Métricas para Classificação

INFORMAÇÃO,

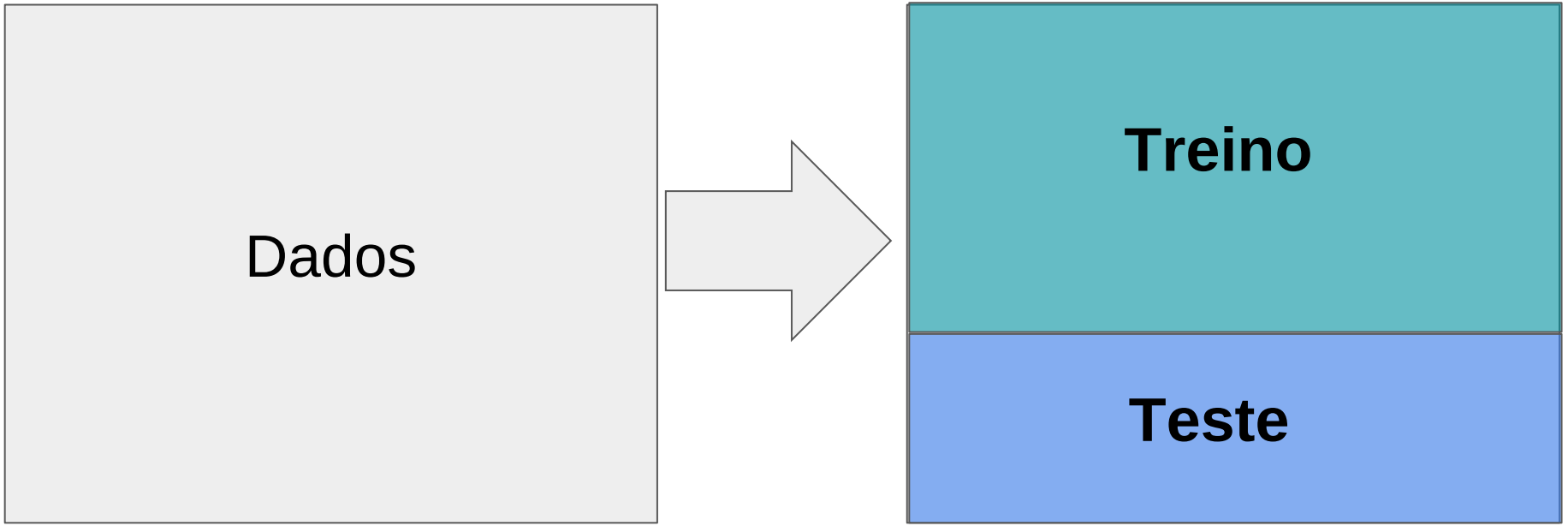
TECNOLOGIA

& INOVAÇÃO

Risco

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

Data Splitting



Risco

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

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

OUTRAS MÉTRICAS

EXEMPLO: Doença rara

Poucos pacientes com Y=1

Classificador trivial: g(x)=0

EXEMPLO: Doença rara

Matriz de confusão

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

EXEMPLO: Doença rara

- **Sensibilidade/Recall:** $S = VP/(VP + FN)$ (dos pacientes doentes, quantos foram corretamente identificados?)
- **Especificidade:** $E = VN/(VN + FP)$ (dos pacientes não doentes, quantos foram corretamente identificados?)
- **Valor preditivo positivo/Precision:** $VPP = VP/(VP + FP)$ (dos pacientes classificados como doentes, quantos foram corretamente identificados?)
- **Valor preditivo negativo:** $VPN = VN/(VN + FN)$ (dos pacientes classificados como não doentes, quantos foram corretamente identificados?)
- **Estatística F1:** $F1 = \frac{2}{1/S + 1/VPP}$

