CT 720 Tópicos em Aprendizagem de Máquina e Classificação de Padrões

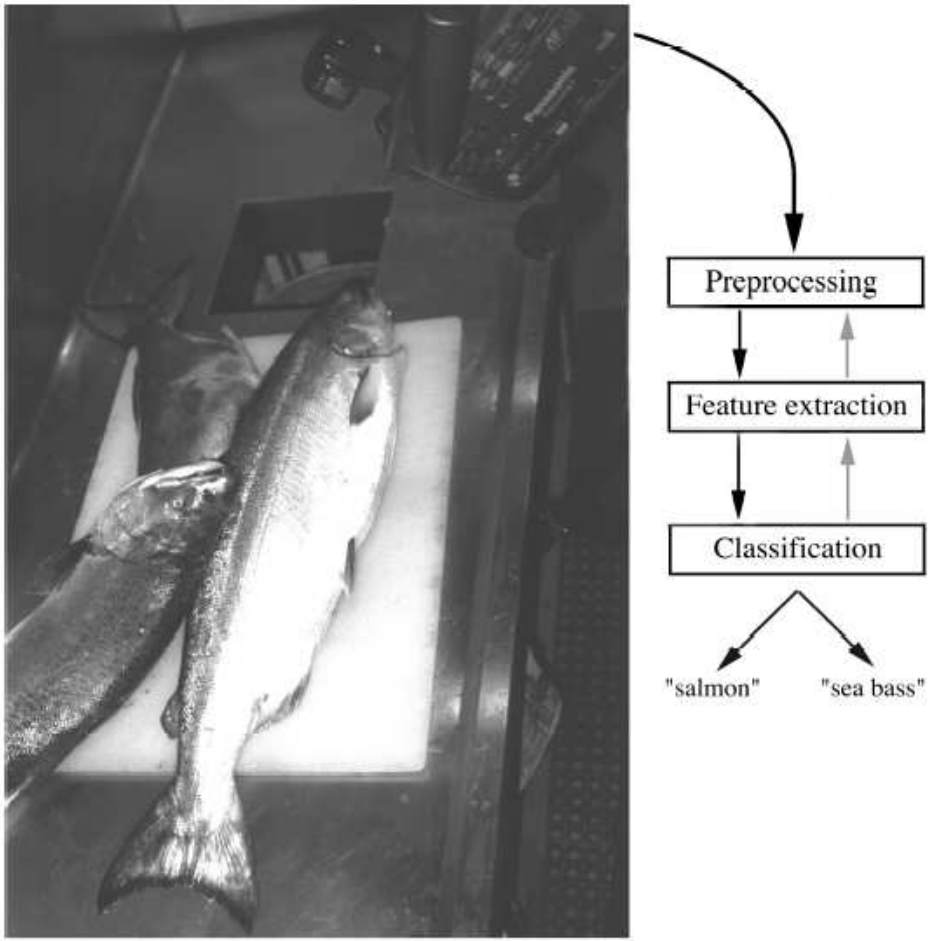
2-Teoria Bayesiana de Decisão

Conteúdo

1. Introdução
2. Teoria Bayesiana de decisão: atributos contínuos
3. Classificação com taxa de erro mínima
4. Funções de discriminação e classificadores
5. Densidade normal
6. Funções de discriminação para densidade normal
7. Teoria Bayesiana de decisão: atributos discretos
8. Redes Bayesianas
9. Resumo

1-Introdução

- Teoria Bayesiana de decisão
  - abordagem estatística para reconhecimento padrões
  - assume problema de decisão formulado probabilisticamente
  - todos valores relevantes de probabilidade conhecidos
  - identificação de sequências DNA
- Este capítulo
  - apresenta fundamentos da teoria
  - teoria: formaliza procedimentos intuitivos


Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"

- Previsão próximo tipo peixe?

  ω = estado : variável aleatória

  $\omega = \omega_1$ *sea bass*

  $\omega = \omega_2$ *salmon*

- Assumindo

  $P(\omega_1)$: probabilidade *a priori* próximo tipo é *sea bass*

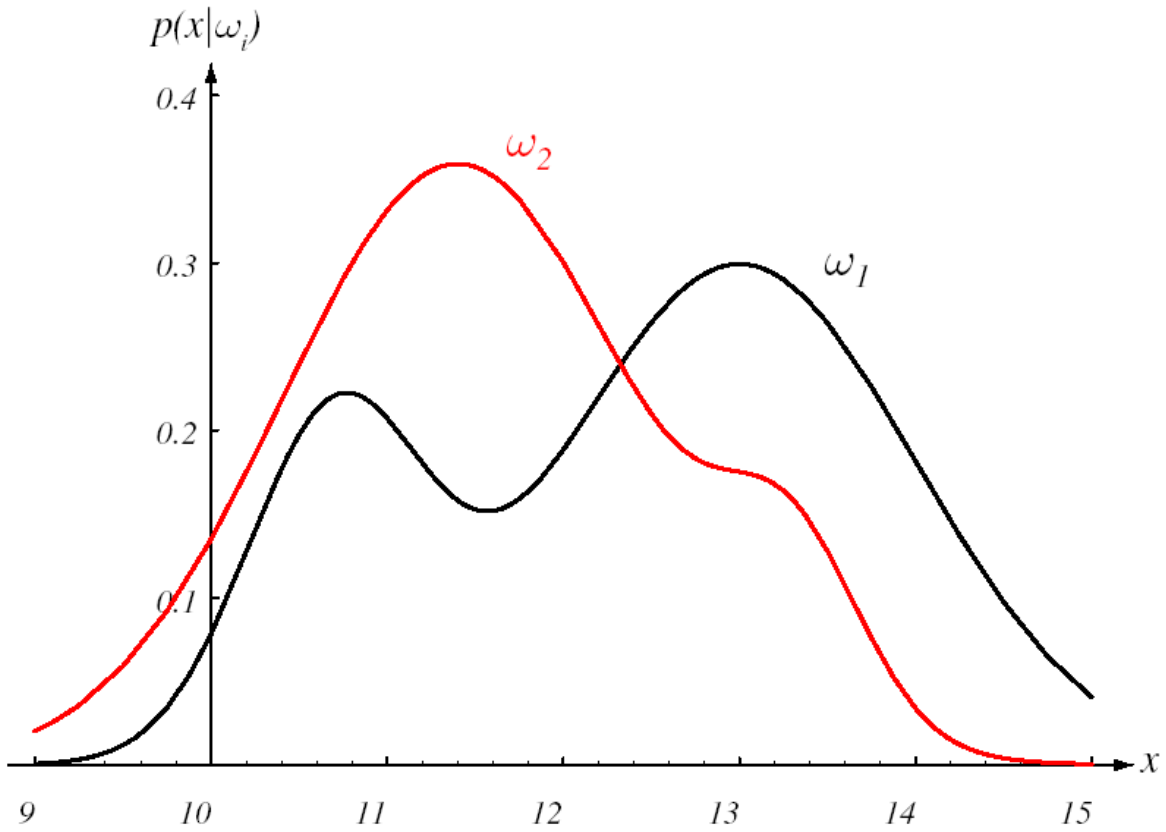
  $P(\omega_1)$: probabilidade *a priori* próximo tipo é *salmon*

  $P(\omega_1) + P(\omega_2) = 1$

- Decidir tipo próximo peixe
  - classificação incorreta: mesmo custo
  - informação disponíveis: somente $P(\omega_1)$ e $P(\omega_2)$

- Regra de decisão
  - decidir ω_1 se $P(\omega_1) > P(\omega_2)$; caso contrário decidir ω_2
  - regra: mesma decisão, mesmo sabendo que existem 2 tipos
  - desempenho depende da escolha de $P(\omega_1)$ e $P(\omega_2)$

- Na prática
  - temos mais informação
  - e.g. medida luminosidade (x)
  - densidade probabilidade condicional de classe $p(x/\omega)$
  - função densidade de probabilidade de x dado estado ω
  - $p(x/\omega_1)$ e $p(x/\omega_2)$ descrevem diferenças de luminosidade



- Supor que conhecemos:
  - $P(\omega_1)$ e $P(\omega_2)$
  - $p(x|\omega_1)$ e $p(x|\omega_2)$
  - medida de luminosidade x
- Como x influencia a escolha correta do tipo?
  - estimativa correta do estado ?
  - como decidir o tipo (classe)

- Classificação usando luminosidade como atributo

$$p(\omega_j, x) = P(\omega_j \mid x)\, p(x) = p(x \mid \omega_j) P(\omega_j)$$

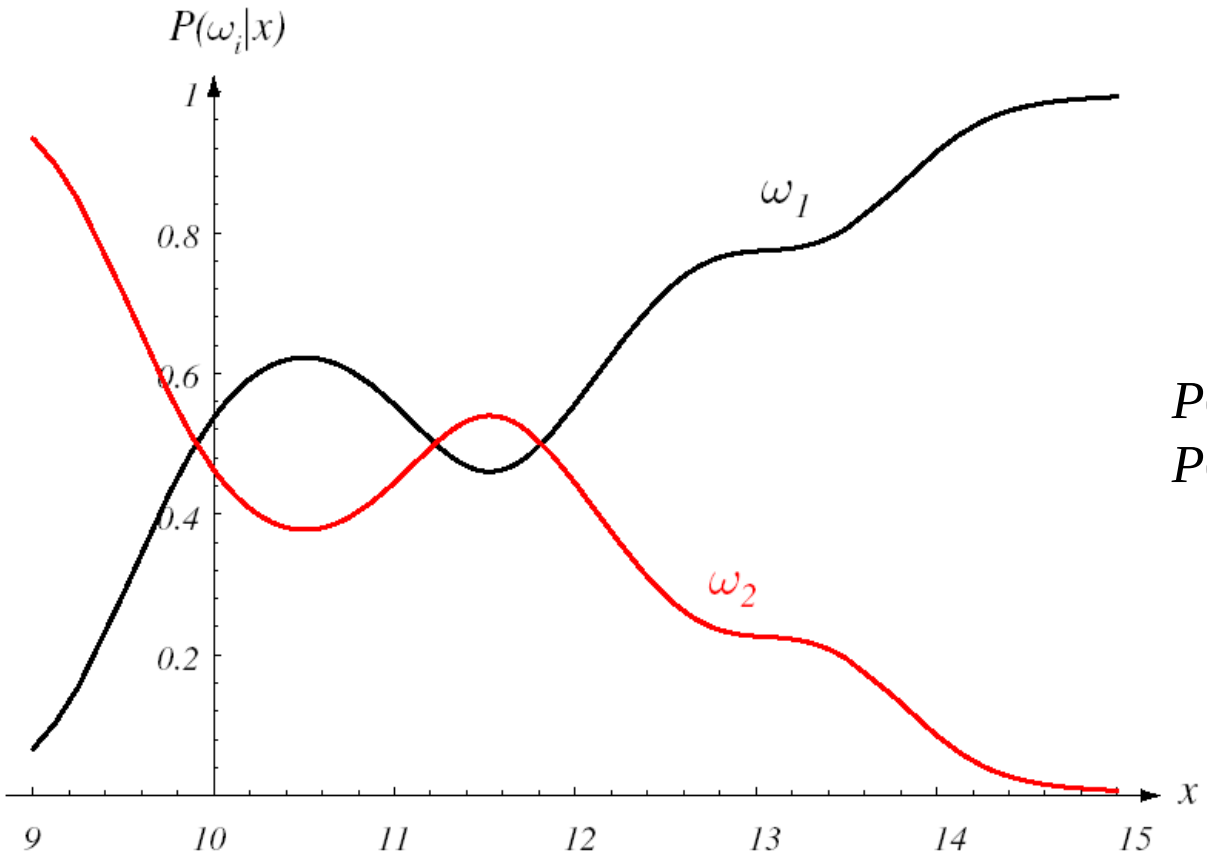
$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

Bayes

$$p(x) = \sum_{j=1}^{2} p(x \mid \omega_j) P(\omega_j)$$

$$P(causa / efeito) = \frac{p(efeito / causa) P(causa)}{P(efeito)}$$

$$posterior = \frac{likelyhood \times prior}{evidence}$$



$P(\omega_1) = 2/3$
$P(\omega_2) = 1/3$

Regra de decisão de Bayes

- decidir ω_1 se $P(\omega_1|\, x) > P(\omega_2|\, x)$; caso contrário decidir ω_2 (*)

- Regra de Bayes minimiza a probabilidade de erro

$$P(error \mid x) = \begin{cases} P(\omega_1 \mid x) & \text{se decidimos } \omega_2 \\ P(\omega_2 \mid x) & \text{se decidimos } \omega_1 \end{cases}$$

  - para uma observação x minimiza-se a probabilidade de erro:
    - decidindo ω_1 se $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; ou ω_2 caso contrário
  - esta regra minimiza a probabilidade de erro para qualquer x, em média?

– em média a probabilidade do erro é

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error \mid x)p(x)dx$$

– logo a integral $P(error)$ é mínima se $P(error/x)$ é mínima, isto é, se

- decidir ω_1 se $P(\omega_1 | x) > P(\omega_2 | x)$; caso contrário decidir ω_2

– $P(error/x) = \min [P(\omega_1 | x) , P(\omega_2 | x)]$

Regra de decisão de Bayes (alternativa)

- decidir ω_1 se $p(x|\omega_1)P(\omega_1) > p(x|\omega_2) P(\omega_2)$; caso contrário decidir ω_2

2-Teoria Bayesiana de decisão: atributos contínuos

- Generalização
  - uso de vários atributos
  - mais de dois estados
  - outras decisões e não só classificação
  - critério mais geral que probabilidade de erro

- Notação

$\mathbf{x}$: atributo, $\mathbf{x} \in \mathbf{R}^d$

$\mathbf{R}^d$: espaço (Euclideano) de atributos

$\{\omega_1, ..., \omega_c\}$: conjunto (finito) de c estados (categorias)

$\{\alpha_1, ..., \alpha_a\}$: conjunto (finito) de a decisões (ações)

$\lambda(\alpha_i, \omega_j)$: *loss function* = custo decisão α_i quando em ω_j

- Regra de Bayes

$$P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j) P(\omega_j)}{p(\mathbf{x})}$$

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j) P(\omega_j)$$

– supor observação **x** e ação α_i correspondente

– se estado verdadeiro é ω_j, então custo associado é $\lambda(\alpha_i, \omega_j)$

– valor esperado do custo da ação α_i será:

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

Risco condicional

– dado **x,** que ação α_i minimiza o risco condicional, $\forall$**x** ?

– solução (ótima): regra de Bayes !

