



# **Universidade Federal da Paraíba**

Coordenação do Curso de Ciência de Dados e  
Inteligência Artificial



## **Validação**

Prof. Gilberto Farias

# Roteiro

- Conceito de validação
- Cross Validation

# Validação X Regularização

## Regularização

$$E_{out} = E_{in} + \underbrace{\textit{penalidade do overfit}}_{\text{regularização estima este valor}}$$

## Validação

$$\underbrace{E_{out}}_{\text{validação estima este valor}} = E_{in} + \textit{penalidade do overfit}$$

## Relembrando a estimativa de erro

E um ponto  $(x, y)$  fora da amostra, o erro é  $e(h(x), y)$

Erro quadrático:  $(h(x) - y)^2$

Erro binário:  $\llbracket h(x), y \rrbracket$

$$\mathbb{E}[e(h(x), y)] = E_{out}(h)$$

$$\text{var}[e(h(x), y)] = \sigma^2$$

## De um ponto para o conjunto de validação

No conjunto validação  $(x_1, y_1), \dots, (x_K, y_K)$ , o erro é  $E_{val} = \frac{1}{K} \sum_{k=1}^K e(h(x_k), y_k)$

$$\mathbb{E}[E_{val}(h)] = \frac{1}{K} \sum_{k=1}^K \mathbb{E}[e(h(x), y)] = E_{out}(h)$$

$$\text{var}[E_{val}(h)] = \frac{1}{K^2} \sum_{k=1}^K \text{var}[e(h(x), y)] = \frac{\sigma^2}{K}$$

$$E_{val} = E_{out} \pm O\left(\frac{1}{\sqrt{K}}\right)$$

# $K$ é tirado de $N$

Dado o conjunto  $D = (x_1, y_1), \dots, (x_N, y_N)$

$K$  pontos  $\rightarrow$  validação

$\underbrace{\hspace{1.5cm}}$

$D_{val}$

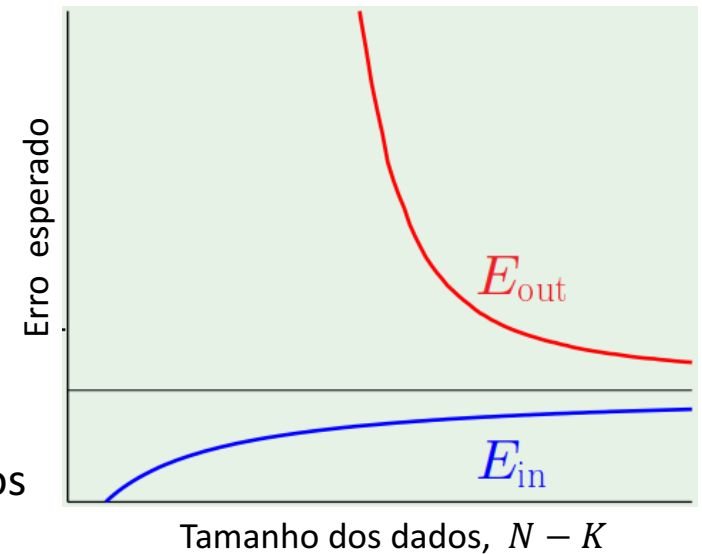
$N - K$  pontos  $\rightarrow$  treino

$\underbrace{\hspace{1.5cm}}$

$D_{train}$

$O\left(\frac{1}{\sqrt{K}}\right)$ :  $K$  pequeno  $\rightarrow E_{val}(h)$  mal estimado

E o  $K$  grande??  $\rightarrow E_{train}(h)$  e  $E_{val}(h)$  mal estimados



Regra de ouro :

