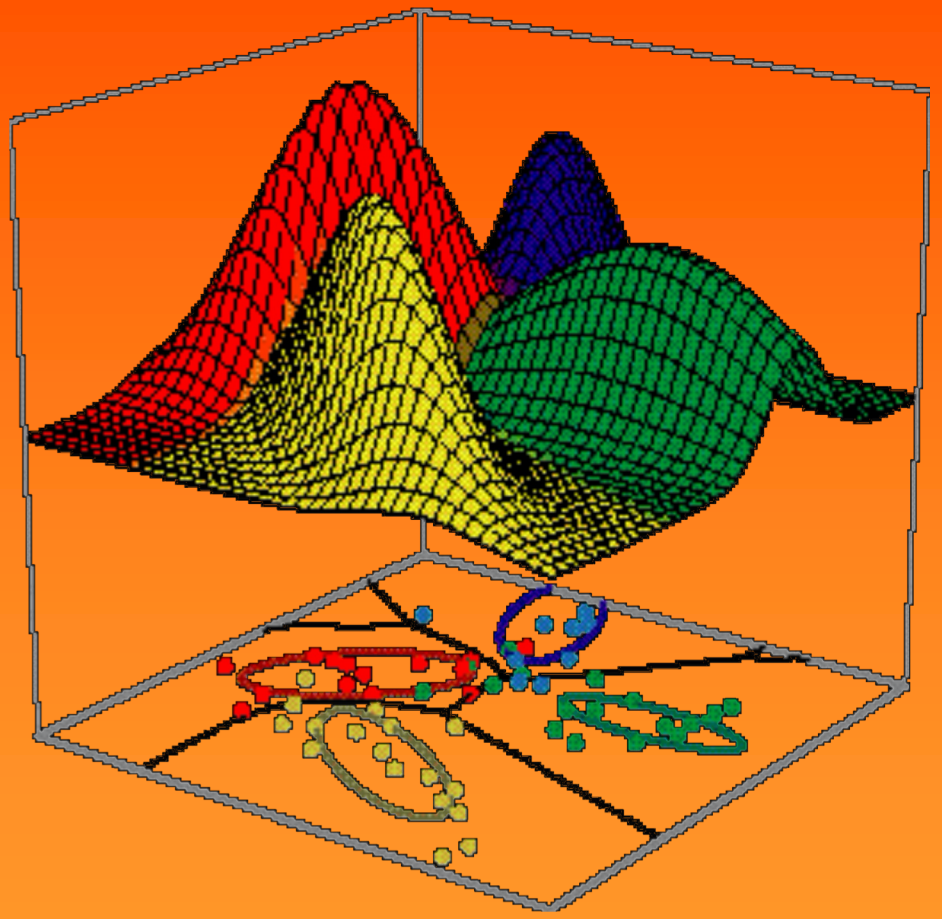


Pattern Classification

All materials in these slides were taken from Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000
with the permission of the authors and the publisher

Chapter 2 (part 3)
Bayesian Decision Theory (Sections 2-6,2-9)

- Discriminant Functions for the Normal Density

- Bayes Decision Theory – Discrete Features

Quiz on coming thursday

Syllabuss - lectures till coming monday

Discriminant Functions for the Normal Density

- We saw that the minimum error-rate classification can be achieved by the discriminant function

  $g_i(x) = \ln P(x \mid \omega_i) + \ln P(\omega_i)$

- Case of multivariate normal

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t \sum_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

*) $\Sigma_i = \sigma^2 I$

$g_i(x) = w^T x + b, \quad w = \frac{\mu_i}{\sigma^2} \quad b = -\frac{\mu_i^T \mu_i}{2\sigma^2}$

