2\beta_1 x_0 x_1 + 2\hat{\beta}_2 x_0 x_1 + x_0 y_1 + x_1 y_2$$

$\frac{\partial L(\beta)}{\partial \beta_2} \rightarrow (2\hat{\beta}_2 x_{12}^2 - 2x_{12}y_1 + 2\beta_1 x_{11}x_{12}) +$

$(2\hat{\beta}_2 x_{22}^2 - 2x_{22}y_2 + 2\hat{\beta}_1 x_{21}x_{22}) + 2\lambda\hat{\beta}_2 = 0$

$\Downarrow$

$$\left.\begin{array}{l} x_{11} = x_{12} = x_0 \\ x_{21} = x_{22} = x_1 \end{array}\right\} \rightarrow (\hat{\beta}_2 x_0^2 - x_0 y_1 + \hat{\beta}_1 x_1^2) + (\hat{\beta}_2 x_1^2 - x_1 y_2 + \hat{\beta}_1 x_1^2) + \lambda\hat{\beta}_2 = 0$$

$\Downarrow$

$$\Rightarrow \lambda\beta_2 = 2\beta_1 x_0 x_1 + 2\hat{\beta}_2 x_0 x_1 + x_0 y_1 + x_1 y_2$$

$$\Rightarrow \lambda\beta_1 = \lambda\beta_2 \rightarrow \boxed{\beta_1 = \beta_2}$$

/ /

c) lasso: should be minimize

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

d) in lasso regression we have absolute for β and can't do partial derivation. so other possible solutions like convex optimization may be useful.

7] a) $D=\{x[1],\dots,x[n]\} \to L(\sigma^2)=p(x[1])\times\dots\times p(x[n])$

likelihood $=\prod_{i=1}^{n} p(x[i])$

$$=\left(\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j+\varepsilon_i\right)^n \to \ell(x)=\log L(\mu)$$

b) prior $\to P(\beta)=\frac{1}{2b}\exp(-|\beta|/b)$

$\to$ posterior = prior x Liklihood $=\frac{1}{2b}\exp(-|\beta|/b)\times\log(\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j)^n$

c) I can't solve this section.

d) posterior = prior x likelihood $=\left(\frac{1}{\sqrt{2\pi c}}\right)^n \exp\left(-\frac{1}{2c^2}\sum_{i=1}^{n}(x[i]-\mu)^2\right)$
$\log$ (prior: $N(0,c)$)

$$\times\log(\beta_0+\sum_{j=1}^{p} x_{ij}\beta_j)^n$$

e) I can't solve this.

/ /

problem 3 : Show that the ridge hat matrix is not a projection matrix.

projection matrix : → 1) idenpotent → $H^2 = H$

- the ridge hat matrix → proof : $H^2 \neq H$ $\quad (H = X(X^TX + \lambda I)^{-1}X^T)$

$\Downarrow$ $\quad X = UDV^T$

$H^2 - H \neq 0$

$\Downarrow$

$H(H - I) \neq 0$

$H = UDV^T(VD^TU^T \cdot UDV^T + \lambda I)^{-1} VD^TU^T$

$UU^T \cdot DD^T(D^TD + \lambda I)^{-1}\left(\left[UU^T \cdot DD^T(D^TD + \lambda I)^{-1}\right] - I\right) \neq 0$

$UU^T \cdot DD^T(D^TD + \lambda I)^{-1} \cdot UU^T \cdot DD^T(D^TD + \lambda I)^{-1} - UU^T DD^T(D^TD + \lambda I)^{-1}$

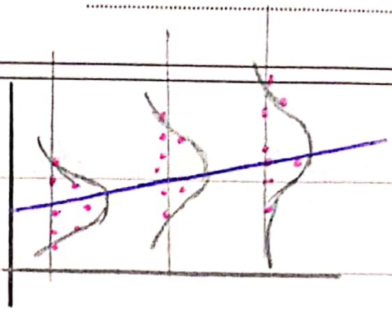
$UU^T \underbrace{DD^T(D^TD)^{-1}}_{I}(I + \lambda I)^{-1} \cdot \underbrace{DD^T(D^TD)^{-1}}_{I}(I + \lambda I)^{-1} - UU^T(I + \lambda I)^{-1}$

$= UU^T(1+\lambda)^2 - UU^T(1+\lambda) \neq 0 \rightarrow (1+\lambda)^2 \neq (1+\lambda)$

$\Downarrow$

$\boxed{H^2 \neq H}$

problem 4 : weighted linear regression →



- derive the optimal solution $\hat{\beta}_n$ for the weighted loss function:

$$L(\beta) = \frac{1}{2}\sum_{i=1}^{n} w[i] \times (y[i] - \beta^T x[i])^2$$

$$\Rightarrow L(\beta) = W^T (Y - X\beta)^T (Y - X\beta)$$

$$= W^T (Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta)$$

$$= W^T Y^T Y - 2 W^T \beta^T X^T Y + W^T \beta^T X^T X \beta$$

$$\rightarrow \hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, L(\beta) \rightarrow \frac{\partial L(\beta)}{\partial \beta} = -2 W^T X^T Y + 2 W^T X^T X \beta$$

$$\Rightarrow \hat{\beta} = \frac{W^T X^T Y}{W^T X^T X} = \boxed{(W^T X^T X)^{-1} W^T X^T Y}$$

$$\Rightarrow \text{Hessian} \rightarrow \nabla^2 L(\beta) = 2 W^T X^T X$$

/ /

problem 5 : Maximum likelihood estimation of multinomial distribution.

$x \to$ random variable $\to x_1, x_2, \dots, x_K$

have K parameters $\to \Pi = (\pi_1, \dots, \pi_K)$, $P(x = x_K) = \pi_K$

subject to $\to \sum \pi_K = 1 \to \sum \pi_K - 1 = 0$

-observed data $\to (n_1, \dots, n_K) \to n_K$: the number of times the value x_K appears in the data

prove the MLE for π_k is $\frac{n_K}{n}$ where $n = \sum n_K$

$P(D|\theta)$

$$P(n_1, \dots, n_K | \pi_1, \dots, \pi_K) = \binom{n}{n_1, n_2, n_K} \pi_1^{n_1} \pi_1^{n_2} \dots \pi_K^{n_K} = \frac{n!}{\prod n_i!} \prod \pi_i^{n_i}$$

$$l(\overbrace{\pi_1, \dots, \pi_K}^{\theta}) = \log n! - \sum_{i=1}^{k} \log n_i! + \sum_{i=1}^{K} n_i \log \pi_i$$

$$\underbrace{L(\pi_1, \dots, \pi_K, \lambda)}_{\text{lagrangian}} = l(\pi_1, \dots, \pi_K) + \lambda\left(\sum_{i=1}^{k} \pi_i - 1\right)$$

$$= \log n! - \sum_{i=1}^{k} \log n_i! + \sum_{i=1}^{k} n_i \log \pi_i + \lambda\left(\sum_{i=1}^{k} \pi_i - 1\right)$$

$$\frac{\partial L(\pi_1, \dots, \pi_K, \lambda)}{\partial \pi} = 0 \to \sum_{i=1}^{k} n_i \frac{1}{\pi_i \operatorname{Ln} 10} + \lambda = 0 \Rightarrow \lambda = -\sum \frac{n_i}{\pi_i \operatorname{Ln} 10}$$

$$\frac{\partial L(\pi_1, \dots, \pi_K, \lambda)}{\partial \lambda} = 0 \to \sum_{i=1}^{k} \pi_i - 1 = 0 \to \sum \pi_i = 1$$