

# Avaliação

Disciplina: Mineração de Dados

Prof. Braian Varjão

# Agenda



1. Métodos de amostragem;
2. Medidas de desempenho para tarefas de classificação;
3. Comparação de múltiplos classificadores.

# Amostragem

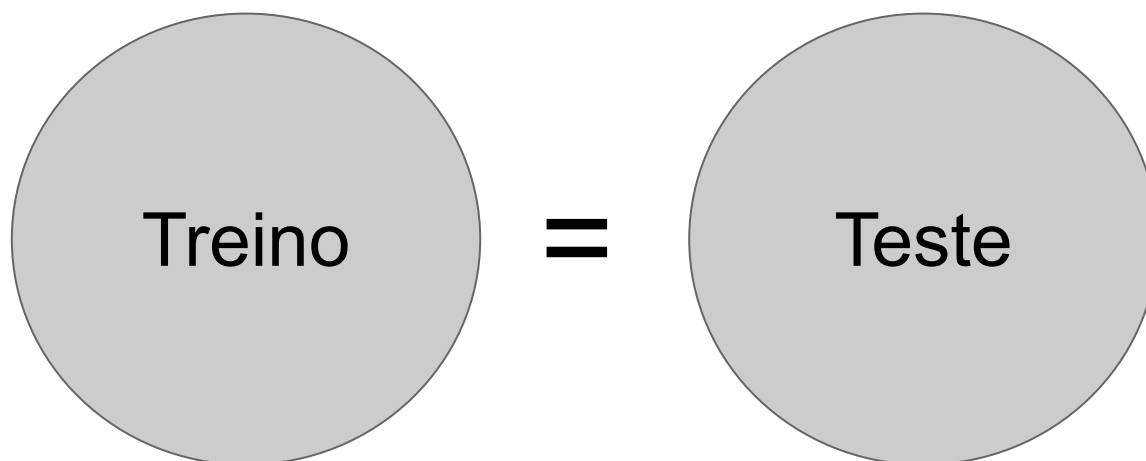
# Métodos de amostragem



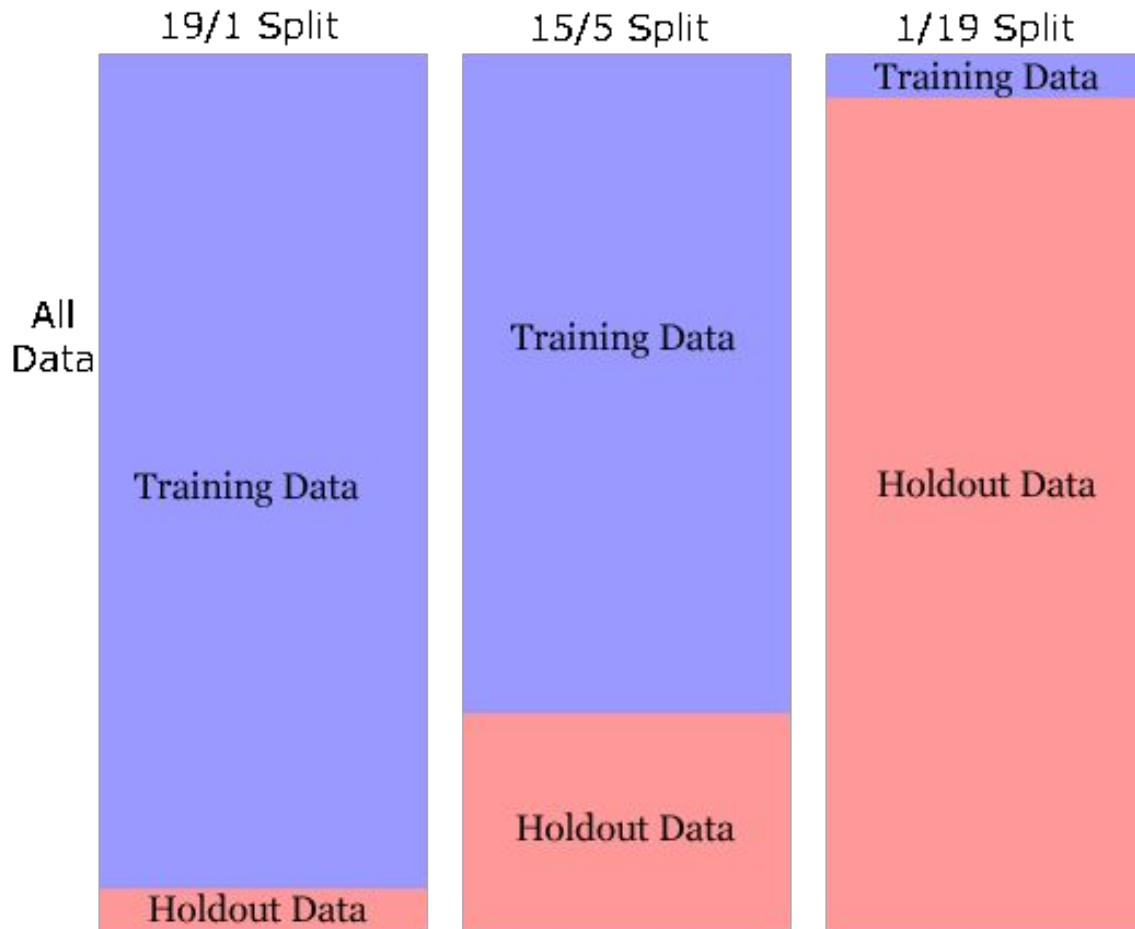
Métodos para estimar o desempenho real de um classificador:

