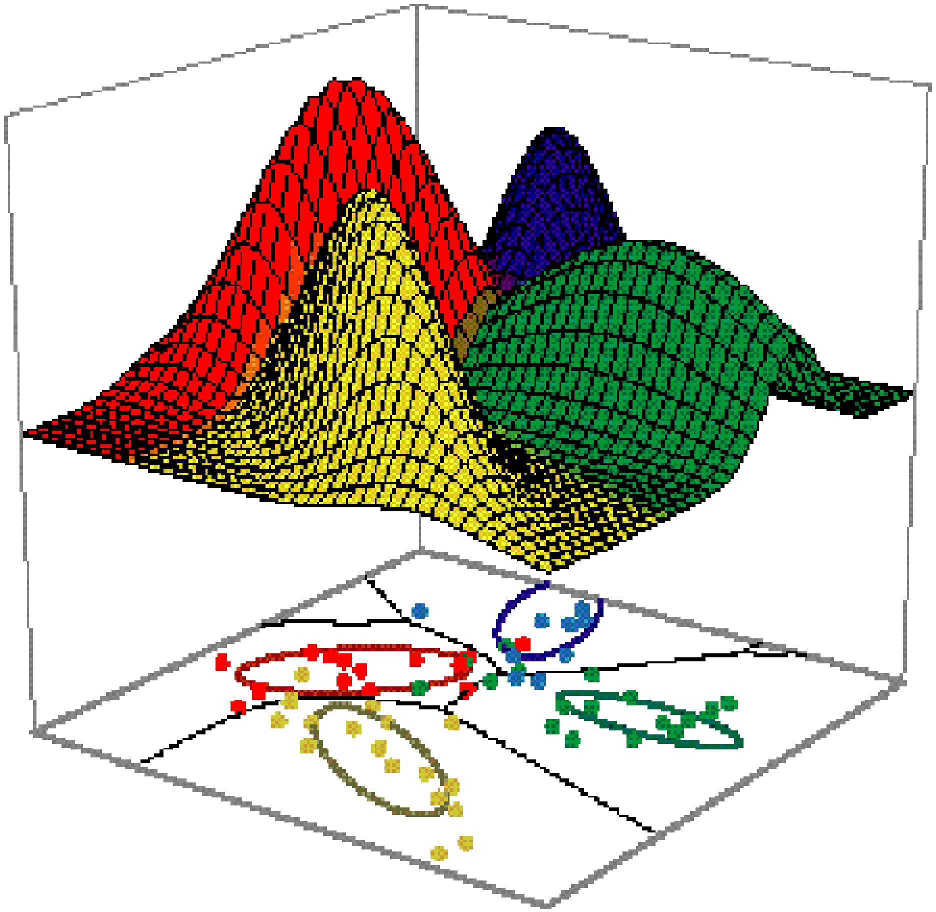


Bayes Optimal Classifier

Classification

Goal: Construct a **predictor** $f : X \rightarrow Y$ to minimize a risk (performance measure) $R(f)$

Sports
Science
News

Features, X **Labels, Y**

$$R(f) = P(f(X) \neq Y)$$ **Probability of Error**

Bayes optimal rule

Ideal goal: Construct **prediction rule** $f^*: \mathcal{X} \rightarrow \mathcal{Y}$

