



*Universidade Federal do Paraná*  
*Laboratório de Estatística e Geoinformação - LEG*



# Métodos de reamostragem

Eduardo Vargas Ferreira

# Função custo

- **Matriz de confusão:** é um layout de tabela que permite a visualização do desempenho do algoritmo.

		Diagnosticado	
		Doente	Saudável
Paciente	Doente	1000	200
	Saudável	800	8000

		Avaliado	
		Spam	Não-Spam
E-mail	Spam	100	170
	Não-Spam	30	700

- Acurácia:** é a razão entre as predições corretas pelo total.

		Diagnosticado	
		Doente	Saudável
Paciente	Doente	1000	200
	Saudável	800	8000

$$\text{Acurácia} = \frac{1000 + 8000}{10000} = 90\%$$

		Avaliado	
		Spam	Não-Spam
E-mail	Spam	100	170
	Não-Spam	30	700

$$\text{Acurácia} = \frac{100 + 700}{1000} = 80\%$$

- **Precisão:** é o número de verdadeiros positivos, dividido pelo número de positivos estimados pelo modelo.

		Diagnosticado	
		Doente	Saudável
Paciente	Doente	1000	200
	Saudável	800	8000

$$\text{Precisão} = \frac{1000}{1000 + 800} = 55,6\%$$

		Avaliado	
		Spam	Não-Spam
E-mail	Spam	100	170
	Não-Spam	30	700

$$\text{Precisão} = \frac{100}{100 + 30} = 76,9\%$$

- Recall:** dado que o estado verdadeiro é positivo, qual a proporção de verdadeiro positivo.

		Diagnosticado	
		Doente	Saudável
Paciente	Doente	1000	200
	Saudável	800	8000

$$\text{Recall} = \frac{1000}{1000 + 200} = 83,3\%$$

		Avaliado	
		Spam	Não-Spam
E-mail	Spam	100	170
	Não-Spam	30	700

$$\text{Recall} = \frac{100}{100 + 170} = 37\%$$

- **$F_1$  score:** é um compromisso entre a **Precisão** e o **Recall** através de uma média harmônica.

$$\begin{array}{c} Y \\ | \\ \text{Média aritmética} = \frac{x+y}{2} \\ | \\ \text{Média harmônica} = 2 \cdot \frac{x \cdot y}{x+y} \\ | \\ X \end{array} \quad F_1 \text{ score} = 2 \times \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}}$$

- **$F_\beta$  score:** é uma generalização do  $F_1$  score, em que  $\beta$  representa a influência da **Precisão** no resultado final.

$$F_\beta \text{ score} = (1 + \beta^2) \frac{\text{Precisão} \cdot \text{Recall}}{\beta^2 \cdot \text{Precisão} + \text{Recall}}$$

- **Soma de quadrados dos desvios (SQD)**

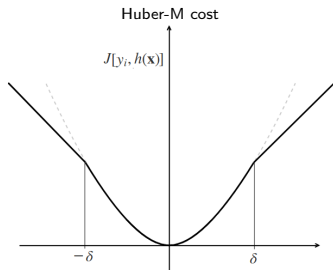
$$J[y_i, h(\mathbf{x})] = \frac{1}{n} \sum_{i=1}^n [y_i - h(x_i)]^2$$

- **Soma dos desvios absolutos (SDA)**

$$J[y_i, h(\mathbf{x})] = \frac{1}{n} \sum_{i=1}^n |y_i - h(x_i)|$$

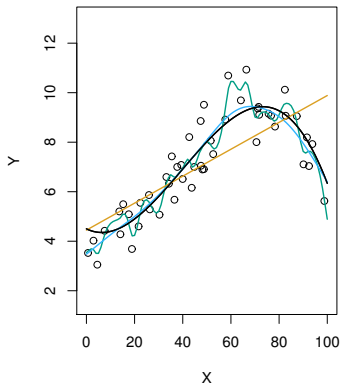
- **Huber-M cost**

$$J[y_i, h(\mathbf{x})] = \frac{1}{n} \sum_{i=1}^n \begin{cases} \frac{1}{2} [y_i - h(x_i)]^2, & \text{para } |y - h(\mathbf{x}_i)| \leq \delta, \\ \delta |y_i - h(x_i)| - \frac{1}{2} \delta^2, & \text{caso contrário.} \end{cases}$$



