Mineração de Dados

Aula 7: Classificação

Rafael Izbicki

Revisão

Um problema de classificação é um problema de predição em que Y é qualitativo.

Em um problema de classificação, é comum se usar

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}(Y \neq g(\mathbf{X})),$$

para medir o risco de um classificador g.

A função g que minimza $R(g)$ é dada por

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \mathbb{P}(Y = c|\mathbf{x})$$

Tal classificador é conhecido como classificador de Bayes (não confundir com teorema de Bayes)

Como no caso de regressão, não conhecemos tal g, mas podemos estimá-la usando

$$\widehat{g}(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{\mathbb{P}}(Y = c|\mathbf{x})$$

Como no caso de regressão, não conhecemos tal g, mas podemos estimá-la usando

$$\widehat{g}(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{\mathbb{P}}(Y = c|\mathbf{x})$$

A estimativa $\widehat{\mathbb{P}}(Y = c|\mathbf{x})$ pode ser obtida usando-se, por exemplo, regressão logística ou regressão linear.

Como no problema de regressão, não assumimos que esta relação (logística ou linear) é válida, mas apenas que ela criar bons classificadores.

Vimos também uma outra abordagem.

Esta consiste em usar o Teorema de Bayes:

$$\mathbb{P}(Y = c|\mathbf{x}) = \frac{f(\mathbf{x}|Y = c)\mathbb{P}(Y = c)}{\sum_{s \in \mathcal{X}} f(\mathbf{x}|Y = s)\mathbb{P}(Y = s)}$$

Estimamos então tanto $f(\mathbf{x}|Y = c)$ quanto $\mathbb{P}(Y = c)$.

A probabilidade $\mathbb{P}(Y = c)$ pode ser facilmente estimada utilizando-se as proporções amostrais de cada categoria de Y.

Já para estimar $f(\mathbf{x}|Y = c)$, é comum fazermos alguma suposição sobre esta distribuição.

Aqui vimos o Naive Bayes, que assume independência condicional das covariáveis:

$$f(\mathbf{x}|Y = s) = f(x_1, \ldots, x_p|Y = s) = \prod_{j=1}^{d} f(x_j|Y = s),$$

Finalmente, vimos que podemos selecionar modelos utilizando-se validação cruzada.

Para tanto, podemos utilizar o seguinte estimador do risco:

$$\widehat{R}(g) := \frac{1}{m} \sum_{k=1}^{m} \mathbb{I}(Y_k' \neq g(\mathbf{X}_k')),$$

Na aula de hoje vamos estudar dois novos métodos para estimar g.

Veremos também outras formas de avaliar o quão razoável uma função g é.

Antes disso . . .

O que muda quando há várias categorias?

Quando temos várias categorias, $g(\mathbf{x})$ ótimo é dado por

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \mathbb{P}(Y = c|\mathbf{x})$$

Para cada c, podemos estimar $\mathbb{P}(Y = c|\mathbf{x})$ usando, e.g, uma regressão logística.

Para tanto, basta estimar $\mathbb{P}(Z = 1|\mathbf{x})$, onde $Z = \mathbb{I}(Y = c)$

Assim, fazemos $|\mathcal{C}|$ regressões logísticas, e então definimos

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{\mathbb{P}}(Y = c|\mathbf{x})$$

Claro, podemos usar outro método que não um regressão logística para estimar cada uma dessas probabilidades.

Como avaliar quão bom g é?

Vimos que a função de risco é dada por

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}\left(Y \neq g(\mathbf{X})\right),$$

Nem sempre tal função nos traz toda informação sobre g.

Um exemplo: suponha que Y indica se uma pessoa tem uma certa **doença rara**, e que, portanto, na nossa amostra i.i.d., temos poucos pacientes com $Y = 1$.

O classificador $g(x) \equiv 0$ terá erro baixo, mas sua performance deixa a desejar.

Na prática, para evitar esse tipo de situação, é comum considerar matrizes de confusão. Ex:

Examplo de matriz de confusão

	Valor verdadeiro	
Valor Predito	Y=0	Y=1
Y=0	VN	FN
Y=1	FP	VP

V: Verdadeiro / F: Falso
P: Positivo / N: Negativo

Com base nessa tabela, define-se

- Sensibilidade: VP/(VP+FN) (dos pacientes doentes, quantos foram corretamente identificados?)
- Especificidade: VN/(VN+FP) (dos pacientes não doentes, quantos foram corretamente identificados?)

	Valor verdadeiro	
Valor Predito	Y=0	Y=1
Y=0	VN	FN
Y=1	FP	VP

V: Verdadeiro / F: Falso
P: Positivo / N: Negativo

- Sensibilidade: VP/(VP+FN) (dos pacientes doentes, quantos foram corretamente identificados?)
- Especificidade: VN/(VN+FP) (dos pacientes não doentes, quantos foram corretamente identificados?)

Para o classificador $g(x) \equiv 0$, temos

Sensibilidade=0; Especificidade=1

Apesar da especificidade ser alta, a sensibilidade é muito baixa. Isso indica que o classificador na realidade é ruim. (obs: a soma dessas duas quantidades não é necessariamente 1).

Assim, na prática é recomendável olhar para sensibilidade e especificidade simultaneamente ao risco estimado.

Atenção: é importante calcular os valores de VP, FN, VN e FP usando-se uma amostra de teste ou validação.

Um outro problema relacionado a isso: se $Y = 1$ é raro, em geral teremos $\mathbb{P}(Y = 1|x)$ baixo.

Assim, usar um corte de $1/2$ pode não ajudar: a regra $g(x) = \mathbb{I}(\mathbb{P}(Y = 1|x) \geq 1/2)$ nos levará a sempre decidir por $Y = 0$, mesmo se essas probabilidades estiverem bem estimadas.

Na prática, para evitar isso é comum buscar cortes diferentes de $1/2$, i.e., buscam-se regras do tipo

$$g(x) = \mathbb{I}(\mathbb{P}(Y = 1|x) \geq K)$$

para diferentes cortes K

Curva ROC

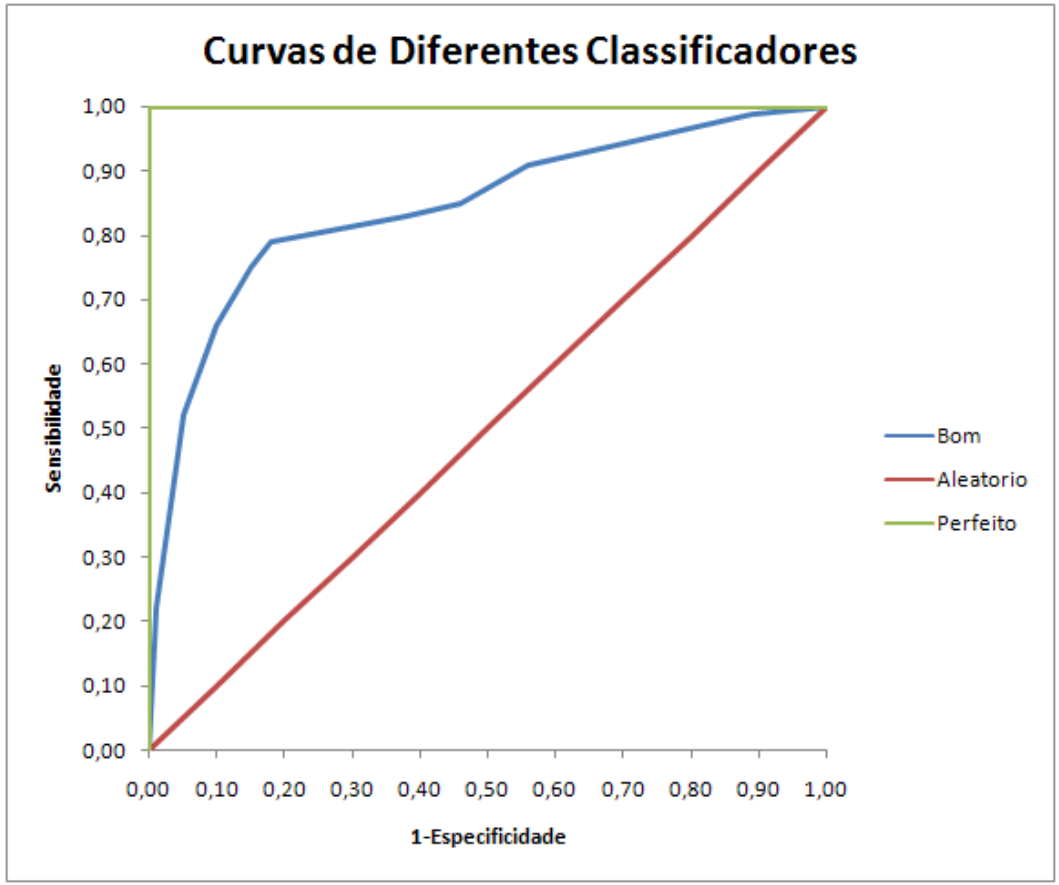


Gráfico da Sensibilidade vs 1-Especificidade para diferentes K's.