$$f^* = \arg\min_f \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$$

Bayes optimal rule

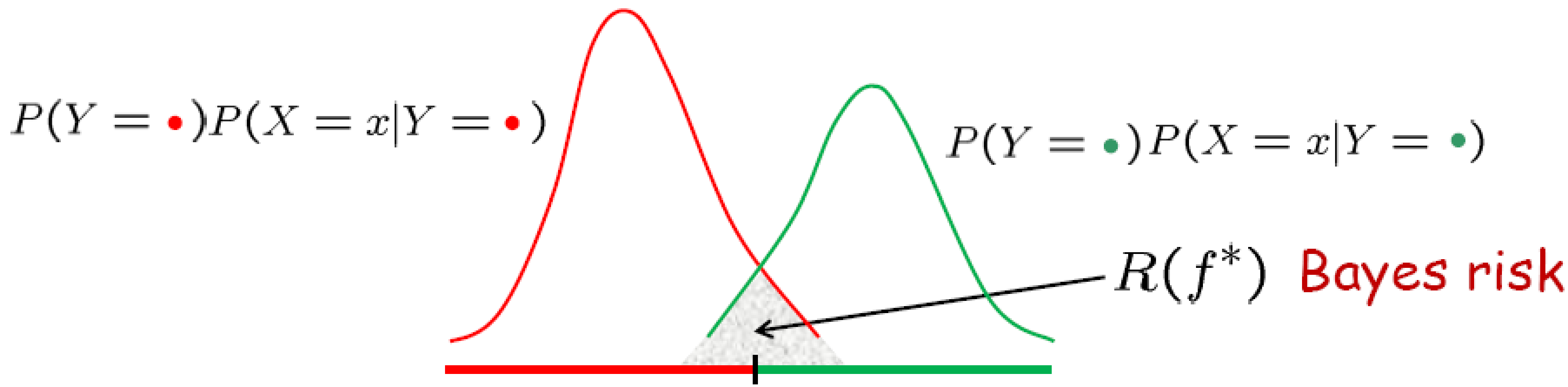
Best possible performance:

Bayes Risk $\quad R(f^*) \leq R(f)$ for all f

BUT… Optimal rule is not computable - depends on unknown P_{XY} !

Optimal Classification

Optimal predictor: $f^* = \arg\min_f P(f(X) \neq Y)$
(Bayes classifier)



$$f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$$

- Even the optimal classifier makes mistakes R(f*) > 0
- Optimal classifier depends on **unknown** distribution P_{XY}

Optimal Classifier

Bayes Rule: $$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$P(Y = y|X = x) \;=\; \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

Optimal classifier:

$$\begin{aligned} f^*(x) &= \arg\max_{Y=y} P(Y = y|X = x) \\ &= \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional density}}\underbrace{P(Y = y)}_{\text{Class prior}} \end{aligned}$$

Equivalent Rules

- *If* $P(\omega_1 / x) > P(\omega_2 / x)$, $\omega = \omega_1$
- *If* $P(x / \omega_1)\, P(\omega_1) > P(x / \omega_2)\, P(\omega_2)$, $\omega = \omega_1$
- If $l(x) = \dfrac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \dfrac{P(\omega_2)}{P(\omega_1)}$, $\omega = \omega_1$
- If

$$h(x) = -\ln[l(x)]$$

$$= -\ln[P(x \mid \omega_1)] + \ln[P(x \mid \omega_2)] < \ln[\frac{P(\omega_1)}{P(\omega_2)}]$$

Example

$P(\omega_1)=0.9$, $P(\omega_2)=0.1$

$P(x / \omega_1)=0.2$, $P(x / \omega_2)=0.4$,

x ? ω_1 or ω_2

For C classes

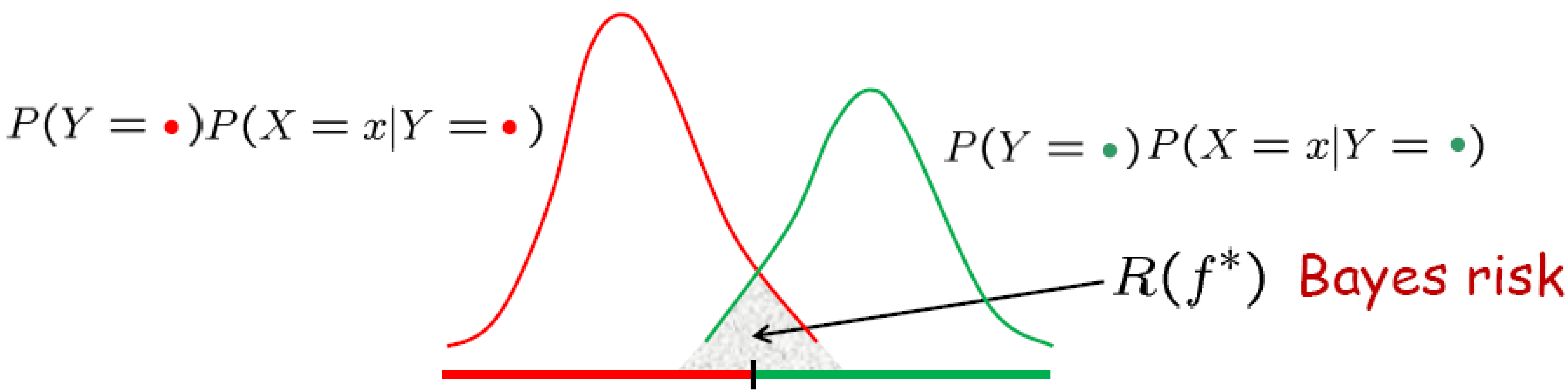
- If $P(\omega_i \mid x) = \max_{j=1,\cdots,c} P(\omega_j \mid x)$, $\omega = \omega_i$
- If $P(x \mid \omega_i)P(\omega_i) = \max_{j=1,\cdots,c} P(x \mid \omega_j)P(\omega_j)$, $\omega = \omega_i$

