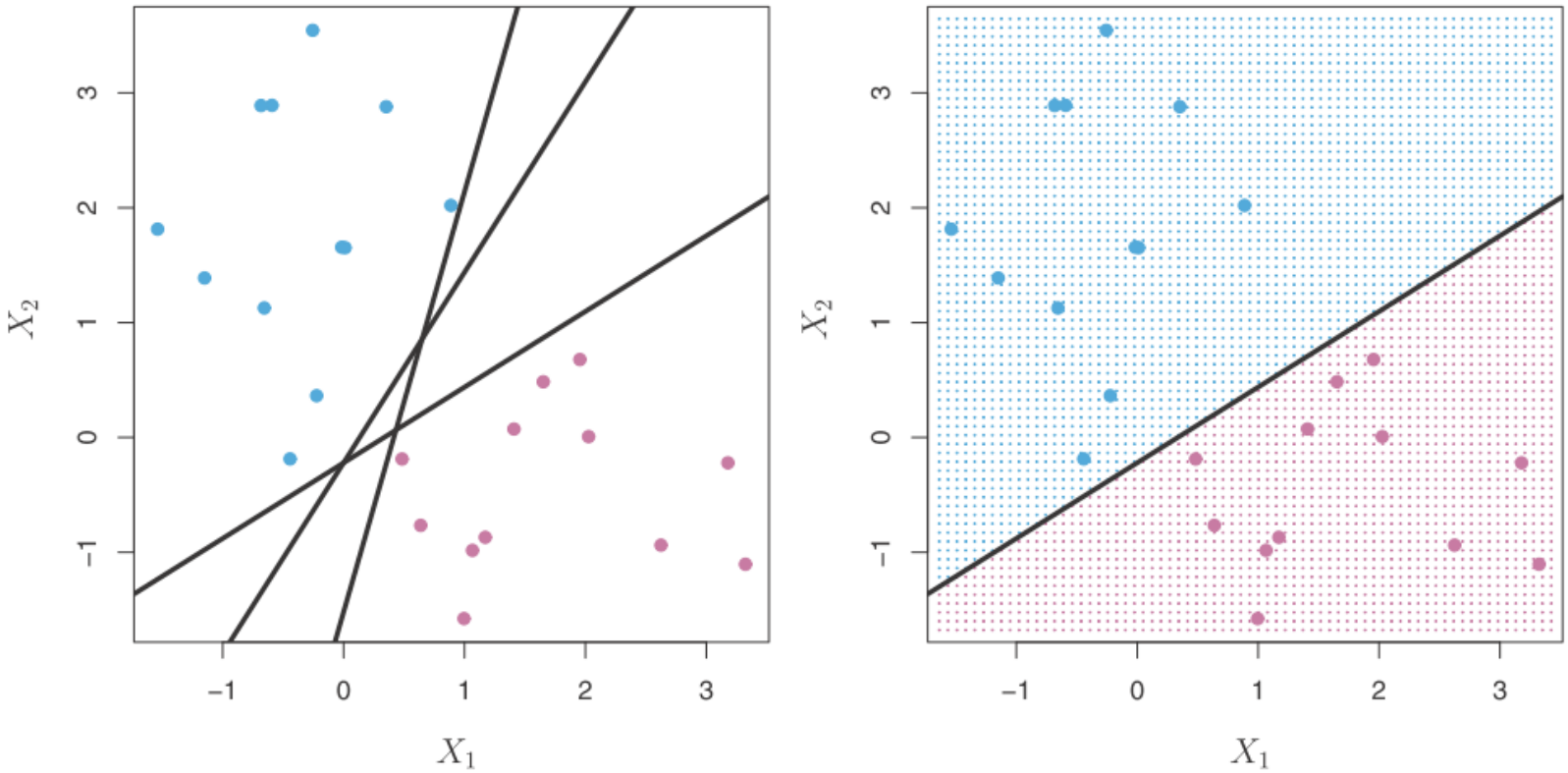
Reconhecimento de Padrões
Teoria da Decisão e Funções Discriminantes

Prof. George von Borries
Departamento de Estatística
Universidade de Brasília

1 - 2024

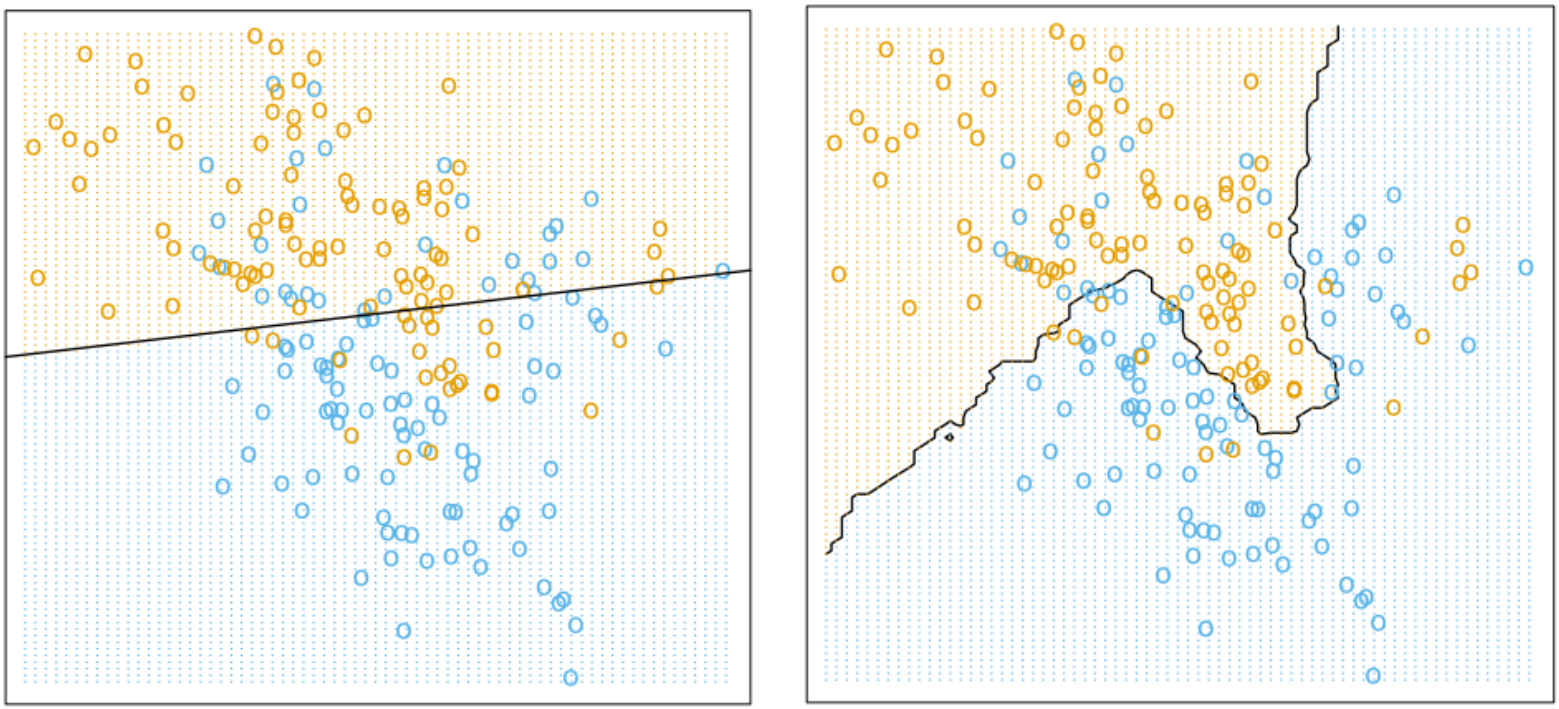
Objetivo

Encontrar funções que melhor separem grupos ou populações.



