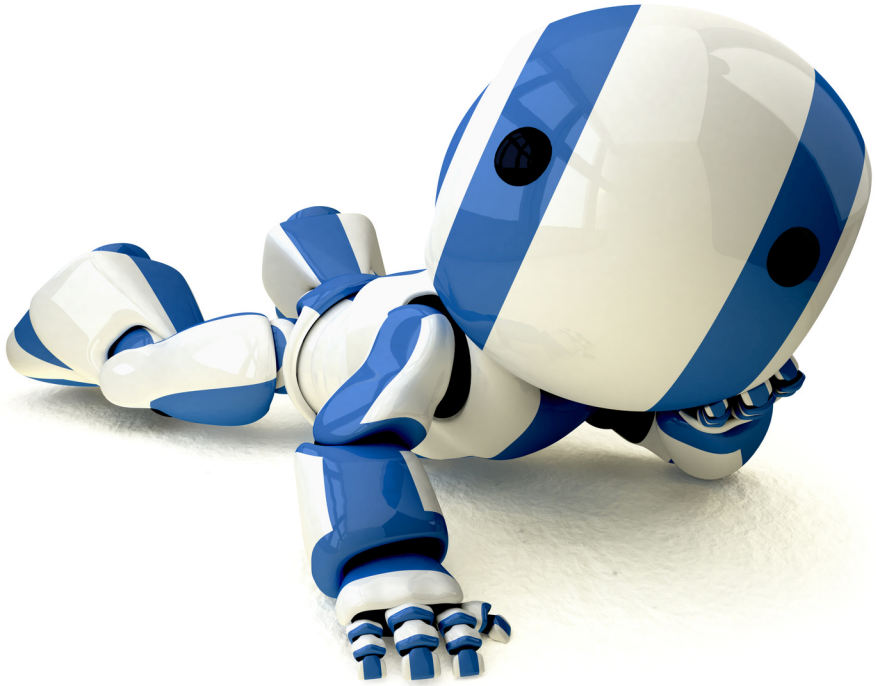
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
CURSO DE EXTENSÃO EM DATA SCIENCE

Aprendizado de Máquina Supervisionado

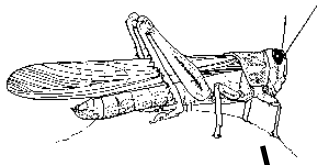


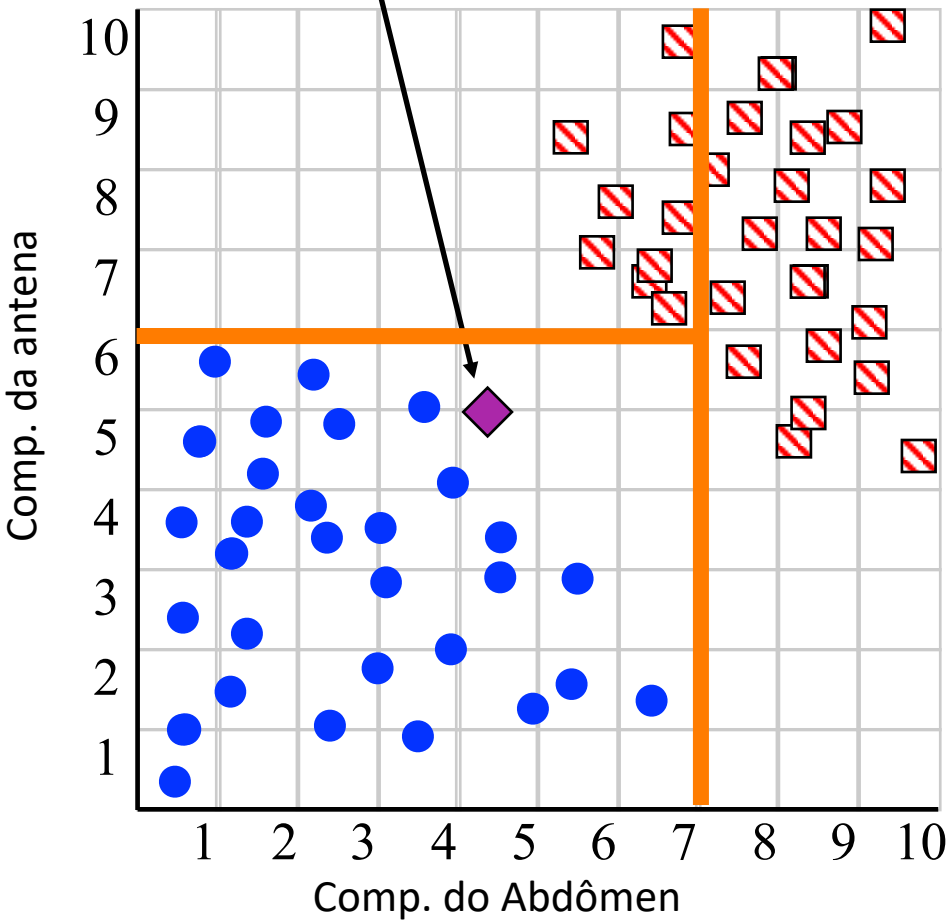
Avaliação de Desempenho de Modelos Supervisionados

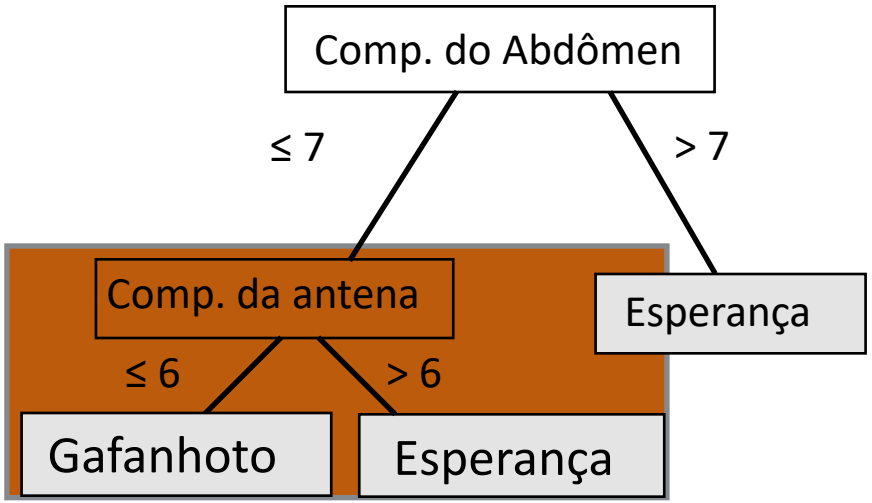
Prof. Dr. Rodrigo C. Barros

BUSINESS INTELLIGENCE AND
MACHINE LEARNING RESEARCH GROUP

Aula Passada







Aula de Hoje

- Protocolos para Avaliação de Desempenho
  - *Holdout*
  - *Random Subsampling*
  - *Cross-Validation*
  - *Bootstrap*
- Medidas para Avaliação de Classificadores
  - Matriz de Confusão
  - Curvas ROC

Desempenho de Modelos Supervisionados

- Espera-se de um classificador/regressor que ele apresente desempenho adequado para dados **não vistos**
  - Poder de generalização

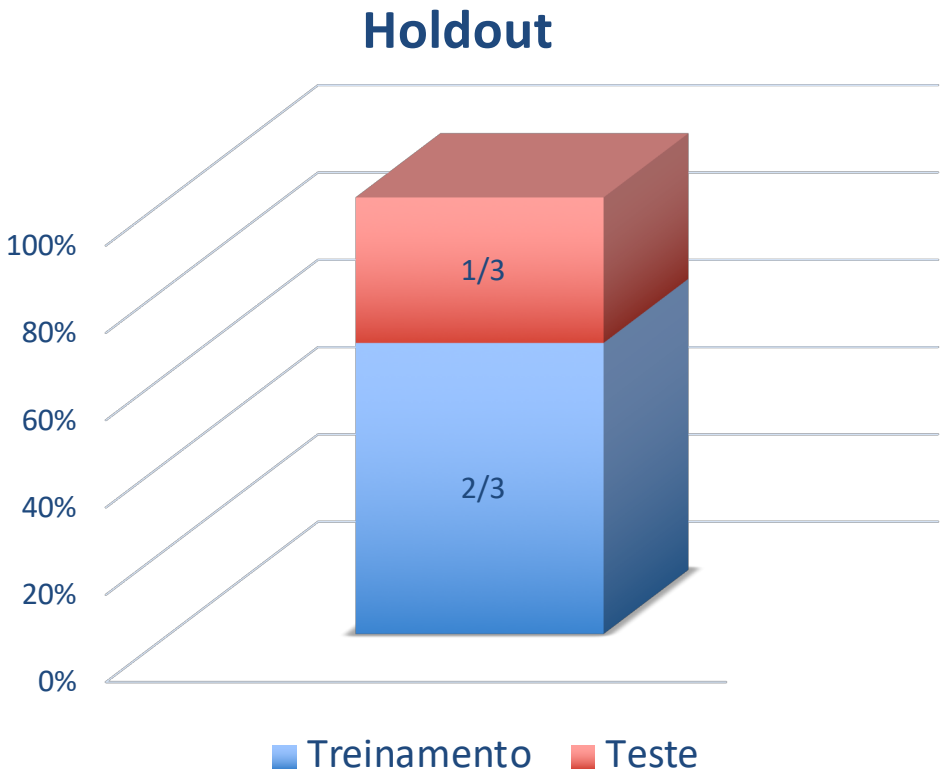
- Para estimarmos de maneira correta o desempenho do modelo, precisamos seguir um protocolo bem definido
  - Separar dados cujo atributo alvo é conhecido em dois conjuntos mutuamente exclusivos: **treinamento e teste**
  - **Jamais** avaliar o desempenho de um modelo em dados utilizados para seu treinamento, sob pena de superestimar o desempenho do modelo

Protocolos para Avaliação de Desempenho

- Existem diferentes protocolos para realizar a separação dos dados disponíveis em conjuntos de treinamento e teste
  - *Holdout*
  - *Random Subsampling*
  - *Cross-Validation*
    - *Leave-one-out*
  - *Bootstrap*

Holdout

- Também conhecido como *split-sample*
- Técnica mais simples para divisão de dados
- Faz uma única partição (aleatória) da amostra em:
  - Conjunto de treinamento
    - Geralmente 1/2 ou 2/3 dos dados
  - Conjunto paras teste
    - Dados restantes