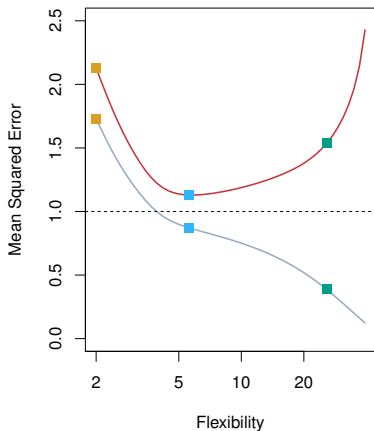
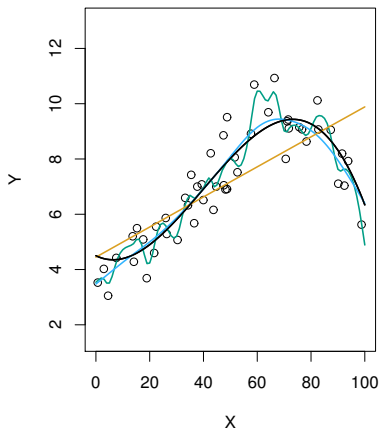


- Suponha que estamos interessados em estudar a relação entre  $X$  e  $Y$ ;



- Podemos definir varias funções,  $h(x)$ . Mas, qual fornece a melhor predição? **Resposta:** a que apresentar menor custo (ou risco).

# Exemplo simulado



Exercícios



Treinamento

Simulado



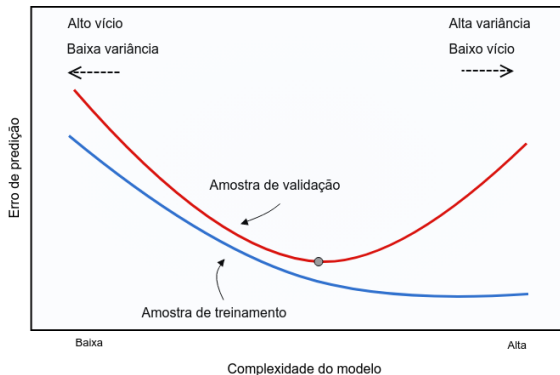
Validação

Vestibular



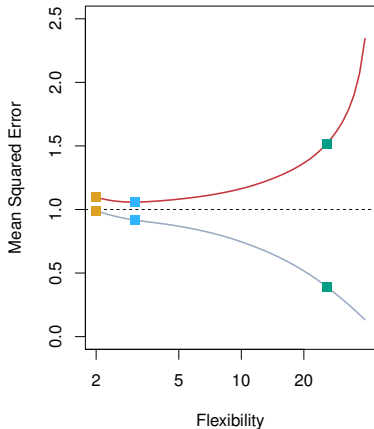
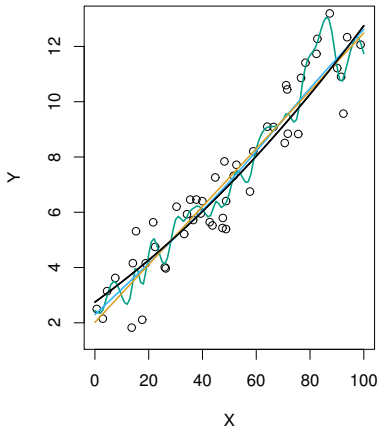
Teste

- **Erro do treino:** é calculado mediante aplicação do método estatístico nos dados de treino;

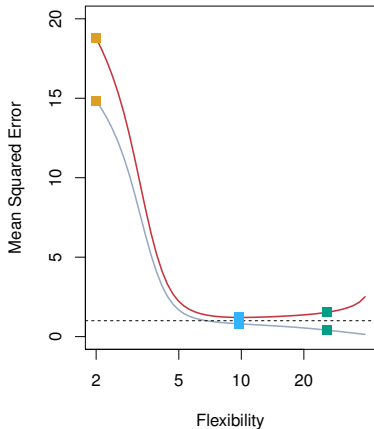
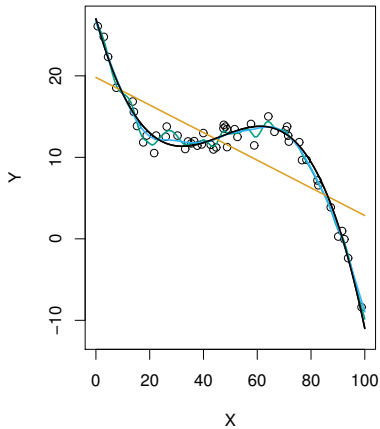


- **Erro de validação:** média do erro resultante da predição de uma nova observação (que não fazia parte dos dados de treino).

