References

1. Pattern Classification
Duda, Hart, Stork

2. Probabilistic Theory of Pattern Recognition
Devroye, Gyorfi, Lugosi

3. Pattern recognition and Machine Learning
Chris Bishop

$\mathcal{X} \to$ Instance space

$\mathcal{Y} \to$ label space

Multi-Category Classification

Classifier

$h: \mathcal{X} \to \{1, 2, \ldots, M\}$

$i \in \{1, \ldots, M\}$

Prior $\longrightarrow$ $P_i = P(Y = i)$

Class conditional Distribution

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

[Densities / p.m.f]

$$\eta_i(x) = P(Y = i \mid X = x)$$

$$P(Y = i \mid X = x) = \frac{p_i P(X = x \mid Y = i)}{\sum_k p_k P(X = x \mid Y = k)}$$

Bayes Classifier

x

$$h^*(x) = i \quad \text{if } P(Y = i \mid X = x) > P(Y = j \mid X = x)$$

for all $j \neq i$

Bayes Error rate

$$E_{Y|X=x}\left[\ell(h(x), Y)\right] = \min_{i \in \{1,\dots,M\}} r_i(x)$$

$$r_i(x) = E_{Y|X=x}\, \ell(i, Y)$$

$$\min_{h \in \mathcal{H}} R(h) \qquad R(h) = P(h(X) \neq Y)$$

$$R(h) = E\left(\ell(h(X), Y)\right)$$

$$\tilde{h} = \underset{h}{\operatorname{argmin}}\, E_{Y|X}\, \ell(h(X), Y)$$

For every x choose $\tilde{h}(x) = i$

if $r_i(x) < r_j(x)$ for all $j \neq i$

choose $\ell(i,i) = 0 \quad i \in \{1, \dots, M\}$

$\ell(i,j) = 1 \quad i \neq j$

$\eta_i(x) = P(Y = i \mid X = x)$

$$\therefore r_i(x) = \sum_{k=1}^{M} \ell(i,k)\,\eta_k(x)$$

$$= \ell(i,i)\,\eta_i(x) + \sum_{k \neq i}^{M} \ell(i,k)\,\eta_k(x)$$

$$\therefore r_i(x) = \sum_{k \neq i}^{M} \eta_k(x) = 1 - \eta_i(x)$$

$$\tilde{h}(x) = i \quad \text{if} \quad r_i(x) < r_j(x)$$

for all $j \neq i$

$$1 - \eta_i(x) < 1 - \eta_j(x)$$

$$\eta_j(x) < \eta_i(x)$$

for all $j \neq i$

$$\tilde{h}(x) = h^*(x)$$

For M category problem

define Discriminant functions

$$g_i : \mathcal{X} \to \mathbb{R} \qquad i \in \{1, \dots, M\}$$

If $g_i(x) > g_j(x)$ for all $j \neq i$

Then $h(x) = i$

Let $f : [0,1] \to \mathbb{R}$ be any monotonic increasing function

$g_i(x) = f(\eta_i(x))$ can be a discriminant function

$x \in \mathbb{R}^d$

$$P(X = x \mid Y = i) = N(x \mid \mu_i, C_i)$$

$$\eta_i(x) = \frac{P_i N(x \mid \mu_i, C_i)}{P(x)}$$

$$C_i = C \qquad i \in \{1, \dots, M\}$$

$$\eta_i(x) = \frac{P_i N(x \mid \mu_i, C)}{P(x)}$$

$$f(z) = \log z \qquad 0 \leq z \leq 1$$

$$\log \eta_i(x) = \log P_i + \log N(x \mid \mu_i, C) - \log P(x)$$

$$= \log P_i + \log \frac{1}{(\sqrt{2\pi})^{n} |C|^{1/2}} - \frac{1}{2}(x - \mu_i)^T C^{-1} (x - \mu_i) - \log P(x)$$

$$g_i(x) = \log P_i - \frac{1}{2}(x - \mu_i)^T C^{-1} (x - \mu_i)$$

If

$$g_i(x) > g_j(x) \quad j \neq i$$

$$h^*(x) = i \qquad j \in \{1, \ldots, M\}$$

If $M = 2$

$$h^*(x) = 1 \qquad g_1(x) - g_2(x) > 0$$

$$= 2 \qquad g_1(x) - g_2(x) < 0$$

$$h^*(x) = \text{sign}(g(x)) \qquad g(x) = g_1(x) - g_2(x)$$

$$\log p_1 - \frac{1}{2}(x - \mu_1)^T C^{-1}(x - \mu_1)$$
$$- \log p_2 + \frac{1}{2}(x - \mu_2)^T C^{-1}(x - \mu_2)$$