Atenção: em problemas de classificação, recomenda-se que $p_{tr}(C_j) \approx p_{test}(C_j)\ \forall C_j \in Y$ (*holdout* **estratificado**)

Holdout

- Não é recomendado se dados não forem **abundantes** (ex: milhares de objetos)
- Caso aplicado em pequenos volumes de dados:
  - Poucos objetos são utilizados no treinamento
  - Modelo torna-se sensível à divisão realizada
    - Quanto menor o conjunto de treinamento, maior a variância (instabilidade / sensibilidade) do modelo obtido
    - Quanto menor o conjunto de teste, menos confiável é a estimativa de desempenho preditivo sobre dados não vistos
    - A solução para este cenário é utilizar métodos de re-amostragem

Métodos de Re-Amostragem

- Utilizam várias partições do conjunto original de dados para criar os conjuntos de treinamento e de teste
  - *Random subsampling*
  - *Cross-validation*
    - *Leave-one-out*
  - *Bootstrap*

Random Subsampling

- Múltiplas execuções de *holdout*
  - **Várias partições (p)** de treinamento e teste são escolhidas de maneira aleatória
  - $X_{tr} \cap X_{test} = \emptyset$
  - Medida de erro é calculada para cada partição
  - Erro de generalização estimado é a média dos erros para as diferentes partições
- Permite uma estimativa de erro mais realista
  - Porém, não há controle do número de vezes que cada objeto é utilizado nos conjuntos de treinamento e teste

Random Subsampling

- Exemplo
  - Suponha a existência dos seguintes objetos com valores de atributo alvo conhecido

$$X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(5)}, \mathbf{x}^{(6)}, \mathbf{x}^{(7)}, \mathbf{x}^{(8)}\}$$

    - Random subsampling com $p = 3$ e divisão 50%

	Treinamento	**Teste**	**Erro**
P_1	$\mathbf{x}^{(2)}, \mathbf{x}^{(4)}, \mathbf{x}^{(6)}, \mathbf{x}^{(7)}$	$\mathbf{x}^{(5)}, \mathbf{x}^{(8)}, \mathbf{x}^{(1)}, \mathbf{x}^{(3)}$	e_1
P_2	$\mathbf{x}^{(5)}, \mathbf{x}^{(3)}, \mathbf{x}^{(4)}, \mathbf{x}^{(8)}$	$\mathbf{x}^{(1)}, \mathbf{x}^{(7)}, \mathbf{x}^{(6)}, \mathbf{x}^{(2)}$	e_2
P_3	$\mathbf{x}^{(3)}, \mathbf{x}^{(7)}, \mathbf{x}^{(5)}, \mathbf{x}^{(4)}$	$\mathbf{x}^{(2)}, \mathbf{x}^{(8)}, \mathbf{x}^{(1)}, \mathbf{x}^{(6)}$	e_3
		Erro de generalização estimado:	$\frac{e_1 + e_2 + e_3}{p}$

Cross-Validation

- Validação cruzada
- Classe de métodos para estimativa da taxa de erro de generalização
  - *k-fold cross-validation*
    - Cada objeto participa o mesmo número de vezes do treinamento ($k - 1$ vezes)
    - Cada objeto participa o mesmo número de vezes do teste (1 vez)

Cross-Validation

- O conjunto de dados é dividido em k partições mutuamente exclusivas
  - A cada iteração, $k - 1$ partições são utilizadas para treinar o modelo
    - A partição restante é utilizada para testar o modelo
  - Erro estimado é a média dos erros das partições
  - Exemplo típico: **10-*fold cross-validation***