É comum escolher K que maximize "Sensibilidade+Especificidade"

AUC: Area Under the Curve. Critério comum para escolher classificador sem estabelecer um corte pré-fixado.

Análise Discriminante

Vamos lembrar o Teorema de Bayes mais uma vez:

$$\mathbb{P}(Y = c|\mathbf{x}) = \frac{f(\mathbf{x}|Y = c)\mathbb{P}(Y = c)}{\sum_{s \in \mathcal{X}} f(\mathbf{x}|Y = s)\mathbb{P}(Y = s)}$$

Vimos que o Naive Bayes criar um classificador com base neste teorema, e assumindo-se independência condicional:

$$f(\mathbf{x}|Y = s) = f(x_1, \ldots, x_p|Y = s) = \prod_{j=1}^{d} f(x_j|Y = s)$$

Que outras suposições podem ser feitas?

Análise Discriminante

Na análise discriminante, supomos que o vetor $\mathbf{X}$, dado Y, tem distribuição normal multivariada.

Existem duas formas de análise discriminante mais comuns:

- Análise Discriminante Linear
- Análise Discriminante Quadrática

Análise Discriminante Linear

Assumimos que

$$\mathbf{X} = (X_1, \ldots, X_d) | Y = c \sim \mathit{Normal}(\mu_c, \Sigma),$$

i.e.,

$$f(\mathbf{x}|Y = c) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-(\mathbf{x}-\mu_c)'\Sigma^{-1}(\mathbf{x}-\mu_c)}$$

Podemos estimar esses parâmetros pelo método da máxima verossimilhança:

$$\widehat{\mu_c} = \frac{1}{|\mathcal{C}_c|} \sum_{k \in \mathcal{C}_c} \mathbf{x}_k; \qquad \widehat{\Sigma} = \frac{1}{n} \sum_{c \in \mathcal{C}} \sum_{k \in \mathcal{C}_c} (\mathbf{x}_k - \widehat{\mu_c})(\mathbf{x}_k - \widehat{\mu_c})'$$

onde $\mathcal{C}_c = \{j = 1, \ldots, n : Y_j = c\}$

Para o caso binário, $g(x) = 1$ se, e só se,

$$\frac{\widehat{\mathbb{P}}(Y=1|x)}{\widehat{\mathbb{P}}(Y=0|x)} \geq K \iff \frac{\widehat{f}(\mathbf{x}|Y=1)\widehat{\mathbb{P}}(Y=1)}{\widehat{f}(\mathbf{x}|Y=0)\widehat{\mathbb{P}}(Y=0)} \geq K \iff$$

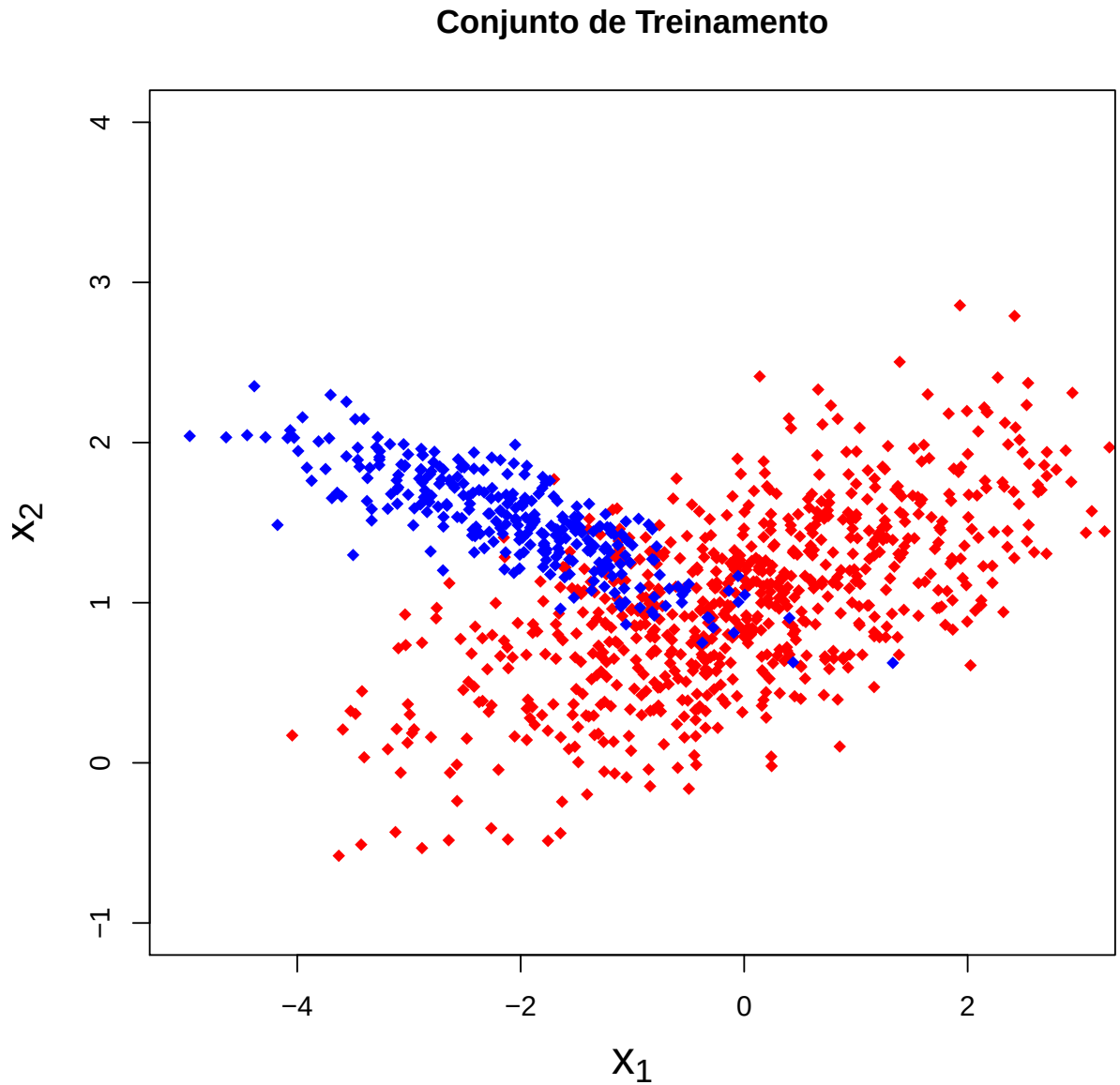
$$\log \widehat{f}(\mathbf{x}|Y=1) - \log \widehat{f}(\mathbf{x}|Y=0) \geq \log K + \log \widehat{\mathbb{P}}(Y=0) - \log \widehat{\mathbb{P}}(Y=1)$$

$$\iff -(\mathbf{x} - \widehat{\mu}_1)'\widehat{\Sigma}^{-1}(\mathbf{x} - \widehat{\mu}_1) + (\mathbf{x} - \widehat{\mu}_0)'\widehat{\Sigma}^{-1}(\mathbf{x} - \widehat{\mu}_0) \geq K'$$

