

(a)

Bayes' theorem gives (eq. 4.7)

$$P(b=k | x=x) = \frac{f_k(x) \pi_k}{\sum_{k=1}^K f_k(x) \pi_k} \propto f_k(x) \pi_k$$

The optimal classification rule (given a 0-1 loss function)

$$\hat{G}(x) = \operatorname{argmax}_k P(b=k | x=x)$$

The denominator is the same for all classes k , and does therefore not affect our classification.Also, we can maximize $\log P(b=k | x=x)$.

$$\begin{aligned} \hat{G}(x) &= \operatorname{argmax}_k f_k(x) \pi_k \\ &= \operatorname{argmax}_k \{ \log f_k(x) + \log \pi_k \} \end{aligned}$$

In linear discriminant analysis (LDA) we assume $x | b=k \sim N(\mu_k, \Sigma_k)$ multivariate normal

$$f_k(x) = (\pi_k)^{-\frac{1}{2}} |\Sigma_k|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right\}$$

with $\Sigma_k = \Sigma$ for all k .Take $\log$ and collect terms that do not depend on the class k ,

$$\begin{aligned} \log(f_k(x) \pi_k) &= -\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) + \ln \pi_k + C \\ &= -\frac{1}{2} \left[\underbrace{x^T \Sigma^{-1} x}_{1 \times 1} - \underbrace{x^T \Sigma^{-1} \mu_k}_{1 \times p \quad p \times 1} - \underbrace{\mu_k^T \Sigma^{-1} x}_{p \times 1 \quad 1 \times p} + \underbrace{\mu_k^T \Sigma^{-1} \mu_k}_{p \times 1 \quad 1 \times p} \right] + \ln \pi_k + C \\ &= \underbrace{x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k}_{\delta_k(x), \text{ the linear discriminant function}} + \underbrace{(\ln \pi_k - \frac{1}{2} x^T \Sigma^{-1} x)}_{\text{independent of } k} \end{aligned}$$

Where

$$\operatorname{argmax}_k P(b=k | x=x) = \operatorname{argmax}_k \delta_k(x)$$

In the two-class problem ($K=2$), we assign to class 2 if

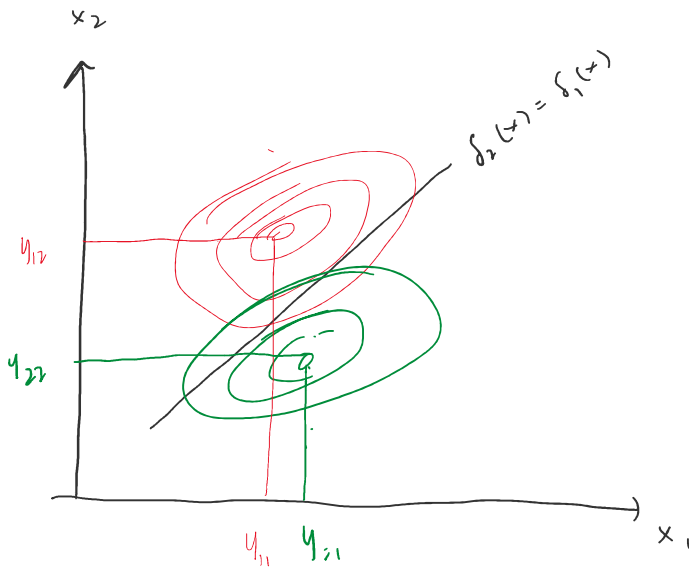
$$f_2(x) > f_1(x)$$

$$f_2(x) - f_1(x) > 0$$

$$x^T \Sigma^{-1} (\mu_2 - \mu_1) - \left(\frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 \right) + (\ln \pi_2 - \ln \pi_1) > 0$$

which we can write as

$$x^T a > b$$



(c)

$$\hat{\Sigma}_B \beta = (\mu_2 - \mu_1) \underbrace{(\mu_2 - \mu_1)^T}_{1 \times 1} \beta$$

$$\left[(N-2)I_p + \frac{N_1 N_2}{N} \underbrace{\Sigma^{-1} \hat{\Sigma}_B}_{1 \times 1} \right] \beta = N \Sigma^{-1} (\mu_2 - \mu_1)^T$$

$$\Rightarrow \hat{\beta} \propto \Sigma^{-1} (\mu_2 - \mu_1)^T$$

This means that the direction of (slope if $p=2$) the associated decision boundaries

is the same.

The value of the intercept are
generally different, leading to different decision rules.