- Regra (estratégia) de decisão

  – $\alpha(\mathbf{x})$: fornece a decisão para cada valor de $\mathbf{x}$
  – para cada $\mathbf{x}$, $\alpha(\mathbf{x})$ assume um dos a valores $\alpha_1, ..., \alpha_a$
  – risco global: risco associado com uma estratégia de decisão

$$R = \int R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$ Risco global

  – se $\alpha(\mathbf{x})$ é tal que o valor de $R(\alpha_i(\mathbf{x}))$ é o menor possível $\forall \mathbf{x}$, então risco global é minimizado; isto motiva o seguinte:

- Regra de Bayes

para minimizar risco global:

1 – calcular

$$R(\alpha_i \mid \mathbf{x}) = \sum_{i=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x}), \quad i = 1, \ldots, a$$

2 – selecionar ação α_i que minimiza $R(\alpha_i|\mathbf{x})$

R^* = risco de Bayes

- Exemplo com duas classes

  – α_1 : decide que estado verdadeiro é ω_1
  – α_2 : decide que estado verdadeiro é ω_2
  – $\lambda_{ij} = \lambda(\alpha_i, \omega_j)$ custo quando decisão $\Rightarrow \omega_i$ mas verdadeiro é ω_j

  risco condicional

$$R(\alpha_1 \mid \mathbf{x}) = \lambda_{11} P(\omega_1 \mid \mathbf{x}) + \lambda_{12} P(\omega_2 \mid \mathbf{x})$$
$$R(\alpha_2 \mid \mathbf{x}) = \lambda_{21} P(\omega_1 \mid \mathbf{x}) + \lambda_{22} P(\omega_2 \mid \mathbf{x})$$

  - decidir ω_1 se $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$

– alternativamente, em termos das probabilidades *a posteriori*

- decidir ω_1 se $(\lambda_{21} - \lambda_{11})P(\omega_1|\,\mathbf{x}) > (\lambda_{12} - \lambda_{22})\,P(\omega_2|\,\mathbf{x})$

– em geral $\lambda_{21} > \lambda_{11}$ e $\lambda_{12} > \lambda_{22}$

– utilizando Bayes, probabilidades *a priori* e densidades condicionais

$$\frac{P(\mathbf{x}\,|\,\omega_1)}{P(\mathbf{x}\,|\,\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})}\frac{P(\omega_2)}{P(\omega_1)}$$

Razão de verosimilhança

3-Classificação com taxa de erro mínima

$$\lambda(\alpha_i \mid \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

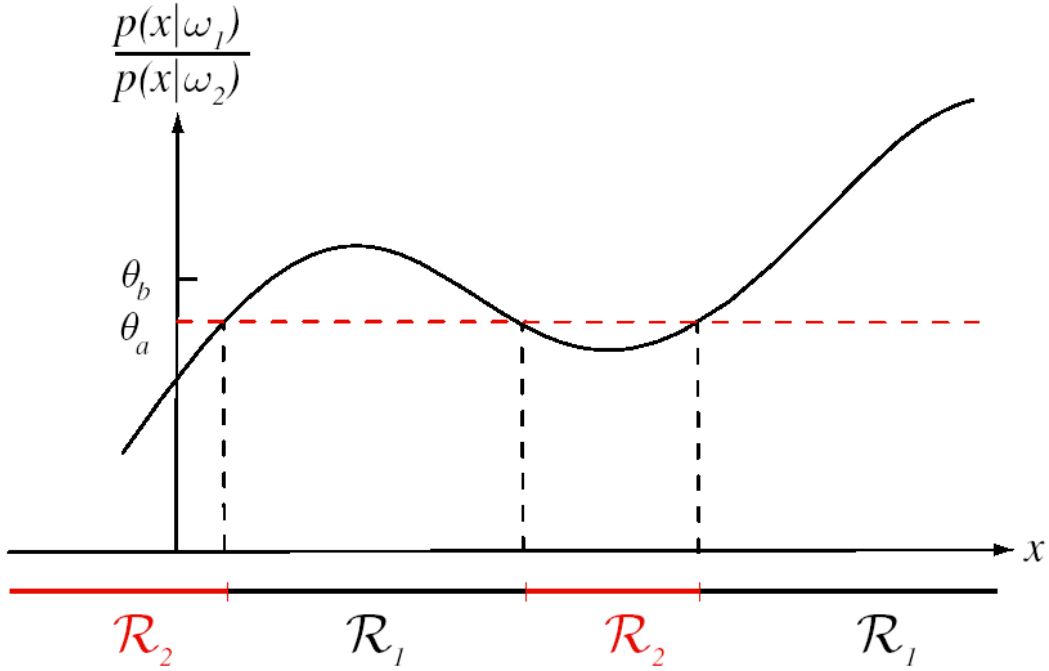
$$\begin{aligned} R(\alpha_i \mid \mathbf{x}) &= \sum_{i=1}^{c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x}) \\ &= \sum_{i \neq j} P(\omega_j \mid \mathbf{x}) \\ &= 1 - P(\omega_i \mid \mathbf{x}) \end{aligned}$$

- Regra Bayes para taxa de erro mínima

  – minimizar risco: selecionar ação que minimiza risco condicional

  – minimizar taxa de erro de classificação:

    - decidir ω_i se $P(\omega_i \mid \mathbf{x}) > P(\omega_j \mid \mathbf{x}) \quad \forall\, i, j$ (ver *)



4-Funções discriminação e classificadores

- Função discriminação

  $g_i(\mathbf{x})$, $i = 1,..., c$

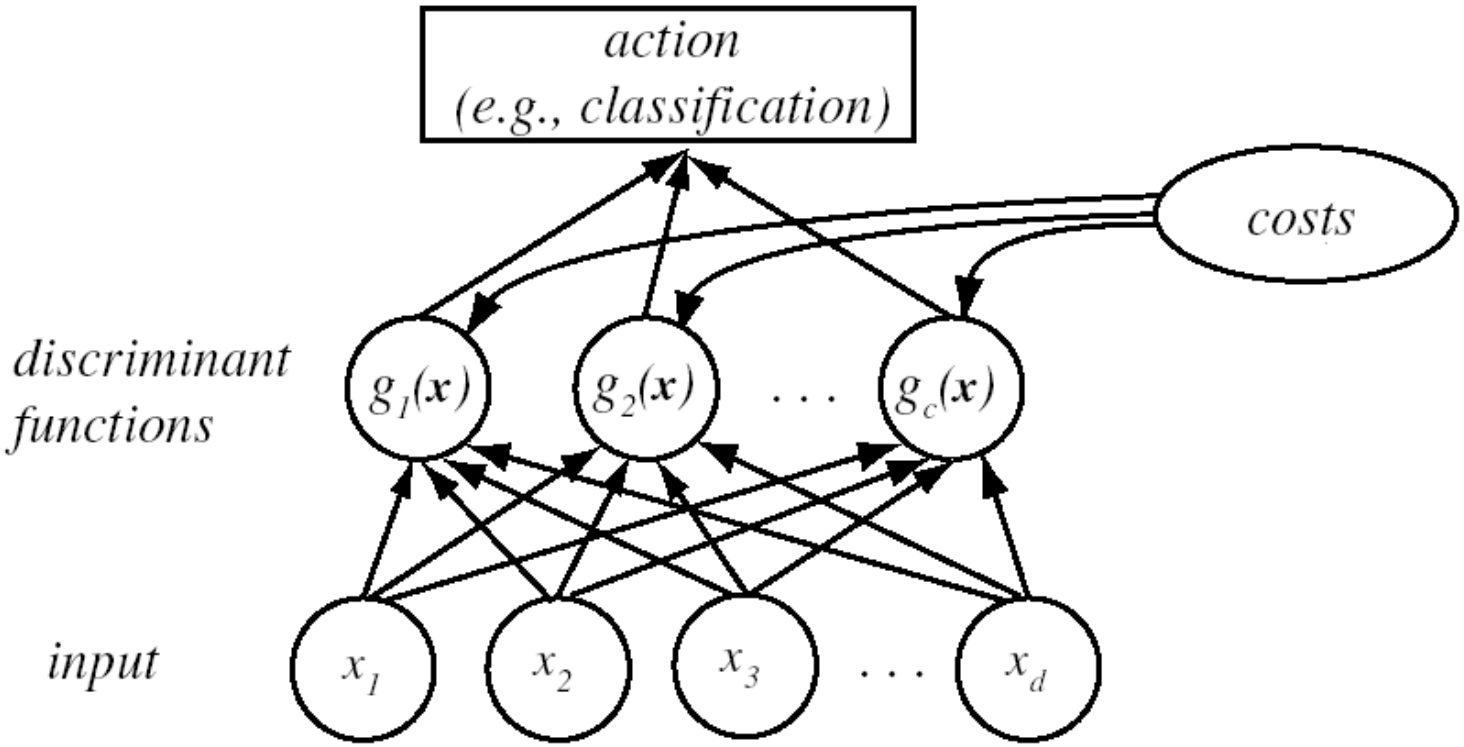
  classificador atribui classe ω_i a $\mathbf{x}$ se $g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall\, j \neq i$

  Exemplos:

  $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$ risco condicional mínimo

  $g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$ erro classificação mínimo

Estrutura funcional de classificadores estatísticos



- Propriedades

  – funções discriminação não são únicas
  – transformações: $f(g_i(\mathbf{x}))$, f monotônica crescente
  – simplificação analítica e computacional
  – exemplos de classificadores (erro classificação mínimo):

$$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x} \mid \omega_j)P(\omega_j)}$$

$$g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)P(\omega_i)$$

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

– formas funções diferentes, mas regras de decisão equivalentes

– efeito: dividir o espaço de atributos em c regiões distintas

- se $g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \forall\, j \neq i$ então $\mathbf{x} \in \mathcal{R}_i$

– $\mathcal{R}_i \cap \mathcal{R}_j = \emptyset$

– $\mathcal{R}_1$,..., $\mathcal{R}_c$ formam uma partição do espaço de atributos

– exemplo: duas categorias (classes)

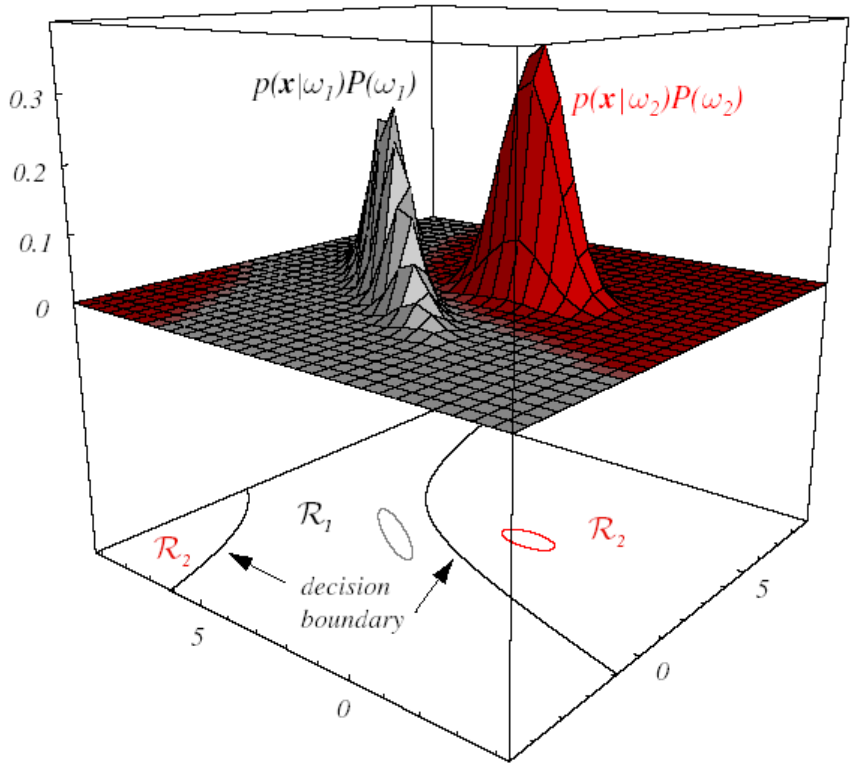
- atribuir ω_1 a $\mathbf{x}$ se $g_1(\mathbf{x}) > g_2(\mathbf{x}) \quad \forall\, j \neq i$

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$

- atribuir ω_1 a $\mathbf{x}$ se $g(\mathbf{x}) > 0$ (***)

$$g(\mathbf{x}) = P(\omega_1 \mid \mathbf{x}) - P(\omega_2 \mid \mathbf{x})$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \quad (**)$$