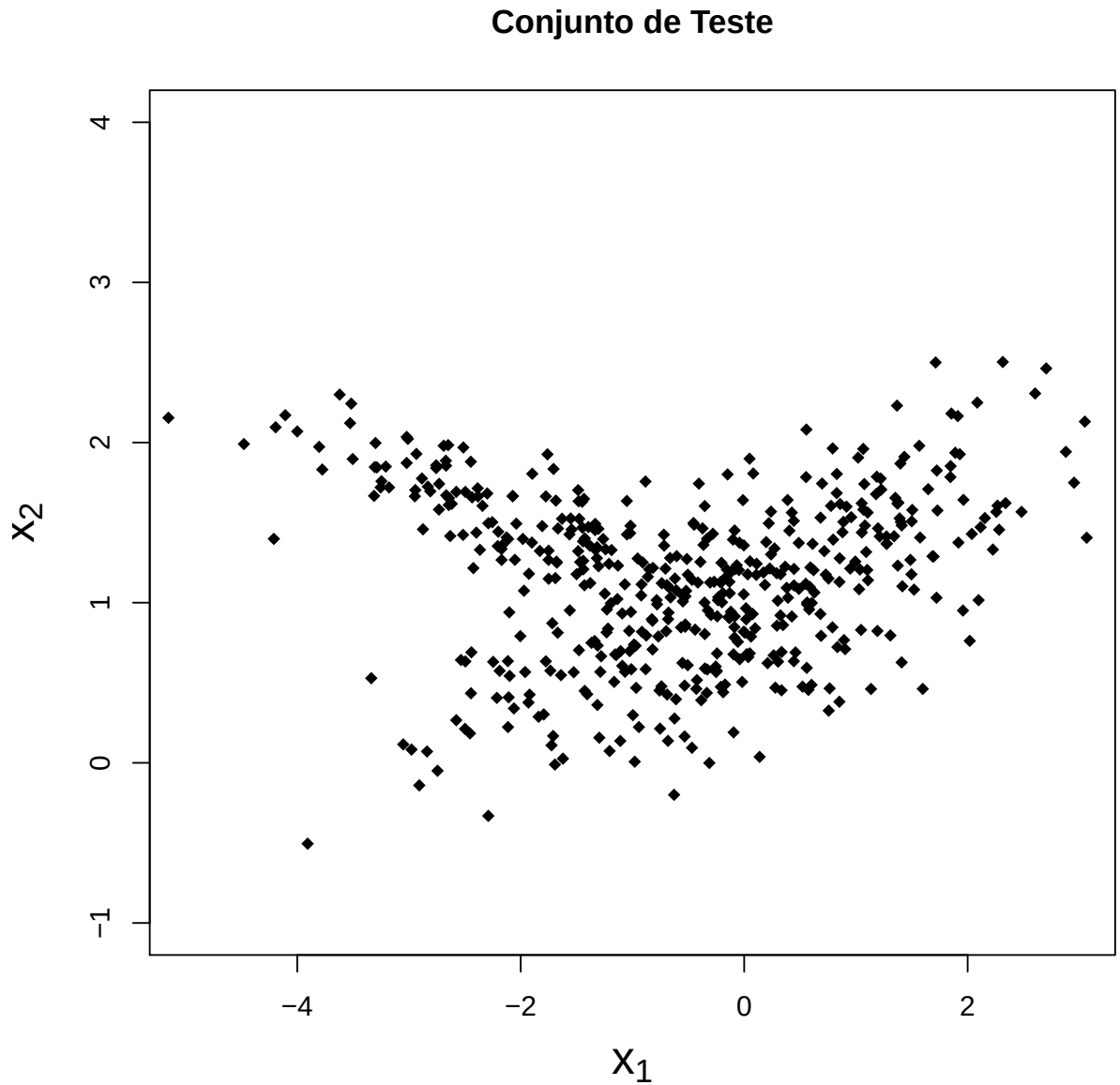
$$\iff +2\mathbf{x}'\widehat{\Sigma}^{-1}\widehat{\mu}_1 - \widehat{\mu}_1'\widehat{\Sigma}^{-1}\widehat{\mu}_1 - 2\mathbf{x}'\widehat{\Sigma}^{-1}\widehat{\mu}_0 + \widehat{\mu}_0'\widehat{\Sigma}^{-1}\widehat{\mu}_0 \geq K'$$

Trata-se da equação de um **hiperplano**, por isso o nome "Análise Discriminante **Linear**"