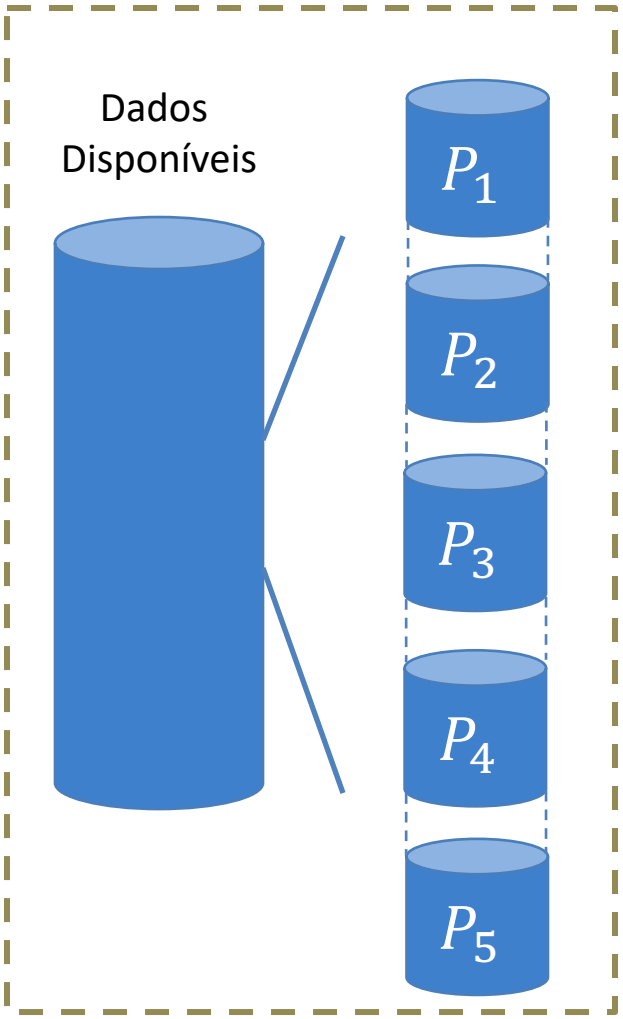
Cross-Validation

- Ex: 5-*fold cross-validation*

Dados
Disponíveis

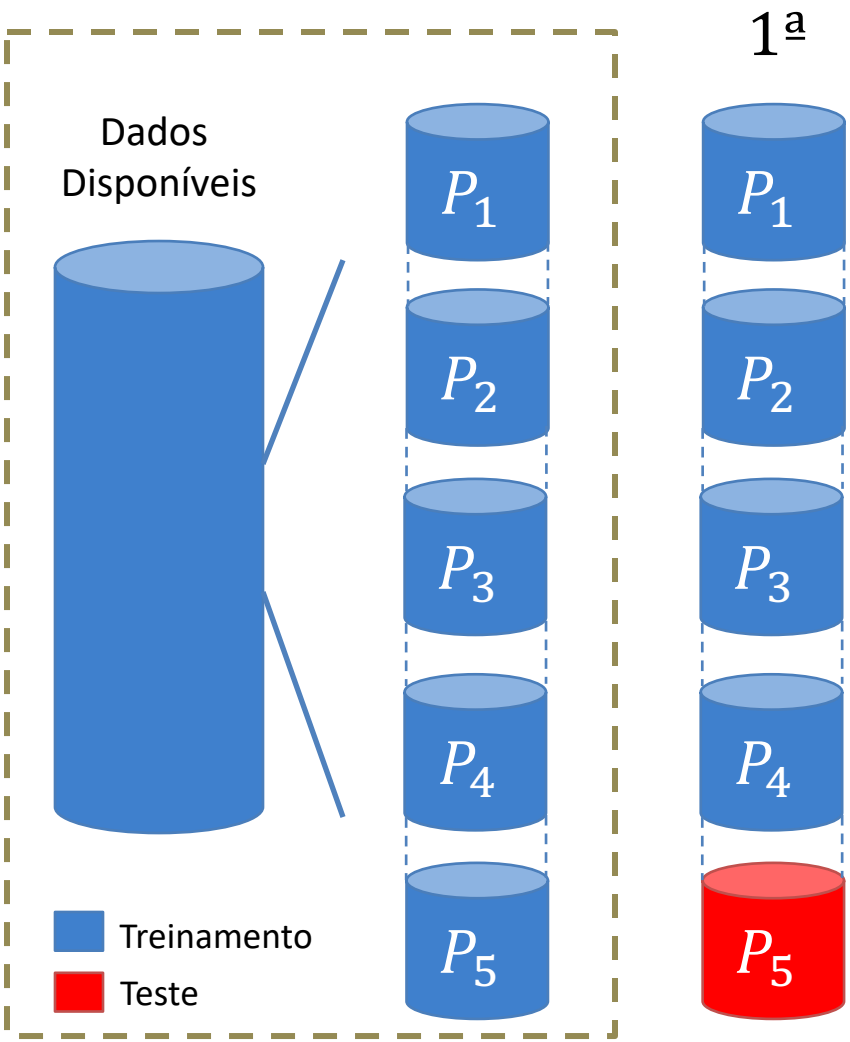
Cross-Validation

- Ex: 5-*fold cross-validation*



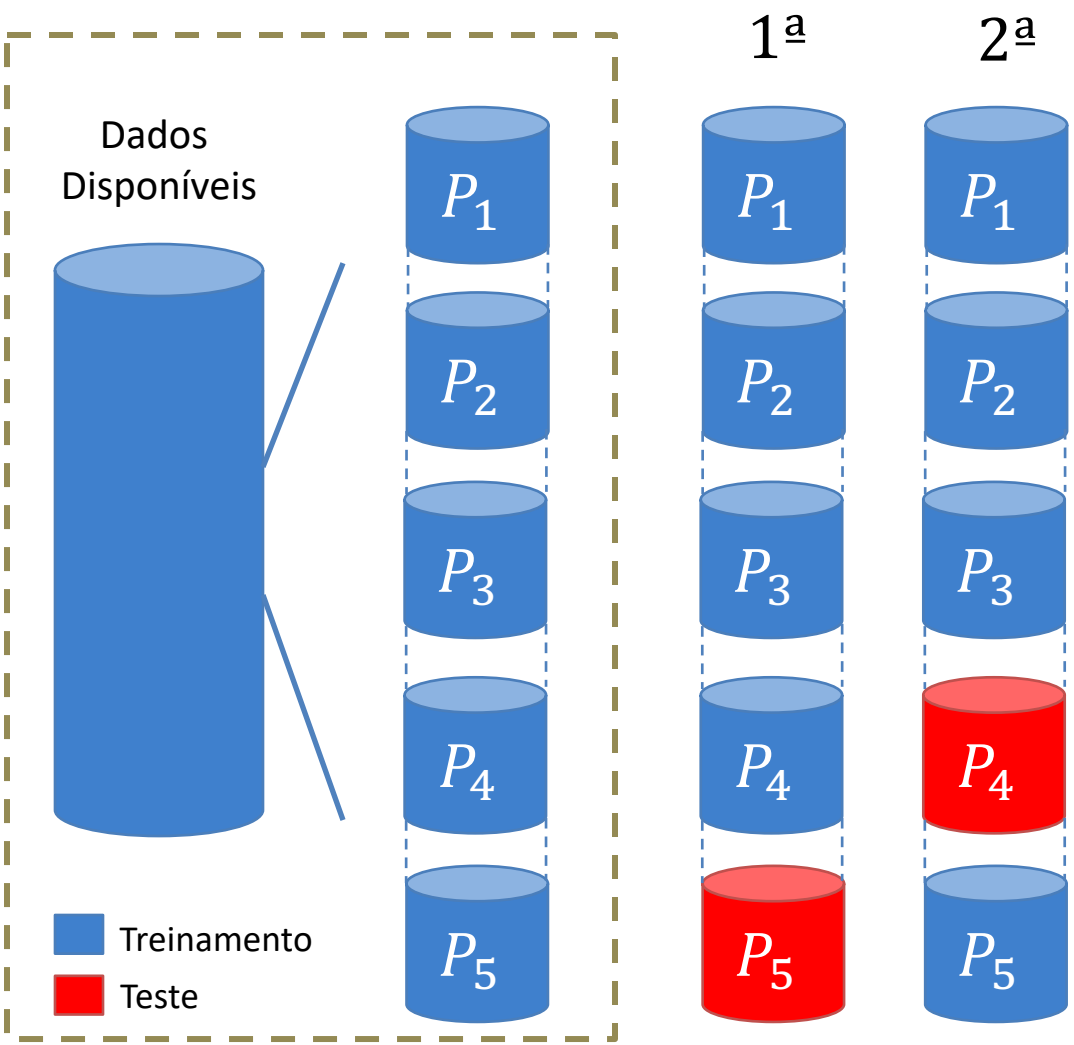
Cross-Validation

- Ex: 5-*fold cross-validation*



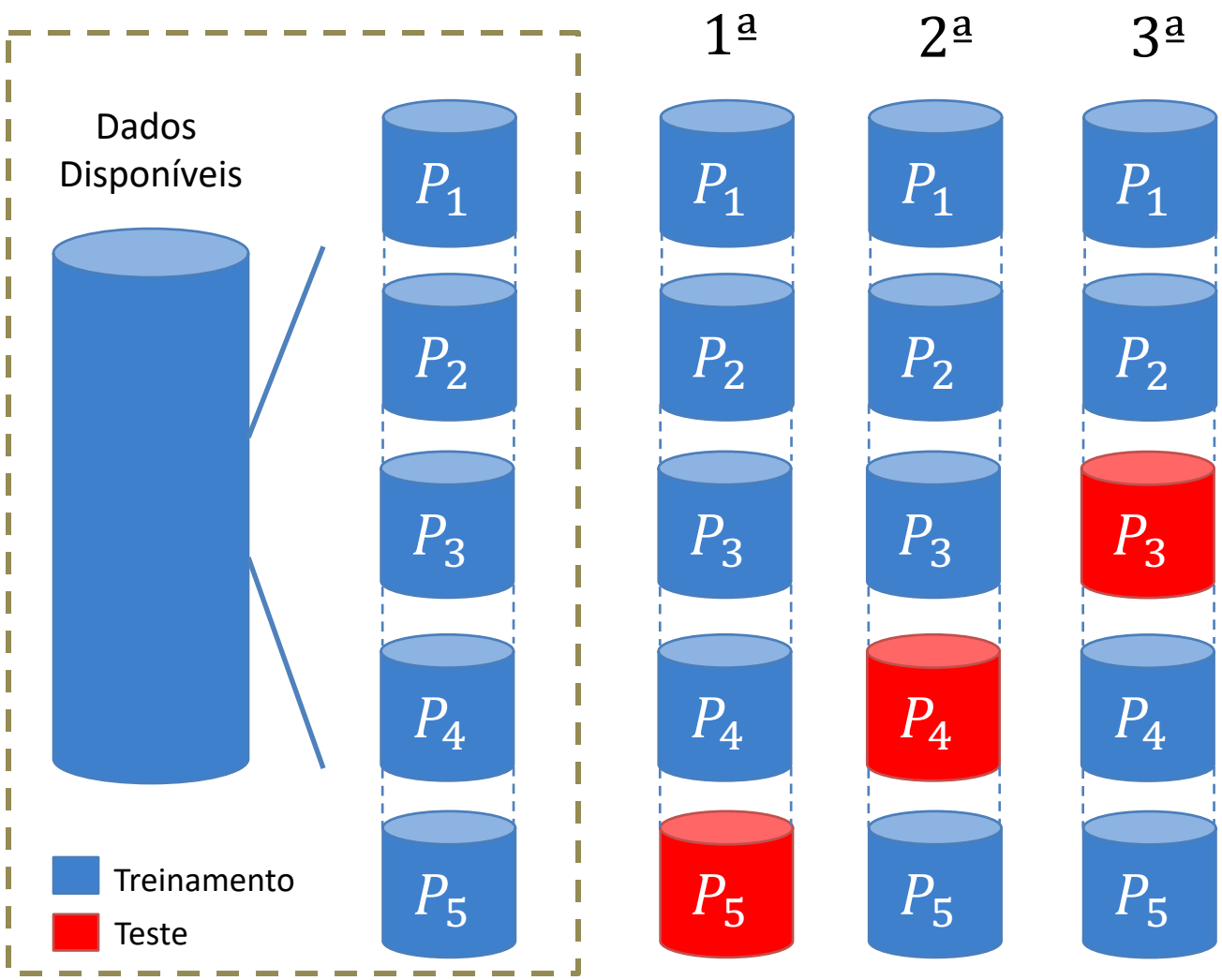
Cross-Validation

- Ex: 5-*fold cross-validation*



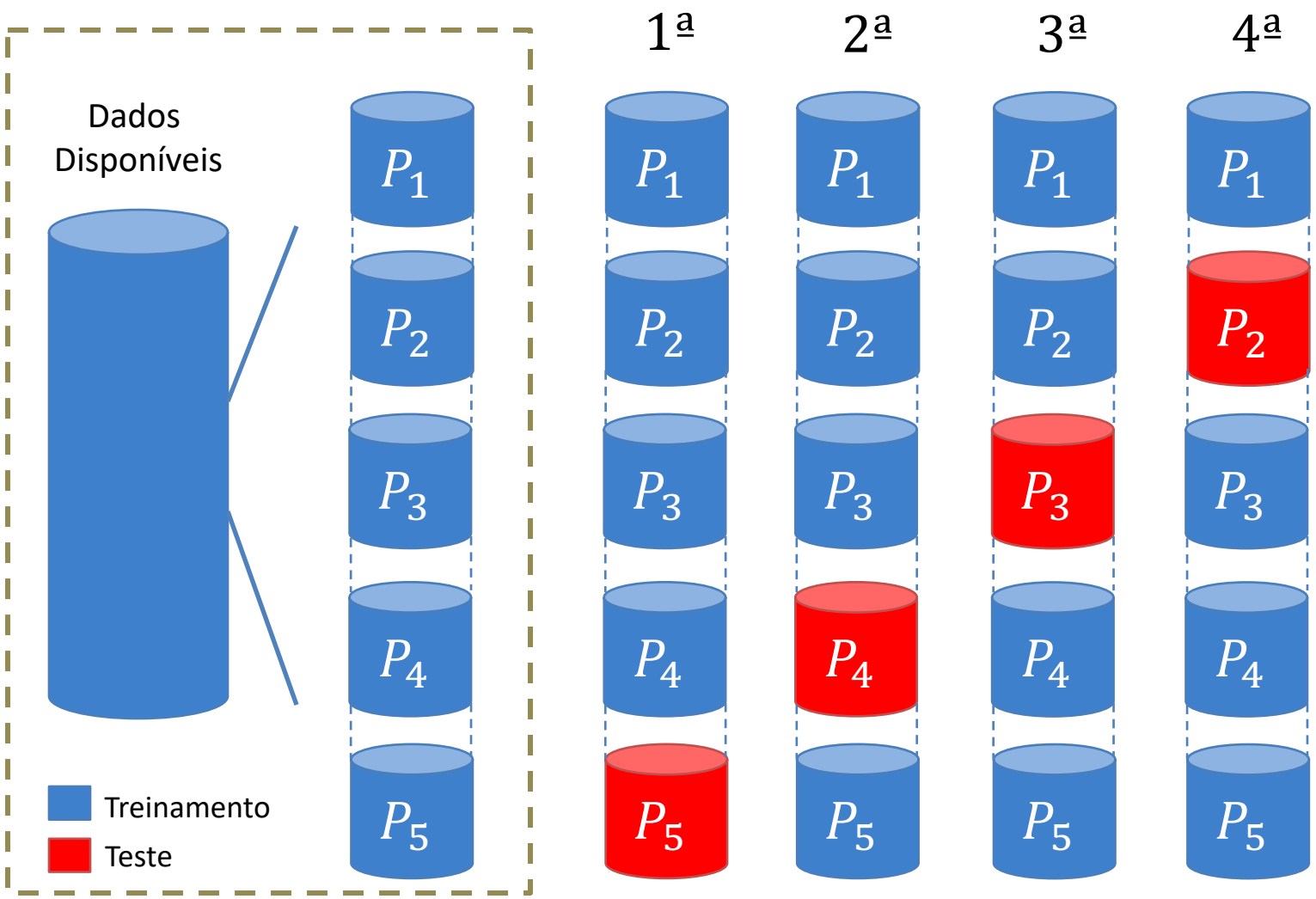
Cross-Validation

- Ex: 5-*fold cross-validation*



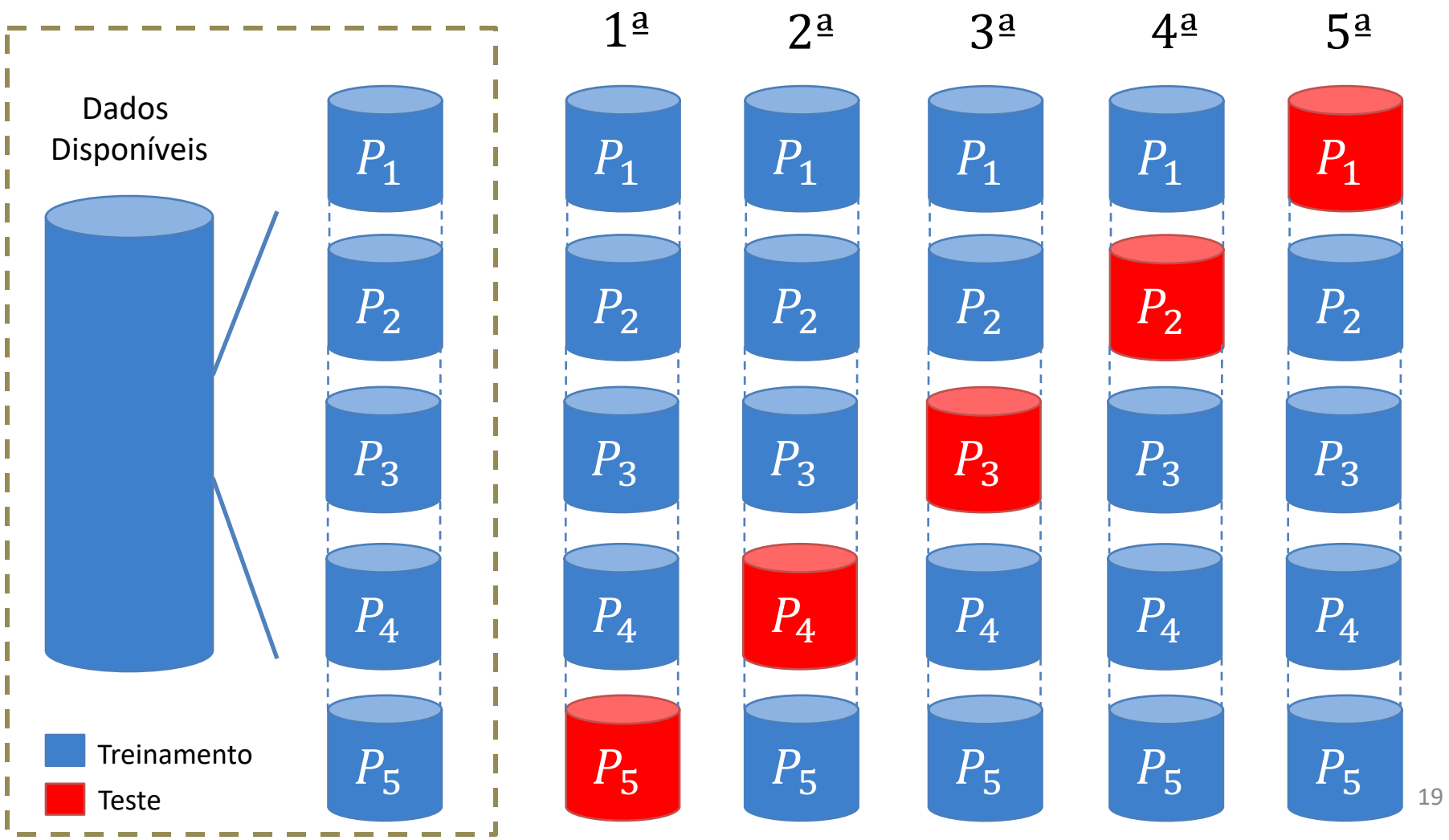
Cross-Validation

- Ex: 5-*fold cross-validation*



Cross-Validation

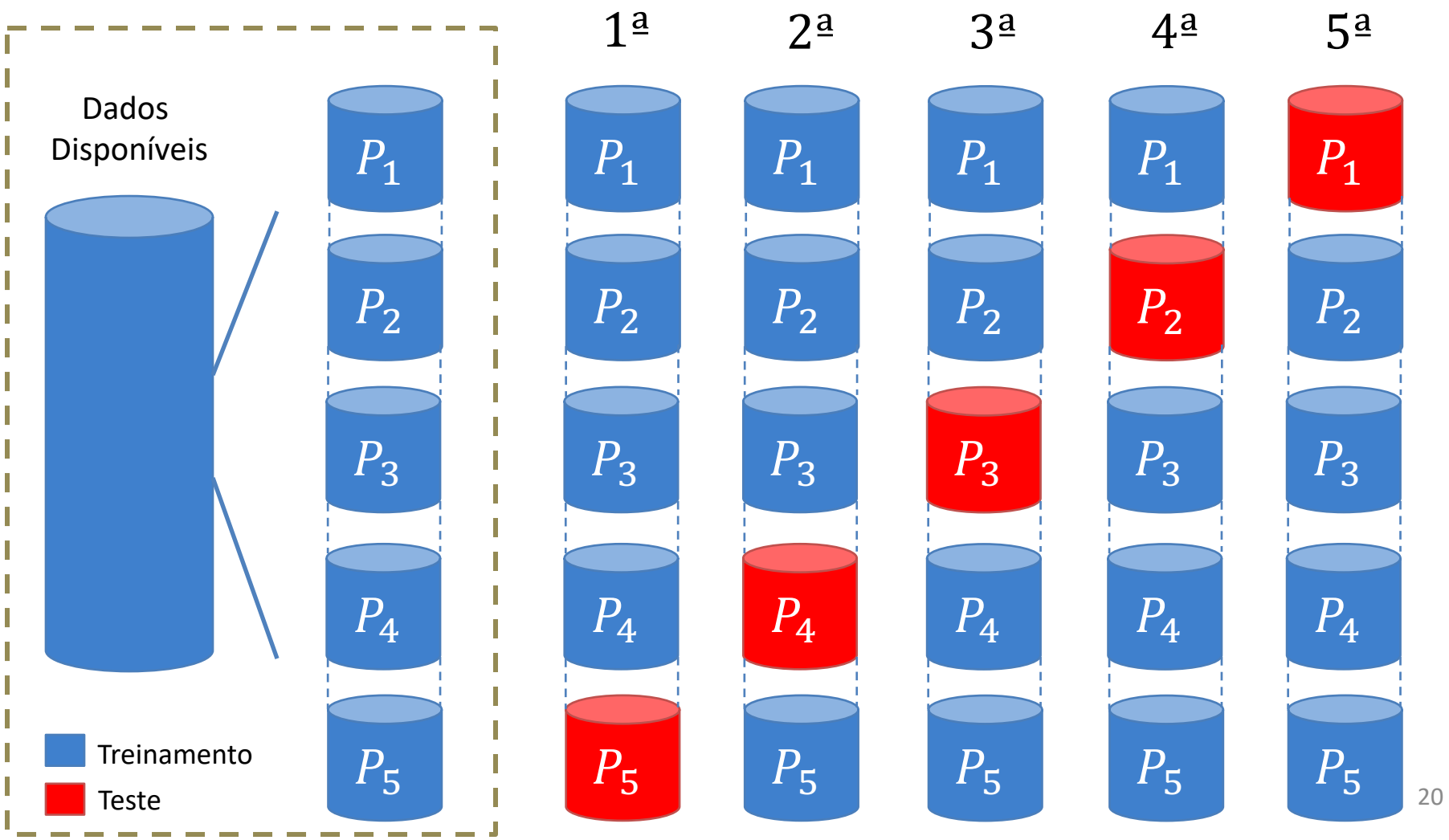
- Ex: 5-*fold cross-validation*



Cross-Validation

- Ex: 5-*fold cross-validation*

Para classificação, recomendado que seja estratificado!



Leave-one-out Cross-Validation

- *Leave-one-out* (LOO)
  - Caso particular de CV onde $k = N$
  - Cada iteração utiliza $(N - 1)$ objetos para treinar e apenas 1 objeto para teste
  - Assim como em k-fold CV, o erro estimado é dado pela média dos (N) erros de teste
  - Computacionalmente caro!!
    - Geralmente utilizado para pequenos conjuntos de objetos
    - Inviável para grandes conjuntos de dados
  - Gera estimativa de erro não-tendenciosa
    - Média das estimativas tende ao verdadeiro erro de generalização
    - Artigos científicos indicam que 10-fold CV aproxima LOO

5×2 *Cross-Validation*

- Amostragens de conjuntos de treinamento e teste com mesmo tamanho
- Equivalente a realizar 5 vezes CV de 2 folds variando o gerador de números aleatórios

```
Seja um  conjunto de N objetos
Para i=1 até 5
    Dividir N aleatoriamente em duas metades
    Usar metade 1 para treinamento e metade 2 para teste
