Exercise 4.2

$x \in \mathbb{R}^p$ Y has 2 classes

size N_1 size N_2

target code: $-\frac{N}{N_1}$ target code: $\frac{N}{N_2}$

(a) LDA classifies to class 2 if $\delta_2(x) > \delta_1(x)$

$$x^T\Sigma^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_2^T\Sigma^{-1}\hat{\mu}_2 + \log\frac{N_2}{N} >$$

$$x^T\Sigma^{-1}\hat{\mu}_1 - \frac{1}{2}\hat{\mu}_1^T\Sigma^{-1}\hat{\mu}_1 + \log\frac{N_1}{N}$$

$$x^T\Sigma^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^T\Sigma^{-1}(\hat{\mu}_2 - \hat{\mu}_1) - \log\frac{N_2}{N_1}$$

(b) $\min \sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2$

Y = target vector, target code $-\frac{N}{N_1}$, $\frac{N}{N_2}$

$Y = t_1 U_1 + t_2 U_2$ $t_1 = -\frac{N}{N_1}$, $t_2 = \frac{N}{N_2}$

$\mathbf{1} = U_1 + U_2$

U_1	$+U_2$	$=\mathbf{1}$
1	0	1
0	1	1
0	1	1
1	0	1

Note that $\hat{\mu}_i = \frac{X^T U_i}{N_i}$ $\Longrightarrow$ $X^T U_1 = N_1\hat{\mu}_1$

$X^T U_2 = N_2\hat{\mu}_2$

Therefore $X^T Y = X^T(t_1 U_1 + t_2 U_2) = t_1 N_1\hat{\mu}_1 + t_2 N_2\hat{\mu}_2$

Our goal is to minimize $(Y-\beta_0\mathbf{1}-X\beta)^T(Y-\beta_0\mathbf{1}-X\beta)$

$$\begin{aligned}\frac{\partial RSS}{\partial \beta_0}&\quad -(Y-\beta_0\mathbf{1}-X\beta)^T\mathbf{1}-\mathbf{1}^T(Y-\beta_0\mathbf{1}-X\beta)=0\\ \frac{\partial RSS}{\partial \beta}&\quad -(Y-\beta_0\mathbf{1}-X\beta)^TX-X^T(Y-\beta_0\mathbf{1}-X\beta)=0\end{aligned}$$

$$\rightarrow\begin{cases}\cancel{2}\beta_0\overbrace{\mathbf{1}^T\mathbf{1}}^{N}-\cancel{2}\mathbf{1}^T(Y-X\beta)=0\\ \cancel{2}X^TX\beta-\cancel{2}X^TY+\cancel{2}\beta_0X^T\mathbf{1}=0\end{cases}$$

$$\beta_0=\frac{1}{N}\mathbf{1}^T(Y-X\beta)$$

$$\rightarrow X^TX\beta-X^TY+\frac{1}{N}X^T\mathbf{1}\mathbf{1}^TY-\frac{1}{N}X^T\mathbf{1}\mathbf{1}^TX\beta=0$$

$$\underbrace{\left(X^TX-\frac{1}{N}X^T\mathbf{1}\mathbf{1}^TX\right)\beta}_{LHS}=\underbrace{X^TY-\frac{1}{N}X^T\mathbf{1}\mathbf{1}^TY}_{RHS}$$

$$\begin{aligned}RHS&=t_1N_1\hat{\mu}_1+t_2N_2\hat{\mu}_2-\frac{1}{N}\left(X^T(U_1+U_2)\right)(U_1+U_2)^TY\\
&=t_1N_1\hat{\mu}_1+t_2N_2\hat{\mu}_2-\frac{1}{N}\left(X^TU_1+X^TU_2\right)(U_1+U_2)^T(t_1U_1+t_2U_2)\\
&=t_1N_1\hat{\mu}_1+t_2N_2\hat{\mu}_2-\frac{1}{N}\left(N_1\hat{\mu}_1+N_2\hat{\mu}_2\right)\Big(t_1U_1^TU_1+\underbrace{\cancel{t_1U_2^TU_1}}_{=0}+\underbrace{\cancel{t_2U_1^TU_2}}_{=0}+t_2U_2^TU_2\\
&=t_1N_1\hat{\mu}_1+t_2N_2\hat{\mu}_2-\frac{1}{N}\left(N_1\hat{\mu}_1+N_2\hat{\mu}_2\right)\left(t_1N_1+t_2N_2\right)\\
&=t_1N_1\hat{\mu}_1+t_2N_2\hat{\mu}_2-\frac{N_1^2}{N}t_1\hat{\mu}_1-\frac{N_1N_2}{N}\hat{\mu}_1t_2-\frac{N_1N_2}{N}\hat{\mu}_2t_1-\frac{N_2^2}{N}t_2\hat{\mu}_2\\
&=t_1\hat{\mu}_1\left(N_1-\frac{N_1^2}{N}\right)+t_2\hat{\mu}_2\left(N_2-\frac{N_2^2}{N}\right)-\frac{N_1N_2}{N}\hat{\mu}_1t_2-\frac{N_1N_2}{N}\hat{\mu}_2t_1\\
&=t_1\hat{\mu}_1\left(\frac{N_1(N_1+N_2)-\cancel{N_1^2}}{N}\right)+t_2\hat{\mu}_2\left(\frac{N_2(N_1+N_2)-\cancel{N_2^2}}{N}\right)-\frac{N_1N_2}{N}\hat{\mu}_1t_2-\frac{N_1N_2}{N}\hat{\mu}_2t_1\\
&=\frac{N_1N_2}{N}(t_1-t_2)(\hat{\mu}_1-\hat{\mu}_2)\end{aligned}$$