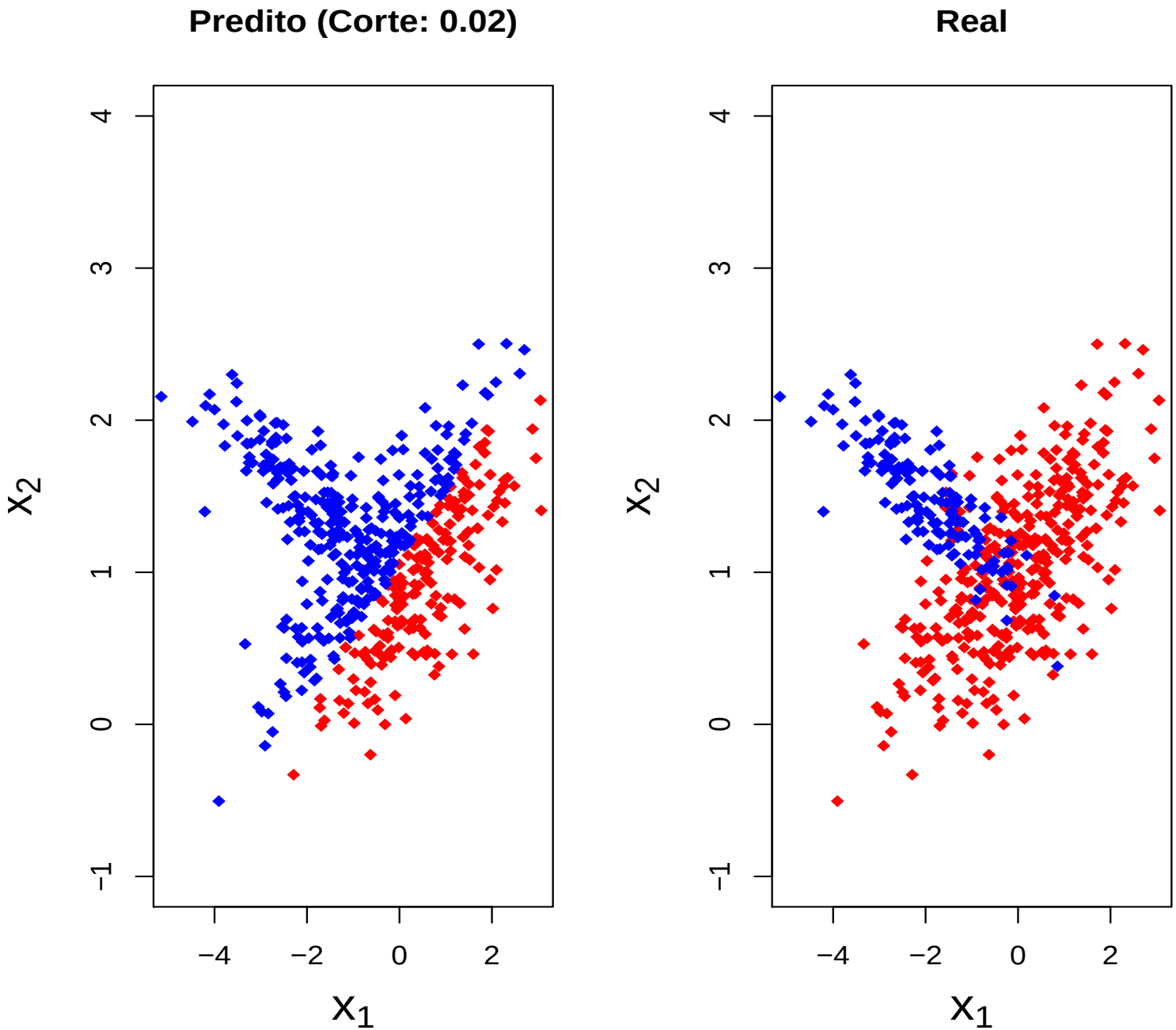
Exemplo



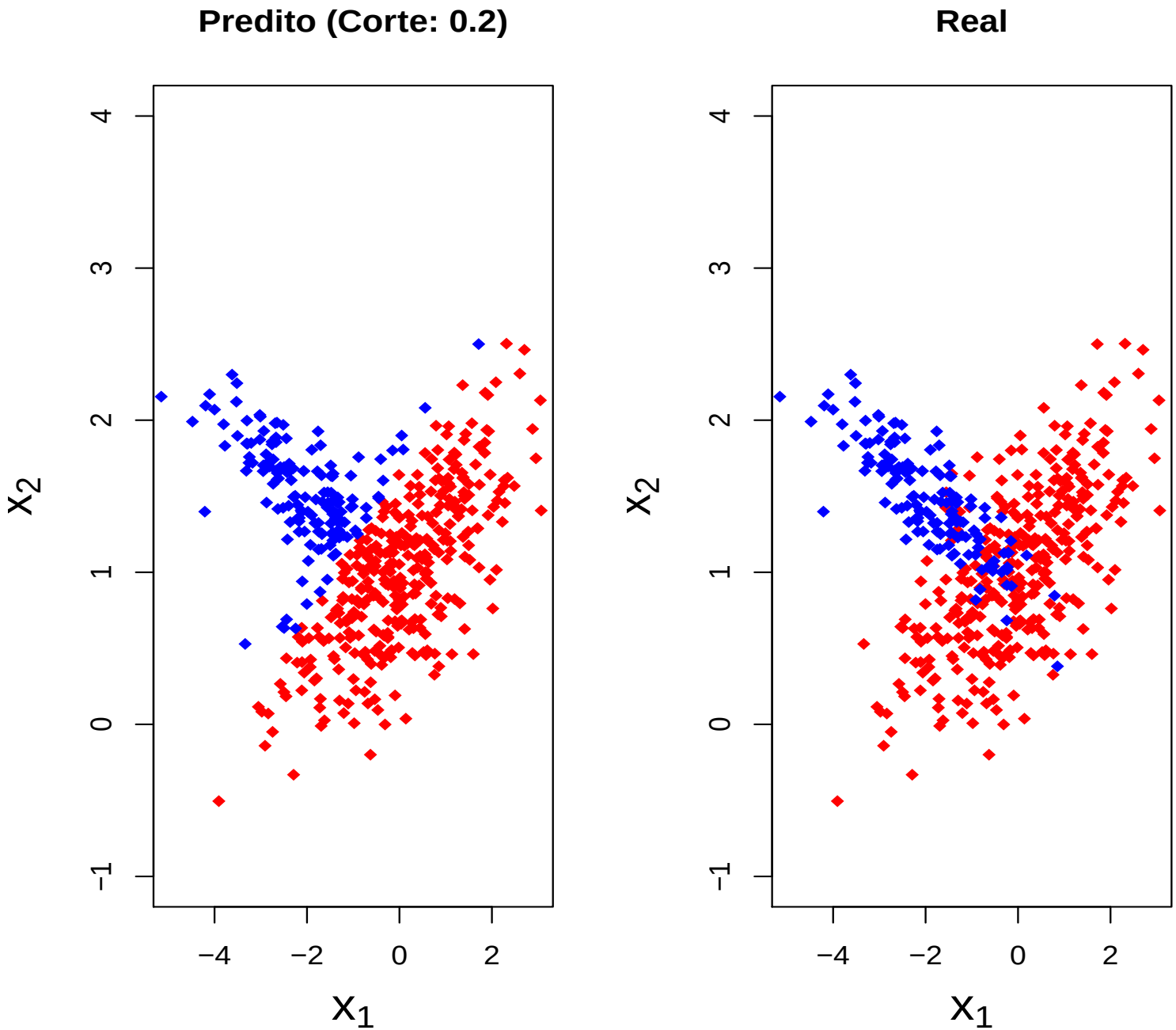
Exemplo



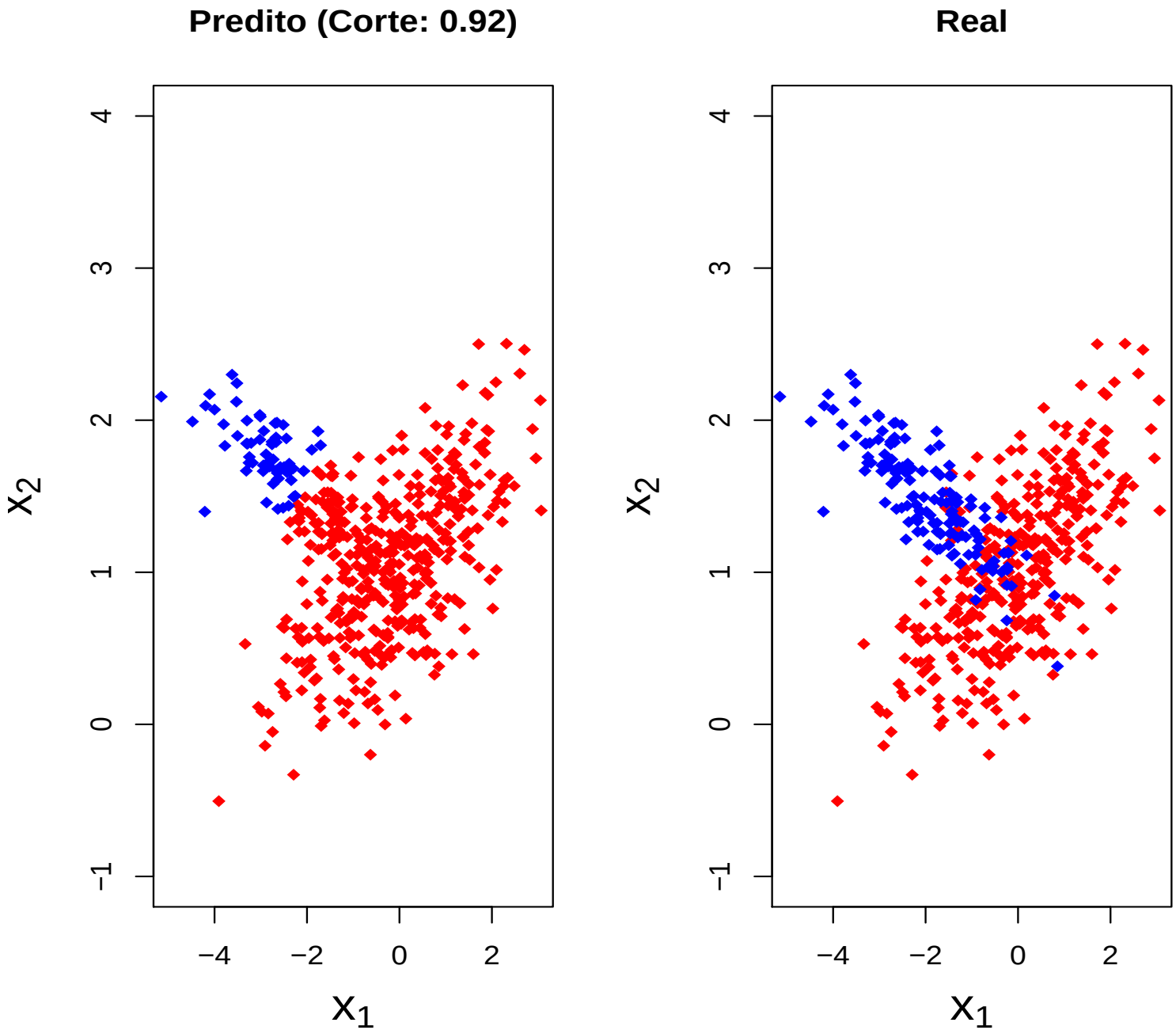
Exemplo



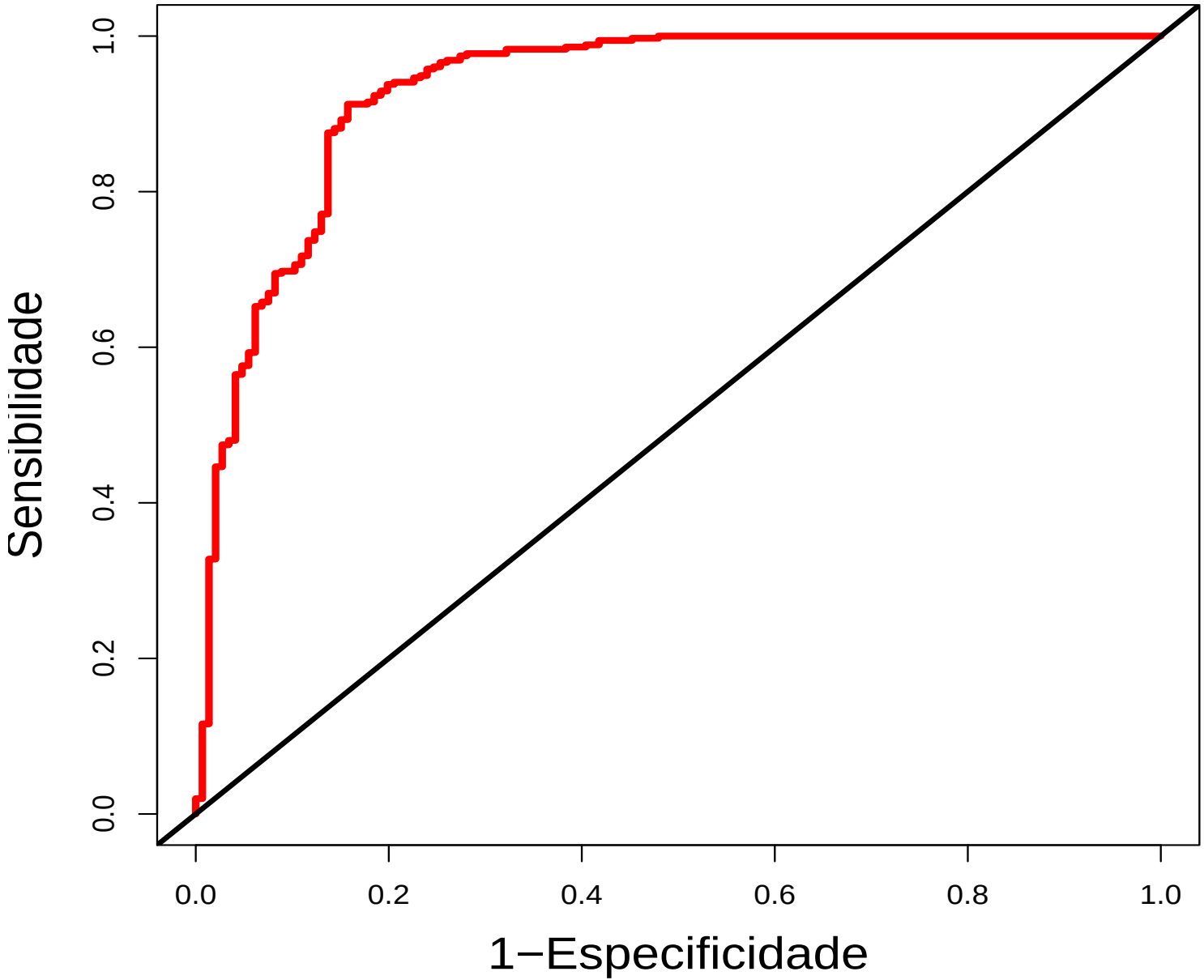
Exemplo



Exemplo



Exemplo


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1−Especificidade
0.0
0.2
0.4
0.6
0.8
1.0

No R:
Análise Discriminante Linear:

```
> library (MASS)
> lda.fit = lda(x=xTreino, grouping=yTreino)
> lda.pred = predict(lda.fit, newdata=xNovo)
> lda.pred$posterior
```

Note que não necessariamente acreditamos na suposição de normalidade

Análise Discriminante Quadrática

Assumimos que

$$\mathbf{X} = (X_1, \ldots, X_d) | Y = c \sim \mathit{Normal}(\mu_c, \Sigma_c),$$

i.e.,

$$f(\mathbf{x}|Y = c) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_c|}} e^{-(\mathbf{x}-\mu_c)'\Sigma_c^{-1}(\mathbf{x}-\mu_c)}$$