# Exemplo simulado

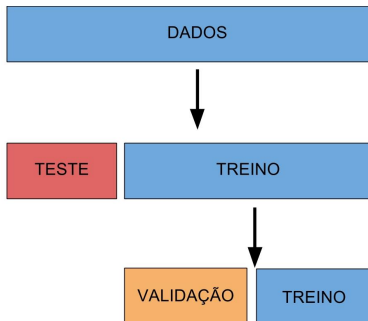


# Exemplo simulado



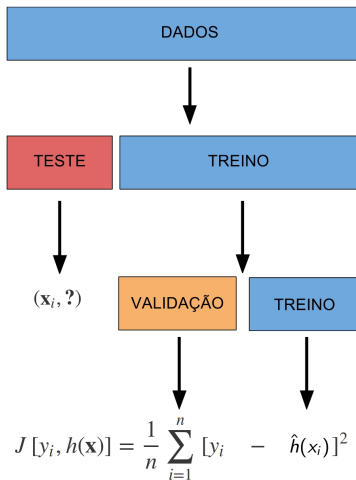
# Validação holdout

- Nesta abordagem, dividimos os dados em apenas duas partes: **treinamento** e **validação**;



- O erro resultante dos dados de validação fornece uma estimativa do **erro do teste**, baseando-se em determinado indicador.

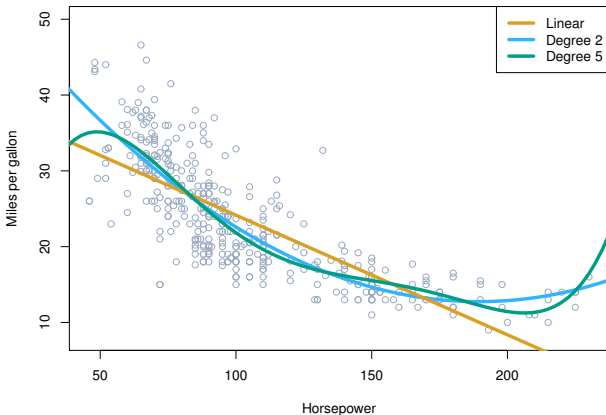




# Exemplo: Auto data set



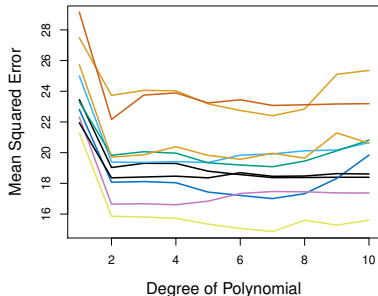
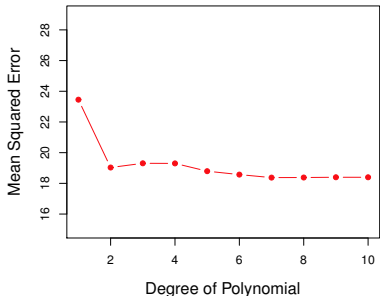
- Neste exemplo, estamos avaliando a relação entre consumo de combustível e potência do automóvel.



# Exemplo: Auto data set



- Separamos aleatoriamente as 392 observações em duas amostras: treinamento (com 196 dados) e validação (196 dados);



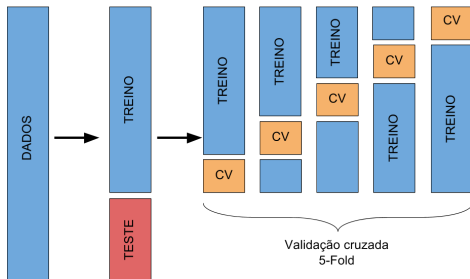
- O gráfico da esquerda temos **divisão única** e da direita **divisão múltipla**.