*) $g_1(x) = g_2(x) \rightarrow$ for two category

*) $\Sigma_i = \Sigma \rightarrow$ not in the form identity

$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \ln P(w_i)$

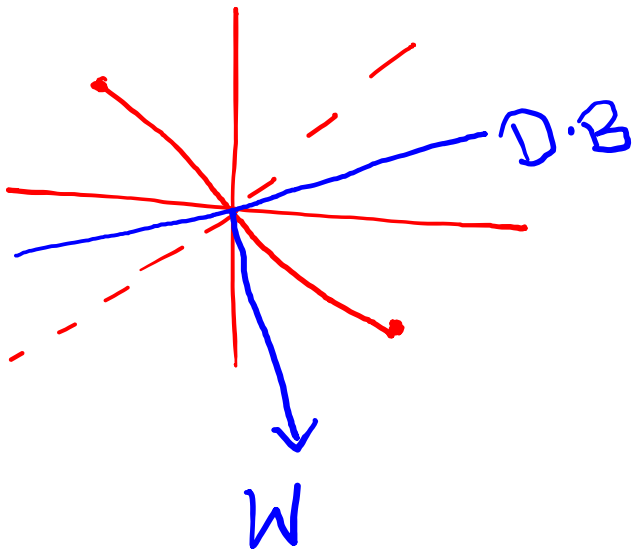
$= -\frac{1}{2}\left\{ x^T \Sigma_i^{-1} x - 2\mu_i^T \Sigma_i^{-1} x + \mu_i^T \Sigma_i^{-1} \mu_i \right) + \ln P(w_i)$

$= \mu_i^T \Sigma^{-1} x - \frac{1}{2}\mu_i^T \Sigma^{-1} \mu_i + \ln P(w_i)$

$= w^T x + b \quad w = \Sigma^{-1}\mu_i \quad b = -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \ln P(w_i)$

D.B $g_i(x) = g_j(x) \equiv w^T x + b$

$w = \Sigma^{-1}(\mu_i - \mu_j)$



*) Σ_i → arbitrary and different for all classes.

- Case $\Sigma_i = \sigma^2.I$ (I stands for the identity matrix)

$$g_i(x) = w_i^t x + w_{i0} \text{ (linear discriminant function)}$$

where :

$$w_i = \frac{\mu_i}{\sigma^2}; \; w_{i0} = -\frac{1}{2\sigma^2}\mu_i^t \mu_i + \ln P(\omega_i)$$

(ω_{i0} is called the threshold for the ith category!)

- A classifier that uses linear discriminant functions is called "a linear machine"

- The decision surfaces for a linear machine are pieces of hyperplanes defined by:

$$g_i(x) = g_j(x)$$

- Case Σ_i = arbitrary
  - The covariance matrices are different for each category

$$g_i(x) = x^t W_i x + w_i^t x + w_{i0}$$

where :

$$W_i = -\frac{1}{2}\Sigma_i^{-1}$$

$$w_i = \Sigma_i^{-1}\mu_i$$

$$w_{i0} = -\frac{1}{2}\mu_i^t\Sigma_i^{-1}\mu_i - \frac{1}{2}ln|\Sigma_i| + ln\,P(\omega_i)$$

(Hyperquadrics which are: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, hyperhyperboloids)

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0},$$

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1},$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i).$$

Assume two different classes – ω_1 & ω_2

Both having multi-variate Gaussian pdf.

