Bayes Optimal Classifier

Renato Assunção

DCC-UFMG

2020

Classificação em duas classes

- Vamos começar com a situação mais simples: duas classes
- Indivíduos são amostrados de uma certa população.
- Esta população é particionada em duas classes disjuntas: pop_1 (denotada π_1) e pop_2 (denotada π_2)
- As duas classes representam uma partição da população:
  - Todo indivíduo pertence a uma das duas subpopulações.
  - Nenhum indivíduo pertence a duas classes ao mesmo tempo.

Exemplos

- Risco de Crédito: Empresas tomadoras de crédito em um banco: $\pi_1 \rightarrow$ créditos bons; $\pi_2 \rightarrow$ créditos ruins
- Crânios em um sítio arqueológico: $\pi_1 \rightarrow$ homens; $\pi_2 \rightarrow$ mulheres
- Saúde: Pessoas com úlcera (π_1) e pessoas sem úlcera (π_2)
- Saúde: Mulheres com (π_1) ou sem (π_2) câncer de mama
- Análise de textos de dois participantes do movimento de independência dos EUA: π_1 : James Madison ou π_2 : Alexander Hamilton.
- Duas espécies da flor Iris: π_1: Iris setosa; π_2: Iris virginica.
- Usuários de um website: π_1: clicam num certo anúncio e π_2, não clicam
- Alunos de um curso online: π_1 : evadem e π_2 completam o curso

Features

- Indivíduos que pertencem a uma de duas sub-populações: π_1 ou π_2.
- Em cada indivíduo, medimos um conjunto de p variáveis (ou features):

$$\mathbf{X} = (X_1, X_2, \ldots, X_p)$$

- Com base nas medições em $\mathbf{X}$, queremos **inferir** se $\mathbf{X} \in \pi_1$ ou se $\mathbf{X} \in \pi_2$.
- Queremos descobrir (ou aprender) uma regra matemática $g(\mathbf{X})$ que prediga se o indivíduo pertence à classe π_1 ou a π_2.
- Esta regra será usada para predizer a classe de novos itens para os quais sabemos $\mathbf{X}$ mas não sabemos a sua classe.

- Para construir uma regra de classificação de novos itens, usamos uma amostra com as classes conhecidas (amostra rotulada com a classe):

Item	Classe ou População	Variáveis X_1 X_2 ... X_p
1	π_1	$X_{1,1}$ $X_{1,2}$ $X_{1,3}$ ... $X_{1,p}$
2	π_1	$X_{2,1}$ $X_{2,2}$ $X_{2,3}$ ... $X_{2,p}$
$\vdots$	$\vdots$	$\vdots$
m_1	π_1	$X_{m_1,1}$ $X_{m_1,2}$ $X_{m_1,3}$ ... $X_{m_1,p}$
1	π_2	$X_{m_1+1,1}$ $X_{m_1+1,2}$ $X_{m_1+1,3}$ ... $X_{m_1+1,p}$
2	π_2	$X_{m_1+2,1}$ $X_{m_1+2,2}$ $X_{m_1+2,3}$ ... $X_{m_1+2,p}$
$\vdots$	$\vdots$	$\vdots$
m_2	π_2	$X_{m_1+m_2,1}$ $X_{m_1+m_2,2}$ $X_{m_1+m_2,3}$ ... $X_{m_1+m_2,p}$
Novo Item	?????	X_1^* X_2^* X_3^* ... X_p^*

- Novo item: conhecemos $\mathbf{X}$ mas não conhecemos a sua classe.
- X_1^* X_2^* X_3^* ... X_p^* $\rightarrow$ valores conhecidos, efetivamente observados.
- ????? $\rightarrow$ queremos inferir a classe do novo item

Exemplos

Populações π_1 e π_2	Variáveis X_1 ... X_p
Risco de Crédito: Empresas tomadoras de crédito em um banco $\pi_1 \rightarrow$ créditos bons $\pi_2 \rightarrow$ créditos ruins	- % do empréstimo frente ao faturamento anual da empresa - tempo como cliente - n^o de empréstimos anteriores pagos a tempo - saldo mensal
Crânios em um sítio arqueológico $\pi_1 \rightarrow$ homens $\pi_2 \rightarrow$ mulheres	- Circunferência - Largura - Altura
Pessoas com úlcera ou sem úlcera	- Medidas de grau de ansiedade - Grau de perfeccionismo - Grau de sentimento de culpa - Grau de dependência
Textos de James Madison ou Alexander Hamilton	- Frequências de palavras distintas e comprimento das sentenças

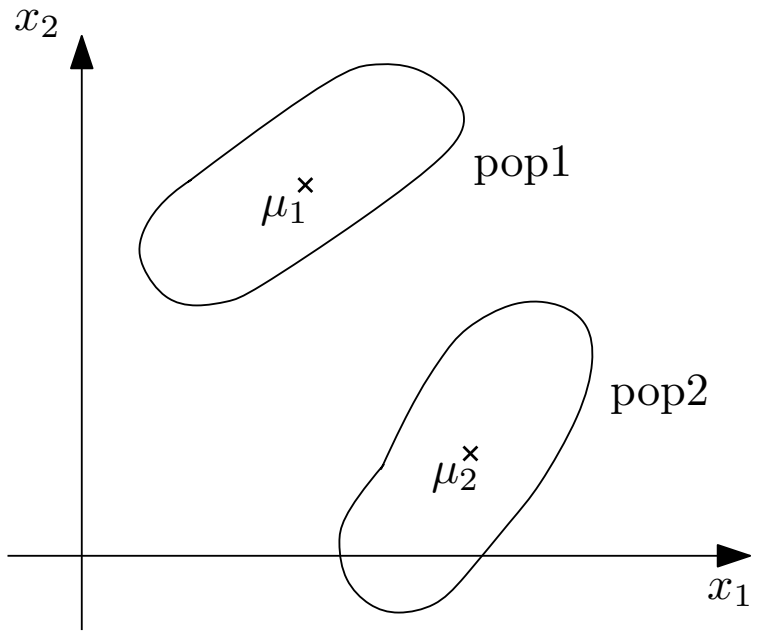
Exemplos

Populações π_1 e π_2	Variáveis X_1 ... X_p
Duas espécies de flor	- Comprimento da pétala - Largura da pétala - Comprimento da sépala - Largura da sépala
Usuários que clicam e não clicam em um anúncio	- Posição do anúncio na página - Tamanho do anúncio - Tem imagem? - Número de palavras
Alunos que evadem e que completam um curso online	- Nota do exame de entrada no curso - Medidas de motivação a partir de questionário na entrada - Renda familiar - Idade

Por que precisamos predizer a classe de um item novo?

- Classe pode ser conhecida apenas no futuro.
  - Ex.: Risco de crédito: No momento em que o crédito é solicitado, não sabemos se o crédito do Indivíduo é bom ou ruim.
- Informação sobre a classe não é conhecida com certeza.
  - Crânios arqueológicos danificados.
- Obter a classe pode implicar em destruir o item.
  - Ex.: Queremos classificar um paciente chegando ao pronto socorro com lesão na cabeça como UTI ou não-UTI, com base em algumas medidas rápidas. Esperar para saber com certeza se deve ir para UTI pode significar esperar demais.

- Cada uma das populações possui uma distribuição conjunta para as p variáveis:
- $\mathbf{X} = (X_1, X_2, ..., X_p)$
- População 1
  $(\mathbf{X}| \in \pi_1) \sim f_1(\mathbf{x})$
- População 2
  $(X| \in \pi_2) \sim f_2(\mathbf{x})$



Por exemplo: $f_1(\mathbf{x}) = N_p(\underset{px1}{\boldsymbol{\mu}_1}, \underset{pxp}{\Sigma_1})$ e $f_2(\mathbf{x}) = N_p(\underset{px1}{\boldsymbol{\mu}_2}, \underset{pxp}{\Sigma_2})$
(mas poderia ser qualquer outra distribuição).

- Com base na amostra rotulada (classes conhecidas), podemos obter estimativas dos valores esperados $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$ das distribuições $f_1(\mathbf{x})$ e de $f_2(\mathbf{x})$ simplesmente tomando a mé dia aritmética de cada variável dentro de cada classe.

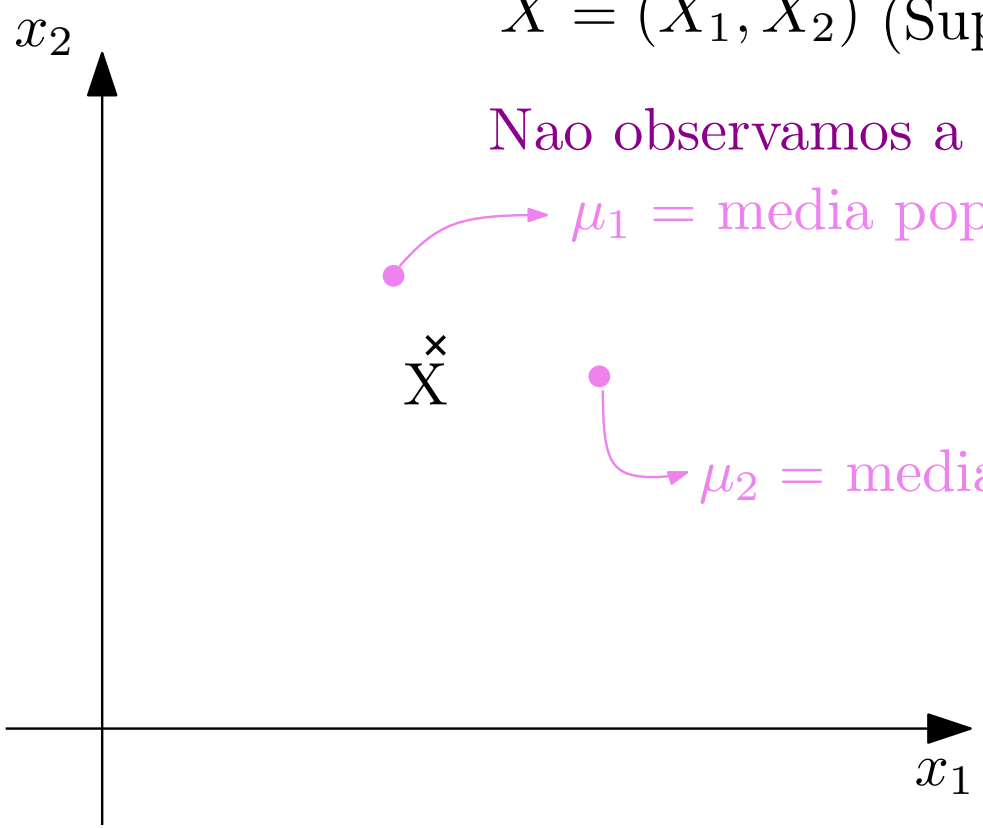
Item	Classe ou População	Variáveis X_1 X_2 ... X_p
1	π_1	$X_{1,1}$ $X_{1,2}$ $X_{1,3}$ ... $X_{1,p}$
2	π_1	$X_{2,1}$ $X_{2,2}$ $X_{2,3}$ ... $X_{2,p}$
$\vdots$	$\vdots$	$\vdots$
m_1	π_1	$X_{m_1,1}$ $X_{m_1,2}$ $X_{m_1,3}$ ... $X_{m_1,p}$
	Médias das p vars	$\bar{x}_{11}$ $\bar{x}_{12}$ $\bar{x}_{13}$ ... $\rightarrow$ vetor $\bar{\mathbf{x}}_1 = \hat{\mu}_1$
1	π_2	$X_{m_1+1,1}$ $X_{m_1+1,2}$ $X_{m_1+1,3}$ ... $X_{m_1+1,p}$
2	π_2	$X_{m_1+2,1}$ $X_{m_1+2,2}$ $X_{m_1+2,3}$ ... $X_{m_1+2,p}$
$\vdots$	$\vdots$	$\vdots$
m_2	π_2	$X_{m_1+m_2,1}$ $X_{m_1+m_2,2}$ $X_{m_1+m_2,3}$ ... $X_{m_1+m_2,p}$
	Médias das p vars	$\bar{x}_{21}$ $\bar{x}_{22}$ $\bar{x}_{23}$ ... $\rightarrow$ vetor $\bar{\mathbf{x}}_2 = \hat{\mu}_2$

- Podemos também estimar as matrizes $p \times p$ de covariância $\mathbf{\Sigma}_1$ e $\mathbf{\Sigma}_2$ com as amostras rotuladas.
- Por exemplo, para a classe 1, estimamos as p variâncias $\sigma^2_{11}, \sigma^2_{1,2}, \ldots, \sigma^2_{1,p}$ através das variâncias amostrais $s^2_{11}, s^2_{1,2}, \ldots, s^2_{1,p}$
- A covariância c_{12} entre a variável i e a variável j (da classe 1) é estimada por $s_{1i} s_{1j} r_{1,ij}$ usando os desvios-padrão de cada variável e a correlação $r_{1,ij}$

