1. $Y = f(X) + \varepsilon$

(a) Show $E[(Y-\hat{f}(X;D))^2] = Var(\varepsilon) + E[(f(X) - E[\hat{f}(X;D)|X])^2]$

$+ E[(\hat{f}(X;D) - E[\hat{f}(X;D)|X])^2]$

Sol: $E[(Y-\hat{f}(X;D))^2] = E[(Y - f(X) + f(X) - \hat{f}(X;D))^2]$

$= \overset{①}{E}[(Y-f(X))^2] + \overset{②}{E}[(f(X) - \hat{f}(X;D))^2]$

$\overset{③}{+2}E[(Y-f(X))(f(X) - \hat{f}(X;D))]$

① $E[(Y-f(X))^2] = Var(\varepsilon)$

② $E[(f(X)-\hat{f}(X;D))^2] = E[(f(X) - E_D[\hat{f}(X;D)|X] + E_D[\hat{f}(X;D)|X] - \hat{f}(X;D))^2]$

$= E[(f(X) - E_D[\hat{f}(X;D)|X)^2]$ — bias

$+ E[(E_D[\hat{f}(X;D)|X] - \hat{f}(X;D))^2]$ — variance

③ $2E[(Y-f(X))(f(X) - \hat{f}(X;D))]$

$= E[E[(Y-f(X))(f(X) - \hat{f}(X;D))]|X;D]$

$= E[(f(X) - f(X))(f(X) - \hat{f}(X;D))]$

$= 0$

① + ② + ③

$\Rightarrow E[(Y-\hat{f}(X;D))^2] = Var(\varepsilon) + E[f(X) - E[\hat{f}(X;D)|X)^2]$

$+ E[(\hat{f}(X;D) - E[\hat{f}(X;D)|X)^2]$

(b) As the flexibility increase, the model has small bias and high variance.

2、(a) $L = \|y - \underset{\sim}{X}\beta\|^2 + \lambda\|\beta\|^2 = (y - \underset{\sim}{X}\beta)^T(y - \underset{\sim}{X}\beta) + \lambda\beta^T\beta$

$$\frac{\partial L}{\partial \beta} = 0 \Rightarrow -2\underset{\sim}{X}^T y + 2\underset{\sim}{X}^T\underset{\sim}{X}\beta + 2\lambda\beta = 0$$

$$\Rightarrow \hat{\beta} = (\underset{\sim}{X}^T\underset{\sim}{X} + \lambda I)^{-1}\underset{\sim}{X}^T y$$

(b) $SE(\hat{\beta})^2 = E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T]$

$$= E[((X^TX+\lambda I)^{-1}X^Ty - \beta)((X^TX+\lambda I)^{-1}X^Ty-\beta)^T]$$

$$= E[((X^TX+\lambda I)^{-1}X^T(X\beta+\varepsilon)-\beta)((X^TX+\lambda I)^{-1}X^T(X\beta+\varepsilon)-\beta)^T]$$

$$= E\Big[\Big((X^TX+\lambda I)^{-1}X^TX\beta + (X^TX+\lambda I)^{-1}X^T\varepsilon - \beta\Big)\Big((X^TX+\lambda I)^{-1}X^TX\beta + (X^TX+\lambda I)^{-1}X^T\varepsilon - \beta\Big)^T\Big]$$

$\because X^TX = aI_p$

$$= E\left[\left(\frac{a}{a+\lambda}\beta + \frac{1}{a+\lambda}X^T\varepsilon - \beta\right)\left(\frac{a}{a+\lambda}\beta + \frac{1}{a+\lambda}X^T\varepsilon - \varepsilon\right)^T\right]$$

$$= E\left[\left(\frac{1}{a+\lambda}(-\lambda\beta + X^T\varepsilon)\right)\left(\frac{1}{a+\lambda}(-\lambda\beta+X^T\varepsilon)\right)^T\right]$$

$$= \left(\frac{1}{a+\lambda}\right)^2\left(\lambda^2\beta\beta^T + a\sigma^2 I_p\right)$$

$$\rightarrow SE(\beta_j) = \frac{1}{a_j+\lambda}\left(\lambda^2\{\beta\beta^T\}_{jj} + a_j\sigma^2\right)$$

Scanned by CamScanner

Q3:-

Given dataset $D = \{(x_1 y_1), (x_2 y_2), (x_3 y_3)\}$
$(1,2), (3,7), (5,8)$

3-fold CV:-

Divide "D" into three groups

D_1, D_2, D_3 of size $n/k = 3/3 = 1$

Let $D_1 = \{(x_1 y_1)\} \rightarrow$ validation set ; $D/D_1 = \{\overset{(3,7)}{(x_2 y_2)}, \overset{(5,8)}{(x_3, y_3)}\}$ as training set

$\hat{\beta}_1 = 0.5$; $\hat{\beta}_0 = 5.50$

$$mse_j = \frac{1}{n_j} \sum_{i \in (x_i, y_i) \in D_j} (y_i - \hat{f}(x_i | D/D_j))^2$$

$\therefore mse_1 = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_1)^2$

$= (2 - 5.50 - 0.5 \times 1)^2 = 16$

Let $D_2 = \{\overset{(3,7)}{(x_2, y_2)}\}$; $D/D_2 = \{\overset{(1,2)}{(x_1 y_1)}, \overset{(5,8)}{(x_3 y_3)}\}$