- ▷ Resubstituição (Resubstitution);
- ▷ Validação simples (Holdout);
- ▷ Validação cruzada (r-fold cross-validation);
- ▷ Validação cruzada estratificada (r-fold stratified cross-validation);
- ▷ Leave-one-out.

# Resubstitution



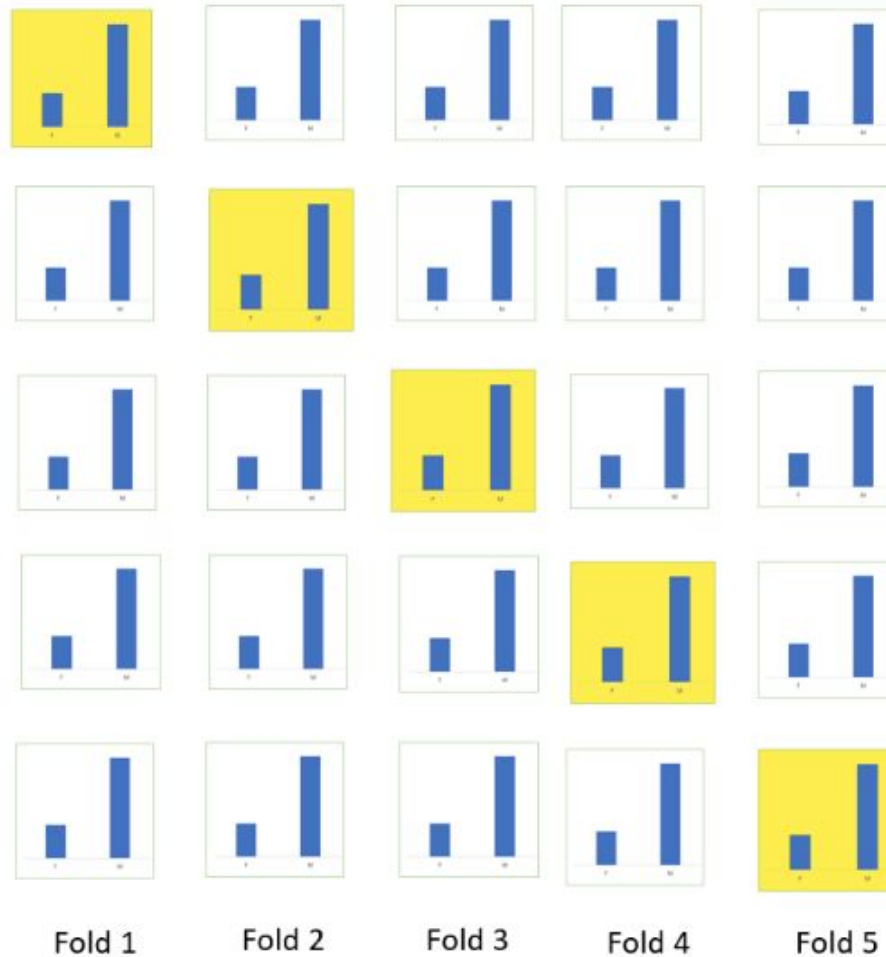
# Houdout



# Cross-validation

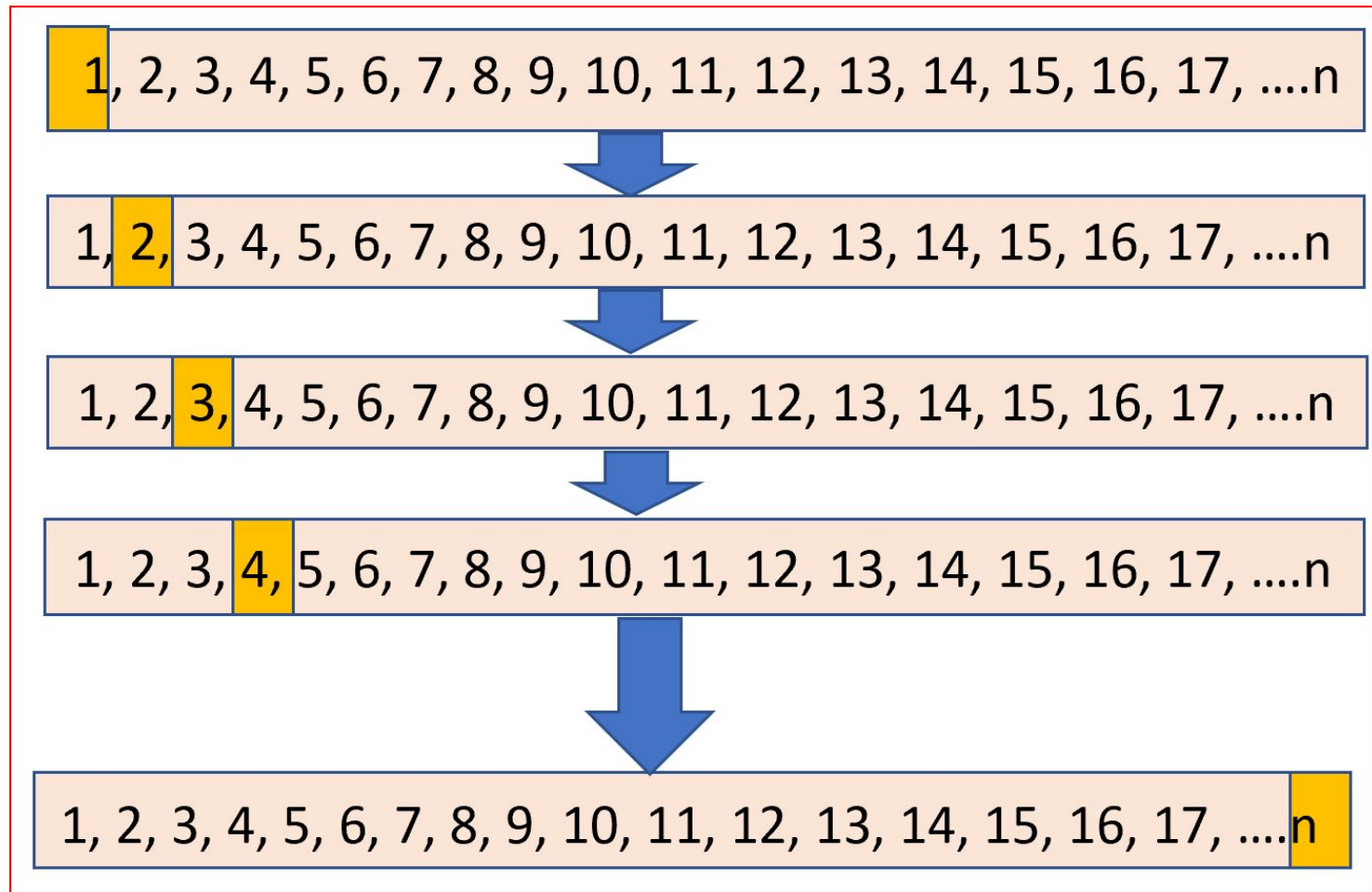


# Stratified cross-validation





# Leave-one-out



# Medidas de Desempenho


## Classificação

# Matriz de confusão

		Predito / Classificado	
		Positivo	Negativo
		Positivo	Negativo
Real	Positivo	<b>TP</b> Verdadeiros Positivos	<b>FN</b> Falsos Negativos
	Negativo	<b>FP</b> Falsos Positivos	<b>TN</b> Verdadeiros Negativos

# Exercício 1

## Detecção de fraude



Conjunto de teste:

300 operações fraudulentas (rótulo 1)

9700 operações não fraudulentas (rótulo 0)

Se o seu classificador acerta 9000 da classe negativa e 100 da classe positiva, como fica a sua matriz de confusão?

