Regressão

Regressão

- A regressão é um técnica para modelar e analisar dados que são compostos por uma variável dependente (a variável de resposta) e uma ou mais variáveis independentes (as variáveis de entrada).
- Usado em predição, inferência, teste de hipóteses . . .

Regressão: Formulação Matemática

y resposta

$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$ entradas

y processo verdadeiro

$\hat{y}$ aproximação

Necessário:

- Função paramétrica $\hat{y} = w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \ldots$
- Conjunto com N amostras de entrada $\mathcal{X} = \{\mathbf{x}[1], \ldots, \mathbf{x}[N]\}$
- e as correspondentes N amostras de saída $\mathcal{Y} = \{y[1], \ldots, y[N]\}$

Objectivo:

estimar os parâmetros $w_1, w_2, \ldots$

Como:

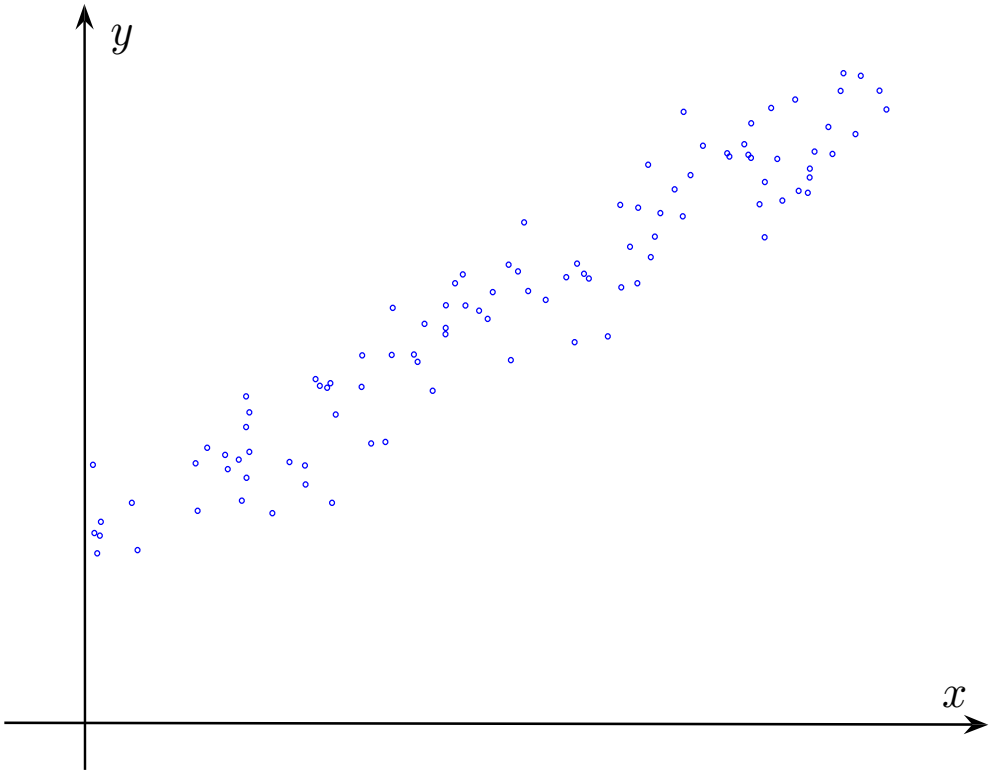
Minimizar a potência do erro

$$\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N}(y[n] - \hat{y}[n])^2$$

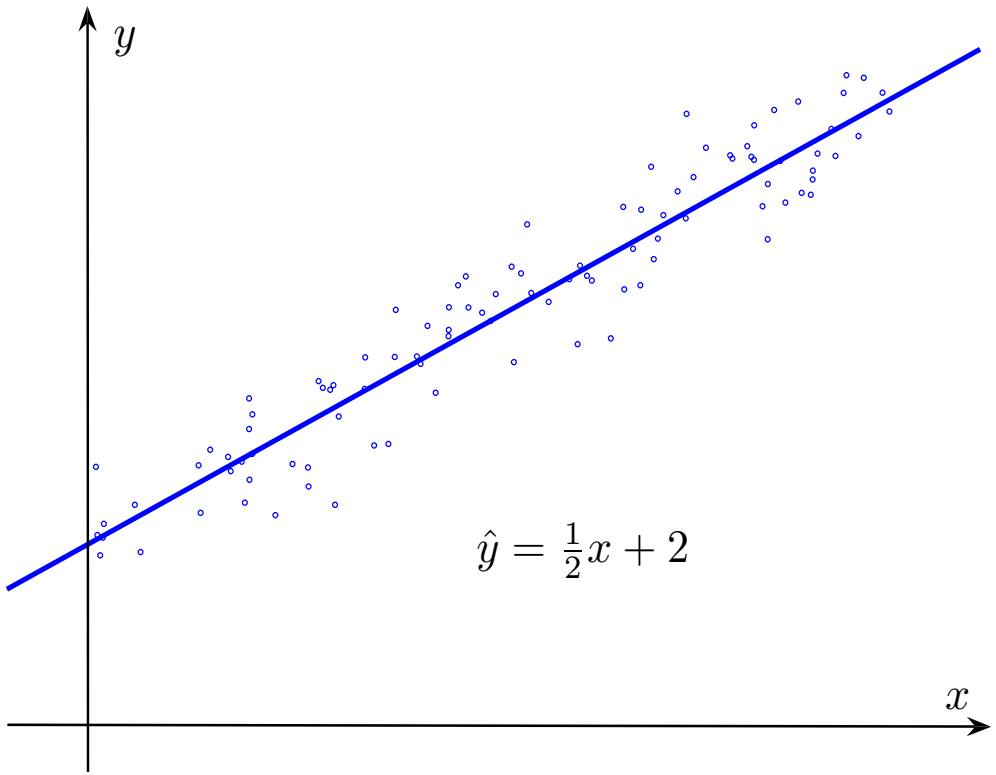
Pressupostos:

- O ruído ϵ é uma variável aleatória de média nula.
- Amostras do ruído são independentes entre si.
- A potência do ruído é constante.
- As variáveis de entrada não têm ruído