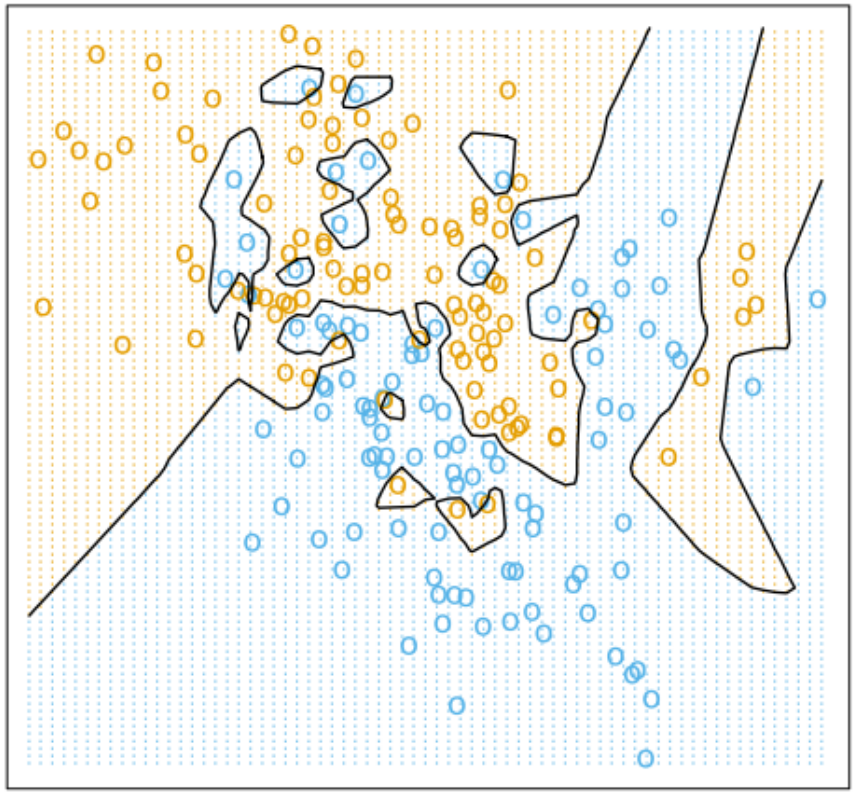
Fonte: James, Witten, Hastie, Tibshirani, 2017.

As funções discriminantes podem ser lineares ou não lineares.



Fonte: Hastie, Tibshirani e Friedman, 2009.

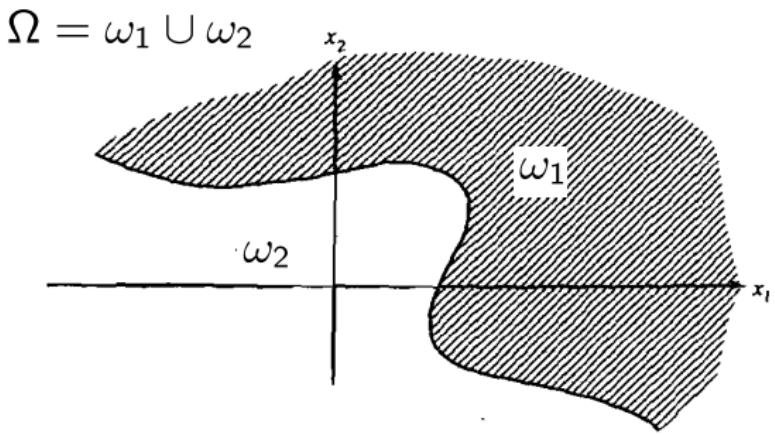
E em alguns casos os discriminantes podem ser mais elaborados.



Fonte: Hastie, Tibshirani e Friedman, 2009.

Elementos de Teoria da Decisão

- Seja Ω o espaço amostral formado pelas classes ω_1 e ω_2 tal que $\Omega = \omega_1 \cup \omega_2$.



- Esta ideia pode ser generalizada para C classes $\omega_1, \ldots, \omega_c$ com probabilidades (a priori) $p(\omega_1), \ldots, p(\omega_c)$ conhecidas.
- Seja $\mathbf{x} = [\mathrm{x}_1 \;\; \mathrm{x}_2 \;\; \ldots \;\; \mathrm{x}_p]$ formado por p características (*features*) e

  ω_j : a classe j que pode ser descrita por $f_j(\mathbf{x}) = p(\mathbf{x}|\omega_j)$, i.e.,
  a função de densidade condicionada a classe ω_j.

1 Regra de Decisão de Bayes para Erro Mínimo

- Na ausência de informação, classificamos um objeto na classe ω_j de maior probabilidade, i.e., se $p(\omega_j) > p(\omega_k)$ para todo $k \neq j$ em $\{1, \ldots, c\}$.
- Na presença de informação $\mathbf{x}$, alocamos $\mathbf{x}$ a classe ω_j se $p(\omega_j|\mathbf{x}) > p(\omega_k|\mathbf{x})$ para qualquer $k \neq j$ em $\{1, \ldots, c\}$.
- As probabilidades a posteriori $p(\omega_j|\mathbf{x})$ podem ser expressas em função de $p(\omega_j)$ e densidades condicionais as classes ω_j, $j \in \{1, \ldots, c\}$, através do Teorema de Bayes, i.e.

$$\mathrm{p}(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)p(\omega_j)}{p(\mathbf{x})}$$

em que $p(\mathbf{x}) = \sum_{i=1}^{c} p(\mathbf{x}|\omega_i)p(\omega_i)$.

- A **Regra de Bayes para Erro Mínimo** aloca $\mathbf{x}$ a classe ω_j se

$$p(\mathbf{x}|\omega_j)p(\omega_j) > p(\mathbf{x}|\omega_k)p(\omega_k)$$

para todo $k \neq j$ em $\{1, \ldots, c\}$.

- Note que a mesma regra pode ser escrita como

$$\frac{\mathrm{p}(\omega_j|\mathbf{x})}{\mathrm{p}(\omega_k|\mathbf{x})} > 1,$$

para todo $k \neq j$ em $\{1, \ldots, c\}$.

- Ou também

$$\frac{p(\mathbf{x}|\omega_j)}{p(\mathbf{x}|\omega_k)} > \frac{p(\omega_k)}{p(\omega_j)}$$

para todo $k \neq j$ em $\{1, \ldots, c\}$.

