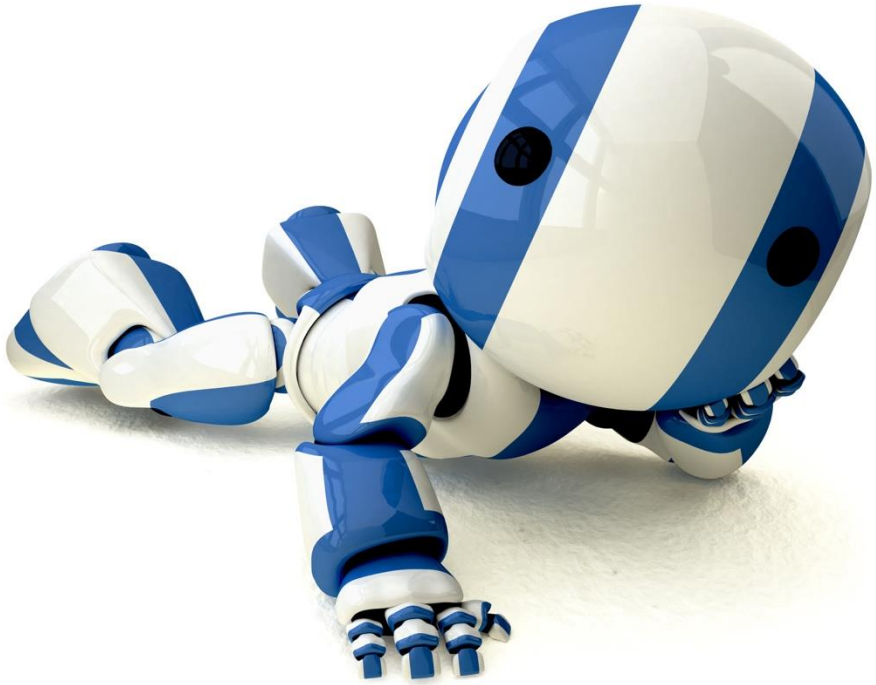
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA

Aprendizado de Máquina

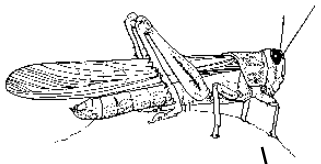


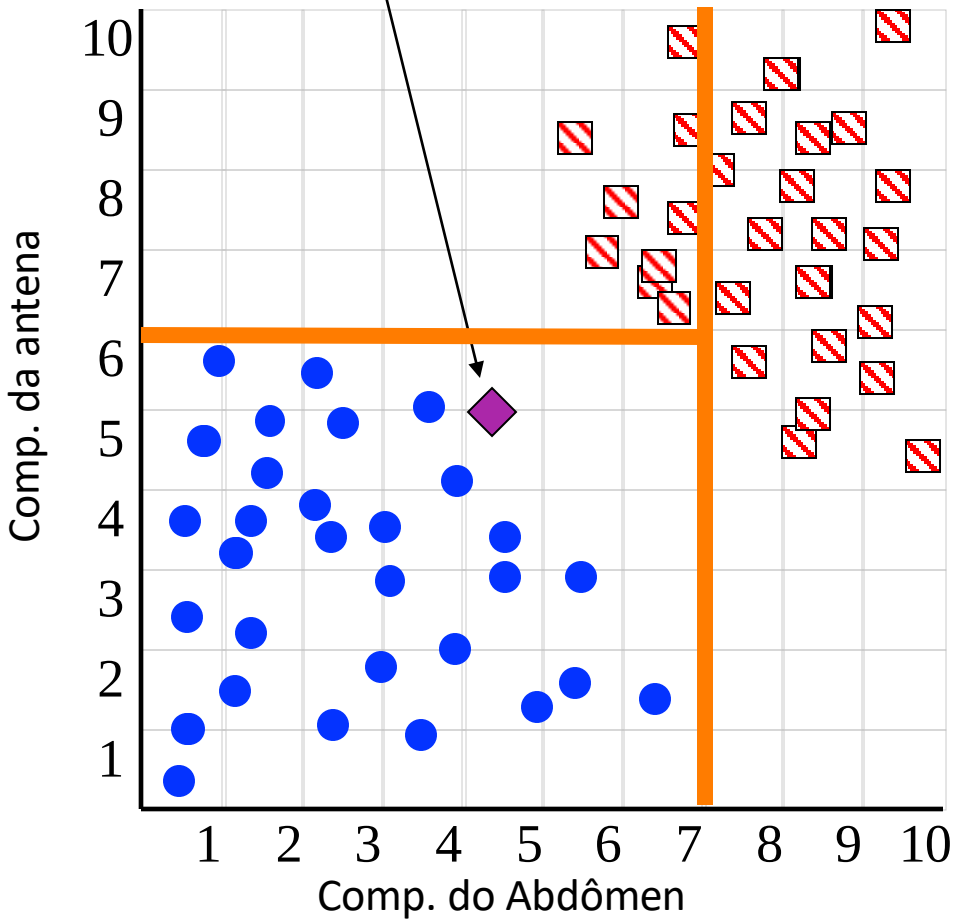
Aprendizado Supervisionado IV
Avaliação de Desempenho de Modelos Supervisionados