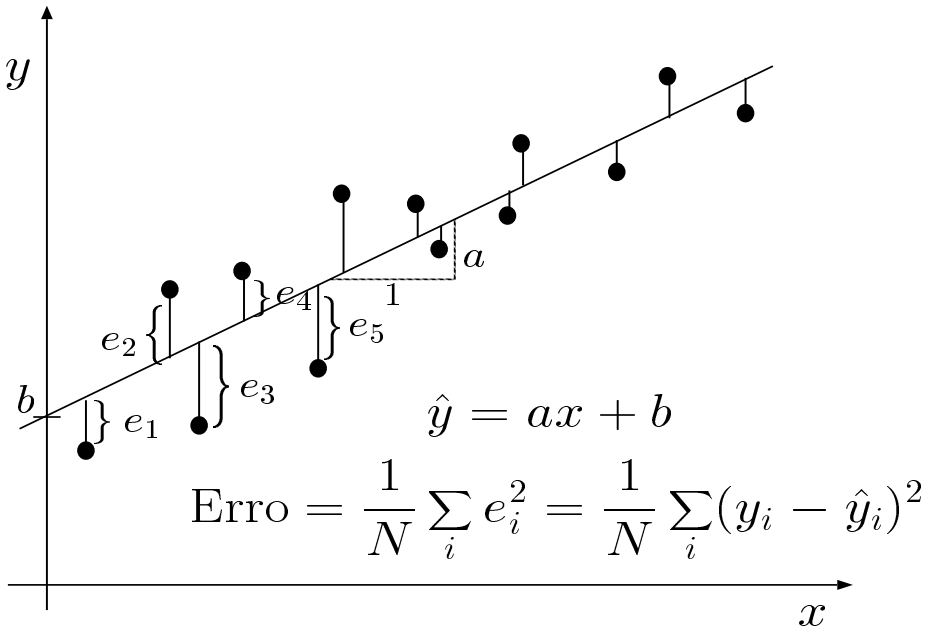
Modelo Simples



Modelo Simples



Regressão Linear: equação de recta



Equação de recta:

$$\hat{y} = ax + b$$
$$y = \alpha x + \beta + \epsilon$$

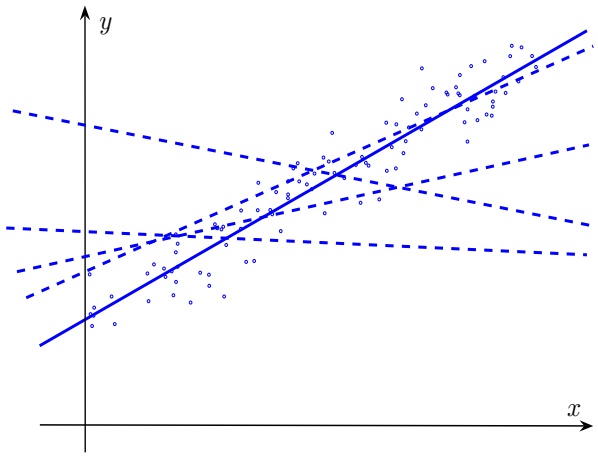
Objectivo: estimar os parâmetros (a e b) da recta que melhor descreve os dados

Método: minimizar o erro quadrático médio $\mathcal{E}$ (distância^2 vertical dos pontos à recta)

Erro: $e = y - (ax + b) = y - \hat{y}$

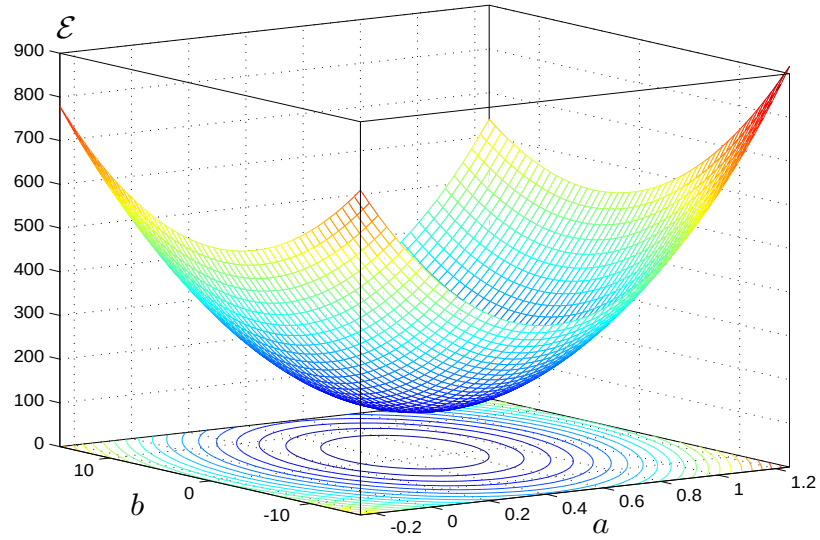
Erro quadrático: $\mathcal{E} = \frac{1}{N}\sum_{n=1}^{N} e[n]^2 = \frac{1}{N}\sum_{n=1}^{N} (y[n] - (ax[n] + b))^2$

Regressão Linear: equação de recta



Pontos + rectas com diferentes valores de a e b.

$\Longrightarrow$



$\mathcal{E}$ para vários valores de a e b. $\mathcal{E}$ é uma parábola: um único mínimo! Pode-se resolver analiticamente.

Mínimos Quadrados

Derivar $\mathcal{E}$ em relação a e a b, e igualar a zero

Corresponde a minimizar a função do erro quadrático $\mathcal{E}$ para os dois conjuntos $\mathcal{X}$ e $\mathcal{Y}$ (N amostras cada um):

$$\frac{\partial \mathcal{E}}{\partial a} = 0$$

$$0 = \frac{\partial}{\partial a}\frac{1}{N}\sum_{n}^{N}(y[n] - ax[n] - b)^2$$

$$0 = -\frac{2}{N}\sum_{n}^{N}(y[n]x[n] - ax[n]^2 - bx[n])$$

$$0 = \frac{1}{N}\sum_{n}^{N} y[n]x[n] - \frac{a}{N}\sum_{n}^{N} x[n]^2 - \frac{b}{N}\sum_{n}^{N} x[n]$$

$$\frac{\partial \mathcal{E}}{\partial b} = 0$$

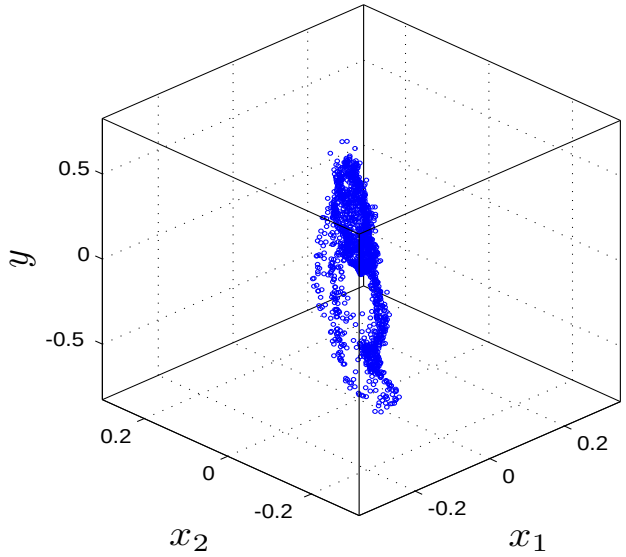
$$0 = \frac{\partial}{\partial b}\frac{1}{N}\sum_{n}^{N}(y[n] - ax[n] - b)^2$$