- Para duas classes,

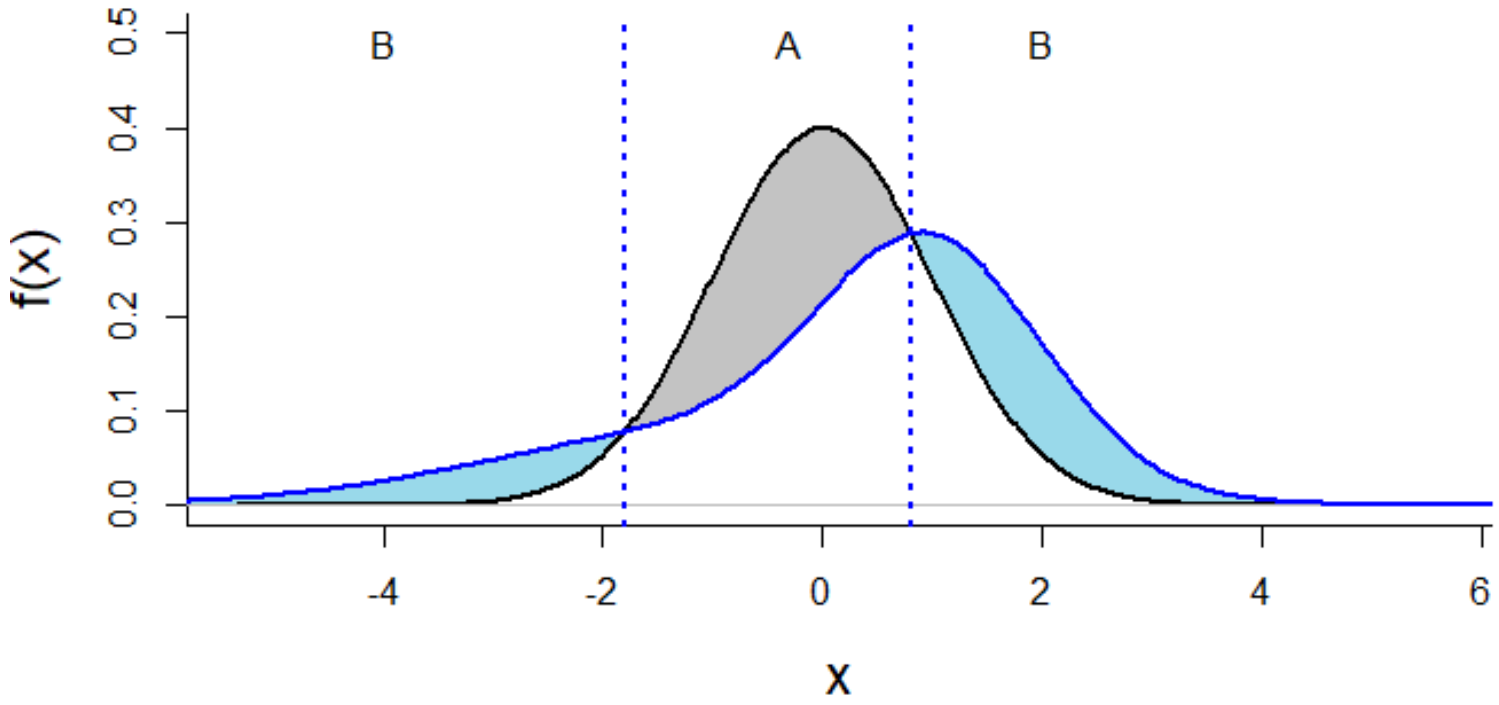
$$\ell_r(\mathbf{x}) = \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{p(\omega_2)}{p(\omega_1)}$$

implica em $\mathbf{x} \in \omega_1$.

Exemplo:

Duas populações com densidades
$f_1(\mathbf{x}) \sim \mathrm{N}(0,1)$ e $f_2(\mathbf{x}) \sim 0.6\mathrm{N}(1,1) + 0.4\mathrm{N}(-1,2)$.

Assumindo $p(\omega_1) = p(\omega_2)$, classificamos como $f_1(\mathbf{x})$ toda observação no intervalo A e como $f_2(\mathbf{x})$ toda observação nos intervalos B.



- A regra de decisão

$$p(\mathbf{x}|\omega_j)p(\omega_j) > p(\mathbf{x}|\omega_k)p(\omega_k)$$

para todo $k \neq j$ em $\{1, \ldots, c\}$, minimiza o erro de alocação. Note que

$$p(\text{erro}) = \sum_{i=1}^{c} p(\text{erro}|\omega_i)p(\omega_i) \quad \text{e} \quad p(\text{erro}|\omega_\text{i}) = \int_{\bar{\Omega}_i} p(\mathbf{x}|\omega_i)d\mathbf{x}.$$

Assim,

$$\begin{aligned} p(\text{erro}) &= \sum_{i=1}^{c} \int_{\bar{\Omega}_i} p(\mathbf{x}|\omega_i)p(\omega_i)d\mathbf{x} = \sum_{i=1}^{c} p(\omega_i)\left(1 - \int_{\Omega_i} p(\mathbf{x}|\omega_i)d\mathbf{x}\right) \\ &= 1 - \sum_{i=1}^{c} p(\omega_i) \int_{\Omega_i} p(\mathbf{x}|\omega_i)d\mathbf{x} \qquad (1) \end{aligned}$$

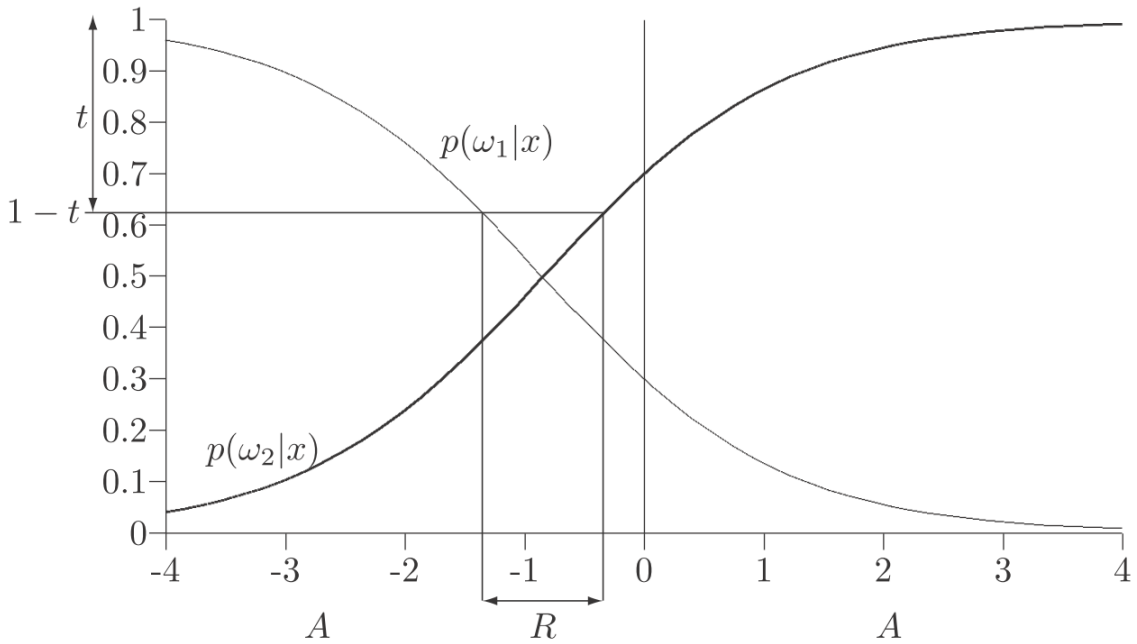
Minimizar (1) é o mesmo que maximizar $\sum_{i=1}^{c} p(\omega_i) \int_{\Omega_i} p(\mathbf{x}|\omega_i)d\mathbf{x}$, i.e, a probabilidade de uma classificação correta. O erro de Bayes é então

$$\boxed{e_\text{B} = 1 - \int \max_i [p(\omega_i)p(\mathbf{x}|\omega_i)]d\mathbf{x}}$$

2 Regra de Decisão de Bayes quando existe incerteza

- Em regiões próximas de limites de decisão existe incerteza sobre a classe ω_i que um objeto deve ser alocado. Neste caso, é possível adiar uma decisão de forma a reduzir o erro de alocação.
- Considere R a região de rejeição (incerteza sobre classificação ou não decisão) e A a região de aceitação (ou classificação). Definindo t como ponto limite (*threshold*),

$$R = \left\{ \mathbf{x} \mid 1 - \max_i p(\omega_i|\mathbf{x}) > t \right\} \quad \text{e} \quad A = \left\{ \mathbf{x} \mid 1 - \max_i p(\omega_i|\mathbf{x}) \leq t \right\}$$



- Considerando c como o número de classes, pode-se mostrar que se $1 - t \leq \frac{1}{c}$ ou $t \geq \frac{c-1}{c}$, a região R será vazia. ($\checkmark$)
- A probabilidade de classificação correta, $c(t)$, é uma função do limiar t,

$$c(t) = \int_A \max_i [p(\omega_i)p(\mathbf{x}|\omega_i)]d\mathbf{x}$$

- e a probabilidade de rejeitar (incerteza) será

$$r(t) = \int_R p(\mathbf{x})d\mathbf{x}.$$

- Assim, a taxa de erro será $e(t) = 1 - c(t) - r(t)$.

3 Regra de Decisão de Bayes para Risco Mínimo

- Em vez de minimizar a probabilidade de cometer um erro, podemos considerar a minimização do risco. Neste caso existe um custo associado ao erro de classificação.

Exemplo:

Considere um sensor que faz um automóvel reduzir a velocidade e até parar quando identifica um obstáculo a frente.