$$P(c) = \sum_{j=1}^{c} \int_{R_j} P(x \mid \omega_j)P(\omega_j)dx$$

$$P(e)=1-P(c)$$

Optimal Classification

Optimal predictor: $f^* = \arg\min_f P(f(X) \neq Y)$
(Bayes classifier)



$$f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$$

- Even the optimal classifier makes mistakes R(f*) > 0
- Optimal classifier depends on **unknown** distribution P_{XY}

Mini Risk-based Bayes

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j \mid x)$$

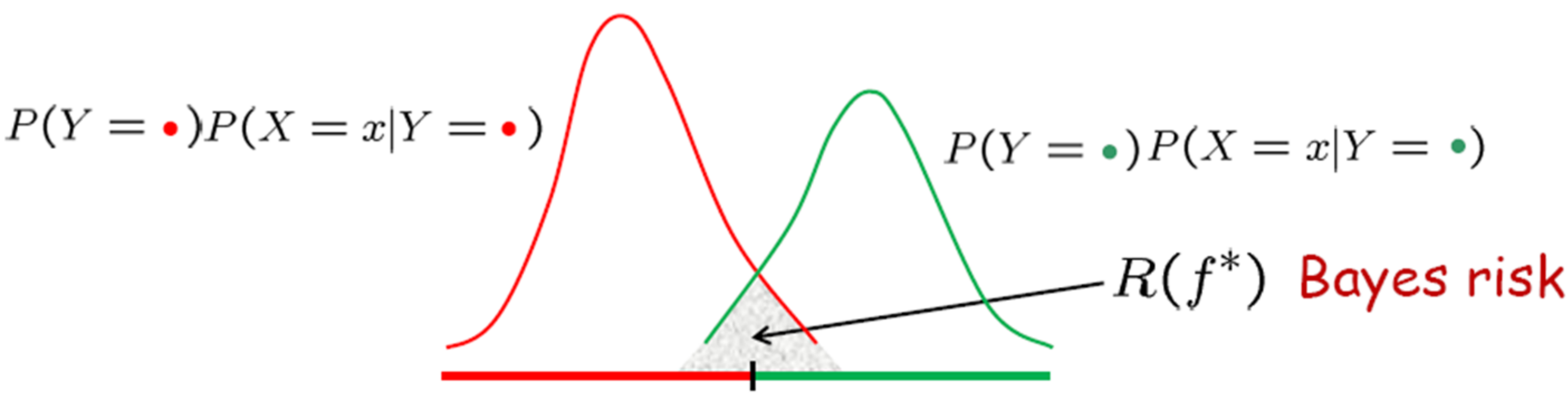
$$\begin{array}{cc} & \begin{array}{cc} \omega_1 & \omega_2 \end{array} \\ \begin{array}{c} \alpha_1 \\ \alpha_2 \end{array} & \begin{bmatrix} \lambda(\alpha_1,\omega_1) & \lambda(\alpha_1,\omega_2) \\ \lambda(\alpha_2,\omega_1) & \lambda(\alpha_2,\omega_2) \end{bmatrix} \end{array}$$

- If $R(\alpha_1 / x) < R(\alpha_2 / x)$, $\alpha = \alpha_1$
- If $(\lambda_{21} - \lambda_{11})P(\omega_1 / x)$

  $> (\lambda_{12} - \lambda_{22})P(\omega_2 / x)$, $\alpha = \alpha_1$

Mini Risk-based Bayes

- If $(\lambda_{21} - \lambda_{11}) P(x \mid \omega_1)\, P(\omega_1) > (\lambda_{12} - \lambda_{22}) P(x \mid \omega_2)\, P(\omega_2)$, $\alpha = \alpha_1$
- If $\dfrac{P(x \mid \omega_1)}{P(x \mid \omega_2)} > \dfrac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \dfrac{P(\omega_2)}{P(\omega_1)}$, $\alpha = \alpha_1$