$$0 = -\frac{2}{N}\sum_{n}^{N}(y[n] - ax[n] - b)$$

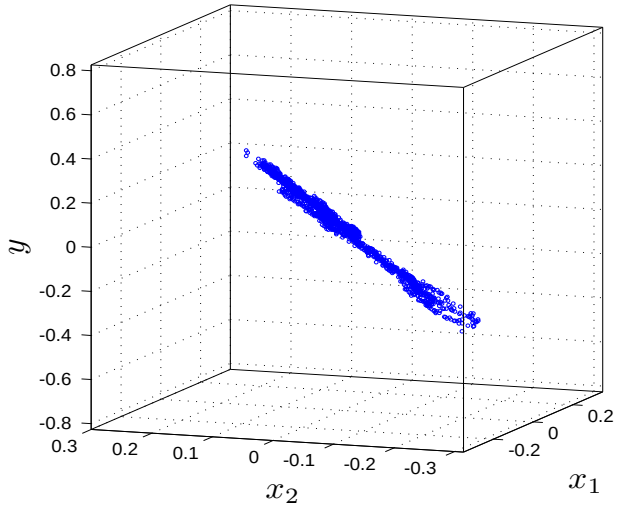
$$0 = \frac{1}{N}\sum_{n}^{N} y[n] - \frac{a}{N}\sum_{n}^{N} x[n] - b$$

Regressão Linear: generalização

Pontos num plano:



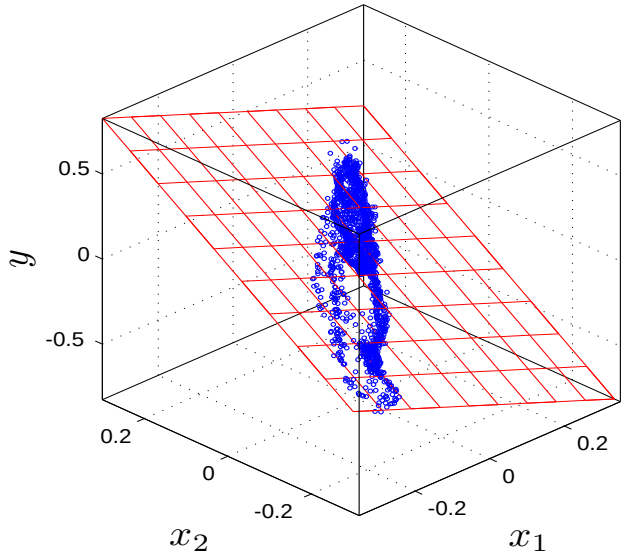
Pontos



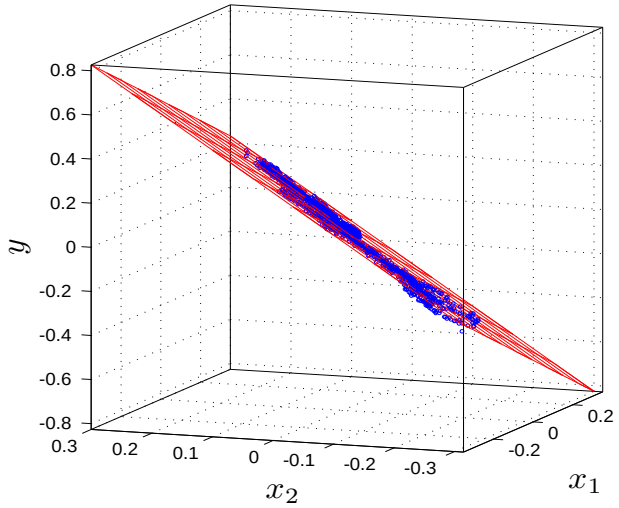
Outra perspectiva

Regressão Linear: generalização

Pontos num plano:



Pontos + plano



Outra perspectiva

Regressão Linear: generalização

Modelo: $\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d = w_0 + \sum_{i=1}^{d} w_i x_i$

Notação vectorial:

$$\hat{y} = \mathbf{w}^\top \mathbf{x} = \mathbf{x}^\top \mathbf{w}$$

onde: $\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$ $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$

Regressão Linear: generalização

Modelo: $\hat{y} = \mathbf{w}^\top \mathbf{x}$

Erro Quadrático: $\mathcal{E} = \frac{1}{N} \sum_{n=1}^{N} \left(y[n] - \mathbf{x}[n]^\top \mathbf{w} \right)^2 = \frac{1}{N} \sum_{n=1}^{N} e[n]^2$

Sistema de equações:

$$\frac{\partial \mathcal{E}}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left\{ \frac{1}{N} \sum_{n=1}^{N} \left(y[n] - \mathbf{x}[n]^\top \mathbf{w} \right)^2 \right\} = 0$$

$$\Leftrightarrow \frac{1}{N} \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}} \left\{ \left(y[n] - \mathbf{x}[n]^\top \mathbf{w} \right)^2 \right\} = 0$$

$$\Leftrightarrow \frac{1}{N} \sum_{n=1}^{N} -2 \left(y[n] - \mathbf{x}[n]^\top \mathbf{w} \right) \underbrace{\frac{\partial \left\{ \mathbf{x}[n]^\top \mathbf{w} \right\}}{\partial \mathbf{w}}}_{=\mathbf{x}[n]} = 0$$

Regressão Linear: generalização

Modelo: $\hat{y} = \mathbf{w}^\top \mathbf{x}$

Erro Quadrático: $\mathcal{E} = \frac{1}{N} \sum_{n=1}^{N} \left(y[n] - \mathbf{x}[n]^\top \mathbf{w} \right)^2 = \frac{1}{N} \sum_{n=1}^{N} e[n]^2$