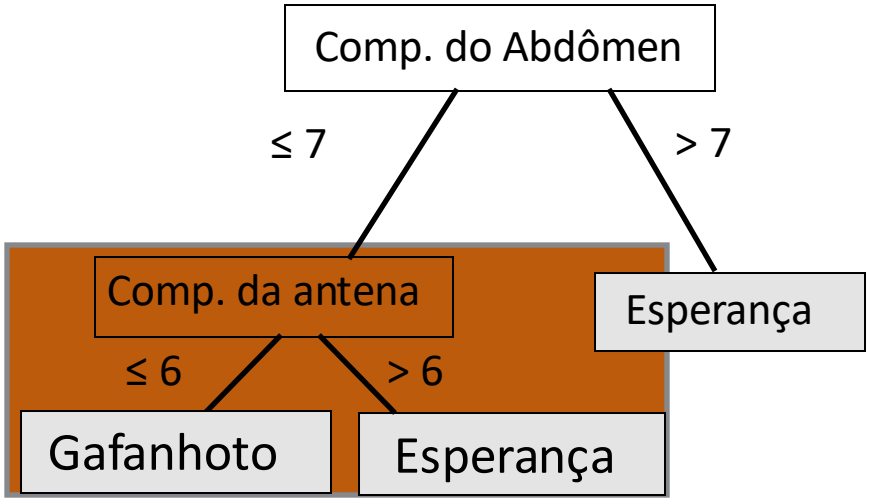
Prof. Me. Otávio Parraga

MALTA
Machine Learning Theory and Applications Lab

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

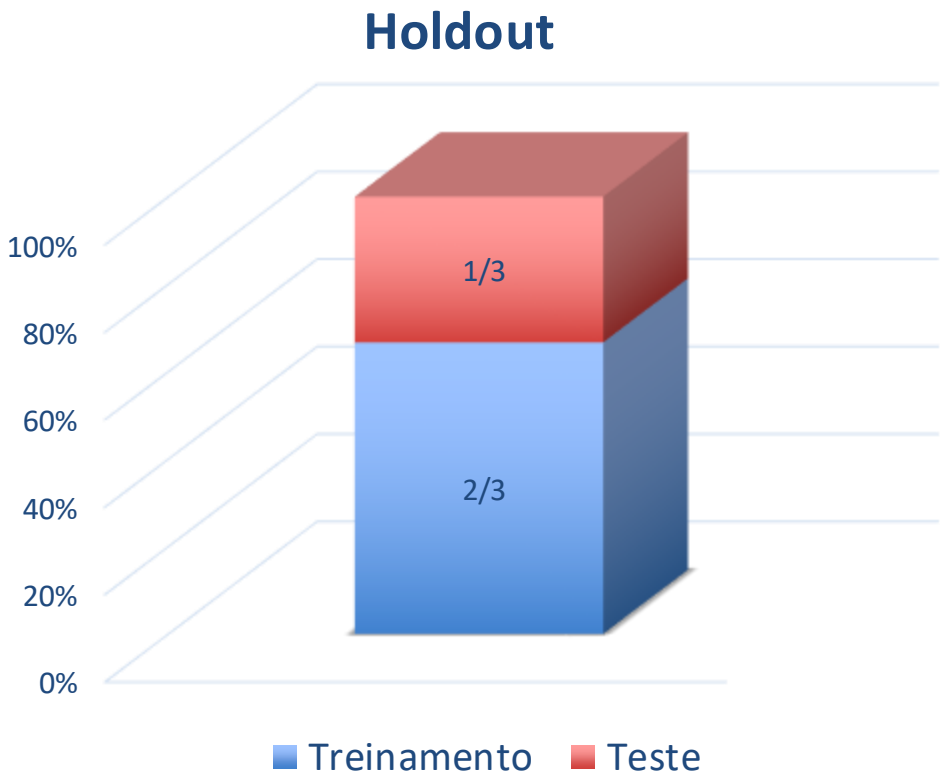
- Para estimarmos de maneira correta o desempenho do modelo, precisamos seguir um protocolo bem definido
  - Separar dados cujo atributo alvo é conhecido em dois conjuntos mutuamente exclusivos: **treinamento e teste**
  - **Jamais** avaliar o desempenho de um modelo em dados utilizados para seu treinamento,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

Holdout

- Também conhecido como *split-sample*
- Técnica mais simples para divisão de dados
- Faz uma única divisão (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 divisões 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)