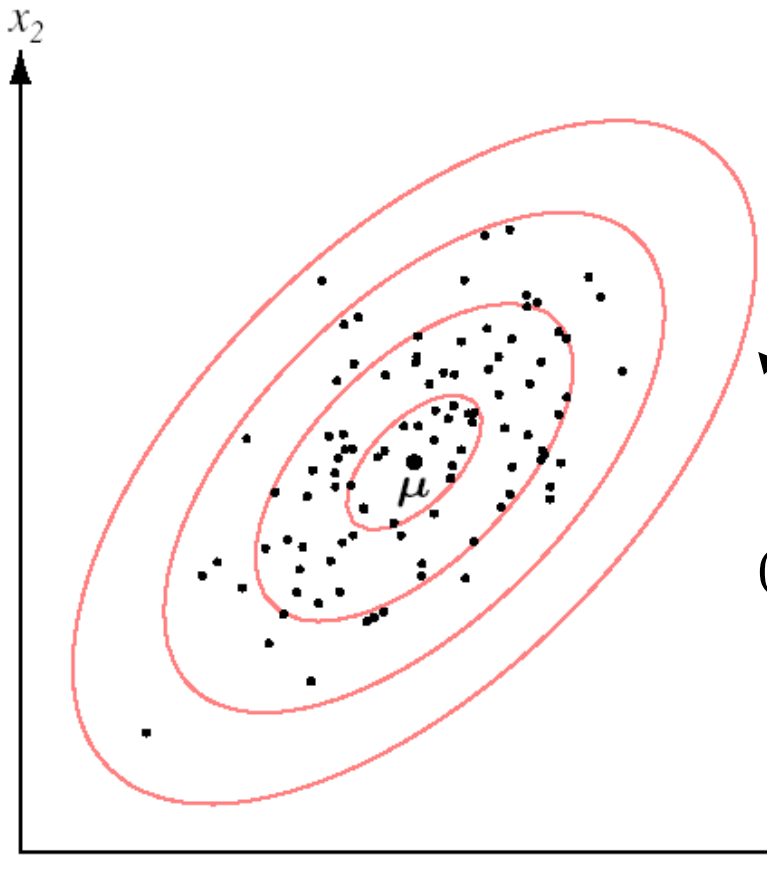
5-Densidade normal

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

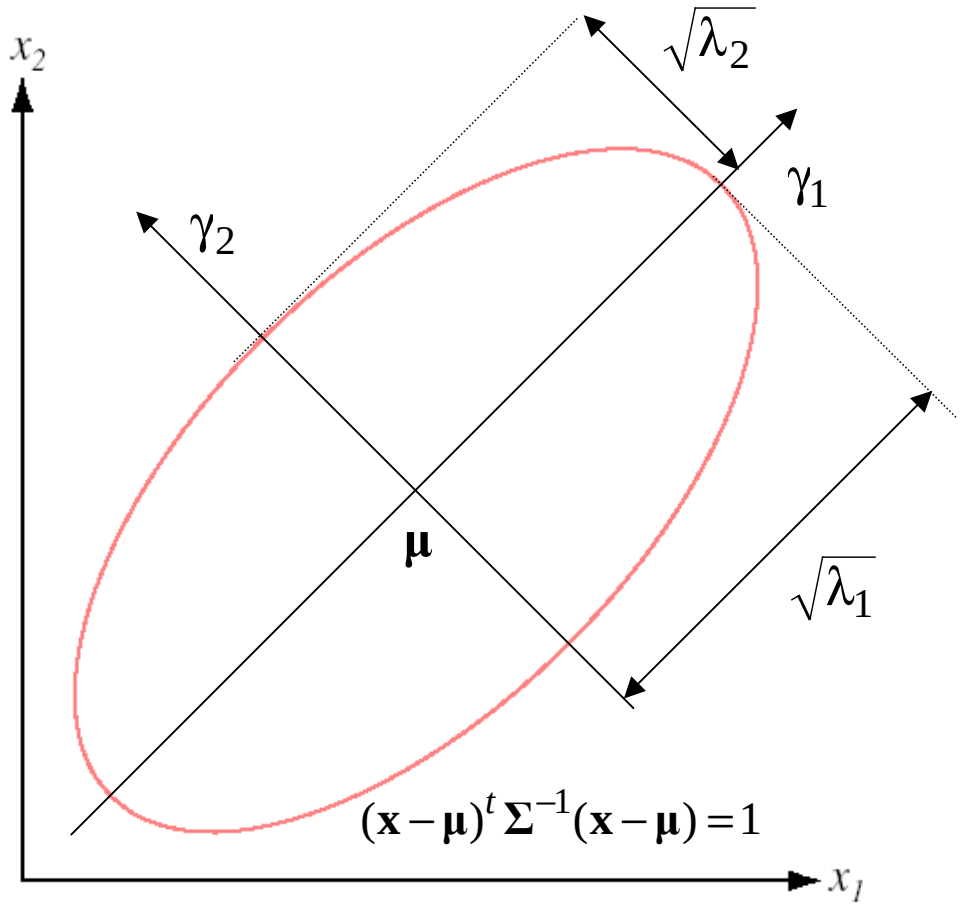
$$\boldsymbol{\mu} = E[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x}\, p(\mathbf{x})\, d\mathbf{x}, \qquad \mu_i = E[x_i]$$

$$\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int_{-\infty}^{\infty} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t\, p(\mathbf{x})\, d\mathbf{x}$$

$$\boldsymbol{\Sigma} = [\sigma_{ij}], \quad \sigma_{ij} = E[(x_i-\mu_i)(x_j-\mu_j)], \quad \boldsymbol{\Sigma} > 0$$



$$(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = cte$$



- Transformações lineares

  – combinações lineares de variáveis aleatórias normais são normais

  $p(\mathbf{x}) = N(\mu, \Sigma)$

  $\mathbf{A}$ $(d \times k)$

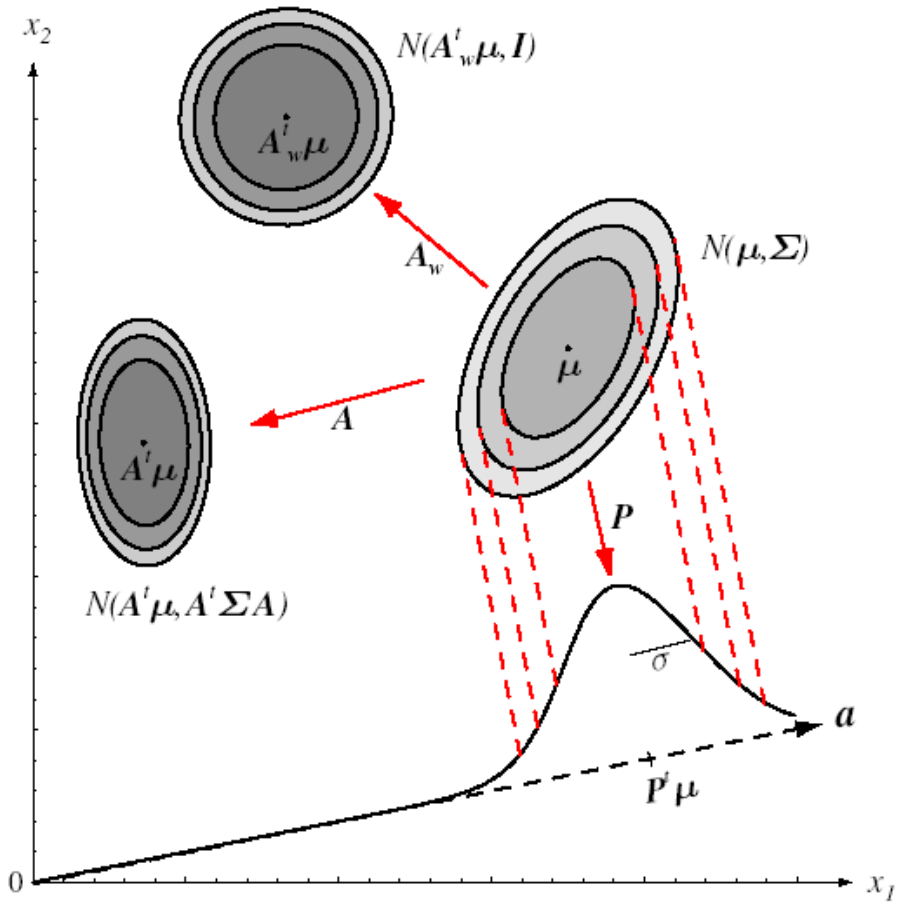
  $\mathbf{y} = \mathbf{A}^t\mathbf{x}$ $(k \times 1)$

  $p(\mathbf{y}) = N(\mathbf{A}^t\mu, \mathbf{A}^t\Sigma\mathbf{A})$

  – se $\mathbf{A} = \mathbf{a}$ $(d \times 1)$, $k = 1$, $\|\mathbf{a}\| = 1$ então

  $y = \mathbf{a}^t\mathbf{x}$ (escalar que representa projeção de $\mathbf{x}$ ao longo de $\mathbf{a}$)

  $\mathbf{a}^t\Sigma\mathbf{a}$ variância da projeção de $\mathbf{x}$ ao longo de $\mathbf{a}$



- Observações

$$r^2 = (\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Distância de Mahalanobis

$$V = V_d \, |\boldsymbol{\Sigma}|^{1/2} \, r^d$$

Volume hiperelipsóide

$$V_d = \begin{cases} \pi^{d/2}/(d/2)! & d \ \ par \\ 2^d \pi^{(d-1)/2}(d-1/2)!/d! & d \ \ impar \end{cases}$$

Volume hiperesfera

6-Funções discriminação p/ densidade normal

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

$$p(\mathbf{x} \mid \omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

- Caso 1: $\sum_i = \sigma^2\mathbf{I}$

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x}-\boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i] + \ln P(\omega_i)$$

termo quadrático é o mesmo $\forall\ i$

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

Máquina linear

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i\,, \qquad w_{io} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i)$$

Limiar (*bias*) para *i*-ésima classe

$$g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$$

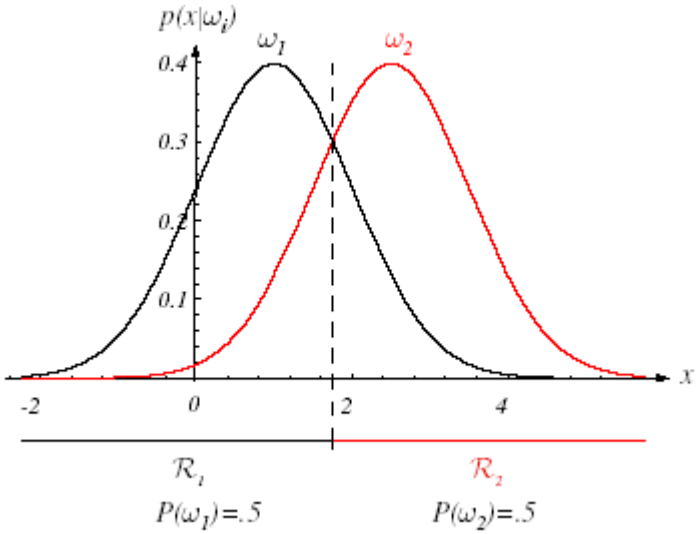
$$\mathbf{w}^t(\mathbf{x} - \mathbf{x}_o) = 0$$

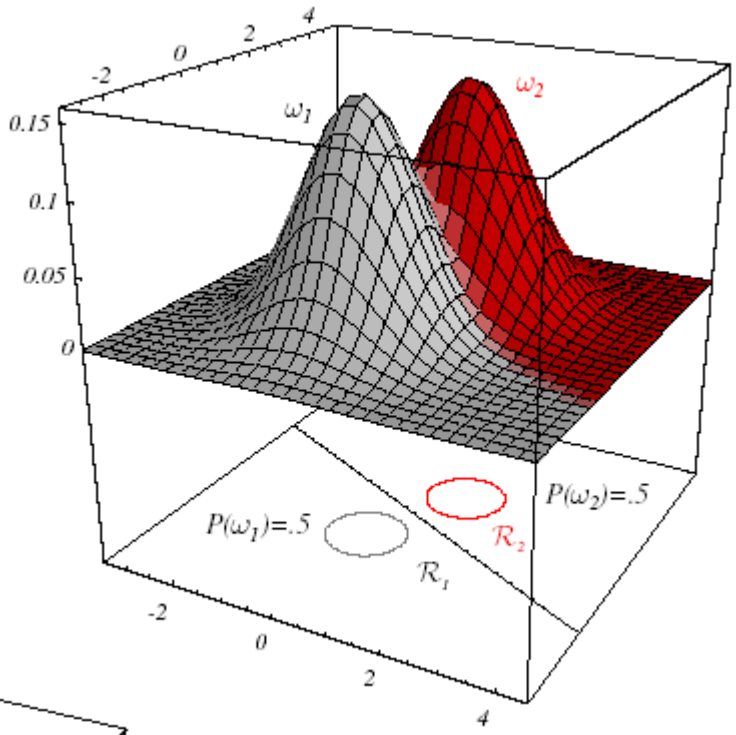
$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

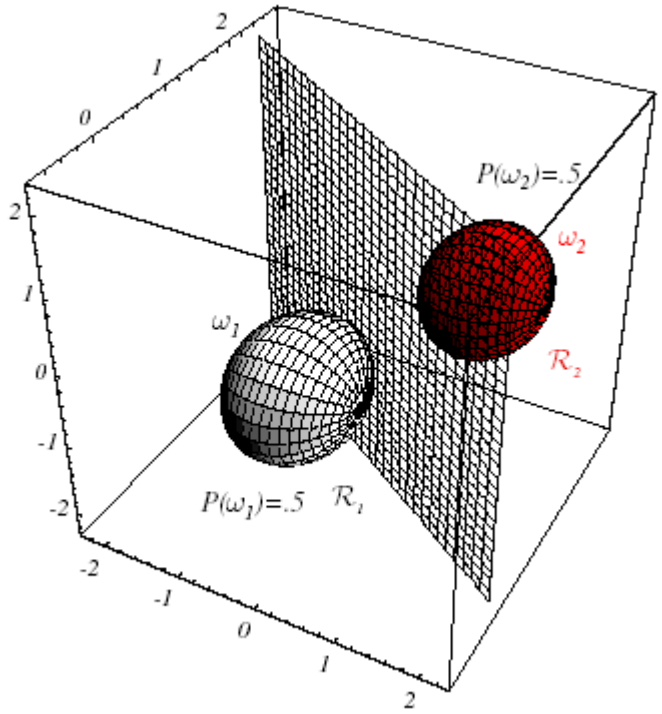
Hiperplano separa $\mathcal{R}_i$ e $\mathcal{R}_j$ passa por $\mathbf{x}_o$ e é ortogonal à reta que une as médias

$$\mathbf{x}_o = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\left\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

$$P(\omega_i) = P(\omega_j) \quad \Rightarrow \quad \mathbf{x}_o = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)$$


$p(x|\omega_i)$
ω_1
ω_2
0.4
0.3
0.2
0.1
x
-2
0
2
4
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_1)=.5$
$P(\omega_2)=.5$


0.15
0.1
0.05
0
ω_1
ω_2
$P(\omega_1)=.5$
$P(\omega_2)=.5$
$\mathcal{R}_1$
$\mathcal{R}_2$