Sistema de equações:

$$\frac{\partial \mathcal{E}}{\partial \mathbf{w}} = 0$$

$$\Leftrightarrow \underbrace{\frac{1}{N} \sum_{n=1}^{N} y[n]\mathbf{x}[n]}_{(d+1)\times 1} - \underbrace{\frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{x}[n]\mathbf{x}[n]^\top \right)}_{(d+1)\times(d+1)} \mathbf{w} = 0$$

$$\Leftrightarrow \mathbf{w}_{\text{opt}} = \left(\frac{1}{N} \sum_{n=1}^{N} \mathbf{x}[n]\mathbf{x}[n]^\top \right)^{-1} \left(\frac{1}{N} \sum_{n=1}^{N} y[n]\mathbf{x}[n] \right)$$

$$\Leftrightarrow \mathbf{w}_{\text{opt}} = \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{r}_{\mathbf{x}y}$$

Regressão Linear: generalização

Modelo: $\hat{y} = \mathbf{w}^\top \mathbf{x}$

Erro Quadrático: $\mathcal{E} = \frac{1}{N} \sum_{n=1}^{N} \left(y[n] - \mathbf{x}[n]^\top \mathbf{w} \right)^2 = \frac{1}{N} \sum_{n=1}^{N} e[n]^2$

Sistema de equações: $\frac{\partial \mathcal{E}}{\partial \mathbf{w}} = 0 \Leftrightarrow \mathbf{w}_{\text{opt}} = \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{r}_{\mathbf{x}y}$

$$\mathbf{X} = \underbrace{\begin{bmatrix} \mathbf{x}[1] & \mathbf{x}[2] & \dots & \mathbf{x}[N] \end{bmatrix}}_{\text{matriz de } (d+1)\times N} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1[1] & x_1[2] & x_1[3] & \dots & x_1[N] \\ x_2[1] & x_2[2] & x_2[3] & \dots & x_2[N] \\ \vdots & & & \ddots & \vdots \\ x_d[1] & x_d[2] & x_d[3] & \dots & x_d[N] \end{bmatrix}$$

$$\mathbf{Y} = \underbrace{\begin{bmatrix} y[1] & y[2] & \dots & y[N] \end{bmatrix}}_{\text{matriz de } 1\times N} \text{ e } \hat{\mathbf{Y}} = \mathbf{w}\mathbf{X}$$

$$\mathbf{w}_{\text{opt}} = \left(\mathbf{X}\mathbf{X}^\top \right)^{-1} \mathbf{X}\, \mathbf{Y}^\top$$

Regressão: modelos não lineares

- Generalização da regressão linear: as entradas são pré-processadas por funções não lineares $\Phi(\mathbf{x})$:
  $\hat{y} = w_0 + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \ldots + w_p\phi_p(\mathbf{x})$

$$\hat{y} = w_0 + \sum_{j=1}^{p} w_j\phi_j(\mathbf{x}) = \underbrace{\begin{bmatrix} w_0 & w_1 & w_2 & \ldots & w_p \end{bmatrix}}_{\mathbf{w}^\top} \underbrace{\begin{bmatrix} 1 \\ \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \\ \vdots \\ \phi_p(\mathbf{x}) \end{bmatrix}}_{\Phi(\mathbf{x})} = \mathbf{w}^\top\Phi(\mathbf{x})$$

Regressão: modelos não lineares

- Generalização da regressão linear: as entradas são pré-processadas por funções não lineares $\Phi(\mathbf{x})$:
  $\hat{y} = w_0 + w_1\phi_1(\mathbf{x}) + w_2\phi_2(\mathbf{x}) + \ldots + w_p\phi_p(\mathbf{x})$
- Continua a ser um modelo linear em $\mathbf{w}$!
- Erro é uma função quadrática (parábola: um único mínimo)
- Solução: mínimos quadráticos

Regressão: modelos não lineares

- Modelo: $\hat{y} = w_0 + \sum_{j=1}^{p} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x})$

1. Regressão linear básica $(\mathbb{R} \to \mathbb{R})$: $\hat{y} = w_0 + w_1 x$
2. Regressão linear $(\mathbb{R}^d \to \mathbb{R})$: $\hat{y} = \mathbf{w}\top\mathbf{x}$
3. Regressão polinomial $(\mathbb{R} \to \mathbb{R})$:

$$\hat{y} = w_0 + w_1 x + w_2 x^2 + \ldots + w_p x^p = \mathbf{w}^\top \Phi(\mathbf{x})$$

com: $\Phi(\mathbf{x}) = [1 \; x \; x^2 \; \ldots \; x^p]^\top$

4. Regressão polinomial $(\mathbb{R}^d \to \mathbb{R})$:

$$\begin{aligned} \hat{y} &= w_0 + w_1 x_1 + w_2 x_2^2 + w_3 x_2 x_k^2 x_d + \ldots + w_n x_1^2 x_2^5 + \ldots \\ &= \mathbf{w}^\top \Phi(\mathbf{x}) \\ \Phi(\mathbf{x}) &= [1 \;\; x_1 \;\; x_2^2 \;\; x_2 x_k^2 x_d \ldots \;\; x_1^2 x_2^5 \;\; \ldots]^\top \end{aligned}$$

Regressão: modelos não lineares

- Modelo: $\hat{y} = w_0 + \sum_{j=1}^{p} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x})$
- Funções de base $\phi_j(\cdot)$ permitem modelar comportamentos não lineares dos dados.
- Polinómios (ex. $\phi_j(x) = x^j$) são funções globais (afectam todo o espaço dos dados). Existem várias funções (locais e globais):
  - *Splines (polinómios locais)*
  - Gaussianas: $\phi_j(x) = \exp\left(\frac{1}{2\sigma^2}(x - \mu_j)^2\right)$
  - Sigmóides: $\phi_j(x) = \frac{1}{1 + \exp\frac{x-\mu_j}{\sigma}}$
  - FFTs, *wavelets*,...

Regressão: modelos não lineares

- Modelo: $\hat{y} = w_0 + \sum_{j=1}^{p} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x})$
- Erro quadrático médio: $\mathcal{E} = \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{X}} (y - \mathbf{w}^\top \Phi(\mathbf{x}))^2$
- Sistema de equações:

$$\mathbf{w}_{\text{opt}} = \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{r}_{\mathbf{x}y}$$

com: $\mathbf{r}_{\mathbf{x}y} = \frac{1}{N} \sum_{n=1}^{N} y[n] \Phi(\mathbf{x}[n])$ $\mathbf{R}_{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^{N} \Phi(\mathbf{x}[n]) \Phi(\mathbf{x}[n])^\top$

Regressão: modelos não lineares

- Modelo: $\hat{y} = w_0 + \sum_{j=1}^{p} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x})$
- Erro quadrático médio: $\mathcal{E} = \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{X}} (y - \mathbf{w}^\top \Phi(\mathbf{x}))^2$
- Sistema de equações:

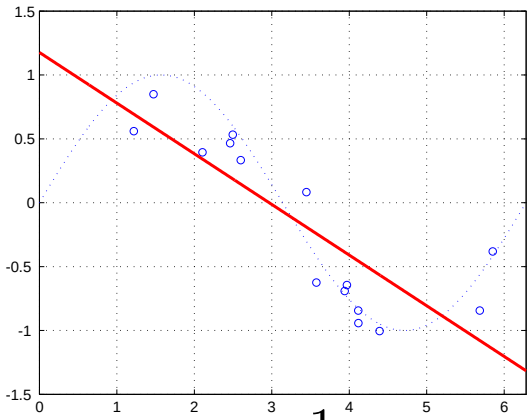
$$\mathbf{\Psi} = [\Phi(\mathbf{x}[1]), \Phi(\mathbf{x}[2]), \cdots, \Phi(\mathbf{x}[N])] = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \phi(\mathbf{x}_1[1]) & \phi(\mathbf{x}_1[2]) & \cdots & \phi(\mathbf{x}_1[N]) \\ \phi(\mathbf{x}_2[1]) & \phi(\mathbf{x}_2[2]) & \cdots & \phi(\mathbf{x}_1[N]) \\ \vdots & \vdots & \ddots & \vdots \\ \phi(\mathbf{x}_d[1]) & \phi(\mathbf{x}_d[2]) & \cdots & \phi(\mathbf{x}_d[N]) \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} y[1] & y[2] & \cdots & y[N] \end{bmatrix}$$

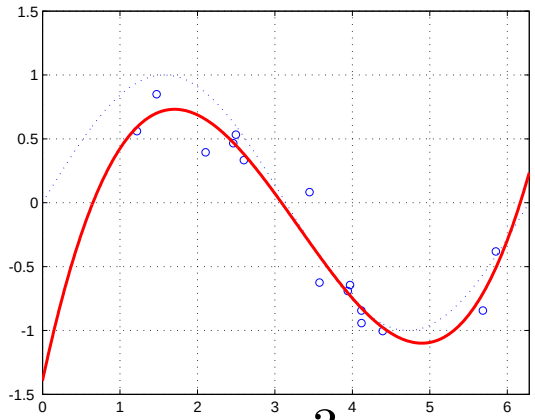
$$\mathbf{w}_{\text{opt}} = \mathbf{R}_{\mathbf{x}}^{-1} \mathbf{r}_{\mathbf{x}y} = \left(\mathbf{\Psi} \mathbf{\Psi}^\top \right)^{-1} \mathbf{\Psi} \mathbf{Y}^\top$$

Regressão: modelos não lineares

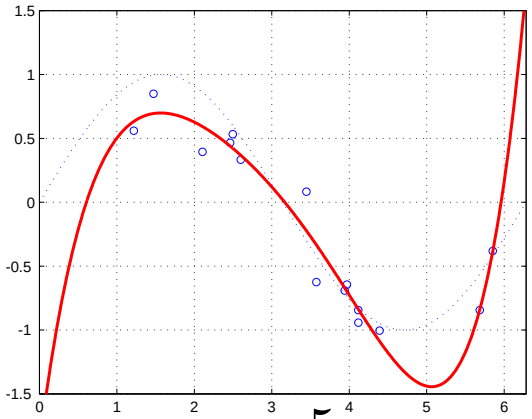
Regressão polinomial: $\hat{y} = w_0 + w_1 x + w_2 x^2 + \ldots + w_p x^p$



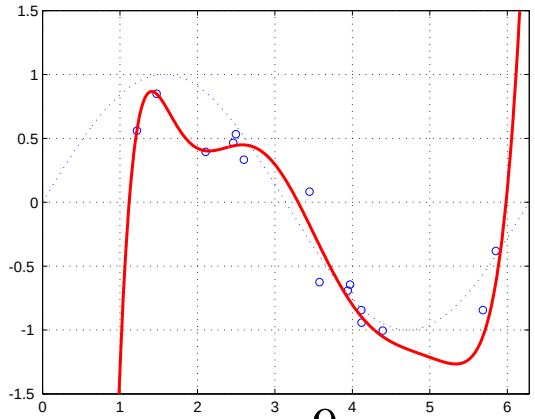
$p = 1$



$p = 3$



$p = 5$

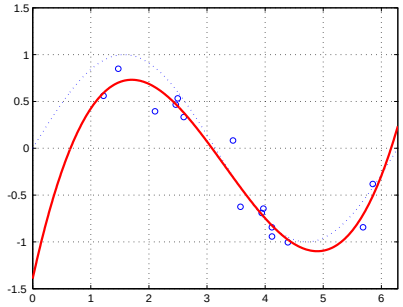


$p = 9$

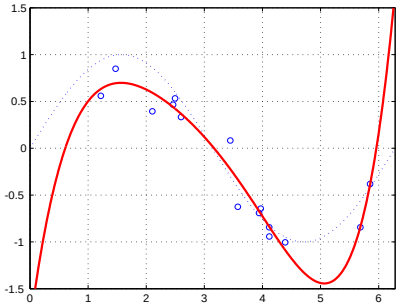
Conjunto com 15 pontos

Regressão: modelos não lineares

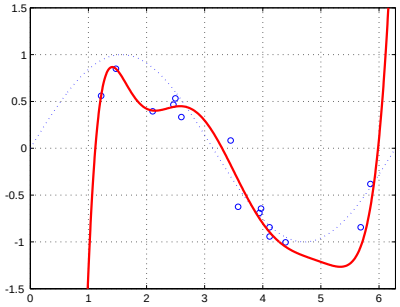
Regressão polinomial: $\hat{y} = w_0 + w_1 x + w_2 x^2 + \ldots + w_p x^p$



