1. According to lagrange multiplier,

$$l(\lambda) = f(x) - \lambda\, g(x)$$

$$f(a) = a^T B a$$

$$g(a) = a^T W a - 1 = 0$$ (equating $g(a) = 0$ according to the law)

$$l(\lambda) = a^T B a - \lambda(a^T W a - 1)$$

$$\frac{dl}{da} = \frac{d}{da}(a^T B a) - \lambda \frac{d}{da}(a^T W a - 1)$$

$$= (1 \times Ba) + a^T (B \times 1) - \lambda(1 \times Wa + a^T (W \times 1 - 0))$$

$$= Ba + a^T B - \lambda(Wa + a^T W)$$

$$= Ba + B^T a - \lambda(Wa + W^T a)$$ [B & W are co-variance & symmetrical]

$$= a(B + B^T) - \lambda(W + W^T) a$$

To find the maxima, let it equal to 0.

$$a(B + B^T) - \lambda a (W + W^T) = 0$$

$[(W + W^T)]^{-1}(B + B^T) a = \lambda a$. Assuming W & B are symmetric, this is a standard eigenvalue problem.

2(a). According to equation 4.33 from the textbook for the LDA rule to classify to class 2, the ratio of ~~pro~~ posterior probability of N_2 & N_1 should be greater than 1.

$$\log \frac{P_r(G=K \mid X=x)}{P_r(G=K \mid X=x)} > 0 \quad \text{where, } K=N_2 \text{ \& } K=N_1$$

$$\log \frac{N_2}{N_1} - \frac{1}{2}(\mu_2+\mu_1)^T \Sigma^{-1}(\mu_2-\mu_1) + x^T \Sigma^{-1}(\mu_2-\mu_1) > 0$$

$$x^T \Sigma^{-1}(\mu_2-\mu_1) > \frac{1}{2}(\mu_2+\mu_1)^T \Sigma^{-1}(\mu_2-\mu_1) - \log(N_2/N_1)$$

Therefore, substituting the target, we get

$$x^T \Sigma^{-1}(\mu_2-\mu_1) > \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T \Sigma^{-1} + \log\frac{N_1}{N} - \log\frac{N_2}{N}$$

4.2(b) In order to minimize the given expression, it is important to satisfy the normal equation which is $X^T X \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} = X^T y$ — eq(1)

Expanding the left hand side of the equation for $X^T X$

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_{N_1} & x_{N_1+1} & x_{N_1+2} & \cdots & x_{N_1+N_2} \end{bmatrix} \begin{bmatrix} 1 & x_1^T \\ 1 & x_2^T \\ \vdots & \vdots \\ 1 & x_N^T \\ 1 & x_{N_1+1}^T \\ 1 & x_{N_1+2}^T \\ \vdots & \vdots \\ 1 & x_{N_1+N_2}^T \end{bmatrix}$$

Multiplying the two matrix we get:

$$\begin{bmatrix} N & \sum_{i=1}^{N} x_i^T \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i x_i^T \end{bmatrix} \text{———} (2)$$

when our response is coded as $-N/N_1$ & $+\frac{N}{N_2}$ for the classes, the right hand side of eq(1) can be written as:

$$\begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_{N_1} & x_{N_1+1} & x_{N_1+2} & \cdots & x_{N_1+N_2} \end{bmatrix} \begin{bmatrix} -N/N_1 \\ -N/N_1 \\ \vdots \\ -N/N_1 \\ N/N_2 \\ N/N_2 \\ \vdots \\ N/N_2 \end{bmatrix}$$

After multiplying the two matrices we get:

$$\begin{bmatrix} N_1\left(-\frac{N}{N_1}\right) + N_2\left(\frac{N}{N_2}\right) \\ \left(\sum_{i=1}^{N_1} x_i\right)\left(-N/N_1\right) + \left(\sum_{i=N_1+1}^{N} x_i\right)\left(N/N_2\right) \end{bmatrix} = \begin{bmatrix} 0 \\ -N\mu_1 + N\mu_2 \end{bmatrix}$$

By introducing class specific means we get.

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N_1} x_i + \sum_{i=N_1+1}^{N} x_i = N_1\mu_1 + N_2\mu_2 \quad (3)$$

By Pooling all of the samples for both class, $K=2$,
we can estimate the covariance matrix $\hat{\Sigma}$

$$\hat{\Sigma} = \frac{1}{N-K}\sum_{K=1}^{K}\sum_{i:g_i=K}(x_i-\mu_K)(x_i-\mu_K)^T$$

$$= \frac{1}{N-2}\left[\sum_{i:g_i=1}(x_i-\mu_1)(x_i-\mu_1)^T + \sum_{i:g=2}(x_i-\mu_2)(x_i-\mu_2)^T\right]$$

$$= \frac{1}{N-2}\left[\sum_{i:g_i=1} x_i x_i^T - N_1\mu_1\mu_1^T + \sum_{i:g_i=1} x_i x_i^T - N_2\mu_2\mu_2^T\right]$$