$\hat{\beta}_1 = 1.5$ $\qquad \hat{\beta}_0 = 0.5$

$\therefore mse_2 = (7 - 0.5 - 1.5 \times 3)^2 = 4$

Let $D_3 = \{\overset{(5,8)}{(x_3, y_3)}\}$; $D/D_3 = \{\overset{(1,2)}{(x_1, y_1)} \overset{(3,7)}{(x_2, y_2)}\}$

$\hat{\beta}_1 = 2.5$ $\qquad \hat{\beta}_0 = -0.5$

$\therefore mse_3 = (8 + 0.5 - 2.5 \times 5)^2 = 16$

$\therefore CV_{(3)} = \sum_{j=1}^{3} \frac{n_j}{n} mse_j \Rightarrow \frac{1}{3}[16 + 4 + 16] \Rightarrow 12$ ✱

Scanned by CamScanner

4. (a)

$$l(\beta_0,\beta_1) = \prod_{i=1}^{n} P(Y=y_i \mid X=x_i ; \beta_0,\beta_1)$$

$$= \prod_{i:y_i=1}^{n} p(x_i;\beta_0,\beta_1) \prod_{i:y_i=-1}^{n} (1-p(x_i;\beta_0,\beta_1))$$

$$\Rightarrow \log l(\beta_0,\beta_1) = \sum_{i=1}^{n}\left[\tfrac{1}{2}(y_i+1)\log p(x_i;\beta_0,\beta_1) + \tfrac{1}{2}(1-y_i)\log(1-p(x_i;\beta_0,\beta_1))\right]$$

$$= \sum_{i=1}^{n}\left[\tfrac{1}{2}(y_i+1)\log\frac{1}{1+e^{-y_i(\beta_0+\beta_1 x_i)}} + \tfrac{1}{2}(1-y_i)\log\frac{1}{1+e^{-y_i(\beta_0+\beta_1 x_i)}}\right]$$

$$= \sum_{i=1}^{n}\log\frac{1}{1+e^{-y_i(\beta_0+\beta_1 x_i)}}$$

$$\Rightarrow \hat{\beta} = \arg\max_{\beta}\sum_{i=1}^{n}\log\frac{1}{1+e^{-y_i(\beta_0+\beta_1 x_i)}}$$

$$= \arg\min_{\beta}\sum_{i=1}^{n}\log(1+e^{-y_i(\beta_0+\beta_1 x_i)})$$

$$\Rightarrow \hat{\beta} = \arg\min_{\beta}\sum_{i=1}^{n}\log(1+e^{-y_i(\beta_0+\beta_i x_i)})$$

(b)

$$\hat{Y} = \arg\max_{k\in\{1,\dots,K\}} P(Y=k \mid X=x)$$

$$= \arg\max_{k\in\{1,\dots,K\}} \frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}$$

$$= \arg\max_{k\in\{1,\dots,K\}} \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}}\exp\left(\frac{-1}{2\sigma_k^2}(x-\mu_k)^2\right)}{\sum_{l=1}^{K}\pi_l\frac{1}{\sqrt{2\pi\sigma_l^2}}\exp\left(\frac{-1}{2\sigma_l^2}(x-\mu_l)^2\right)}$$

$$= \arg\max_{k\in\{1,\dots,K\}} \log\pi_k - \frac{1}{2\sigma_k^2}(x-\mu_k)^2 - \frac{1}{2}\log\sigma_k^2$$

Q.5(a) given predictors (age, study, hw)

Backward stepwise selection algorithm

1> $p=3$; M_3 (full model): (age + study + hw) RSS 14.4

2> for M_2: choose the best (ie with smallest RSS)

among $\underbrace{(\text{age + study})}_{RSS:\ 18.2}$; $\underbrace{(\text{age + hw})}_{17.4}$; $\underbrace{(\text{study + hw})}_{17.6}$

$\therefore$ M_2 = (age + hw) RSS: 17.4

3> for M_1 choose the best among age, hw

M_1 = (hw) RSS: 22.3

Q.5(b) $C_p = \frac{1}{n}\left[RSS + 2(p+1)\hat{\sigma}^2\right]$, $n=10$

where

$\hat{\sigma}^2$ = Estimate of variance of error ϵ (with full model

$= \frac{RSS}{n-p-1} = \frac{14.4}{10-4} = \frac{14.4}{6} = 2.4$

"3"

for M_3 :

$$C_p = \frac{1}{10}\left[14.4 + 2(3+1)\times 2.4\right] = 3.66$$

for M_2 :

$$C_p = \frac{1}{10}\left[17.4 + 2(2+1)\times 2.4\right] = 3.18$$

for M_1 :

$$C_p = \frac{1}{10}\left[22.3 + 2(1+1)\times 2.4\right] = 3.19$$

choose the model that has smallest "C_p"

from above, Model $M_2 = (age + hw)$ has least C_p

$\therefore$ M_2 is the best model

Question 6:

a) Cross validation has less bias and tends not to over estimate the test error rate as much as validation set does, because CV repeatedly fit the statistical learning method using entire training data, where as in validation set approach uses part of of the size of original data set.

b) By using Linear model, $p(x) = \beta_0 + \beta_1 x$ may go beyond $[0, 1]$

$\therefore$ we must model $p(x)$ using a function that gives outputs between 0 and 1 $\forall$ x.

logistic function is one as such.

c) $t_j = \dfrac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)}$ should be large;

which measures how far $\hat{\beta}_j$ is to 0 relative to $\hat{SE}(\hat{\beta}_j)$.

Scanned by CamScanner