Example

$$P(\omega_1) = 0.9,\quad P(\omega_2) = 0.1$$

$$P(x / \omega_1) = 0.2,\quad P(x / \omega_1) = 0.4$$

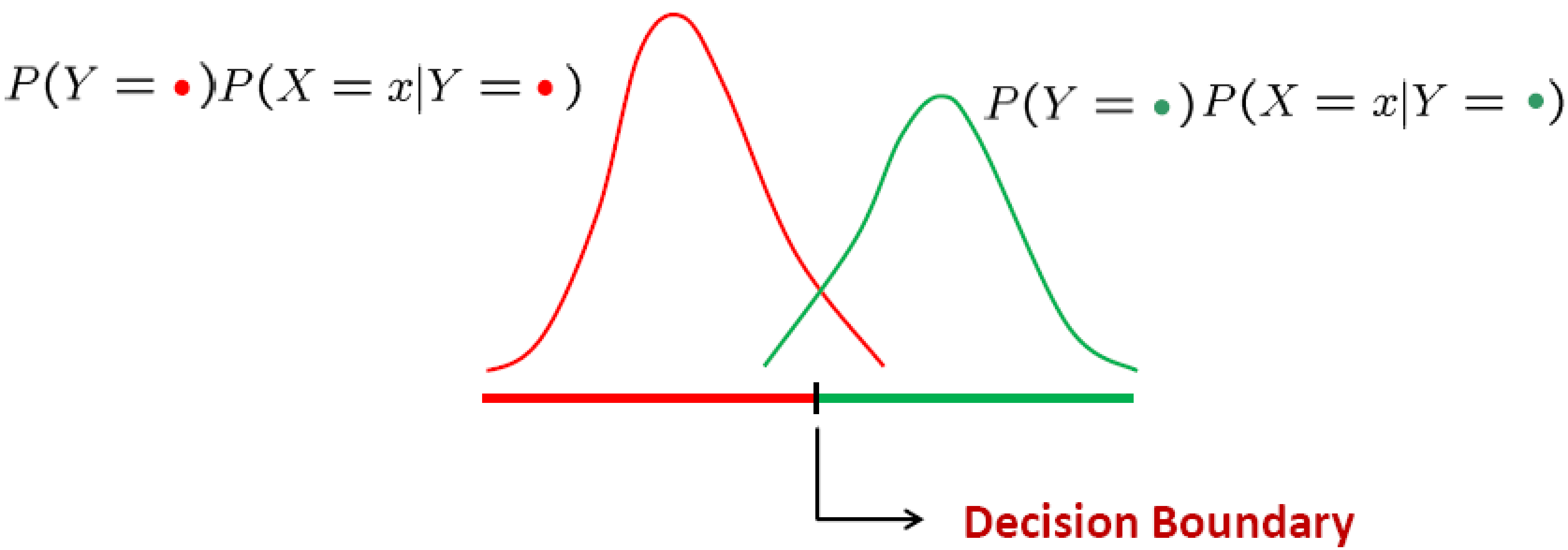
$$\lambda = \begin{bmatrix} 0 & 6 \\ 1 & 0 \end{bmatrix}$$

x ? ω_1 or ω_2

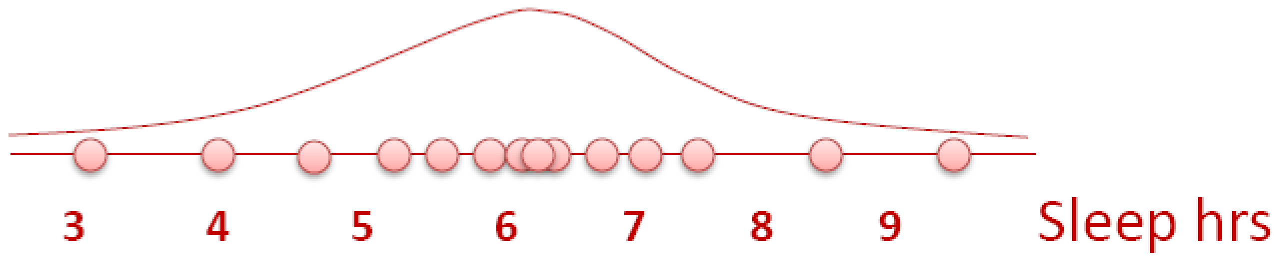
Example: 1-d Decision Boundaries

- Gaussian class conditional densities (1-dimension/feature)

