PMR5251 - Avaliação do Comportamento Mecânico de Materiais Utilizando uma Abordagem de ML

MACHINE LEARNING: REGRESSION

Izabel F. Machado
Larissa Driemeier

NOSSAS AULAS

Data	Assunto	Conteúdo principal
01/10	Introdução ao Aprendizado de Máquinas	Teoria conceitual
15/10	Redes Neurais	Teoria e Prática
22/10	Regressão	Teoria e Prática
12/11	Classificação	Teoria e Prática

AULA DE HOJE

- Máxima verossimilhança
- O que é regressão
- Suposições de Gauss Markov
- Regressão Linear
  - Simples
  - Múltipla
  - Polinomial

MÁXIMA VEROSSIMILHANÇA

O que é isso?

MÁXIMA VEROSSIMILHANÇA

Vamos supor que você está analisando o que os clientes da sua confeitaria estão achando de seu novo bolo. Você consegue as seguintes respostas:

Cliente	Você gostou do bolo?
1	Sim
2	Não
3	Sim
4	Não
5	Sim
6	Sim
7	Sim
8	Sim
9	Não
10	Sim

A distribuição de Bernoulli é a distribuição de probabilidade discreta de uma variável aleatória que obtém uma saída binária: 1 com probabilidade p e 0 com probabilidade $(1-p)$.

$$P(X=x)=\begin{cases} p & se\ x=1 \\ 1-p & se\ x=0 \end{cases}$$

MODELO DE BERNOULLI

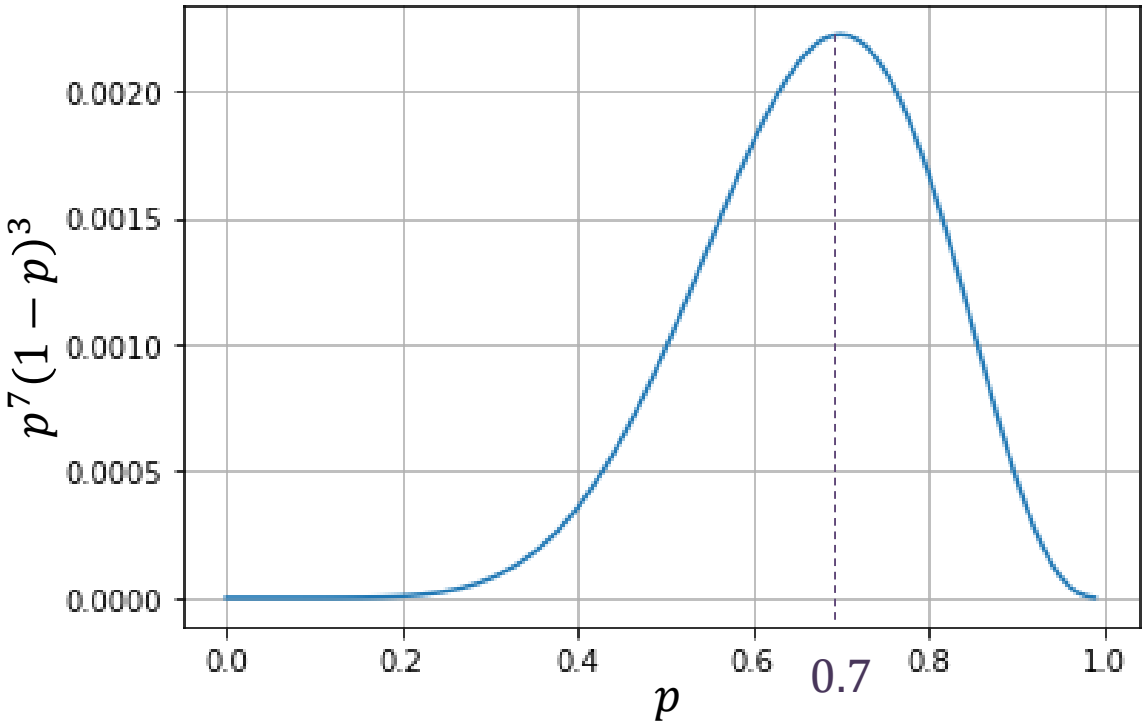
Para os dados que temos....

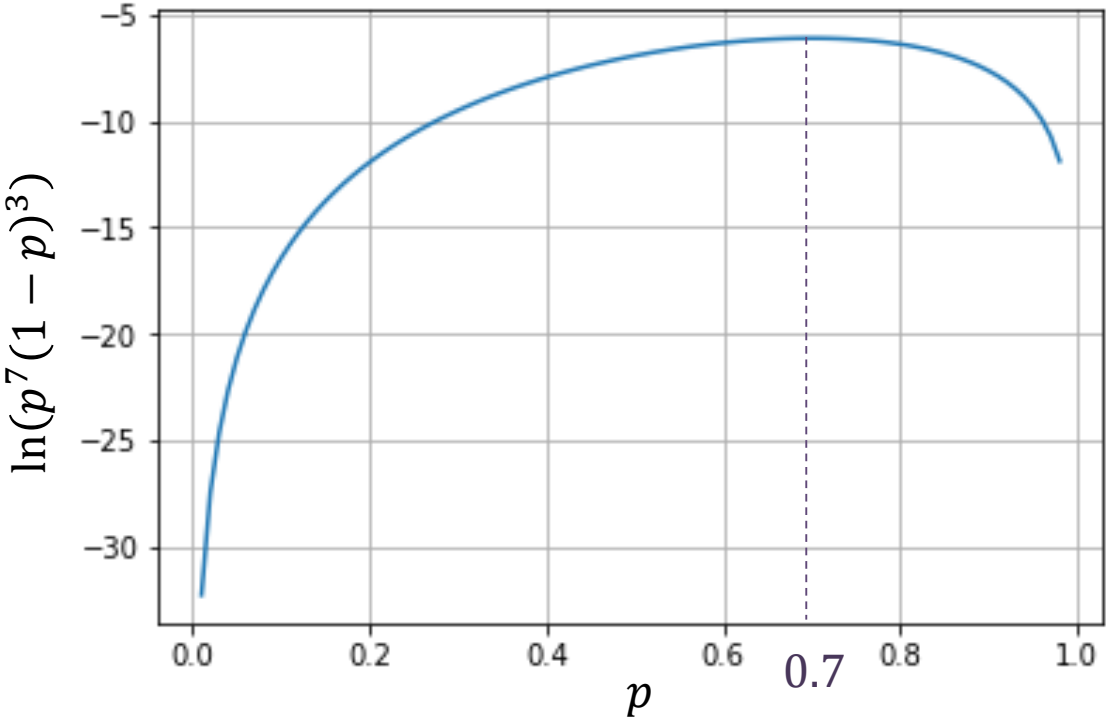
$$p \times (1-p) \times p \times (1-p) \times p \times p \times p \times p \times (1-p) \times p$$

$$p^7(1-p)^3$$

$$\hat{p} = \arg\max_{\boldsymbol{p}} p^7(1-p)^3$$

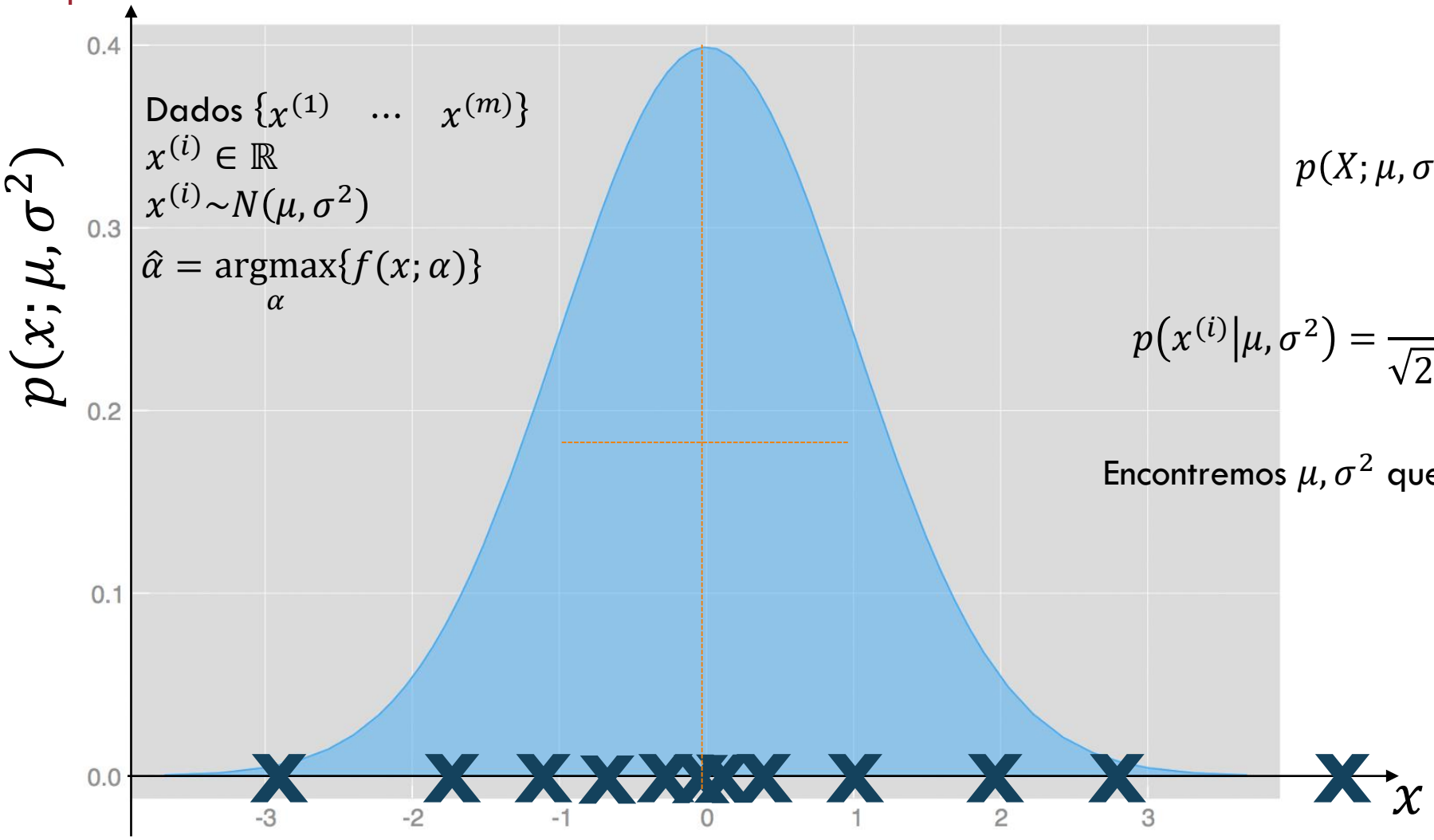
MÁXIMA VERISSIMILHANÇA?





$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\omega}) = \arg\max_{\boldsymbol{\omega}} \sum_{i=1}^{m} \ln p\left(y^{(i)}\middle|\boldsymbol{x}^{(i)};\boldsymbol{\omega}\right)$$

OUTRO EXEMPLO



$$p(X;\mu,\sigma^2) = \prod_{i=1}^{m} p\left(x^{(i)}\middle|\mu,\sigma^2\right)$$

$$p\left(x^{(i)}\middle|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=0}^{m}\left(x^{(i)}-\mu\right)^2\right)$$

Encontremos μ, σ^2 que maximizem $p\left(x^{(i)}\middle|\mu,\sigma^2\right)$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)}-\mu\right)^2$$

PARA VOCÊ ENTENDER PASSO A PASSO

$$L(\mu,\sigma^2) = \underset{\mu,\sigma^2}{\operatorname{argmax}} \log\left[\frac{1}{(2\pi\sigma^2)^{m/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=0}^{m}\left(x^{(i)}-\mu\right)^2\right)\right]$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}}\left[-\frac{m}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=0}^{m}\left(x^{(i)}-\mu\right)^2\right]$$

$$\hat{\mu} \rightarrow \frac{\partial}{\partial\mu}L(\mu,\sigma^2) = 0$$

$$-\frac{1}{2\sigma^2}\sum_{i=0}^{m} -\left(x^{(i)}-\hat{\mu}\right) = 0$$

$$\sum_{i=0}^{m}\left(x^{(i)}\right) = m\hat{\mu}$$

$$\hat{\mu} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\hat{\sigma}^2 \rightarrow \frac{\partial}{\partial\sigma^2}L(\mu,\sigma^2)$$

$$= -\frac{m}{2}\frac{2\pi\sigma}{2\pi\sigma^2}\ln 2\pi\sigma^2 + \frac{\sigma}{2\sigma^4}\sum_{i=0}^{m}\left(x^{(i)}-\mu\right)^2$$

$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)}-\mu\right)^2$$

REGRESSÃO LINEAR

O QUE É REGRESSÃO?

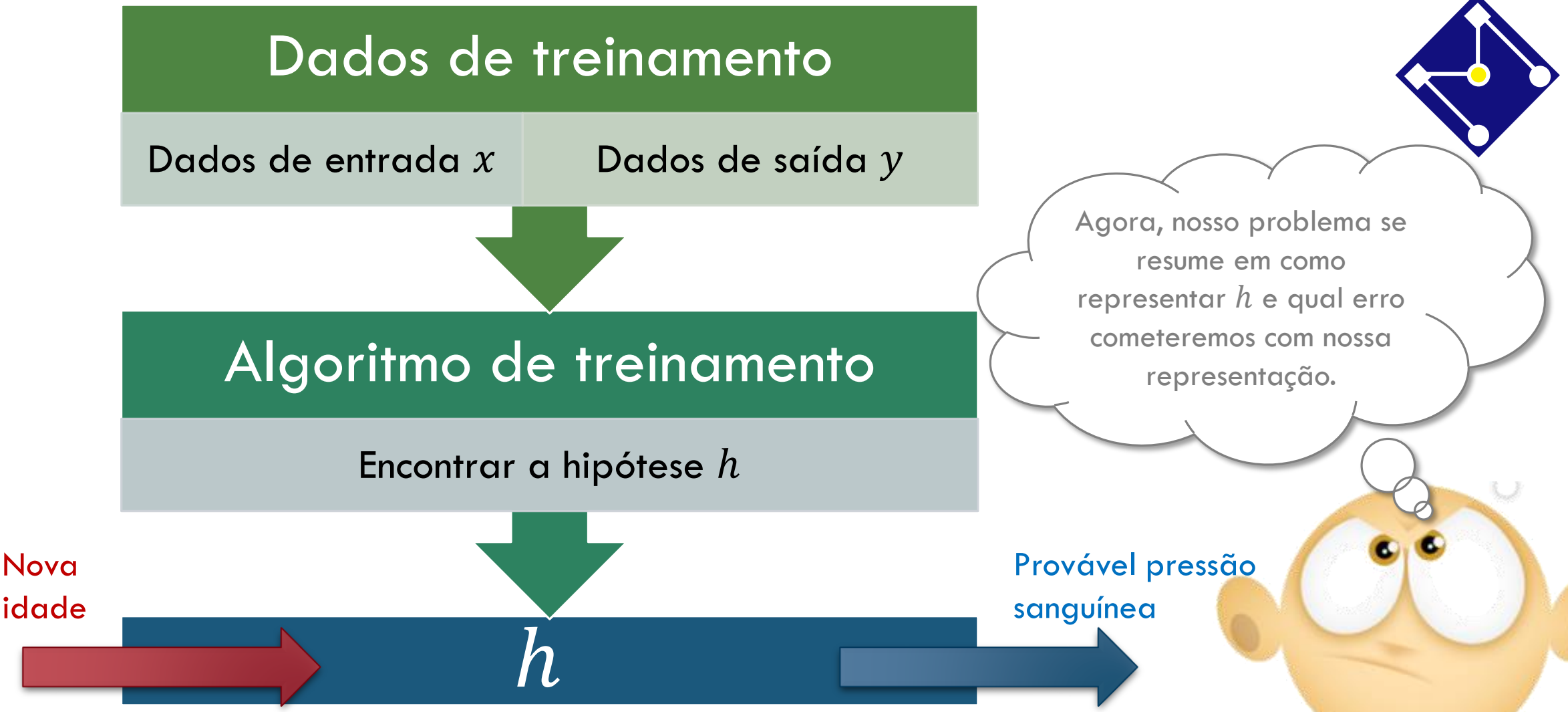
Regressão refere-se a prever a saída de uma **variável numérica (dependente)** a partir de um conjunto de uma ou mais **variáveis independentes** . Uma equação de regressão é usada em estatística para descobrir qual relação existe , e se existe, entre conjuntos de dados.