$$p(x|\omega_1) \sim N(\mu_1, \Sigma_1) \qquad x \in R^d$$

$$p(x|\omega_2) \sim N(\mu_2, \Sigma_2)$$

$$\mu_1 \neq \mu_2, \quad \Sigma_1 \neq \Sigma_2$$

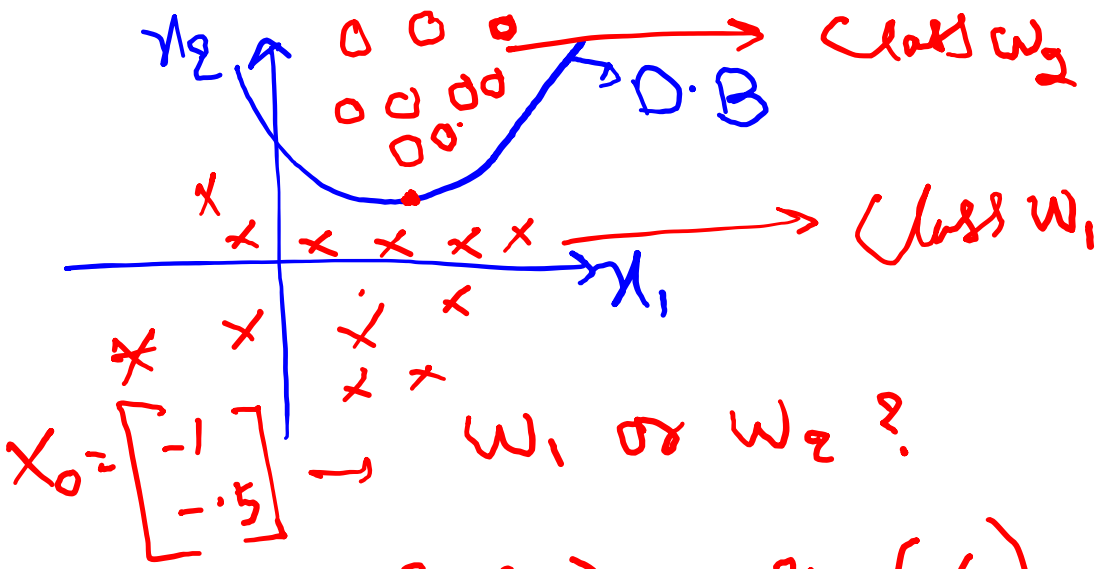
$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

The inverse matrices are then,

$$\boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

We assume equal prior probabilities, $P(\omega_1) = P(\omega_2) = 0.5$, and substitute these into the general form for a discriminant, Eqs. 64 – 67, setting $g_1(\mathbf{x}) = g_2(\mathbf{x})$ to obtain the decision boundary:

$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2.$$ → D.B.



6th. slot timer.

7th → crossbod.

$g_1(x_0)$ $g_2(x_0)$

$d = 10^6, 10^4$ $g_1(x) = g_2(x)$

w_1 → Take $ML

w_2 → Dont Take $ML →

X

Grades, Instructor, TA, has changed. $d=4$ → seek opinion from bo.

x_1 x_2 x_3 x_4 5th.

Demo

Use LDA when cov matrices are not same?

$$\Sigma_i = \Sigma$$

$$\Sigma = \frac{n_1\Sigma_1 + n_2\Sigma_2 + \cdots + n_c\Sigma_c}{n_1 + n_2 + \cdots + n_c}$$

$n_1 \rightarrow$ no. of samples in class 1

$\vdots$

$n_c \rightarrow \vdots \quad \cdots \quad \vdots$ class c

Bayes Decision Theory – Discrete Features

- Components of x are binary or integer valued, x can take only one of m discrete values

$$v_1, v_2, \ldots, v_m$$

- Case of independent binary features in 2 category problem

Let $x = [x_1, x_2, \ldots, x_d]^t$ where each x_i is either 0 or 1, with probabilities:

$$p_i = P(x_i = 1 \mid \omega_1)$$

$$q_i = P(x_i = 1 \mid \omega_2)$$

Category / classes.

$$p_i = P_\theta(x_i = 1 \mid \omega_1)$$

$$1 - p_i = P_\theta(x_i = 0 \mid \omega_1)$$

~~$P(x_i) =$~~ $= p_i^{x_i} * (1-p_i)^{1-x_i}$ ✓

$$P(x_i \mid \omega_1) = p_i \delta(x_i) + (1-p_i)\delta(1-x_i) \quad \times$$

$$= p_i \delta(1-x_i) + (1-p_i)\delta(x_i) \quad ✓$$

$\delta \rightarrow$ discretes. $\delta(0) = 1$

$\delta(\neq 0) = 0$

$$P(x_i|\omega_1) = p_i^{x_i}(1-p_i)^{1-x_i}$$

$$P(x|\omega_1) = P(x=[x_1, x_2 \cdots x_i, \cdots x_d]^T | \omega_1)$$

$$x=[x_1, x_2 \cdots x_d]^T$$

*) Assumption → Conditional independence.