Caso 1: Automóvel circulando em vias de baixa velocidade.
O custo do erro de identificação de um obstáculo (quando não existe) é menor que o custo de erro de não identificação do obstáculo.

Caso 2: Automóvel circulando em vias de alta velocidade.
O custo do erro de identificação de um obstáculo (quando não existe) é maior que o custo de erro de não identificação do obstáculo.

- Seja λ_{ij} = custo de alocar um padrão $\mathbf{x}$ a ω_i quando $\mathbf{x} \in \omega_j$.
  λ_{ij} pode ser um valor monetário, tempo, avaliação de qualidade, etc. Geralmente é atribuído de forma subjetiva por um especialista.
- O **risco condicional** de alocar um padrão $\mathbf{x}$ a ω_i é definido por

$$r_c^i(\mathbf{x}) = \sum_{j=1}^{c} \lambda_{ji} p(\omega_j|\mathbf{x}).$$

- O **risco médio** na região Ω_i é

$$r_m^i = \int_{\Omega_i} r_c^i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \int_{\Omega_i} \sum_{j=1}^{c} \lambda_{ij} p(\omega_j|\mathbf{x}) p(\mathbf{x}) d\mathbf{x},$$

- e o risco total ou custo esperado será

$$r_t = \sum_{i=1}^{c} r_m^i = \sum_{i=1}^{c} \int_{\Omega_i} \sum_{j=1}^{c} \lambda_{ji} p(\omega_j|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}.$$

- O risco total

$$r_t = \sum_{i=1}^{c} r_m^i = \sum_{i=1}^{c} \int_{\Omega_i} \sum_{j=1}^{c} \lambda_{ji} p(\omega_j|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

é minimizado se a região Ω_i é escolhida de tal forma que

$$\sum_{j=1}^{c} \lambda_{ji} p(\omega_j|\mathbf{x}) p(\mathbf{x}) \leq \sum_{j=1}^{c} \lambda_{jk} p(\omega_j|\mathbf{x}) p(\mathbf{x}) \quad k = 1, \ldots, c$$

tal que $\mathbf{x} \in \Omega_i$.

- A Regra de decisão de Bayes para risco mínimo será

$$r^* = \int_{\mathbf{x}} \min_{i=1,\ldots,c} \lambda_{ji} p(\omega_j|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}.$$

- No caso de custos iguais, em que $\lambda_{ij} = 1$ se $i \neq j$ e $\lambda_{ij} = 0$ se $i = j$, a regra de decisão será alocar $\mathbf{x}$ a classe ω_i se

$$\sum_{j=1}^{c} p(\omega_j|\mathbf{x}) p(\mathbf{x}) - p(\omega_i|\mathbf{x}) p(\mathbf{x}) \leq \sum_{j=1}^{c} p(\omega_j|\mathbf{x}) p(\mathbf{x}) - p(\omega_k|\mathbf{x}) p(\mathbf{x})$$

para todo $k \in \{1, \ldots, c\}$, i.e.,

$$p(\mathbf{x}|\omega_i) p(\omega_i) \geq p(\mathbf{x}|\omega_k) p(\omega_k) \quad k = 1, \ldots, c.$$

4 Regra de Decisão de Neyman-Pearson

- Considere uma classe de interesse ω_1 e outra classe ω_2.

  Dois erros podem ser cometidos:

  $\epsilon_1 = \int_{\Omega_2} p(\mathbf{x}|\omega_1)d\mathbf{x} =$ probabilidade de erro do Tipo I

  $\epsilon_2 = \int_{\Omega_1} p(\mathbf{x}|\omega_2)d\mathbf{x} =$ probabilidade de erro do Tipo II
- A regra de decisão de Neyman-Pearson deseja minimizar o erro ϵ_1 sujeito a ϵ_2 fixo $(= \epsilon_0)$.
- Em algumas áreas, chamamos de ω_1 a classe com evento positivo e ω_2 a classe com evento negativo. Neste caso ϵ_1 é taxa de falsos negativos (FNR) e ϵ_2 a taxa de falsos positvos (FPR).

Exemplo: Detecção de Sinais em Radar

Detecção de sinal ω_1 na presença de ruído ω_2.
ϵ_1 = sinal disponível mas identificado como ruído (*missed detection*).
ϵ_2 = ruído é identificado como sinal (*false alarm*).
Procura-se minimizar

$$r = \int_{\Omega_2} p(\mathbf{x}|\omega_1)d\mathbf{x} + \lambda \left\{ \int_{\Omega_1} p(\mathbf{x}|\omega_2)d\mathbf{x} - \epsilon_0 \right\}$$

em que λ é o multiplicador de Lagrange e ϵ_0 a a taxa especificada de falsos alarmes.

- r pode ser escrito como

$$r = (1 - \lambda\epsilon_0) + \int_{\Omega_1} \{\lambda p(\mathbf{x}|\omega_2)d\mathbf{x} - p(\mathbf{x}|\omega_1)d\mathbf{x}\}$$

minimizado se escolhermos Ω_1 tal que o integrando seja negativo, i.e.,

$$\lambda p(\mathbf{x}|\omega_2)d\mathbf{x} - p(\mathbf{x}|\omega_1) < 0,$$

sendo assim $\mathbf{x} \in \Omega_1$.

- r será minimizado se escolhermos Ω_1 tal que o integrando seja negativo, i.e.,

$$\lambda p(\mathbf{x}|\omega_2)d\mathbf{x} - p(\mathbf{x}|\omega_1) < 0,$$

  sendo assim $\mathbf{x} \in \Omega_1$.

- Em termos de razão de verossimilhança,

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \lambda$$

  quando $\mathbf{x} \in \Omega_1$.

- λ é escolhido em termos da taxa especificada ϵ_0 de falsos alarmes (erro do Tipo II). Geralmente λ é obtido numericamente.

- A performance da regra de decisão pode ser resumida pela **Curva ROC** ou Curva Característica de Operação (*Receiver Operating Characteristic*).

Curva ROC (Receiver Operating Characteristic): é um gráfico da sensibilidade em função de (1 − Especificidade) para diferentes valores de λ[1].

Sensibilidade (Se): Taxa de Verdadeiros Positivos (TPR), i.e., a probabilidade de um teste ser positivo, dado que o evento de interesse ocorre.

TPR indicando se o teste é sensível para a detecção do sinal (evento de interesse).

$1 - \mathrm{TPR} = \mathrm{FPR}$ (Taxa de Falsos Positivos) é a taxa de sinais não detectados.

Especificidade (Es): é a Taxa de Verdadeiros Negativos (TNR), i.e., a probabilidade do teste ser negativo, dado que o evento de interesse não ocorre.

TNR indicando se o teste é específico para o sinal em questão.

$1 - \mathrm{TNR} = \mathrm{FNR}$ (Taxa de Falsos Negativos) é a taxa de falsos alarmes.

[1] Esta curva resume o poder preditivo do teste e foi desenvolvida na 2^a Guerra Mundial para detecção de sinais.

Considere a seguinte tabela de contingência 2×2 para identificação de um sinal (avião detectado) ou ruído (interferência de uma ave).

		Classe	
		ω_1 (Ex. Sinal)	ω_2 (Ex. Ruído)
Classificação	$\hat{\omega}_1$	$n_{\text{TP}} = n(\hat{\omega}_1 \vert \omega_1)$	$n_{\text{FP}} = n(\hat{\omega}_1 \vert \omega_2)$
	$\hat{\omega}_2$	$n_{\text{FN}} = n(\hat{\omega}_2 \vert \omega_1)$	$n_{\text{TN}} = n(\hat{\omega}_2 \vert \omega_2)$
		n_1	n_0

sendo a classe de interesse (no caso sinal) considerada como evento positivo, temos

n_{TP} e n_{FP}: número de verdadeiros e falsos positivos e
n_{FN} e n_{TN}: número de falsos e verdadeiros negativos.