$$P(X = x|Y = y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right)$$



Gaussian Distribution

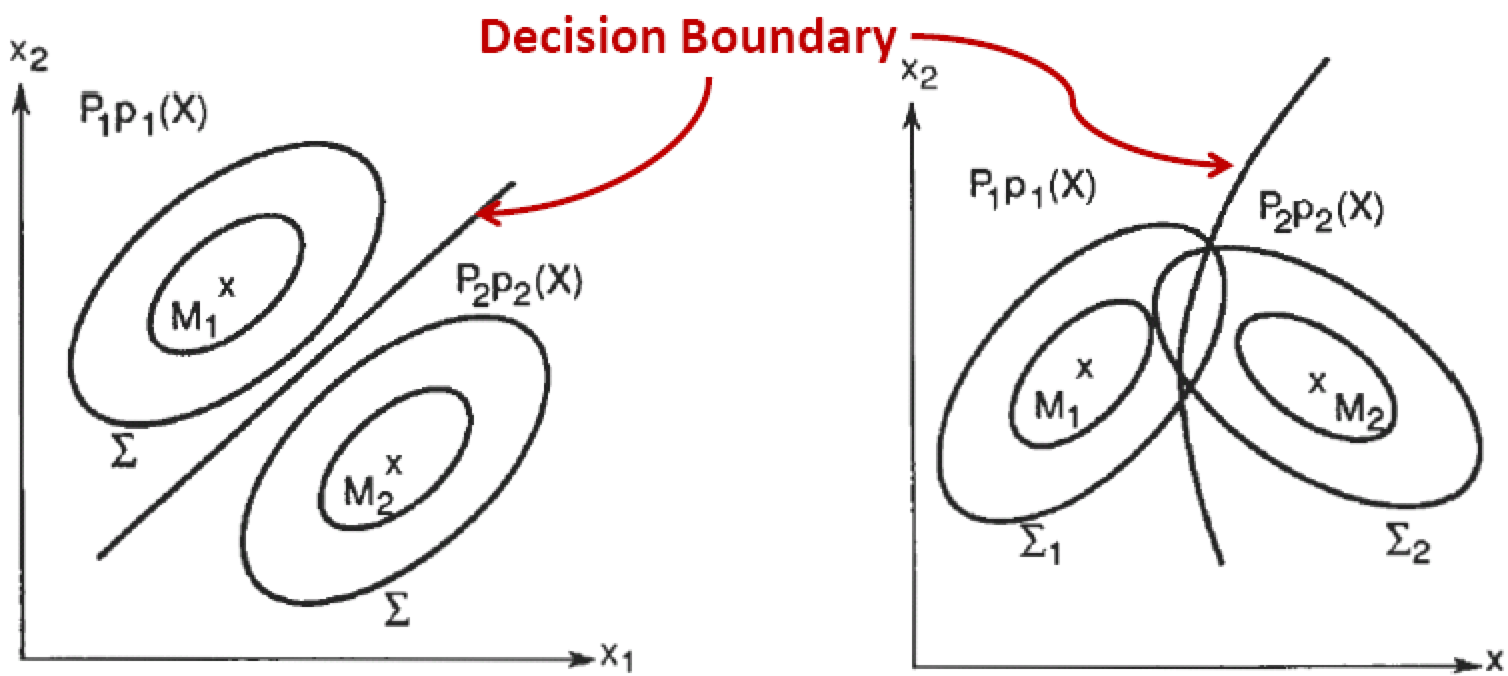
Data, D = 

- Parameters: μ – mean, σ^2 - variance
- Sleep hrs are **i.i.d.**:
  - **Independent** events
  - **Identically distributed** according to Gaussian distribution

