MTH9898_HW2_Chenyu_Zhao

北京大學
PEKING UNIVERSITY

3.2 1) $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ $\quad X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ & \vdots & & \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}$ $\quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$

$\hat\beta = (X^TX)^{-1}X^Ty$ $\qquad Var\,\hat\beta = (X^TX)^{-1} Var(\varepsilon) = (X^TX)^{-1}\sigma^2$

$\hat y_0 = X_0^T\hat\beta$ $\qquad Var(\hat y_0) = X_0^T Var(\hat\beta) X_0 = X_0^T (X^TX)^{-1} X_0 \sigma^2$

$\because y|X \sim N$ $\quad \therefore \hat\beta \sim N$ $\quad \therefore \hat y_0 \sim N$

$\therefore$ 95% confidence interval for $\hat y_0$ $\quad [\hat y_0 - 1.96\sqrt{Var(\hat y_0)}, \hat y + 1.96\sqrt{Var(\hat y_0)}]$

which is $[\, X_0^T(X^TX)^{-1}X^Ty - 1.96\sqrt{x_0^T(X^TX)^{-1}x_0\sigma^2},\ X_0^T(X^TX)^{-1}X^Ty + 1.96\sqrt{x_0^T(X^TX)^{-1}x_0\sigma^2}\,]$

2) from 3.15

$$C_\beta = \{\beta \mid (\hat\beta-\beta)^T X^TX(\hat\beta-\beta) \le \hat\sigma^2 \chi^{2\,(1-\alpha)}_{p+1}\}$$

$$\therefore \quad X_{left} = \underset{\beta\in C_\beta}{\arg\max}\ \beta^T x$$

$$X_{right} = \underset{\beta\in C_\beta}{\arg\min}\ \beta^T x$$

The first interval should be wider. Because the first one treat every beta as independent normal distribution while the second one treat betas as joint-normal distribution. The second one is more accurate.

3.12 $L = (Y-X\beta)^T(Y-X\beta) + \lambda\beta^T\beta$

$\nabla L = 2X^TX\beta - 2X^TY + 2\lambda\beta \equiv 0$

$\hat{\beta} = (X^TX+\lambda I)^{-1}X^Ty$

$\tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I_{p\times p} \end{bmatrix}$ $\tilde{Y} = \begin{bmatrix} Y \\ 0_{p\times 1} \end{bmatrix}$ $\tilde{\beta} = (\tilde{X}^T\tilde{X})\tilde{X}^T\tilde{Y}$

$\tilde{X}^T\tilde{X} = \begin{bmatrix} X \\ \sqrt{\lambda} I_{p\times p} \end{bmatrix}^T \begin{bmatrix} X \\ \sqrt{\lambda} I_{p\times p} \end{bmatrix} = X^TX + \lambda I$

$\tilde{X}^T\tilde{Y} = [X^T, \sqrt{\lambda} I_{p\times p}] \begin{bmatrix} Y \\ 0 \end{bmatrix} = X^TY$

$\therefore \tilde{\beta} = (X^TX+\lambda I)X^TY$

$\tilde{\beta} = \hat{\beta}$

4.3 original space: $P(G=k \mid X=x) = \dfrac{f_k(x)\pi_k}{\sum^k f_l(x)\pi_l}$

$f_k(x) = N(\mu_k, \Sigma)$

$\log \dfrac{P(G=k \mid X=x)}{P(G=l \mid X=x)} = \log\dfrac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k+\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l) + X^T\Sigma^{-1}(\mu_k-\mu_l)$

$\therefore \delta_k(x) = X^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k$

after transformation, π_k remains same

$f_k(\hat{y}) = N(B^T\mu_k, B^T\Sigma B)$ $\hat{y} = B^Tx$

$\delta_k'(x) = X^TB(B^T\Sigma B)^{-1}B^T\mu_k - \frac{1}{2}\mu_k^TB(B^T\Sigma B)^{-1}B^T\mu_k + \log\pi_k$

$= \delta_k(x)$ if $B(B^T\Sigma B^{-1})B^T = \Sigma^{-1}$

indeed, because $B(B^T\Sigma B)^{-1}B^T\Sigma B = B$

$\therefore B(B^T\Sigma B)^{-1}B^T\Sigma = I$ $\therefore B(B^T\Sigma B)^{-1}B^T = \Sigma^{-1}$

$\therefore$ LDA using $\hat{y}$ should be same with using X