## Validação cruzada por $k - fold$

- O método consiste em dividir os dados em  $K$  partes iguais. Ajusta-se o modelo com  $K - 1$  partes, e uma é destinada para às predições;



- Isto é feito para  $k = 1 : K$ , em seguida os resultados são combinados;



- Sejam as  $K$  partes denotadas por  $C_1, C_2, \dots, C_K$ , em que  $C_k$  representa o índice da  $k$ -ésima parte;

1	2	3	...	K
Validação	Treino	Treino	...	Treino

- Considere que temos  $n_k$  observações na parte  $k$ : se  $n$  é múltiplo de  $K$ , então  $n_k = n/K$ . Calcule:

$$\begin{aligned} CV_{(K)} &= \sum_{k=1}^K \frac{n_k}{n} EQM_k \\ &= \frac{n_1}{n} EQM_1 + \frac{n_2}{n} EQM_2 + \frac{n_3}{n} EQM_3 + \dots + \frac{n_K}{n} EQM_K \end{aligned}$$

- Em que  $EQM_k = \sum_{i \in C_k} [y_i - \hat{h}(x_i)]^2 / n_k$ .

- Sejam as  $K$  partes denotadas por  $C_1, C_2, \dots, C_K$ , em que  $C_k$  representa o índice da  $k$ -ésima parte;

1	2	3	...	K
Treino	Validação	Treino	...	Treino

- Considere que temos  $n_k$  observações na parte  $k$ : se  $n$  é múltiplo de  $K$ , então  $n_k = n/K$ . Calcule:

$$\begin{aligned} CV_{(K)} &= \sum_{k=1}^K \frac{n_k}{n} EQM_k \\ &= \frac{n_1}{n} EQM_1 + \frac{n_2}{n} EQM_2 + \frac{n_3}{n} EQM_3 + \dots + \frac{n_K}{n} EQM_K \end{aligned}$$

- Em que  $EQM_k = \sum_{i \in C_k} [y_i - \hat{h}(x_i)]^2 / n_k$ .

- Sejam as  $K$  partes denotadas por  $C_1, C_2, \dots, C_K$ , em que  $C_k$  representa o índice da  $k$ -ésima parte;

1	2	3	...	K
Treino	Treino	Validação	...	Treino

- Considere que temos  $n_k$  observações na parte  $k$ : se  $n$  é múltiplo de  $K$ , então  $n_k = n/K$ . Calcule:

$$\begin{aligned} CV_{(K)} &= \sum_{k=1}^K \frac{n_k}{n} EQM_k \\ &= \frac{n_1}{n} EQM_1 + \frac{n_2}{n} EQM_2 + \frac{n_3}{n} EQM_3 + \dots + \frac{n_K}{n} EQM_K \end{aligned}$$

- Em que  $EQM_k = \sum_{i \in C_k} [y_i - \hat{h}(x_i)]^2 / n_k$ .



- Sejam as  $K$  partes denotadas por  $C_1, C_2, \dots, C_K$ , em que  $C_k$  representa o índice da  $k$ -ésima parte;

1	2	3	...	K
Treino	Treino	Treino	...	Validação

- Considere que temos  $n_k$  observações na parte  $k$ : se  $n$  é múltiplo de  $K$ , então  $n_k = n/K$ . Calcule:

$$\begin{aligned} CV_{(K)} &= \sum_{k=1}^K \frac{n_k}{n} EQM_k \\ &= \frac{n_1}{n} EQM_1 + \frac{n_2}{n} EQM_2 + \frac{n_3}{n} EQM_3 + \dots + \frac{n_K}{n} EQM_K \end{aligned}$$

- Em que  $EQM_k = \sum_{i \in C_k} [y_i - \hat{h}(x_i)]^2 / n_k$ .