Example: 2-d Decision Boundaries

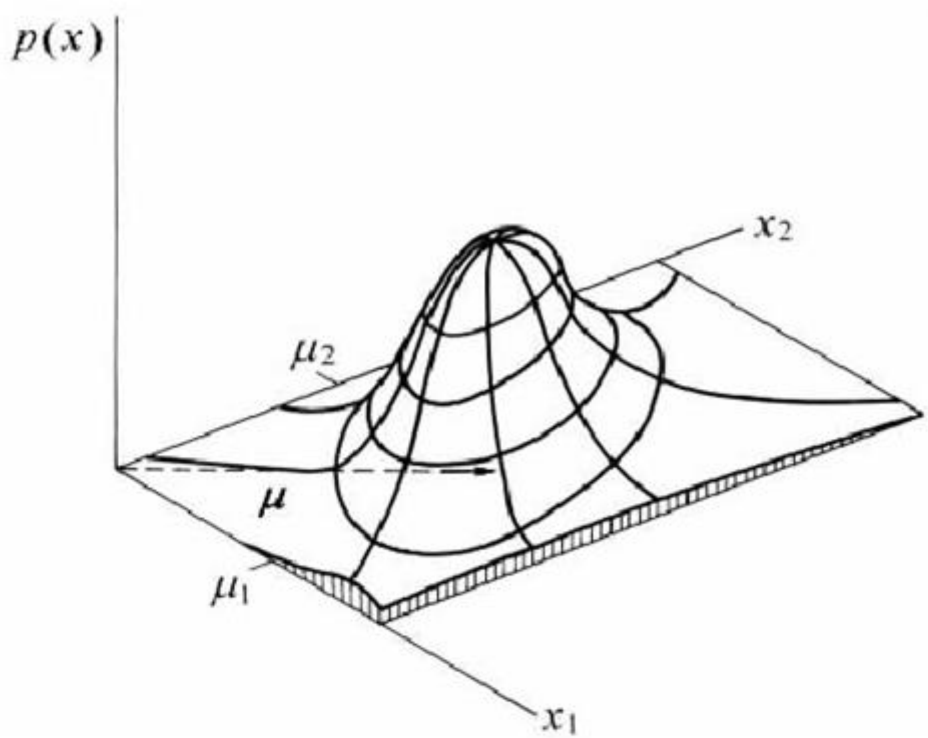
- Gaussian class conditional densities (2-dimensions/features)

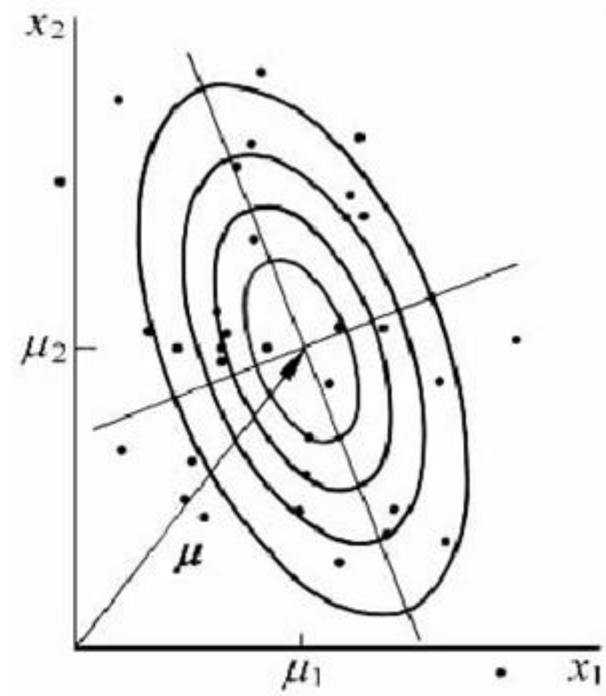
$$P(X = x|Y = y) = \frac{1}{\sqrt{2\pi|\Sigma_y|}} \exp\left(-\frac{(x-\mu_y)'\Sigma_y^{-1}(x-\mu_y)}{2} \right)$$



Properties of Multivariate Gaussian (I)

- $P(x) \sim N(\mu, \Sigma)$





Properties of Multivariate Gaussian (II)

- hyper-elliptical surface of constant probability density for a Gaussian, i.e. $(x-\mu)^t \Sigma^{-1}(x-\mu)=$constant
- Noncorrelation=independence
- Marginal distribution is Gaussian
- Conditional distribution is also Gaussian
- Linear transformation is still Gaussian
- Linear combination is still Gaussian

Discriminant function and decision boundary

- Discriminant function

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

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

- Decision boundary *$g_i(x) = g_j(x)$*

i.e.