Em Aprendizado de Máquina, essa equação é obtida através de um *algoritmo de treinamento* utilizando os m dados $\left(x^{(i)}, y^{(i)}\right)$.

A análise de regressão é uma maneira de classificar matematicamente quais dessas variáveis realmente têm impacto. Responde às perguntas: *Quais fatores são mais importantes? O que podemos ignorar? Como esses fatores interagem entre si? E, talvez o mais importante, até que ponto estamos certos sobre todos esses fatores?*

FEATURES **TARGET**

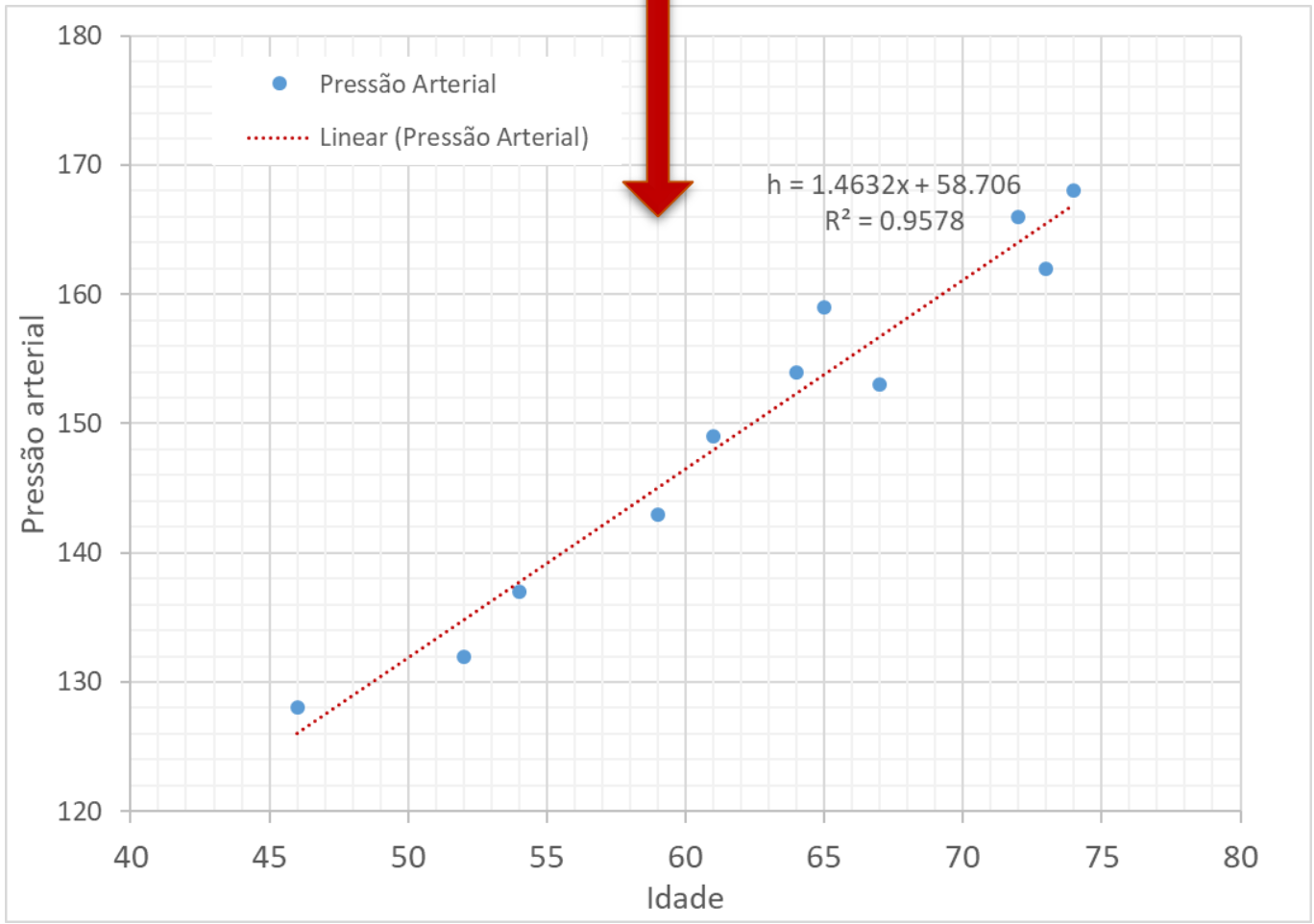
x: Variável independente		y: variável dependente
Idade	**Peso (Kg)**	**Pressão Arterial**
52	78.5	132
59	83.5	143
67	88.0	153
73	95.7	162
64	88.9	154
74	99.8	168
54	85.3	137
61	85.3	149
65	93.9	159
46	75.7	128
72	98.4	166


Dados de treinamento
Dados de entrada x
Dados de saída y
Algoritmo de treinamento
Encontrar a hipótese h
Nova idade
h
Provável pressão sanguínea

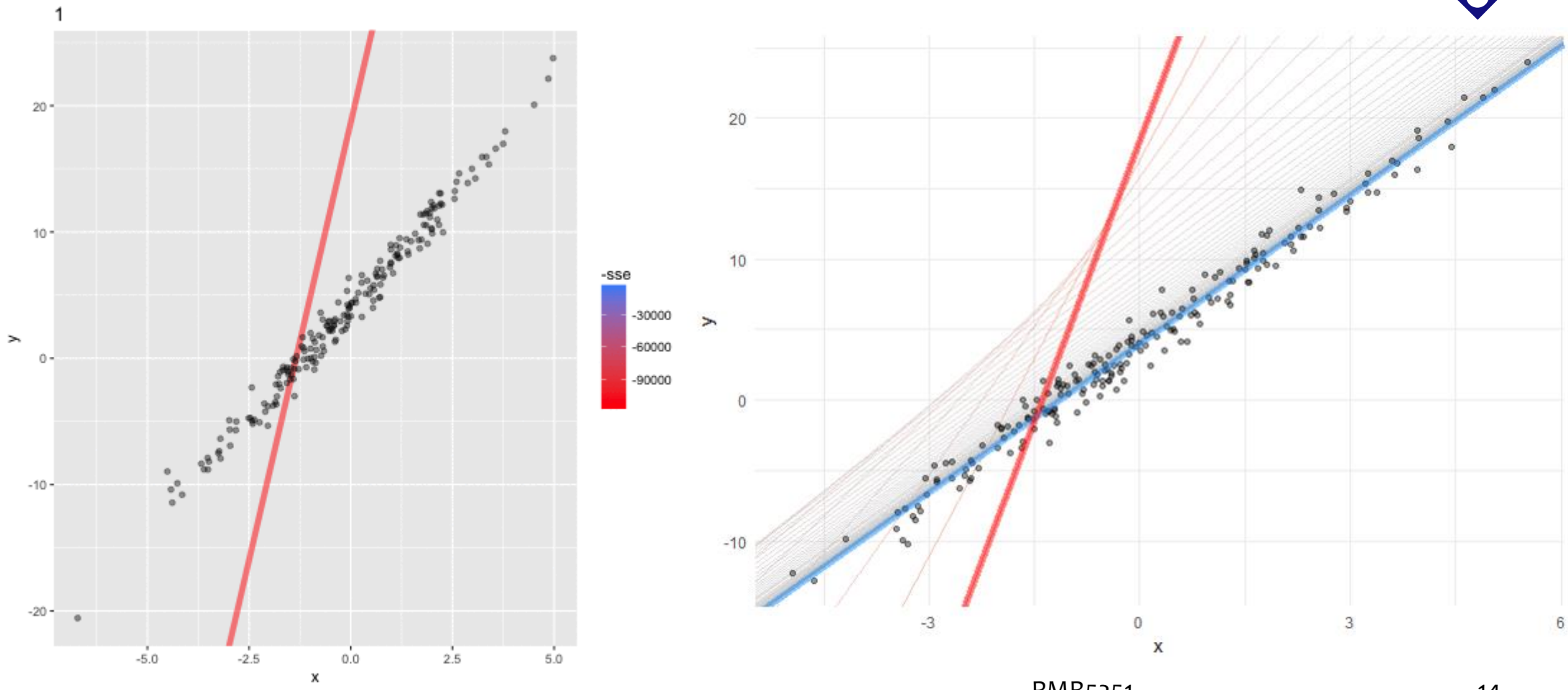
Agora, nosso problema se resume em como representar h e qual erro cometeremos com nossa representação.

HIPÓTESE h

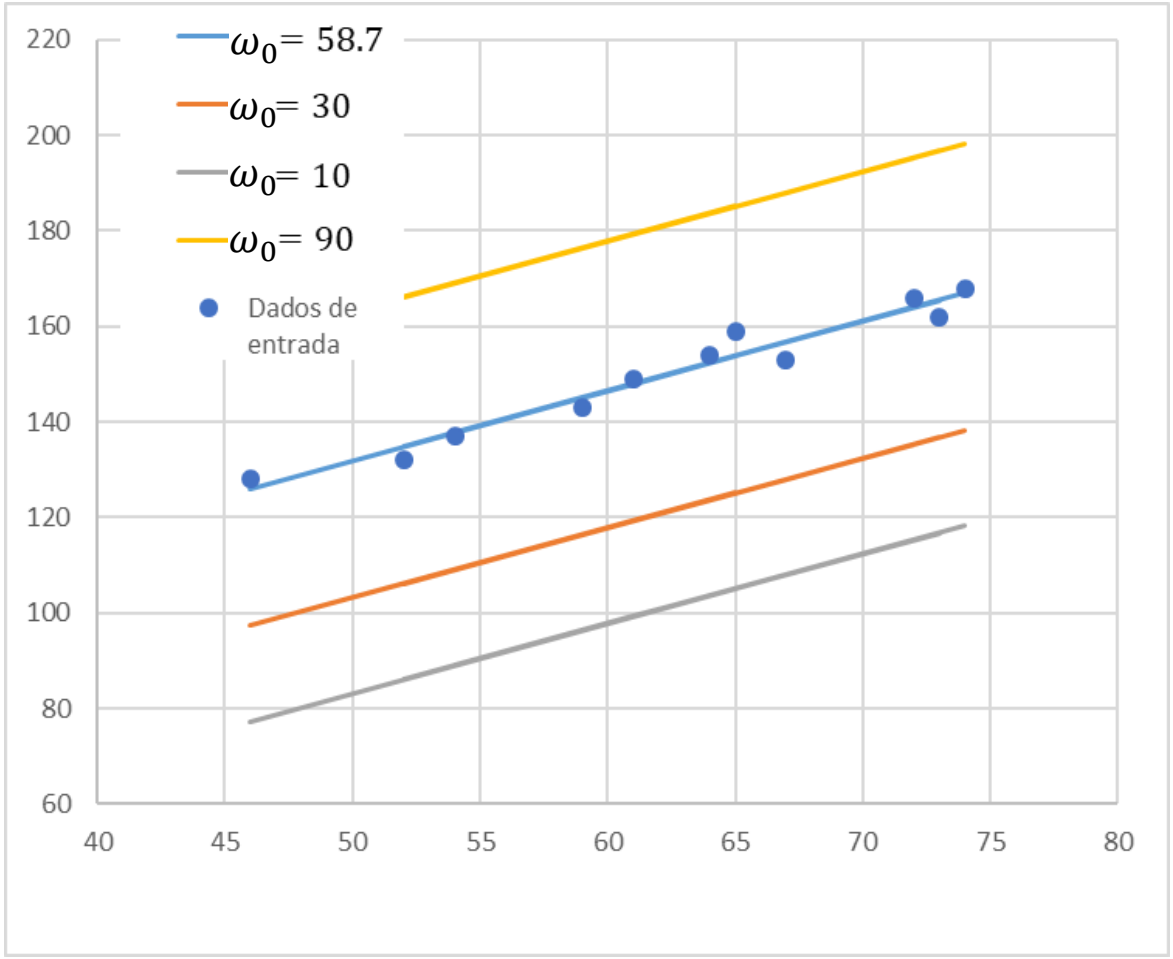
Hipótese h mapeia x em y.