Mesma suposição que a Linear, mas com variâncias diferentes em cada grupo.

Podemos estimar esses parâmetros pelo método da máxima verossimilhança:

$$\widehat{\mu_c} = \frac{1}{|\mathcal{C}_c|} \sum_{k \in \mathcal{C}_c} \mathbf{X}_k; \qquad \widehat{\Sigma}_c = \frac{1}{|\mathcal{C}_c|} \sum_{k \in \mathcal{C}_c} (\mathbf{x}_k - \widehat{\mu_c})(\mathbf{x}_k - \widehat{\mu_c})'$$

onde $\mathcal{C}_c = \{j = 1, \ldots, n : Y_j = c\}$

Para o caso binário, $g(x) = 1$ se, e só se,

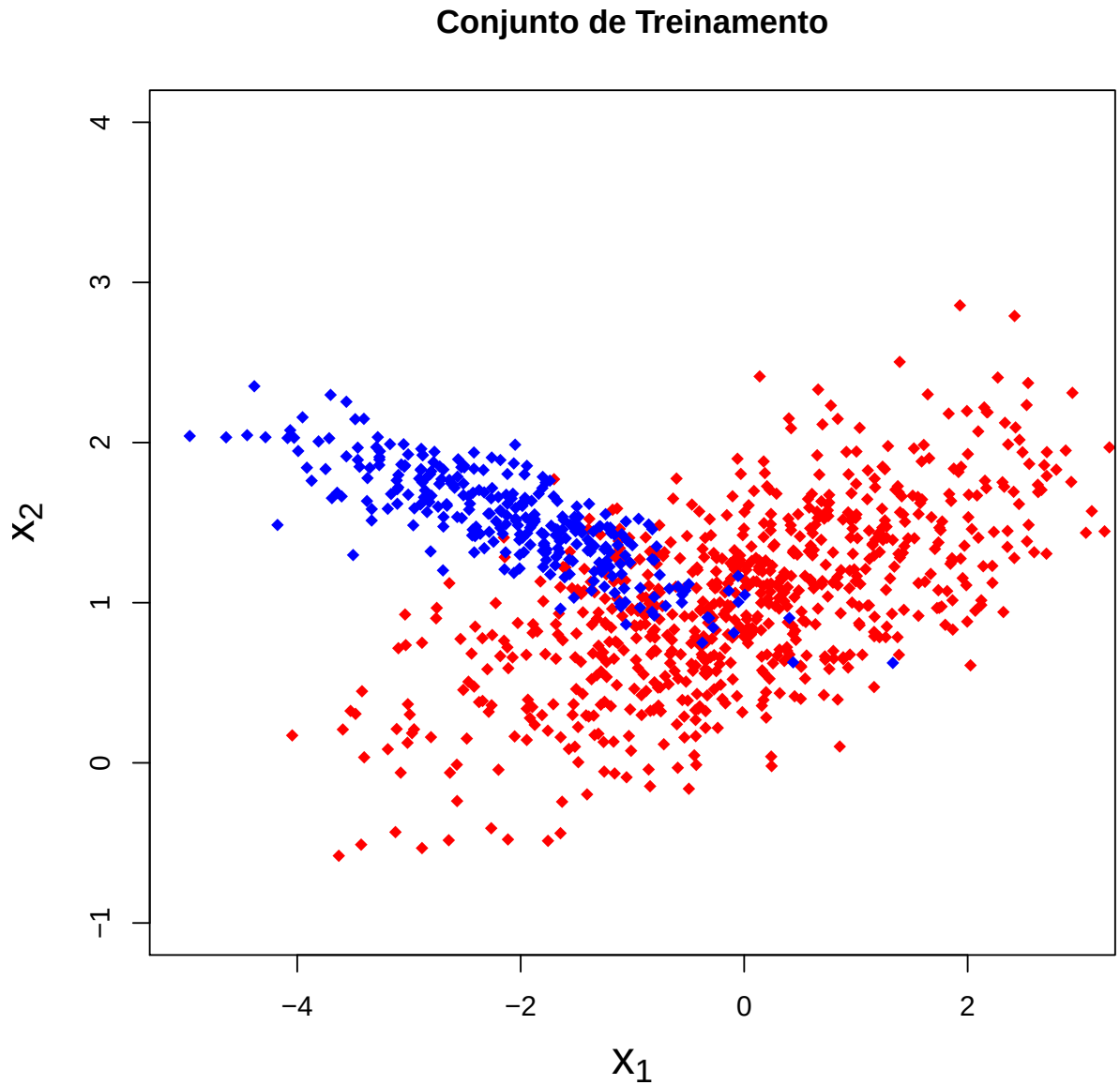
$$\frac{\widehat{\mathbb{P}}(Y=1|x)}{\widehat{\mathbb{P}}(Y=0|x)} \geq K \iff \frac{\widehat{f}(\mathbf{x}|Y=1)\widehat{\mathbb{P}}(Y=1)}{\widehat{f}(\mathbf{x}|Y=0)\widehat{\mathbb{P}}(Y=0)} \geq K \iff$$

$$\log \widehat{f}(\mathbf{x}|Y=1) - \log \widehat{f}(\mathbf{x}|Y=0) \geq \log K + \log \widehat{\mathbb{P}}(Y=0) - \log \widehat{\mathbb{P}}(Y=1)$$