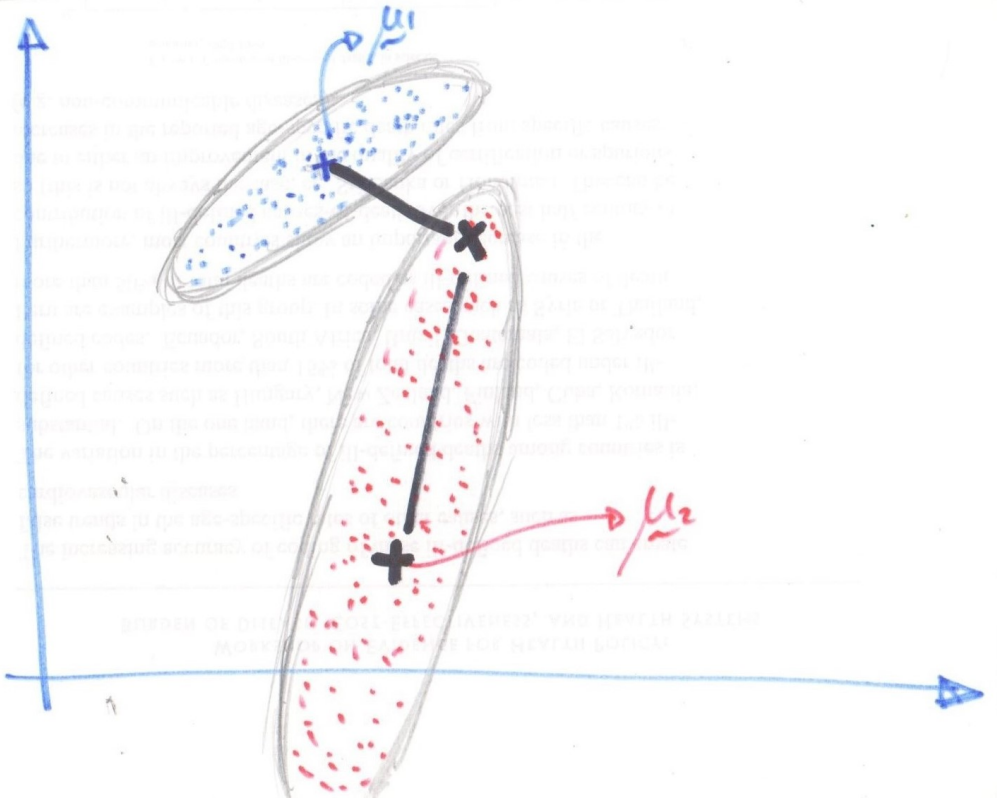
Item	Classe ou População	Variáveis X_1 X_2 ... X_p
1	π_1	$X_{1,1}$ $X_{1,2}$ $X_{1,3}$ ... $X_{1,p}$
2	π_1	$X_{2,1}$ $X_{2,2}$ $X_{2,3}$ ... $X_{2,p}$
$\vdots$	$\vdots$	$\vdots$
m_1	π_1	$X_{m_1,1}$ $X_{m_1,2}$ $X_{m_1,3}$ ... $X_{m_1,p}$
Variância amostral das p vars		$s^2_{11}, s^2_{1,2}, \ldots, s^2_{1,p}$

- Novo item

$$X = (X_1, X_2) \text{ (Supondo apenas } p = 2 \text{ variaveis)}$$



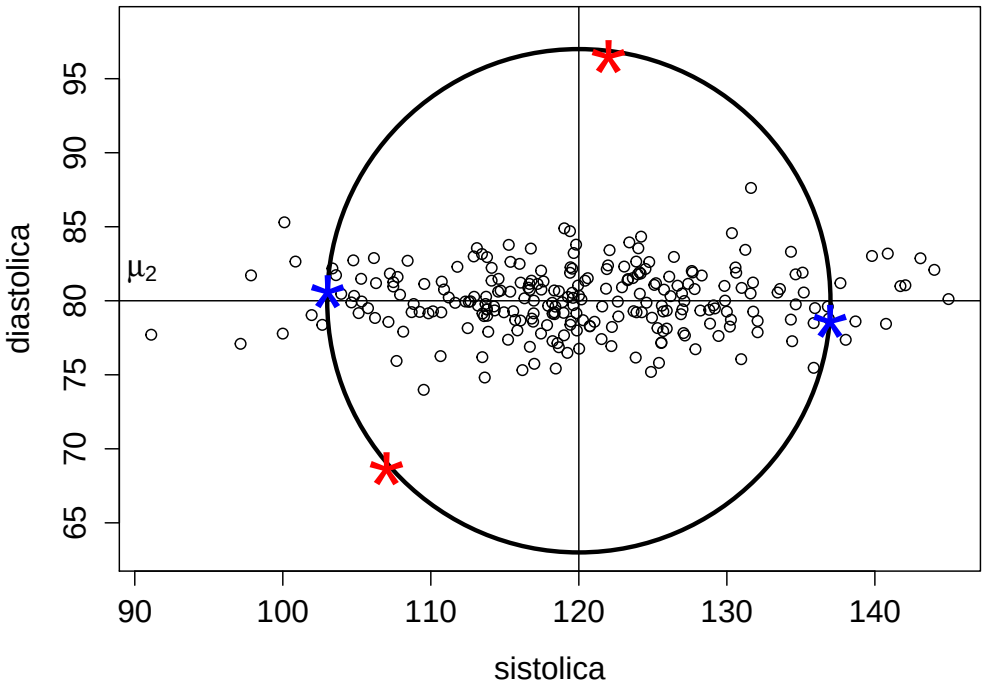
- Olhar a distância do novo item **X** aos vetores $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2 \Rightarrow$ parece razoável alocar **X** à população 1, pois a distância entre **X** e $\boldsymbol{\mu}_1$ é menor que entre **X** e $\boldsymbol{\mu}_2$.



- Distância Euclidiana de **X** à μ_1 é menor que sua distância a μ_2.

- No entanto, **X** parece pertencer à população 2!
- Precisamos levar em conta as correlações.
- Precisamos olhar a distância estatística ou a distância de Mahalanobis do novo item **X** a cada um dos centros μ_1 e μ_2.

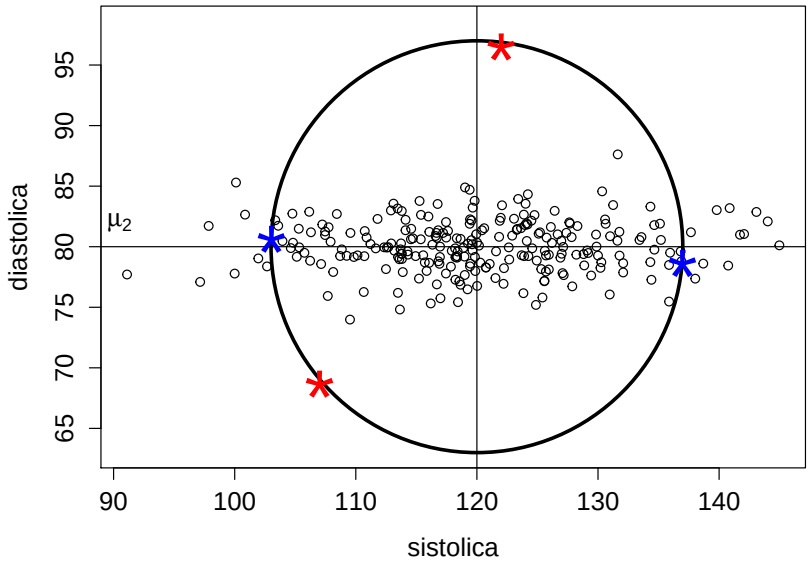
Distância Estatística



- Distância Euclidiana não é a melhor maneira de medir distâncias entre vetores aleatórios.
- Os pontos coloridos de vermelho e azul abaixo possuem a mesma distância euclidiana ao centro da nuvem de pontos estatísticos.

- Centro da nuvem representa o valor médio de cada variável, o perfil "médio" desta população estatística.

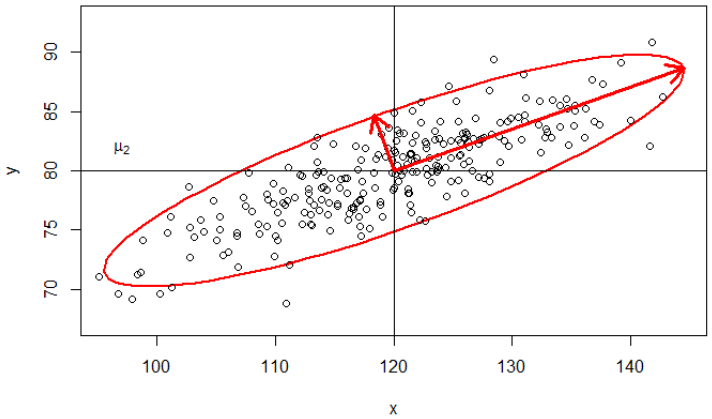
Distância Estatística



- INTUITIVAMENTE, os pontos vermelhos estão ESTATISTICAMENTE mais distantes do centro da nuvem que os pontos azuis.

- Como criar uma medida de distância matemática que incorpore esta intuição?
- Distância de Mahalanobis: leva em conta as variâncias de cada variável.

Distância de Mahalanobis



- Mahalanobis também leva em conta as correlações entre as variáveis.
- Pontos na elipse inclinada estão à mesma distância do centro da nuvem.

- As correlações entre as variáveis $\rightarrow$ inclinação da elipse.
- Obs: eixos das elipses são os componentes principais!!
- Para mais detalhes, ver capítulo 13 do meu livro em: `https://homepages.dcc.ufmg.br/~assuncao/EstatCC/FECD.pdf`

Distância Euclidiana em duas dimensões

- Queremos a distância entre um ponto arbitrário $\mathbf{x} = (x_1, x_2)$ e o ponto $\boldsymbol{\mu} = (\mu_1, \mu_2)$ que representa os valores esperados das variáveis X_1 e X_2.

$$
\begin{aligned}
d^2(\boldsymbol{x}, \boldsymbol{\mu}) &= (x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 \\
&= (x_1 - \mu_1, x_2 - \mu_2) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \\
&= \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix}^t \mathbf{I} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \\
&= (\boldsymbol{x} - \boldsymbol{\mu})^t \, \mathbf{I} \, (\boldsymbol{x} - \boldsymbol{\mu})
\end{aligned}
$$

- Esta forma matricial é uma maneira complicada, um tanto pedante, de escrever a distância.
- Entretanto ela é uma representação muito útil: o caso genérico vai ficar MUITO simples nesta notação matricial.

Distância Euclidiana em p dimensões

- Queremos a distância entre um ponto arbitrário $\mathbf{x} = (x_1, x_2, \ldots, x_p) \in \mathbb{R}^p$ e o ponto $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p) \in \mathbb{R}^p$ que representa os valores esperados das variáveis $X_1, X_2, \ldots, X_p$.
- Em p dimensões:

$$
\begin{aligned}
d^2(\boldsymbol{x}, \boldsymbol{\mu}) &= (x_1 - \mu_1)^2 + (x_2 - \mu_2)^2 + \ldots + (x_p - \mu_p)^2 \\
&= (x_1 - \mu_1, x_2 - \mu_2, \ldots, x_p - \mu_p) \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ & & \ldots & & \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix} \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \ldots \\ x_p - \mu_p \end{pmatrix} \\
&= (\boldsymbol{x} - \boldsymbol{\mu})^t \, \mathbf{I} \, (\boldsymbol{x} - \boldsymbol{\mu})
\end{aligned}
$$

Distância Euclidiana em p dimensões

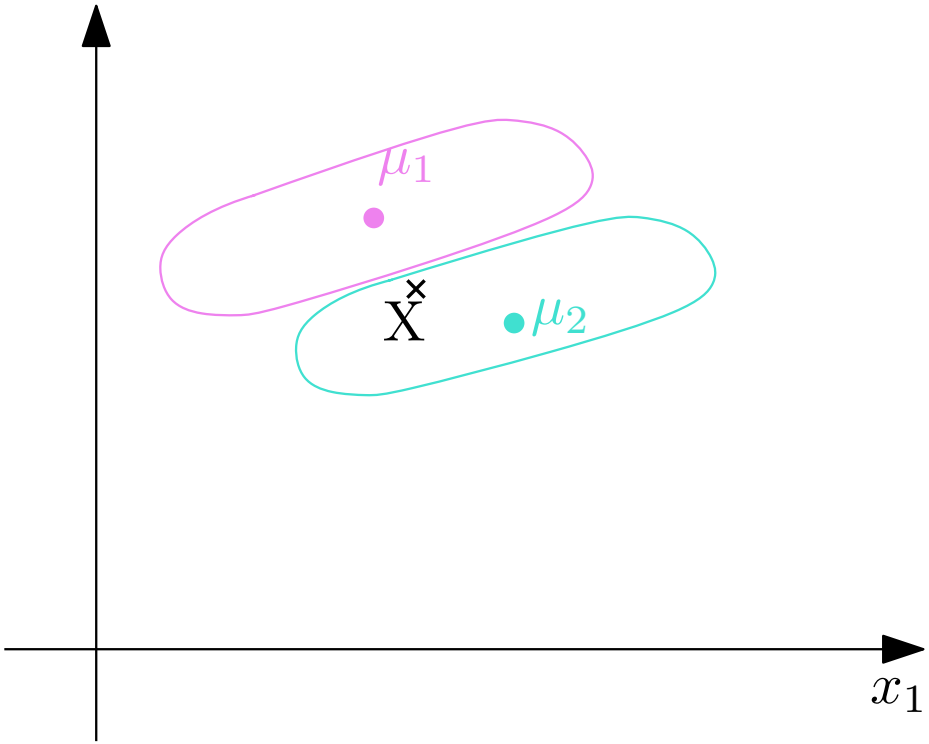
- A representação matricial com p dimensões é a mesma daquela com duas dimensões.
- A distância estatística (de Mahalanobis) substitui a matriz $p \times p$ identidade $\mathbf{I}$ na expressão acima por $\mathbf{\Sigma}^{-1}$, a inversa da matriz de covariância, também de dimensão $p \times p$.
- Para uma explicação intuitiva e bem detalhada, ver o capítulo 13 do meu livro em:
  `https://homepages.dcc.ufmg.br/~assuncao/EstatCC/FECD.pdf`