# Exercício 1 - Resposta

		Predito / Classificado	
		Positivo	Negativo
Real	Positivo	<b>TP = 100</b> Verdadeiros Positivos	<b>FN = 200</b> Falsos Negativos
	Negativo	<b>FP = 700</b> Falsos Positivos	<b>TN = 9000</b> Verdadeiros Negativos

# Acurácia

		Predito / Classificado	
		Positivo	Negativo
Real	Positivo	TP = 100	FN = 200
	Negativo	FP = 700	TN = 9000

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

Qual a taxa geral de acertos?

$$\text{ACC} = (100 + 9000) / (100 + 9000 + 700 + 200) = 0,910$$

# Revocação

## Taxa de verdadeiros positivos

		Predito / Classificado	
		Positivo	Negativo
Real	Positivo	TP = 100	FN = 200
	Negativo	FP = 700	TN = 9000

$$\text{Recall} = \frac{tp}{tp + fn}$$

Qual o percentual de casos fraudulentos corretamente identificados?

$$RC = 100 / (100 + 200) = 0,333$$

# Precisão

		Predito / Classificado	
		Positivo	Negativo
Real	Positivo	TP = 100	FN = 200
	Negativo	FP = 700	TN = 9000

$$\text{Precision} = \frac{tp}{tp + fp}$$

Qual o percentual de casos fraudulentos identificados que realmente eram casos fraudulentos?

$$PR = 100 / (100 + 700) = 0,125$$



# Qual medida utilizar?



Classificação de artigos entre as categorias de uma revista digital.

Identificação de pessoas com covid.

Identificação de investimentos promissores.

# F $\beta$

$$F_{\beta} = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

$\beta > 1 \rightarrow \text{revocação} > \text{precisão}$

$0 < \beta < 1 \rightarrow \text{revocação} < \text{precisão}$

$$F_{1,0} = 2 \times (0,125 \times 0,333) / (1 \times 0,125 + 0,333) =$$

**0,182**

$$F_{0,5} = 1,25 \times (0,125 \times 0,333) / (0,25 \times 0,125 + 0,333) =$$

**0,143**

# Como avaliar problemas multi-classe?

# Problemas multi-classe

		Predicted			
		A	B	C	
True labels	A	2	2	0	4
	B	1	2	0	3
	C	0	0	3	3
		3	4	3	Total

A	
TP = 2	FN = 2
FP = 1	TN = 5

B	
TP = 2	FN = 1
FP = 2	TN = 5

C	
TP = 3	FN = 0
FP = 0	TN = 7