- Sejam as  $K$  partes denotadas por  $C_1, C_2, \dots, C_K$ , em que  $C_k$  representa o índice da  $k$ -ésima parte;

1	2	3	...	K
Treino	Treino	Treino	...	Validação

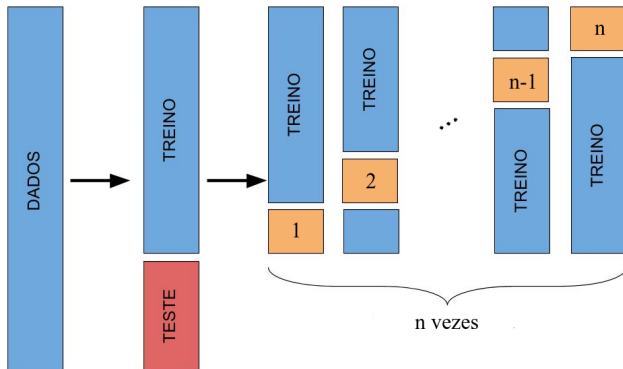
- Considere que temos  $n_k$  observações na parte  $k$ : se  $n$  é múltiplo de  $K$ , então  $n_k = n/K$ . Calcule:

$$\begin{aligned} CV_{(K)} &= \sum_{k=1}^K \frac{n_k}{n} Err_k \\ &= \frac{n_1}{n} Err_1 + \frac{n_2}{n} Err_2 + \frac{n_3}{n} Err_3 + \dots + \frac{n_K}{n} Err_K \end{aligned}$$

- Em que  $Err_k = \sum_{i \in C_k} \mathbb{1}(y_i \neq \hat{y}_i) / n_k$ , e  $\hat{y}_i$  é a classificação da  $i$ -ésima observação.

# Validação cruzada leave-one-out

# Validação cruzada leave-one-out (LOOCV)

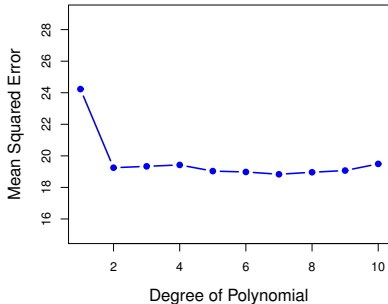


# Exemplo: Auto data set

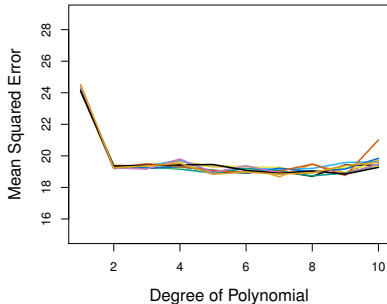


- O gráfico da direita apresenta 9 diferentes validações cruzadas 10 – fold. Em cada uma, temos uma nova partição dos dados.