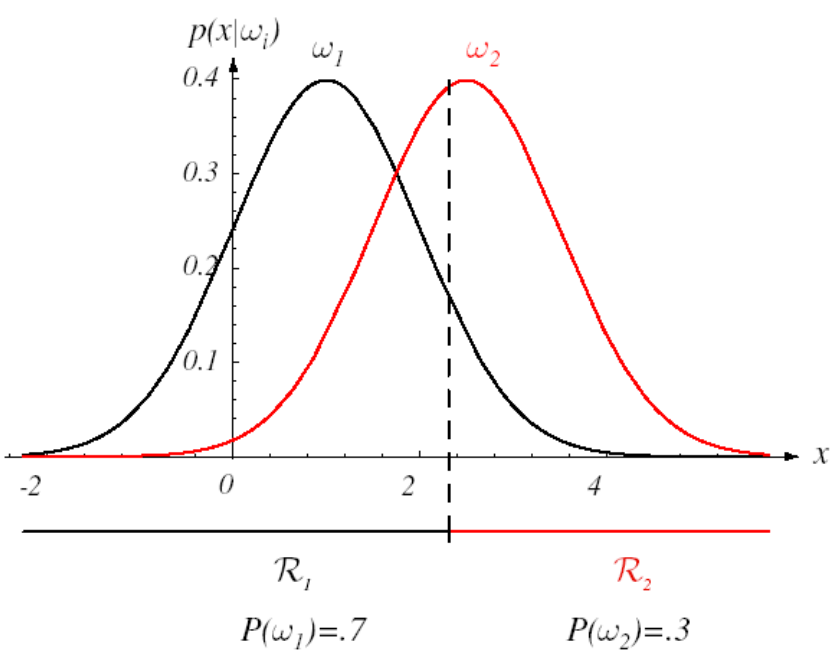

$P(\omega_2)=.5$
ω_2
ω_1
$\mathcal{R}_2$
$P(\omega_1)=.5$
$\mathcal{R}_1$

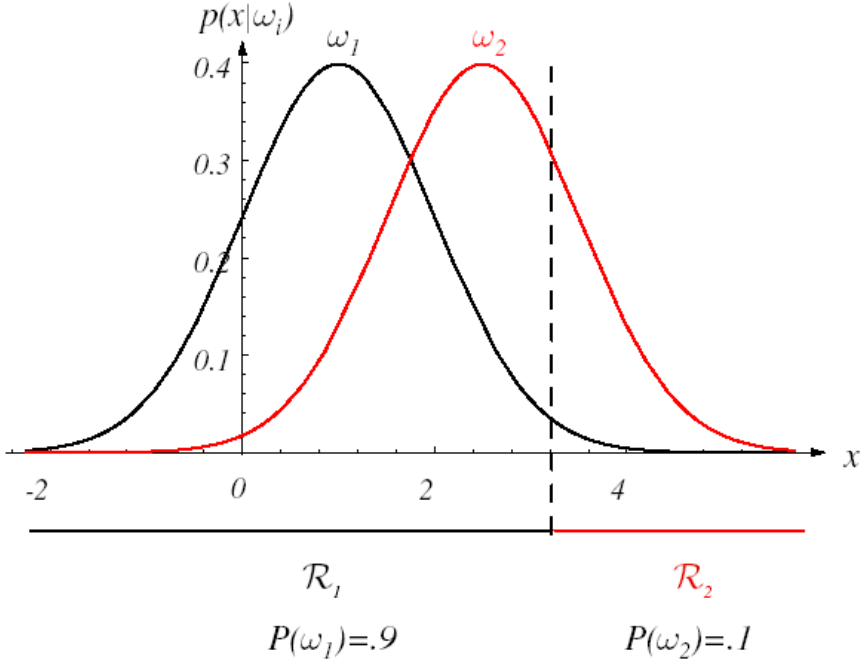
$P(\omega_i) \neq P(\omega_j)$

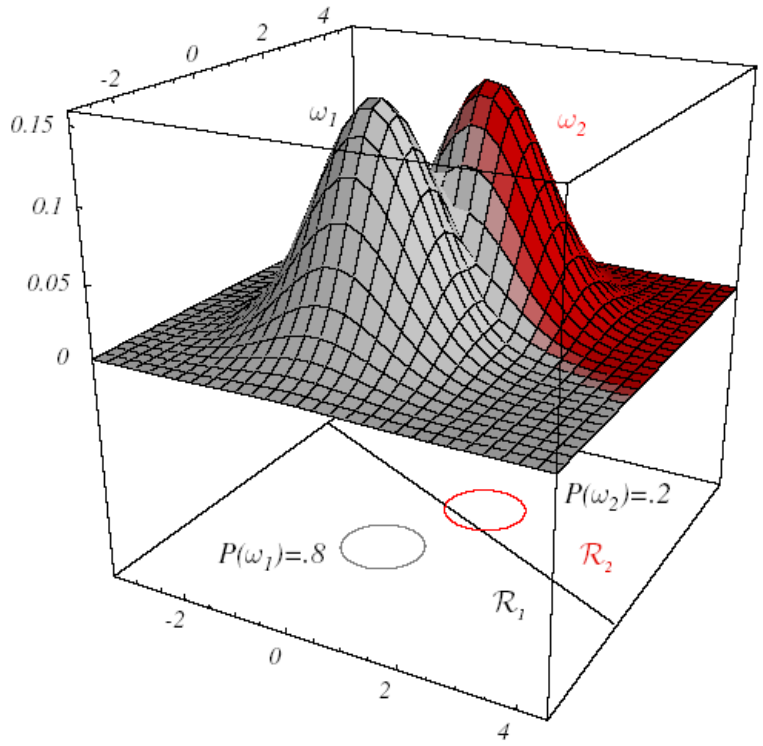
$\mathbf{x}_o$ se afasta da média mais provável

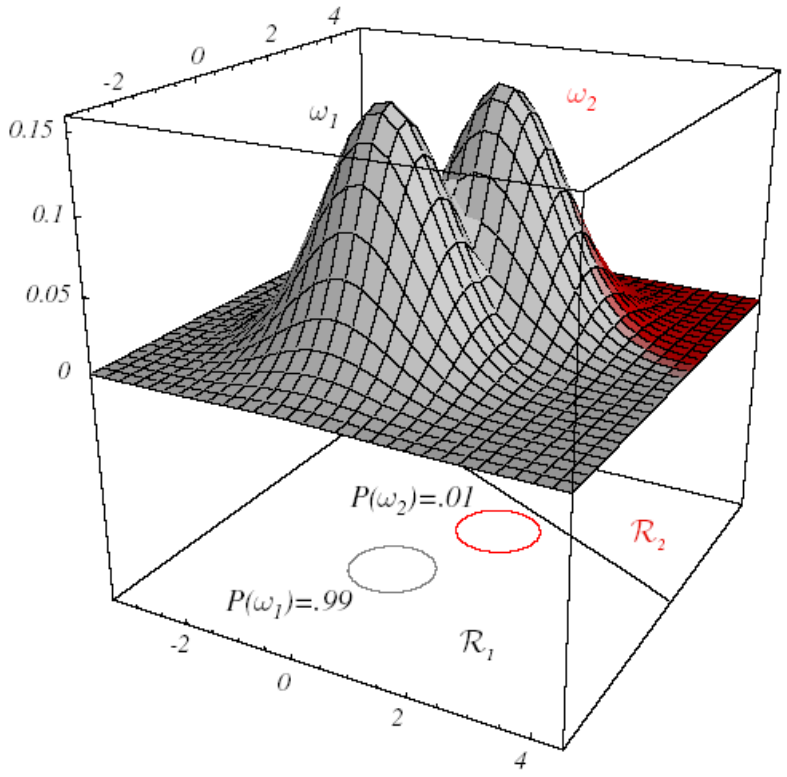
$P(\omega_i) = P(\omega_j), \quad \forall i, j$ termo $\ln P(\omega_i)$ irrelevante

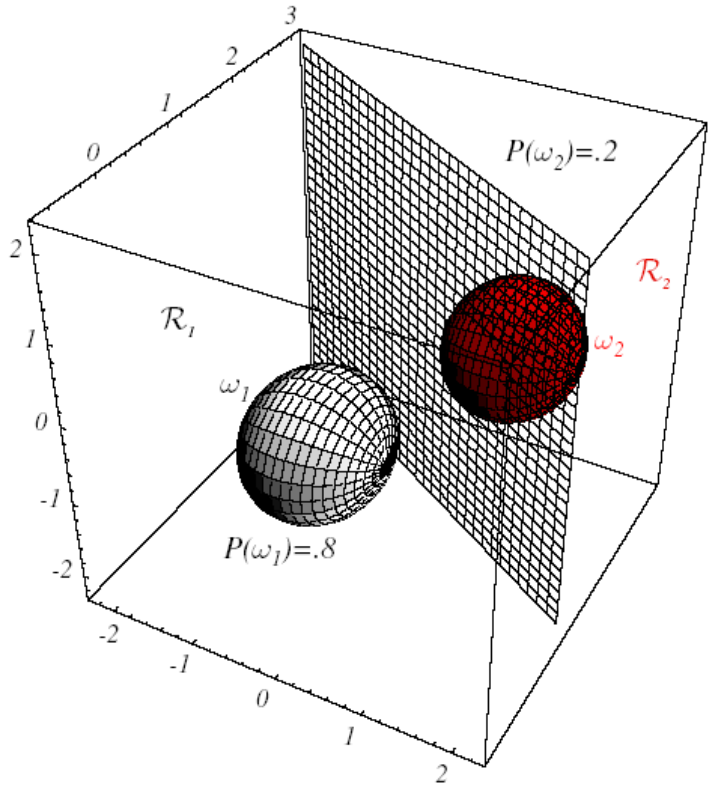
$g_i(\mathbf{x}) = \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$ Classificador distância mínima

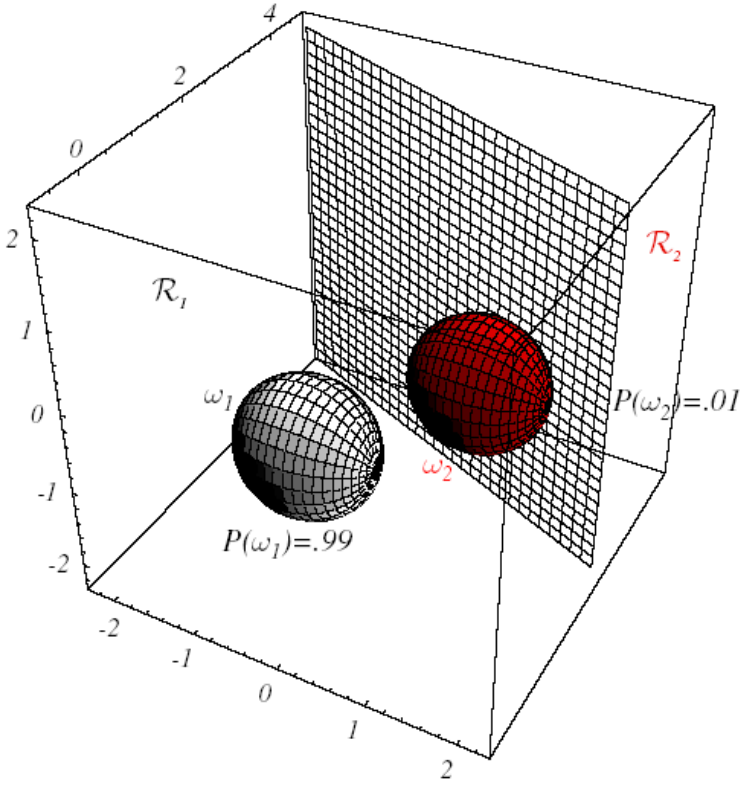

$p(x|\omega_i)$
ω_1
ω_2
0.4
0.3
0.2
0.1
-2
0
2
4
x
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_1)=.7$
$P(\omega_2)=.3$


$p(x|\omega_i)$
ω_1
ω_2
0.4
0.3
0.2
0.1
-2
0
2
4
x
$\mathcal{R}_1$
$\mathcal{R}_2$
$P(\omega_1)=.9$
$P(\omega_2)=.1$


0.15
0.1
0.05
0
-2
0
2
4
ω_1
ω_2
$P(\omega_2)=.2$
$P(\omega_1)=.8$
$\mathcal{R}_2$
$\mathcal{R}_1$


0.15
0.1
0.05
0
-2
0
2
4
ω_1
ω_2
$P(\omega_2)=.01$
$P(\omega_1)=.99$
$\mathcal{R}_2$
$\mathcal{R}_1$


3
2
1
0
2
1
0
-1
-2
-2
-1
0
1
2
$P(\omega_2)$=.2
$\mathcal{R}_2$
$\mathcal{R}_1$
ω_2
ω_1
$P(\omega_1)$=.8


4
2
0
2
1
0
-1
-2
-2
-1
0
1
2
$\mathcal{R}_2$
$\mathcal{R}_1$
ω_1
$P(\omega_2)$=.01
ω_2
$P(\omega_1)$=.99

- Caso 2: $\Sigma_i = \Sigma$

independente de i

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + \ln P(\omega_i)$$

se $P(\omega_i) = P(\omega_j), \quad \forall i, j$ termo $\ln P(\omega_i)$ irrelevante, logo:

$$g_i(\mathbf{x}) = \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

Classificador distância (Mahalanobis) mínima

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

eliminando $\mathbf{x}^t \sum^{-1}\mathbf{x}$ da expansão de $(\mathbf{x}-\boldsymbol{\mu})^t \sum^{-1}(\mathbf{x}-\boldsymbol{\mu})$ (independe de i)

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{io}$$ Máquina linear

$$\mathbf{w}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i \qquad w_{io} = -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i)$$

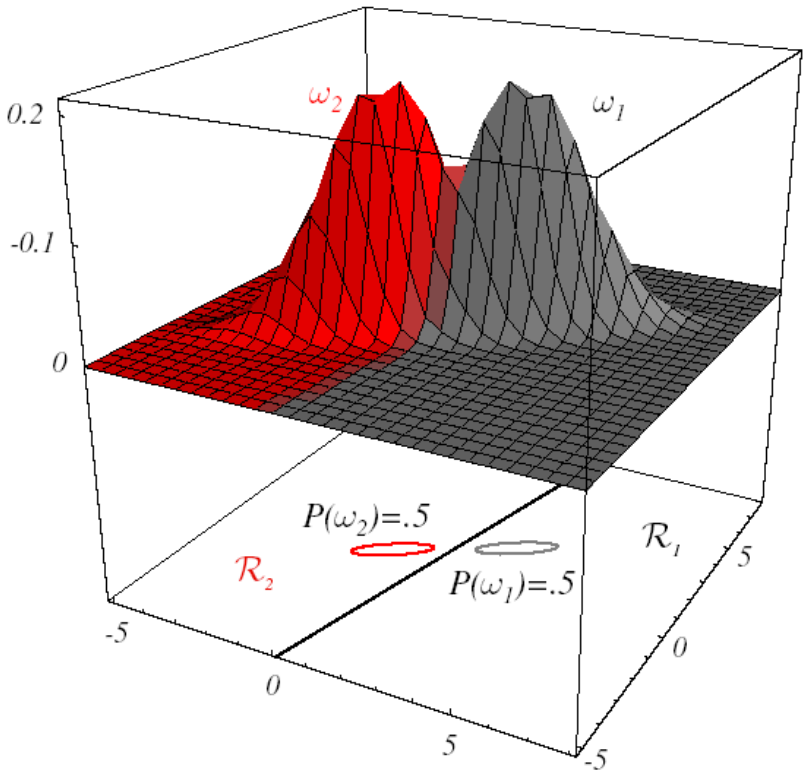
se $\mathcal{R}_i$ e $\mathcal{R}_j$ são contíguas, a superfície de decisão é um hiperplano

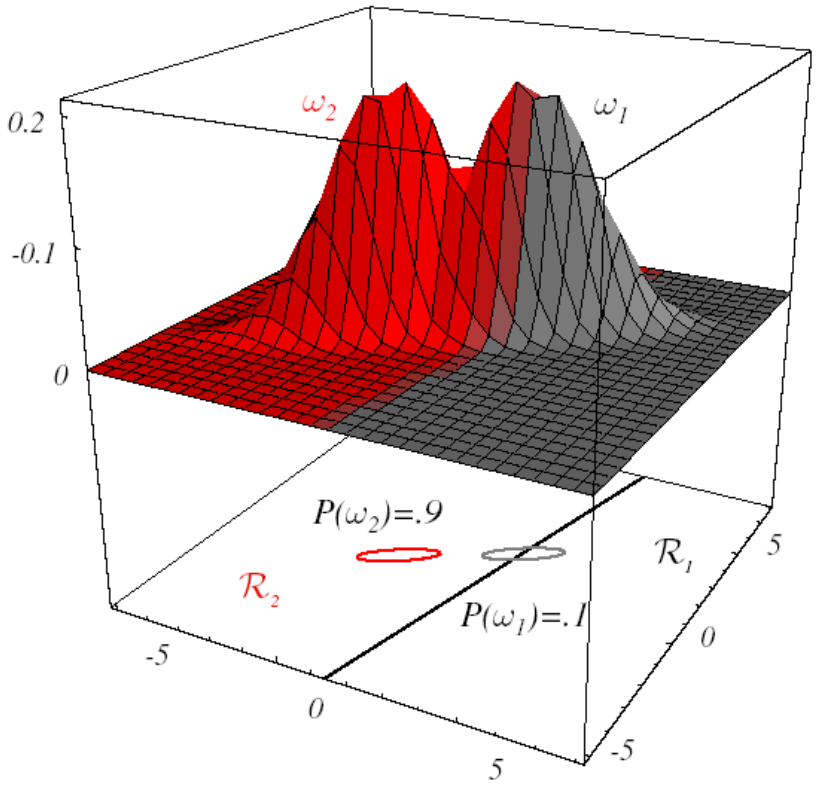
$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_o)=0$$