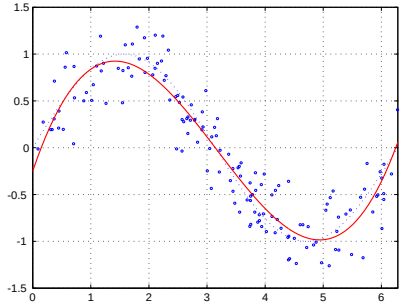
$p = 3$

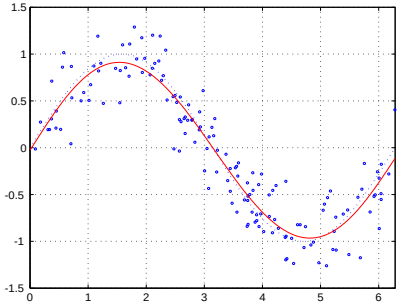


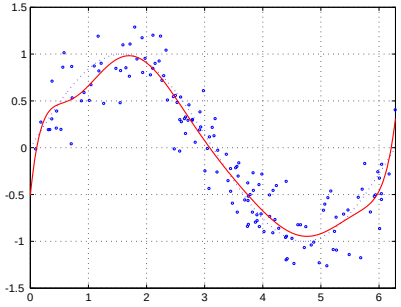
$p = 5$



$p = 9$





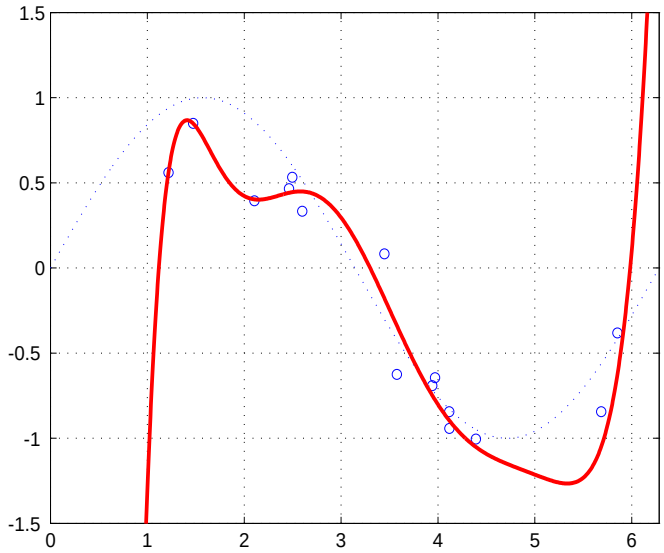


1ª linha com 15 pontos e a 2ª com 150

- Mais pontos ajuda! Modelos complexos melhor comportados
- Factoide: 10 pontos por parâmetro

Regressão: capacidade de generalização

- Generalização é fundamental para obter um bom desempenho com novos dados
- Sobre aprendizagem resulta numa fraca capacidade de generalização

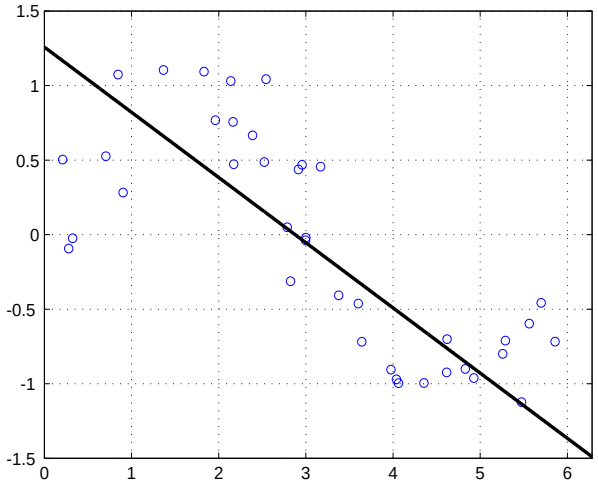


←sobre aprendizagem!

Regressão: capacidade de generalização

Regularização

- Ao aumentar a ordem do modelo, também aumentamos o valor dos coeficientes w_i



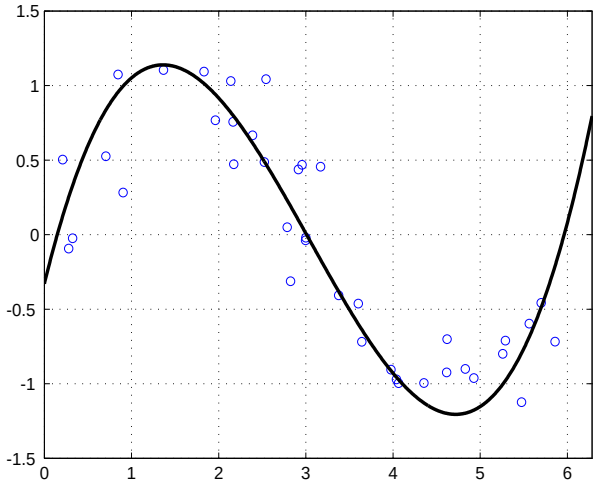
$p = 1$

	$p=1$	$p=3$	$p=5$	$p=9$
w_0	1.26			
w_1	−0.39			
w_2				
w_3				
w_4				
w_5				
w_6				
w_7				
w_8				
w_9				

Regressão: capacidade de generalização

Regularização

- Ao aumentar a ordem do modelo, também aumentamos o valor dos coeficientes w_i



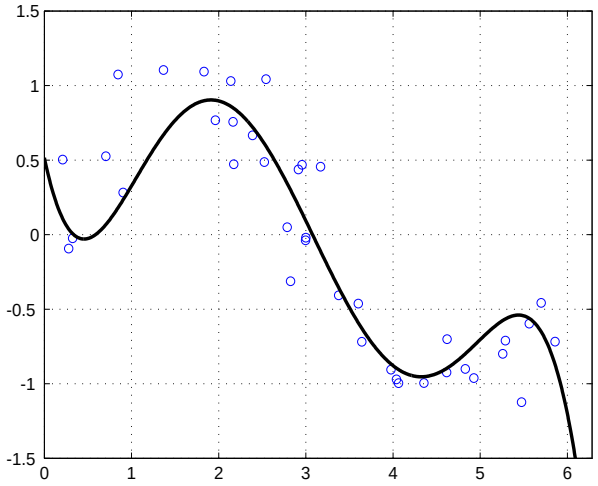
$p = 3$

	$p=1$	$p=3$	$p=5$	$p=9$
w_0	0.96	−0.33		
w_1	−0.35	2.39		
w_2		−1.13		
w_3		0.12		
w_4				
w_5				
w_6				
w_7				
w_8				
w_9				

Regressão: capacidade de generalização

Regularização

- Ao aumentar a ordem do modelo, também aumentamos o valor dos coeficientes w_i



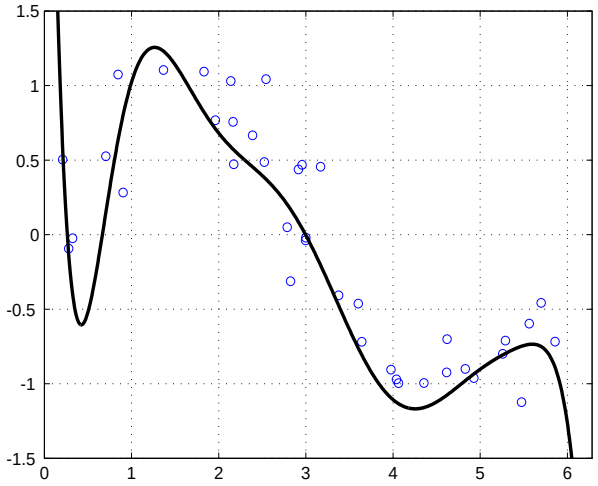
$p = 5$

	$p=1$	$p=3$	$p=5$	$p=9$
w_0	0.96	−0.60	0.52	
w_1	−0.35	2.37	−2.76	
w_2		−1.05	4.32	
w_3		0.11	−2.13	
w_4			0.41	
w_5			-0.03	
w_6				
w_7				
w_8				
w_9				