$$-\frac{1}{2}[(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) - (x-\mu_j)^t \Sigma_j^{-1}(x-\mu_j)] - \frac{1}{2}\ln\frac{\left|\Sigma_i\right|}{\left|\Sigma_j\right|} + \ln\frac{P(\omega_i)}{P(\omega_j)} = 0$$

Case 1: $\Sigma_i = \sigma^2 I$

$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^t(x-\mu_i) + \ln P(\omega_i)$$

$$= -\frac{1}{2\sigma^2}(x^t x - 2\mu_i^t x + \mu_i^t \mu_i) + \ln P(\omega_i)$$

Linear discriminant function:

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

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

It is a function that is a linear combination of the components of x where w is the weight vector and w_0 the bias

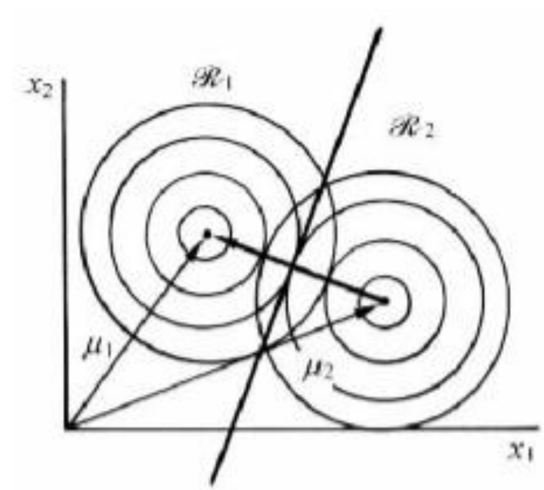
Decision boundary:

$$w^t(x - x_0) = 0$$

where $w = \mu_i - \mu_j;$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\left\|\mu_i - \mu_j\right\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

- $P(\omega_i) = P(\omega_j)$
- $P(\omega_i) \neq P(\omega_j)$

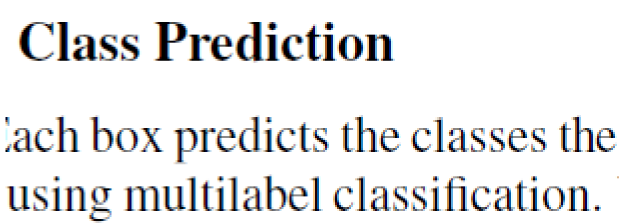


Minimum distance classifier

Discriminant function: $g_i(x) = -\|x - \mu_i\|^2$

$$g_i(x) = \max_{j=1,\cdots,c} g_j(x) \Rightarrow \omega = \omega_i$$

Each mean vector is thought of as being an ideal prototype or template for patterns in its class (template-matching procedure)


Class Prediction

ach box predicts the classes the using multilabel classification.

Case 2: $\Sigma_i = \Sigma$

Linear discriminant function:

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

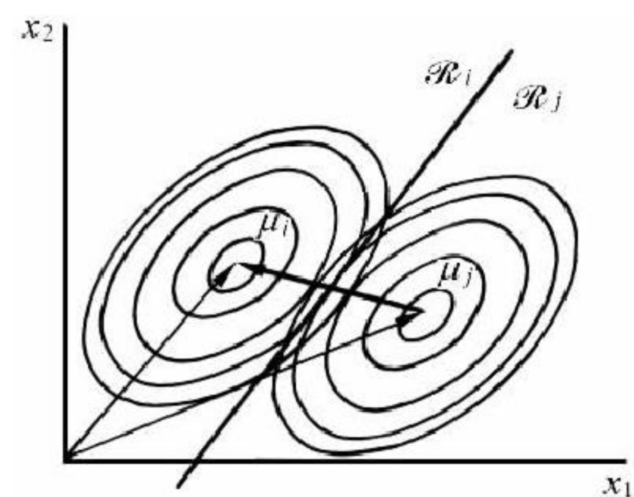
$$\text{where } w_i = \Sigma^{-1}\mu_i; \quad w_{i0} = -\frac{1}{2}\mu_i^t \Sigma^{-1}\mu_i + \ln P(\omega_i)$$

Decision boundary:

$$w^t(x - x_0) = 0$$

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

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\mu_i-\mu_j)^t \Sigma^{-1}(\mu_i-\mu_j)}(\mu_i - \mu_j)$$