Mahalanobis

- $\mathbb{E}(X) = \underset{px1}{\boldsymbol{\mu}} =$ vetor com os valores esperados de cada uma das p variáveis
- $\mathbb{V}(X) = \underset{pxp}{\boldsymbol{\Sigma}} =$ matriz de variâncias e covariâncias do vetor $\mathbf{X}$

$$d^2_{\boldsymbol{\Sigma}}(\mathbf{X}, \boldsymbol{\mu}) = (\mathbf{X} - \boldsymbol{\mu})^t \, \boldsymbol{\Sigma}^{-1} \, (\mathbf{X} - \boldsymbol{\mu})$$

$$d^2(X, \mu) = \boxed{(X - \mu)^t} \cdot \boxed{\Sigma^{-1}} \cdot \boxed{\begin{matrix} X \\ - \\ \mu \end{matrix}}$$



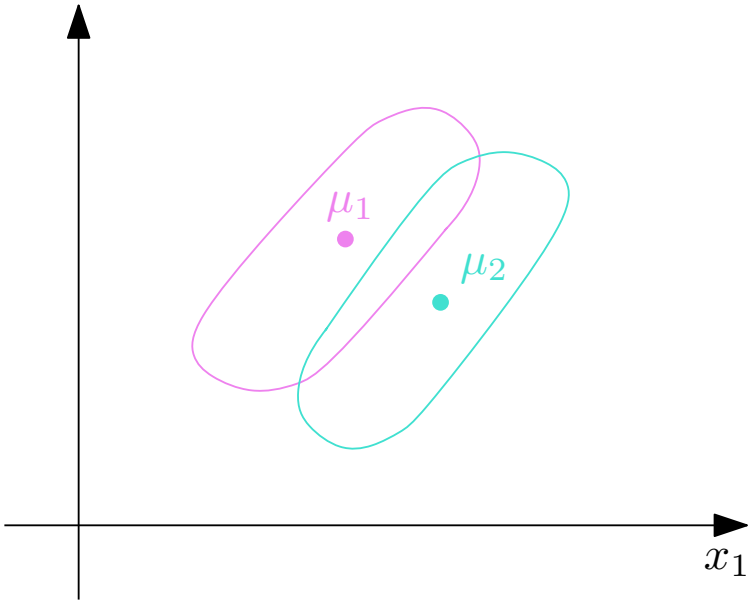
- $d_1^2 = d_{\mathbf{\Sigma}_1}^2(\mathbf{X}, \boldsymbol{\mu}_1)$
  $= (\mathbf{X} - \boldsymbol{\mu}_1)^t \mathbf{\Sigma}_1^{-1} (\mathbf{X} - \boldsymbol{\mu}_1)$
- $d_2^2 = d_{\mathbf{\Sigma}_2}^2(\mathbf{X}, \boldsymbol{\mu}_2)$
  $= (\mathbf{X} - \mu_2)^t \mathbf{\Sigma}_2^{-1} (\mathbf{X} - \boldsymbol{\mu}_2)$

Uma Regra de Classificação Inicial

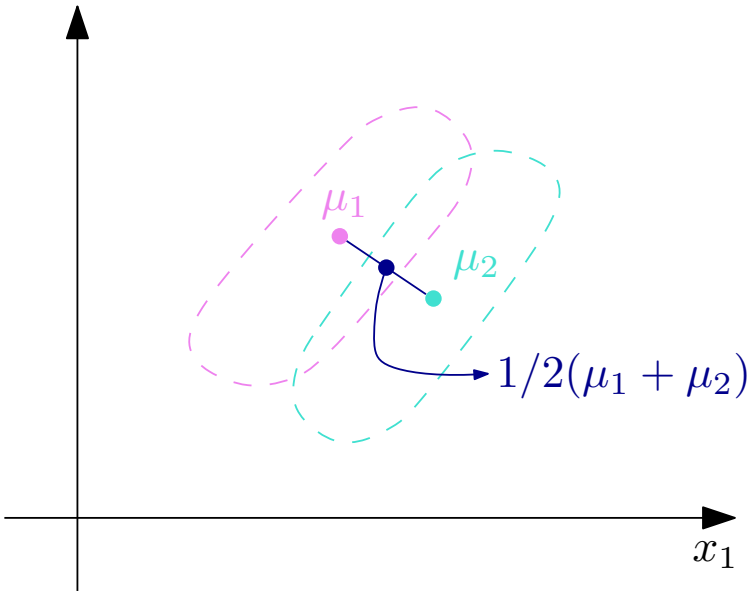
- Aloque $\mathbf{X}$ à população com a menor distância de Mahalanobis d^2.

- Isto é,
  - Se $d^2_{\mathbf{\Sigma_1}}(\mathbf{X}, \boldsymbol{\mu}_1) < d^2_{\mathbf{\Sigma_2}}(\mathbf{X}, \mu_2) \Rightarrow$ aloque $\mathbf{X}$ à pop1;
  - Caso contrário, aloque $\mathbf{X}$ à pop2.

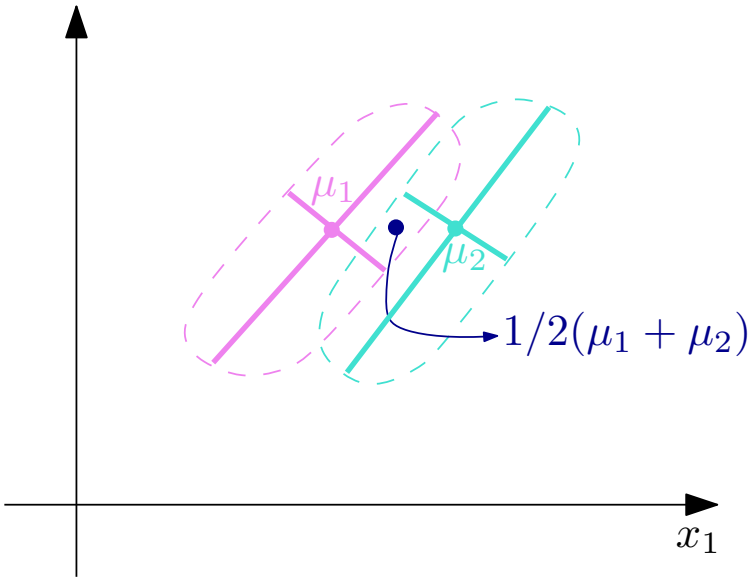
- Espaço $\mathbb{R}^p$ é particionado em duas regiões:
  - $R_1 = \left\{x \in \mathbb{R}^p \mid d^2_{\mathbf{\Sigma_1}}(\mathbf{X}, \boldsymbol{\mu}_1) < d^2_{\mathbf{\Sigma_2}}(\mathbf{X}, \mu_2)\right\}$
  - $R_2 = \mathbb{R}^p - R_1 =$ pontos que serão alocados à pop2.



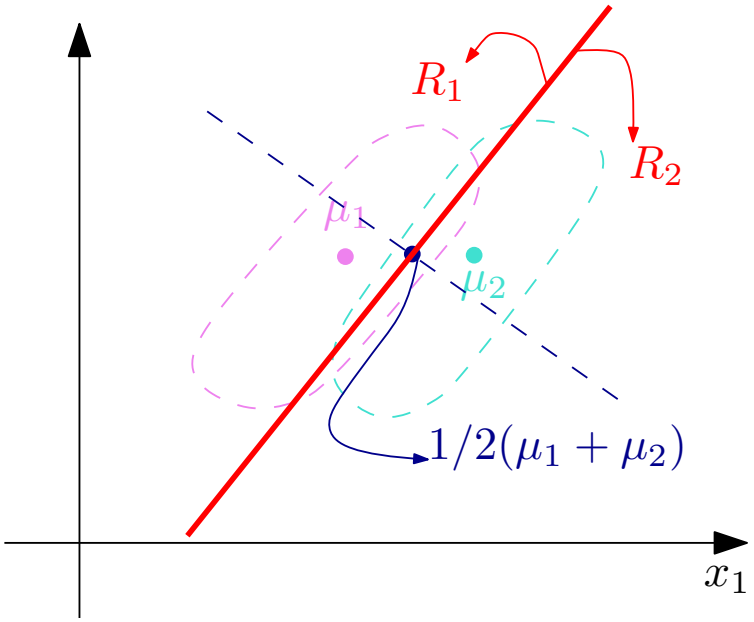
- Quais são essas duas regiões?
- A seguir uma visão intuitiva.
- Resultado mais rigoroso vem a seguir.



- Obtenha o perfil médio das duas populações.



- Considere a estrutura de correlação entre as variáveis de cada grupo (mais a frente, veremos os detalhes disso)



$\Leftarrow$ as duas regiões.

- Outra maneira de ver a regra de classificação:

$$\begin{cases} \pi_1 = pop_1 \\ \pi_2 = pop_2 \end{cases}$$

- $f_1(\mathbf{x}) =$ densidade do vetor $\mathbf{X}$ se $\in \pi_1$
- $f_2(\mathbf{x}) =$ densidade do vetor $\mathbf{X}$ se $\in \pi_2$

- Vamos supor que, dentro de cada classe, os dados $\mathbf{X}$ sigam uma distribuição gaussiana:

$$\mathbf{X} \sim \begin{cases} N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), & se \in \pi_1 \\ N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), & se \in \pi_2 \end{cases}$$

- No caso gaussiano, temos

$$f_1(\mathbf{x}) = \underbrace{\left[\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_1|^{1/2}}\right]}_{c_1 = \text{ constante em } \mathbf{x}} \exp\left(-\frac{1}{2}\underbrace{(\mathbf{x}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}_{\text{distância de Mahalanobis}}\right)$$

$$f_2(\mathbf{x}) = \underbrace{\left[\frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_2|^{1/2}}\right]}_{c_2 = \text{ constante em } \mathbf{x}} \exp\left(-\frac{1}{2}\underbrace{(\mathbf{x}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}_{\text{distância de Mahalanobis}}\right)$$

- Tomando a razão das densidades no mesmo ponto $\mathbf{x}$:

$$\begin{aligned}\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{c_1 \ \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^t\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right)}{c_2 \ \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^t\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right)} \\ &= \exp\left(-\frac{1}{2}(d^2_{\boldsymbol{\Sigma}_1}(\mathbf{x},\boldsymbol{\mu}_1) - d^2_{\boldsymbol{\Sigma}_2}(\mathbf{x},\boldsymbol{\mu}_2)\right)\end{aligned}$$

- Vamos escolher um threshold $M \geq 1$. Veja que:

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > M \iff \underbrace{d^2_{\boldsymbol{\Sigma}_1}(\mathbf{x},\boldsymbol{\mu}_1) < d^2_{\boldsymbol{\Sigma}_2}(\mathbf{x},\boldsymbol{\mu}_2) + K}_{\text{condição para alocar a } \pi_1}$$

onde $K = \log(c_1/(Mc_2))$.

Caso $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$

- Se $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, temos $c_1 = c_2$. Além disso, tomando $M = 1$, a regra fica simplemente

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > 1 \iff \underbrace{d^2_{\mathbf{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) < d^2_{\mathbf{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2)}_{\text{condição para alocar a } \pi_1}$$

- Isto é, no caso gaussiano com covariâncias iguais, o conjunto R_1 dos pontos $\mathbf{x} \in \mathbb{R}^p$ tais que $f_1(\mathbf{x}) > f_2(\mathbf{x})$ é o mesmo conjunto de pontos em que $d^2_{\mathbf{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) < d^2_{\mathbf{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2)$
- No caso geral, com $M = 1$, o conjunto R_1 dos pontos e que $f_1(\mathbf{x}) > f_2(\mathbf{x})$ é o mesmo que pedir a distância de Mahalanobis $d^2_{\mathbf{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1)$ menor que a distância $d^2_{\mathbf{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2)$ acrescida da constante $\log(c_1/c_2)$.

Saindo do caso gaussiano

- Assim, no caso gaussiano, definir a região de classificação à π_1 usando a razão de densidades é matematicamente equivalente a definir usando as distâncias de Mahalanobis.

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > 1 \iff \underbrace{d^2_{\boldsymbol{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) < d^2_{\boldsymbol{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2) + K}_{\text{condição para alocar a } \pi_1}$$

- E quando $\mathbf{X}$ não seguir uma gaussiana?
- O que devemos usar para definir R_1?
- Faz diferença?

Caso não-gaussiano

- Conhecemos a densidade de $\mathbf{X}$ em cada classe: $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$.
- **Método 1:** Imitando o que descobrimos no caso gaussiano, podemos escolher R_1 como sendo o seguinte conjunto de pontos do $\mathbb{R}^p$:

$$R_1 = \{\mathbf{x} \in \mathbb{R}^p \text{ tais que } f_1(\mathbf{x}) > f_2(\mathbf{x})\}$$

- **Método 2:** Outra opção, quando as amostras forem grandes:
  - com as amostras rotuladas, obtemos boas aproximições para os vetores de valores esperados $\boldsymbol{\mu}_1$ e $\boldsymbol{\mu}_2$.
  - Obtemos boas estimativas das matrizes $p \times p$ de covariância $\boldsymbol{\Sigma}_1$ e $\boldsymbol{\Sigma}_2$.
  - Para cada pontos $\mathbf{x} \in \mathbb{R}^p$ podemos calcular as duas distâncias de Mahalanobis: $d^2_{\boldsymbol{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1)$ e $d^2_{\boldsymbol{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2)$.
  - Calcule a constante $K = \log(c_1/c_2)$ que envolve os determinantes de $\boldsymbol{\Sigma}_1$ e de $\boldsymbol{\Sigma}_2$.
  - Escolhemos R_1 como sendo o seguinte conjunto de pontos:

$$R_1 = \{\mathbf{x} \in \mathbb{R}^p \text{ tais que } d^2_{\boldsymbol{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) < d^2_{\boldsymbol{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2) + K\}$$

Caso não-gaussiano

- Os dois métodos coincidem no caso gaussiano, gerando a mesma região R_1, mas não no caso não-gaussiano.
- Qual dos dois métodos é o melhor no caso não-gaussiano?
- É melhor usar a razão de densidades ou a distância de Mahalanobis?
- Existiria um terceiro método (árvore de classificação, por exemplo) melhor que estes dois métodos?
- Talvez este terceiro método possa ser usado até no caso gaussiano também.
- Existirá um método imbatível, insuperável, o melhor de todos os possíveis e imagináveis, por mais criativos que sejamos?
- De forma surpreendente, podemos responder SIM a esta questão. E ainda saberemos que método ótimo é este.