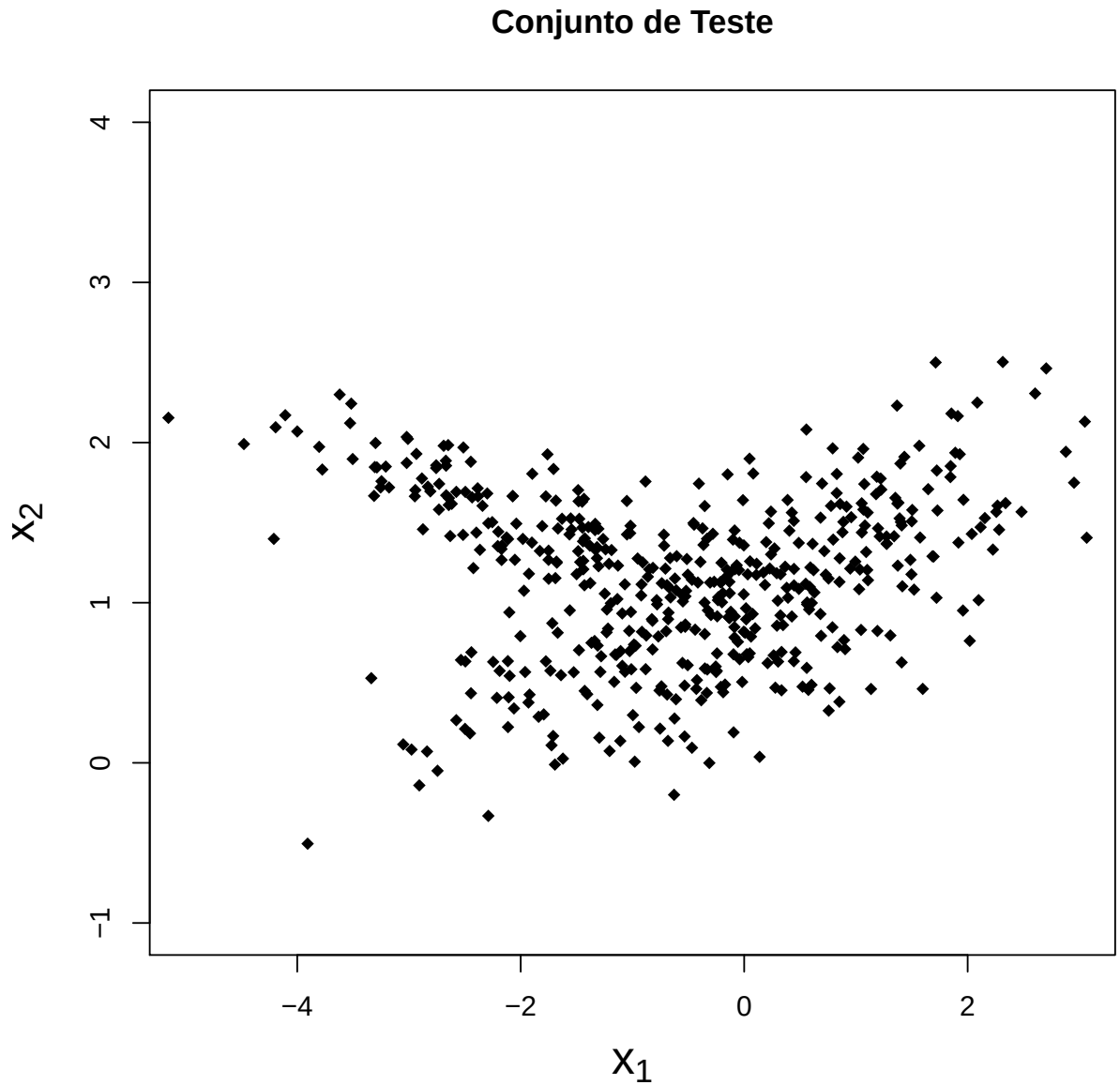
$$\iff -(\mathbf{x} - \widehat{\mu}_1)' \widehat{\Sigma_1}^{-1} (\mathbf{x} - \widehat{\mu}_1) + (\mathbf{x} - \widehat{\mu}_0)' \widehat{\Sigma_0}^{-1} (\mathbf{x} - \widehat{\mu}_0) \geq K'$$

Trata-se da equação quadrática, por isso o nome "Análise Discriminante **Quadrática**"