Assim,

$$\begin{aligned}
\text{Sensibilidade} &= n_{\text{TP}}/n_1 = p(\hat{\omega}_1 \mid \omega_1) = p(\hat{\text{Y}} = 1 \mid \text{Y} = 1); \\
\text{Especificidade} &= n_{\text{TN}}/n_0 = p(\hat{\omega}_2 \mid \omega_2) = p(\hat{\text{Y}} = 0 \mid \text{Y} = 0); \\
\text{1 - Especificidade} &= 1 - n_{\text{TN}}/n_0 = n_{\text{FP}}/n_0 = p(\hat{\omega}_1 \mid \omega_2) = p(\hat{\text{Y}} = 1 \vert \text{Y} = 0).
\end{aligned}$$

A discriminação perfeita implica modelo com $n_{\text{TP}} = n_1$ e $n_{\text{TN}} = n_0$, i.e., $\text{Se} = \text{Es} = 1$.

Seja $0 \leq \lambda \leq 1$, um ponto de corte tal que:

- $\hat{\pi}_i > \lambda \Rightarrow$ predizer resposta como sinal (ω_1) ou
- $\hat{\pi}_i \leq \lambda \Rightarrow$ predizer resposta como ruído (ω_2),

para $i = 1, \ldots, n$.

Etapas da obtenção da curva ROC: considerando valores de $\lambda \in (0, 1)$[2],

1. Realizar a alocação para cada $\mathbf{x}_i$;
2. Obter $\hat{\pi}_i$, $i = 1, \ldots, n$;
3. Comparar $\hat{\pi}_i$ com λ e classificar a observação como sinal ou ruído.
4. Montar a tabela de classificação de acordo com os valores reais e preditos para $i = 1, \ldots, n$.
5. Obter Se, Es e $1 - $ Es.
6. Construir a curva ROC plotando os valores em (5.) para cada λ.

[2] Note que: $\lambda = 0$ significa classificar todos os eventos como ω_1 e $\lambda = 1$ significa classificar todos os eventos como ruído.

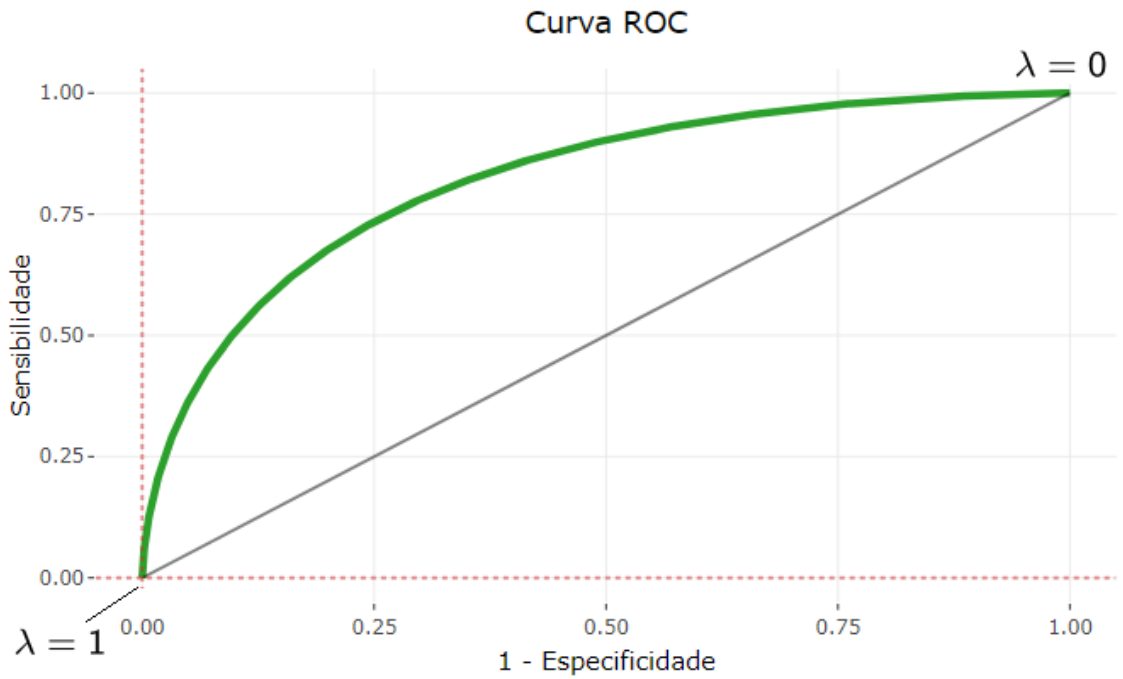
- Casos extremos (para qualquer modelo):

$$\begin{array}{lllllll} \lambda = 1: & n_{TP} = 0 & n_{FP} = 0 & n_{FN} = n_1 & n_{TN} = n_0 & \Rightarrow \text{Se} = 0 & \text{Es} = 1 \\ \lambda = 0: & n_{TP} = n_1 & n_{FP} = n_0 & n_{FN} = 0 & n_{TN} = 0 & \Rightarrow \text{Se} = 1 & \text{Es} = 0 \end{array}$$

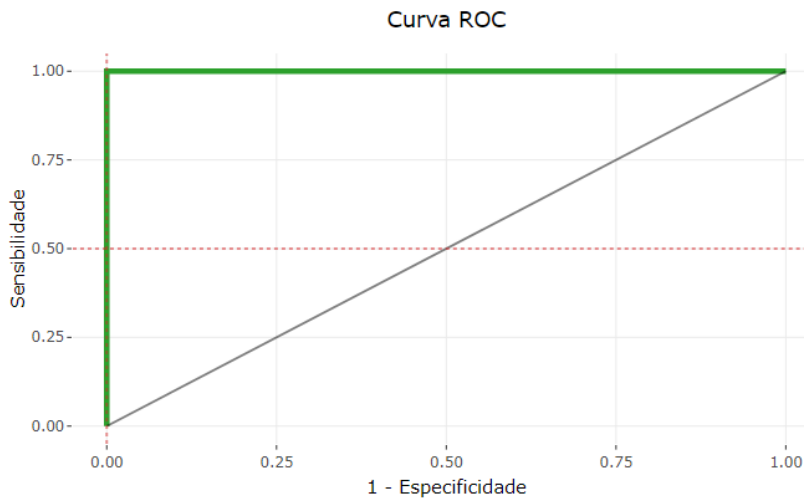
- Sensibilidade e (1 - Especificidade) tem foco nos casos previstos, i.e.,

$$\begin{aligned} p(\hat{\omega}_1 \mid \omega_1) &= Se \\ p(\hat{\omega}_1 \mid \omega_2) &= 1 - p(\hat{\omega}_2 \mid \omega_2) = 1 - Es \end{aligned}$$

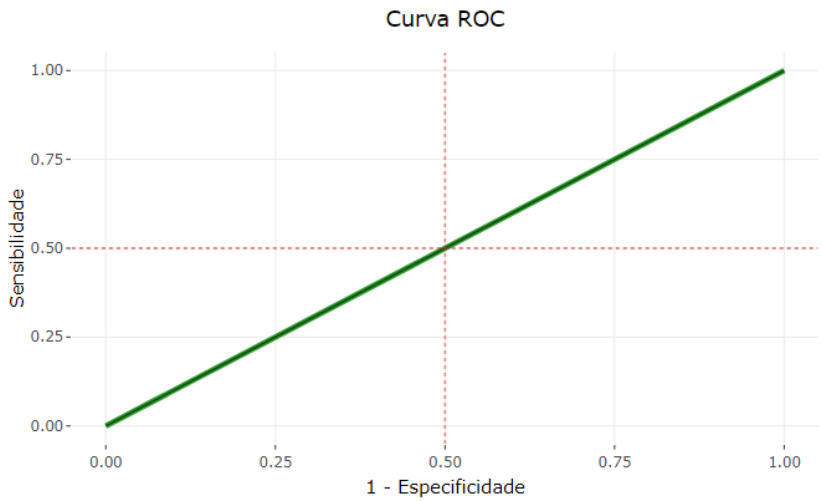
- Uma boa discriminação implica em $Se > 1 - \text{Es}$.
- A curva ROC considera Se e $(1 - \text{Es})$ para diferentes valores de λ