Scanned by CamScanner

Following equation (2)

$$\sum_{i=1}^{N} x_i x_i^T = (N-2)\hat{\Sigma} + N_1 \mu_1 \mu_1^T + N_2 \mu_2 \mu_2^T$$

Writing both side of (1) as linear system

$$x^T x \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix} = X^T y$$

$$\begin{bmatrix} N & N_1 \mu_1^T + N_2 \mu_2^T \\ N_1 \mu_1 + N_2 \mu_2 & (N-2)\hat{\Sigma} + N_1 \mu_1 \mu_1^T + N_2 \mu_2 \mu_2^T \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix}$$

(3)

$$= \begin{bmatrix} 0 \\ -N\mu_2 + N\mu_2 \end{bmatrix} \quad \text{—— (4)}$$

$$N\beta_0 + (N_2 \mu_1^T + N_2 \mu_2^T)\beta = 0$$

Solving for β_0 in terms of β.

$$\beta_0 = \left(-\frac{N_1}{N}\mu_1^T - \frac{N_2}{N}\mu_2^T\right)\beta$$

—— (5)

Putting this in eq (4)

$$\left(N_1\mu_1 + N_2\mu_2\right)\left(-\frac{N_1}{N}\mu_1^T - \frac{N_2}{N}\mu_2^T\right)\beta +$$

$$\left((N-2)\hat{\Sigma} + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T\right)\beta = N\left(\mu_2 - \mu_1\right)$$

Outer product terms

$$= -\frac{N_1^2}{N}\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T -$$

$$\frac{N_2^2}{N}\cancel{N_2}\mu_2\mu_2^T + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T$$

$$= \left(-\frac{N_1^2}{N} + N_1\right)\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T + \left(-\frac{N_2^2}{N} + N_2\right)$$

$$\cdot\ \mu_2\mu_2^T$$

$$= \frac{N_1}{N}\left(-N_1 + N\right)\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T + \frac{N_2}{N}\left(-N_2 + N\right)\mu_2\mu_2^T$$

$$= \frac{N_1N_2}{N}\mu_1\mu_1^T - \frac{2N_1N_2}{N}\mu_1\mu_2^T + \frac{N_2N_1}{N}\mu_2\mu_2^T$$

$$= \frac{N_1N_2}{N}\left(\mu_1\mu_1^T - 2\mu_1\mu_2 - \mu_2\mu_2\right) = \frac{N_1N_2}{N}\left(\mu_1 - \mu_2\right)\left(\mu_1 - \mu_2\right)^T.$$

Here we have $N_1 + N_2 = N$. If we introduce the matrix

$$\hat{\Sigma}_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$$

Equation for β, $\left[(N-2)\hat{\Sigma} + \frac{N_1 N_2}{N}\hat{\Sigma}_B\right]\beta = N(\mu_2 - \mu_1)$. — (6)

2(c) Since $\hat{\Sigma}_B \beta$ is $(\mu_2 - \mu_1)(\mu_2 - \mu_1)^T \beta$ and Product $(\mu_2 - \mu_1)^T \beta$ is scalar, vector direction of $\Sigma_B \beta$ is given by $\mu_2 - \mu_1$, so eq 6 as ~~to~~ both the right hand side & the term $\frac{N_1 N_2}{N}\hat{\Sigma}_B$ are in the direction of $\mu_2 - \mu_1$ the sol^n β must be proportional to $\Sigma^{-1}(\mu_2 - \mu_1)$

2(d) If we use $x_i \in N$ was arbitrary & distinct, it proves the point.

2(e) If U_i is the n element vector with j^{th} element.

our target value Y as $t_1 U_1 + t_2 U_2$.

$U_1 + U_2 = 1$. Our estimates $\hat{\mu}_1, \hat{\mu}_2$ as

$x^T U_i = N_i \hat{\mu}_i$ & $X^T Y = t_1 N_1 \hat{\mu}_1 + t_2 N_2 \hat{\mu}_2$

Solving for 2(a) we get.

$$\hat{\beta}_0 = \frac{1}{N} 1^T (Y - X\hat{\beta})$$
$$= -\frac{1}{N} (N_1 \mu_1^T + N_2 \mu_2^T) \hat{\beta}$$

we can than write our predicted value

$$\hat{f}(x) = \hat{\beta}_0 + \beta^T x \cancel{}$$
$$= \frac{1}{N} (N x^T - N_1 \mu_1^T - N_2 \mu_2^T) \hat{\beta}$$
$$= \frac{1}{N} (N x^T - N_1 \mu_1^T - N_2 \mu_2^T) \lambda \Sigma^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$$

for some $\lambda \in R$, classification rule is $\hat{f}(x) > 0$

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

$$x^T \Sigma^{-1} (\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{N} (N_1 \hat{\mu}_1^T + N_2 \mu_2^T) \Sigma^{-1} (\hat{\mu}_2 - \hat{\mu}_1)$$

$\therefore$ which different than LDA rule unless $N_1 = N_2$

Scanned by CamScanner