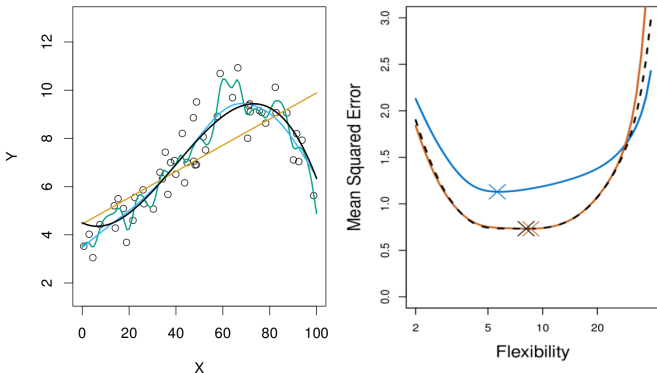
LOOCV



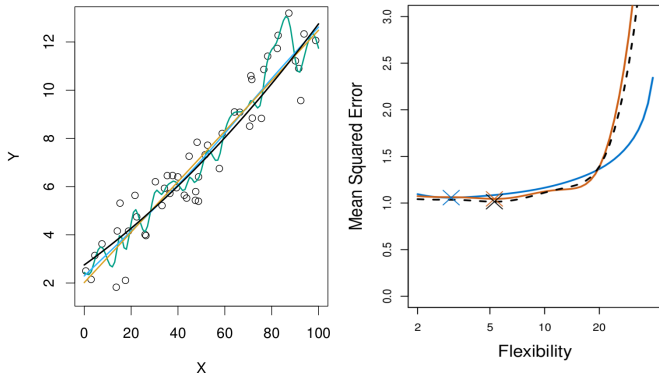
10-fold CV



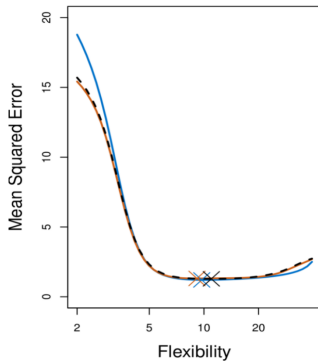
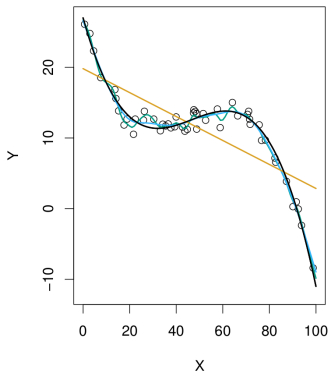
- O gráfico abaixo apresenta a verdadeira curva do EQM em azul, a estimativa LOOCV pontilhada e 10 – *fold* em laranja;



- O gráfico abaixo apresenta a verdadeira curva do EQM em azul, a estimativa LOOCV pontilhada e 10 – *fold* em laranja;



- O gráfico abaixo apresenta a verdadeira curva do EQM em azul, a estimativa LOOCV pontilhada e 10 – *fold* em laranja;





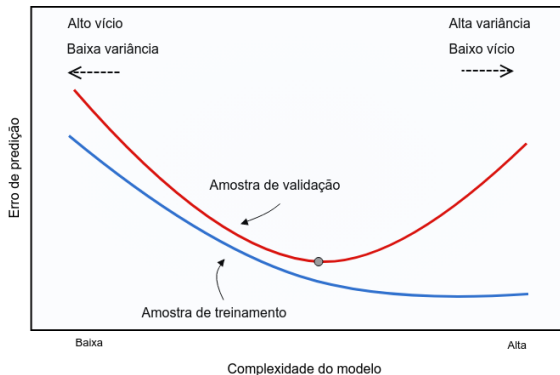
# **Bias-Variance Trade-Off**

# Bias-Variance Trade-Off



- Um bom desempenho no conjunto de teste requer um baixo erro quadrático médio. Porém, note que

$$E [y_0 - h(x_0)]^2 = \text{Var} [h(x_0)] + \text{Vício} [h(x_0)]^2 + \text{Var}(\varepsilon).$$



- Um bom desempenho no conjunto de teste requer um baixo erro quadrático médio. Porém, note que

$$E [y_0 - h(x_0)]^2 = \text{Var} [h(x_0)] + \text{Vício} [h(x_0)]^2 + \text{Var}(\varepsilon).$$

## Variância

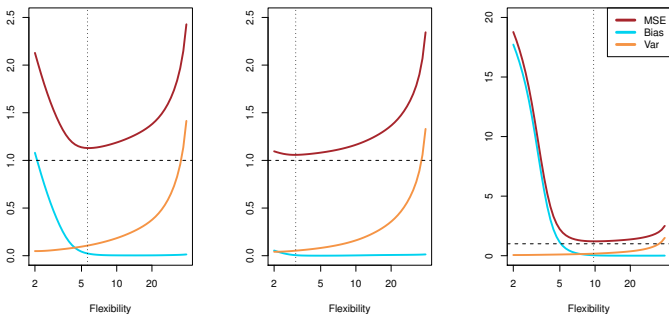
- ★ Refere-se ao quanto  $h(x_0)$  muda quando a estimamos utilizando diferentes dados de treino;
- ★ Em geral, quanto mais flexível o modelo, maior a variância.

## Vício

- ★ Refere-se ao erro de aproximar um problema real (extremamente complicado) por uma função simples;
- ★ Em geral, quanto mais simples o modelo, maior o vício.