Matriz de confusão

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

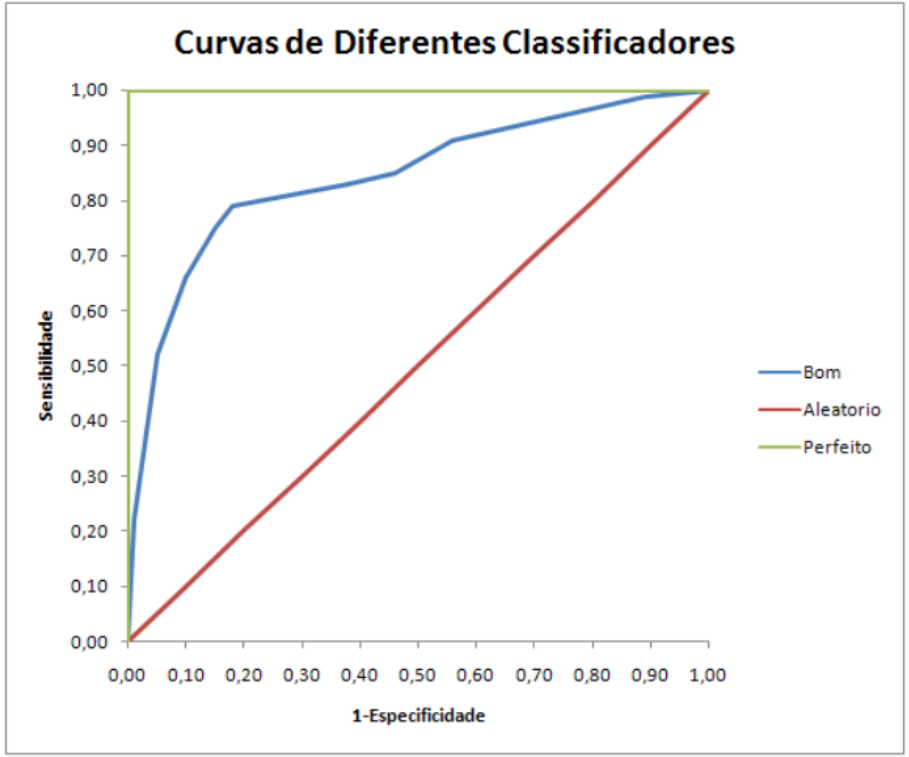
Problema relacionado

$Y = 1$ é raro $\Rightarrow \mathbb{P}(Y = 1|\mathbf{x})$ baixo

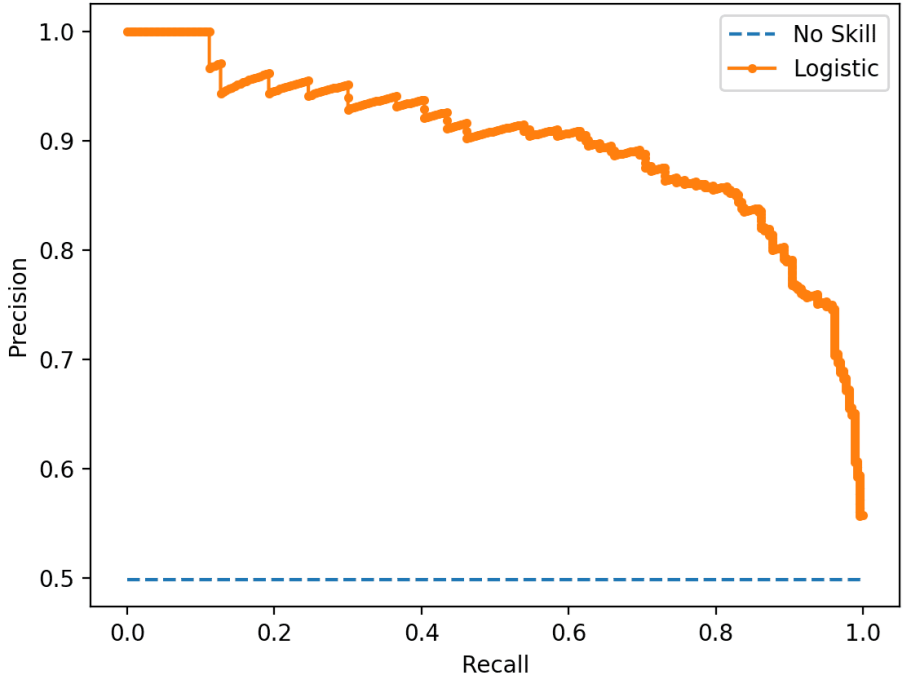
$g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) \geq 1/2) = 0$ para quase todo $\mathbf{x}$

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

Curva ROC


Curvas de Diferentes Classificadores
Sensibilidade
1,00
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00
0,00 0,10 0,20 0,30 0,40 0,50 0,60 0,70 0,80 0,90 1,00
1-Especificidade
Bom
Aleatorio
Perfeito

Precision-Recall Curve



Assimetria nos erros

O problema é o desbalanceamento?

Python