$$LHS = X^T X - \frac{1}{N} X^T 1 1^T X$$

$$\Sigma = \sum_{k=1}^{K} \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N-2)$$

$$(N-2)\Sigma = X^T X - N_1 \hat{\mu}_1 \hat{\mu}_1^T - N_2 \hat{\mu}_2 \hat{\mu}_2^T$$

$$\text{so } X^T X = (N-2)\Sigma + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T$$

$$\frac{1}{N} X^T 1 1^T X = \frac{1}{N} X^T (U_1 + U_2)(U_1 + U_2)^T X$$

$$= \frac{1}{N} \left(N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2\right)\left(N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2\right)^T$$

$$= \frac{N_1^2 \hat{\mu}_1 \hat{\mu}_1^T}{N} + \frac{N_1 N_2}{N} \hat{\mu}_1 \hat{\mu}_2^T + \frac{N_1 N_2}{N} \hat{\mu}_2 \hat{\mu}_1^T + \frac{N_2^2}{N} \hat{\mu}_2 \hat{\mu}_2^T$$

$$= (N-2)\Sigma + N_1 \hat{\mu}_1 \hat{\mu}_1^T + N_2 \hat{\mu}_2 \hat{\mu}_2^T - \frac{N_1^2}{N} \hat{\mu}_1 \hat{\mu}_1^T - \frac{N_1 N_2}{N} \hat{\mu}_1 \hat{\mu}_2^T - \frac{N_1 N_2}{N} \hat{\mu}_2 \hat{\mu}_1^T - \frac{N_2^2}{N} \hat{\mu}_2 \hat{\mu}_2^T$$

$$= (N-2)\Sigma + \left(\frac{N_1(N_1 + N_2) - \cancel{N_1^2}}{N}\right) \hat{\mu}_1 \hat{\mu}_1^T + \left(\frac{N_2(N_1 + N_2) - \cancel{N_2^2}}{N}\right) \hat{\mu}_2 \hat{\mu}_2^T + {}$$
$$- \frac{N_1 N_2}{N} \hat{\mu}_1 \hat{\mu}_2^T - \frac{N_1 N_2}{N} \hat{\mu}_2 \hat{\mu}_1^T$$

$$\Sigma_B = \frac{N_1 N_2}{N^2} (\hat{\mu}_2 - \hat{\mu}_1)^T (\hat{\mu}_2 - \hat{\mu}_1)$$

$$= (N-2)\Sigma + \frac{N}{N} \frac{N_1 N_2}{N} (\hat{\mu}_2 - \hat{\mu}_1)^T (\hat{\mu}_2 - \hat{\mu}_1)$$

$$= (N-2)\Sigma + N \Sigma_B$$

$$\left[(N-2)\Sigma + N \Sigma_B\right]\beta = \frac{N_1 N_2}{N} (t_1 - t_2)(\hat{\mu}_1 - \hat{\mu}_2)$$

$$= \frac{N_1 N_2}{N} \left(-\frac{N}{N_1} - \frac{N}{N_2}\right)(\hat{\mu}_1 - \hat{\mu}_2)$$

$$= \frac{N_1 \cancel{N_2}}{\cancel{N}} \left(- \frac{\cancel{N}(N_1 + N_2)}{\cancel{N_1 N_2}}\right)(\hat{\mu}_1 - \hat{\mu}_2)$$

$$= N(\hat{\mu}_2 - \hat{\mu}_1)$$

(c) Note that $\hat{\Sigma}_B \beta = (\hat{\mu}_2 - \hat{\mu}_1) \underbrace{\frac{N_1 N_2}{N^2} (\hat{\mu}_2 - \hat{\mu}_1)^T \beta}_{\text{scalar}}$, it goes in the direction of $\checkmark$