INTUIÇÃO DE NOSSO PROBLEMA DE REGRESSÃO



Fonte: https://towardsdatascience.com/machine-learning-fundamentals-via-linear-regression-41a5d11f5220

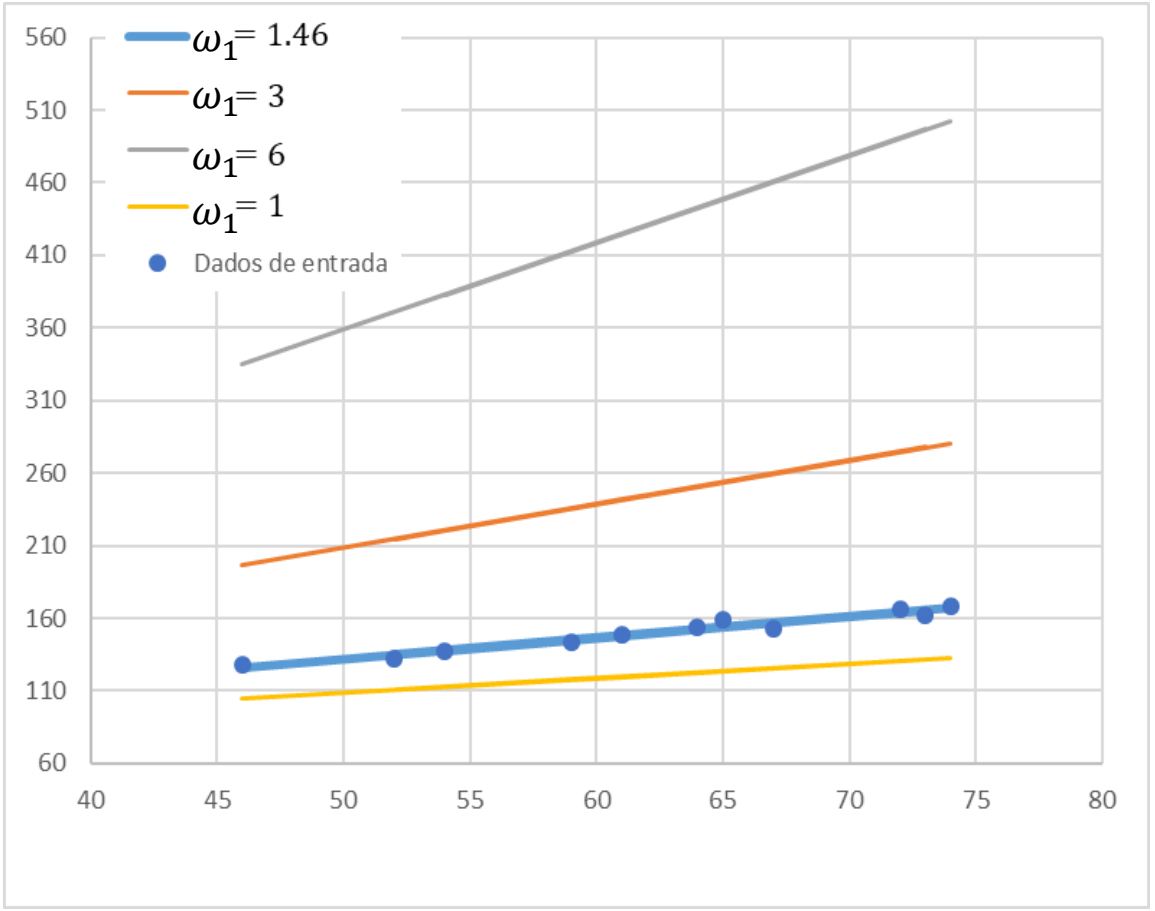


ω_0 e ω_1 serão definidos através dos meus m dados.

De que forma????

ω_0 e ω_1 devem ser tais que cometerei o menor erro possível quando uso a hipótese $h(x) = \hat{y}$ para prever y.

$$h(x) = \hat{y} = \omega_0 + \omega_1 x$$



MAS, COMO EU MEÇO O QUANTO ESTOU ERRANDO???

Define-se o resíduo no conjunto de dados i como a distância entre a resposta da minha hipótese, $\hat{y}^{(i)}$, e a resposta exata $y^{(i)}$,

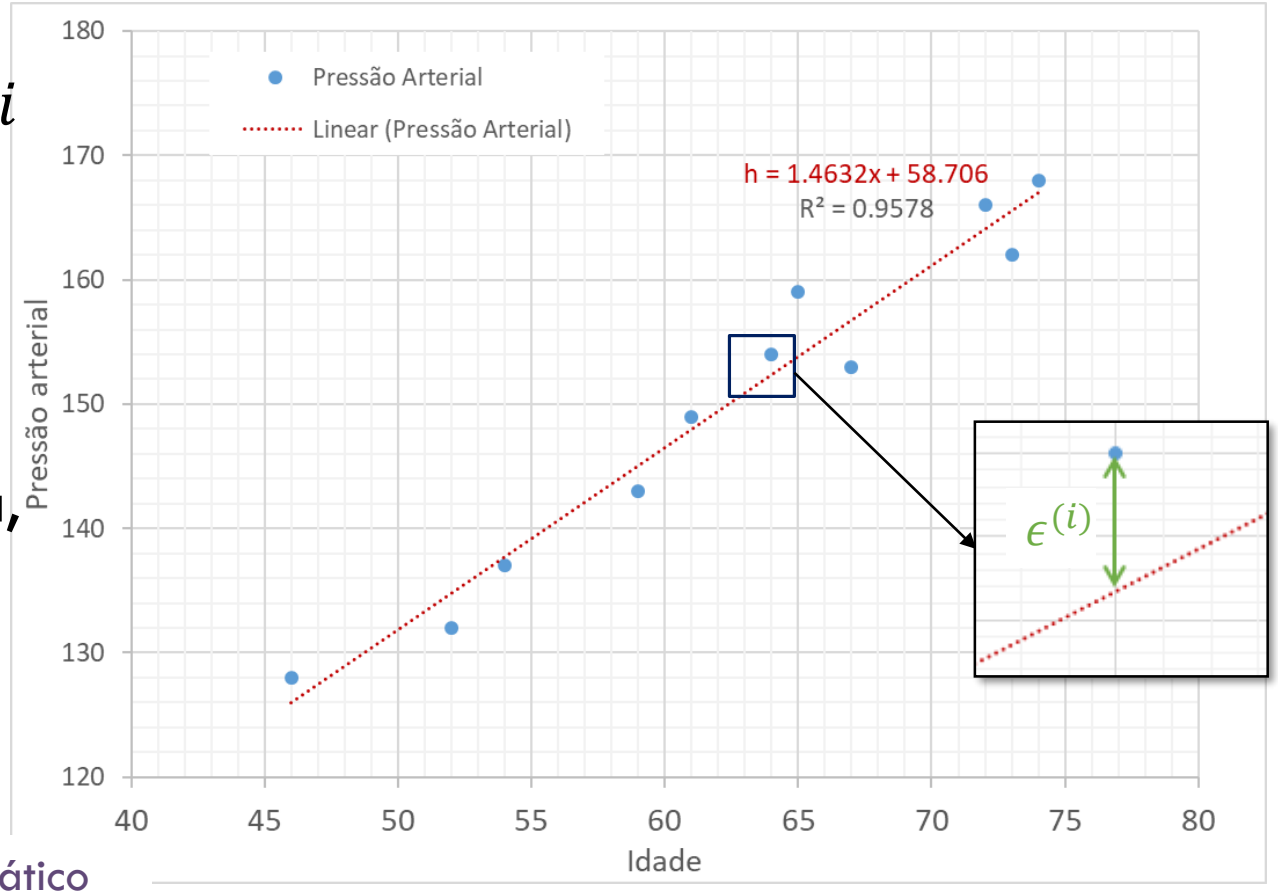
$$e^{(i)} = y^{(i)} - \hat{y}^{(i)}$$

O resíduo quadrático de todo o problema, denominado aqui como função **CUSTO**, será definido como,

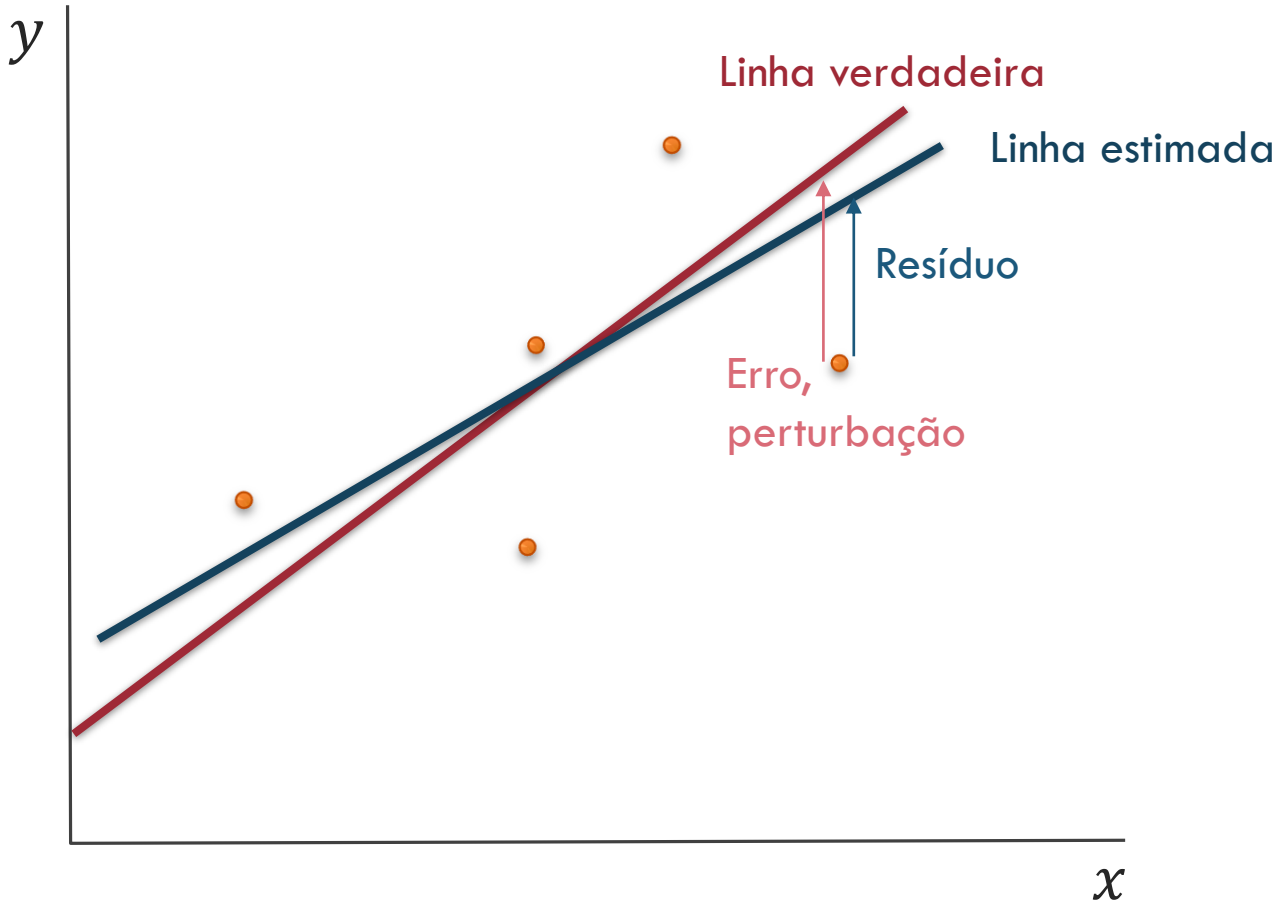
$$J(\omega_0, \omega_1) = \frac{1}{m}\sum_{i=1}^{m}\left[e^{(i)}\right]^2$$

Função Erro Quadrático

Função Erro Quadrático Total (EQT)

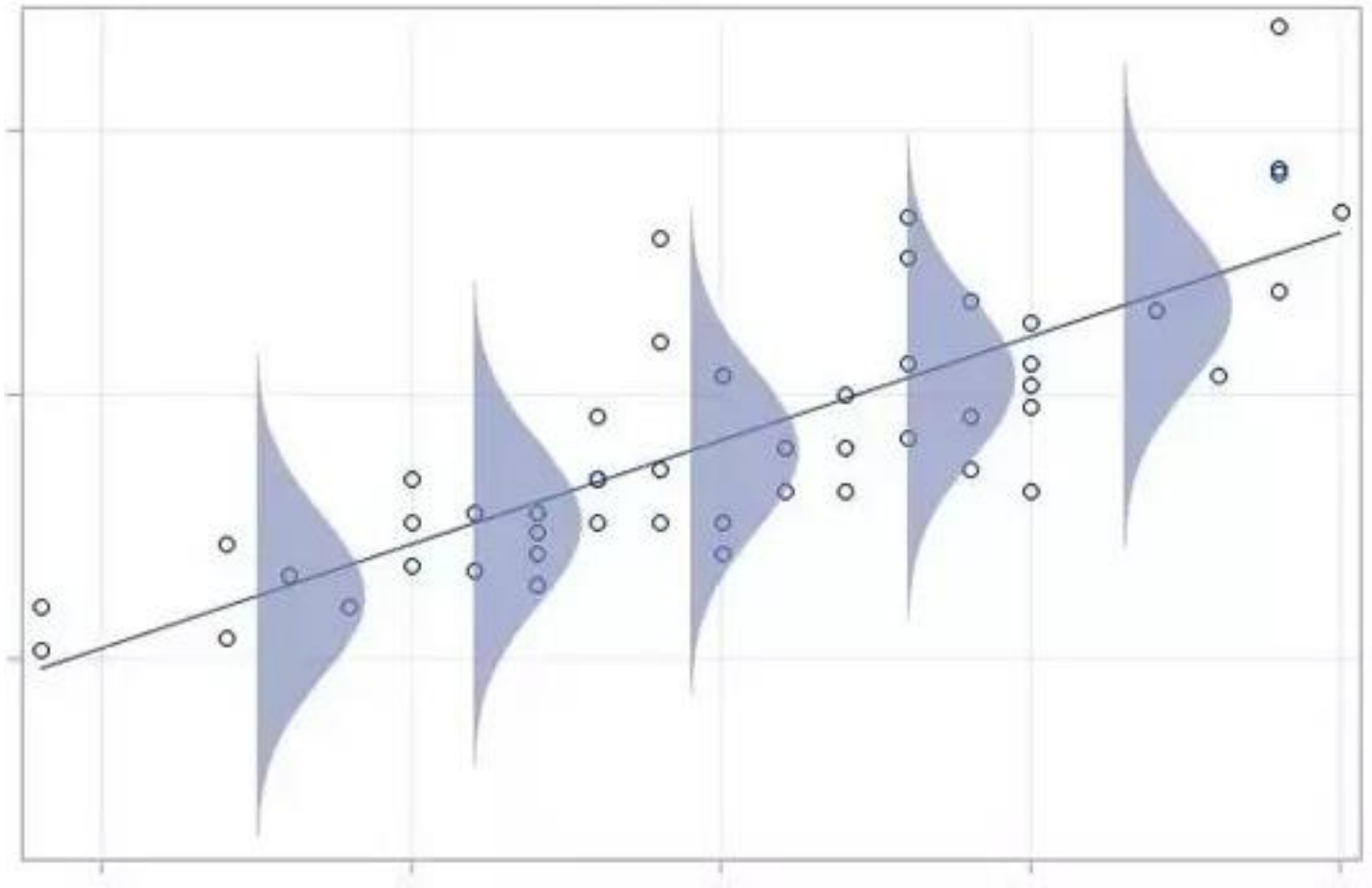


ERRO VS RESÍDUO



O resíduo é calculado após a execução do modelo de regressão e é a diferença entre os valores observados e os valores estimados.

O erro do conjunto de dados é a diferença entre os valores observados e os valores verdadeiros, não observáveis.



Assume-se que o erro tem uma variância constante e é normalmente distribuído.

TEOREMA DE GAUSS MARKOV

Ao usar estimadores não viesados, ié, $E\left(\epsilon^{(i)}\right) = 0$, nos modelos de regressão, garantimos que, pelo menos em média, estimamos o parâmetro verdadeiro.