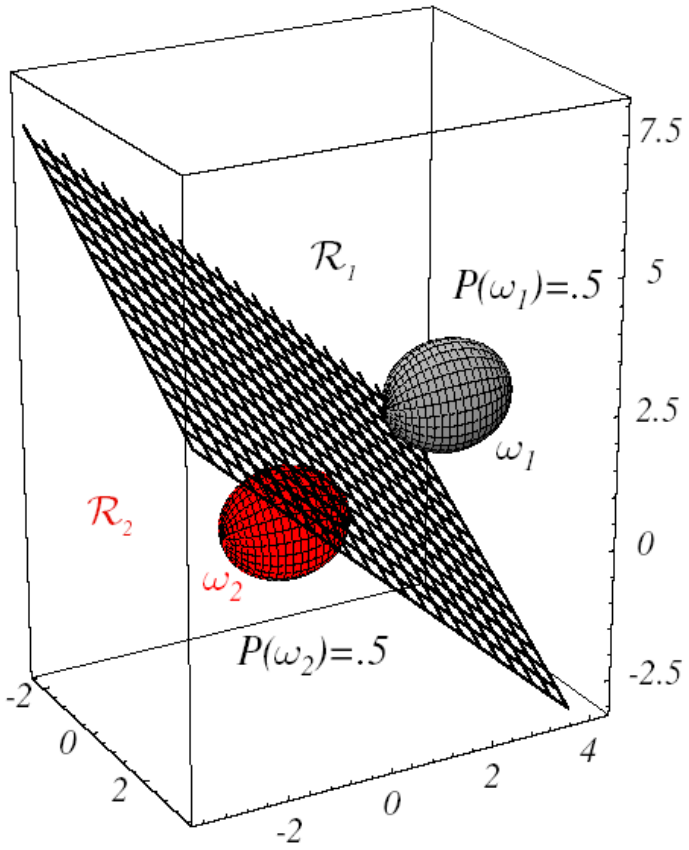
$$\mathbf{w}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

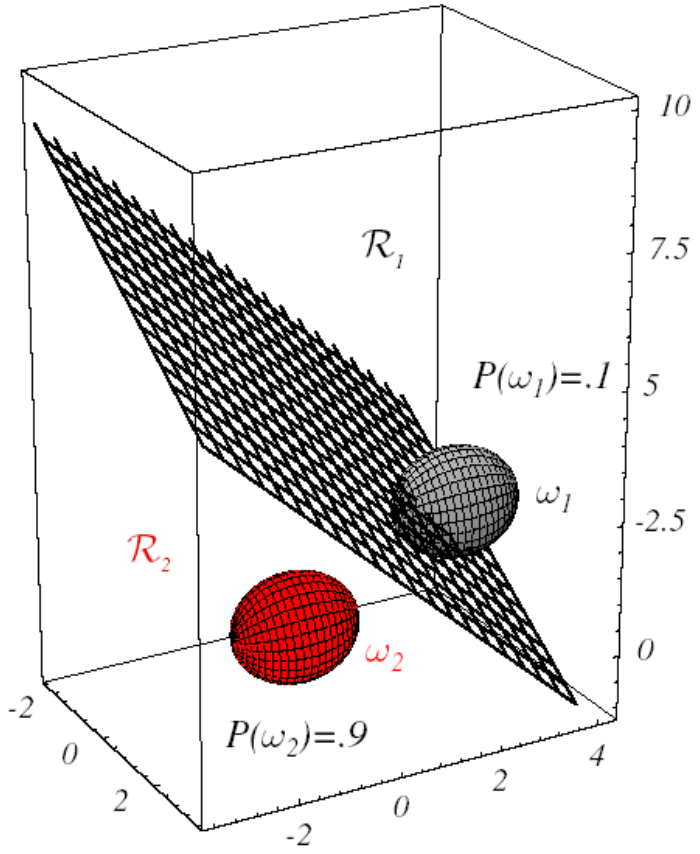
$$\mathbf{x}_o=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\ln[P(\omega_i)/P(\omega_j)]}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\Sigma^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$$

– hiperplano separa $\mathcal{R}_i$ e $\mathcal{R}_j$ não é ortogonal à reta que une as médias
– hiperplano intercepta esta reta em $\mathbf{x}_o$
– se $P(\omega_i) = P(\omega_j)$, $\forall i, j$ então está no meio das médias
– se $P(\omega_i) \neq P(\omega_j)$, então hiperplano se afasta da média mais provável


0.2
-0.1
0
ω_2
ω_1
$P(\omega_2)$=.5
$P(\omega_1)$=.5
$\mathcal{R}_1$
$\mathcal{R}_2$
-5
0
5
-5
0
5


0.2
-0.1
0
ω_2
ω_1
$P(\omega_2)$=.9
$P(\omega_1)$=.1
$\mathcal{R}_1$
$\mathcal{R}_2$
-5
0
5
-5
0
5


$\mathcal{R}_1$
$P(\omega_1)=.5$
ω_1
$\mathcal{R}_2$
ω_2
$P(\omega_2)=.5$
7.5
5
2.5
0
-2.5
-2
0
2
-2
0
2
4


$\mathcal{R}_1$
$P(\omega_1)=.1$
ω_1
$\mathcal{R}_2$
ω_2
$P(\omega_2)=.9$
10
7.5
5
-2.5
0
-2
0
2
-2
0
2
4

- Caso 3: $\sum_i$ = arbitrária

único termo independente de *i*

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

expandindo

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

Classificador quadrático

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} \qquad \mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$$

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

– caso com duas classes

– superfícies de decisão são hiperquadráticas

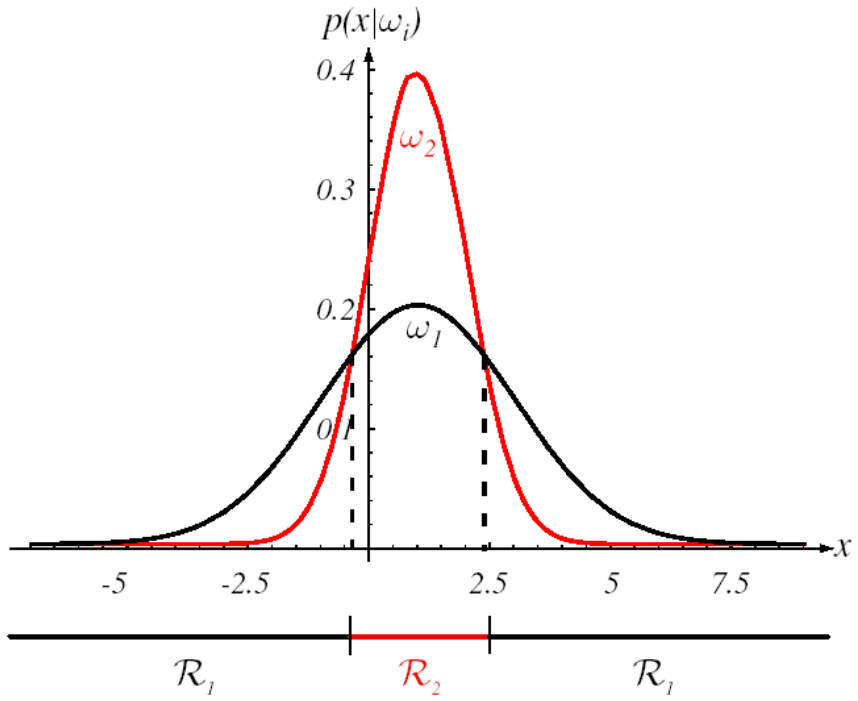
  • hiperplanos

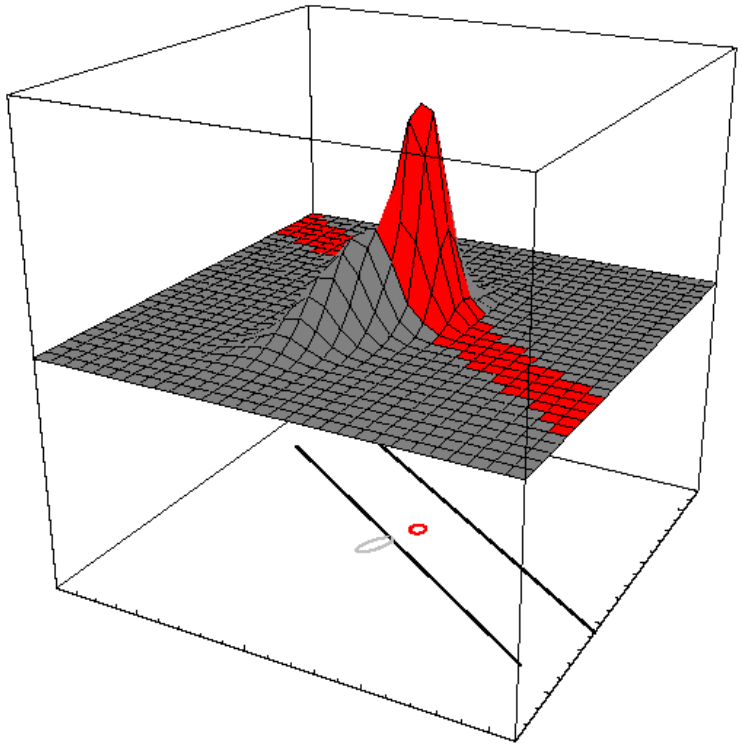
  • hiperesferas

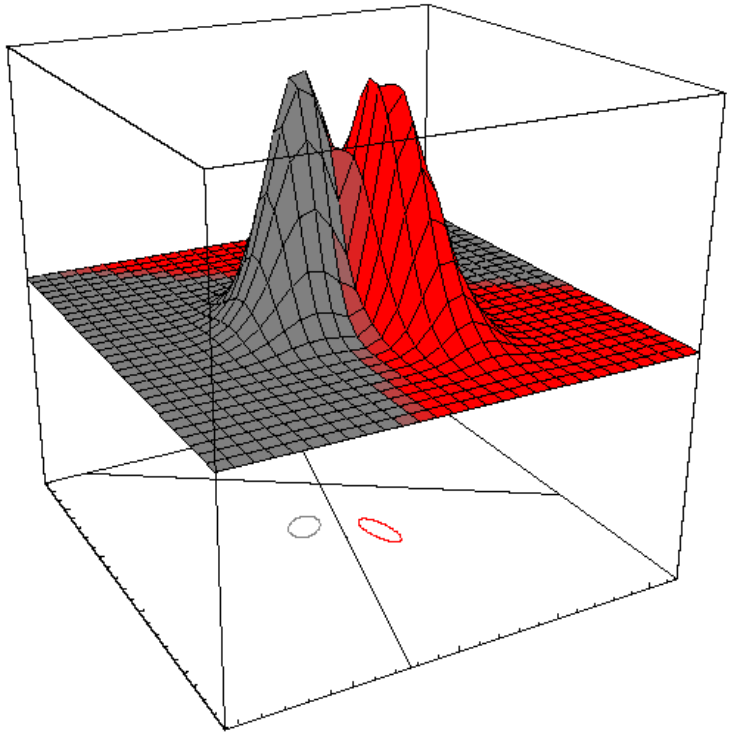
  • hiperelipsóides

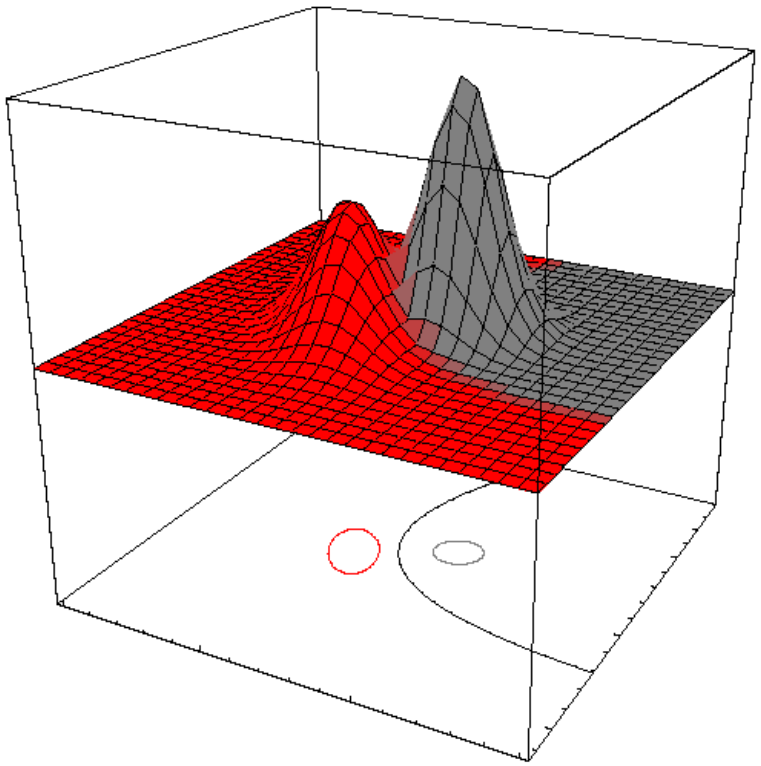
  • hiperparabolóides

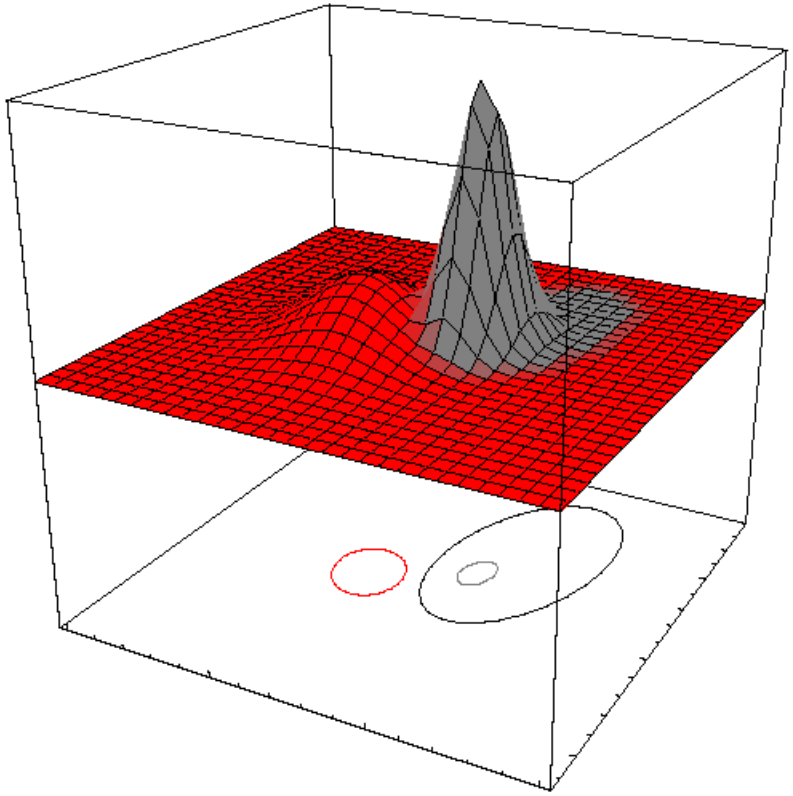
– regiões (decisão) não necessariamente conectadas