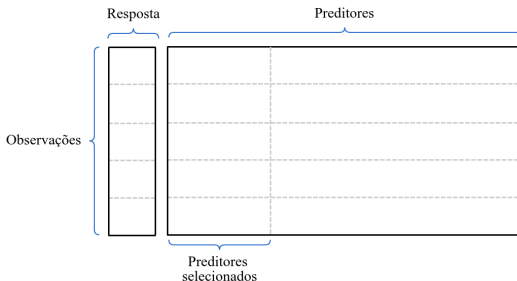
- Um bom desempenho no conjunto de teste requer um baixo erro quadrático médio. Porém, note que

$$E [y_0 - h(x_0)]^2 = \text{Var} [h(x_0)] + \text{Vício} [h(x_0)]^2 + \text{Var}(\varepsilon).$$



## **Validação cruzada: certo e errado**

- Considere um classificador aplicado aos dados de duas classes:
  1. Começando com 5000 preditores e amostra de tamanho 50, filtramos os 100 preditores com maior correlação entre as classes;
  2. Aplicamos um classificador utilizando somente os 100 preditores.

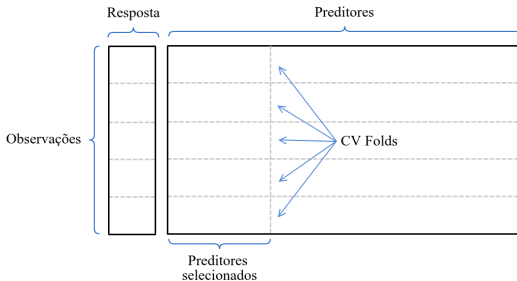


- Como podemos estimar o desempenho do teste para este classificador?  
**Validação cruzada**

# Validação cruzada: abordagem errada!



- Podemos aplicar validação cruzada no Passo 2, esquecendo o Passo 1 (não incorporando o fato de termos eliminado 4900 preditores)? **Não!**

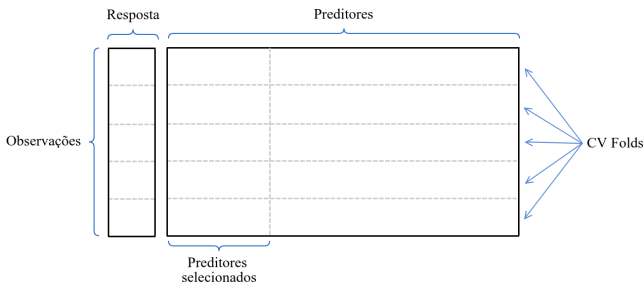


- Isso seria ignorar o fato de que no Passo 1 o procedimento já viu os rótulos de treinamento, e aprendeu com isso.

# Validação cruzada: abordagem certa!



- Podemos aplicar validação cruzada no Passo 2, esquecendo o Passo 1 (não incorporando o fato de termos eliminado 4900 preditores)? **Não!**



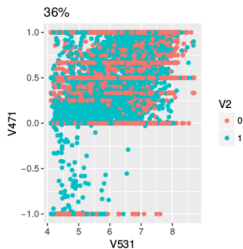
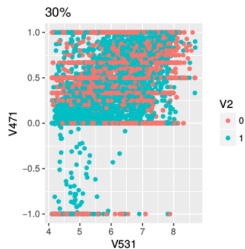
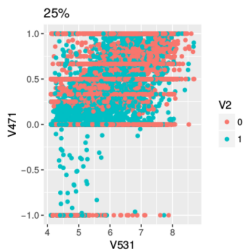
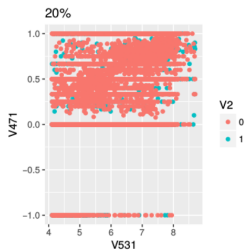
- Isso seria ignorar o fato de que no Passo 1 o procedimento já viu os rótulos de treinamento, e aprendeu com isso.