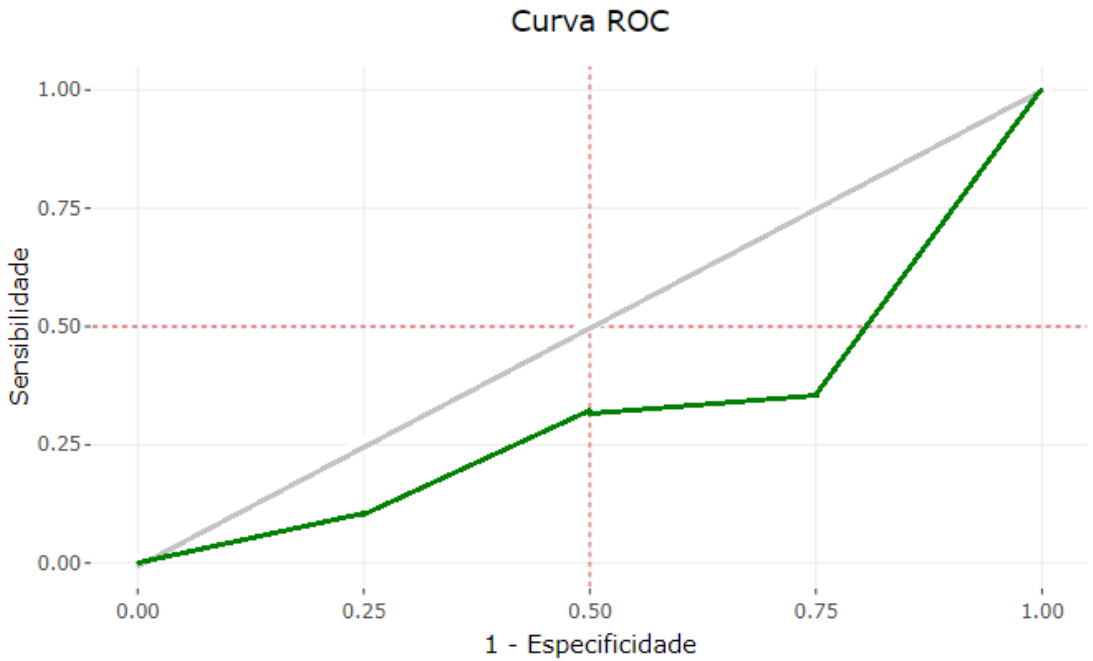
- Modelo com discriminação perfeita:



- Modelo sem poder de discriminação:

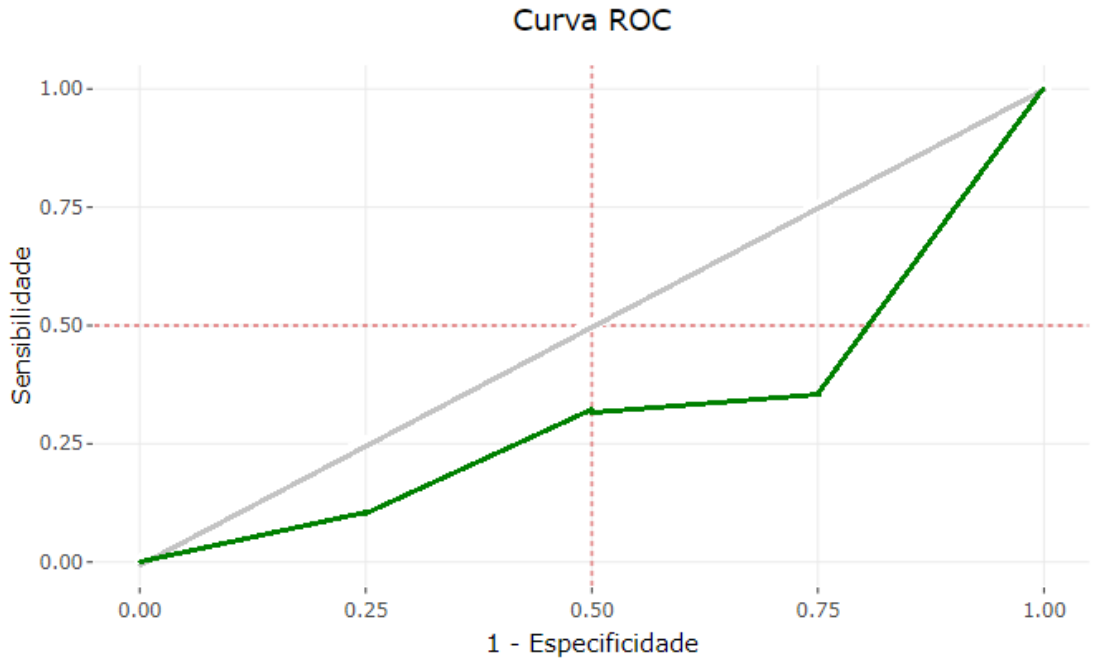


Pergunta 1: É possível obter a seguinte curva ROC?



Resposta:

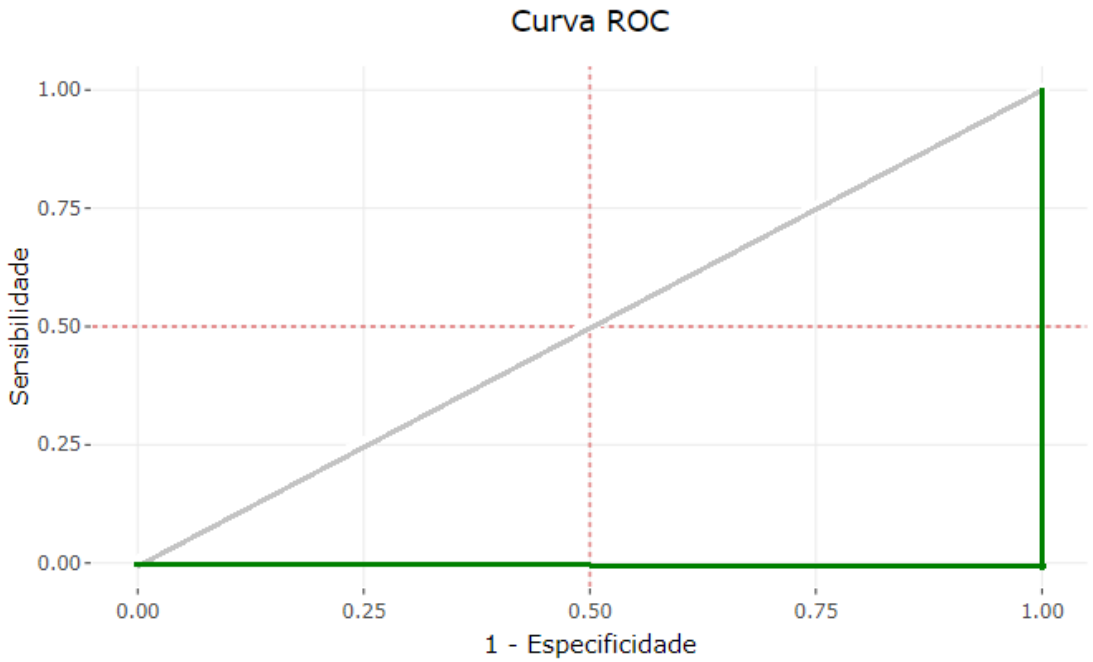
Pergunta 1: É possível obter a seguinte curva ROC?



Resposta:

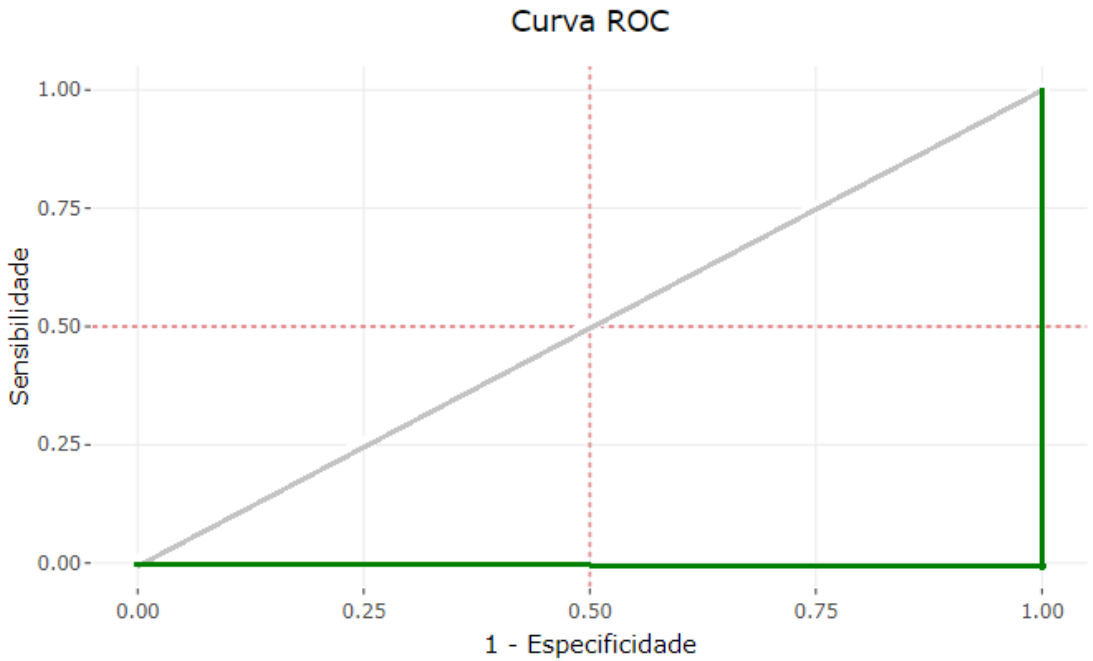
Sim. Neste caso, $p(\hat{\omega}_2 \mid \omega_2) > p(\hat{\omega}_1 \mid \omega_1)$, i.e., ruído é melhor identificado que sinal. Um resultado que não é de muita utilidade.

Pergunta 2: É possível obter a seguinte curva ROC?



Resposta:

Pergunta 2: É possível obter a seguinte curva ROC?



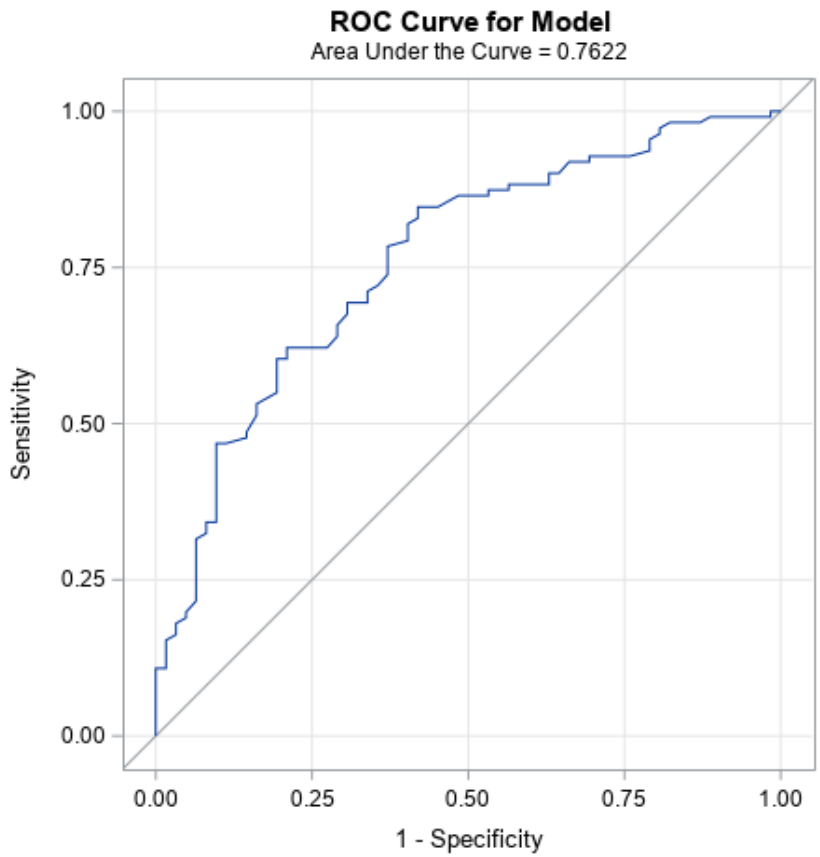
Resposta:

Apesar de ser um resultado possível, não é um resultado razoável. Isto porque, neste caso, todas as observações seriam previstas como ruídos. Não precisamos de um modelo de classificação para esta situação!

AUC (*Area Under Curve*):

- Na comparação de modelos podemos utlizar a área abaixo da curva ROC (AUC) como medida de qualidade.

Exemplo:



Referência:

- Artigo: von Borries G. e Quadros, A.V.C. (2022) ROC App: An Application to Understand ROC Curves. Brazilian Journal of Biometrics. ISSN: 2764-5290.

  Aplicativo para auxiliar a compreensão da Curva ROC.

  - Artigo: https://biometria.ufla.br/index.php/BBJ/article/view/566/339
  - Aplicativo Shiny: https://gfvonborries.shinyapps.io/roc_app/
  - Código R: https://github.com/GvBorries/ROCApp

Funções Discriminantes

- **Classificação através de funções densidade:** a teoria da decisão via regra de Bayes exige conhecimento da densidade $p(\mathbf{x}|\omega_i)$ através de
  - Estimação paramétrica de densidades. Exemplo: estimação dos parâmetros da distribuição normal (suposição) em cada uma das classes, através dos dados observados.
  - Estimação não paramétrica de densidades. Exemplo: estimação de densidade via núcleos (*kernel density estimation*).

- **Classificação através de funções discriminantes:** neste caso são feitas suposições sobre a função discriminante. Uma função discrimininante é uma função do padrão de $\mathbf{x}$ que resulta na regra de classificação.

Exemplo:

Função discriminante $h(\mathbf{x})$ tal que
$h(\mathbf{x}) > k \Rightarrow x \in \omega_1$ e
$h(\mathbf{x}) < k \Rightarrow x \in \omega_2$.
$h(\mathbf{x}) = k \Rightarrow$ padrão designado por sorteio, no caso de duas classes.

- Asim como no caso da regra de decisão de Bayes para erro mínimo, uma função discriminante ótima para duas classes implica

$$h(\mathbf{x}) = \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)}$$

com $k = p(\omega_2)/p(\omega_1)$.

- As funções discriminantes não são únicas, i.e., se f é uma função monótona,
$g(\mathbf{x}) = f(h(\mathbf{x})) > k' \Rightarrow x \in \omega_1 \quad \text{e} \quad g(\mathbf{x}) = f(h(\mathbf{x})) < k' \Rightarrow x \in \omega_2.$
- Para c classes, o padrão é alocado a classe com maior discriminante, i.e.,

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \Rightarrow \mathbf{x} \in \omega_i \quad j = 1, \ldots, c,\ j \neq i.$$

- A diferença básica para a teoria da decisão é que concentramos na forma da função discriminante e não na imposição de uma distribuição (mas que pode ocorrer mesmo assim).

Exemplo:

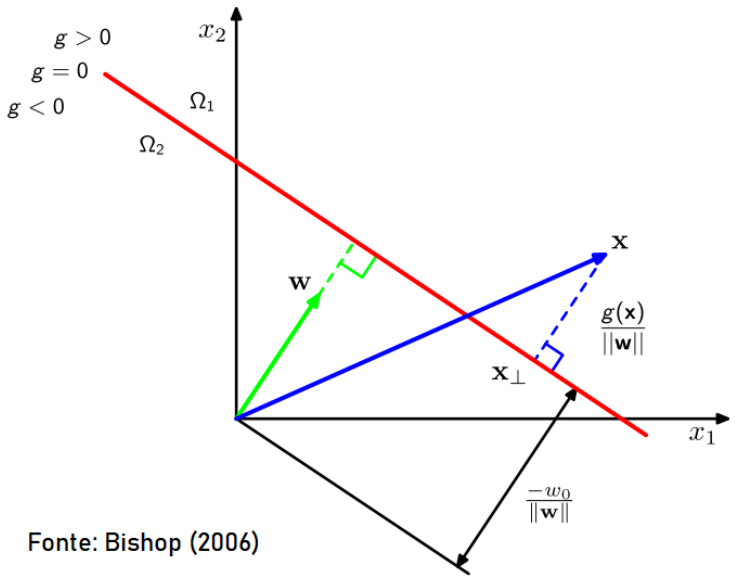
Não assume uma distribuição específica:
Função Discriminante de Fisher para $g > 2$ populações.
Assume populações normais:
Discriminante Linear para duas populações $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

1 Função Discriminante Linear

- Seja $\mathbf{x} = \{x_1, \ldots, x_p\}^T$. Uma função discriminante linear tem forma

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \sum_{i=1}^{p} w_i x_i + w_0.$$

- Discriminante linear com vetor de pesos $\mathbf{w}$ e limite (*threshold*) w_0.
- Representa a equação de um hiperplano com vetor unitário na direção de $\mathbf{w}$ e distância da origem igual a $|w_0|/||\mathbf{w}||$, perpendicular ao hiperplano.
- O valor da função discriminante para o padrão $\mathbf{x}$ é uma medida da distância de $\mathbf{x}$ perpendicular ao hiperplano.



Fonte: Bishop (2006)

Exemplo: Discriminante Linear para duas Populações $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \quad i = 1, 2.$$

Então

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}}{\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\}}$$

e

$$\begin{aligned} \ell(\mathbf{x}) = \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \overbrace{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}}^{\text{Disc. Linear de Fisher}} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \\ &= \mathbf{w}^{\mathsf{T}}\mathbf{x} + w_0. \end{aligned}$$

$\Rightarrow$ Se $\ell(\mathbf{x}) > 0$ alocar $\mathbf{x}$ a π_1.

Exemplo: Classificador de Distância Mínima

Suponha um grupo de pontos (*prototype points*) $\mathbf{p}_1, \ldots, \mathbf{p}_c$ representantes das classses $\omega_1, \ldots, \omega_c$.

Para cada ponto, a distância Euclideana é $|\mathbf{x} - \mathbf{p}_i|^2 = \mathbf{x}^\mathsf{T}\mathbf{x} - 2\mathbf{x}^\mathsf{T}\mathbf{p}_i + \mathbf{p}_i^\mathsf{T}\mathbf{p}_i$.

A distância mínima de classificação é obtida pelo maior $\mathbf{x}^\mathsf{T}\mathbf{p}_i - \frac{1}{2}\mathbf{p}_i^\mathsf{T}\mathbf{p}_i = \mathbf{w}_i^\mathsf{T}\mathbf{x} + w_{i0}$ (função discriminante linear).

Nota: $\mathbf{p}_i$ pode ser a média da classe i.

2 Função Discriminante Linear em Partes (*piecewise*)

- Muitas vezes não é possível separar as classes com uma função linear.

(a) (b)

- Suponha que existem n_i *prototypes* na classe ω_i, $\mathbf{p}_i^1, \ldots, \mathbf{p}_i^{n_i}$, $i = 1, \ldots, c$.
- A função discriminante da classe ω_i será

$$g_i(\mathbf{x}) = \max_{j=1,\ldots,n_i} g_i^j(\mathbf{x}) = \max_{j=1,\ldots,n_i} \left\{ \mathbf{x}^T \mathbf{p}_i^j - \frac{1}{2} \mathbf{p}_i^{j^T} \mathbf{p}_i^j \right\} \quad j = 1, \ldots, n_i; \quad i = 1, \ldots, c.$$

- O padrão $\mathbf{x}$ é alocado a classe com maior $g_i(\mathbf{x})$.
- Este método divide o espaço em $\sum_{i=1}^{c} n_i$ regiões que são conhecidas como "mosaico"(*tesselation*) de Dirichelet.
- A regra de decisão do vizinho mais próximo é um caso particular em que cada padrão no grupo de treinamento é considerado um vetor *prototypes*.

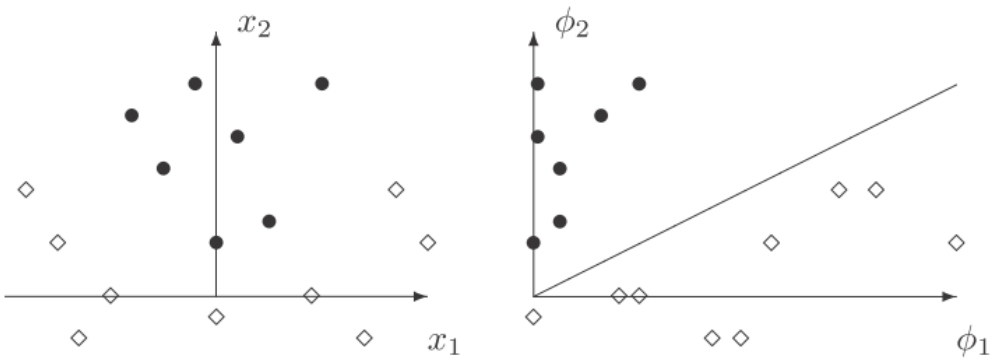
3 Função Discriminante Linear Generalizada (*phi machine*)

- Função discriminante da forma

$$g(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\phi + w_0$$

  sendo $\phi = (\phi_1(\mathbf{x}), \ldots, \phi_D(\mathbf{x}))^{\mathsf{T}}$ uma função vertorial de $\mathbf{x}$.
- Isto define a função generalizada, que é função de transformações de transformaões das variáveis originais. Se, por exemplo, $\phi_i(\mathbf{x}) = x_i$ e $D = p$, então a função se reduz a função linear discriminante.

Exemplo: Classificador de Distância Mínima



Exemplo de transformação não linear que permite a discriminação linear. No caso, $\phi_1(\mathbf{x}) = x_1^2$ e $\phi_2(\mathbf{x}) = x_2$.

Table 1.1 Discriminant functions, ϕ.

Discriminant function	Mathematical form, $\phi_i(\boldsymbol{x})$
Linear	$\phi_i(\boldsymbol{x}) = x_i, i = 1, \ldots, p$
Quadratic	$\phi_i(\boldsymbol{x}) = x_{k_1}^{l_1} x_{k_2}^{l_2}, i = 1, \ldots, (p+1)(p+2)/2 - 1$ $l_1, l_2 = 0$ or 1; $k_1, k_2 = 1, \ldots, p$ l_1, l_2 not both zero
νth order polynomial	$\phi_i(\boldsymbol{x}) = x_{k_1}^{l_1} \ldots x_{k_\nu}^{l_\nu}, i = 1, \ldots, \binom{p+\nu}{\nu} - 1$ $l_1, \ldots, l_\nu = 0$ or 1; $k_1, \ldots, k_\nu = 1, \ldots, p$ l_i not all zero
Radial basis function	$\phi_i(\boldsymbol{x}) = \phi(\lvert\boldsymbol{x} - \boldsymbol{v}_i\rvert)$ for centre $\boldsymbol{v}_i$ and function ϕ
Multilayer perceptron	$\phi_i(\boldsymbol{x}) = f(\boldsymbol{x}^T \boldsymbol{v}_i + v_{i0})$ for direction $\boldsymbol{v}_i$ and offset v_{i0}. f is the logistic function, $f(z) = 1/(1 + \exp(-z))$

(Webb e Copsey, 2011)