Regressão: capacidade de generalização

Regularização

- Ao aumentar a ordem do modelo, também aumentamos o valor dos coeficientes w_i



$p = 9$

	$p=1$	$p=3$	$p=5$	$p=9$
w_0	0.96	−0.60	0.07	6.3
w_1	−0.35	2.37	−0.60	−45.2
w_2		−1.05	2.27	117.5
w_3		0.11	−1.35	−120.2
w_4			0.28	75.3
w_5			−0.02	−28.5
w_6				6.7
w_7				−0.9
w_8				0.1
w_9				0.0

Regressão: capacidade de generalização

Regularização

- Penalizar pesos com valor elevado - incluir termo de penalização na função do erro

$$\mathcal{E}_{\mathrm{Tot}} = \mathcal{E}_{\mathrm{Q}}(\mathbf{x}, \mathbf{w}) + \lambda \mathcal{E}_{\mathrm{W}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} \left(y[n] - \mathbf{w}^{\top} \Phi(\mathbf{x}[n]) \right)^2 + \lambda \mathcal{E}_{\mathrm{W}}(\mathbf{w})$$

onde λ é um parâmetro que controla a influência do termo de regularização na função total do erro.

Regressão: capacidade de generalização

Regularização

- Penalizar pesos com valor elevado - incluir termo de penalização na função do erro
- Soma do valor quadrático dos coeficientes:

$$\mathcal{E}_{\mathrm{W}}(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^\top\mathbf{w}$$

$$\mathcal{E}_{\mathrm{Tot}} = \frac{1}{N}\sum_{n=1}^{N}\left(y[n] - \mathbf{w}^\top\Phi(\mathbf{x}[n])\right)^2 + \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w}$$

- $\dfrac{\partial\mathcal{E}_{\mathrm{Tot}}}{\partial\mathbf{w}} = 0 \iff \mathbf{w} = \left(\lambda\mathbf{I} + \mathbf{\Psi}\mathbf{\Psi}^\top\right)^{-1}\mathbf{\Psi}\mathbf{Y}^\top$
- Também conhecido por *weight decay* e por *ridge regression*

Regressão: capacidade de generalização

Regularização

- Penalizar pesos com valor elevado - incluir termo de penalização na função do erro
- Outras formas de regularização:

$$\mathcal{E}_{\text{Tot}} = \frac{1}{N}\sum_{n=1}^{N}\left(y[n] - \mathbf{w}^{\top}\Phi(\mathbf{x}[n])\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{p}|w_j|^q$$

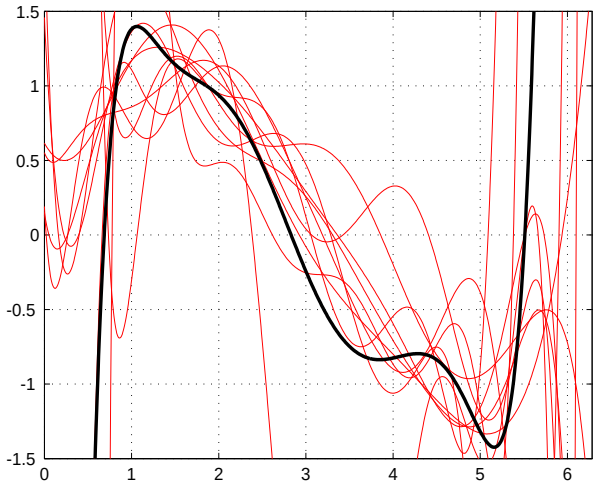
com $q > 0$

- Solução analítica através de multiplicadores de Lagrange.
- $q = 2$ regualrizador quadrático, $q = 1$ *lasso*

Regressão: capacidade de generalização

Regularização

- Penalizar pesos com valor elevado - incluir termo de penalização na função do erro



$p = 9$

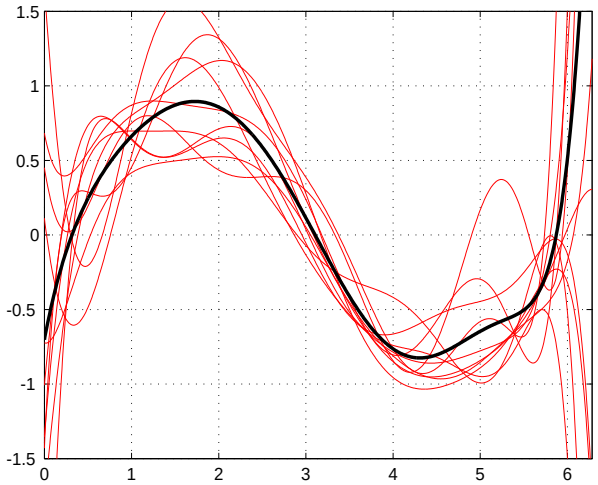
$\mathbf{w}$ calculado com um média de 10 estimações, com conjuntos de 15 pontos

	$\lambda = 0$	$\lambda = 10^{-5}$	$\lambda = 1$
w_0	-44.2		
w_1	170.4		
w_2	-266.0		
w_3	228.8		
w_4	-120.3		
w_5	40.5		
w_6	-8.8		
w_7	1.2		
w_8	-0.01		
w_9	0		

Regressão: capacidade de generalização

Regularização

- Penalizar pesos com valor elevado - incluir termo de penalização na função do erro



$p = 9$

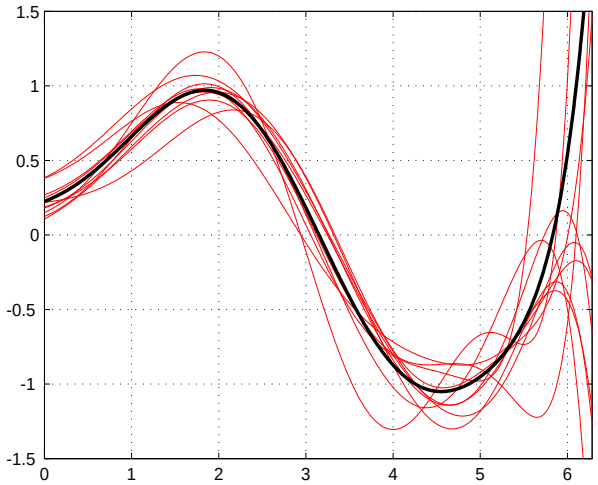
$\mathbf{w}$ calculado com um média de 10 estimações, com conjuntos de 15 pontos