Ao comparar diferentes estimadores não viesados, é, portanto, interessante saber qual deles tem a maior precisão possível.

O teorema de Gauss Markov nos diz que se um certo conjunto de suposições for atendido, a estimativa de mínimos quadrados ordinários para coeficientes de regressão fornece a mais baixa variância de amostragem dentro da classe dos estimadores lineares não enviesados (BLUE, do inglês Best Linear Unbiased Estimate) possível.

SUPOSIÇÕES DE GAUSS-MARKOV

1. **Linearidade:** os parâmetros que estimamos usando o método OLS devem ser lineares.
2. **Aleatoriedade:** nossos dados devem ter sido amostrados aleatoriamente na população.
3. **Distúrbio com média nula:** valor esperado do termo erro é zero para todas as observações $E\left(\epsilon^{(i)}\right) = 0$
4. **Distúrbios com covariância nula:** Cada termo de erro é independentemente distribuído e não correlacionado $Cov\left(\epsilon^{(i)}, \epsilon^{(j)}\right) = 0, i \neq j$
5. **Exogeneidade:** os regressores $\boldsymbol{x}^{(i)}$ não são correlacionados com o termo de perturbação $Cov\left(\boldsymbol{x}^{(i)}, \epsilon^{(i)}\right) = 0$
6. **Homocedasticidade:** a variância do distúrbio é constante em $\boldsymbol{x}$ e no tempo. $\sigma\left(\epsilon^{(i)}\right) = \left(\epsilon^{(i)^2}\right) = \sigma_\epsilon^2 = constante$

MODELOS LINEARES

Considere o conjunto de m dados de treinamento que compreende as variáveis independentes $\boldsymbol{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \cdots, x_n^{(i)}\right)$ e as variáveis-alvo correspondentes $y^{(i)}$.
Se assumirmos uma relação linear,

$$y^{(i)} = \omega_0 + \omega_1 x_1^{(i)} x + \omega_2 x_2^{(i)} + \omega_3 x_3^{(i)} + \cdots + \omega_n x_n^{(i)} + \epsilon = \boldsymbol{x}^{(i)T} \boldsymbol{\omega} + \epsilon$$

onde $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Isso significa essencialmente que nossos dados têm uma relação linear que é corrompida pelo ruído gaussiano aleatório que tem média zero e variação constante.

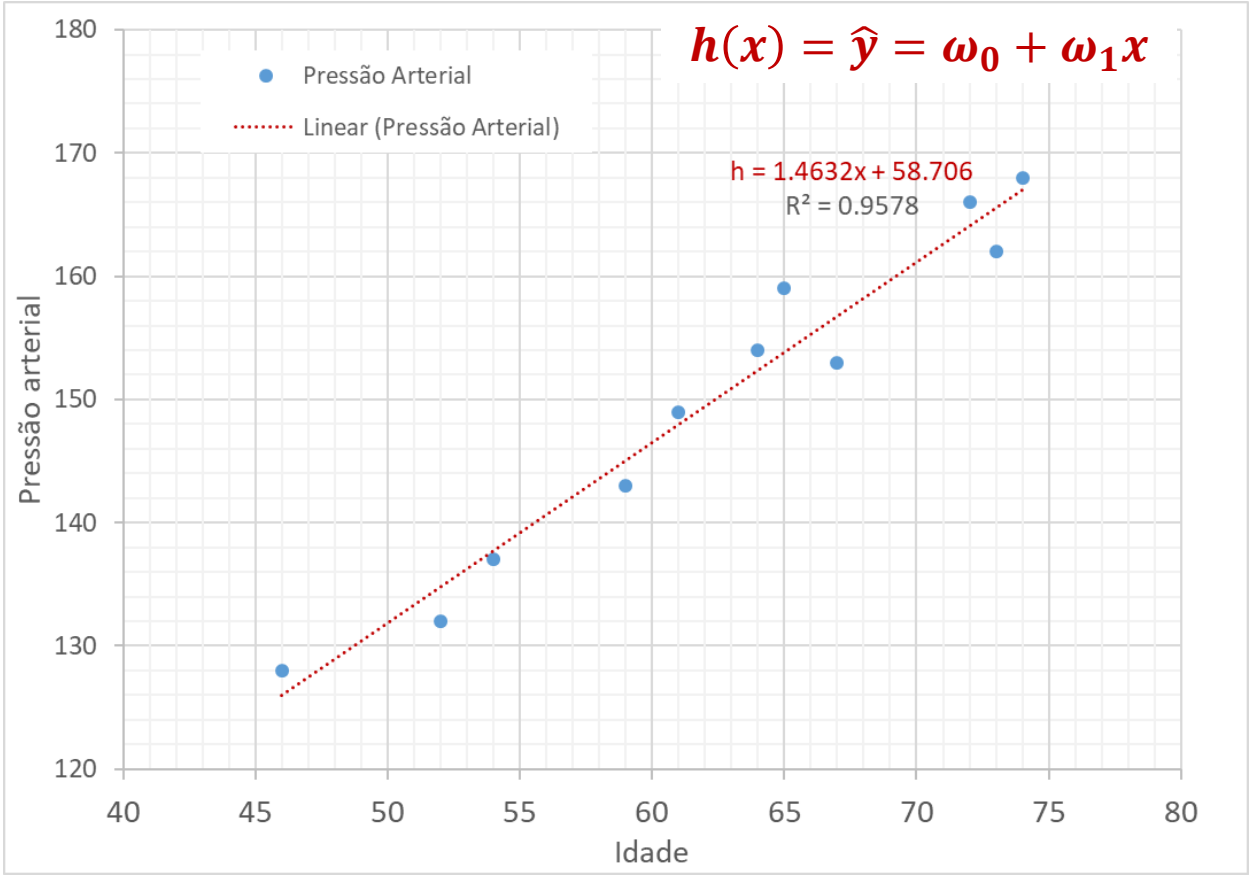
Isso tem a implicação de que $y^{(i)}$ é uma variável aleatória gaussiana e podemos calcular sua expectativa e variação:

$$E[y^{(i)}] = E[x^{(i)T}\omega + \epsilon^{(i)}] = x^{(i)T}\omega$$
$$Var[y^{(i)}] = Var[x^{(i)T}\omega + \epsilon^{(i)}] = \sigma^2$$

REGRESSÃO LINEAR SIMPLES

Tenho $m = 11$ dados de treinamento (•), onde x é a variável única de entrada $(n = 1)$ e y é a variável de saída.

m	Idade (variável de entrada)	Pressão Arterial (variável de saída)
1	52	132
2	59	143
3	67	153
4	73	162
5	64	154
6	74	168
7	54	137
8	61	149
9	65	159
10	46	128
11	72	166



PROBLEMA LINEAR SIMPLES

$\boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \end{bmatrix}$ ω_0, ω_1 arranjados em um vetor $\boldsymbol{\omega}$

HIPÓTESE: $h(x) = \hat{y} = \omega_0 + \omega_1 x = \boldsymbol{\omega}^T \boldsymbol{x}$ → $\boldsymbol{x} = \begin{bmatrix} x_0 = 1 \\ x_1 \end{bmatrix}$ $x_0 = 1, x_1$ arranjados em um vetor $\boldsymbol{x}$

FUNÇÃO CUSTO: $J(\omega_0, \omega_1) = \frac{1}{2m} \sum_{i=1}^{m} \left[e^{(i)} \right]^2, e^{(i)} = \hat{y}^{(i)} - y^{(i)}$

GOL: $\min_{(\omega_0, \omega_1)} J(\omega_0, \omega_1)$

"A problem well stated is a problem half solved."

Charles F. Kettering

ESTIMADORES DE MÍNIMOS QUADRADOS ORDINÁRIOS

$$\frac{\partial EQT}{\partial w_0} = 2\sum_i \left(y^{(i)} - w_0 - w_1 x^{(i)}\right)(-1) = 0$$

$$\frac{\partial EQT}{\partial w_1} = 2\sum_i \left(y^{(i)} - w_0 - w_1 x^{(i)}\right)(-x^{(i)}) = 0$$

$$w_1 = \frac{s_{xy}}{s_{xx}}$$

$$w_0 = \bar{y} - w_1 \bar{x}$$

$$\bar{x} = \frac{1}{n}\sum_i x^{(i)}$$

$$\bar{y} = \frac{1}{n}\sum_i y^{(i)}$$

$$s_{xx} = \sum_i (x^{(i)} - \bar{x})^2$$

$$s_{yy} = \sum_i (y^{(i)} - \bar{y})^2$$

$$s_{xy} = \sum_i (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$$

COEFICIENTE DE CORRELAÇÃO DE PEARSON r

$$\sum_i \left[\epsilon^{(i)}\right]^2 = s_{yy}\left(1 - \frac{s_{xy}^2}{s_{xx}s_{yy}}\right)$$

$$r = \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}}$$

O *coeficiente de correlação de Pearson* varia de -1 a 1 e mede o quão bem a variável dependente pode ser explicada por um modelo linear da variável dependente. Valores mais próximos de zero significam um modelo linear menos explicativo. Valores mais próximos de 1 ou -1 significam um modelo linear mais explicativo.

ALGORITMO DE GRADIENTE DESCENDENTE

Atribua um valor inicial, $\boldsymbol{\omega}^{(0)}$ para o vetor de parâmetros → $\boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \end{bmatrix}$ ω_k arranjados em um vetor $\boldsymbol{\omega}$

Atribua um valor arbitrariamente pequeno para uma constante $\varepsilon > 0$ $(1e^{-4}?)$,

Defina α e $i = 0$

Calcule $\nabla J\left(\boldsymbol{\omega}^{(0)}\right)$ → $\nabla J(\boldsymbol{\omega}) = \begin{bmatrix} \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_0} & \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_1} \end{bmatrix}$

Enquanto $\left\|\nabla J\left(\boldsymbol{\omega}^{(0)}\right)\right\| > \varepsilon$:

$\omega_k^{(i+1)} = \omega_k^{(i)} - \alpha \nabla J_k\left(\boldsymbol{\omega}^{(i)}\right), \qquad k = 0,1$

$i \mathrel{+}= 1$

ALGORITMO

repetir até convergência{

$$\omega_k^{(i+1)} = \omega_k^{(i)} - \alpha \frac{\partial}{\partial \omega_k} J\left(\boldsymbol{\omega}^{(i)}\right)$$

}

Update simultâneo:

$$aux_0 = \omega_0 - \alpha \frac{\partial}{\partial \omega_0} J(\omega_0, \omega_1)$$

$$aux_1 = \omega_1 - \alpha \frac{\partial}{\partial \omega_1} J(\omega_0, \omega_1)$$

$$\omega_0, \omega_1 = aux_0, aux_1$$

$\nabla J(\boldsymbol{\omega})$

$$J(\omega_0, \omega_1) = J(\boldsymbol{\omega}) = \frac{1}{2m}\sum_{i=1}^{m}\left[\left(\omega_0 + \omega_1 x^{(i)}\right) - y^{(i)}\right]^2 \quad \rightarrow \quad \nabla J_k = \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_k} = ?$$

$$j = 0 \quad \rightarrow \quad \frac{\partial J(\omega_0, \omega_1)}{\partial \omega_0} = \frac{1}{m}\sum_{i=1}^{m}\left[\left(\omega_0 + \omega_1 x^{(i)}\right) - y^{(i)}\right] = \frac{1}{m}\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - y^{(i)}\right]$$

$$j = 1 \quad \rightarrow \quad \frac{\partial J(\omega_0, \omega_1)}{\partial \omega_1} = \frac{1}{m}\sum_{i=1}^{m}\left[\left(\omega_0 + \omega_1 x^{(i)}\right) - y^{(i)}\right] x^{(i)} = \frac{1}{m}\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - y^{(i)}\right] x^{(i)}$$

ALGORITMO DE GRADIENTE DESCENDENTE

Repita até convergência {

$$\omega_0^{(i+1)} = \omega_0^{(i)} - \alpha \frac{1}{m} \sum_{i=1}^{m} \left[h\left(x^{(i)}\right) - y^{(i)} \right]$$

$$\omega_1^{(i+1)} = \omega_1^{(i)} - \alpha \frac{1}{m} \sum_{i=1}^{m} \left[h\left(x^{(i)}\right) - y^{(i)} \right] x^{(i)}$$

}

"Batch": todos os exemplos de treinamento são usados em cada passo do gradiente descendente

Update ω_0, ω_1 simultaneamente!

REGRESSÃO LINEAR MÚLTIPLA

HIPÓTESE: $h(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) = \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n = \boldsymbol{\omega}^T \boldsymbol{x}$

FUNÇÃO CUSTO: $J(\boldsymbol{\omega}) = \frac{1}{2m} \sum_{i=1}^{m} \left[h\left(x^{(i)}\right) - y^{(i)} \right]^2$

GOL: $\min_{(\boldsymbol{\omega})} J(\boldsymbol{\omega})$

$$\boldsymbol{x} = \begin{bmatrix} x_0 = 1 \\ x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$$

x_k arranjados em um vetor $\boldsymbol{x}$

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \cdots \\ \omega_n \end{bmatrix}$$

ω_k arranjados em um vetor $\boldsymbol{\omega}$

m	Idade	Peso (Kg)	Pressão Arterial
1	52	78.5	132
2	59	83.5	143
3	67	88.0	153
4	73	95.7	162
5	64	88.9	154
6	74	99.8	168
7	54	85.3	137
8	61	85.3	149
9	65	93.9	159
10	46	75.7	128
11	72	98.4	166

m: número de dados de entrada (11)
n: número de features (01: idade, 02:peso)

$x_j^{(i)}$ é o valor da *feature* j do i-ésimo conjunto de entradas

$$x_2^{(7)} = 85.3$$

$$x_1^{(4)} = 73$$

REGRESSÃO LINEAR MÚLTIPLA...

$$\arg\min_{\mathbf{w}} \| \mathbf{e} \|^2 = \| \mathbf{y} - \mathbf{X}\mathbf{w} \|_2^2$$

$$\begin{aligned}\mathcal{E} &= \mathbf{y}^T\mathbf{y} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} \\ &= \mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}\end{aligned}$$

$$\frac{\partial \mathcal{E}}{\partial \mathbf{w}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\mathbf{w} = 0$$

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$$

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

PORQUE USAR GD AO INVÉS DA SOLUÇÃO ANALÍTICA $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$???

- você está considerando possíveis mudanças ou generalizações do modelo, adicionando termos mais complexos ou algum método de regularização. Enfim, você precisa de um método mais genérico porque você não sabe muito sobre o futuro do código e do modelo;
- solução analítica é mais cara computacionalmente, e você precisa de eficiência;
- solução analítica requer mais memória, o que você não tem;
- solução analítica é difícil de implementar e você precisa de um código simples e fácil!