Case 3: $\Sigma_i \neq \Sigma_j$

Discriminant function:

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

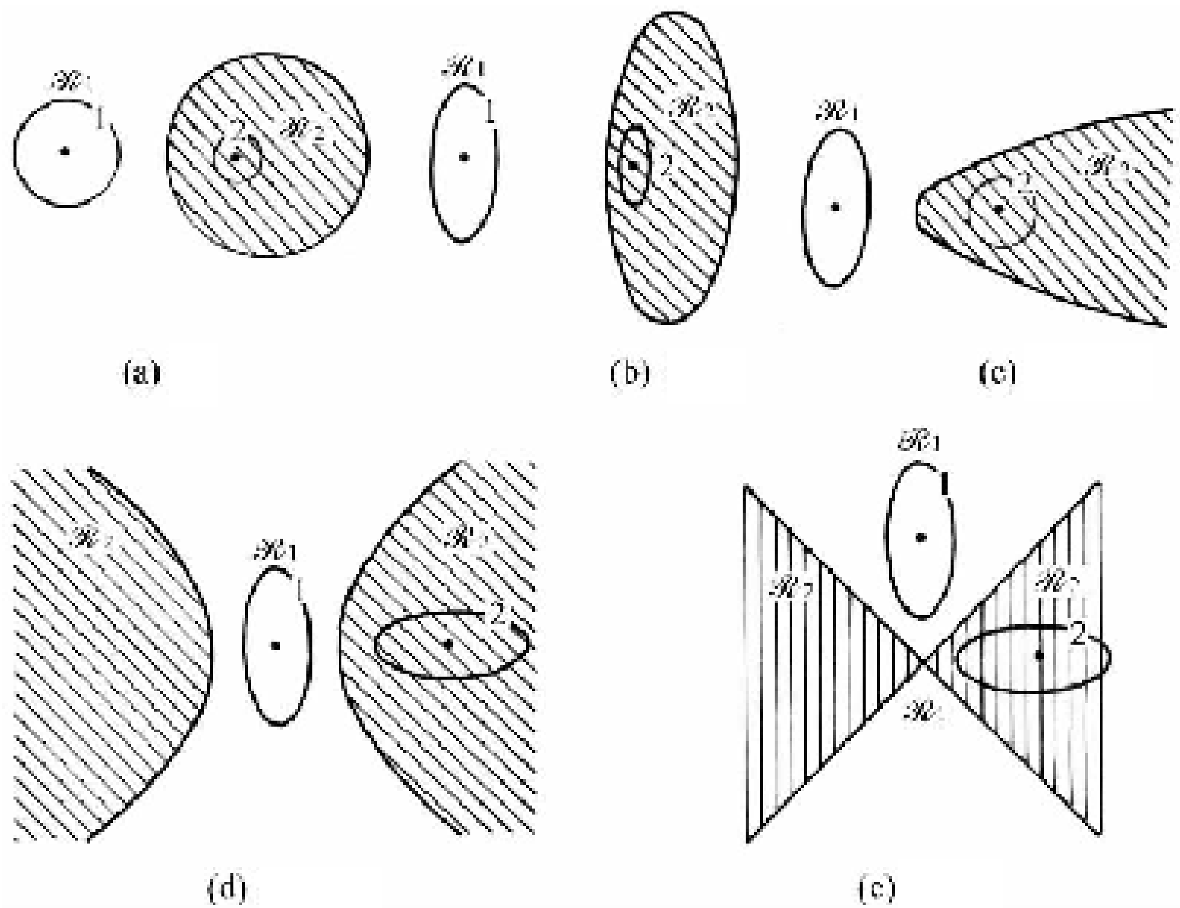
$$\text{where } W_i = -\frac{1}{2}\Sigma_i^{-1}$$

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

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

Decision boundary:

$$x^t (W_i - W_j) x + (w_i - w_j)^t x + w_{i0} - w_{j0} = 0$$



Parameters Learning

Optimal classifier:

$$f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$$

$$= \arg\max_{Y=y} \underbrace{P(X = x|Y = y)}_{\text{Class conditional density}} \underbrace{P(Y = y)}_{\text{Class prior}}$$

Need to know Prior P(Y = y) for all y

Likelihood P(X=x|Y = y) for all x,y