	$\lambda = 0$	$\lambda = 10^{-5}$	$\lambda = 1$
w_0	-44.2	-0.70	
w_1	170.4	3.11	
w_2	-266.0	-3.72	
w_3	228.8	3.49	
w_4	-120.3	-2.20	
w_5	40.5	0.90	
w_6	-8.8	-0.24	
w_7	1.2	0.04	
w_8	-0.01	0	
w_9	0	0	

Regressão: capacidade de generalização

Regularização

- Penalizar pesos com valor elevado - incluir termo de penalização na função do erro



$p = 9$

$\mathbf{w}$ calculado com um média de 10 estimações, com conjuntos de 15 pontos

	$\lambda = 0$	$\lambda = 10^{-5}$	$\lambda = 1$
w_0	-44.2	-0.70	0.22
w_1	170.4	3.11	0.23
w_2	-266.0	-3.72	0.18
w_3	228.8	3.49	0.10
w_4	-120.3	-2.20	-0.03
w_5	40.5	0.90	-0.08
w_6	-8.8	-0.24	0.04
w_7	1.2	0.04	-0.01
w_8	-0.01	0	0
w_9	0	0	0

Compromisso: Variância vs Polarização

- Modelo $\mathcal{F}(\mathbf{x}, \mathbf{w})$ estimado com N amostras de $\mathbf{x}$ (modelo estimado com o par de conjuntos $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$)
- Qual o erro se tivéssemos um número infinito de conjuntos $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_\infty$?

Notação:

$\mathbf{x}$	entradas
$y = \mathcal{H}(\mathbf{x}) + \epsilon$	processo verdadeiro
ϵ	ruído gaussiano média nula: $\mathcal{N}(0, \sigma^2)$
$\hat{y} = \mathcal{F}(\mathbf{x}, \mathbf{w})$	aproximação obtida com o conjunto $\mathcal{D}$

- Erro no conjunto $\mathcal{D}_k$: $\mathcal{E}_k = \dfrac{1}{N} \displaystyle\sum_{\mathbf{x}, y \in \mathcal{D}_k} (y - \mathcal{F}(\mathbf{x}, \mathbf{w}))^2$
- Erro em todos os conjuntos:

$$\mathcal{E}_{\text{Tot}} = \lim_{K \to \infty} \frac{1}{K} \sum_{k=1}^{K} \mathcal{E}_k = \mathbb{E}\left\{(y - \mathcal{F}(\mathbf{x}, \mathbf{w}))^2\right\}$$

Compromisso: Variância vs Polarização

- Erro médio:

$$\begin{aligned}\mathcal{E}_{\text{Tot}} &= \iint_{\mathbf{x},y} (y - \underbrace{\mathcal{F}(\mathbf{x},\mathbf{w})}_{\hat{y}})^2 p(y,\mathbf{x}) d\mathbf{x} dy \\ &= \mathbb{E}\left\{(y-\hat{y})^2\right\} = \mathbb{E}\{(\underbrace{y - \mathcal{H}(\mathbf{x})}_{\epsilon} + \mathcal{H}(\mathbf{x}) - \hat{y})^2\}\end{aligned}$$

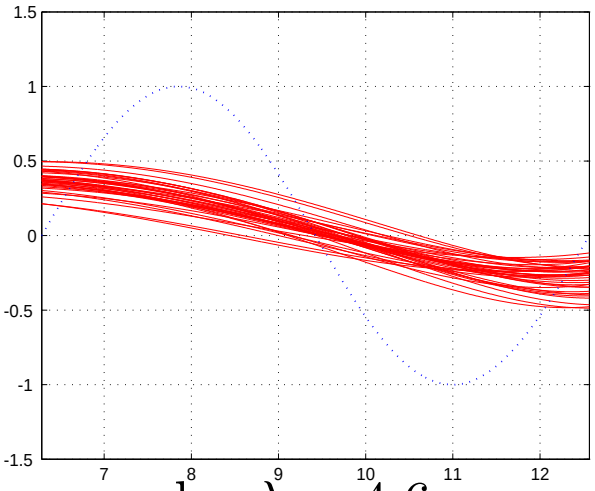
  Com algumas manipulações resulta:

$$\begin{aligned}\mathcal{E}_{\text{Tot}} &= \mathbb{E}\left\{\epsilon^2\right\} + \mathbb{E}\left\{(y - \mathbb{E}\{\hat{y}\})^2\right\} + \mathbb{E}\left\{(\hat{y} - \mathbb{E}\{\hat{y}\})^2\right\} \\ &= \mathbb{V}\left\{\text{ruído}\right\} + \text{polarização}^2 + \mathbb{V}\left\{\hat{y}\right\}\end{aligned}$$

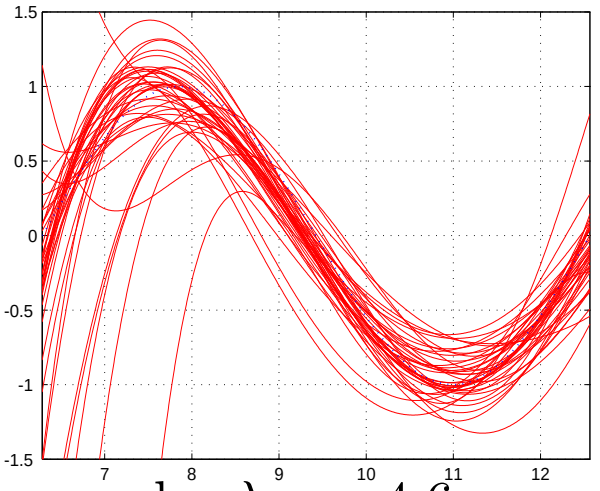
- O erro é a soma da variância do ruído (não podemos controlar) mais a polarização (ao quadrado) do modelo mais a variância do modelo.

Compromisso: Variância vs Polarização

- Erro médio $= \mathbb{V}\{\text{ruído}\} + \text{polarização}^2 + \mathbb{V}\{\hat{y}\}$
- Compromisso variância vs polarização:
  - λ elevados resultam, numa polarização elevada
  - λ pequenos resultam numa variância elevada
  - O melhor modelo é o que consegue obter um balanço óptimo entre os dois casos



$\log \lambda = 4.6$



$\log \lambda = -4.6$