ALGORITMO DE GRADIENTE DESCENDENTE

Atribua um valor inicial, $\boldsymbol{\omega}^{(0)}$ para o vetor de parâmetros

Atribua um valor arbitrariamente pequeno para uma constante $\varepsilon > 0$ $(1e^{-4}?)$,

Defina α e $i = 0$

Calcule $\nabla J\left(\boldsymbol{\omega}^{(0)}\right)$

Enquanto $\left\|\nabla J\left(\boldsymbol{\omega}^{(0)}\right)\right\| > \varepsilon$:

$$\omega_k^{(i+1)} = \omega_k^{(i)} - \alpha \nabla J_k\left(\boldsymbol{\omega}^{(i)}\right), \qquad j = 0,1,\dots,n$$

$$i \mathrel{+}= 1$$

$\nabla J(\boldsymbol{\omega})$

$$J(\boldsymbol{\omega}) = \frac{1}{2m}\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - y^{(i)}\right]^2 \quad \rightarrow \quad \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_j} = ?$$

$$j = 0 \quad \rightarrow \quad \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_0} = \frac{1}{m}\sum_{i=1}^{m}[h(x^{(i)}) - y^{(i)}]$$

$$j \neq 0 \quad \rightarrow \quad \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_j} = \frac{1}{m}\sum_{i=1}^{m}[h(x^{(i)}) - y^{(i)}]\, x_j^{(i)}$$

Vamos ao Notebook

$m = 6, n = 2$

x_1	x_2	y
4	1	2
2	8	−14
1	0	1
3	2	−1
1	4	−7
6	7	−8

$$\omega_0^{(0)} = 0, \omega_1^{(0)} = -0.017, \omega_2^{(0)} = -0.048$$

$$h = \hat{y} = 0x_0 - 0.017x_1 - 0.048x_2$$

$$h = \hat{y} = [0 \quad -0.017 \quad -0.048] \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_0} = \frac{1}{m}\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - y^{(i)}\right] = \frac{1}{6}\left\{\left[[0 \quad -0.017 \quad -0.048]\begin{bmatrix}1\\4\\1\end{bmatrix} - 2\right] + \left[[0 \quad -0.017 \quad -0.048]\begin{bmatrix}1\\2\\8\end{bmatrix} - (-14)\right] + \cdots + \left[[0 \quad -0.017 \quad -0.048]\begin{bmatrix}1\\6\\7\end{bmatrix} - (-8)\right]\right\}$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_1} = \frac{1}{m}\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - y^{(i)}\right]x_1^{(i)} = \frac{1}{6}\left\{\left[[0 \quad -0.017 \quad -0.048]\begin{bmatrix}1\\4\\1\end{bmatrix} - 2\right]4 + \left[[0 \quad -0.017 \quad -0.048]\begin{bmatrix}1\\2\\8\end{bmatrix} - (-14)\right]2 + \cdots + \left[[0 \quad -0.017 \quad -0.048]\begin{bmatrix}1\\6\\7\end{bmatrix} - (-8)\right]6\right\}$$

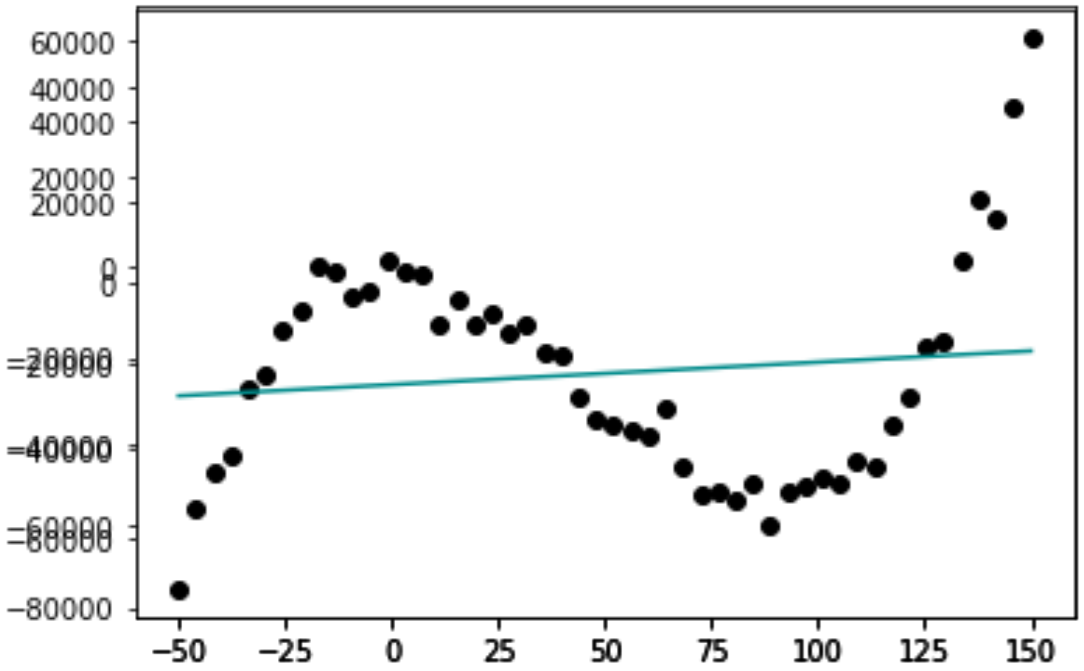
$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_2} = \frac{1}{m}\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - y^{(i)}\right]x_2^{(i)} = \frac{1}{6}\left\{\left[[0 \quad -0.017 \quad -0.048]\begin{bmatrix}1\\4\\1\end{bmatrix} - 2\right]1 + \left[[0 \quad -0.017 \quad -0.048]\begin{bmatrix}1\\2\\8\end{bmatrix} - (-14)\right]8 + \cdots + \left[[0 \quad -0.017 \quad -0.048]\begin{bmatrix}1\\6\\7\end{bmatrix} - (-8)\right]7\right\}$$

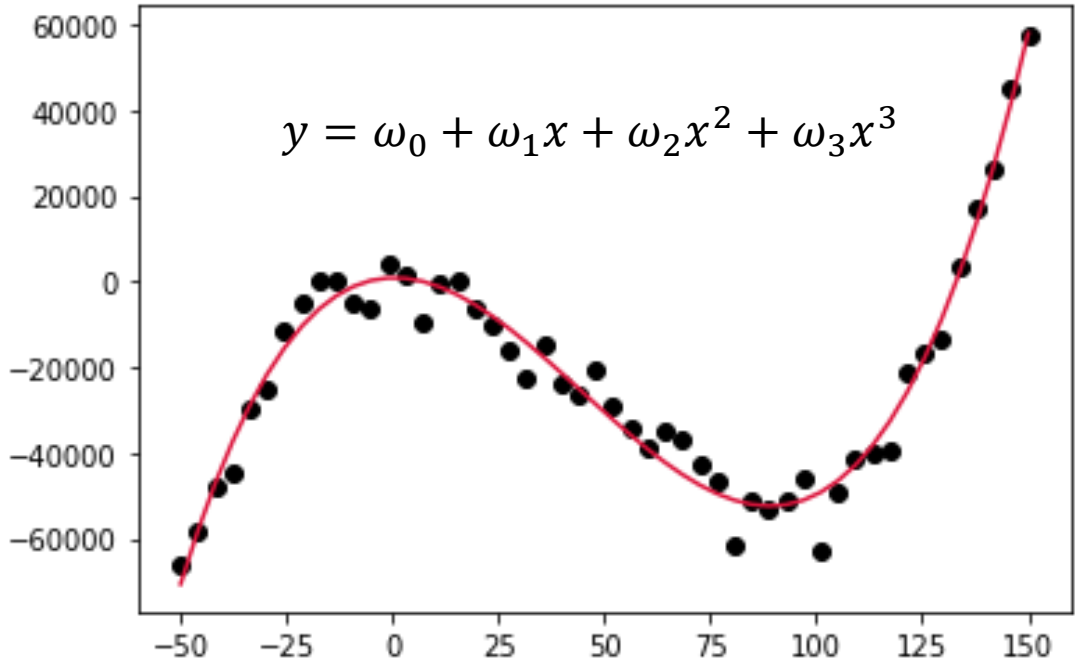
$$\omega_0^{(1)} = \omega_0^{(0)} - \alpha\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_0}, \omega_1^{(1)} = \omega_1^{(0)} - \alpha\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_1}, \omega_2^{(1)} = \omega_2^{(0)} - \alpha\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_2}$$

REGRESSÃO POLINOMIAL

REGRESSÃO POLINOMIAL

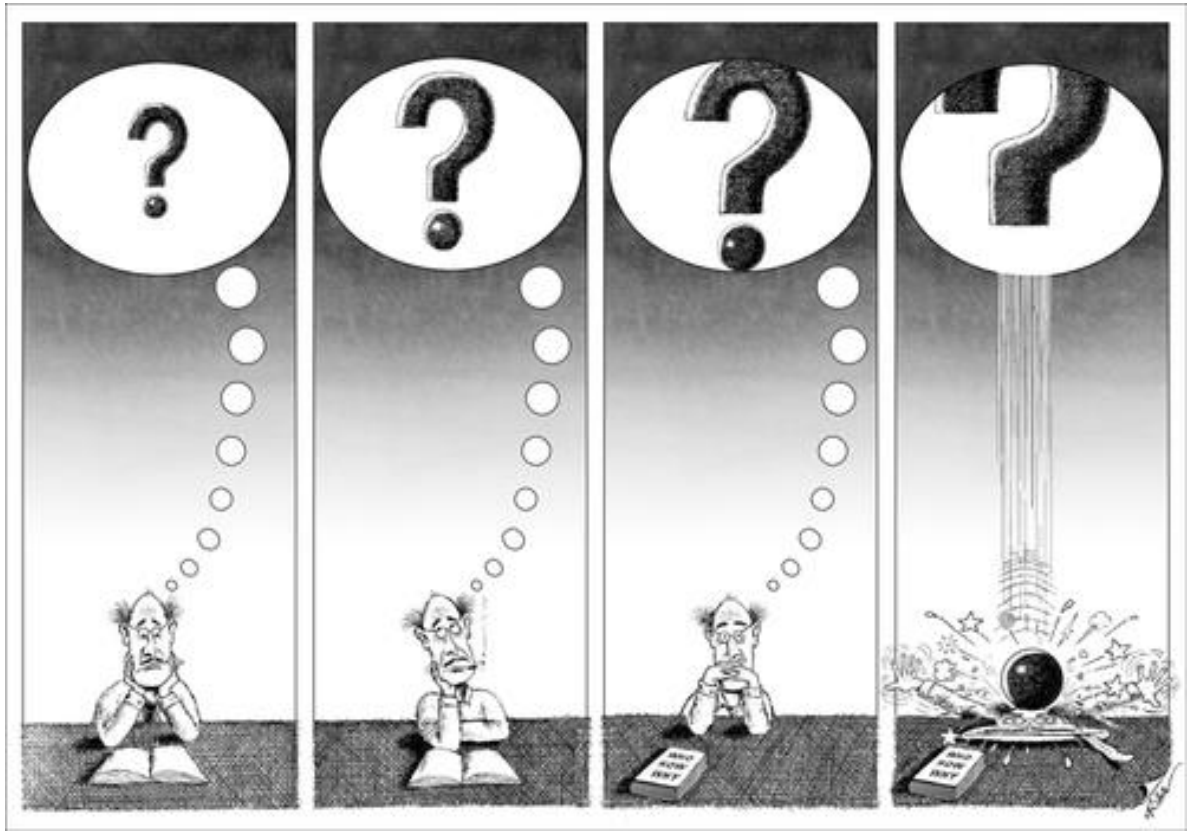
$$y = \omega_0 + \omega_1 x + \omega_2 x^2 + \omega_3 x^3 + \cdots + \omega_n x^n + \varepsilon$$





Regressão polinomial é uma forma de regressão linear na qual a relação entre a variável independente x e a variável dependente y é modelada como um polinômio de n-ésimo grau. A regressão polinomial modela uma relação não linear entre o valor de x e a média condicional correspondente de y, $E(y|x)$.

Regressão **linear** polinomial ???



Regressão Linear Simples

$$\hat{y} = \omega_0 + \omega_1 x$$

LINEARIDADE ESTÁ NOS PESOS ω
x SÃO CONSTANTES, DADOS DE ENTRADA.

Regressão Linear Múltipla

$$\hat{y} = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \cdots + \omega_n x_n$$