O caso geral para classificação

- Na verdade, o problema que vamos resolver é mais geral do que o que consideramos até agora.
- Uma situação mais geral:
  - Custo de classificação errada pode variar de acordo com a classe.
  - Uma das populações pode ser muito mais frequente do que a outra
  - A distribuição pode não ser gaussiana
- Exemplo de risco de crédito:
  - Cliente solicita empréstimo no banco
  - Queremos saber, no momento do empréstimo, se ele é um bom risco (pagará no prazo) ou um mau risco.
  - Nos baseamos em várias características (features) medidas no momento do empréstimo:
    - idade, sexo, tempo como cliente, saldo médio,
    - % do empréstimo em relação ao saldo,
    - já pegou empréstimo antes?

- Custo de classificação em uma matriz:

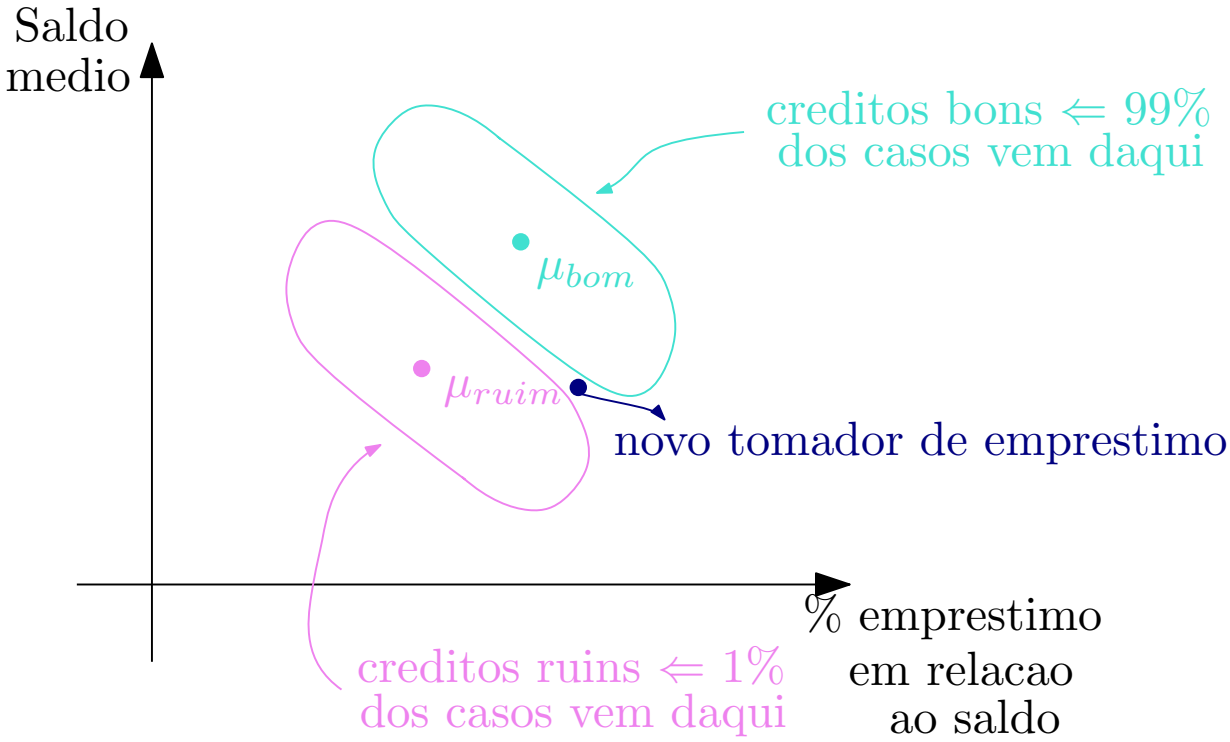
		classificado em π_1	classificado em π_2
População Verdadeira	π_1 Bom crédito	custo $= 0$	$c(2\| \in \pi_1)$
	π_2 Mau crédito	$c(1\| \in \pi_2)$	custo $= 0$

- $c(2| \in \pi_1)$ = custo de classificar como mau crédito um bom pagador;
  = custo de perder um bom cliente;
  = perder o pequeno ganho a ser obtido por juros do empréstimo.
- $c(1| \in \pi_2)$ = custo de classificar como bom crédito um mau pagador;
  = custo de perder todo \$\$ emprestado;
  = perder todo o valor emprestado
- Em geral, nesse problema $c(1| \in \pi_2) >> c(2| \in \pi_1)$

- Um outro exemplo típico: um paciente entra no pronto-socorro com um traumatismo craniano causado por uma queda (comum entre idosos e crianças), um acidente com moto ou esportes ou uma agressão física.
- Não é uma situação rara: ocorrem 50 casos por 10 mil habitantes nos EUA por ano, com 2,5 milhões atendimentos em pronto-socorros, 282 mil internações hospitalares e 56 mil mortes.
- A decisão mais importante é se devemos levar o paciente imediatamente para a UTI ou se ele deve ficar sob observação.
- As primeiras horas após a lesão ocorrer são decisivas.

- Os custos de uma decisão errada são bem diferentes:
  - Levar para a UTI imediatamente mas sem necessidade gasta recursos do hospital que poderiam ser usados de outra forma.
  - Deixar sob observação um paciente que necessitava de tratamento intensivo pode significar sua morte.
- Os custos muito diferentes têm impacto numa regra de classificação: se quisermos minimizar o custo esperado de uma decisão ruim, devemos levar em conta esses custos muito diferentes.
- Como fazer isso?
- O segundo ponto que queremos considerar é o tamanho desbalanceado das duas populações

- No mercado de risco de crédito, maus pagadores são muito mais raros do que bons pagadores.



- E daí?
- Aonde você classificaria um novo item se $d^2_{\mathbf{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) = d^2_{\mathbf{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2)$???

- Suponha que os custos de classificação incorreta sejam iguais: $c(1| \in \pi_2) = c(2| \in \pi_1)$
- Se $d^2_{\mathbf{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) = d^2_{\mathbf{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2)$, estamos dizendo que não existe evidência nas variáveis em $\mathbf{x}$ para saber se $\in \pi_1$ ou se $\in \pi_2$
- O ponto $\mathbf{x}$ está igualmente distante das duas populações.
- Resta uma informação *a priori*, que não está no novo caso $\mathbf{x}$: é que, com alta probabilidade (0.99), o novo caso $\mathbf{x}$ vem de π_1.
- A chance de um novo casos qualquer vir de π_2 é muito pequena (1% apenas). Então:
  - se os custos são os mesmos
  - se o novo caso está igualmente distante de π_1 e π_2
  - parece razoável usar a informação *adicional* de que existem muito mais casos em π_1 do que em π_2 e alocar em π_1.
- Como misturar custos e probabilidades *a priori* no caso geral?

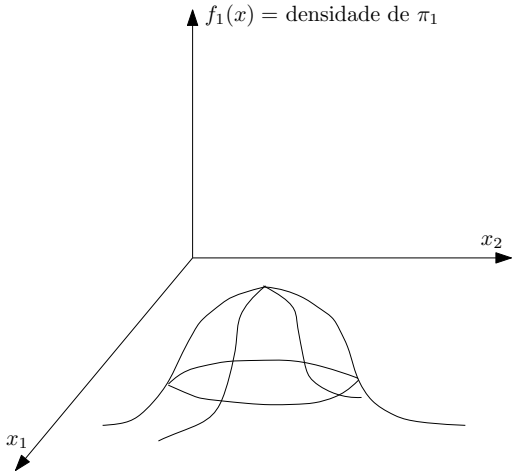
- Um terceiro ponto a ser considerado:
- A distribuição dos dados pode não ser gaussiana.
- No caso gaussiano, como

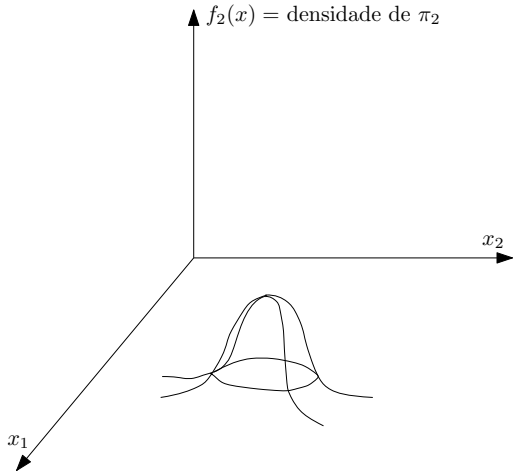
$$f(\mathbf{x}) = c^{te} \cdot exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right) ,$$

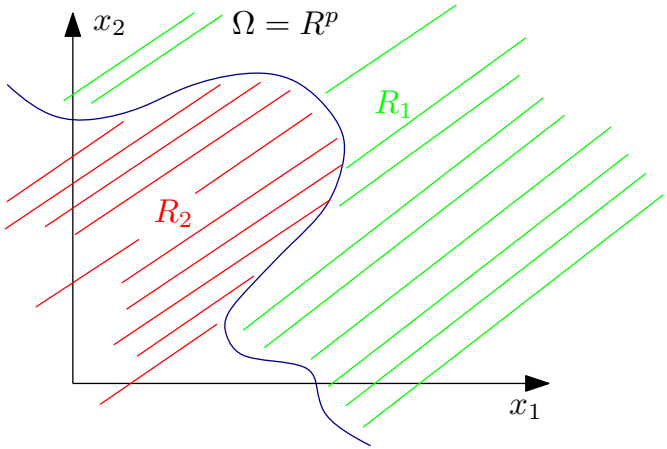
  comparar distâncias de Mahalanobis é equivalente a comparar duas densidades de probabilidades. Os dois métodos são, na verdade, um só. Mas não sabemos se existe outro melhor que este.
- Vamos considerar o caso de distribuições $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$ arbitrárias. Não estaremos restritos a densidades gaussianas nem vamos nos limitar a olhar apenas as distâncias de Mahalanobis.
- E vamos descobrir a melhor regra de classificação: não existe nada melhor que este novo classificador.
- Ele é o *classificador ótimo de Bayes* (optimal Bayes classifier).

Expected cost of misclassification (ECM)

Temos duas densidades, $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$, e uma regra de classificação: $\mathbb{R}^p = R_1 \cup R_2$.



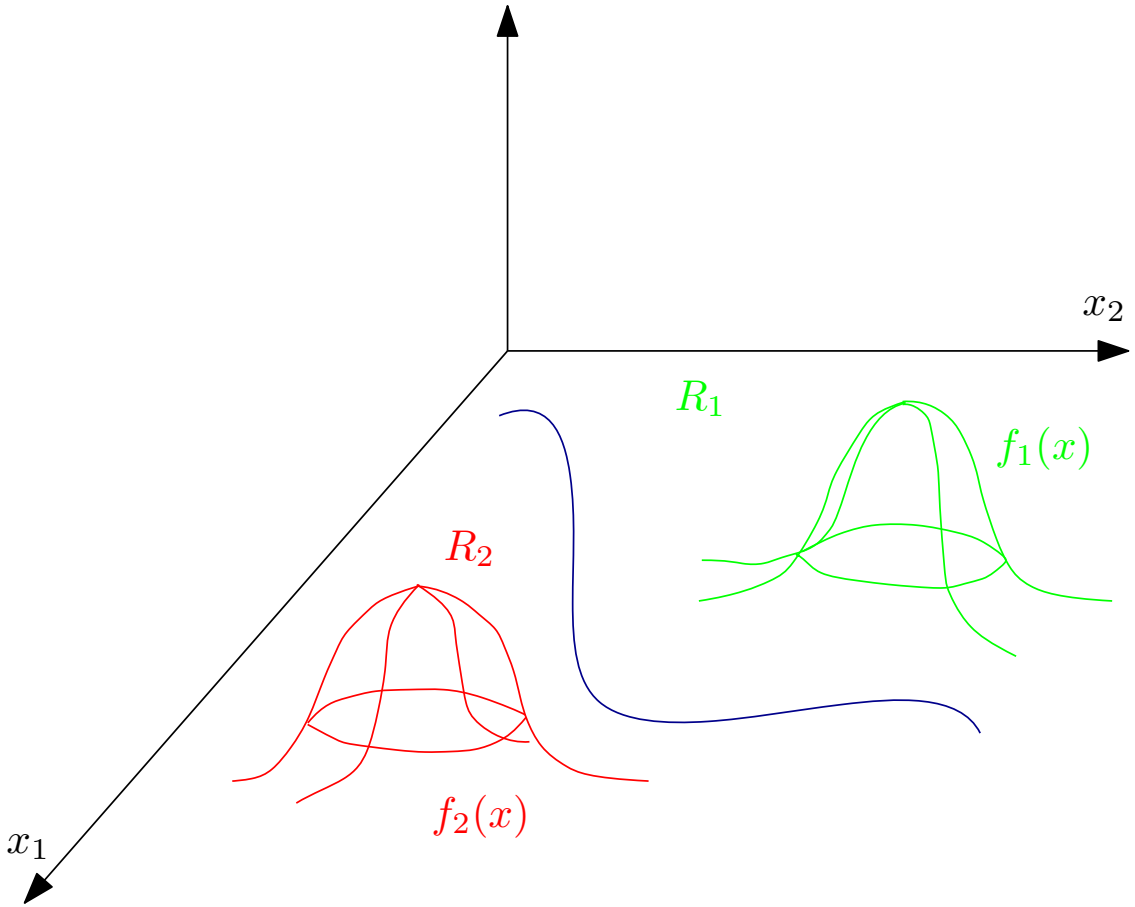




$x \in R_1 \Rightarrow$ classificado em π_1

$x \in R_2 \Rightarrow$ classificado em π_2

- As duas densidades: $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$.
- R_1 e R_2 são definidas por alguma regra de classificação (regra que não é necessariamente boa).