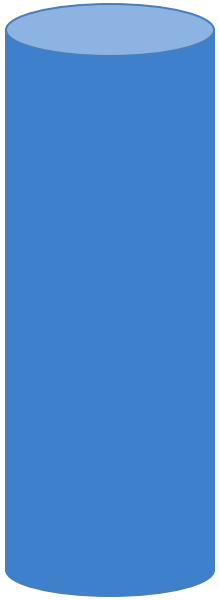
Cross-Validation

- 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

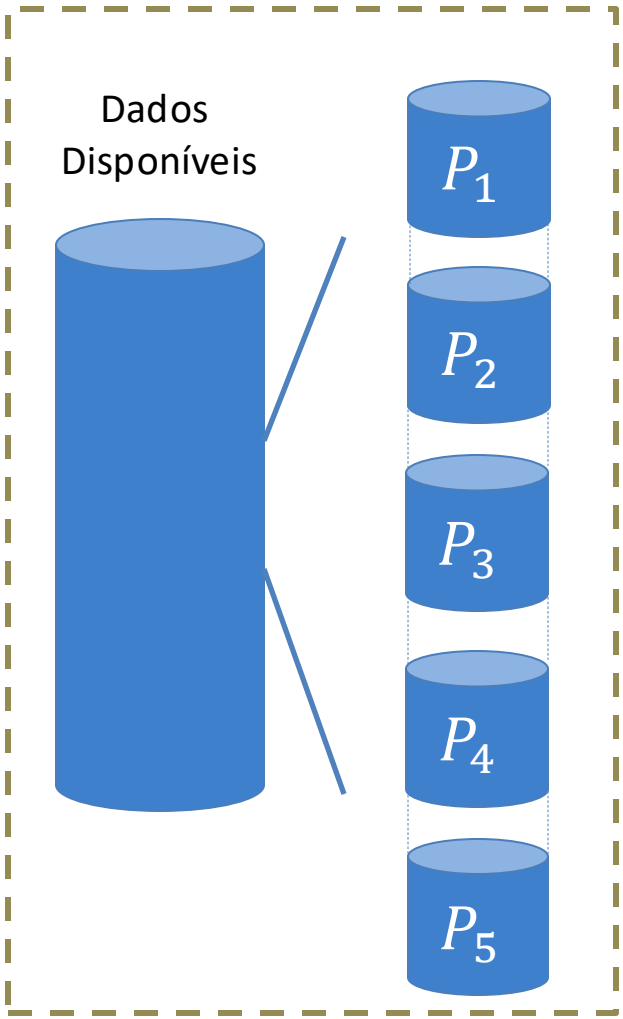
- Ex: 5-*fold cross-validation*