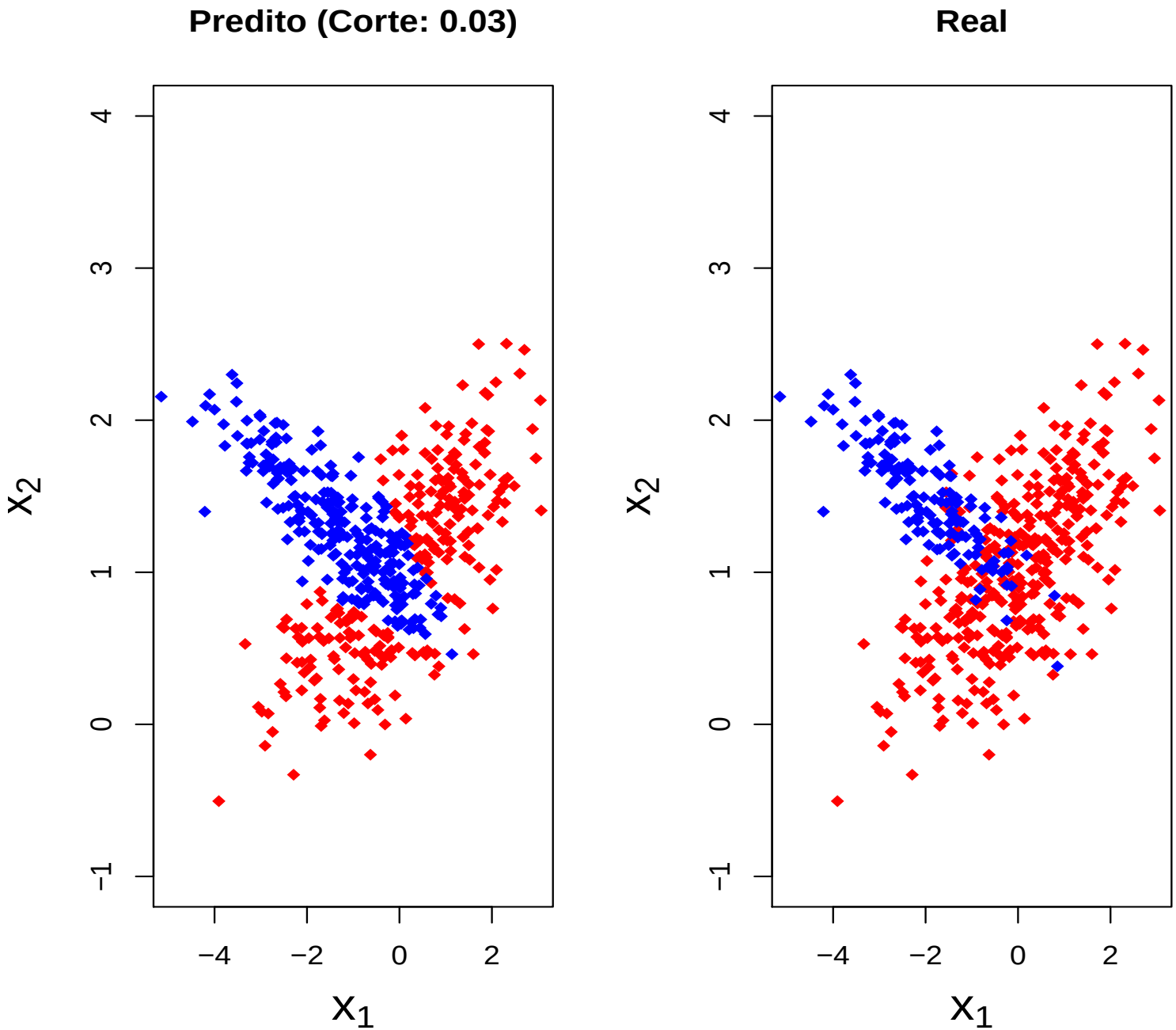
Exemplo



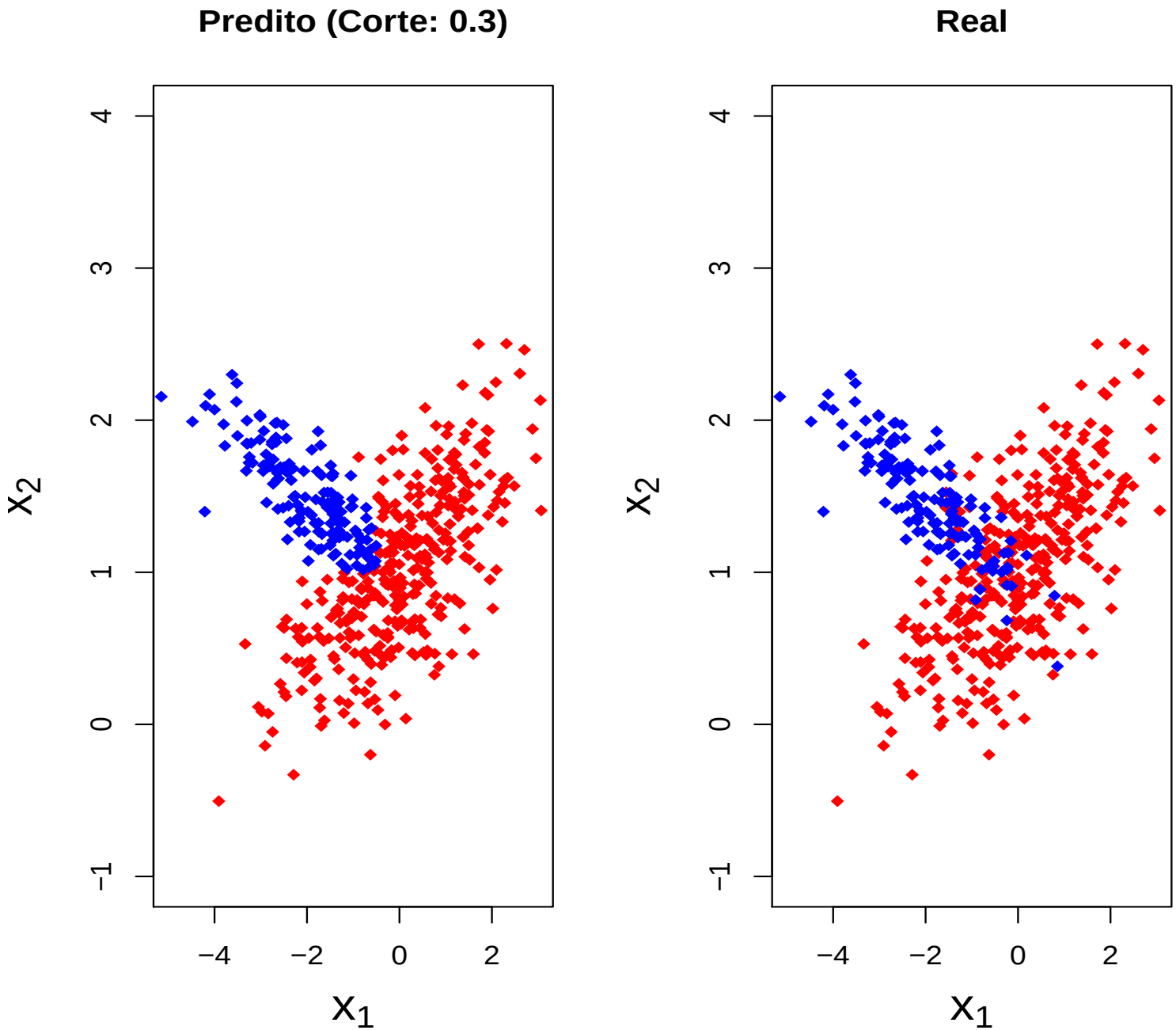
Exemplo



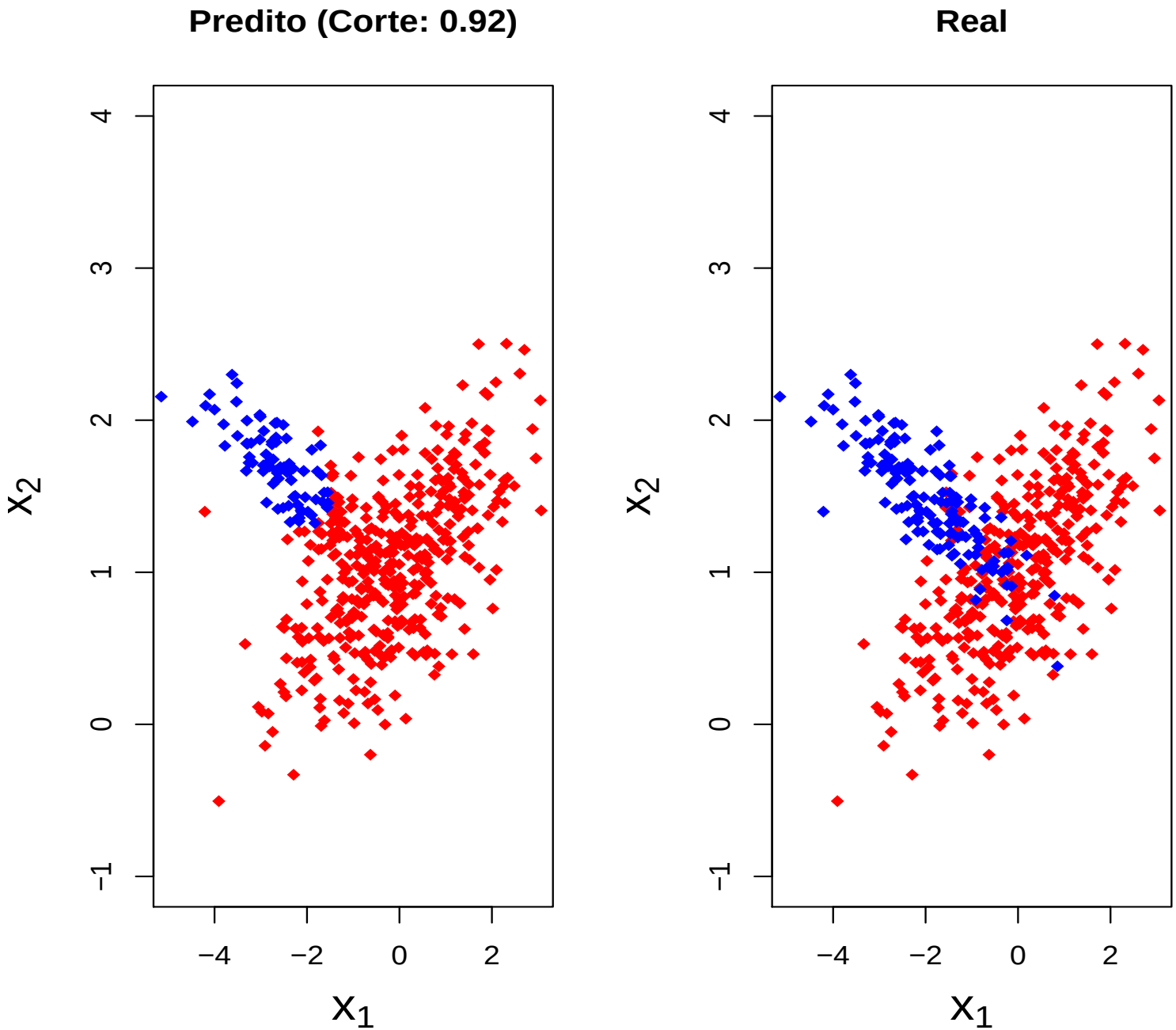
Exemplo



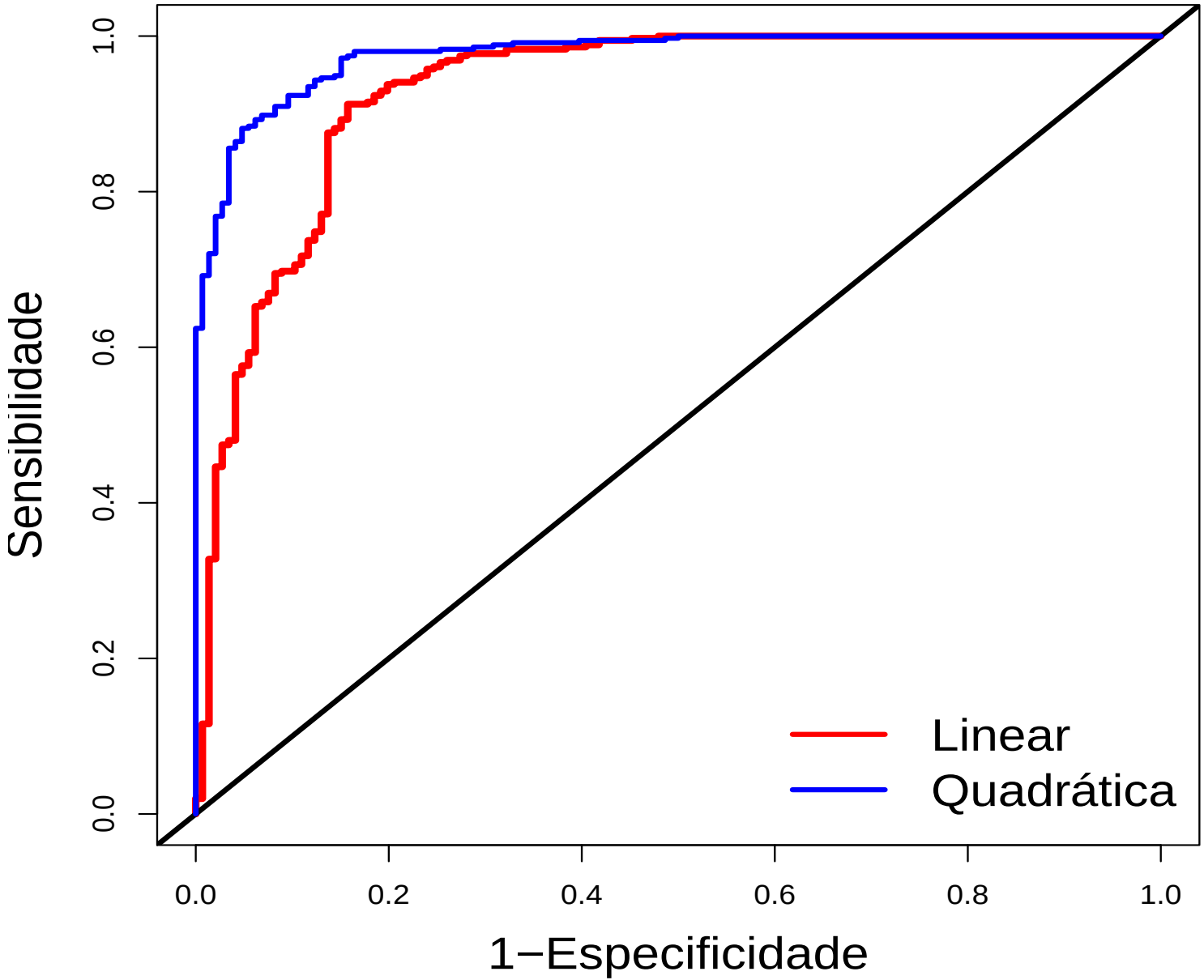
Exemplo



Exemplo



Exemplo


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1−Especificidade
0.0
0.2
0.4
0.6
0.8
1.0
Linear
Quadrática

No R

Análise Discriminante Quadrática: Igual à linear, mas usar `qda` ao invés de `lda`:

```
> library (MASS)
> qda.fit = qda(x=xTreino, grouping=yTreino)
> qda.pred = predict(qda.fit, newdata=xNovo)
> qda.pred$posterior
```

Revisão

Vimos que a função de risco é dada por

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}\left(Y \neq g(\mathbf{X})\right),$$

Nem sempre tal função nos traz toda informação sobre g. É comum considerar matrizes de confusão.

Examplo de matriz de confusão

	Valor verdadeiro	
Valor Predito	Y=0	Y=1
Y=0	VN	FN
Y=1	FP	VP

V: Verdadeiro / F: Falso
P: Positivo / N: Negativo

Com base nessa tabela, define-se

- Sensibilidade: VP/(VP+FN) (dos pacientes doentes, quantos foram corretamente identificados?)
- Especificidade: VN/(VN+FP) (dos pacientes não doentes, quantos foram corretamente identificados?)

Um outro problema relacionado a isso: se $Y = 1$ é raro, em geral teremos $\mathbb{P}(Y = 1|x)$ baixo.

Assim, usar um corte de $1/2$ pode não ajudar: a regra $g(x) = \mathbb{I}(\mathbb{P}(Y = 1|x) \geq 1/2)$ nos levará a sempre decidir por $Y = 0$, mesmo se essas probabilidades estiverem bem estimadas.

Na prática, para evitar isso é comum buscar cortes diferentes de $1/2$, i.e., buscam-se regras do tipo

$$g(x) = \mathbb{I}(\mathbb{P}(Y = 1|x) \geq K)$$