Both the LHS and the RHS of

$$[(N-2)\Sigma + N\Sigma_B]\beta = N(\hat{\mu}_2 - \hat{\mu}_1)$$

go in the direction of $(\hat{\mu}_2 - \hat{\mu}_1)$, so the solution must be proportional to $\Sigma^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$

$$\lambda \Sigma^{-1}(\hat{\mu}_2 - \hat{\mu}_1)$$

(d) t_1 and t_2 were chosen arbitrarily, so the result holds

(e)
$$
\begin{aligned}
\hat{\beta}_0 &= \frac{1}{N} \mathbb{1}^T (Y - X\hat{\beta}) \\
&\overset{!}{=} \frac{1}{N} (U_1 + U_2)^T (t_1 U_1 + t_2 U_2) - \frac{1}{N} (U_1 + U_2)^T X \hat{\beta} \\
&\overset{!}{=} \frac{1}{N} (t_1 \underbrace{U_1^T U_1}_{=N_1} + t_2 \underbrace{U_1^T U_2}_{=0} + t_1 \underbrace{U_2^T U_1}_{=0} + t_2 \underbrace{U_2^T U_2}_{=N_2}) - \frac{1}{N} (U_1^T X + U_2^T X) \hat{\beta} \\
&\overset{!}{=} \frac{1}{N} \left(-\frac{N}{\cancel{N_1}} \cdot \cancel{N_1} + \frac{N}{\cancel{N_2}} \cdot \cancel{N_2}\right) - \frac{1}{N} (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) \beta \\
&= -\frac{1}{N} (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) \hat{\beta}
\end{aligned}
$$

$$
\begin{aligned}
\hat{f}(x) &= \hat{\beta}_0 + x^T \hat{\beta} \\
&\overset{!}{=} -\frac{1}{N} (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) \hat{\beta} + x^T \hat{\beta} \\
&\overset{!}{=} \frac{1}{N} (N x^T - N_1 \hat{\mu}_1 - N_2 \hat{\mu}_2) \hat{\beta} \\
&\overset{!}{=} \frac{1}{N} (N x^T - N_1 \hat{\mu}_1 - N_2 \hat{\mu}_2) \lambda \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) \qquad \lambda \in \mathbb{R}
\end{aligned}
$$

$$\hat{f}(x) > 0 \iff N x^T \lambda \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) > (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) \lambda \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$$

$$x^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{N} (N_1 \hat{\mu}_1 + N_2 \hat{\mu}_2) \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$$

LDA (4.11): $x^T \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2} (\hat{\mu}_1 + \hat{\mu}_2) \hat{\Sigma}^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$

The two rules are equal if $N_1 = N_2 = \frac{N}{2}$

Linear Discriminant Analysis

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

$$\Pr[G=k \mid X=x] = \frac{f_k(x)\,\hat{\pi}_k}{\sum_{\ell=1}^{K} f_\ell(x)\,\hat{\pi}_\ell}$$

$f_k(x)$ Gaussian $\rightarrow$ LDA, QDA

$\Sigma_k = \Sigma \quad \forall k \rightarrow$ LDA

When we do NOT assume $\Sigma_k = \Sigma \; \forall k$, then QDA

$$\delta_k(x) = -\frac{1}{2} \log|\Sigma_k| - \frac{1}{2} (x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k) + \log \hat{\pi}_k$$

QDA the quadratic term does NOT simplify

- estimates for QDA are the same as those for LDA, but

$$\hat{\Sigma}_k = \sum_{g_i = k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N_k - 1)$$

- similar techniques difference in # of parameters to estimate
  - for LDA, # pars $= (K-1) \times (p-1)$ $\rightarrow$ # parameters to estimate for each difference

    $(K-1)$: all differences to contrast $\delta_k - \delta_1$
  - for QDA, # pars $= (K-1) \times [\{p(p+3)/2\} + 1]$

Note: both methods perform quite well in a large number of situations
- data support only linear (or quadratic) decision boundaries
- Gaussian models are stable

Regularized Discriminant Analysis

- idea: create a sort of compromise between LDA and QDA
- we allow differences among the covariance matrices, but we shrink them toward Σ ← similar to ridge

$$\hat{\Sigma}_k(\alpha) = \alpha \hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$$

where $\alpha \in [0;1]$ should be chosen (i.e., by cross-validation)
α controls the amount of shrinkage

$\alpha = 0 \longrightarrow$ LDA
$\alpha = 1 \longrightarrow$ QDA

- further possibility is to shrink $\hat{\Sigma}$ toward $\sigma^2 I$

$$\hat{\Sigma} = \gamma \hat{\Sigma} + (1-\gamma)\sigma^2 I$$

where $\gamma \in [0;1)$ has a similar meaning than α

- Combining

$$\hat{\Sigma}_k(\alpha,\gamma) = \alpha \hat{\Sigma}_k + (1-\alpha)\left[\gamma \Sigma + (1-\gamma)\sigma^2 I\right]$$

we obtain a general family for the covariate matrix, which depends on α, γ

Reduced-rank LDA

Fisher: find the best combination $z = aX$ such that the between-class variance is maximized relative to the within-class variance

Total variance $T = \underline{B} + \underline{W}$

x_i from class k
g_i class of x_i

$$T = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T$$

$$= \frac{1}{N}\sum_{k=1}^{K}\sum_{g_i=k}\left((x_i - \bar{x}_k) + (\bar{x}_k - \bar{x})\right)\left((x_i - \bar{x}_k) + (\bar{x}_k - \bar{x})\right)^T$$

$$= \frac{1}{N}\Bigg[\sum_{k=1}^{K}\sum_{g_i=k}(x_i - \bar{x}_k)(x_i - \bar{x}_k) + \sum_{k=1}^{K}\Big[\underbrace{\sum_{g_i=k}(x_i - \bar{x}_k)}_{=0}\Big](\bar{x}_k - \bar{x})^T +$$

$$+ \sum_{k=1}^{K}(\bar{x}_k - \bar{x})\underbrace{\sum_{g_i=k}(x_i - \bar{x}_k)^T}_{=0} + \sum_{k=1}^{K}\sum_{g_i=k}(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T\Bigg]$$

$$= \underbrace{\frac{1}{N}\sum_{k=1}^{K}\sum_{g_i=k}(x_i - \bar{x}_k)(x_i - \bar{x}_k)^T}_{\text{within-class variance } W} + \underbrace{\frac{1}{N}\sum_{k=1}^{K}N_k(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T}_{\text{between-class variance } B}$$

a: $a^T B a$ is maximized
$a^T W a$ is minimized

a_1 : $\arg\max_a \dfrac{a^T B a}{a^T W a}$ $\rightsquigarrow$ $\max_a a^T B a$ subject to $a^T W a = 1$

→ generalized eigenvalue problem, a corresponds to the largest eigenvalue of $W^{-1}B$

a_2 : $a_2 \perp a_1$ $\quad \max_a \dfrac{a^T B a}{a^T W a}$

a_3 ...

$a_1, a_2, \ldots$ are called:
- discriminant coordinates
- canonical variates

Motivations
- initially data reduction for visualization
- can be seen as a restricted classification rule

the centroids lie in the L-dimensional subspace of $\mathbb{R}^p$

Logistic regression

- model the posterior probabilities of the K classes, s.t.
  - linear functions in x
  - sum of them $= 1$
  - they $\in [0;1]$

Logistic regression models

$$\log \frac{Pr[G=1 \mid X=x]}{Pr[G=K \mid X=x]} = \beta_{10} + \beta_1^T x$$

$$\log \frac{Pr[G=2 \mid X=x]}{Pr[G=K \mid X=x]} = \beta_{20} + \beta_2^T x$$

$$\vdots$$

$$\log \frac{Pr[G=K-1 \mid X=x]}{Pr[G=K \mid X=x]} = \beta_{K-1\,0} + \beta_{K-1}^T x$$

- specifies $K-1$ log-odd
- based on the logit transformation

$K=2$ ≥ 2

$$\log \frac{p}{1-p} = x\beta \iff p = \frac{e^{x\beta}}{1+e^{x\beta}}$$

$$Pr[G=1 \mid X=x] = p = \frac{e^{x\beta^T}}{1+e^{x\beta^T}}$$

$$Pr[G=2 \mid X=x] = 1 - Pr[G=1 \mid X=x] = 1-p = \frac{1}{1+e^{x\beta^T}}$$

$$\frac{e^{x\beta^T}}{1+e^{x\beta^T}} + \frac{1}{1+e^{x\beta^T}} = 1$$

$\beta = \beta_0\, \beta$

$X = (1, x)$

$$\hat{\beta} = \arg\max_{\beta} L(\beta)$$

$$L(\beta) = \prod_{i=1}^{N} f(y_i; p(x_i; \beta))$$

$$= \prod_{i=1}^{N} \cancel{\binom{1}{y_i}}\, p(x_i;\beta)^{y_i} (1 - p(x_i;\beta))^{1-y_i}$$

$$\ell(\beta) = \sum_{i=1}^{N} y_i \log p(x_i;\beta) + (1 - y_i) \log(1 - p(x_i;\beta))$$

$$= \sum_{i=1}^{N} \left[y_i \log \frac{e^{x\beta^T}}{1+e^{x\beta^T}} + \log\left(1 - \frac{e^{x\beta^T}}{1+e^{x\beta^T}}\right) - y \log\left(1 - \frac{e^{x\beta^T}}{1+e^{x\beta^T}}\right)\right]$$

$$= \sum_{i=1}^{N} \left[y_i x_i \beta^T - \cancel{y_i \log(1+e^{x\beta^T})} + \log\left(\frac{1+\cancel{e^{x\beta^T}} - \cancel{e^{x\beta^T}}}{1+e^{x\beta^T}}\right) \cancel{+ y \log\left(\frac{1}{1+e^{x\beta^T}}\right)}\right]$$

$$= \sum_{i=1}^{N} y_i x_i \beta^T - \log(1 + e^{x\beta^T})$$

$$\ell_\beta(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left[y_i x_i - \frac{e^{x\beta^T}}{1+e^{x\beta^T}} x_i \right] = 0$$

$$= \sum_{i=1}^{N} x_i (y_i - p(x_i;\beta)) = 0$$

system of $p+1$ equations not linear in β

→ to find the solution, we can use the Newton-Raphson algorithm

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

$f(x_n) \leftarrow \ell_\beta(\beta)$

$f'(x_n) \leftarrow \ell_{\beta\beta}(\beta) = \dfrac{\partial^2 \ell(\beta)}{\partial\beta\,\partial\beta^T}$

$$\ell_{\beta\beta}(\beta) = -\sum_{i=1}^{N} \left(x_i e^{x_i\beta^T}(1 + e^{x_i\beta^T}) - x_i e^{x_i\beta^T} e^{x_i\beta^T} \right) x_i^T \Big/ (1 + e^{x\beta^T})^2$$

$$= -\sum_{i=1}^{N} x_i \frac{e^{x_i\beta^T}(1 + \cancel{e^{x_i\beta^T}} - \cancel{e^{x_i\beta^T}})}{(1+e^{x_i\beta^T})^2} x_i^T$$

$$= -\sum_{i=1}^{N} x_i \frac{e^{x_i\beta^T}}{(1+e^{x_i\beta^T})^2} x_i^T$$

Newton - Raphson

$$\beta_{new} = \beta_{old}\ \ell_{\beta\beta}(\beta)^{-1}\ell_\beta(\beta)$$

$$\ell_\beta(\beta) = X^T(y-p) \qquad p = \frac{e^{x\beta^T}}{1+e^{x\beta^T}}$$

$$\ell_{\beta\beta}(\beta) = -X^TWX \qquad \text{where } W = \begin{bmatrix} \ddots & & \\ & p(1-p) & \\ & & \ddots \end{bmatrix}$$

$$\underbrace{\frac{e^{x\beta^T}}{1+e^{x\beta^T}}}_{p} \cdot \underbrace{\frac{1}{1+e^{x\beta^T}}}_{(1-p)} = \frac{e^{x\beta^T}}{(1+e^{x\beta^T})^2}$$

β in W and p are β_{old}

$$\beta_{new} = \beta_{old} + \overbrace{(X^TWX)^{-1}}^{\ell_{\beta\beta}(\beta)}\overbrace{X^T(y-p)}^{\ell_\beta(\beta)}$$

$$= (X^TWX)^{-1}X^TW\underbrace{(X\beta_{old} + W^{-1}(y-p))}_{z}$$

$$= (X^TWX)^{-1}X^TWz \longrightarrow \text{weighted least square}$$

- repeat the steps of the Newton-Raphson algorithm until it converges to $\hat{\beta}$

L_1 regularized logistic regression

The L_1 penalty (LASSO) can be applied to the logistic regression as well

$$\hat{\beta} = \arg\min_\beta \left\{-\ell(\beta) + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

$$= \arg\max_\beta \left\{\ell(\beta) - \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

$$= \arg\max_\beta \left\{\sum_{i=1}^{N}\left[y_i(\beta_0+\beta_i^Tx) - \log(1+e^{\beta_0+\beta^Tx})\right] - \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

Exercise:

- try to reproduce Table 3.3 with the data from the South Africa heart disease example

$\hat{\beta}$	LS	bern(α=0.05)	LASSO	RIDGE	LDA	QDA

- read ch 4.4.5