Dhanashree Kharate

1. a. Ridge regression parameter estimate

$$\hat{\beta_r} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \|\beta\|^2 \quad (1)$$

Show that

$$\hat{\beta_r} = [X^T X + \lambda I]^{-1} X^T Y$$

$$\hat{\beta}_r = \arg\min_{\beta} (Y - \beta^T X)^T (Y - \beta^T X) + \lambda \beta^T \beta$$

minimizing above equation

$$= \frac{\partial}{\partial \beta} (Y - \beta^T X)^T (Y - \beta^T X)$$

$\because \nabla_w (b^T \bar{w}^T A \bar{w}) = 2A\bar{w}$

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

$\because \cancel{\nabla_w (\bar{w}^T b) = b}$

$\therefore \bar{w}^T \bar{b} = (\bar{w}^T b)^T = b^T \bar{w}$

$$\frac{\partial}{\partial \beta} \lambda \beta^T \beta = 2\lambda\beta$$

$$0 = -\cancel{2}X^T (Y - \beta^T X) + \cancel{2}\lambda\beta$$

$$0 = -X^T Y + X^T \beta^T X + \lambda \beta$$

$$X^T Y = (\cancel{X^T} X + \cancel{\lambda I})$$

$$\hat{\beta}_r = \beta (X^T X + \lambda I)^{-1} X^T Y$$

b. $\hat{\beta} = [X^T X]^{-1} X^T Y$

$\hat{\beta}_r$ can be viewed as MLE ?

Yes, it can be when $\lambda = 0$ $p < 2$

$$\hat{\beta}_r = [X^T X + \lambda I]^{-1} X^T Y$$

$$= [X^T X] X^T Y$$

$$||\hat{\beta}_r|| < ||\hat{\beta}||$$

$$\hat{\beta} = [X^T X]^{-1} X^T Y$$

$$= \frac{X^T Y}{X^T X}$$

$$\hat{\beta} = X^T Y \qquad X^T X = I$$

$$\hat{\beta}_r = (X^T X + \lambda I)^{-1} X^T Y$$

$$= \frac{1}{1+\lambda} X^T Y$$

$$\hat{\beta}_r = \frac{1}{1+\lambda} \hat{\beta}$$

Scanned by CamScanner

1. c.

S.T. $\hat{\beta}_{MAP} = \hat{\beta}_r$

$$\hat{\beta}_{MAP} = \arg\max_\beta f_{\beta|y}(\beta|y)$$

$$p\ f_{\beta|y}(\beta|y) = \text{constant} \times f_{y|\beta}(y|\beta)\, f_\beta(\beta)$$

$$\hat{\beta}^{ridge} = \arg\min_\beta \sum_{n=1}^{N} \frac{1}{2}(y_n - \beta^T x_n)^2 + \lambda \sum_{i=1}^{p} \beta_i^2$$

$$\hat{\beta}^{MAP} = \arg\max_\beta \{\log P(\beta | x_{1:N}, y_{1:N}, \lambda)\}$$

$$= \arg\max_\beta \{\log (P(y_{1:N}|x_{1:N}, \beta) \sum_{i=1}^{p} P(\beta_i/\lambda))\}$$

$$= \arg\max_\beta \{\log P(y_{1:N}|x_{1:N}, \beta) + \sum_{i=1}^{p} \log(\beta_i/\lambda)\}$$

$$P(y_{1:N}|x_{1:N}, \beta) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{\frac{-1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \beta^T x_n)^2\right\}$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{\frac{-1}{2\sigma^2} RSS(\beta)\right\}$$

$$\arg\max_\beta \{\log P(y_{1:N}/x_{1:N}, \beta)\} = \arg\max_\beta \{-RSS(\beta)\}$$

$$\beta_i \sim N(0, 1/2\lambda) \quad P(\beta_i/\lambda) = \frac{1}{\sqrt{2\pi/\lambda}} \exp\left(\frac{-\beta_i^2 \lambda}{2}\right)$$

Scanned by CamScanner

variance of β is $\lambda/2$.

$\lambda \uparrow$ MAP estimate diverges from MLE

d. $\hat{\beta}_{LASSO} = \underset{\beta}{\arg\min} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2 + \lambda \|\beta\|$

Lasso L_1 penalty is equivalent to assuming Laplace distribution of β values

$$p(\beta_i) \propto \exp\{\lambda |\beta_i|\}$$

e. i) $\|\hat{\beta}_r\| < \|\hat{\beta}\|$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$= \frac{X^T Y}{X^T X}$$

$$\hat{\beta} = X^T Y \qquad \because X^T X = I$$

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$$

$$= \frac{1}{1+\lambda} X^T Y$$

$$\hat{\beta}_{ridge} = \frac{1}{1+\lambda} \hat{\beta}$$

$$\|\hat{\beta}_{ridge}\| < \|\hat{\beta}\|$$

Scanned by CamScanner

e ii) $\mathrm{Cov}[\hat{\beta}_r] < \mathrm{Cov}[\hat{\beta}]$

$$\mathrm{Cov}[\hat{\beta}_r] = \sigma_\epsilon^2 [X^TX + \lambda I]^{-1} [X^TX] [X^TX + \lambda I]^{-1}$$