$$P(x_1, x_2|\omega) = P(x_1|\omega)P(x_2|\omega)$$

$$P(x|\omega_1) = \prod_{i=1}^{d} P(x_i|\omega_1)$$

$$= \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$$

If the classifier assumes this types of Conditional independence. → Naive Bayes class.

$$P(X|\omega_2) = \prod_{i=1}^{d} q_i^{x_i} (1-q_i)^{1-x_i}$$

we have ~~likelihood~~ likelihood for both classes.

$$g_1(X) = ? = P(\omega_2) P(X/\omega_2) + P(\omega_1) P(X/\omega_1)$$

↓ discriminant.

$$\ln P(x|\omega_1) + \ln P(\omega_1)$$

$$g_i(x) = \ln P(\omega_i/x) +$$

$$g_1(x) = \ln \prod_{i=1}^{N} p_i^{x_i} (1-p_i)^{1-x_i} + \ln p(\omega_1)$$

$$= \sum_{i=1}^{N} x_i \ln p_i + (1-x_i) \ln (1-p_i) + \ln p(\omega_1)$$

$g_2(x)$

$g_1(x) = g_2(x) \rightarrow$ D.B

Usually for two category class.

We define. $g(x) = g_1(x) - g_2(x)$

$g(x)$: dichotimizer

ω

- The discriminant function in this case is:

$$g(x) = \sum_{i=1}^{d} w_i x_i + w_0$$

where :

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1,\ldots,d$$

and :

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

decide ω_1 *if* $g(x) > 0$ *and* ω_2 *if* $g(x) \leq 0$

Ex.

- Suppose two categories consist of independent binary features in three dimensions with known feature probabilities. Let us construct the Bayesian decision boundary if P (ω1) = P (ω2) = 0.5 and the individual components obey:
- $p_i = 0.8$ and $q_i = 0.5$ for i = 1,2,3.
- wi = ln .8(1 − .5)/.5(1 − .8) = 1.3863

$g(x) = g_1(x) - g_2(x)$

$= 0 \rightarrow D.B.$

$p_i = P_r(x_i = 1 | \omega_1)$

$q_i = P_r(x_i = 1 | \omega_2)$

$\omega_1 = \ln \frac{0.8 \times 0.5}{0.5 \times 0.2} = \ln 4 = \omega_2 = \omega_3$

$$w_0 = \sum_{i=1}^{3} \ln \frac{.2}{.5}.$$

$$g(x) = 0$$

$$\sum_{i=1}^{3} w_i x_i + w_0 = 0$$

$$w_1 (x_1 + x_2 + x_3) + 3 \ln \frac{2}{5} = 0$$

$$ax + by + cz + d = 0$$

Find the multiclass discriminant corresponding to min probability of error for independent binary valued features.

$\omega_1, \omega_2, \ldots, \omega_c$

$g_i(x) \rightarrow \forall i = 1, \ldots, c$

$$P(x|\omega_1) = \prod_{i=1}^{d} p_i^{x_i} (1-p_i)^{1-x_i}$$

$$p_{ij} = P_r(x_i = 1 | \omega_j)$$

$$p_{i1} = P_r(x_i = 1 | \omega_1) \qquad 1 - p_{i1} = 1 - P_r(x_i = 1 | \omega_1)$$

$$p(x|\omega_j) = \prod_{i=1}^{d} p_{ij}^{x_i} (1-p_{ij})^{1-x_i}$$

$$p(x_i|\omega_j) = p_{ij}^{x_i} (1-p_{ij})^{1-x_i}$$

$$g_j(x) = \ln p(x|\omega_j) + \ln p(\omega_j)$$

$$\forall j = 1, 2, \ldots,$$

$$\max_j g_j(x) \rightarrow \text{map } x \text{ to } j$$