    Usar metade 2 para treinamento e metade 1 para teste
```

- Executar mais do que 5×
  - Sobreposição dos conjuntos se torna tão grande que dificilmente adiciona nova informação
- Executar menos do que 5×
  - Não haverá objetos suficientes para ajustar uma distribuição e testar hipóteses (menos do que 10 partições)

T. G. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. **Neural Computation** (7), 1895-1923, 1998.

Bootstrap

- Funciona melhor que *cross-validation* para conjuntos muito pequenos
- Forma mais simples de *bootstrap*:
  - Em vez de usar sub-conjuntos dos dados, usa **sub-amostras**
    - Cada sub-amostra é amostrada **com reposição** do conjunto total de objetos
    - Cada sub-amostra tem o **mesmo número de objetos** do conjunto **original** e é utilizada para treinamento
    - Objetos **restantes** (não amostrados) são utilizados no **teste**

Bootstrap

- Se conjunto original tem N objetos
  - Amostra de tamanho N tende a ter $\approx 63{,}2\%$ dos objetos originais (demais $\approx 36{,}8\%$ são objetos duplicados)
- Processo é **repetido b vezes**
  - Resultado final = média dos b experimentos
- Existem diversas variações
  - Por exemplo, .632 bootstrap

.632 *Bootstrap*

- Existe intersecção entre as b sub-amostras de teste (cada sub-amostra tem $\approx 36{,}8\%$ dos objetos originais)
- Em vez de utilizar a média do erro nos b experimentos, ponderar o erro por sub-amostra de teste, e_i, juntamente com o erro de treinamento e_t

$$e_{0.632} = \frac{1}{b}\sum_{i=1}^{b}(0.632 \times e_i + 0.368 \times e_t)$$

Bradley Efron and Robert Tibshirani. Improvements on Cross-Validation: The .632+ Bootstrap Method. ***Journal of the American Statistical Association*** (92), 548-560, 1997.

Estimativa de Erro de Classificação

- Principal objetivo de um modelo supervisionado é **prever com sucesso** o valor de saída para objetos ainda não vistos
  - Errar o mínimo possível
- Para **quantificar** o desempenho preditivo (estimado) do modelo criado, existem diversas **medidas** na literatura
  - Cada medida tem um viés... (Teorema do NFL)
  - Para problemas de regressão:
    - Erro quadrático médio (com ou sem raiz)
    - Erro absoluto médio
    - ...
  - Para problemas de classificação:
    - Acurácia/Erro
    - Matriz de Confusão
    - Curvas PR e ROC
    - Kappa
    - ...

Taxa de Classificação Incorreta

- A medida clássica para estimar a taxa de erro de um classificador é denominada de **taxa de classificação incorreta** (*misclassification rate*), ou simplesmente **erro de classificação**
  - Proporção dos objetos de teste que são classificados incorretamente pelo classificador

$$erro = \frac{\#erros}{N_{teste}}$$

  - Usualmente é medida de forma indireta através do seu complemento, a **taxa de classificação correta**:

$$acuracia = \frac{\#acertos}{N_{teste}}$$

    - Acurácia
    - $acuracia = (1 - erro)$

Acurácia

- Do inglês, *Accuracy*
  - Dá **tratamento igual a todas as classes** do problema
  - **Não é** uma medida **adequada** para medir problemas com **classes desbalanceadas**
    - A medida privilegia a classe majoritária
    - Na vasta maioria dos problemas desbalanceados, a classe interessante (prioritária) é a classe rara =(
  - Ex: considere um problema de 2 classes
    - Classe 1 = 9990 objetos
    - Classe 2 = 10 objetos
      - Se modelo prevê apenas classe 1, acurácia será de 9990/10000 = 99.9%
      - Note que tal modelo não é sequer inteligente!!!

Tipos de Erros

- Em classificação binária, é comum nomear os objetos da classe de maior interesse de **positivos (+)**
  - Normalmente a classe rara ou minoritária
  - Demais objetos são nomeados **negativos (–)**
- Em alguns casos, os erros têm igual importância
- Em muitos casos, no entanto, **erros têm prioridades distintas (custos!)** considerando as possíveis consequências
  - Ex: diagnóstico negativo para indivíduo doente

Tipos de Erros

- Existem dois tipos de erro em classificação binária:
  - Classificar objeto negativo como positivo
    - **Falso Positivo** (FP), Alarme Falso
    - Erro do Tipo I
    - Ex: paciente diagnosticado como doente, embora esteja saudável
  - Classificar objeto positivo como negativo
    - **Falso Negativo** (FN)
    - Erro do Tipo II
    - Ex: paciente diagnosticado como saudável, mas está doente

Matriz de Confusão

- Também chamada de **Tabela de Contingência**
  - Permite a extração de **diversas medidas** de desempenho preditivo
  - Pode ser utilizada para distinguir os tipos de erros
  - Pode ser utilizada para problemas binários ou multi-classe

Classe Prevista	Classe Verdadeira		
	A	B	C
A	25	10	0
B	0	40	0
C	5	0	20

Matriz de Confusão

- Também chamada de **Tabela de Contingência**
  - Permite a extração de **diversas medidas** de desempenho preditivo
  - Pode ser utilizada para distinguir os tipos de erros
  - Pode ser utilizada para problemas binários ou multi-classe

Classe Prevista	Classe Verdadeira		
	A	B	C
A	25	10	0
B	0	40	0
C	5	0	20

Diagonal principal: acertos!

Matriz de Confusão

- Também chamada de **Tabela de Contingência**
  - Permite a extração de **diversas medidas** de desempenho preditivo
  - Pode ser utilizada para distinguir os tipos de erros
  - Pode ser utilizada para problemas binários ou multi-classe

Classe Prevista	Classe Verdadeira		
	A	B	C
A	25	10	0
B	0	40	0
C	5	0	20

Valores fora da diagonal principal: erros!

Matriz de Confusão

- Também chamada de **Tabela de Contingência**
  - Permite a extração de **diversas medidas** de desempenho preditivo
  - Pode ser utilizada para distinguir os tipos de erros
  - Pode ser utilizada para problemas binários ou multi-classe

Classe Prevista	Classe Verdadeira		
	A	B	C
A	25	10	0
B	0	40	0
C	5	0	20

Acurácia: $\frac{25 + 40 + 20}{25 + 40 + 20 + 10 + 5} = \frac{85}{100} = 0.85 \text{ ou } 85\%$

Matriz de Confusão Binária

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Matriz de Confusão Binária

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Acurácia:

$$\frac{VP + VN}{VP + VN + FP + FN}$$

Matriz de Confusão Binária

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Acurácia: $\frac{VP + VN}{VP + VN + FP + FN}$

Erro: $\frac{FP + FN}{VP + VN + FP + FN} = (1 - acurácia)$

Matriz de Confusão Binária

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Erro do Tipo I:
(TFP)
(Taxa de Alarmes Falsos)
(Custo)

$$\frac{FP}{FP + VN}$$

Matriz de Confusão Binária

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Erro do Tipo II: (TFN)

$$\frac{FN}{FN + VP}$$

Exercício

- Avalie os 3 classificadores abaixo:

Classe Prevista	Classe Verdadeira	
	P	N
P	25	10
N	45	60

Classe Prevista	Classe Verdadeira	
	P	N
P	70	20
N	15	30

Classe Prevista	Classe Verdadeira	
	P	N
P	70	95
N	30	5

Classificador 1	
Acurácia =	
Erro =	
TFN =	
TFP =	

Classificador 2	
Acurácia =	
Erro =	
TFN =	
TFP =	

Classificador 3	
Acurácia =	
Erro =	
TFN =	
TFP =	

Matriz de Confusão Binária

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Especificidade: (TVN)

$$\frac{VN}{FP + VN} = (1 - TFP)$$

Matriz de Confusão Binária

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Sensibilidade:
(TVP)
(*Recall*, Revocação, Benefício)

$$\frac{VP}{FN + VP} = (1 - TFN)$$

Matriz de Confusão Binária

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	70	40
Negativa	30	60

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Precisão:
(*Precision*)

$$\frac{VP}{FP + VP}$$

Precision x *Recall*

Precisão:
(*Precision*)

$$\frac{VP}{FP + VP}$$

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

O que acontece se um modelo classificar todos exemplos como sendo positivos?

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Revocação:
(Recall)

$$\frac{VP}{FN + VP}$$

Precision x *Recall*

Precisão:
(*Precision*)

$$\frac{VP}{FP + VP}$$

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

O que acontece se um modelo classificar todos exemplos como sendo positivos?

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Revocação:
(Recall)

$$\frac{VP}{\cancel{FN} + VP}$$

máximo

Precision x *Recall*

Precisão:
(*Precision*)

$$\frac{VP}{FP + VP}$$

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

O que acontece se um modelo classificar todos exemplos como sendo positivos?

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

Revocação:
(Recall)

$$\frac{VP}{\cancel{FN} + VP}$$

máximo

F-Measure

- Média harmônica de *precision* e *recall*
  - Também conhecida como F_1 *score* ou F-score

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

Resumo das Medidas Apresentadas

$$\textit{Acurácia} = \frac{VP + VN}{VP + FP + VN + FN}$$

$$\textit{Erro} \\ (1 - Acurácia) = \frac{FP + FN}{VP + FP + VN + FN}$$

$$\text{Especificidade} \\ (\text{TVN}, 1 - TFP) = \frac{VN}{FP + VN}$$

$$\text{TFP} \\ (\text{Erro tipo I, Custo}) = \frac{FP}{FP + VN}$$

$$\textit{Recall} \\ (\text{TVP, Sensibilidade, Benefício}) = \frac{VP}{FN + VP}$$

$$\text{TFN} \\ (\text{Erro tipo II}, \\ 1 - Recall) = \frac{FN}{FN + VP}$$

$$\textit{Precision} = \frac{VP}{VP + FP}$$

$$F_1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

Gráficos ROC

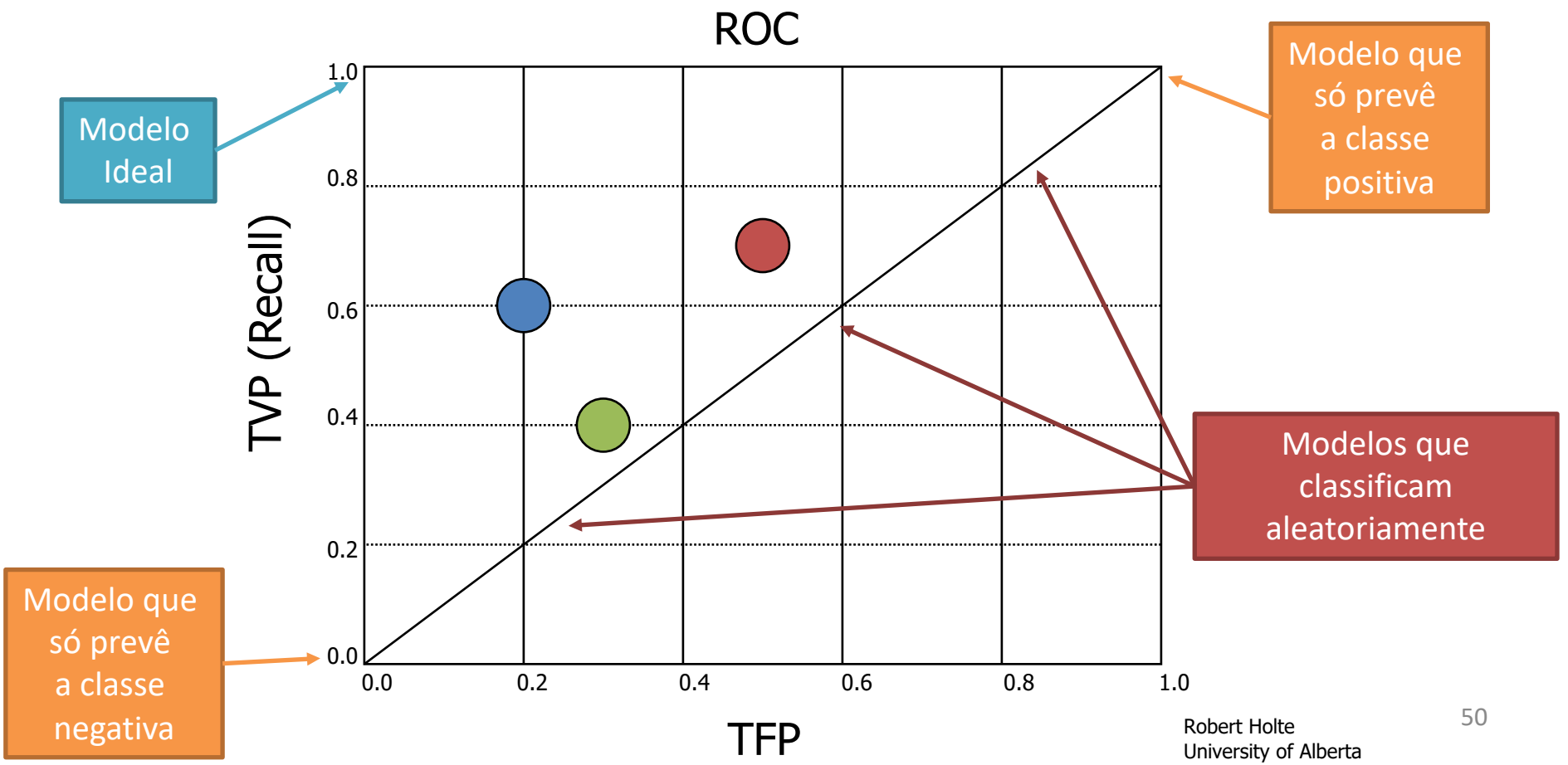
- Do inglês, *Receiver Operating Characteristics*
- Medida de desempenho originária da área de processamento de sinais
  - Muito utilizada na área médica (e na biologia em geral)
  - Mostra relação entre custo (TFP, Erro do Tipo I) e benefício (TVP, *Recall*)
  - Lembre-se que:
    - TFP é a taxa de alarmes falsos (erros na classe negativa, Erro do Tipo I)
    - TVP é a taxa de acertos na classe positiva (1 – Erro do Tipo II)

Imagine a existência de três modelos:

Modelo 1	
TFP	0.3
TVP	0.4

Modelo 2	
TFP	0.5
TVP	0.7

Modelo 3	
TFP	0.2
TVP	0.6



Gráficos ROC

- Resumindo:
  - Classificador **ideal** mais a **noroeste**
  - Classificadores próximos ao **canto inferior esquerdo** são **conservadores**
    - Apenas detectam a classe positiva com forte evidência
    - Portanto, cometem poucos FPs
  - Classificadores próximos ao **canto superior direito** são **liberais**
    - Detectam a classe positiva com pouca evidência
    - Correm o risco de alta taxa de alarme falso
  - Classificadores ao **redor da linha central** tem comportamento **similar** ao esperado de **classificação aleatória**

Gráficos ROC

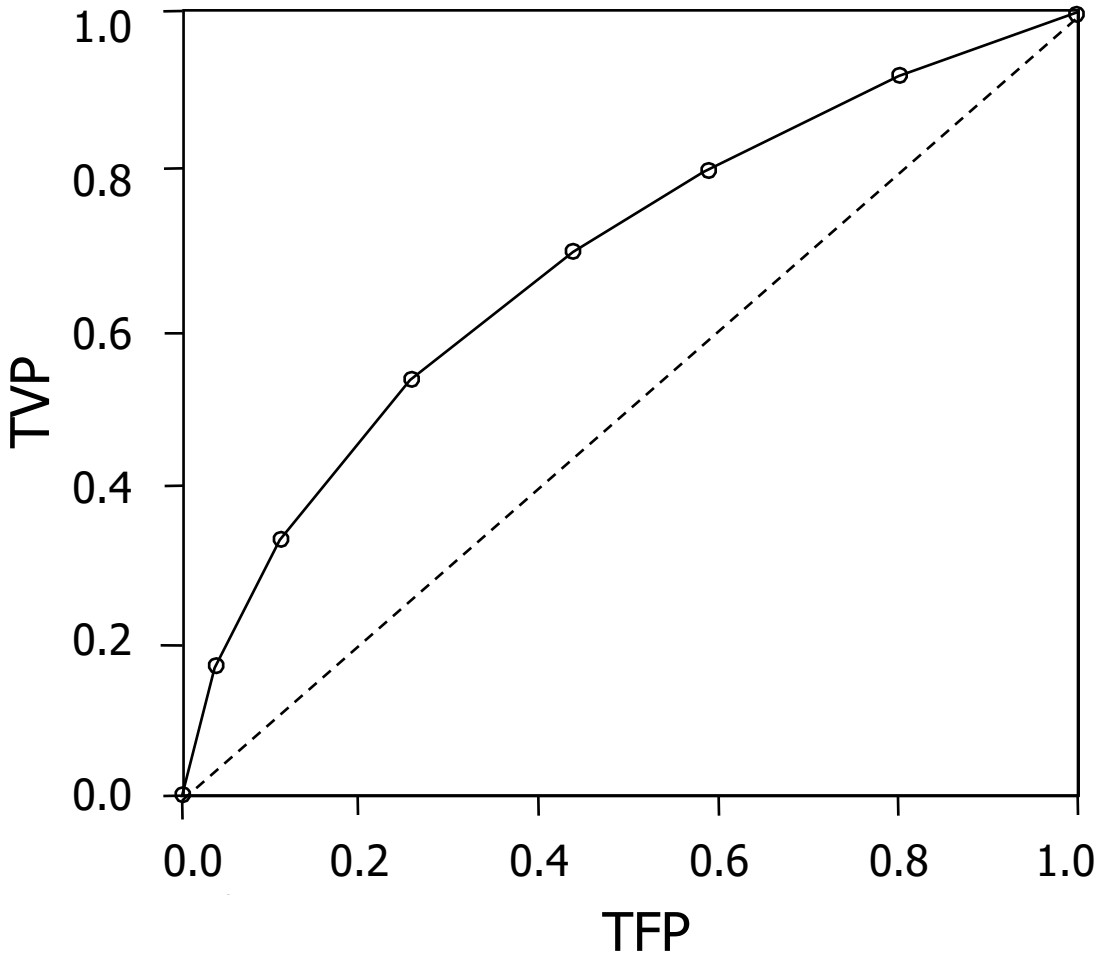
- Alguns modelos possuem **saída discreta**
  - Árvores de Decisão, regras, SVMs…
    - Atribuem cada objeto a uma das classes
    - Produzem um **único ponto no gráfico ROC**
- Outros modelos geram como saída um **escore** (e.g., **probabilidade**) associado a cada classe
  - Naïve Bayes, Redes Neurais...
  - Permitem gerar uma **curva no gráfico ROC**
- Curvas ROC permitem uma **melhor comparação de classificadores**
  - São insensíveis a mudanças na distribuição das classes no conjunto de teste

Curvas ROC

- Algoritmos que geram escores:
  - Diferentes valores de **limiar** para os escores associados à classe positiva podem ser utilizados para **gerar um classificador (modelo)**
    - Cada valor de limiar gera diferentes valores de TVP e TFP, correspondendo a um **ponto distinto** no gráfico ROC
    - Ligação (**interpolação**) dos pontos gera uma curva ROC

Curvas ROC

- 8 limiares utilizados gerando 8 pontos no gráfico ROC
- Os pontos são interpolados para gerar a curva ROC



Curvas ROC

Objeto	Classe Real	Escore (Classe +)
$\mathbf{x}^{(6)}$	$+$	0.9
$\mathbf{x}^{(3)}$	$+$	0.8
$\mathbf{x}^{(2)}$	$-$	0.7
$\mathbf{x}^{(9)}$	$+$	0.6
$\mathbf{x}^{(5)}$	$+$	0.6
$\mathbf{x}^{(1)}$	$-$	0.5
$\mathbf{x}^{(7)}$	$-$	0.3
$\mathbf{x}^{(8)}$	$-$	0.2
$\mathbf{x}^{(4)}$	$-$	0.2
$\mathbf{x}^{(10)}$	$-$	0.1

1. Ordenar objetos em ordem decrescente de escore para a classe positiva $(+)$
2. Para cada limiar de decisão θ:
   i. Classificar todos os objetos
   ii. Calcular VP, VN, FP, FN
   iii. Calcular TVP e TFP e plotar ponto no gráfico ROC

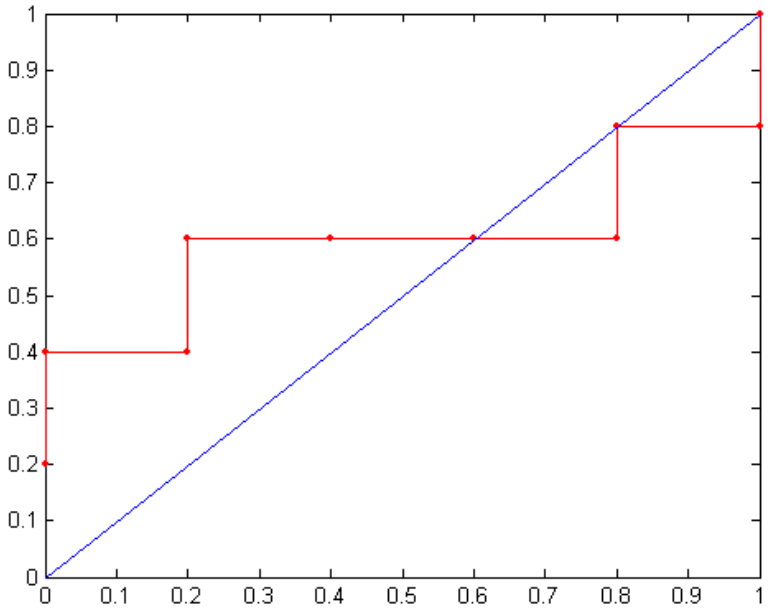
$$Classe = \begin{cases} escore \geq \theta: + \\ escore < \theta: - \end{cases}$$

Curvas ROC

Tan et al. 2005

	Class	+	-	+	-	-	-	+	-	+	+	
Threshold ≥		0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
	TP	5	4	4	3	3	3	3	2	2	1	0
	FP	5	5	4	4	3	2	1	1	0	0	0
	TN	0	0	1	1	2	3	4	4	5	5	5
	FN	0	1	1	2	2	2	2	3	3	4	5
→	TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
→	FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

ROC Curve



Curvas ROC

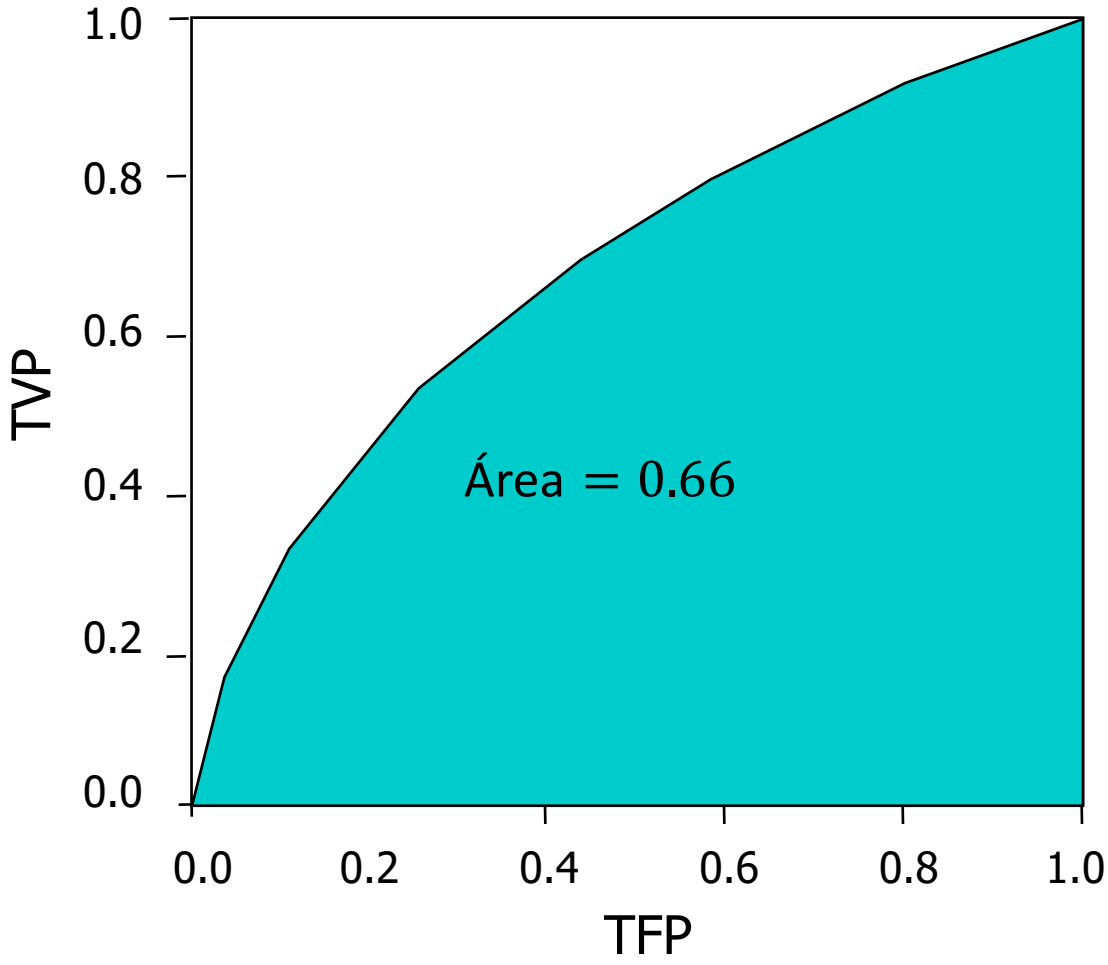
- Algoritmos que geram valores discretos:
  - Podem ser **modificados** para gerar escores
    - Para ADs, pode se utilizar a **fração dos objetos** de treinamento **positivos** do **nó folha** correspondente como **escore**
    - Para k-NN, pode se utilizar a **fração dos k vizinhos** mais próximos que pertencem à classe **positiva** como **escore**
    - Para SVMs, pode se utilizar a **distância** normalizada do **objeto** ao **hiperplano separador** como **escore**
    - …
  - Podem ser **combinados em comitês**
    - Algoritmo é executado sobre **amostragens** do conjunto de **treinamento**, gerando **múltiplos modelos**
    - Cada modelo prevê uma das duas classes ($+$ ou $-$)
    - O escore será a **fração dos modelos** que **previram a classe positiva**

Área sob a Curva ROC (AUC)

- Fornece uma estimativa do desempenho de classificadores
- Valor contínuo no intervalo $[0, 1]$
  - Quanto maior melhor
  - Adição de áreas de sucessivos trapézios

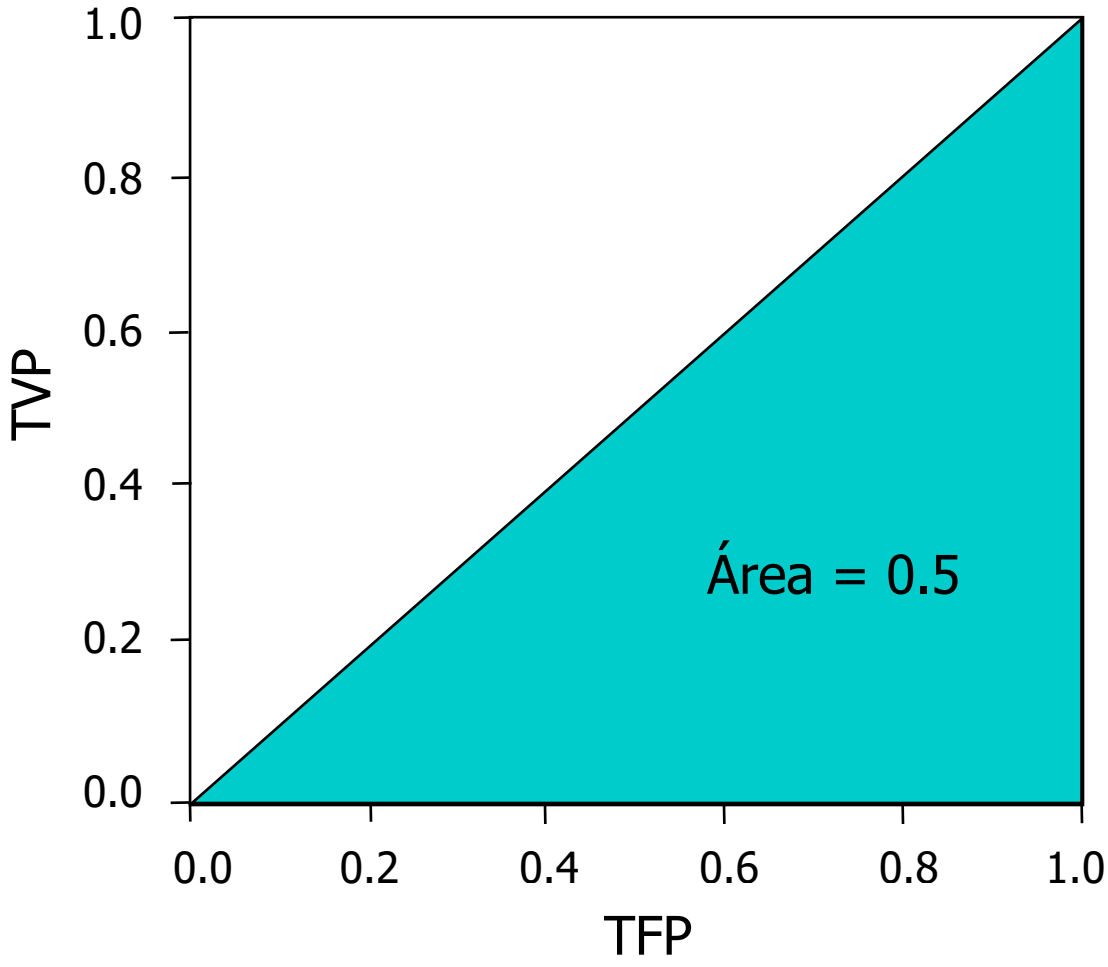
- É possível provar que a AUC equivale à **probabilidade** do modelo **atribuir um escore** $P(+ \mid \mathbf{x})$ **maior** a um **objeto positivo** escolhido aleatoriamente **do que a um objeto negativo** escolhido aleatoriamente

Área sob a Curva ROC (AUC)



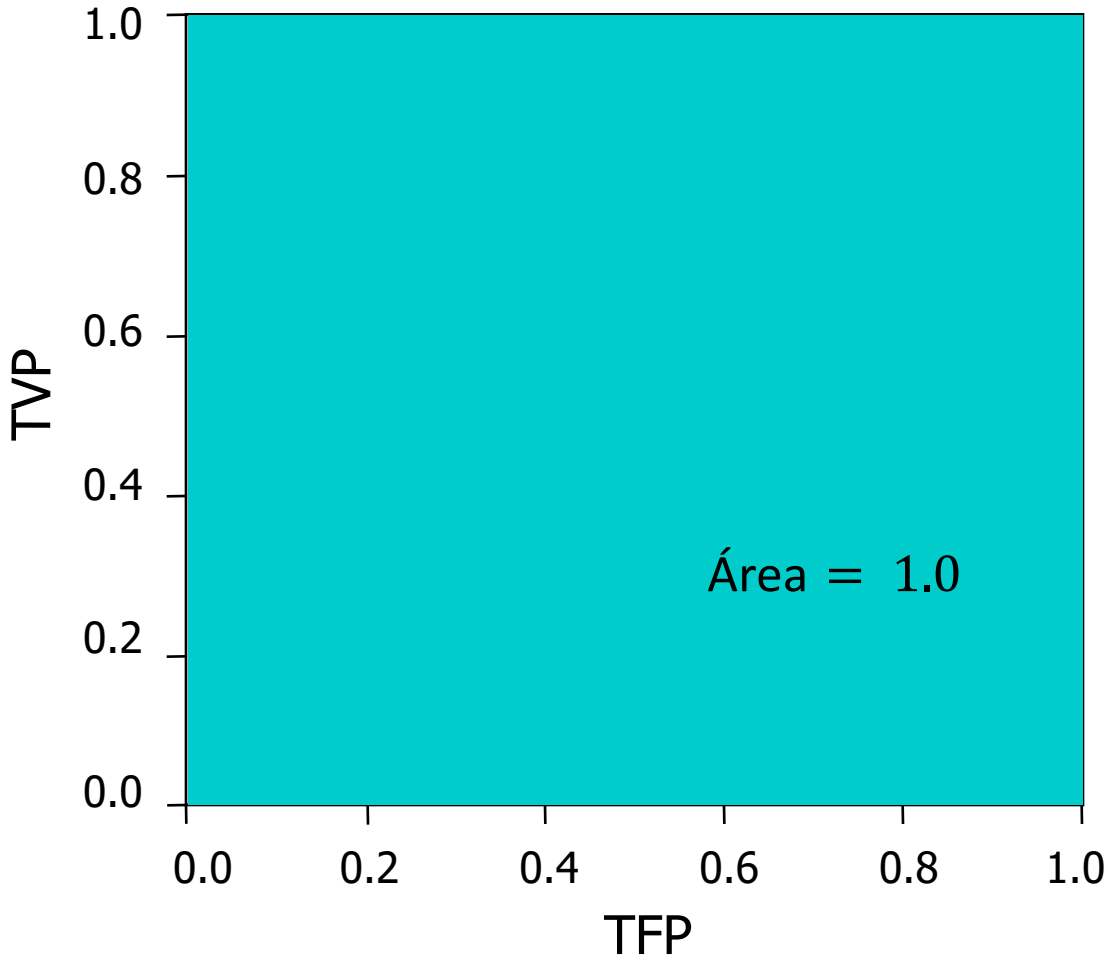
Área Sob Curvas ROC

Modelo Inútil
(Aleatório)