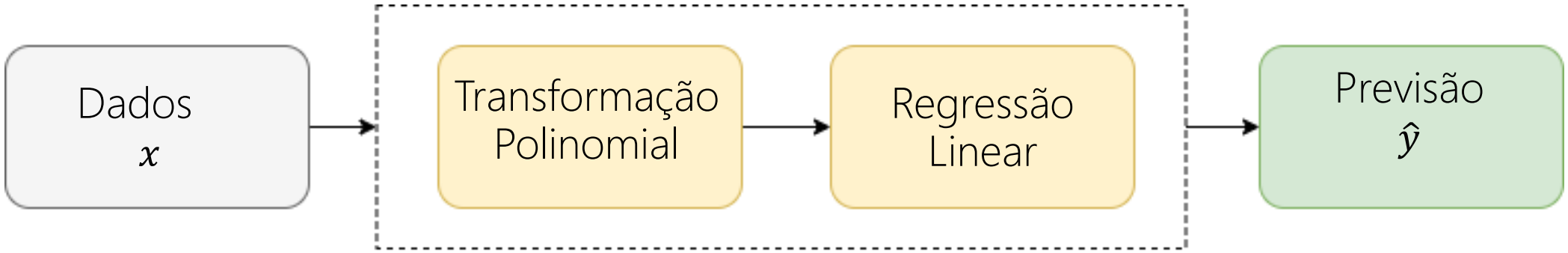
Regressão Linear Polinomial

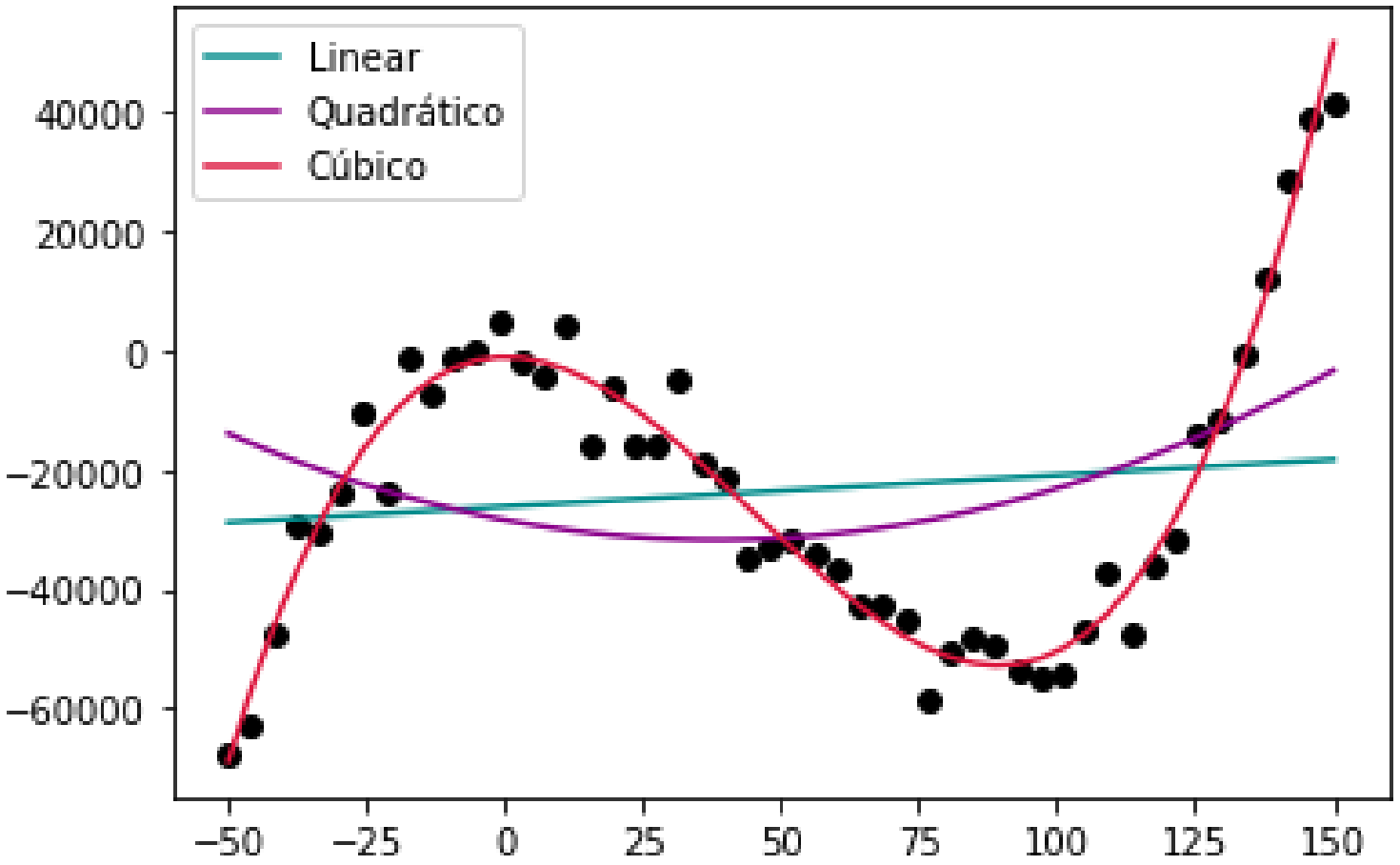
$$\hat{y} = \omega_0 + \omega_1 x + \omega_2 x^2 + \omega_3 x^3 + \cdots + \omega_n x^n$$

$$h(\boldsymbol{x}) = \omega_0 + \omega_1 x + \omega_2 x^2 + \omega_3 x^3 + \cdots + \omega_n x^n + \varepsilon$$

$$h(\boldsymbol{x}) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + \cdots + \omega_n x_n$$

Basta que supor que $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$ e assim por diante até$x_n = x^n$. Inteligente, certo? Portanto, podemos realmente usar modelos de regressão linear para executar regressão polinomial ...





Esse modelo ainda é considerado linear, pois pesos associados às features ainda são lineares. x^2 e x^3 são apenas *features*. No entanto, as curvas que estamos ajustando são de natureza quadrática e cúbica.

PROBLEMA DEFINIDO

HIPÓTESE: $h(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) = \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n = \boldsymbol{\omega}^T \boldsymbol{x}$

FUNÇÃO CUSTO: $J(\boldsymbol{\omega}) = \frac{1}{2m} \sum_{i=1}^{m} \left[h\left(x^{(i)}\right) - y^{(i)} \right]^2$

GOL: $\min_{(\boldsymbol{\omega})} J(\boldsymbol{\omega})$

$$\boldsymbol{x} = \begin{bmatrix} x_0 = 1 \\ x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}$$

x_k arranjados em um vetor $\boldsymbol{x}$

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \cdots \\ \omega_n \end{bmatrix}$$

ω_k arranjados em um vetor $\boldsymbol{\omega}$

O MÉTODO DOS MÍNIMOS QUADRADOS NO AJUSTE DE UM MODELO POLINOMIAL

$$\hat{y}^{(i)} = \omega_0 + \omega_1 x^{(i)} + \omega_2 \left(x^{(i)}\right)^2 + \omega_3 \left(x^{(i)}\right)^3 + \cdots + \omega_n \left(x^{(i)}\right)^n$$

$$J(\boldsymbol{\omega}) \approx \sum_{i=1}^{m} \left[\omega_0 + \omega_1 x^{(i)} + \omega_2 \left(x^{(i)}\right)^2 + \omega_3 \left(x^{(i)}\right)^3 + \cdots + \omega_n \left(x^{(i)}\right)^n - y^{(i)}\right]^2$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_k} = 0$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_0} = m\omega_0 + \omega_1 \sum_{i=1}^{m} x^{(i)} + \omega_2 \sum_{i=1}^{m} \left(x^{(i)}\right)^2 + \cdots + \omega_n \sum_{i=1}^{m} \left(x^{(i)}\right)^n - \sum_{i=1}^{m} y^{(i)} = 0$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_1} = \omega_0 \sum_{i=1}^{m} x^{(i)} + \omega_1 \sum_{i=1}^{m} \left(x^{(i)}\right)^2 + \omega_2 \sum_{i=1}^{m} \left(x^{(i)}\right)^3 + \cdots + \omega_n \sum_{i=1}^{m} \left(x^{(i)}\right)^{n+1} - \sum_{i=1}^{m} x^{(i)} y^{(i)} = 0$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_n} = \omega_0 \sum_{i=1}^{m} \left(x^{(i)}\right)^n + \omega_1 \sum_{i=1}^{m} \left(x^{(i)}\right)^{n+1} + \omega_2 \sum_{i=1}^{m} \left(x^{(i)}\right)^{n+2} + \cdots + \omega_n \sum_{i=1}^{m} \left(x^{(i)}\right)^{2n} - \sum_{i=1}^{m} \left(x^{(i)}\right)^n y^{(i)} = 0$$

FINALMENTE...

$$\begin{bmatrix} m & \sum_{i=1}^{m} x^{(i)} & \cdots & \sum_{i=1}^{m} \left(x^{(i)}\right)^{n} \\ \sum_{i=1}^{m} x^{(i)} & \sum_{i=1}^{m} \left(x^{(i)}\right)^{2} & \cdots & \sum_{i=1}^{m} \left(x^{(i)}\right)^{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{m} \left(x^{(i)}\right)^{n} & \sum_{i=1}^{m} \left(x^{(i)}\right)^{n+1} & \cdots & \sum_{i=1}^{m} \left(x^{(i)}\right)^{2n} \end{bmatrix} \begin{bmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} y^{(i)} \\ \sum_{i=1}^{m} x^{(i)} y^{(i)} \\ \vdots \\ \sum_{i=1}^{m} \left(x^{(i)}\right)^{n} y^{(i)} \end{bmatrix}$$

POR EXEMPLO...

$$\boldsymbol{x} = \begin{bmatrix} x_0 = 1 \\ x \\ x^2 \end{bmatrix}$$

HIPÓTESE: $h(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) = \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2$
$= \omega_0 x_0 + \omega_1 x + \omega_2 x^2 = \boldsymbol{\omega}^T \boldsymbol{x}$

$$\boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \end{bmatrix}$$

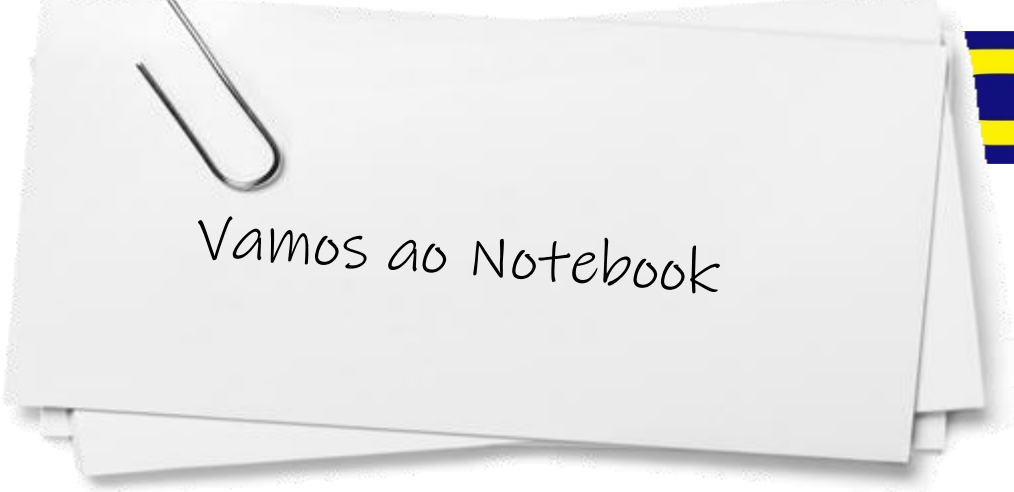
FUNÇÃO CUSTO:
$$J(\boldsymbol{\omega}) = \frac{1}{2m} \sum_{i=1}^{m} \left[\omega_0 x_0 + \omega_1 x^{(i)} + \omega_2 \left(x^{(i)}\right)^2 - y^{(i)} \right]^2$$

GRADIENTE:
$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_0} = \frac{1}{m} \sum_{i=1}^{m} \left[\omega_0 x_0 + \omega_1 x^{(i)} + \omega_2 \left(x^{(i)}\right)^2 - y^{(i)} \right]$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_1} = \frac{1}{m} \sum_{i=1}^{m} \left[\omega_0 x_0 + \omega_1 x^{(i)} + \omega_2 \left(x^{(i)}\right)^2 - y^{(i)} \right] x^{(i)}$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_2} = \frac{1}{m} \sum_{i=1}^{m} \left[\omega_0 x_0 + \omega_1 x^{(i)} + \omega_2 \left(x^{(i)}\right)^2 - y^{(i)} \right] \left(x^{(i)}\right)^2$$

POR EXEMPLO....



$$h(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) = \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2$$
$$= \omega_0 x_0 + \omega_1 x + \omega_2 x^2 = \boldsymbol{\omega}^T \boldsymbol{x}$$

$$J(\boldsymbol{\omega}) = \frac{1}{2m} \sum_{i=1}^{m} \left[\omega_0 x_0 + \omega_1 x^{(i)} + \omega_2 \left(x^{(i)}\right)^2 - y^{(i)} \right]^2$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_0} = \frac{1}{m} \sum_{i=1}^{m} \left[\omega_0 x_0 + \omega_1 x^{(i)} + \omega_2 \left(x^{(i)}\right)^2 - y^{(i)} \right] = m\omega_0 + \omega_1 \sum_{i=1}^{m} x^{(i)} + \omega_2 \sum_{i=1}^{m} \left(x^{(i)}\right)^2 + \sum_{i=1}^{m} y^{(i)}$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_1} = \frac{1}{m} \sum_{i=1}^{m} \left[\omega_0 x_0 + \omega_1 x^{(i)} + \omega_2 \left(x^{(i)}\right)^2 - y^{(i)} \right] x^{(i)} = \omega_0 \sum_{i=1}^{m} x^{(i)} + \omega_1 \sum_{i=1}^{m} \left(x^{(i)}\right)^2 + \omega_2 \sum_{i=1}^{m} \left(x^{(i)}\right)^3 + \sum_{i=1}^{m} x^{(i)} y^{(i)}$$

$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_2} = \frac{1}{m} \sum_{i=1}^{m} \left[\omega_0 x_0 + \omega_1 x^{(i)} + \omega_2 \left(x^{(i)}\right)^2 - y^{(i)} \right] \left(x^{(i)}\right)^2 = \omega_0 \sum_{i=1}^{m} \left(x^{(i)}\right)^2 + \omega_1 \sum_{i=1}^{m} \left(x^{(i)}\right)^3 + \omega_2 \sum_{i=1}^{m} \left(x^{(i)}\right)^4 + \sum_{i=1}^{m} \left(x^{(i)}\right)^2 y^{(i)}$$

ALGORITMO DE GRADIENTE DESCENDENTE

Atribua um valor inicial, $\boldsymbol{\omega}^{(0)}$ para o vetor de parâmetros

Atribua um valor arbitrariamente pequeno para uma constante $\varepsilon > 0$ $(1e^{-4}?)$,

Defina α e $i = 0$

Calcule $\nabla J\left(\boldsymbol{\omega}^{(0)}\right)$

Enquanto $\left\|\nabla J\left(\boldsymbol{\omega}^{(0)}\right)\right\| > \varepsilon$:

$$\omega_k^{(i+1)} = \omega_k^{(i)} - \alpha \nabla J_k\left(\boldsymbol{\omega}^{(i)}\right), \qquad k = 0,1,\ldots,n$$

$$i \mathrel{+}= 1$$

$\nabla J(\boldsymbol{\omega})$

$$J(\boldsymbol{\omega}) = \frac{1}{2m}\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - y^{(i)}\right]^2 \quad \rightarrow \quad \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_k} = ?$$

$$k = 0 \quad \rightarrow \quad \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_0} = \frac{1}{m}\sum_{i=1}^{m}[h(x^{(i)}) - y^{(i)}]$$

$$k \neq 0 \quad \rightarrow \quad \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_k} = \frac{1}{m}\sum_{i=1}^{m}[h(x^{(i)}) - y^{(i)}]\, x_k^{(i)}$$

O QUE PODE DAR ERRADO?