Veja que:

(a) estabelecer uma partição de $R_1 \cup R_2 = \mathbb{R}^p$, com $R_2 = \mathbb{R}^p - R_1$, implica em criar uma regra de classificação:
Regra: Se $\mathbf{x} \in R_1$, aloque $\mathbf{x}$ a π_1. Else, aloque $\mathbf{x}$ a π_2.

(b) estabelecer uma regra de classificação qualquer implica em criar uma partição de $\mathbb{R}^p$:
$R_1 = \{\mathbf{x} \in \mathbb{R}^p |$ a regra aloca $\mathbf{x}$ a $\pi_1\}$
$R_2 = \mathbb{R}^p - R_1$

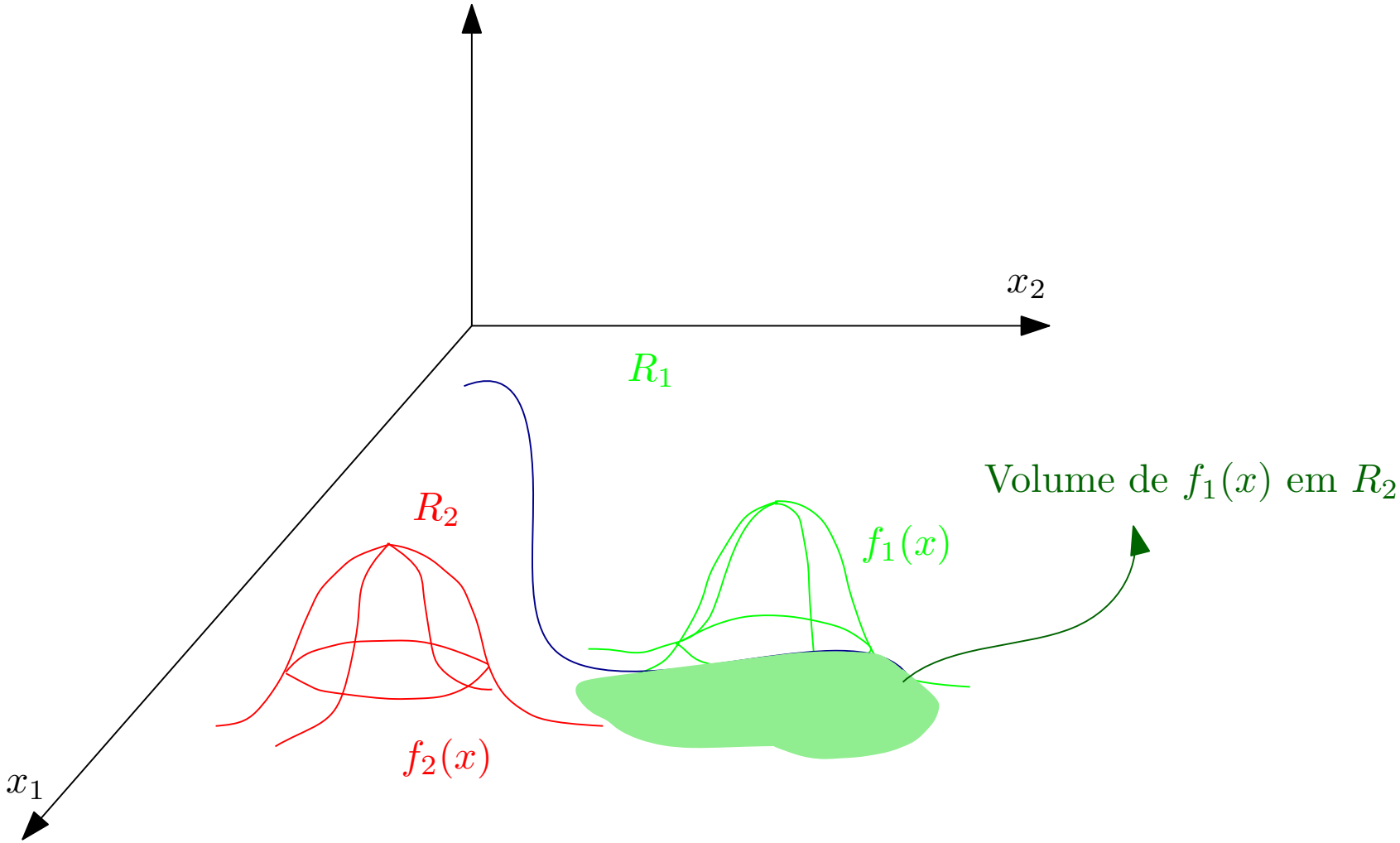
Assim, estabelecer uma regra de classificação baseada em $\mathbf{x} \in \mathbb{R}^p$ é equivalente a estabelecer uma partição $R_1 \cup R_2 = \mathbb{R}^p$.

- Veja que o classificador (ou regra de classificação é uma função matemática, determinística.
- Por exemplo, seja $\mathbf{x} = (x_1, x_2, x_3) = (\text{age, sex, income})$.
- Suponha que $\mathbf{x}_i = (37, \text{FEM}, 25) = \mathbf{x}_j$, duas pessoas com os mesmos três atributos.
- O classificador não muda de valor (ou de classe) diante desses dois exemplos, a classe atribuída será para os dois exemplos.
- O classificador é uma função matemática

$$g(\mathbf{x}) = \begin{cases} \pi_1 & \text{if } \mathbf{x} \in R_1 \\ \pi_2 & \text{if } \mathbf{x} \in R_2 = \mathbb{R}^p - R_1 \end{cases}$$

- Dado um certo exemplo $\mathbf{x}$, a regra vai alocá-lo a uma das duas classes.
- Se tivermos outro exemplo $\mathbf{x}^*$ cujas variáveis tenham os mesmos valores, a classe atribuída a $\mathbf{x}^*$ será a mesma da classe atribuída a $\mathbf{x}$.

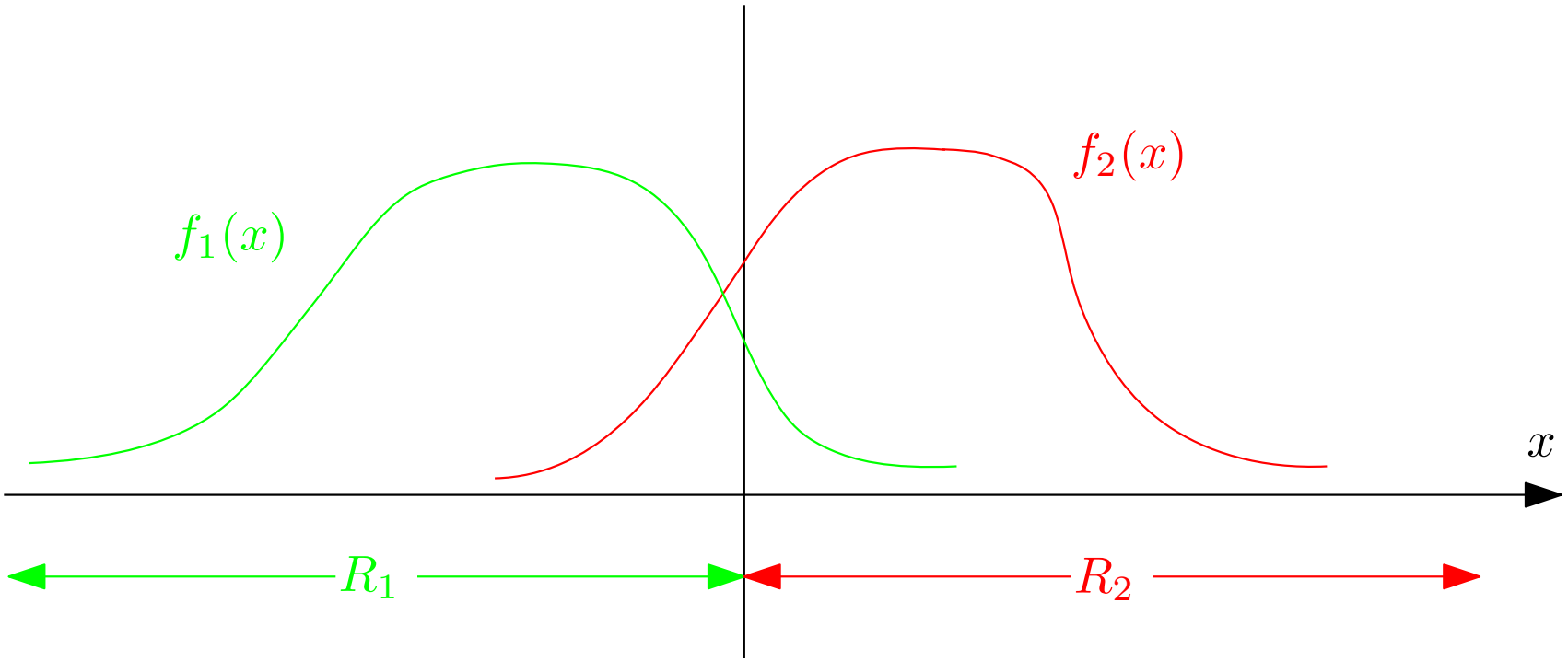
- Probabilidade condicional de classificar um objeto em π_2 quando, de fato, ele é de π_1 é:
- $\mathbb{P}(Class.\ em\ \pi_2 | \in \pi_1) = \mathbb{P}(\mathbf{X} \in R_2 | \in \pi_1) = \int_{R_2} f_1(\mathbf{x}) d\mathbf{x}$



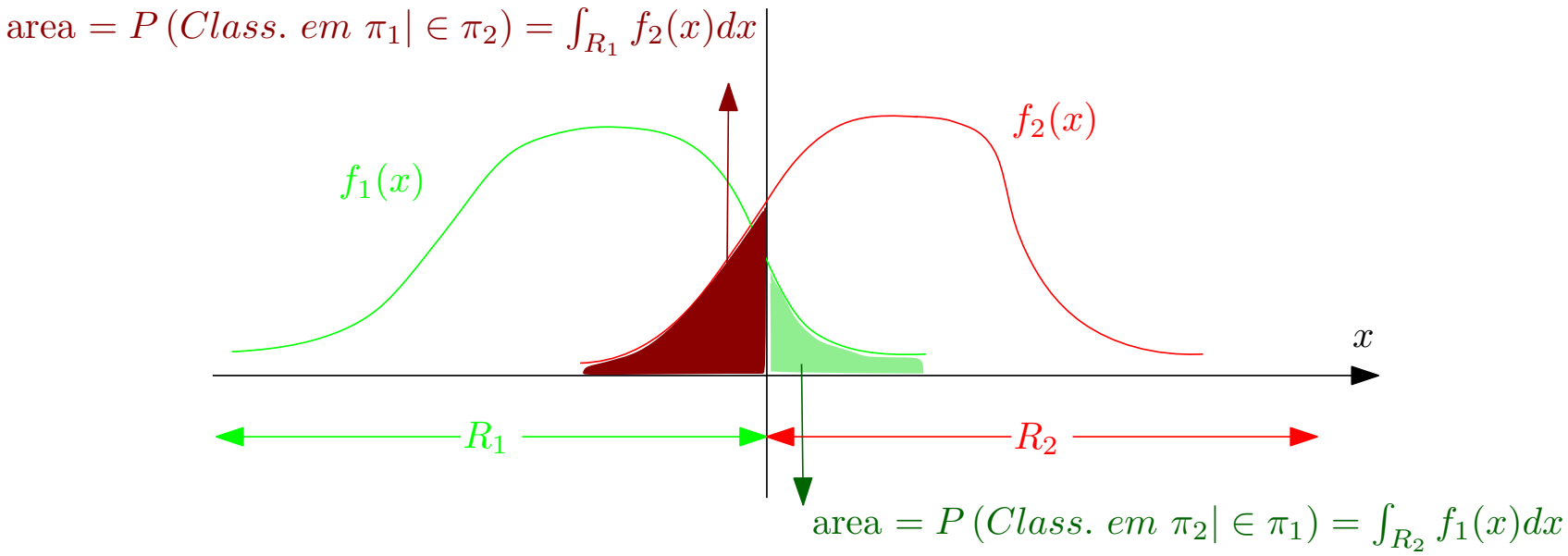
- Similarmente,
- probabilidade de classificar erradamente em π_1 dado que ele é de π_2:
- $\mathbb{P}(Class.\ em\ \pi_1 | \in \pi_2) = \mathbb{P}(\mathbf{X} \in R_1 | \in \pi_2) = \int_{R_1} f_2(\mathbf{x}) d\mathbf{x}$
- A probabilidade desse segundo erro de classificação é o volume (integral) sob $f_2(\mathbf{x})$ na região R_1.