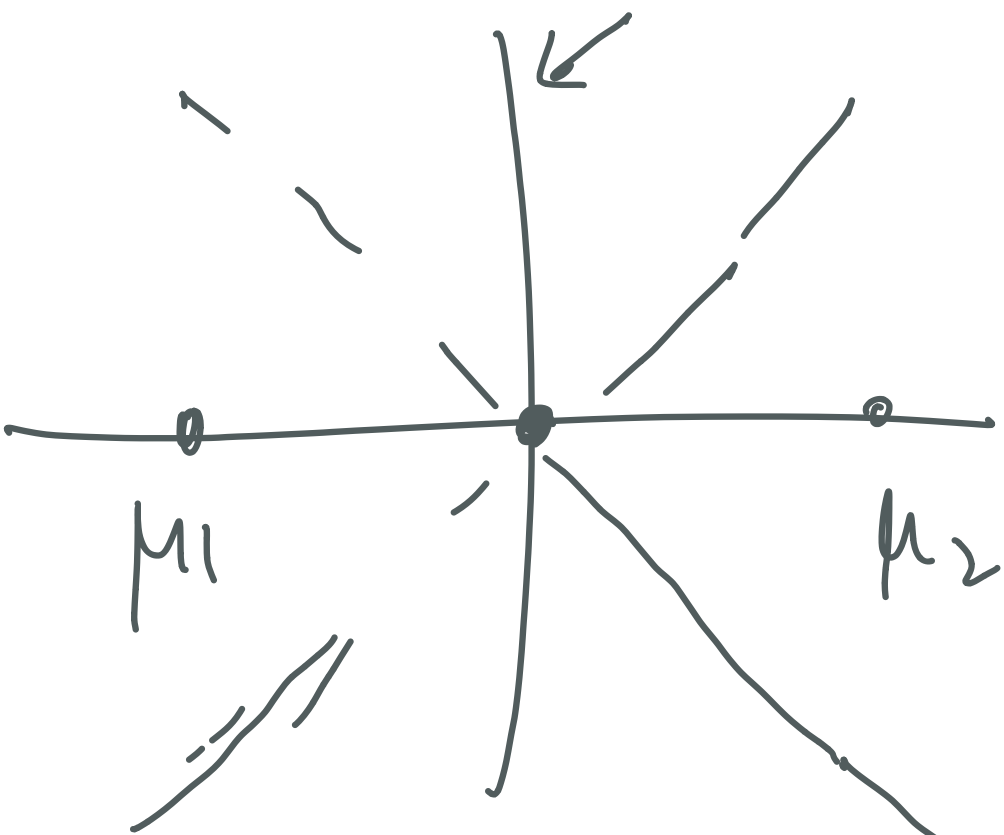
$$= w^T x + b.$$

$P_1 = P_2 = 0.5$

$$h(x) = \text{sign}\left(w^T\left(x - \frac{\mu_1+\mu_2}{2}\right)\right)$$

$$w^T\left(\frac{\mu_1+\mu_2}{2}\right) = -b$$

$$w^T\left(x - \frac{\mu_1+\mu_2}{2}\right) = 0$$



$C = \sigma^2 I$ $\quad$ $w = \frac{1}{\sigma^2}(\mu_1 - \mu_2)$

Let $X \in \mathbb{R}^d$ with p.d.f. $f_X(x)$

$$E(X) = \mu \qquad C_{ij} = E(X_i - \mu_i)(X_j - \mu_j)$$

For any $u \in \mathbb{R}^d$

$$z = u^T X$$

Find $E(z)$, $\mathrm{var}(z)$

$$E(z) = E(u^T X) = u^T \mu$$

$$\mathrm{var}(z) = E(u^T X - E(z))^2 = u^T C u$$

Suppose $\|w\| = 1$ $\qquad \|w\| = \sqrt{w^T w}$

$$C = \{\alpha w \mid \alpha \in \mathbb{R}\} \qquad = \sqrt{\sum_i w_i^2}$$

Projection of any $v \in \mathbb{R}^d$ on C is the point in C which is closest to v

Digression

$$\min_{\alpha \in \mathbb{R}} \|v - \alpha w\|^2 \geq 0 \quad v, w \in \mathbb{R}^d$$

$$= \|v\|^2 - 2\alpha v^T w + \alpha^2 \|w\|^2$$

$$\alpha^* = \frac{v^T w}{\|w\|^2}$$

$$= \|v\|^2 - 2\alpha \alpha^* \|w\|^2 + \alpha^2 \|w\|^2$$

$$= \|v\|^2 - (\alpha^*)^2 \|w\|^2 + \|w\|^2 (\alpha - \alpha^*)^2$$

$$\geq \|v\|^2 - (\alpha^*)^2 \|w\|^2$$

Therefore α^* minimizes $\|v - \alpha w\|^2$

$$\|v\|^2 - (\alpha^*)^2 \|w\|^2 \geq 0$$

$$\|w\|^2 \|v\|^2 \geq (v^T w)^2$$

$\|v - \alpha w\| = 0$ iff $v = tw$

Then for any $v, w \in \mathbb{R}^d$

$$\|w\|^2 \|v\|^2 \geq (v^T w)^2$$

Equality holds iff $w = tv$

0 0 0