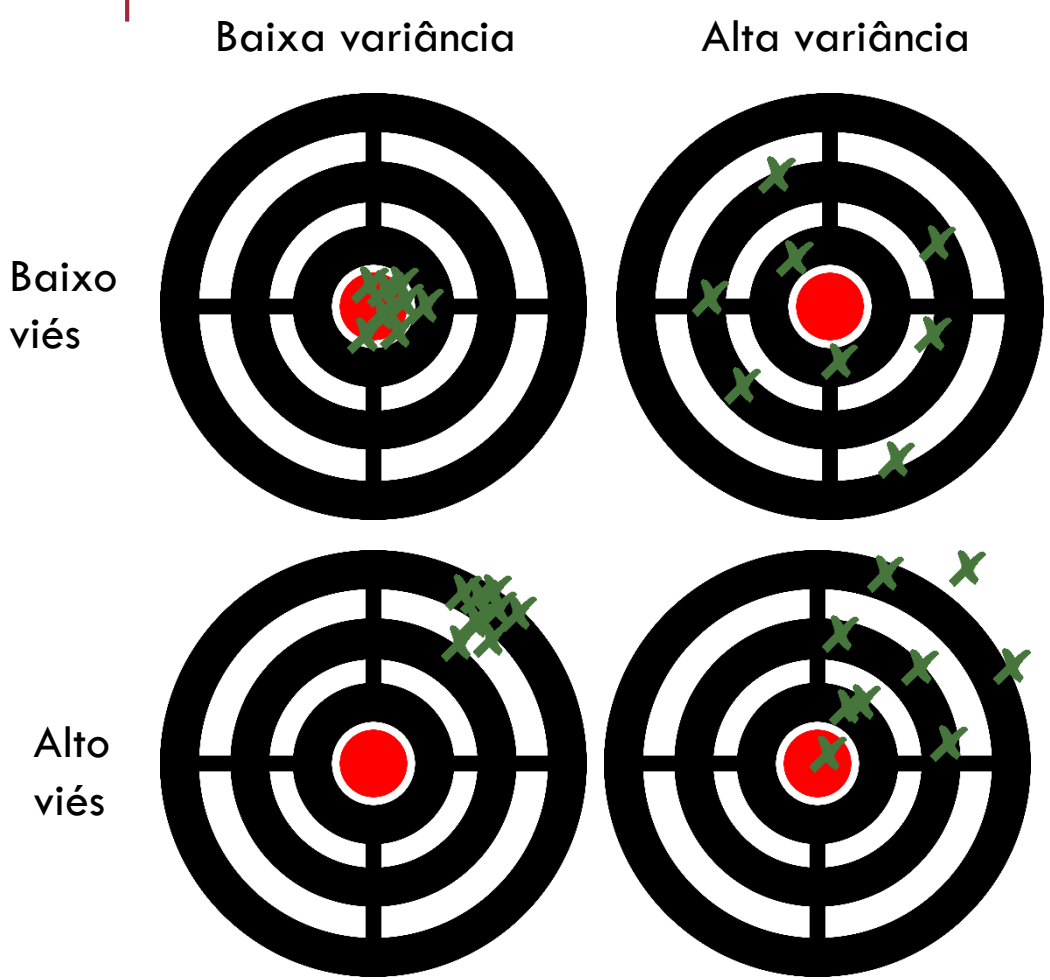
ERRO QUADRÁTICO MÉDIO

A decomposição em variância-viés do erro quadrático médio é extremamente importante se você deseja realmente compreender super ajuste (overfitting), sub ajuste (underfitting) e capacidade do modelo.

$$\mathbb{E}_{\boldsymbol{x},\varepsilon}\left\{\left(\breve{y}(\boldsymbol{x})+\epsilon-h(\boldsymbol{X},\boldsymbol{w})\right)^2\right\}$$

- Dataset e ruído
- Fenômeno
- Ruído
- Aprendido com $\boldsymbol{x}$

VIÉS E VARIÂNCIA



O **viés** representa o erro em relação do valor esperado da predição do modelo com o valor que gostaríamos de predizer.

$$Bias(\hat{y}) = \mathbb{E}[\hat{y} - y]$$

A **variância** mede a distância de cada estimativa pontual à sua média.

$$Var(\hat{y}) = \mathbb{E}[\hat{y}^2] - \mathbb{E}[\hat{y}]^2$$

O erro quadrático médio, que indica a qualidade de um estimador, é **a soma da variância e do quadrado do viés.** Ele mostra a variação total em torno de um valor verdadeiro,

$$EQM = \epsilon^2 + Var(\hat{y}) + \left(Bias(\hat{y})\right)^2$$

EM NOSSO PROBLEMA

Um modelo com ALTO VIÉS aprende relações erradas e gera previsões longe do esperado. O modelo não aprende corretamente com o conjunto de dados, assumindo informações sobre os dados que não são necessariamente corretas. Dessa forma, modelos com alto viés possuem um problema de underfitting.

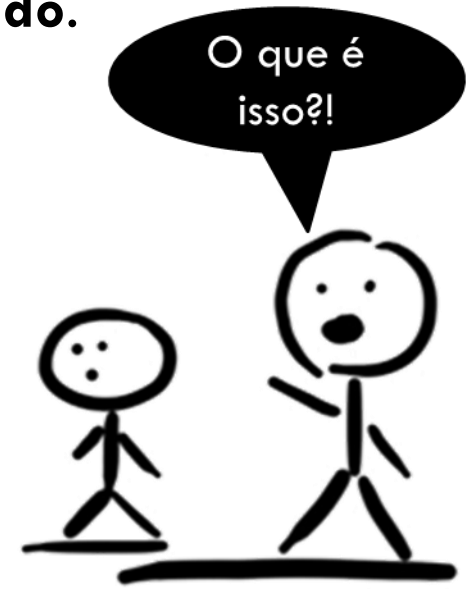
Modelos com alta ALTA VARIÂNCIA focam excessivamente se ajustar aos dados e, inclusive, ao ruído. Assim, esses modelos têm um problema de overfitting, ou seja, se adaptam tão bem ao conjunto de dados que não conseguem generalizar para além dele.

ERRO IRREDUTÍVEL refere-se a pontos fora da curva, exceções. Não é possível observar cada pequeno fator que leva um evento a acontecer, possuímos limitações. Ao elaborar um modelo escolhemos os fatores que são mais relevantes ao nosso problema e deixamos de lado alguns outros.

COMO DETECTAR?

Existem duas maneiras:

1. Se seu conjunto de dados for 2D ou 3D você pode visualizar em gráficos e tentar entender o que, possivelmente, está errado com sua implementação;
2. Traçar **a curva de aprendizado**.



Curvas de aprendizado: curvas de aprendizado calculadas na métrica pela qual os parâmetros do modelo estão sendo otimizados, por exemplo, custo.

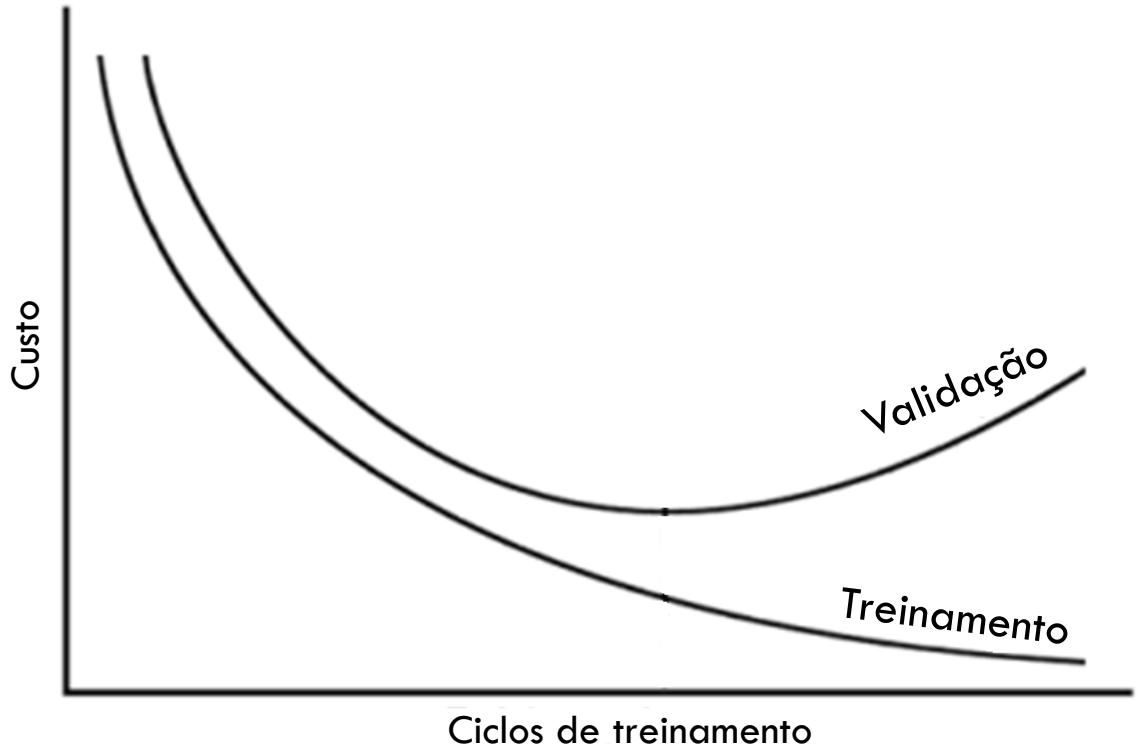
CURVA DE APRENDIZADO

Durante o treinamento de um modelo de aprendizado de máquina, o estado atual do modelo em cada etapa do algoritmo de treinamento pode ser avaliado.

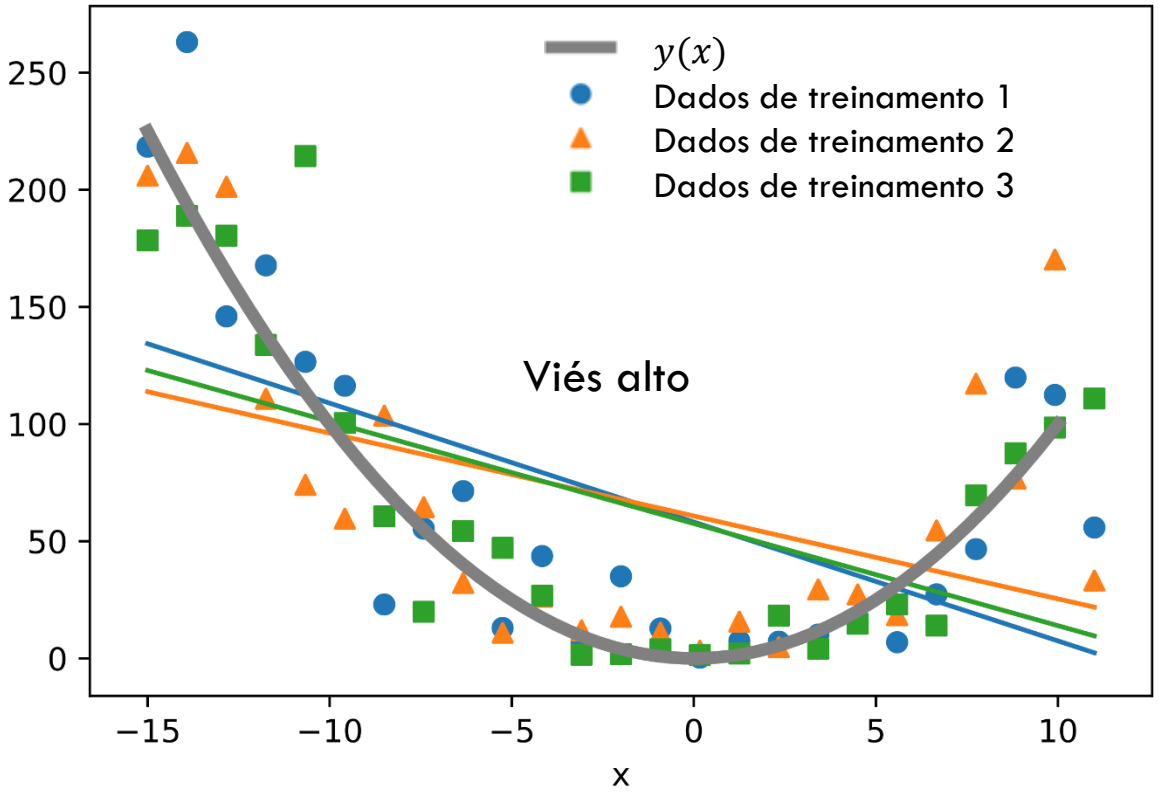
Curva de aprendizado de treinamento: Curva de aprendizado calculada a partir do conjunto de dados de treinamento que fornece uma ideia de quão bem o modelo está **aprendendo**.

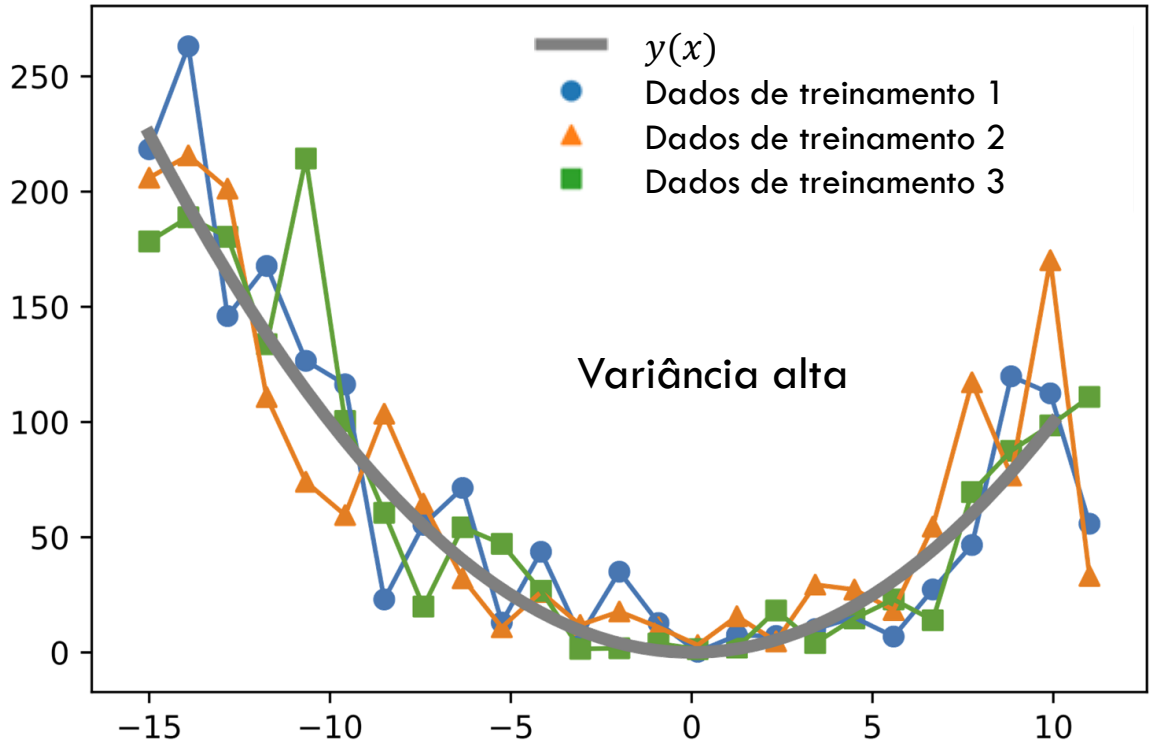
Curva de aprendizado de validação: Curva de aprendizado calculada a partir de um conjunto de dados de validação que fornece uma ideia de quão bem o modelo está **generalizando**.

É comum criar as duas curvas de aprendizado para um modelo durante o treinamento.