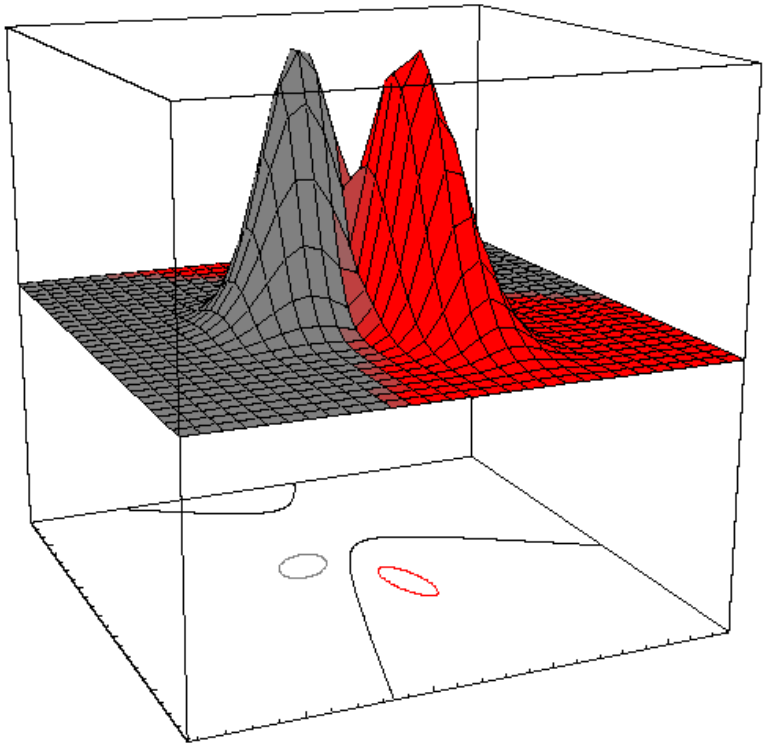


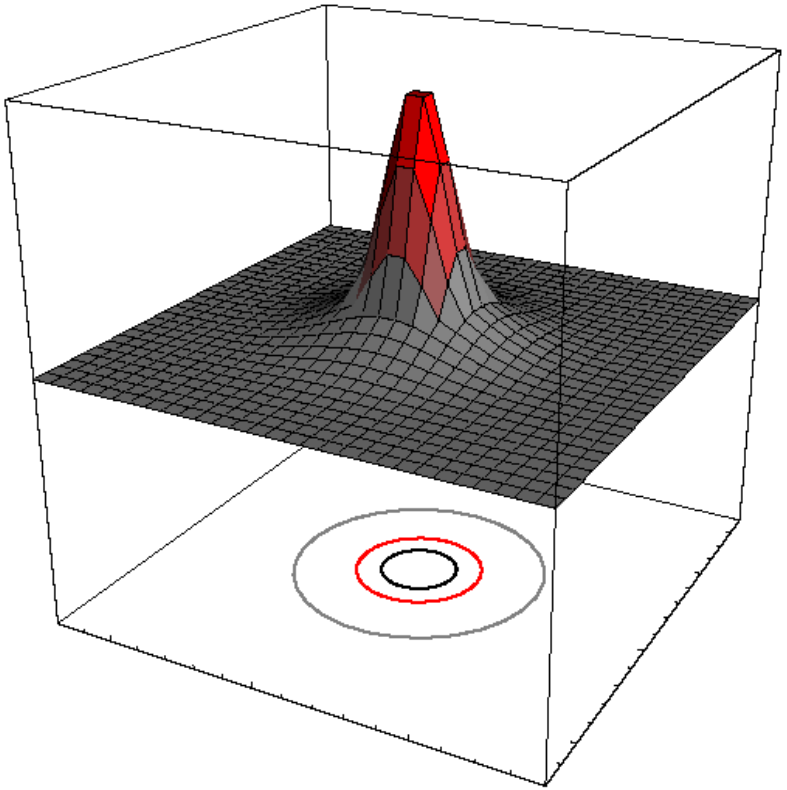


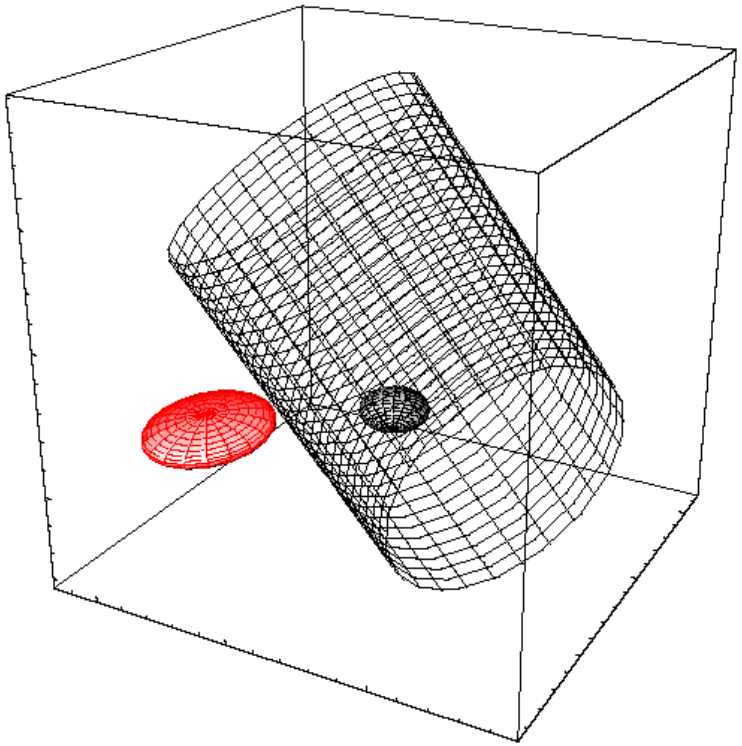


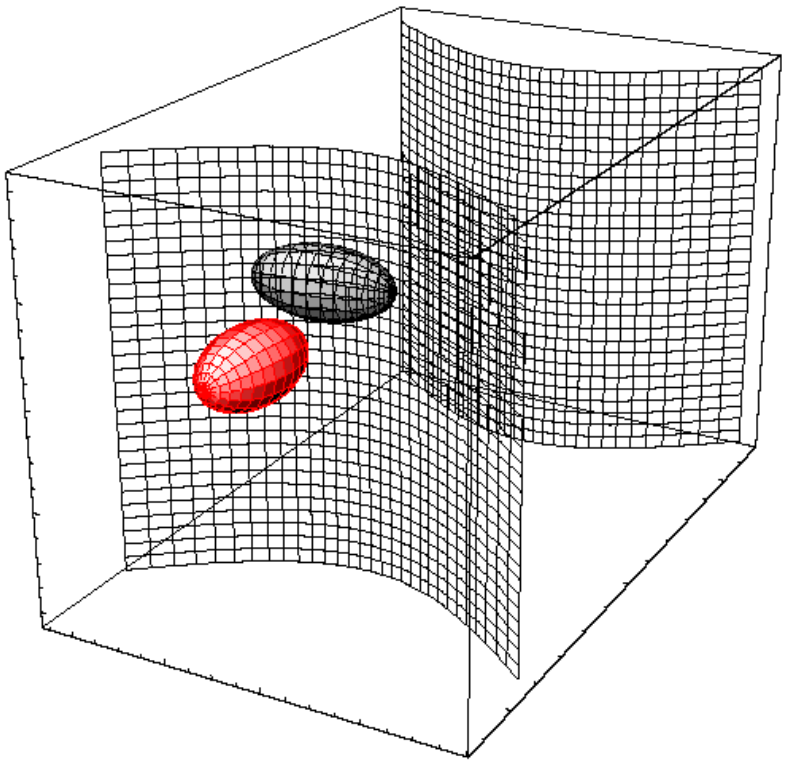


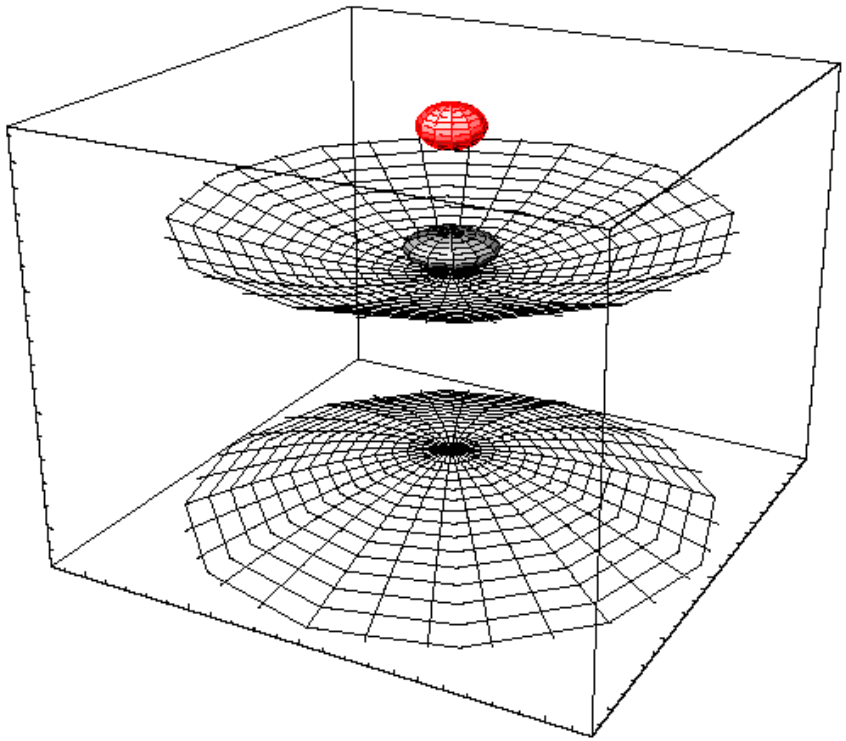


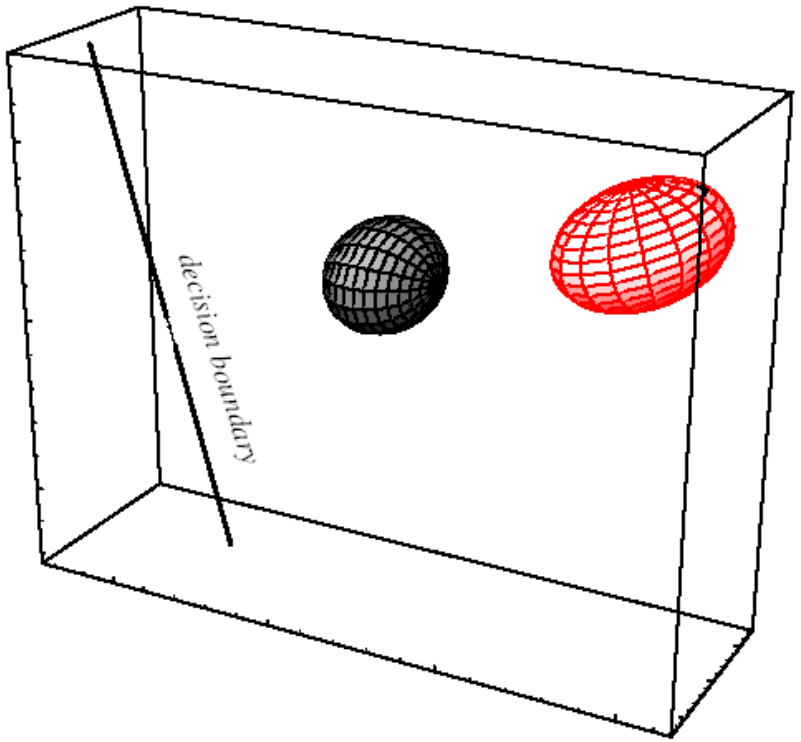




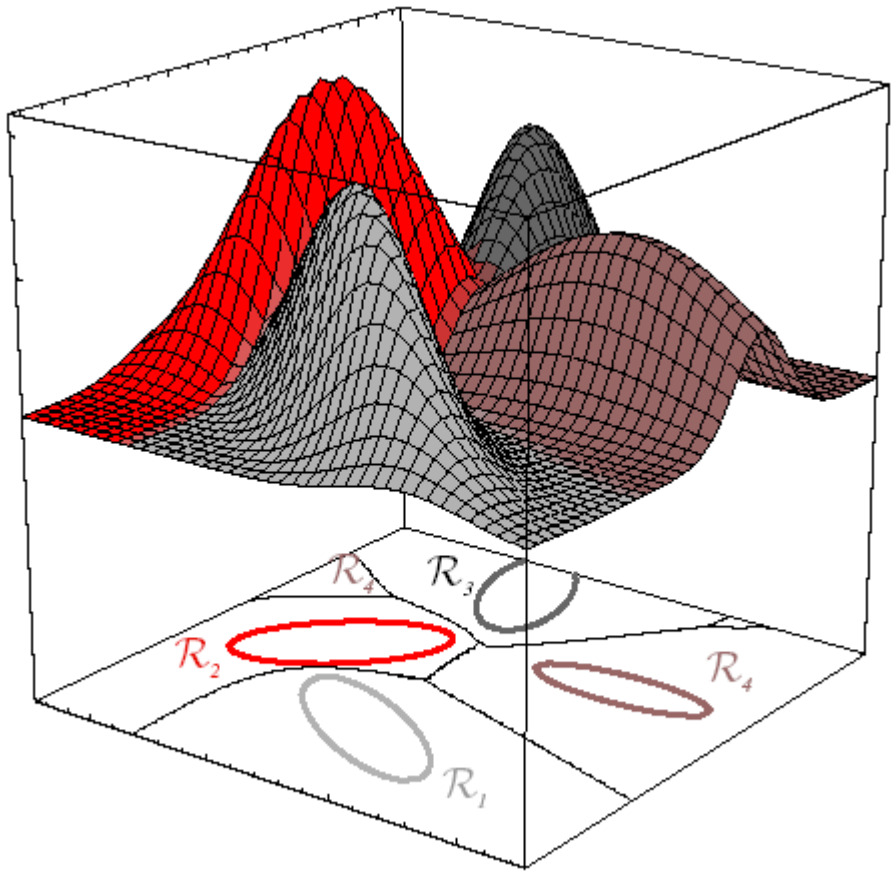




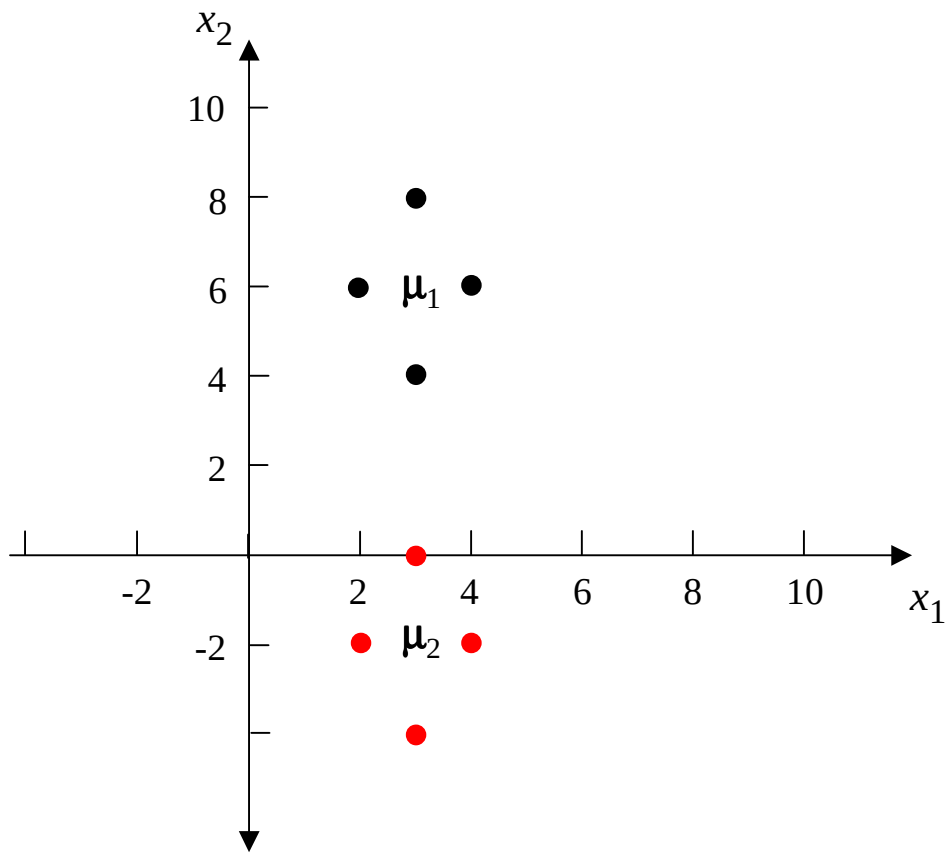



decision boundary

Quatro classes



- Exemplo: região de decisão, dados Gaussianos



$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix} \quad \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1/2 & 0 \\ 0 & 2 \end{bmatrix}$$

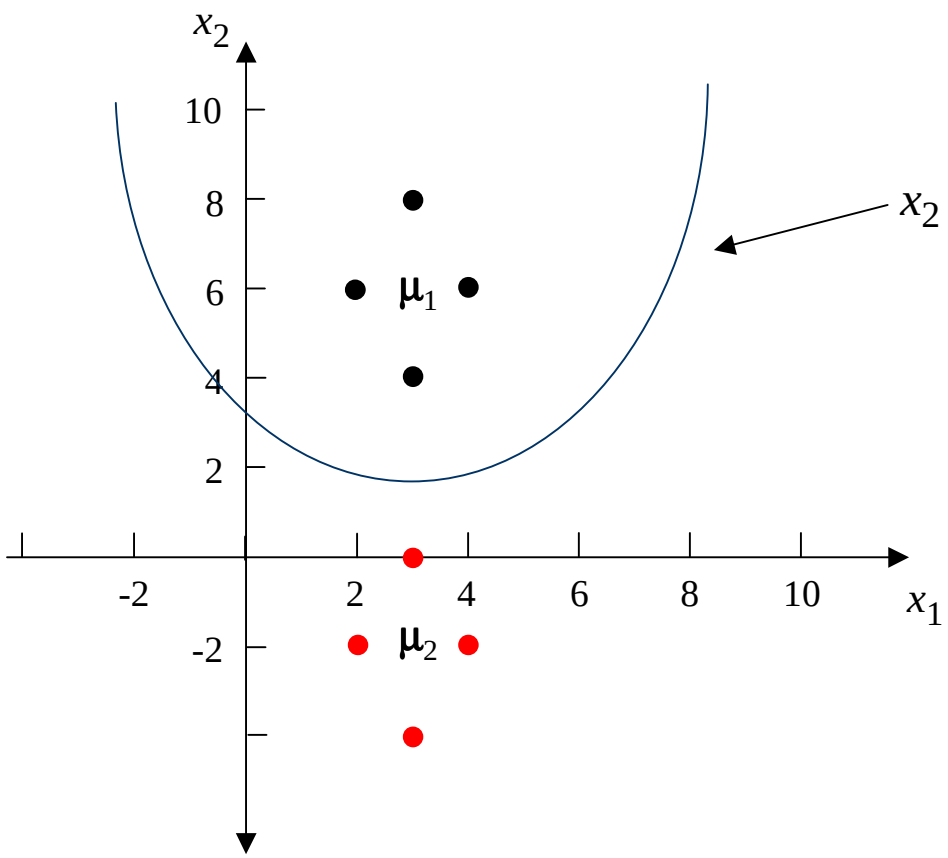
$$\boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

$$P(\omega_1) = P(\omega_2) = 0.5$$

$$g_1(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_1 \mathbf{x} + \mathbf{w}_1^t \mathbf{x} + w_{10}$$

$$g_1(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_1 \mathbf{x} + \mathbf{w}_1^t \mathbf{x} + w_{10}$$

$$g_1(\mathbf{x}) = g_2(\mathbf{x})$$



$$x_2 = 3.514 - 1.125x_1 + 0.1875x_1^2$$

7-Teoria Bayesiana de decisão: atributos discretos

$\mathbf{x} \in \{\mathbf{v}_1, ..., \mathbf{v}_m\}$

$$P(\omega_j \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \omega_j) P(\omega_j)}{P(\mathbf{x})}$$

Regra de Bayes

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

$$\alpha^* = \arg\min_i R(\alpha_i \mid \mathbf{x})$$

Regra de decisão mínimo risco

Regra de Bayes para taxa de erro mínima

- decidir ω_i se $P(\omega_i \mid \mathbf{x}) > P(\omega_j \mid \mathbf{x}) \quad \forall\, i, j$ (ver *)

Funções de discriminação

$$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid \omega_i) P(\omega_i)}{\sum_{j=1}^{c} P(\mathbf{x} \mid \omega_j) P(\omega_j)}$$

$$g_i(\mathbf{x}) = P(\mathbf{x} \mid \omega_i) P(\omega_i)$$

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

- Exemplo: atributos binários independentes
  - duas categorias (classes)
  - cada componente é um valor binário
  - componentes são independentes

$$\mathbf{x} = (x_1, \ldots, x_d)^t, \quad x_i \in \{0,1\}$$