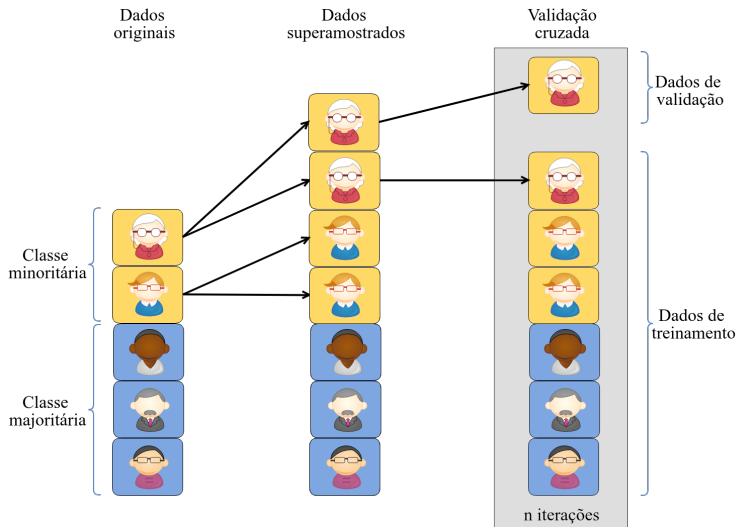
# Exemplo: clientes em atraso



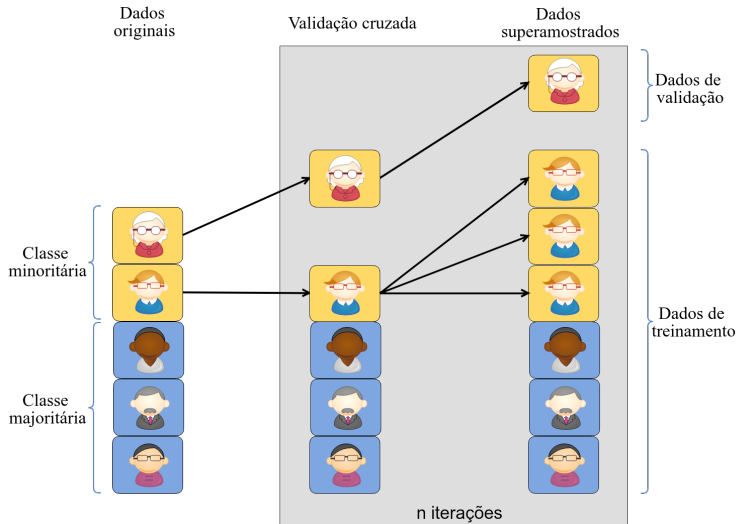
- Criamos exemplos “sintéticos”, superamostrando a classe minoritária.



# Exemplo: abordagem errada!

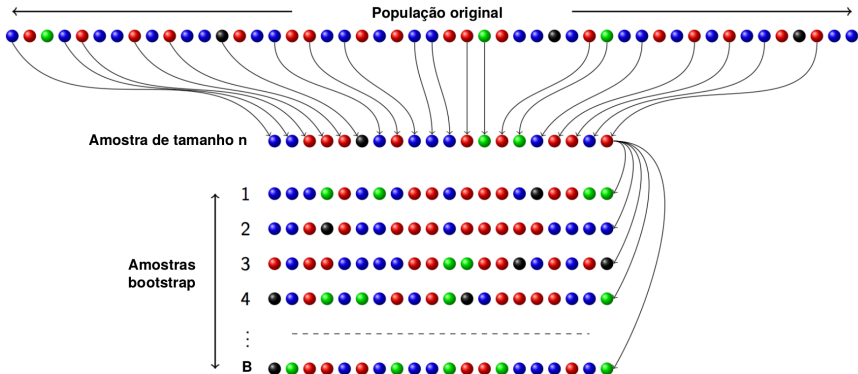


# Exemplo: abordagem certa!

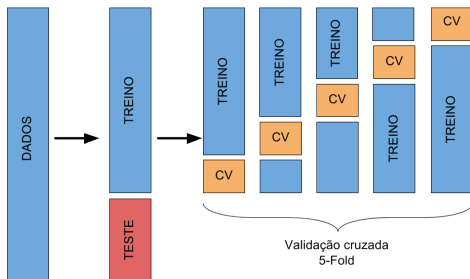


# Bootstrap

# Bootstrap estimando o erro de predição



- Na validação cruzada, o  $K$ -ésimo *fold* de validação é distinto dos demais  $k - 1$  *folds* usados no treinamento;



- Não há **overlap** entre os dados de treino e validação. O que é crucial para seu sucesso. Queremos uma ideia sobre os dados de teste (novos dados);

- James, G., Witten, D., Hastie, T. e Tibshirani, An Introduction to Statistical Learning, 2013;
- Hastie, T., Tibshirani, R. e Friedman, J., The Elements of Statistical Learning, 2009;
- Lantz, B., Machine Learning with R, Packt Publishing, 2013;
- Tan, Steinbach, and Kumar, Introduction to Data Mining, Addison-Wesley, 2005;
- Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani