M4.L1 : Classification

TMA4268 , 28.01.2019

response

$Y \in \{1, 2, .., K\}$ often $\{0,1\}$ — iris species

X covariates $\in \mathbb{R}^p$ — septal length, width

Classification:

$X \to f(X) \to$ predict class membership

and/or give probability that X

belongs to class k $(k = 1, .., K)$

Describe class boundaries $\leftarrow$ focus for discriminant analysis.

Training set

(x_i, Y_i) $i = 1, .., n$

to construct classification rule $\hat{Y}_i$

Test set

to evaluate the rule

· misclassification rate

minimize a 0/1 loss

if $(\hat{Y}_i = Y_i) \to$ loss 0

$(\hat{Y}_i \neq Y_i) \to$ 1

Bayes classifier

Bayes theorem: $P(Y=k \mid X=x) = \dfrac{P(Y=k \cap X=x)}{f(x)}$

class k, covariates

pdf for x

$P(Y=k)$

$$= \frac{\pi_k \, f_k(x)}{\sum_{l=1}^{K} \pi_l \cdot f_l(x)}$$

pdf for x when $Y=k$

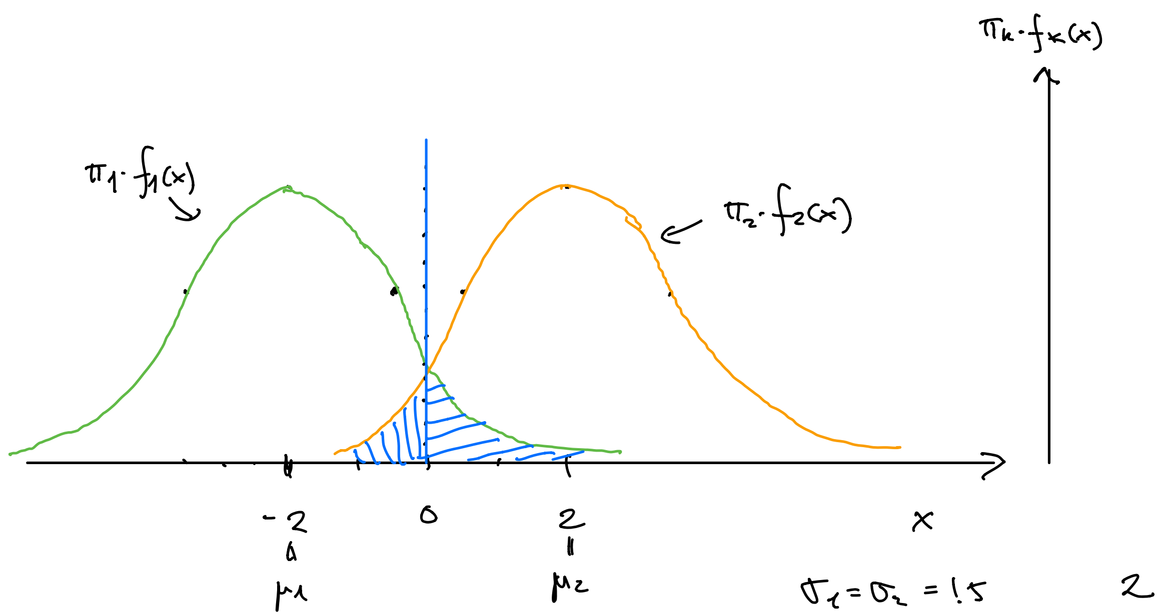
The Bayes classifier assigns a new observation x_0 to the class k where $P(Y=k \mid X=x_0)$ is the largest $(k=1,..,K)$

→ produce Bayes decision boundary

→ Bayes error rate = the best we can do

|| comparable to irreducible error (regression)

Ex: $\pi_1 = \pi_2 = \frac{1}{2}$



Bayes decision boundary

$$P(Y=1 \mid X) = P(Y=2 \mid X)$$

$$\underbrace{\pi_1}_{P(Y=1)} \cdot \frac{f_1(x)}{\cancel{f(x)}} = \frac{\pi_2 \cdot f_2(x)}{\cancel{f(x)}}$$

if $\pi_1 = \pi_2$, class boundary at x such that

$$f_1(x) = f_2(x)$$

$$\cancel{\frac{1}{\sqrt{2\pi}}} \cancel{\frac{1}{\sigma_1}} \cdot \exp\left\{-\frac{1}{2\sigma_1^2}(x-\mu_1)^2\right\} = \cancel{\frac{1}{\sqrt{2\pi}}} \cancel{\frac{1}{\sigma_2}} \exp\left\{-\frac{1}{2\sigma_2^2}(x-\mu_2)^2\right\}$$

$$\downarrow \ln$$

$$-\cancel{\frac{1}{2\sigma^2}}(x-\mu_1)^2 = -\cancel{\frac{1}{2\sigma^2}}(x-\mu_2)^2 \qquad \sigma_1 = \sigma_2 = \sigma$$

$$x^2 - 2\mu_1 x + \mu_1^2 = x^2 - 2\mu_2 x + \mu_2^2$$

$$(\mu_1^2 - \mu_2^2) = 2(\mu_1 - \mu_2)\, x$$

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

$$(\mu_1^2 - \mu_2^2) = (\mu_1 - \mu_2)(\mu_1 + \mu_2)$$

Here: $\mu_1 = -2$, $\mu_2 = 2$ $\Rightarrow$ $x = 0$ boundary.

$$\text{Bayes error} = \frac{1}{2}\left(P(X > 0 \mid Y=1) + P(X < 0 \mid Y=2)\right)$$
$$= \frac{1}{2} \cdot 2 \int_0^\infty f_1(x)\, dx = \dots = 0.09 \qquad 9\%$$

If we get a lower error rate than 9% ... something is wrong!

Classify to class k

with maximal

$P(Y = k \mid X = x)$

diagnostic paradigm

estimate

$P(Y = k \mid X = x)$

directly

$\uparrow$

KNN

logistic regr.

tree

SVM

NN

sampling paradigm

estimate π_k and

$f_k(x)$ then look at

$\pi_k \cdot f_k(x)$

$\uparrow$

LDA / QDA

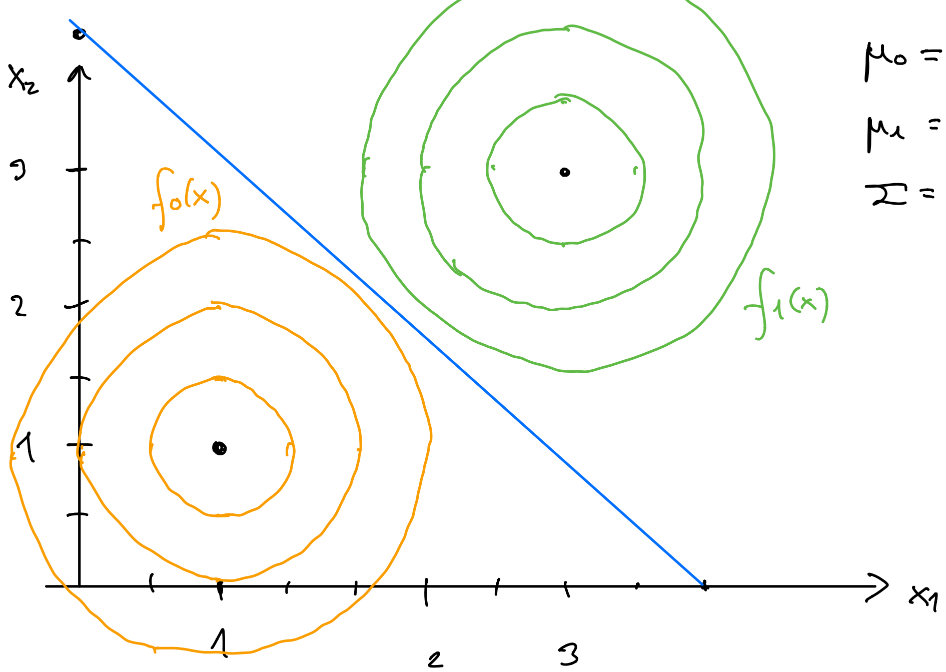
Linear discriminant analysis (LDA)

Assume $f_k(x) = mvN$ $\qquad X$ p-dim

$$= \underbrace{\left[(2\pi)^{-p/2} \det(\Sigma)^{-\frac{1}{2}}\right]}_{const} \cdot \exp\left\{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)\right\}$$

$\uparrow$ assume $\Sigma_k = \Sigma$ for all k

For simplicity $K=2$ classes and $p=2$



Ex:

$\mu_0 = (1,1)^T$

$\mu_1 = (3,3)^T$

$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

$$P(Y=0 \mid X) = P(Y=1 \mid X)$$

$$\frac{\pi_0 \cdot f_0(x)}{\cancel{f(x)}} = \frac{\pi_1 f_1(x)}{\cancel{f(x)}}$$

$$\pi_0 \cdot \cancel{const} \cdot \exp\left\{-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right\} = \pi_1 \cdot \cancel{const} \cdot \exp\left\{-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right\}$$

$\downarrow \ln = \log$

$$\log(\pi_0) - \frac{1}{2}\underbrace{(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)} = \log(\pi_1) - \frac{1}{2}\underbrace{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)}$$

$$\cancel{x^T\Sigma^{-1}}x - 2\mu_0^T\Sigma^{-1}x + \mu_0^T\Sigma^{-1}\mu_0 \qquad \cancel{x^T\Sigma^{-1}} - 2\mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1$$

$$\underbrace{\log\pi_0 - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \mu_0^T\Sigma^{-1}x}_{\delta_0(x)} = \underbrace{\log\pi_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}x}_{\delta_1(x)}$$

discriminant function

For our synthetic dataset setting $\delta_0(x) = \delta_1(x)$ with μ_0, μ_1, Σ as above, and $\pi_0 = \pi_1$ $\rightarrow$ $x_2 = 4 - x_1$ as optimal class boundary

But μ_k and Σ are unknown: use the training set to estimate:

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i$$

$$\underset{p\times p}{\hat{\Sigma}_k} = \frac{1}{n_k-1}\sum_{i:y_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T$$

$$\hat{\Sigma} = \sum_{k=1}^{K}\frac{n_k-1\cdot\hat{\Sigma}_k}{n-K}$$

QDA: $f_k(x) \sim$ mvN but with possibly Σ_k different

To compare methods: look at misclassification rate on test data.

Diagnostic paradigm "$P(Y=j \mid X=x_0)$" directly

KNN-classifier:

number of neighbors to be used

$$\hat{P}(Y=j \mid X=x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$

possible classes

the K closest points to x_0

$$= \frac{\#\text{ in } N_0 \text{ with class } j}{K}$$