$$p_i = \Pr[x_i = 1 | \omega_1)$$

$$q_i = \Pr[x_i = 1 | \omega_2)$$

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

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

– Razão de verosimilhança

$$\frac{P(\mathbf{x}\,|\,\omega_1)}{P(\mathbf{x}\,|\,\omega_2)} = \prod_{i=1}^{d}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$$

$$g(\mathbf{x}) = \ln\frac{P(\mathbf{x}\,|\,\omega_1)}{P(\mathbf{x}\,|\,\omega_2)} + \ln\frac{P(\omega_1)}{P(\omega_2)} \qquad (**)$$

$$g(\mathbf{x}) = \sum_{i=1}^{d}\left[\, x_i \ln\frac{p_i}{q_i} + (1-x_i)\ln\frac{1-p_i}{1-q_i}\,\right] + \ln\frac{P(\omega_1)}{P(\omega_2)} \qquad \text{Linear em } x_i$$

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

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

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

lembrar: decidir ω_1 se $g(\mathbf{x}) > 0$ e ω_2 se $\mathbf{g}(\mathbf{x}) \geq 0$ (***)

- Análise

– $g(\mathbf{x})$ é uma combinação linear (ponderada) das componentes de $\mathbf{x}$

– valor de w_i é a relevância de uma resposta $x_i = 1$ na classificação

– se $p_i = q_i$ então valor de x_i é irrelevante ($w_i = 0$)

– se $p_i > q_i$ então $(1 - p_i) < (1 - q_i)$, $w_i > 0$ e $x_i = 1 \Rightarrow w_i$ votos para ω_1

– se $p_i < q_i$ então $(1 - p_i) > (1 - q_i)$, $w_i < 0$ e $x_i = 1 \Rightarrow |w_i|$votos para ω_2

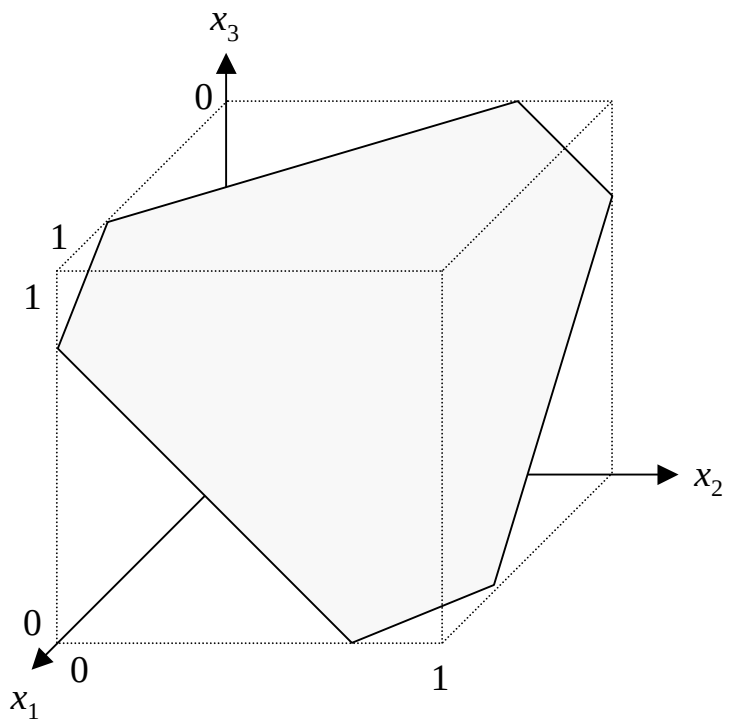
- Exemplo: dados binários 3-d
  - duas categorias (classes)
  - cada componente é um valor binário
  - componentes são independentes

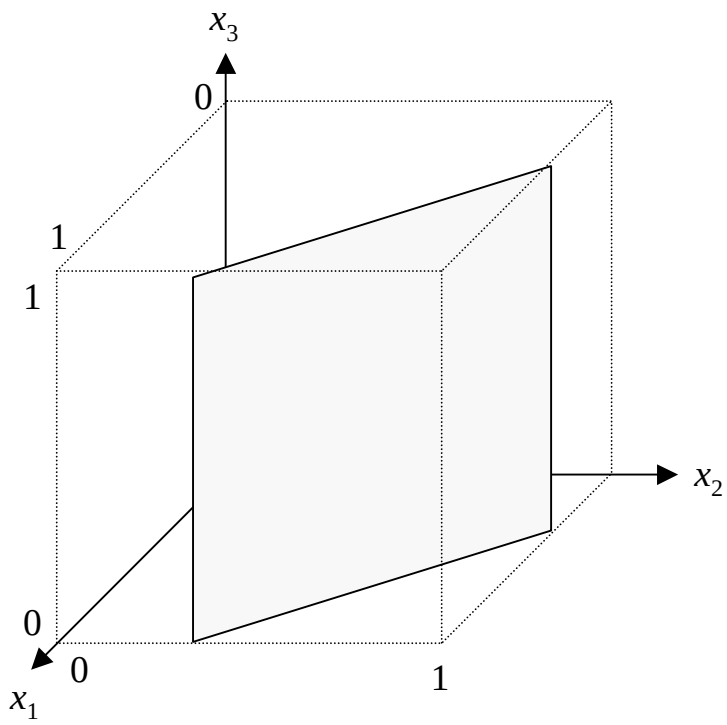
$$P(\omega_1) = P(\omega_2) = 0.5$$

$$p_i = 0.5,\ q_i = 0.8,\ i = 1,2,3$$

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)}, \quad i = 1,\ldots,d \qquad w_i = \ln \frac{0.8(1-0.5)}{0.5(1-0.8)} = 1.3863 \quad i = 1,\ldots,3$$

$$w_0 = \sum_{i=1}^{d} \ln \frac{(1-q_i)}{(1-p_i)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \qquad w_0 = \sum_{i=1}^{3} \ln \frac{(1-0.8)}{(1-0.5)} + \ln \frac{0.5}{0.5} = -1.75$$