$$= \sigma_\epsilon^2 \frac{X^TX}{[X^TX + \lambda I][X^TX + \lambda I]}$$

$\because X^TX = I$ $$= \frac{\sigma_\epsilon^2}{(1+\lambda)(1+\lambda)}$$

$$\mathrm{Cov}[\hat{\beta}_r] = (1+\lambda)^{-2} \sigma_\epsilon^2$$

$$\mathrm{Cov}[\beta] = \sigma^2 [X^TX]^{-1}$$
$$= \sigma^2$$

$$\frac{\sigma_\epsilon^2}{(1+\lambda)^2} < \sigma^2$$

2. $$y_i = \frac{\gamma_0 x_i}{\gamma_1 + x_i} + \epsilon_i$$

$$y_i' = \beta_0 + \beta_1 x_i'$$

$$y_i' = \frac{1}{y_i} \qquad \beta_0 = \frac{1}{\gamma_0} \qquad \beta_1 = \frac{\gamma_1}{\gamma_0} \qquad x_i' = \frac{1}{x_i}$$

$$y_i' = \beta_0 + \beta_1 x_i'$$

$$= \frac{1}{\gamma_0} + \frac{\gamma_1}{\gamma_0}\frac{1}{x_i}$$

$$y_i' = \frac{x_i + \gamma_1}{\gamma_0 x_i}$$

$$= \frac{1 + x_i' \gamma_1}{\gamma_0}$$

$$y_i' = \frac{1 + x_i' \hat{\beta}_1 / \hat{\beta}_0}{1/\hat{\beta}_0}$$

$$\hat{y}_i = \hat{\beta}_0 + x_i' \hat{\beta}_1$$

9. a

$$Err_{in} = \frac{1}{n}\sum_{i=1}^{n} E\left[L(y_i, \hat{g}(x_i))|_{x=x_i}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left[\left[y_i - \hat{g}(x_i)\right]^2_{x=x_i}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} E\left[(y_i - \hat{g}(x_i))(y_i - \hat{g}(x_i))^T|_{x=x_i}\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} (E(y_i) - \hat{g}(x_i))(E(y_i) - \hat{g}(x_i))^T + E\left[(y_i - E[\hat{g}(x_i)])(\hat{y}_i - E[\hat{g}(x_i)])^T\right]$$

Thus,

$$Err_{in} = \frac{1}{n} E[\overline{err}] + \frac{2}{n}\sum_{i=1}^{n} Cov(y_i, \hat{g}(x_i))$$ [from ① and ②

$$\because E[\overline{err}] = \frac{1}{n} E[(y-\hat{y})^T (y-\hat{y})] \quad ①$$

$$\because Cov(y_i, \hat{g}(x_i)) = E\left[(y_i - E[\hat{g}(x_i)])(y_i - E[\hat{g}(x_i)])^T\right] \quad ②$$

①

9.b.
9.Q.b. $y_i = g_i^T \theta + \epsilon_i$

$y_i = g^T(x_i)\theta + \epsilon_i$

In sample test data $y^* = G^*\theta + \epsilon^*$ ①

Training data $\underline{y} = \underline{G}\theta + \underline{\epsilon}$ ②

G is same for ① and ②

Training fits
$$\hat{y} = G\hat{\theta}$$
$$= G[G^TG]^{-1}G^T y$$
$$= G[G^TG]^{-1}G^T[G\theta + \epsilon]$$
$$= G + P\epsilon$$
$\uparrow$ Projection matrix $n \times n$

$$y^* - \hat{y} = G\theta + \epsilon^* - G\theta + P\epsilon$$
$$= \epsilon^* - P\epsilon$$

Training error $y - \hat{y} = [G\theta + \epsilon] - [G\theta + P\epsilon]$
$$= [I - P]\,\epsilon$$

$$Err_{in} = \frac{1}{n} E[(y^* - \hat{y})^T(y^* - \hat{y})]$$
$$= \frac{1}{n} E[(\epsilon^* - P\epsilon)^T(\epsilon^* - P\epsilon)]$$
$$= \frac{1}{n} E[\epsilon^{*T}\epsilon^* - 2\epsilon^{*T}P\epsilon + \epsilon^T P^T P\epsilon]$$

$$= \frac{1}{n}\{n\sigma_\epsilon^2 - 0\} + \frac{1}{n} E\{\epsilon^T P \epsilon\}$$

For projectn matrix $P^T = P$

trace $E\{\sigma^2 I P\}$

σ^2 trace(P)

p

no of regression coefficient

$$= \frac{1}{n}\{n\sigma_\epsilon^2 + \sigma_\epsilon^2 p\}$$

$$Err_{in} = \sigma_\epsilon^2 + \frac{\sigma_\epsilon^2 p}{n} \quad (1)$$

Excepted value of in sample error

$$E[\overline{err}] = \frac{1}{n} E[(y-\hat{y})^T(y-\hat{y})]$$

$$= \frac{1}{n} E[\epsilon^T[I-P]\epsilon]$$

$$= \frac{1}{n} tr\, E[\epsilon\epsilon^T[I-P]\}\}$$

$$= \frac{1}{n} tra\{\sigma_\epsilon^2 I[I-P]\}$$

$$= \frac{\sigma_\epsilon^2}{n} tr[I-P]$$

$$Err_{in} - E[\overline{err}] = \frac{2\sigma_\epsilon^2 p}{n} \quad (2)$$

From (1) $\sigma_\epsilon^2 = \frac{SSE}{n}$ $C_p \equiv \hat{Err}_{in} = \frac{SSE}{n} + \frac{2p\sigma_\epsilon^2}{n}$