Linear Discriminant Analysis (LDA).

Data $\{(y_i, x_i)_{i=1}^{n}\}$

Assumption: $X|Y=0 \sim N_p(\mu_0, \Sigma) =: f_0$. $X|Y=1 \sim N_p(\mu_1, \Sigma) =: f_1$.

Optimal Bayes Classifier $= I\{P(Y=1|X) \geq P(Y=0|X)\}$.

more generally, $\hat{k} = \arg\max_k P(Y=k|X)$

$\log f_0 = \frac{1}{2}\log|\Sigma| + \frac{1}{2}(X-\mu_0)^T\Sigma^{-1}(X-\mu_0)$ $(X|Y=0)$.

$\log f_1 = \frac{1}{2}\log|\Sigma| + \frac{1}{2}(X-\mu_1)^T\Sigma^{-1}(X-\mu_1)$. $(X|Y=1)$

$$P(Y=0|X) = \frac{P(Y=0, X)}{P(X)} = \frac{P(X|Y=0)P(Y=0)}{P(X|Y=0)P(Y=0) + P(X|Y=1)P(Y=1)}$$ Baye's Formulae.

$P(Y=0) = \pi$, $P(Y=1) = 1-\pi$, then,

$$P(Y=0|X) = \frac{\pi f_0}{\pi f_0 + (1-\pi) f_1}. \qquad P(Y=1|X) = \frac{(1-\pi) f_1}{\pi f_0 + (1-\pi) f_1}$$

$$P(Y=1|X) \geq P(Y=0|X) \leftrightarrow \frac{(1-\pi) f_1}{\pi f_0 + (1-\pi) f_1} \geq \frac{1}{2}$$

$$\leftrightarrow \frac{(1-\pi)}{\pi \cdot f_0/f_1 + (1-\pi)} \geq \frac{1}{2} \leftrightarrow (1-\pi) \geq \frac{1}{2}\pi\frac{f_0}{f_1} + \frac{1}{2}(1-\pi)$$

$$\leftrightarrow \frac{1}{2}(1-\pi) \geq \frac{1}{2}\pi \cdot \frac{f_0}{f_1} \leftrightarrow \frac{1-\pi}{\pi} \geq \frac{f_0}{f_1}. \leftrightarrow \log\frac{1-\pi}{\pi} \geq \log\frac{f_0}{f_1}.$$

$$\log\frac{f_0}{f_1} = \log f_0 - \log f_1 = -(X-\mu_0)^T\Sigma^{-1}(X-\mu_0) + (X-\mu_1)^T\Sigma^{-1}(X-\mu_1).$$

$$= -(\mu_0-\mu_1)^T\Sigma^{-1}X - \underbrace{\tfrac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \tfrac{1}{2}\mu_1^T\Sigma^{-1}\mu_1}_{(\#)} \leq \underbrace{\log\frac{1-\pi}{\pi}}_{(\circ)}.$$

$$\leftrightarrow (\mu_1-\mu_0)^T\Sigma^{-1}X \leq c, \quad c = (\circ) - (\#).$$

classify as class 1 if $(\mu_1 - \mu_0)^T \Sigma^{-1} x \le c$

$$c = \log\left(\frac{1-\pi}{\pi}\right) + \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 - \frac{1}{2}\mu_0^T \Sigma^{-1} \mu_0 .$$

$n = n_0 + n_1$.

If we can obtain (from data): $X: (X_1, \dots X_{n_0}, X_{n_0+1}, \dots, X_{n_0+n_1})$
$Y: (0 \dots, 0, 1, \dots, 1)$

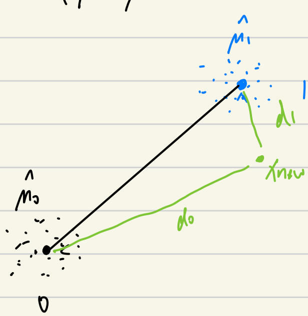
$$\hat{\mu}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} X_i \qquad \hat{\mu}_1 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} X_i . \qquad \hat{\pi} = \frac{n_0}{n}.$$

Let $\begin{cases} X_{ij} = X_j & \text{if } i = 0. \\ X_{ij} = X_{j+n_0} & \text{if } i = 1. \end{cases}$ $\begin{pmatrix} X_1, \dots, X_{n_0} = X_{01}, \dots, X_{0n_0}. \\ X_{n_0+1}, \dots, X_{n_0+n_1} = X_{11}, \dots, X_{1n_1} \end{pmatrix}$

$$\hat{\Sigma} = \frac{1}{n-2}\sum_{j=1}^{2}\sum_{i=1}^{n_j} (X_{ji} - \bar{X}_j)(X_{ji} - \bar{X}_j)^T .$$

We can classify new observation X_{new} by:

$$I\left((\hat{\mu}_1 - \hat{\mu}_0)^T \Sigma^{-1} X_{new} \le \hat{c}\right) = g_{bayes}(X_{new}).$$



$d_0 > d_1$

$\Rightarrow$ X_{new} is "closer" to $\hat{\mu}_1$ than $\hat{\mu}_0$.

$\Rightarrow$ classify as class 1.

Quadratic Discriminant Analysis (QDA).

$X|Y=0 \sim N_p(\mu_0, \Sigma_0)$. $\quad X|Y=1 \sim N_p(\mu_1, \Sigma_1)$.

$\log f_0 = \frac{1}{2}\log|\Sigma_0| + \frac{1}{2}(X-\mu_0)^T\Sigma_0^{-1}(X-\mu_0)$

$\log f_1 = \frac{1}{2}\log|\Sigma_1| + \frac{1}{2}(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)$

$\log f_0 - \log f_1 = \text{Const.} + \frac{1}{2}(X-\mu_0)^T\Sigma_0^{-1}(X-\mu_0) - \frac{1}{2}(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)$

$= \text{Const.} + \frac{1}{2}x^T(\Sigma_0^{-1} - \Sigma_1^{-1})x - x^T\Sigma_0^{-1}\mu_0 + \frac{1}{2}\mu_0^T\Sigma_0^{-1}\mu_0$.

$+ x^T\Sigma_1^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1$.

$= \text{Const.} + \frac{1}{2}x^T(\Sigma_0^{-1} - \Sigma_1^{-1})x - x^T(\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1) \geq \log\frac{1-\pi}{\pi}$

classify as class 1 if $\underbrace{\frac{1}{2}x^T(\Sigma_0^{-1} - \Sigma_1^{-1})x - x^T(\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1)}_{h(x)} \geq c$.

$c = \log\left(\frac{1-\pi}{\pi}\right) - \frac{1}{2}\mu_0^T\Sigma_0^{-1}\mu_0 + \frac{1}{2}\mu_1^T\Sigma_1^{-1}\mu_1 - \frac{1}{2}\log|\Sigma_0| + \frac{1}{2}\log|\Sigma_1|$

For a new observation X_{new},

$g_{bayes}(X_{new}) = I(h(X_{new}) \geq c)$

$$I\left(\log \frac{f_0}{f_1} \geq \frac{1-\pi}{\pi}\right).$$

Nonparametric Estimation of f_0, f_1.

LDA, QDA are restrictive in the way that we have to assume a $N_p(\cdot,\cdot)$

A more flexible approach is to estimate f_0, f_1 using nonparametric estimation.

HW3 Q2, Q3.

- KDE (Kernel Density Estimation).
- Smoothing Spline.

idea: optimal bayes classifier

$$g_{bayes}(X) = I\left\{\log\left(\frac{f_0(X)}{f_1(X)}\right) \geq \log \frac{1-\pi}{\pi}\right\}.$$

1. Obtain $\hat{f_0}$, $\hat{f_1}$, $\hat{\pi} = \frac{n_0}{n}$.

2. $$\hat{g}_{bayes}(X_{new}) = I\left\{\log\left(\frac{\hat{f_0}(X_{new})}{\hat{f_1}(X_{new})}\right) \geq \log\left(\frac{1-\hat{\pi}}{\hat{\pi}}\right)\right\}.$$

KDE: $f(x) \approx \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)$, $K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right)$, $K(\cdot)$ predetermined.

Popular choice of $K(\cdot) = \Phi(\cdot)$ standard normal cdf.

Best choice of $K(\cdot) = E\text{-Kernel} = K(u) = (1-u)^2 I(|u| \leq 1)$

Idea: $$f(x) = \frac{d}{dx}F(x) = \lim_{h\to 0}\frac{F(x+h) - F(x-h)}{2h} = \lim_{h\to 0}\frac{P(x-h \leq X \leq x+h)}{2h}$$

Let $P = F_n$ empirical measure, $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$, then,

$$f(x) = \lim_{h\to 0}\frac{1}{2h}\frac{1}{n}\sum_{i=1}^{n} I(x-h \leq X_i \leq x+h)$$

When h is small, then,

$$f(x) \approx \frac{1}{2nh}\sum_{i=1}^{n} I(-h \leq X_i - x \leq h) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2h}I(|X_i - x| \leq h)$$

$$K_h(\cdot) = \frac{1}{2h}I(|X_i - x| \leq h) \qquad K(\cdot) = \frac{1}{2}I(|X_i - x| \leq 1)$$

$\Rightarrow$ numerical stability: $K(\cdot) = \Phi(\cdot)$.

$$\hat{f}_0(x) = \frac{1}{n_0}\sum_{i=1}^{n_0} K_h(X_{0i} - x)$$

$$\hat{f}_1(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} K_h(X_{1i} - x).$$

What is our h? Hyperparameter selection.

$$\hat{h} = \arg\min_h CV(h)$$