$$K = \frac{N}{5}$$

$\leftarrow K$

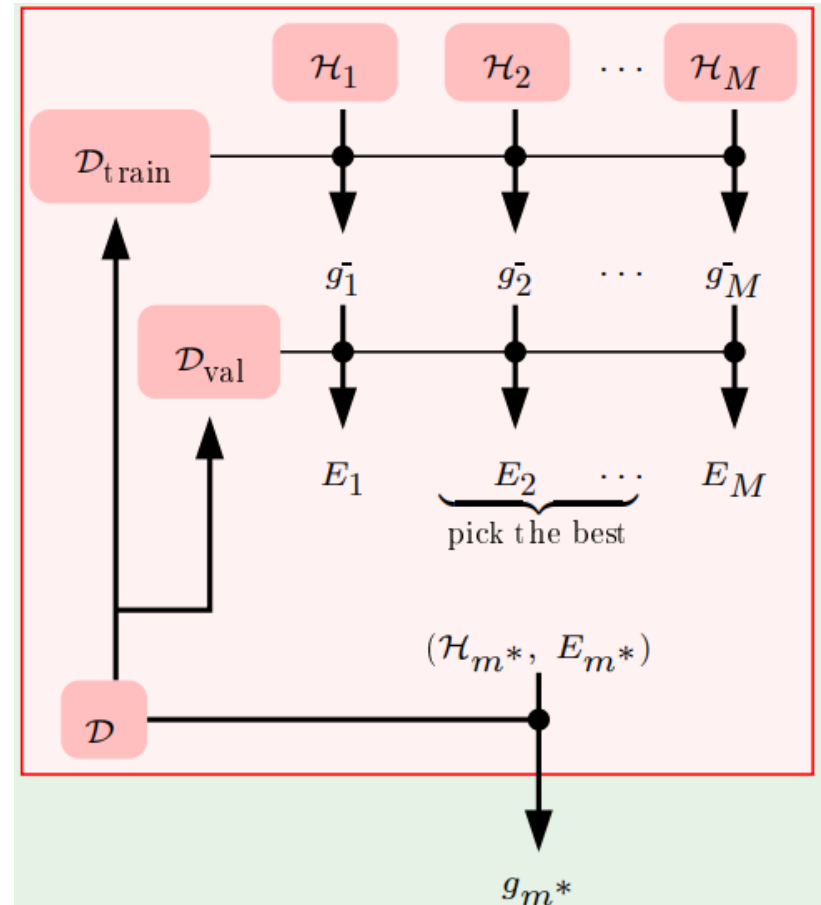
# Usando $D_{val}$ para escolher o modelo

Sejam  $M$  modelos  $\mathcal{H}_1, \dots, \mathcal{H}_M$

Use  $D_{train}$  para aprender  $g_m^-$  de cada modelo

Avalie  $g_m^-$  usando  $D_{val}$  ( $E_m = E_{val}(g_m^-)$ )

Escolha o modelo  $m = m^*$  com o menor  $E_m$



# O dilema do tamanho do $K$

A cadeia de raciocínio

$$E_{out}(g) \approx E_{out}(g^-) \approx E_{val}(g^-)$$

(pequeno  $K$ )

(grande  $K$ )

Podemos ter o  $K$  pequeno e grande ao mesmo tempo??

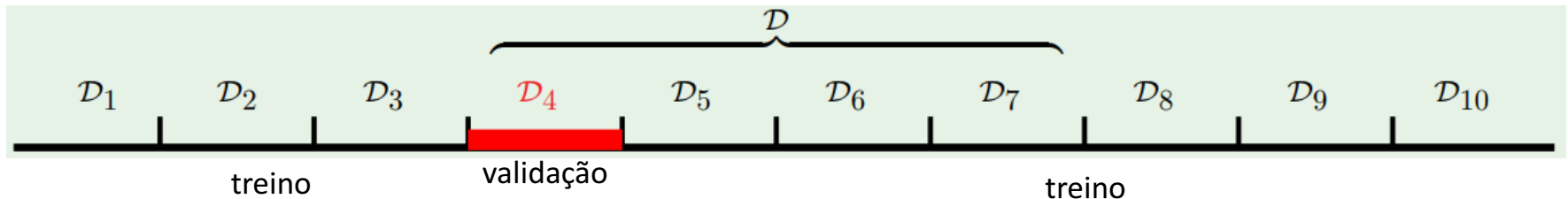


# Cross Validation

Divide a base  $D$  em  $\frac{N}{K}$  pedaços

$\frac{N}{K}$  sessões de treino com  $N - K$  pontos distintos

Valida cada sessão com os  $K$  pontos restantes e computada o  $E_{val}$  médio



10-fold cross validation :  $K = \frac{N}{10}$