Dados Disponíveis



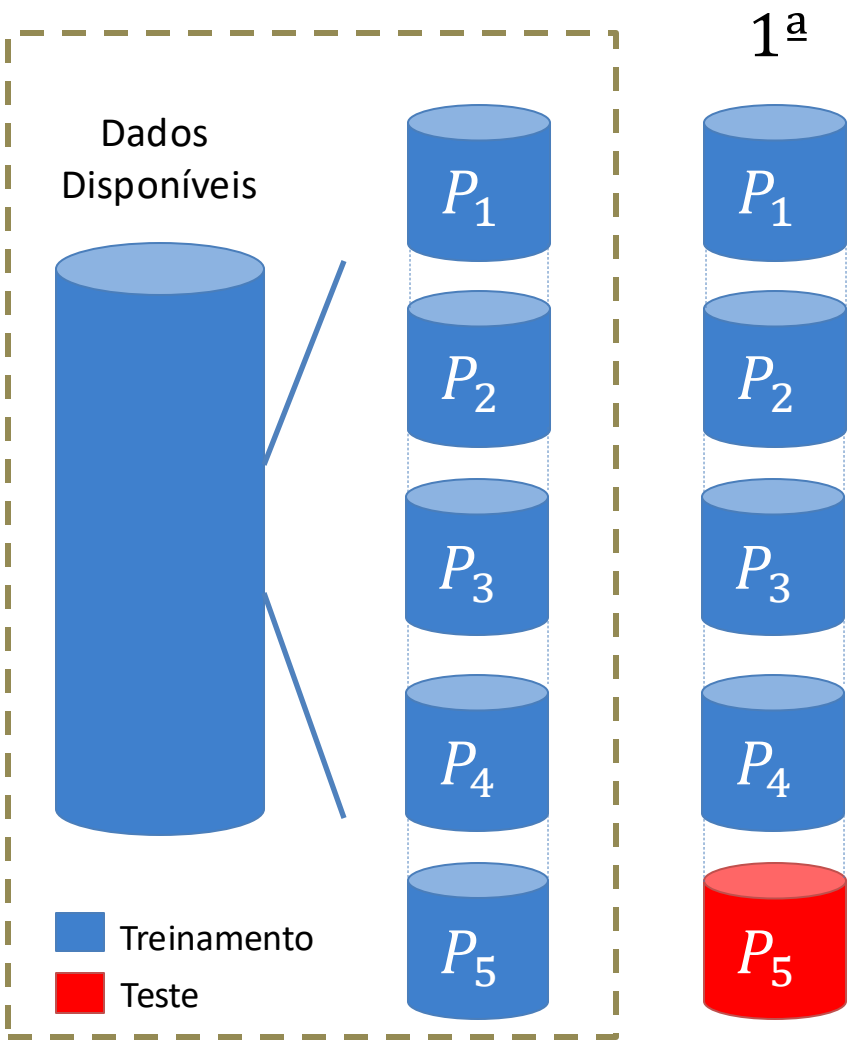
Cross-Validation

- Ex: 5-*fold cross-validation*



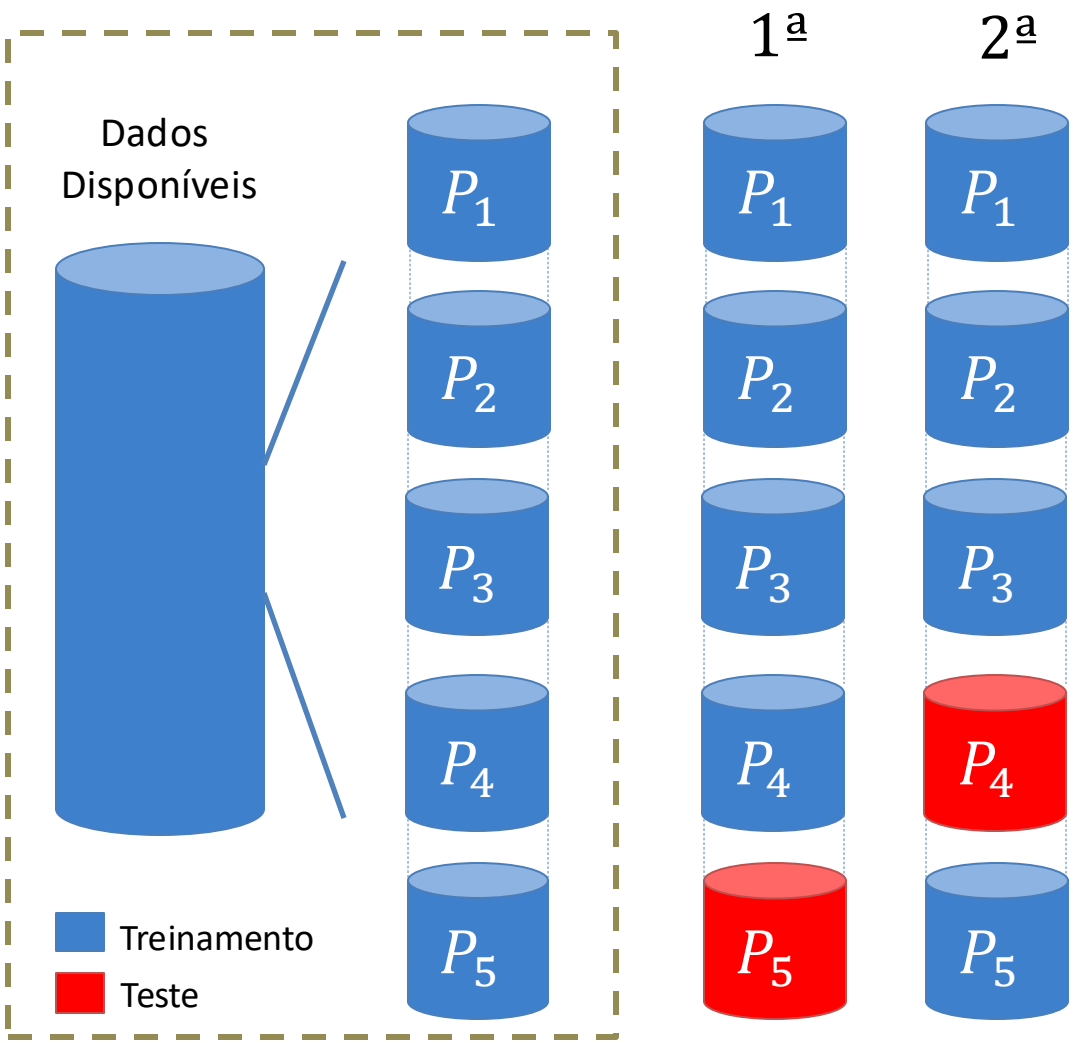
Cross-Validation

- Ex: 5-*fold cross-validation*



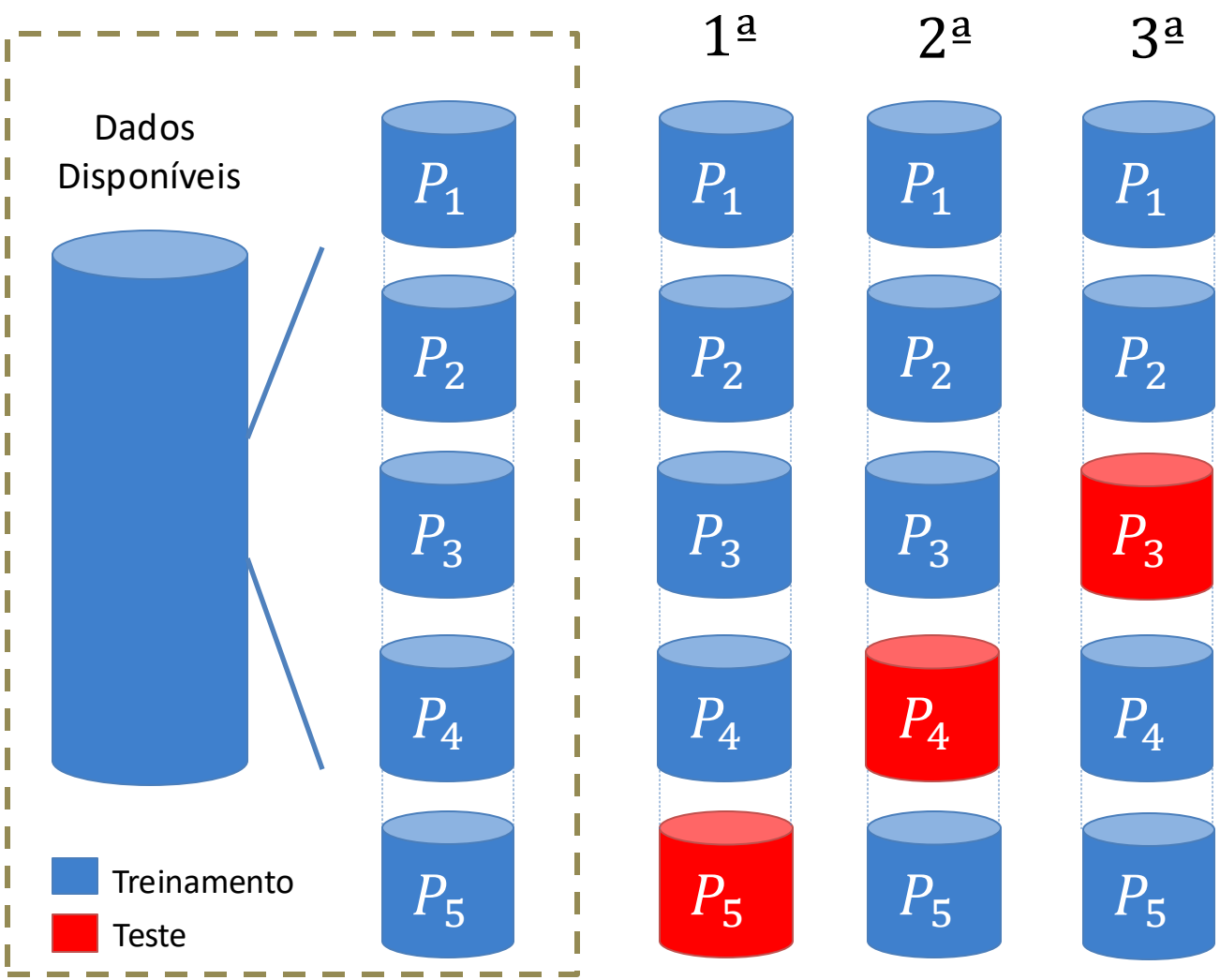
Cross-Validation

- Ex: 5-*fold cross-validation*



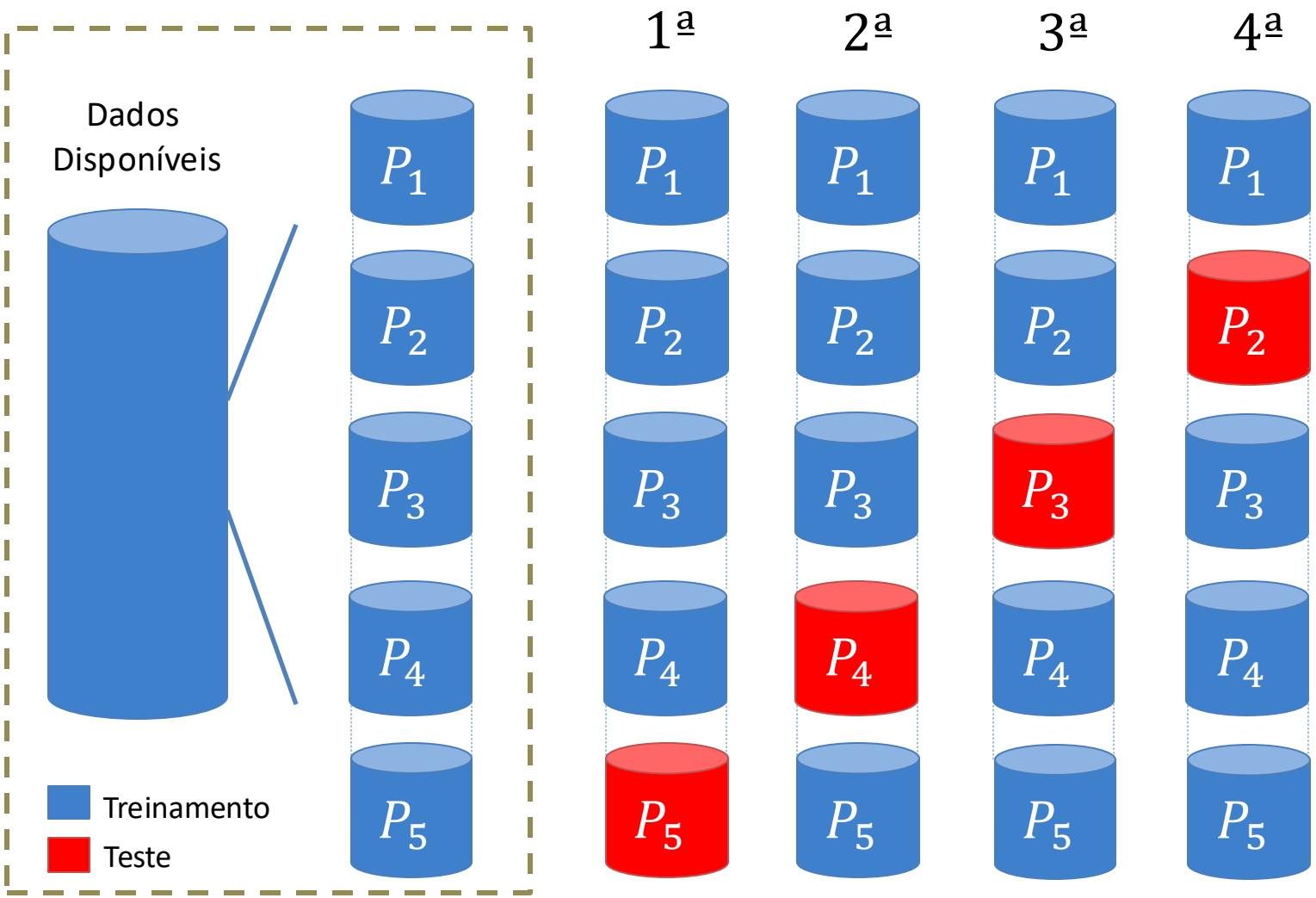
Cross-Validation

- Ex: 5-*fold cross-validation*



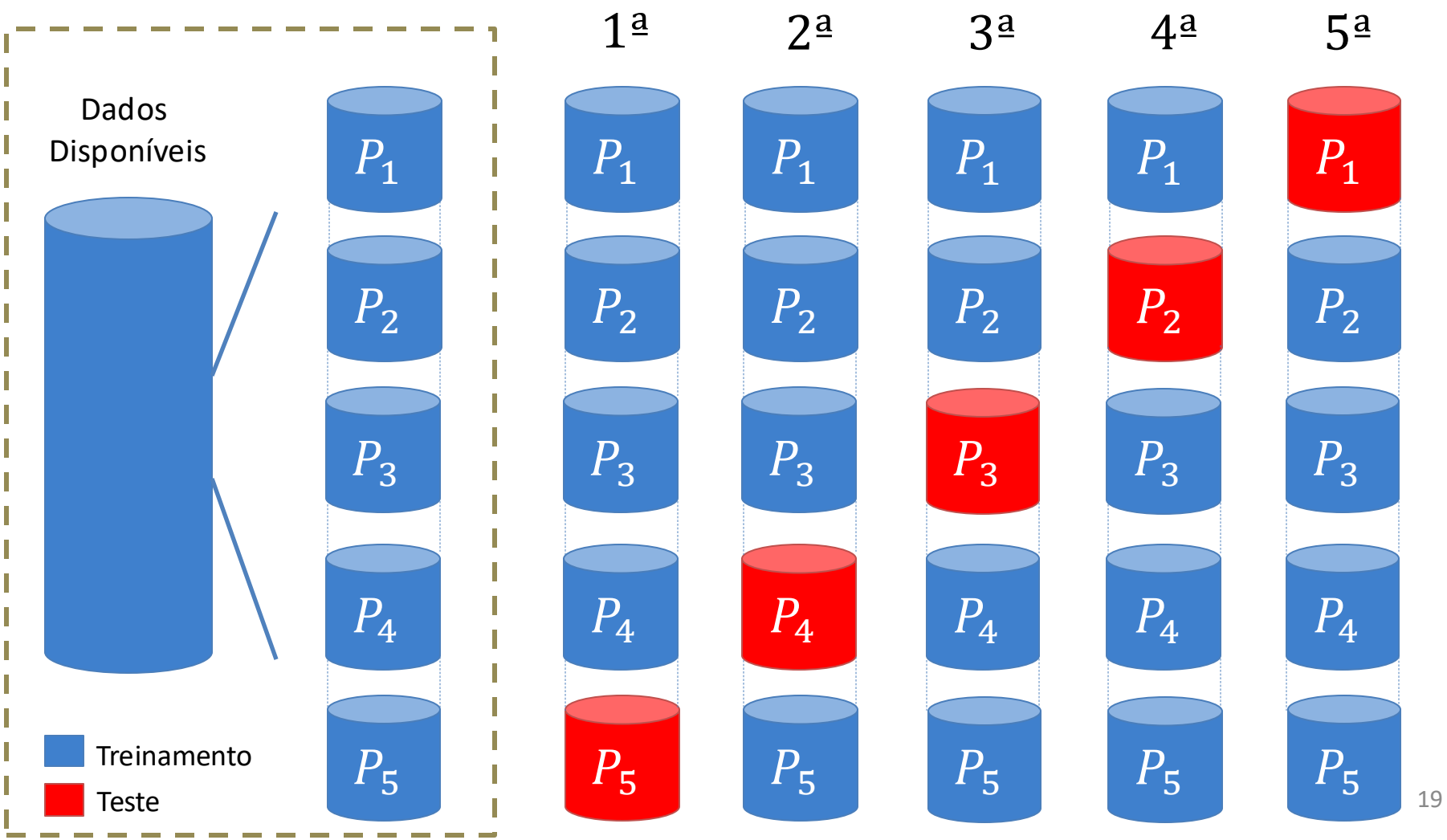
Cross-Validation

- Ex: 5-*fold cross-validation*



Cross-Validation

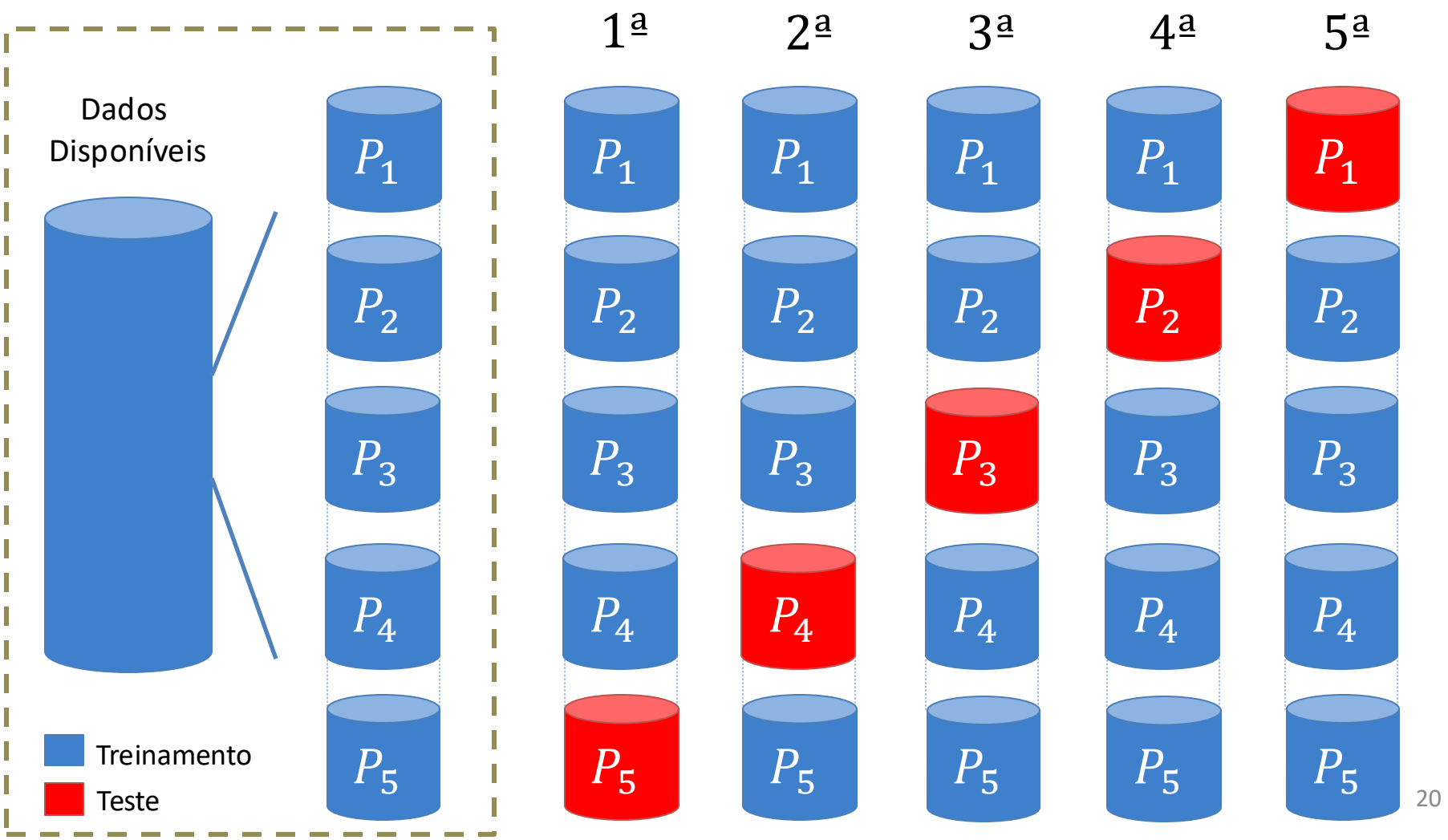
- Ex: 5-*fold cross-validation*



Cross-Validation

- Ex: 5-*fold cross-validation*

Para classificação, recomendado que seja estratificado!



Leave-one-out Cross-Validation

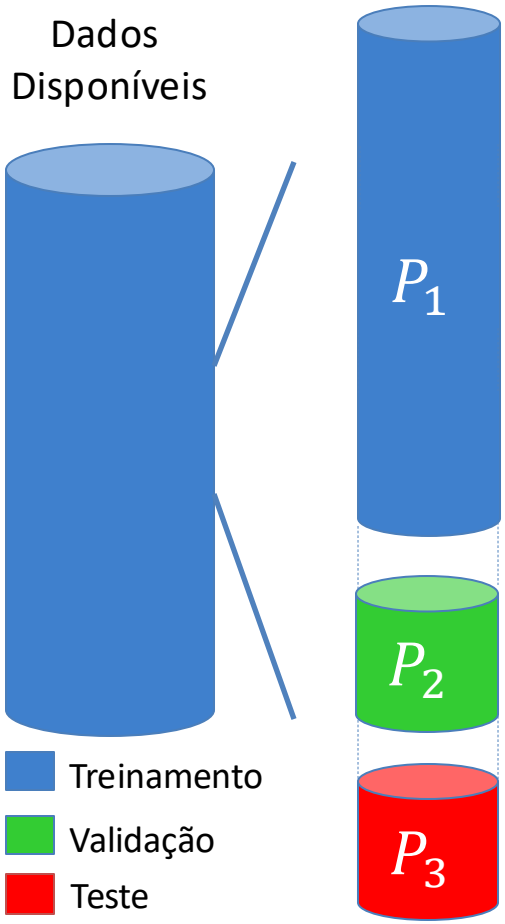
- *Leave-one-out* (LOO)
  - Caso particular de CV onde $k = N$
  - Cada iteração utiliza $(N - 1)$ objetos para treinar e apenas 1 objeto para teste
  - Assim como em k-fold CV, o erro estimado é dado pela média dos (N) erros de teste
  - Computacionalmente caro!!
    - Geralmente utilizado para pequenos 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

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**

Procedimento Padrão

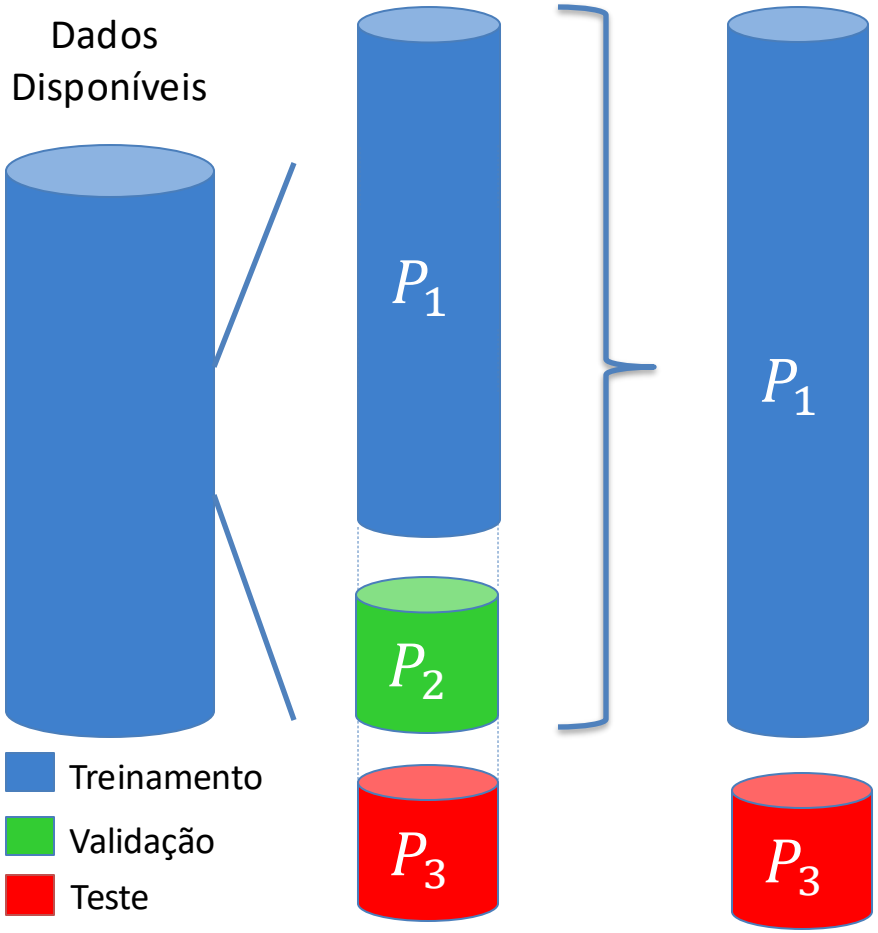
- Idealmente dividimos o nosso conjunto em três porções



- Conjunto P1 é utilizado para treinar o modelo
- Conjunto P2 é onde avaliamos o modelo
  - Procuramos por melhores hiper-parâmetros, técnicas e variáveis
- Conjunto P3 é onde apresentamos o resultado

Procedimento Padrão

- No cenário ideal
- Impraticável para muitos dados (*Deep Learning*)



CROSS VALIDATION

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

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

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

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

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

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

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

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

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

Estimativa de Erro de Regressão

- Para problemas de Regressão, não podemos considerar se um valor é exato ou não

Valor Real	Valor Predito
100	100.01
200	345

100 != 100.01

200 != 345

- Ambos não são exatos, porém um é bem mais próximo do outro!

Estimativa de Erro de Regressão

- Extraímos medidas baseadas no erro que temos:

Valor Real	Valor Predito	Erro
100	100.01	- 0.01
200	345	- 145
124	90	34

+ Próximo

+ Distante

$$RMSE = \frac{1}{N}\sqrt{\sum_{i=1}^{N}(y^{(i)} - \hat{y}^{(i)})^2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y^{(i)} - \hat{y}^{(i)}\right|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

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:

- Rodrigo Coelho Barros (PUCRS)
- 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.