Caso 1-dim

- Vamos ver o caso em que $p = 1$ (uma única variável)
- As densidades $f_1(x)$ e $f_2(x)$, e R_1 e R_2:

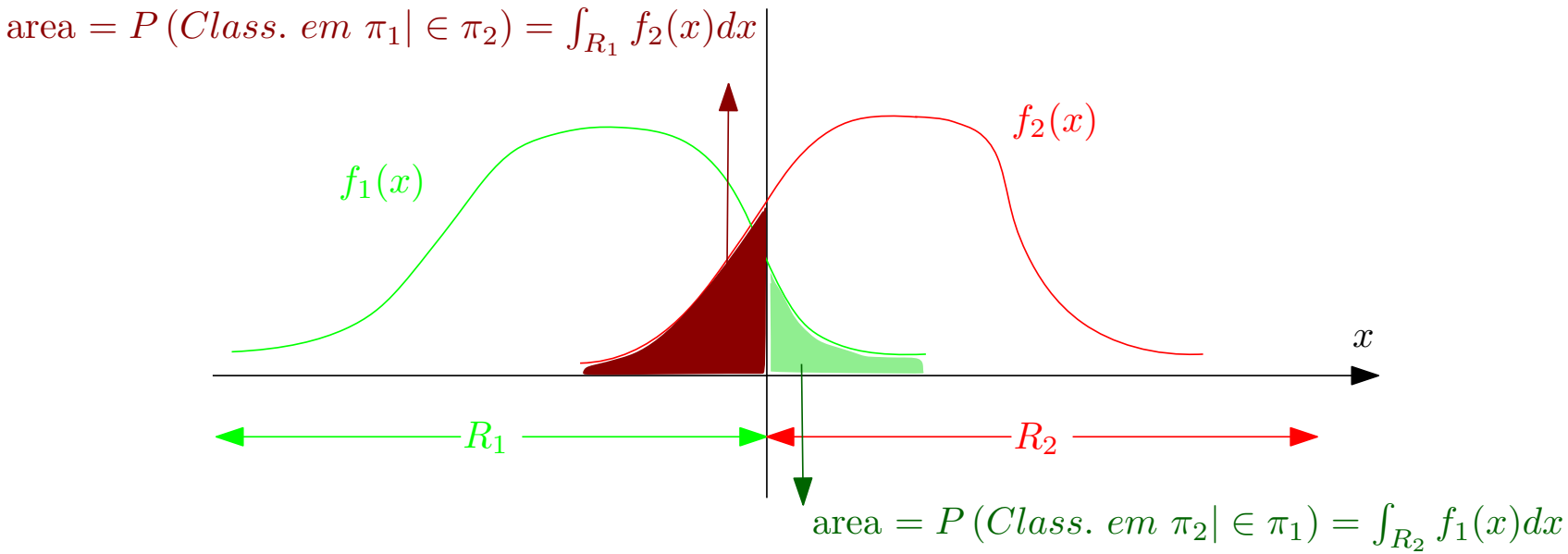


- As duas probabilidades de classificação incorreta:



- Área em vermelho: probab de alocar a π_1 um exemplo vindo de $f_2(x)$ (e portanto, vindo de π_2).
- Área em verde: probab de alocar a π_2 um exemplo vindo de $f_1(x)$ (e portanto, vindo de π_1).

Trade-off

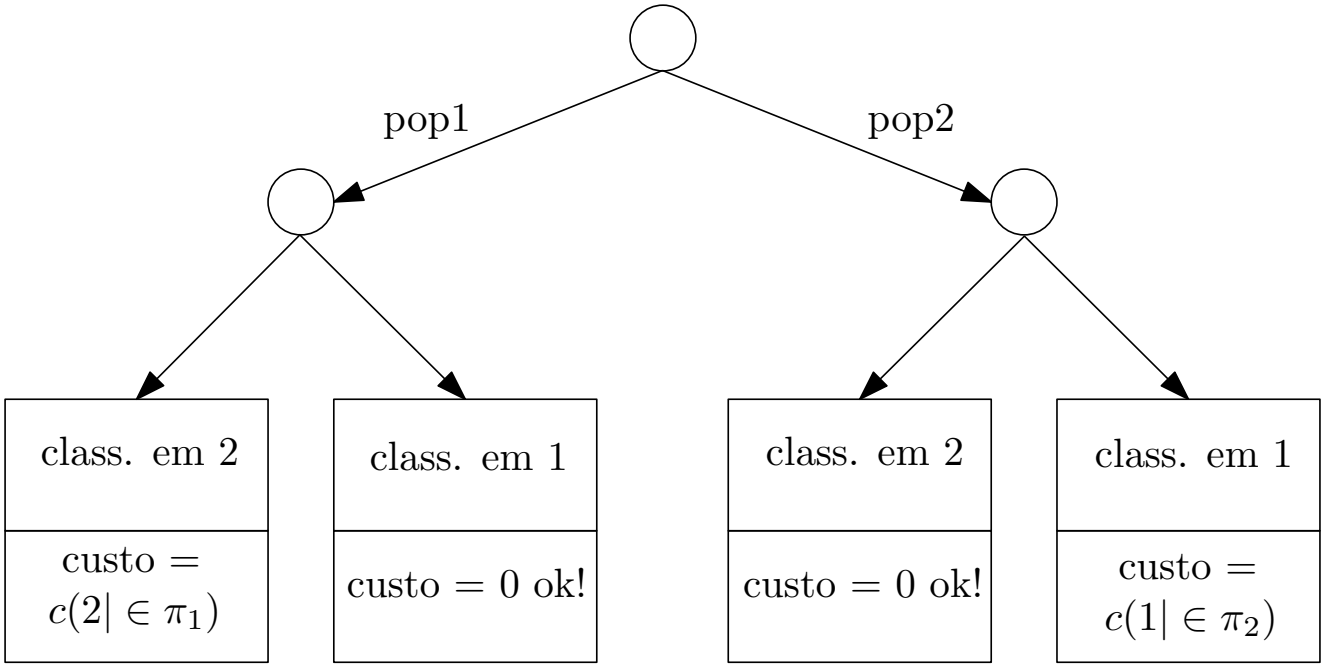


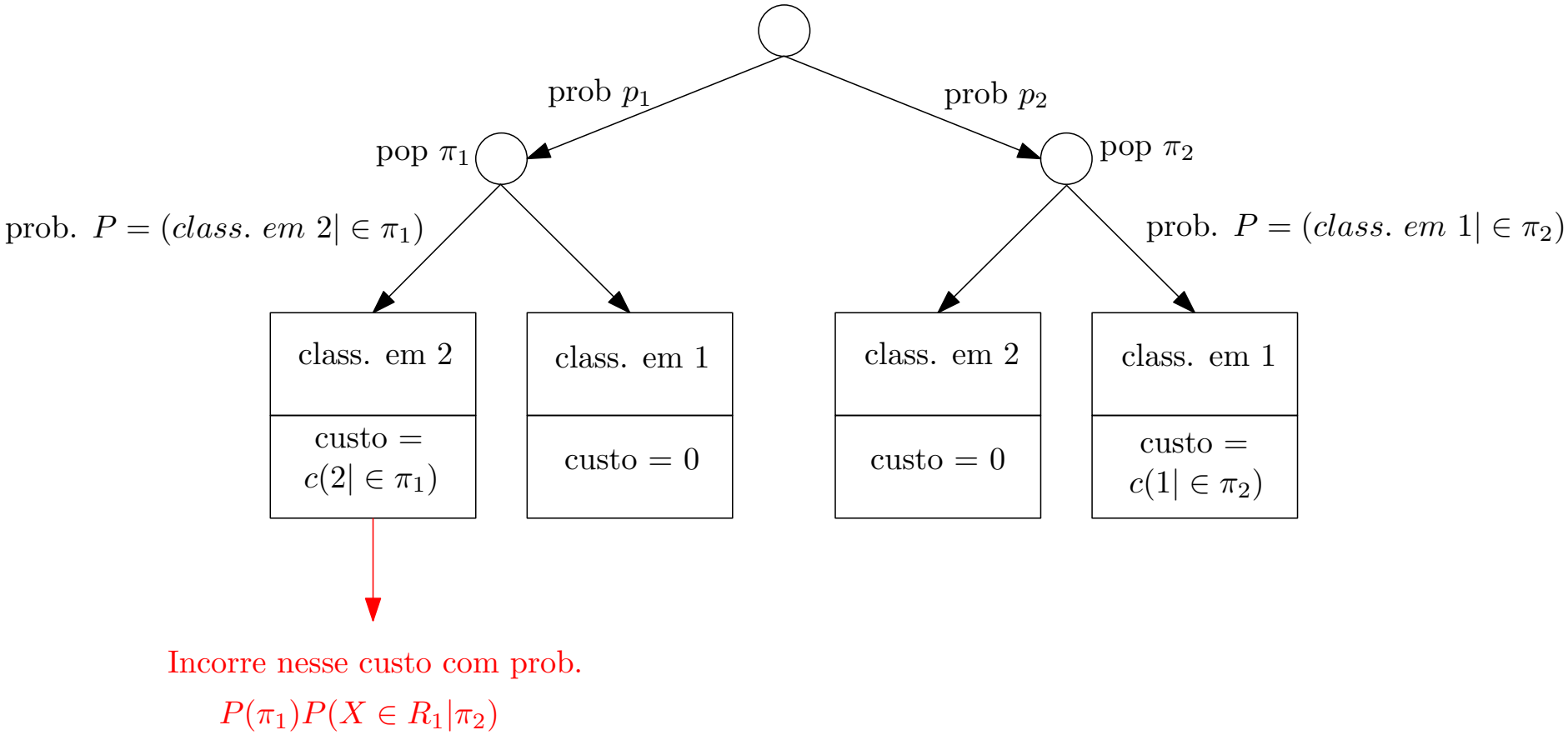
- Veja que quando procuramos diminuir $\mathbb{P}\,(Class.\ em\ \pi_1| \in \pi_2)$ estamos aumentando $\mathbb{P}\,(Class.\ em\ \pi_2| \in \pi_1)$.
- Existe um trade-off entre essas probabilidades de classificação incorreta.
- Como escolher uma boa partição R_1 e R_2 do espaco $\mathbb{R}^p$?
- Como os dois erros possuem custos diferentes, nós vamos minimizar o custo esperado de má classificação.

- Temos também as probabilidades a priori de que os objetos venham de π_1 ou π_2:

$$\begin{aligned} p_1 &= \mathbb{P}(\in \pi_1) \\ p_2 &= \mathbb{P}(\in \pi_2) = 1 - \mathbb{P}(\in \pi_1) = 1 - p_1 \end{aligned}$$

- Quadro geral:





TEXTO VERMELHO ERRADO: Deveria ser $P(\pi_1)P(X \in R_2| \in \pi_1)$.

- Casos novos chegam: alguns nós clasificamos corretamente (com custo zero); outros, classificamos incorretamente (com custo > 0).
- Podemos ter $c(2| \in \pi_1) \neq c(1| \in \pi_2)$
- Nos casos em que erramos, às vezes caímos no custo mais elevado; às vezes, no custo menor.
- É impossível (nos casos realistas) ter uma regra baseada num vetor $\mathbf{x}$ que nunca erre.
- Queremos uma regra de classificação que, em geral (ou, em média) leve a um custo pequeno
- $\Rightarrow$ queremos um custo médio (ou esperado) pequeno.

EMC: Expected misclassification cost

- Custo esperado (ou custo médio) de má-classificação:
- Custo é variável aleatória e possui três valores possíveis: 0, $c(2|\in\pi_1)$ e $c(1|\in\pi_2)$
- Estes custos aleatórios acontecem com certas probabilidades.
- Qual seu valor esperado?

$$
\begin{aligned}
EMC &= \mathbb{E}(\text{cost}) \\
&= 0 \times \mathbb{P}(\text{acertar}) + \text{cost}_1 \times \mathbb{P}(\text{erro 1}) + \text{cost}_2 \times \mathbb{P}(\text{erro 2}) \\
&= c(2|\in\pi_1)\mathbb{P}(\text{vir de } \pi_1 \text{ e errar}) + c(1|\in\pi_2)\mathbb{P}(\text{vir de } \pi_2 \text{ e errar}) \\
&= c(2|\in\pi_1)\mathbb{P}(\mathbf{X}\in R_2|\in\pi_1)\mathbb{P}(\pi_1) + c(1|\in\pi_2)\mathbb{P}(\mathbf{X}\in R_1|\in\pi_2)\mathbb{P}(\pi_2)
\end{aligned}
$$

- $EMC \rightarrow$ é custo esperado de má classificação. (expected misclassification cost).

- Queremos achar as regiões R_1 e R_2 que minimizam o EMC.
- Isto é equivalente a encontrar o classificador que torna o EMC o menor possível.
- Solução:

$$R_1 = \left\{ \mathbf{x} \in \mathbb{R}^p \text{ tais que } \underbrace{\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}}_{(1)} \geq \underbrace{\frac{c(1|\in \pi_2)}{c(2|\in \pi_1)}}_{(2)} \cdot \underbrace{\frac{p_2}{p_1}}_{(3)} \right\}$$

  - (1): razão das densidades das duas classes
  - (2): razão de custos
  - (3): razão de probabilidades a priori

- Prova: Queremos R_1 e R_2 que minimizem EMC:

$$EMC = c(2|\in \pi_1)\mathbb{P}(\mathbf{X} \in R_2|\in \pi_1)\mathbb{P}(\pi_1) + c(1|\in \pi_2)\mathbb{P}(\mathbf{X} \in R_1|\in \pi_2)\mathbb{P}(\pi_2)$$

$\int_{R_2} f_1(x)\, dx$ $\qquad$ $\int_{R_1} f_2(x)\, dx$

- Como $R_1 \cup R_2 = \mathbb{R}^p$ então

$$1 = \int_{\mathbb{R}^p} f_1(\mathbf{x})\, d\mathbf{x} = \int_{R_1} f_1(\mathbf{x})\, d\mathbf{x} + \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x}\,.$$

- e podemos escrever a primeira integral em EMC (em azul) da seguinte forma:

$$\int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} = 1 - \int_{R_1} f_1(\mathbf{x})\, d\mathbf{x}$$

- Vamos agora substituir a integral azul em EMC pela expressão em vermelho.