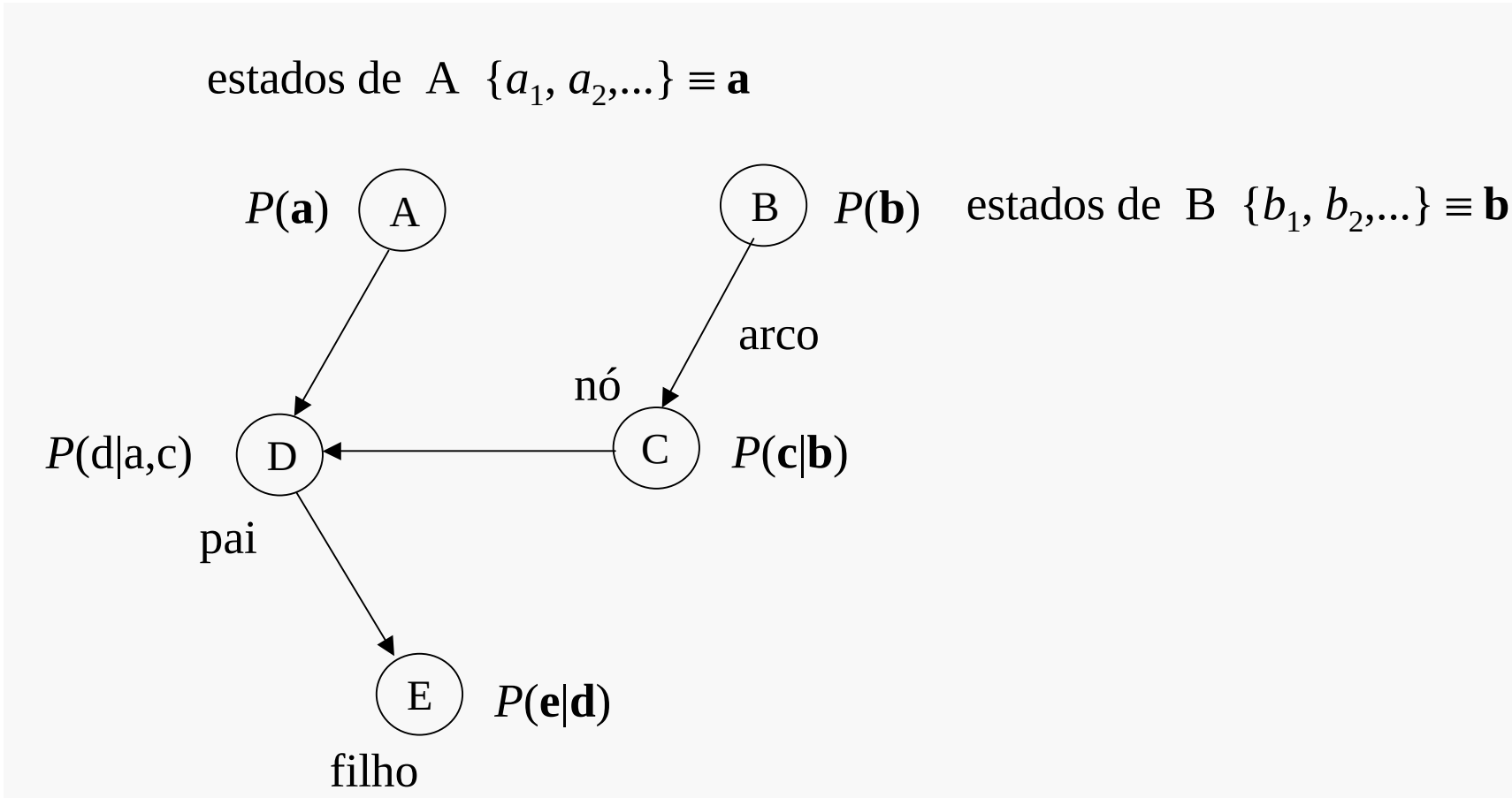


$p_i = 0.8 \quad q_i = 0.5$
$i = 1,2,3$

$p_i = 0.8 \quad q_i = 0.5,\ i = 1,2$
$p_3 = q_3$

8-Redes Bayesianas

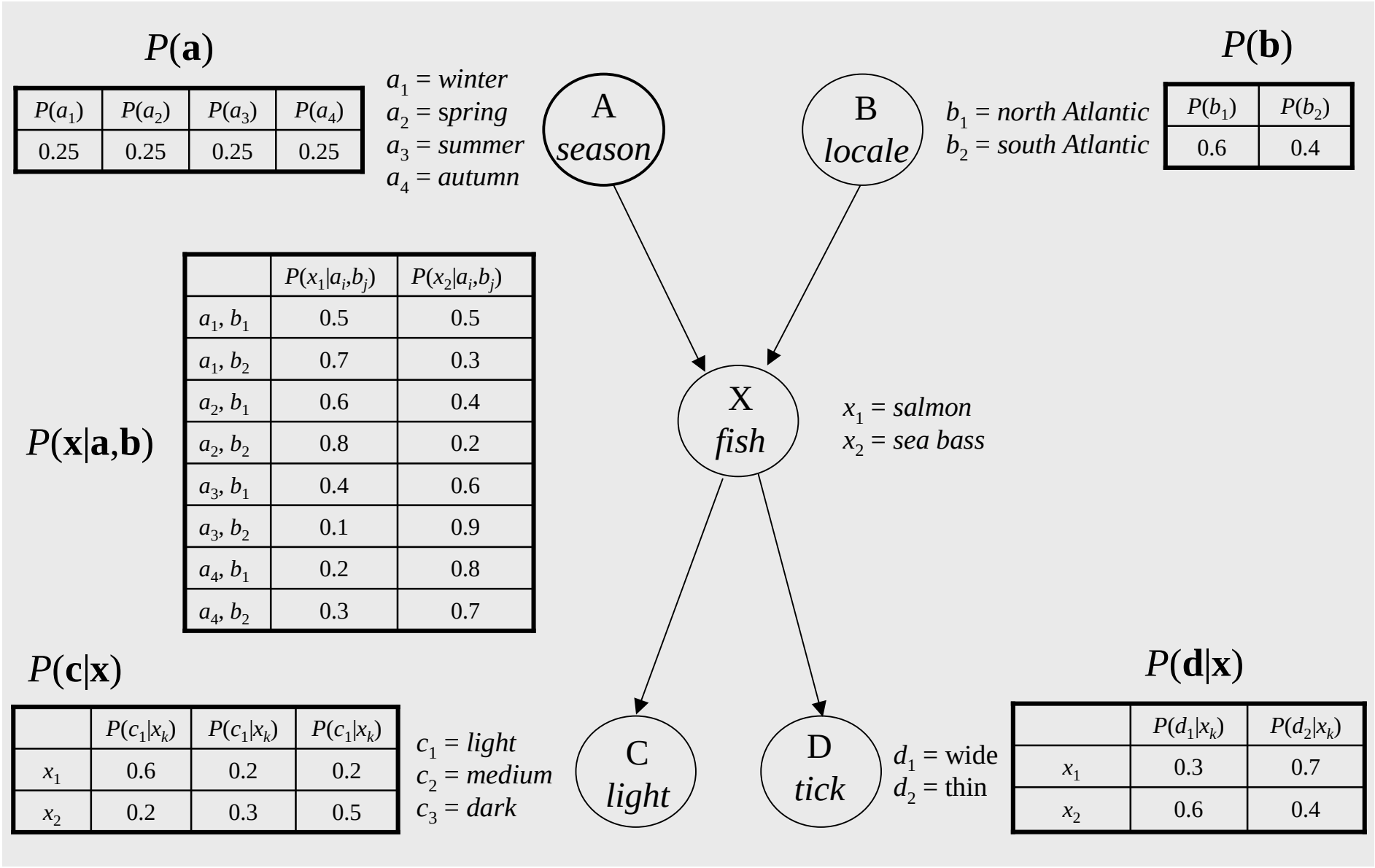
- Conhecimento sobre distribuições
  - parâmetros de distribuições
  - dependência/independência estatística
  - relações causais entre variáveis
- Redes Bayesianas
  - explora informação estrutural no raciocínio com variáveis
  - usa relações probabilísticas entre variáveis
  - assume relações causais



Regra de Bayes + probabilidades condicionais

$\Downarrow$

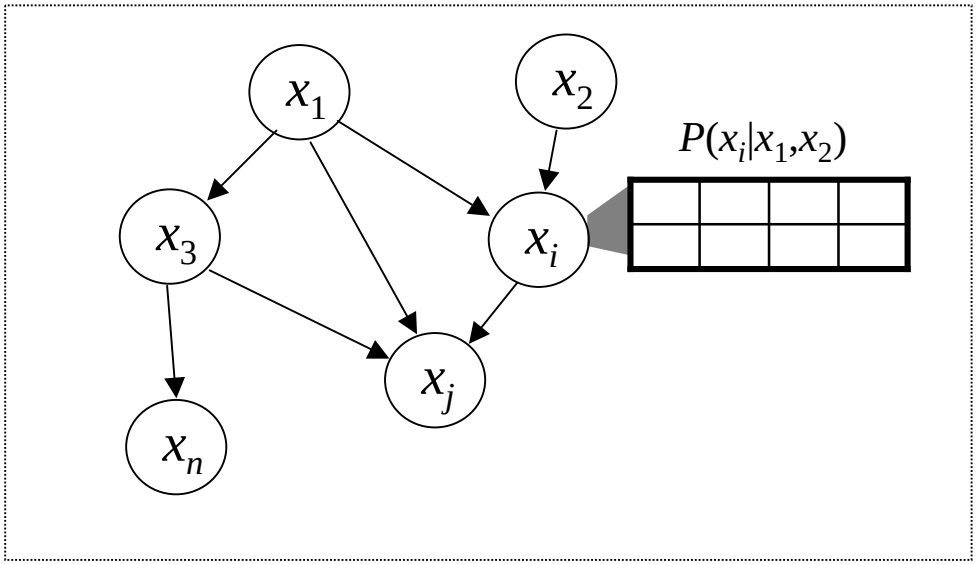
probabilidade conjunta de qualquer configuração de variáveis



$$P(a_3, b_1, x_2, c_3, d_2) = P(a_3)P(b_1)P(x_2 \mid a_3, b_1)P(c_3 \mid x_2)P(d_2 \mid x_2) = 0.25 \times 0.6 \times 0.4 \times 0.5 \times 0.4 = 0.012$$

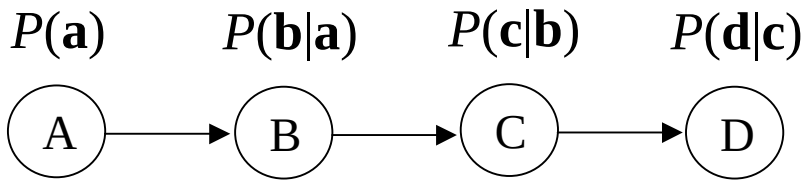
- Redes Bayesianas formalmente

  – grafo acíclico
  – nó: variável aleatória (atributo)
  – arco: efeito, causa (A afeta B → B condicionado a A)
  – cada nó condicionalmente independente dos não descendentes
  – representa probabilidade conjunta das variáveis



$$P(x_1,\ldots,x_n)=\prod_{i=1}^{n}P(x_i \mid PaiDe(x_i))$$

- Exemplo



$$P(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) = P(\mathbf{a})P(\mathbf{b}\,|\,\mathbf{a})P(\mathbf{c}\,|\,\mathbf{b})P(\mathbf{d}\,|\,\mathbf{c})$$

$$P(\mathbf{d}) = \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$$

$$= \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a})P(\mathbf{b}\,|\,\mathbf{a})P(\mathbf{c}\,|\,\mathbf{b})P(\mathbf{d}\,|\,\mathbf{c})$$

$$= \sum_{\mathbf{c}} P(\mathbf{d}\,|\,\mathbf{c}) \sum_{\mathbf{b}} P(\mathbf{c}\,|\,\mathbf{b}) \underbrace{\sum_{\mathbf{a}} P(\mathbf{b}\,|\,\mathbf{a})P(\mathbf{a})}_{P(\mathbf{b})}$$

$$= \sum_{\mathbf{c}} P(\mathbf{d}\,|\,\mathbf{c}) \underbrace{\sum_{\mathbf{b}} P(\mathbf{c}\,|\,\mathbf{b})P(\mathbf{b})}_{P(\mathbf{c})}$$

$$= \underbrace{\sum_{\mathbf{c}} P(\mathbf{d}\,|\,\mathbf{c})P(\mathbf{c})}_{P(\mathbf{d})}$$

– em geral: dados os valores de algumas variáveis (evidência: **e**)

qual é o valor de uma configuração das outras variáveis (**x**) ?

$$P(\mathbf{x} \mid \mathbf{e}) = \frac{P(\mathbf{x}, \mathbf{e})}{P(\mathbf{e})} = \alpha P(\mathbf{x}, \mathbf{e})$$

- Exemplo: *salmon and sea bass*

– probabilidade peixe veio do Atlântico Norte (b_1)
– sabendo que é primavera (*spring* a_2)
– peixe é salmão (*salmon* x_1) claro (*light* c_1)

$P(b_1 \mid a_2, x_1, c_1)$?

- Em classificação (erro mínimo): *salmon or sea bass* ?

  – sabe-se que
    - peixe é claro (c_1)
    - origem é Atlântico Norte (b_2)

  – não se sabe:
    - estação do ano
    - espessura

  – problema de classificação:

$$P(x_1 \mid c_1, b_2) \text{ ?}$$

$$P(x_2 \mid c_1, b_2) \text{ ?}$$

$$P(x_1 \mid c_1, b_2) = \frac{P(x_1, c_1, b_2)}{P(c_1, b_2)} = \alpha \sum_{\mathbf{a},\mathbf{d}} P(x_1, \mathbf{a}, b_2, c_1, \mathbf{d})$$

$$= \alpha \sum_{\mathbf{a},\mathbf{d}} P(\mathbf{a}) P(b_2) P(x_1 \mid \mathbf{a}, b_2) P(c_1 \mid x_1) P(\mathbf{d} \mid x_1)$$

$$= \alpha P(b_2) P(c_1 \mid x_1) \left[\sum_{\mathbf{a}} P(\mathbf{a}) P(x_1 \mid \mathbf{a}, b_2) \right] \left[\sum_{\mathbf{d}} P(\mathbf{d} \mid x_1) \right]$$

$$= \alpha P(b_2) P(c_1 \mid x_1)$$
$$\times [P(a_1) P(x_1 \mid a_1, b_2) + P(a_2) P(x_1 \mid a_2, b_2)$$
$$+ P(a_3) P(x_1 \mid a_3, b_2) + P(a_4) P(x_1 \mid a_4, b_2)]$$
$$\times \underbrace{[P(d_1 \mid x_1) + P(d_2 \mid x_1)]}_{= 1}$$

$$P(x_1 \mid c_1, b_2) = \alpha(0.4)(0.6)[(0.25)(0.7) + (025)(0.8) + (0.25)(0.1) + (0.25)(0.3)]1.0$$

$$P(x_1 \mid c_1, b_2) = \alpha\, 0.114$$

$$P(x_2 \mid c_1, b_2) = \alpha\, 0.066$$

classificação: *salmon* !

- *Naive* Bayes

  – relações de dependência entre atributos desconhecidas

  – neste caso assume-se independência condicional

$$P(\mathbf{x}\,|\,\mathbf{a},\mathbf{b}) = P(\mathbf{x}\,|\,\mathbf{a})P(\mathbf{x}\,|\,\mathbf{b})$$

9-Resumo

- Teoria Bayesiana de decisão é simples
- Regras de decisão
  - minimizar risco: ação que minimiza risco condicional
  - minimizar Pr[erro]: estado que maximiza densidade *a posteriori* $P(\omega_j|x)$
- Superfícies decisão hiperquadráticas no caso Gaussiano
- Redes: relações dependência/independência entre variáveis

Observação

Este material refere-se às notas de aula do curso CT 720 Tópicos Especiais em Aprendizagem de Máquina e Classificação de Padrões da Faculdade de Engenharia Elétrica e de Computação da Unicamp e do Centro Federal de Educação Tecnológica do Estado de Minas Gerais. Não substitui o livro texto, as referências recomendadas e nem as aulas expositivas. Este material não pode ser reproduzido sem autorização prévia dos autores. Quando autorizado, seu uso é exclusivo para atividades de ensino e pesquisa em instituições sem fins lucrativos.