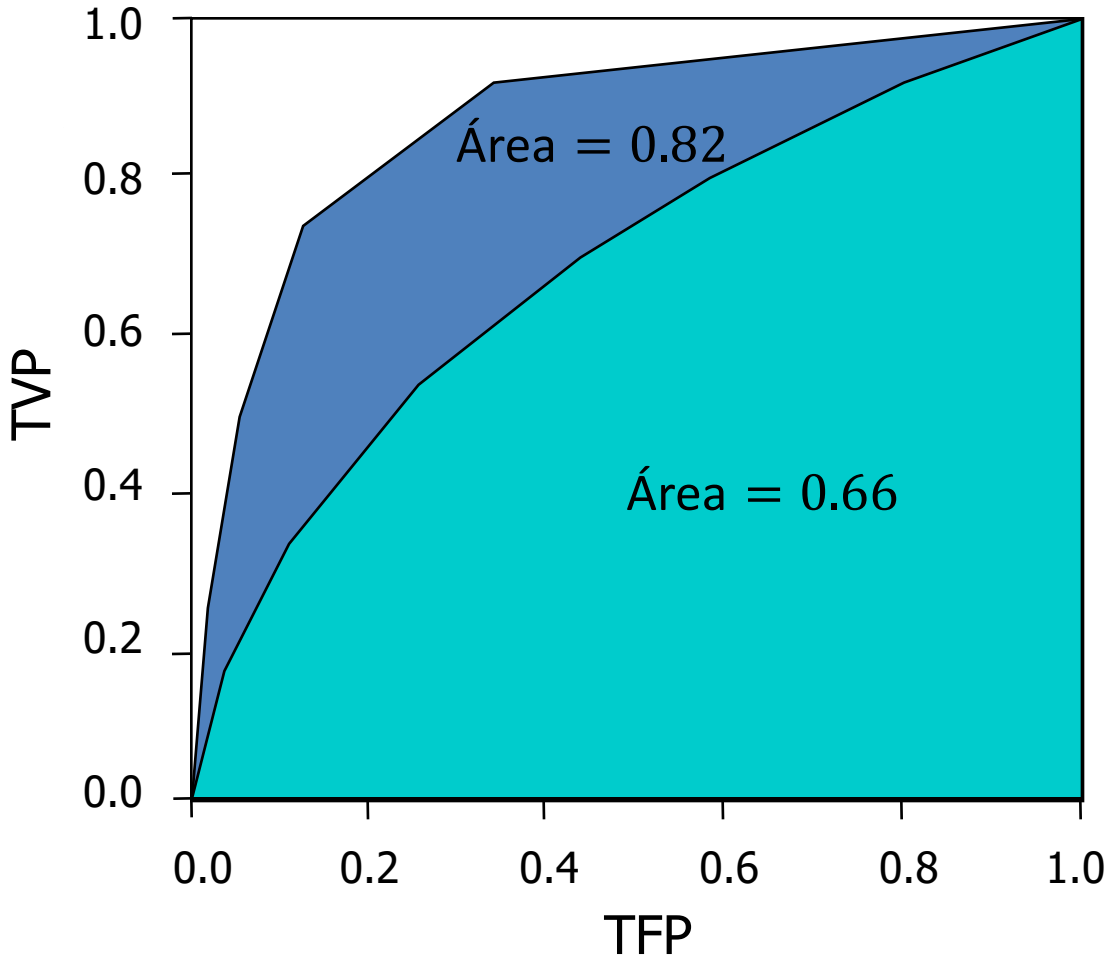


Área sob a Curva ROC (AUC)

Modelo Ideal



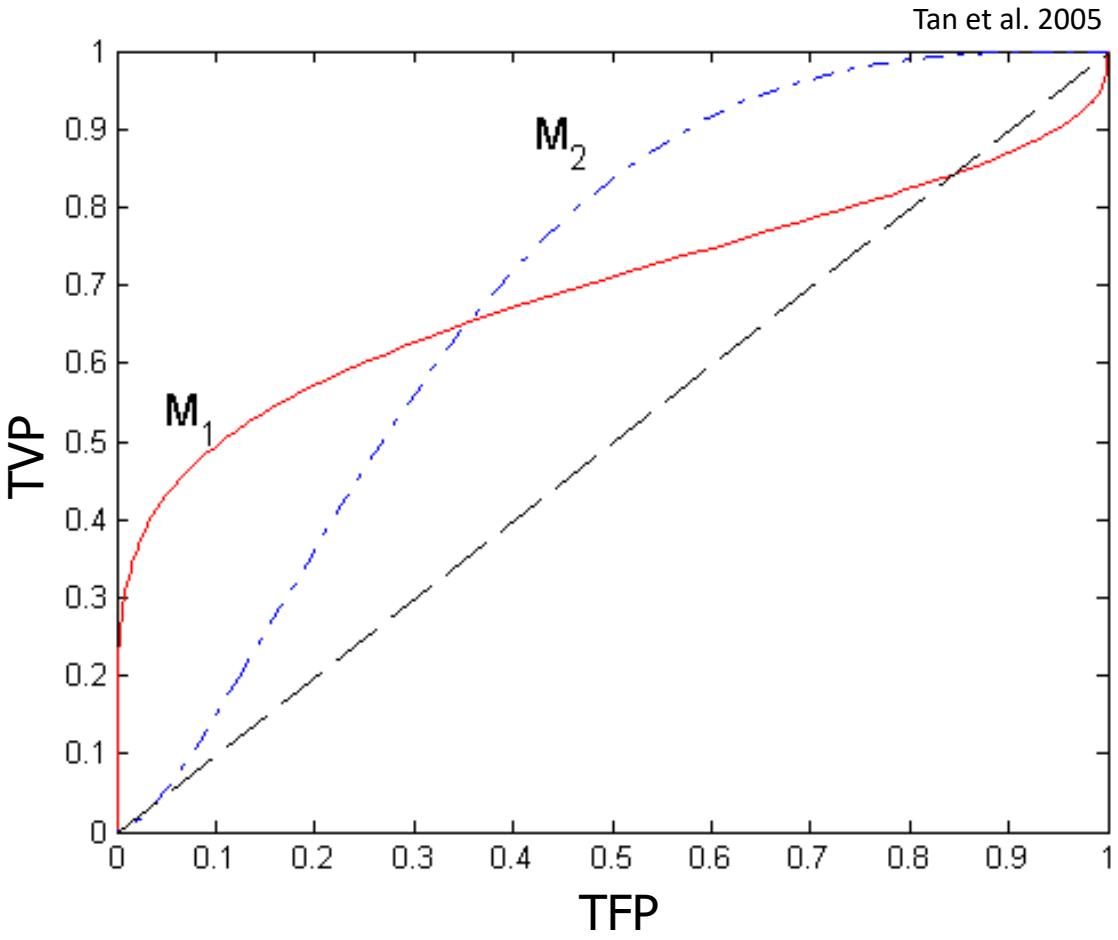
Área sob a Curva ROC (AUC)



Área sob a Curva ROC (AUC)

- **Nota 1:** um modelo com maior AUC pode apresentar AUC pior em trechos da curva...
  - AUC **não deve ser vista** como **critério absoluto**
  - Deve ser vista como medida de desempenho auxiliar as demais, com suas vantagens e desvantagens

Área sob a Curva ROC (AUC)



- Modelos similares em desempenho preditivo
  - M_1 é melhor para cenários conservadores
  - M_2 é melhor para cenários liberais

Área sob a Curva ROC (AUC)

- **Nota 2:**
  - Para **maior confiabilidade da análise**, calcula-se a AUC utilizando-se algum dos procedimentos de avaliação de desempenho vistos anteriormente (e.g., *cross-validation*) para gerar **múltiplas curvas ROC**
    - AUC mais confiável é tomada a partir de algum tipo de **média** das **AUCs** previamente calculadas, **ou** a partir de uma **curva média**
    - A **variância** das curvas também é um fator a ser analisado

Área sob a Curva ROC (AUC)

- **Nota 3:**
  - Distribuição das classes é dada pela proporção entre os valores da 1ª e 2ª colunas da matriz de confusão

Classe Prevista	Classe Verdadeira	
	Positiva	Negativa
Positiva	VP	FP
Negativa	FN	VN

  - Observe que a quantidade de objetos em cada classe não afeta o gráfico ROC
    - TVP e TFP são taxas e calculadas por coluna
      - Objetos da classe negativa não afetam o cálculo de TVP assim como objetos da classe positiva não afetam o cálculo de TFP
  - Logo, gráficos ROC são insensíveis à distribuição das classes
    - Análise robusta ao problema de desbalanceamento!

Área sob a Curva ROC (AUC)

- **Nota 4:**
  - Existem análises ROC para **problemas multi-classe** (mais do que duas classes), porém são muito mais complexas do que para problemas binários
  - Por exemplo, pode-se considerar as **relações ROC** existentes entre cada **par de classes**...
  - Outra opção é considerar as **relações ROC** existentes entre cada classe e as demais classes
    - **Uma classe** é vista como **positiva** e as **demais** como **negativa**

Sugestão de Leituras

- Seções 4.5, 4.6 (Tan et al., 2006)
- Capítulo 9 (Faceli et al., 2011)

Créditos e Referências

Slides adaptados dos originais gentilmente cedidos por:

- André Carvalho (ICMC-USP)
- Ricardo Campello (ICMC-USP)

- Tan, P. N., Steinbach, M., Kumar, V. **Introduction to Data Mining.** Addison-Wesley, 2005. 769 p.

- Faceli et al. **Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina**. LTC, 2011. 378 p.