Trabalhando EMC...

- Temos

$$EMC = c(2| \in \pi_1)\left(1 - \int_{R_1} f_1(x)\ dx)\right)\ \mathbb{P}(\pi_1) + c(1| \in \pi_2) \int_{R_1} f_2(x)\ dx)\ \mathbb{P}(\pi_2)$$

- Agora, as duas integrais possuem a mesma região R_1 de integração e portanto os dois integrandos podem ser colocados sob o mesmo sinal de integral. Isto implica que

$$EMC = c(2| \in \pi_1)\mathbb{P}(\pi_1) + \int_{R_1} (c(1| \in \pi_2)\ \mathbb{P}(\pi_2)\ f_2(x) - c(2| \in \pi_1)\ \mathbb{P}(\pi_1)\ f_1(x))\ \ dx)$$

- Queremos escolher R_1 de forma que EMC seja mínimo.

$$EMC = c(2|\in \pi_1)\mathbb{P}(\pi_1) + \int_{R_1} \underbrace{(c(1|\in \pi_2)\,\mathbb{P}(\pi_2)\,f_2(\mathbf{x}) - c(2|\in \pi_1)\,\mathbb{P}(\pi_1)\,f_1(\mathbf{x}))}_{h(\mathbf{x})}\,d\mathbf{x}$$

- O 1^o termo não envolve R_1.
- Vamos olhar o 2^o termo
- Escolher R_1 é escolher a região em que vamos integrar ("somar') $h(\mathbf{x})$.
- A expressão $h(\mathbf{x})$ não envolve R_1.
- Para alguns $\mathbf{x}$, teremos $h(\mathbf{x}) > 0$; para outros pontos $\mathbf{x}$, teremos $h(\mathbf{x}) < 0$
- Para minimizar EMC, devemos tornar a integral o mais negativa possível.
- Conseguimos isto escolhendo $R_1 = \{\mathbf{x} \in \mathbb{R}^p$ tais que $h(\mathbf{x}) \leq 0\}$.
- Isso minimiza EMC!!

- ECM é minimizado se escolhermos

$$R_1 = \{\mathbf{x} \in \mathbb{R}^p \text{ tais que } h(\mathbf{x}) \leq 0\}$$

$$\begin{aligned} EMC &= c(2|\in \pi_1)\mathbb{P}(\pi_1) + \int_{R_1} \underbrace{\left(c(1|\in \pi_2)\,\mathbb{P}(\pi_2)\,f_2(\mathbf{x}) - c(2|\in \pi_1)\,\mathbb{P}(\pi_1)\,f_1(\mathbf{x})\right)}_{h(\mathbf{x})}\, d\mathbf{x} \\ &= c(2|\in \pi_1)\mathbb{P}(\pi_1) + I \qquad \text{onde } I = \int_{R_1} h(\mathbf{x})\, d\mathbf{x} \end{aligned}$$

onde $I = \int_{R_1} h(\mathbf{x})\, d\mathbf{x}$.

- Veja que $h(\mathbf{x}) \leq 0$ é o mesmo que

$$c(1|\in \pi_2)\,\mathbb{P}(\pi_2)\,f_2(\mathbf{x}) - c(2|\in \pi_1)\,\mathbb{P}(\pi_1)\,f_1(\mathbf{x}) \leq 0$$

- Passando o segundo termo para o outro lado da desigualdade e as posições, temos

$$c(2|\in \pi_1)\,\mathbb{P}(\pi_1)\,f_1(\mathbf{x}) \geq c(1|\in \pi_2)\,\mathbb{P}(\pi_2)\,f_2(\mathbf{x})$$

ou ainda, após rearranjar os termos,

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|\in \pi_2)}{c(2|\in \pi_1)}\,\frac{\mathbb{P}(\pi_2)}{\mathbb{P}(\pi_1)}$$

- Para ficar em paz com esta afirmação, defina $R_1 = \{\mathbf{x} \in \mathbb{R}^p \text{ tais que } h(\mathbf{x}) \leq 0\}$ e seja E_1 o valor do ECM com esta regra de classificação .
- Se $\mathbf{x} \notin R_1$ então $h(\mathbf{x}) > 0$.
- Seja $R_1^* = A \cup R_1$ uma nova regra de classificação (com $A \cap R_1 = \emptyset$) com ECM dado por E_1^*
- Veremos que $E_1 \leq E_1^*$.

- Temos

$$
\begin{aligned}
E_1^* &= c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1^*} h(\mathbf{x})\, d\mathbf{x} \\
&= c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1\cup A} h(\mathbf{x})\, d\mathbf{x} \\
&= c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1} h(\mathbf{x})\, d\mathbf{x} + \int_A \underbrace{h(\mathbf{x})}_{>0}\, d\mathbf{x} \\
&\geq c(2|\in\pi_1)\mathbb{P}(\pi_1) + \int_{R_1} h(\mathbf{x})\, d\mathbf{x} = E_1
\end{aligned}
$$

- Assim, aumentar a região R_1 com qualquer outra região A Inunca será capaz de fazer o ECM ser menor que aquele de R_1.
- Um argumento análogo, mostra que definir uma nova região para a classe 1 subtraindo uma área qualquer de R_1 também nunca leva a um ECM menor (exercício).

Resumo: Optimal Bayes Classifier

- Em cada caso, observamos o vetor aleatório $\mathbf{X}$.
- Existem duas populações: π_1 e π_2.
- Um novo caso vem da pop π_1 com probabilidade p_1 e da pop π_2 com probabilidade $p_2 = 1 - p_1$.
- As densidades de $\mathbf{x}$: $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$.
- Existem dois custos de classificação errada: $c(2| \in \pi_1)$ e $c(1| \in \pi_2)$
- Baseado em $\mathbf{x}$, queremos predizer a sua classe: 1 ou 2.
- Cada regra de classificação tem seu ECM = custo esperado de má-classificação (custo médio se classificarmos vários itens).
- Dentre todas as regras possíveis, aquela que torna mínimo o ECM é: aloque o caso a π_1 caso

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1| \in \pi_2)}{c(2| \in \pi_1)} \frac{\mathbb{P}(\pi_2)}{\mathbb{P}(\pi_1)}$$

Resumo: Optimal Bayes Classifier

- Regra ótima de Bayes:

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \underbrace{\frac{c(1|\in\pi_2)}{c(2|\in\pi_1)}\ \frac{\mathbb{P}(\pi_2)}{\mathbb{P}(\pi_1)}}_{\text{cte. em } \mathbf{x}}$$

- A regra é bem intuitiva, gera um algoritmo muito simples.
- Recebemos um novo caso com atributos no vetor $\mathbf{x}$.
- Qual sua classe? 1 ou 2?
- Calcule $f_1(\mathbf{x})/f_2(\mathbf{x})$, a razão de densidades no ponto $\mathbf{x}$ (aprox, é a razão das "probabilidades" de observar $\mathbf{x}$ em 1 ou 2).
- Se esta razão for "grande", aloque a 1. Razoável, não?
- De fato, se $f_1(\mathbf{x})/f_2(\mathbf{x}) \approx 7$, então a chance de observar $\mathbf{x}$ em 1 é aprox 7 vezes maior que a mesma chance em 2.
- Parece razoável alocar a 1. Mas...

Resumo: Optimal Bayes Classifier

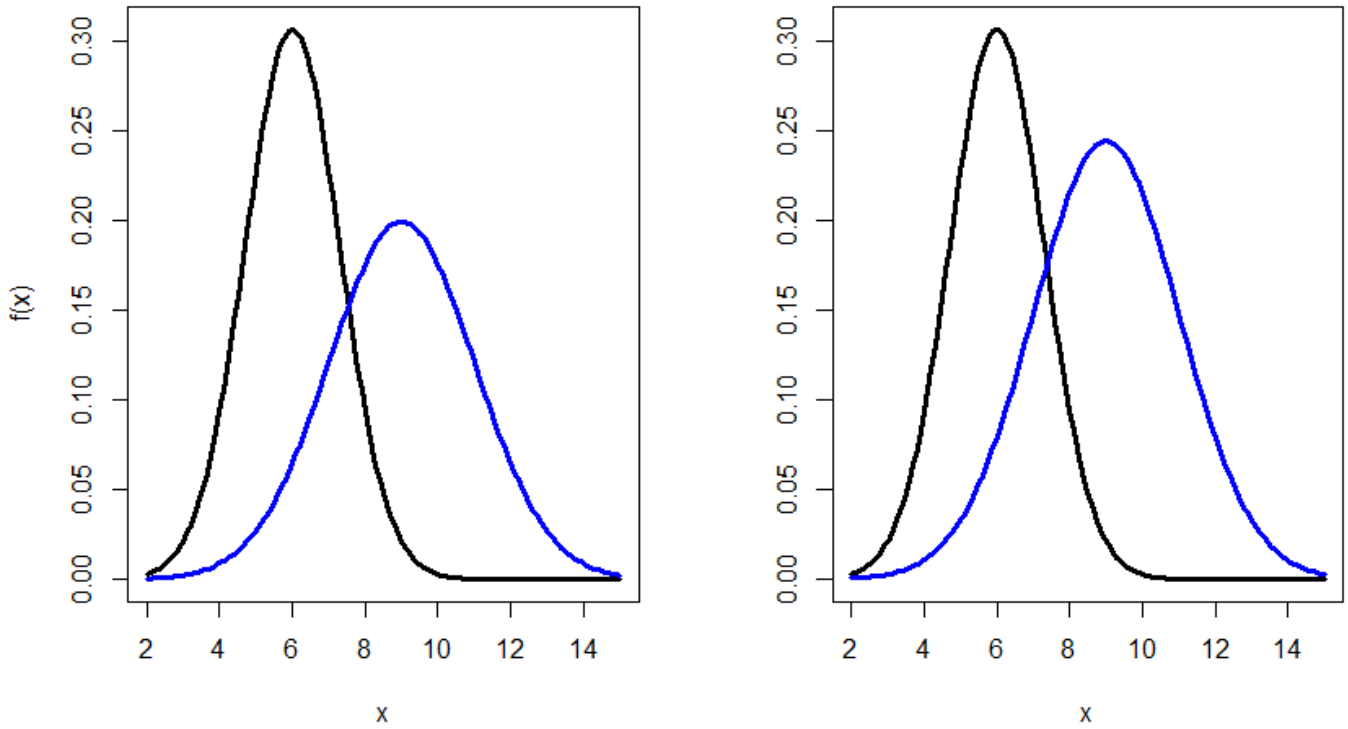
- Por quê não alocar a 1 simplesmente se tivermos $f_1(\mathbf{x})/f_2(\mathbf{x}) > 1$? Isto é, se $f_1(\mathbf{x}) > f_2(\mathbf{x})$, devemos alocar a 1?
- Nem sempre.
- Queremos levar em conta os custos e diferentes frequências das classes na população total.
- Como fazer isto?
- Basta calcular: (a) a razão de custos, (b) a razão de probabilidades *a priori*, e multiplicá-las.
- Este valor não depende de $\mathbf{x}$, é uma constante.
- A beleza não é porque a regra é muito simples. Qualquer um pode bolar uma regra simples.
- A beleza é que a regra é muito simples *e é a melhor possível e imaginável*. Nada pode ser melhor que ela (para reduzir o ECM).

Exemplo uni-dimensional

- Imagine que π_1 é uma classe rara: $p_1 = 0.02$
- É muito pior errar quando o item é de π_1: $c(2| \in \pi_1) = 40c(1| \in \pi_2)$.
- A regra é então alocar a 1 toda vez que $f_1(\mathbf{x})/f_2(\mathbf{x}) \geq (1/40) \times (0.98/0.02) = 1.22$.
- Ou seja, se $f_1(\mathbf{x}) \geq 1.22\ f_2(\mathbf{x})$, aloque a 1.
- Suponha que as duas densidades sejam como a seguir.
- Como encontrar a região de alocação a π_1?

Exemplo uni-dimensional

- Na esquerda, temos o plot de $f_1(x)$ (preto) e $f_2(x)$ (azul).
- Na direita, temos $f_1(x)$ e $1.22 \times f_2(x)$.
- Aloque a 1 toda vez que a curva preta for maior que a curva azul neste plot da direita.