para diferentes cortes K

Curva ROC

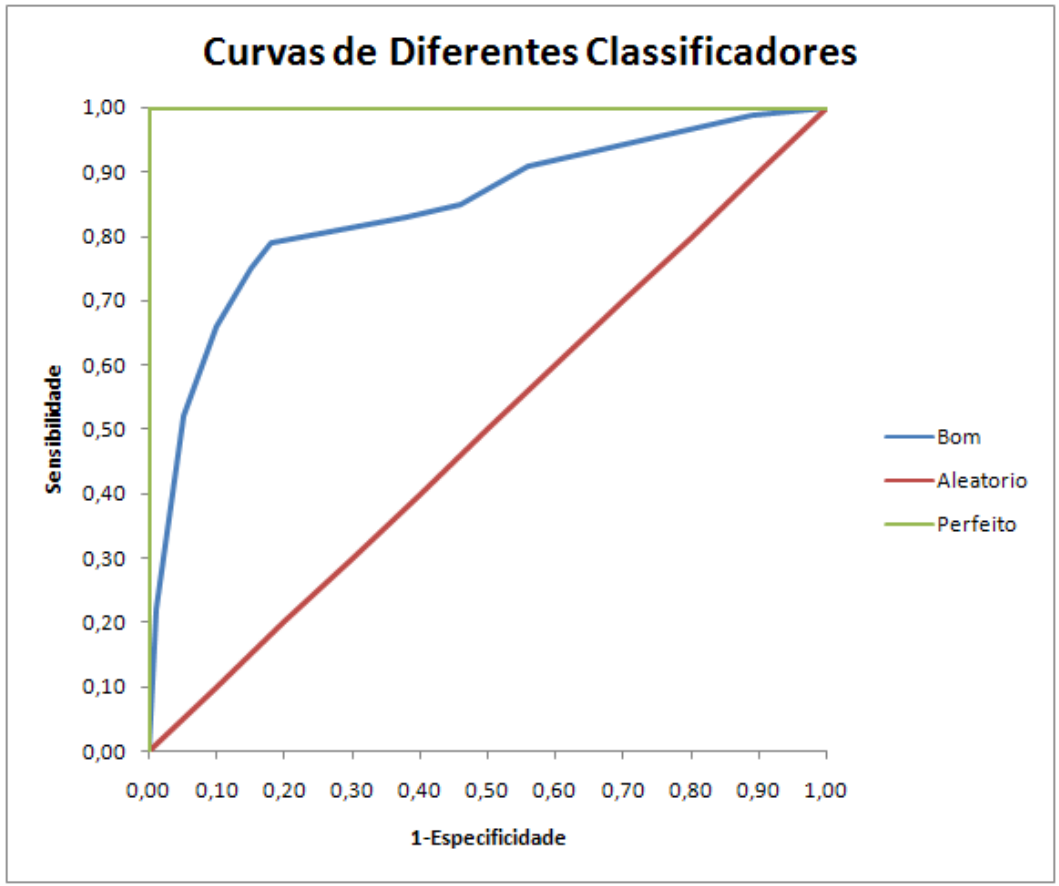


Gráfico da Sensibilidade vs 1-Especificidade para diferentes K's.

É comum escolher K que maximize "Sensibilidade+Especificidade"

Análise Discriminante

Vamos lembrar o Teorema de Bayes mais uma vez:

$$\mathbb{P}(Y = c|\mathbf{x}) = \frac{f(\mathbf{x}|Y = c)\mathbb{P}(Y = c)}{\sum_{s\in\mathcal{X}} f(\mathbf{x}|Y = s)\mathbb{P}(Y = s)}$$

Na análise discriminante, supomos que o vetor $\mathbf{X}$, dado Y, tem distribuição normal multivariada.

Existem duas formas de análise discriminante mais comuns:

- Análise Discriminante Linear
- Análise Discriminante Quadrática

Análise Discriminante Linear

Assumimos que

$$\mathbf{X} = (X_1, \ldots, X_d) | Y = c \sim Normal(\mu_c, \Sigma),$$

i.e.,

$$f(\mathbf{x}|Y = c) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-(\mathbf{x}-\mu_c)' \Sigma^{-1} (\mathbf{x}-\mu_c)}$$

Regra de decisao:

$$2\mathbf{x}' \widehat{\Sigma}^{-1} \widehat{\mu}_1 - 2\mathbf{x}' \widehat{\Sigma}^{-1} \widehat{\mu}_0 \geq K''$$

Análise Discriminante Quadrática

Assumimos que

$$\mathbf{X} = (X_1, \ldots, X_d) | Y = c \sim \mathit{Normal}(\mu_c, \Sigma_c),$$

i.e.,

$$f(\mathbf{x}|Y = c) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_c|}} e^{-(\mathbf{x}-\mu_c)' \Sigma_c^{-1} (\mathbf{x}-\mu_c)}$$

Mesma suposição que a Linear, mas com variâncias diferentes em cada grupo.

Regra de decisao:

$$-(\mathbf{x} - \widehat{\mu}_1)' \widehat{\Sigma_1}^{-1} (\mathbf{x} - \widehat{\mu}_1) + (\mathbf{x} - \widehat{\mu}_0)' \widehat{\Sigma_0}^{-1} (\mathbf{x} - \widehat{\mu}_0) \geq K'$$