Classificação ótima com duas gaussianas

- Densidade de uma $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ no ponto $\mathbf{x} \in \mathbb{R}^p$:

$$
\begin{aligned}
f(\mathbf{x}) &= \underbrace{\left[(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\right]}_{\text{constante em } \mathbf{x}} \exp\left(-\frac{1}{2}\underbrace{(\mathbf{x}-\boldsymbol{\mu})^t\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})}_{\text{Mahalanobis}}\right) \\
&= k\ \exp\left(-\frac{1}{2}d^2_{\boldsymbol{\Sigma}}(\mathbf{x},\boldsymbol{\mu})\right)
\end{aligned}
$$

- Queremos a razão de duas densidades gaussianas, $f_1(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ e $f_2(\mathbf{x}) \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, no mesmo ponto $\mathbf{x}$:

$$
\begin{aligned}
\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{k_1\ \exp\left(-\frac{1}{2}d^2_{\boldsymbol{\Sigma}_1}(\mathbf{x},\boldsymbol{\mu}_1)\right)}{k_2\ \exp\left(-\frac{1}{2}d^2_{\boldsymbol{\Sigma}_2}(\mathbf{x},\boldsymbol{\mu}_2)\right)} \\
&= \frac{k_1}{k_2}\exp\left(-\frac{1}{2}(d^2_{\boldsymbol{\Sigma}_1}(\mathbf{x},\boldsymbol{\mu}_1) - d^2_{\boldsymbol{\Sigma}_2}(\mathbf{x},\boldsymbol{\mu}_2))\right)
\end{aligned}
$$

Classificação ótima com duas gaussianas

- A regra ótima é: aloque a π_1 se

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|\in \pi_2)}{c(2|\in \pi_1)} \frac{\mathbb{P}(\pi_2)}{\mathbb{P}(\pi_1)}$$

- Portanto, aloque a π_1 se

$$\frac{k_1}{k_2} \exp\left(-\frac{1}{2}(d^2_{\mathbf{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) - d^2_{\mathbf{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2))\right) \geq \frac{c(1|\in \pi_2)}{c(2|\in \pi_1)} \frac{\mathbb{P}(\pi_2)}{\mathbb{P}(\pi_1)}$$

- Tomando logs dos dois lados e mudando de lado alguns termos, temos que alocar a π_1 se:

$$d^2_{\mathbf{\Sigma_1}}(\mathbf{x}, \boldsymbol{\mu_1}) \leq d^2_{\mathbf{\Sigma_2}}(\mathbf{x}, \boldsymbol{\mu_2}) + 2\left[\log\left(\frac{c(2|\in \pi_1)}{c(1|\in \pi_2)}\right) + \log\left(\frac{\mathbb{P}(\pi_1)}{\mathbb{P}(\pi_2)}\right) + \log\left(\frac{k_2}{k_1}\right)\right]$$

- Vamos entender um pouco melhor esta fórmula.

Classificação ótima com duas gaussianas

- Repetindo, alocar a π_1 se:

$$d^2_{\Sigma_1}(x, \mu_1) \leq d^2_{\Sigma_2}(x, \mu_2) + 2\log\left(\frac{c(2|\in\pi_1)}{c(1|\in\pi_2)}\right) + 2\log\left(\frac{\mathbb{P}(\pi_1)}{\mathbb{P}(\pi_2)}\right) + 2\log\left(\frac{k_2}{k_1}\right)$$

Mahalanobis Mahalanobis custos prioris covariancias

- Ideia: aloque a π_1 se a distância de Mahalanobis de $\mathbf{x}$ a $\boldsymbol{\mu}_1$ for menor que a distância de Mahalanobis a $\boldsymbol{\mu}_2$ mais ou menos "alguma coisa".
- O "alguma coisa" leva em conta os custos, prioris e estruturas de covariância de cada população.
- Esta fórmula mostra a *melhor* maneira de levar estes aspectos em conta.
- Por exemplo, os custos devem ser analisados em função de sua diferença *relativa* e numa escala log.
- Por exemplo, não é a diferença $c(2|\in\pi_1) - c(1|\in\pi_2)$ que nos interessa, mas sim $c(2|\in\pi_1)/c(1|\in\pi_2)$.

Classificação ótima com duas gaussianas

- Repetindo, alocar a π_1 se:

$$d^2_{\Sigma_1}(x, \mu_1) \leq d^2_{\Sigma_2}(x, \mu_2) + 2\log\left(\frac{c(2|\in\pi_1)}{c(1|\in\pi_2)}\right) + 2\log\left(\frac{\mathbb{P}(\pi_1)}{\mathbb{P}(\pi_2)}\right) + 2\log\left(\frac{k_2}{k_1}\right)$$

Mahalanobis Mahalanobis custos prioris covariancias

- O termo envolvendo os custos dos dois erros desaparece se eles forem iguais.
- Aquele envolvendo as probabilidades *a priori* também desaparece se $\mathbb{P}(\pi_1) = \mathbb{P}(\pi_2)$.
- Do mesmo modo, se $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, o último termo desaparece.
- Neste caso em que todos estes termos desaparecem, a regra ótima simplesmente compara as distâncias de Mahalanobis": alocar a π_1 se

$$d^2_{\mathbf{\Sigma}}(\mathbf{x}, \mu_1) \leq d^2_{\mathbf{\Sigma}}(\mathbf{x}, \mu_2)$$

Classificação ótima com duas gaussianas

- Vamos começar a introduzir os termos adicionais, um de cada vez, para entender seu efeito.
- Suponha que os custos sejam diferentes e que um deles é 100 vezes maior que o outro: $c(2|\in\pi_1) = 100c(1|\in\pi_2)$.
- Isto é, é 100 mais pior alocar um caso de π_1 erradamente que alocar errado um caso de π_2.
- Deveríamos ser *menos propensos* então a alocar um caso a π_2.
- De fato, neste caso, a regra ótima é alocar $\mathbf{x}$ a π_1 se

$$d^2_{\mathbf{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_1) \leq d^2_{\mathbf{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_2) + 2\log(100)$$

- Agora ficou mais fácil alocar um caso a π_1. A distância de Mahalanobis a π_1 nem precisa ser a menor delas agora.
- Lembre-se: esta é a regra ótima, a melhor possível.

Classificação ótima com duas gaussianas

- Do mesmo modo, suponha que a classe 1 seja 100 vezes mais frequente que a classe 2: $\mathbb{P}(\pi_1) = 100\mathbb{P}(\pi_2)$.
- Isto é, quando um caso qualquer aparece, sem considerar o valor de $\mathbf{x}$, sabemos que é 100 mais provável que ele seja de π_1 do que de π_2.
- Novamente, deveríamos ser *menos propensos* então a alocar um caso a π_2.
- Como antes, a regra ótima é alocar $\mathbf{x}$ a π_1 se

$$d^2_{\mathbf{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_1) \leq d^2_{\mathbf{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}_2) + 2\log(100)$$

Classificação ótima com duas gaussianas

- Por último, para ver o efeito do terceiro termo, vamos imaginar que a variável $\mathbf{x}$ é uni-dimensional.
- Assim, $\mathbf{\Sigma}_1 = \sigma_1^2$ e $\mathbf{\Sigma}_2 = \sigma_2^2$, as variâncias de X em cada população.
- Além disso, a distância de Mahalanobis reduz-se a $d^2_{\mathbf{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}) = ((x - \mu)/\sigma)^2$, o desvio padronizado.
- Suponha que uma das populações tenha variância muito maior que a outra: $\sigma_2^2 = (10)^2\sigma_1^2$.
- Isto é, pontos de π_2 espalham-se em torno de sua média μ_2 muito mais que pontos de π_1 em torno de μ_1.
- A regra ótima é alocar $\mathbf{x}$ a π_1 se

$$\left(\frac{x - \mu_1}{\sigma_1}\right)^2 \leq \left(\frac{x - \mu_2}{\sigma_2}\right)^2 + \log(100)$$

- Assim, penalizamos a população com maior dispersão. Pense assim, se as duas distâncias padronizadas forem iguais, o melhor é alocar à π_1, a menos dispersa.

Caso gaussiano com $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$

- Alocar a π_1 se:

$$d^2_{\mathbf{\Sigma_1}}(\mathbf{x}, \boldsymbol{\mu_1}) - d^2_{\mathbf{\Sigma_2}}(\mathbf{x}, \boldsymbol{\mu_2}) \leq 2 \log \left(\frac{c(2| \in \pi_1)\mathbb{P}(\pi_1)}{c(1| \in \pi_2)\mathbb{P}(\pi_2)} \right) + 2 \log \left(\frac{k_2}{k_1} \right)$$

- Se $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$, os seus determinantes também são iguais e $k_2 = k_1$.
- Vamos denotar a constante (em $\mathbf{x}$) do lado esquerdo, envolvendo as probabilidades a priori e os custos de K:

$$K = \log \left(\frac{c(2| \in \pi_1)\mathbb{P}(\pi_1)}{c(1| \in \pi_2)\mathbb{P}(\pi_2)} \right)$$

- A regra ótima no caso gaussiano com covariâncias iguais é alocar a π_1 se

$$d^2_{\mathbf{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) - d^2_{\mathbf{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2) \leq 2K$$

Caso gaussiano com $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$

- Alocar a π_1 se $d^2_{\mathbf{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) - d^2_{\mathbf{\Sigma}_2}(\mathbf{x}, \boldsymbol{\mu}_2) \leq 2K$
- Expandimos a expressão da distância:

$$\begin{aligned} d^2_{\mathbf{\Sigma}_1}(\mathbf{x}, \boldsymbol{\mu}_1) &= (\mathbf{x} - \boldsymbol{\mu}_1)^t \, \mathbf{\Sigma}_1^{-1} \, (\mathbf{x} - \boldsymbol{\mu}_1) \\ &= \mathbf{x}^t \mathbf{\Sigma}_1^{-1} \mathbf{x} + \boldsymbol{\mu}_1^t \mathbf{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - 2\mathbf{x}^t \mathbf{\Sigma}_1^{-1} \boldsymbol{\mu}_1 \end{aligned}$$

- Do mesmo modo, expandimos a outra distância. Cancelamos o termo $\mathbf{x}^t \mathbf{\Sigma}_1^{-1} \mathbf{x}$ encontrando: aloque $\mathbf{x}$ a π_1 se

$$\begin{aligned} 0 &\leq \mathbf{x}^t \underbrace{\mathbf{\Sigma}_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}_{p \times 1, \quad \boldsymbol{\beta}} + \underbrace{(\boldsymbol{\mu}_2^t \mathbf{\Sigma}_2^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1^t \mathbf{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - K)}_{1 \times 1, \quad \alpha} \\ &= \alpha + \boldsymbol{\beta}^t \mathbf{x} \end{aligned}$$

usando que $\boldsymbol{\beta}^t \mathbf{x} = \mathbf{x}^t \boldsymbol{\beta}$.

Caso gaussiano com $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$

- Vamos escrever $\lambda(\mathbf{x}) = \alpha + \boldsymbol{\beta}^t \mathbf{x}$.
- O conjunto de pontos $\mathbf{x} \in \mathbb{R}^p$ tais que $\lambda(\mathbf{x}) = 0$ constitui a fronteira de decisão (decision boundary).
- No caso em que $\mathbf{x} \in \mathbb{R}^2$, esta fronteira é uma linha reta.
- Se $\mathbf{x} \in \mathbb{R}^3$, a fronteira é um plano.
- Passar para o notebook python para ilustrar.

Caso gaussiano com $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$

- E o caso gaussiano com $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$?
- Manipulação matricial da regra ótima leva à conclusão de que a fronteira de decisão é uma parábola, e não mais uma reta.
- A fórmula geral do caso gaussiano, como já sabemos, alocar a π_1 se:

$$d^2_{\mathbf{\Sigma_1}}(\mathbf{x}, \boldsymbol{\mu_1}) - d^2_{\mathbf{\Sigma_2}}(\mathbf{x}, \boldsymbol{\mu_2}) \leq 2 \log \left(\frac{c(2| \in \pi_1)\mathbb{P}(\pi_1)}{c(1| \in \pi_2)\mathbb{P}(\pi_2)} \right) + 2 \log \left(\frac{k_2}{k_1} \right)$$

- Expandindo as fórmulas quadráticas das distâncias, exatamente como fizemos antes, leva a uma expressão simples.

Caso gaussiano com $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$

- Alocar $\mathbf{x}$ a π_1 se:

$$\mathbf{x}^t\mathbf{A}\mathbf{x} - 2\boldsymbol{\beta}^t\mathbf{x} + \alpha \leq 0$$

com $\mathbf{A}$ sendo uma matriz $p \times p$, $\boldsymbol{\beta}$ sendo um vetor coluna $p \times 1$ e α sendo um escalar (um número real). Mais especificamente,

$$\mathbf{A} = \mathbf{\Sigma}_1^{-1} - \mathbf{\Sigma}_2^{-1}$$

$$\boldsymbol{\beta} = \mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2$$

- A constante α é uma expressão um pouco mais longa:

$$\alpha = (\boldsymbol{\mu}_1^t\mathbf{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^t\mathbf{\Sigma}_2^{-1}\boldsymbol{\mu}_2) - 2\log\left(\frac{c(2|\in\pi_1)\mathbb{P}(\pi_1)}{c(1|\in\pi_2)\mathbb{P}(\pi_2)}\right) - 2\log\left(\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|}\right)$$

Caso gaussiano bi-dimensional, com $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$

- No caso em que $\mathbf{x} \in \mathbb{R}^2$, a fronteira ótima de Bayes é uma forma quadrática em x_1 e x_2.
- Isto é, a fronteira de decisão $\lambda(\mathbf{x}) = 0$ (o conjunto de pontos que separa as duas classes) será uma expressão do seguinte tipo:

$$c_1 x_1^2 + c_2 x_2^2 + c_3 x_1 x_2 + c_4 x_1 + c_5 x_2 + c_6 = 0$$

  onde as constantes c_j são determinadas pelos parâmetros das duas distribuições, pelos custos e pelas probabilidades a priori.
- A fronteira $\lambda(\mathbf{x}) = 0$ costuma ser uma curva que lembra o formato de uma parábola (mas não é exatamente uma parábola).
- Ver notebook python para exemplos.

Pros e Cons of Optimal Bayes Classifier

- Precisa conhecer a densidades.
- Se não conhecer, precisa estimá-las e então terá erro de estimacao
- Em principio: estimar via kde (kernel). Entao, a dificuldade será encontrar a regiao em espacos multi-dimensionais.
- Eh otima apenas para ECM. Não otimiza para outras funções objetivo.
- Vantagens: otima; simples; intuitiva;