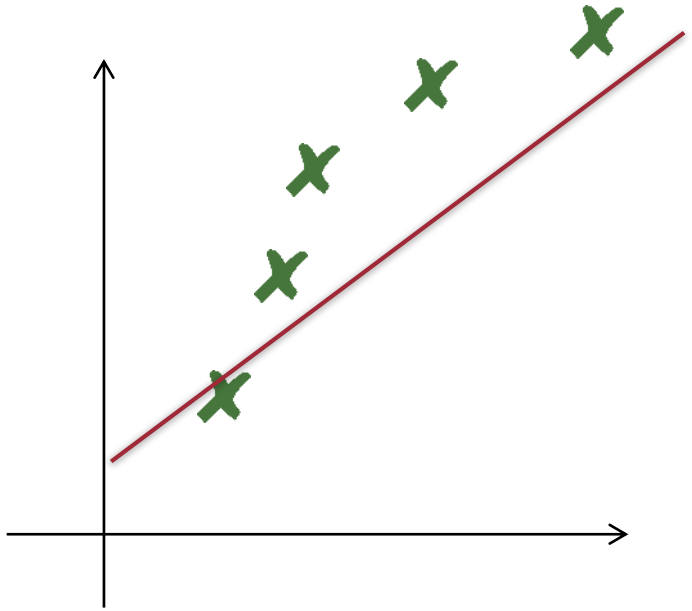
EXEMPLO EM DADOS 2D





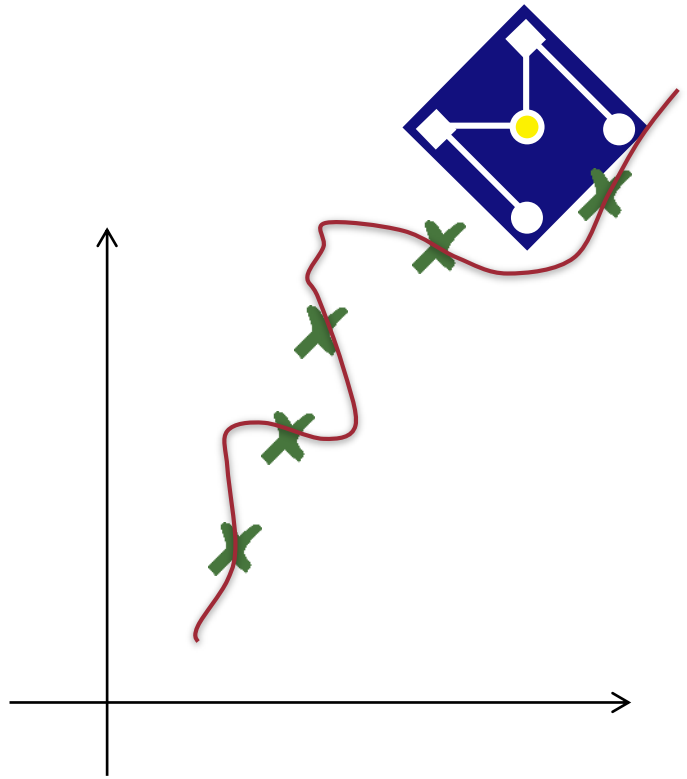
http://rasbt.github.io/mlxtend/user_guide/evaluate/bias_variance_decomp/

Existe um “trade-off”, ou seja, um balanço, entre viés e variância, de forma que, quando se aumenta a complexidade de um modelo, o quadrado de seu viés tem seu valor diminuído, enquanto a variância tem seu valor aumentado.



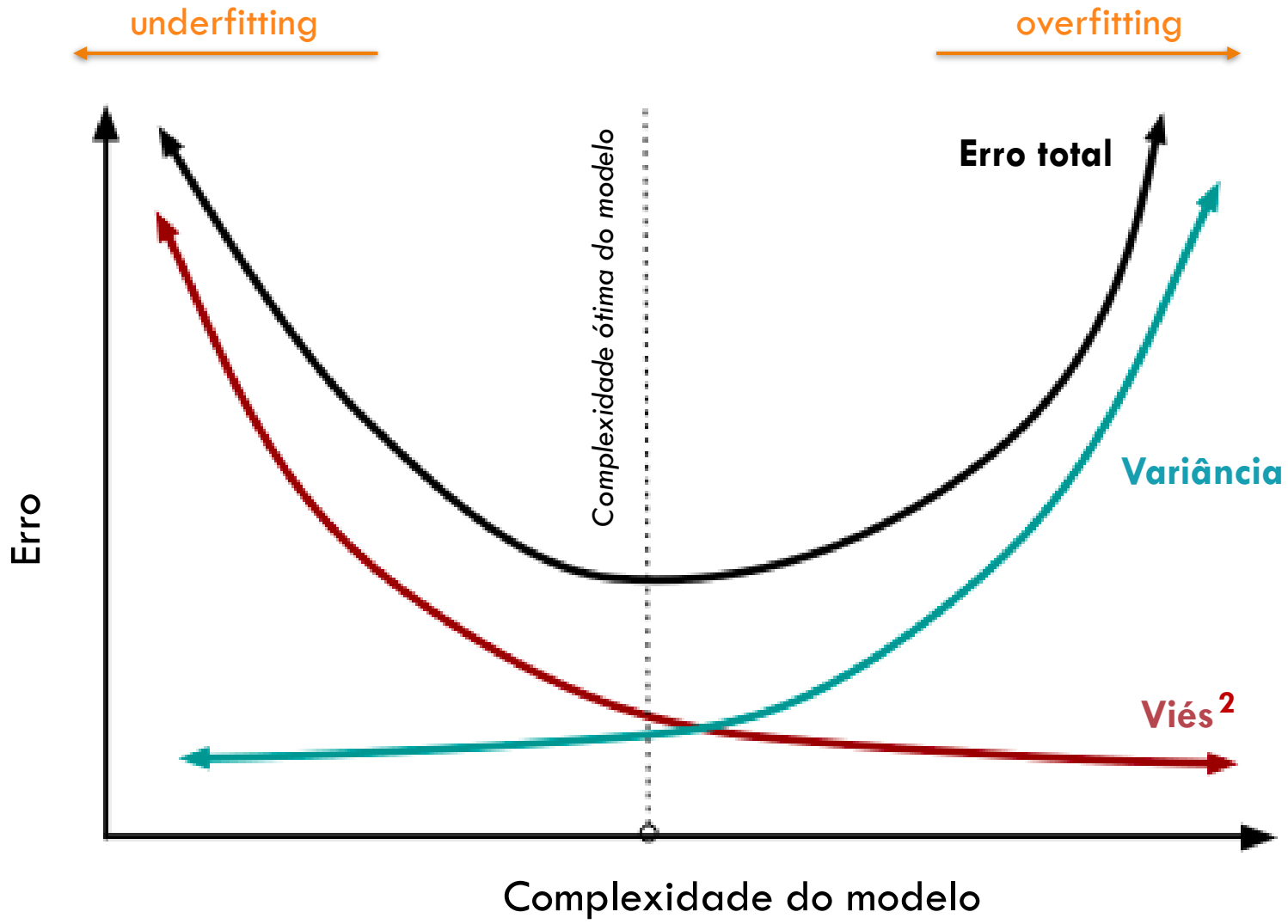
Alto viés (bias): incapacidade do modelo de capturar a verdadeira relação entre variáveis e o objeto a ser predito.

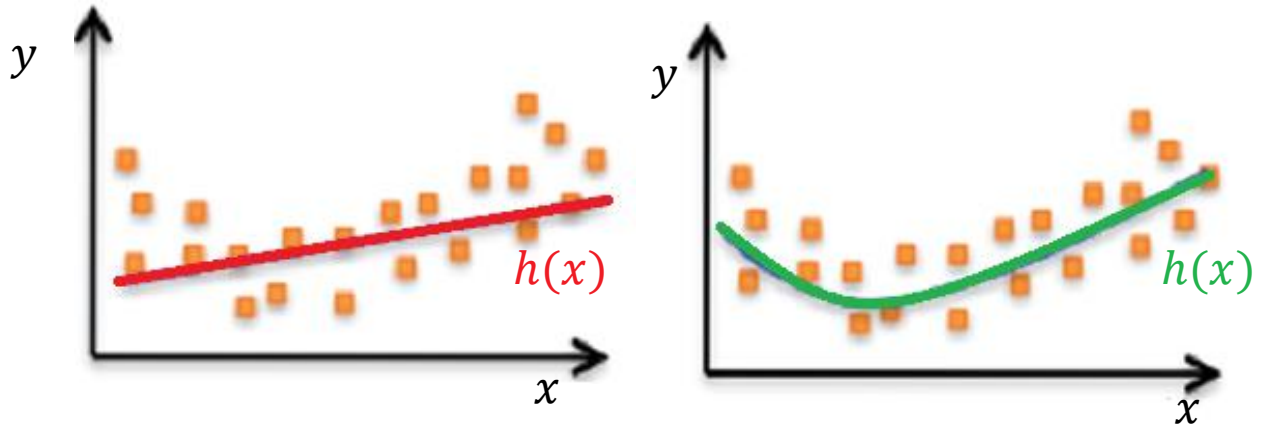
UNDERFITTING

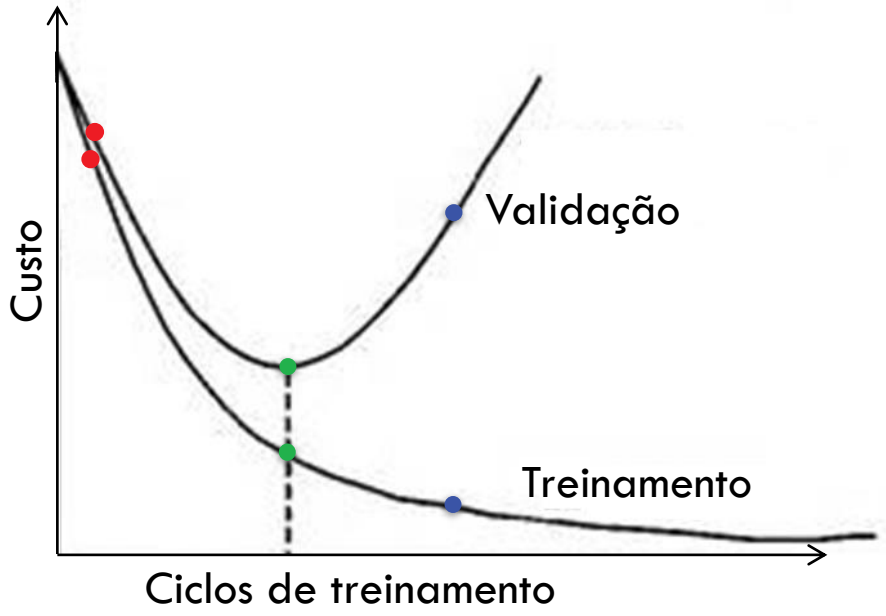


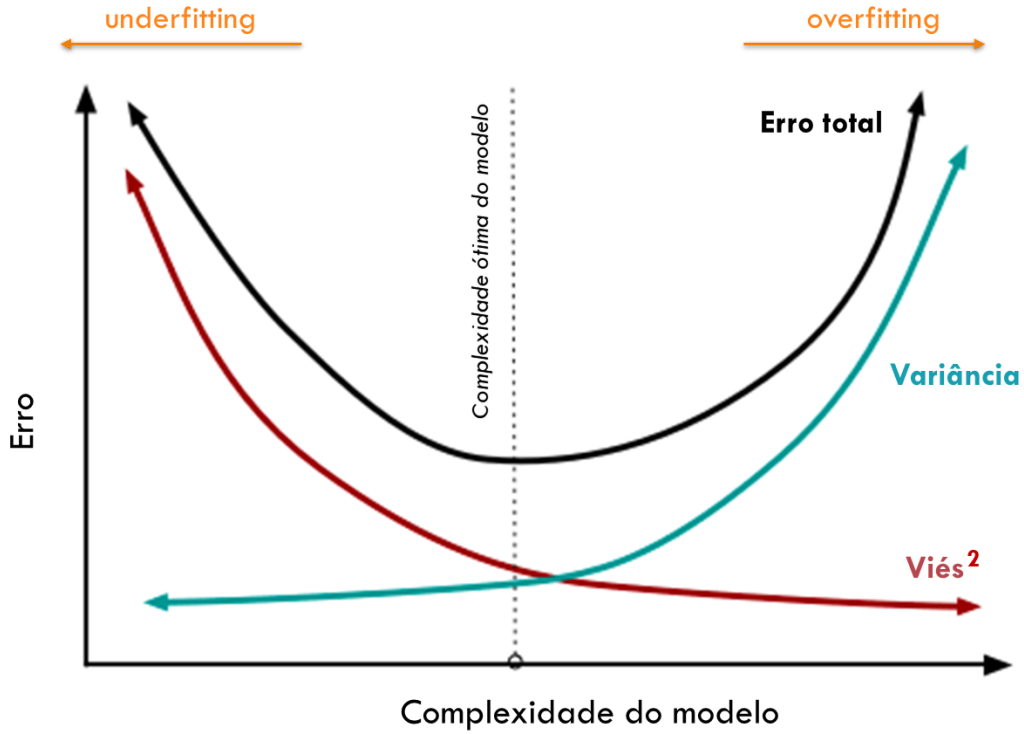
Alta Variância (variance): sensibilidade de um modelo ao ser usado com outro conjunto de dados, diferente do treinamento.

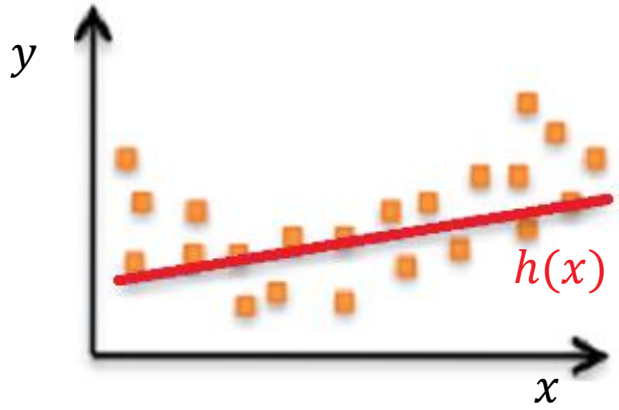
OVERFITTING

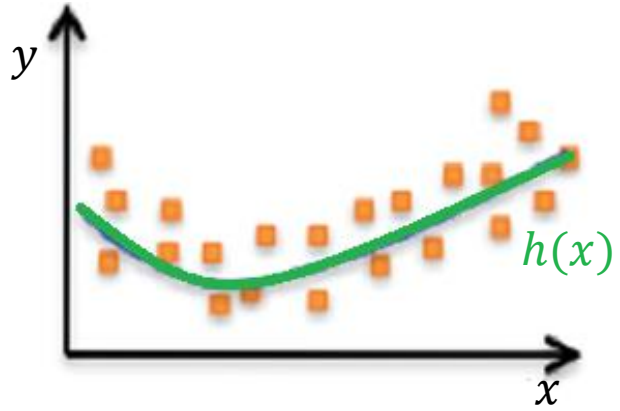

underfitting
overfitting
Erro total
Complexidade ótima do modelo
Variância
Viés 2
Erro
Complexidade do modelo

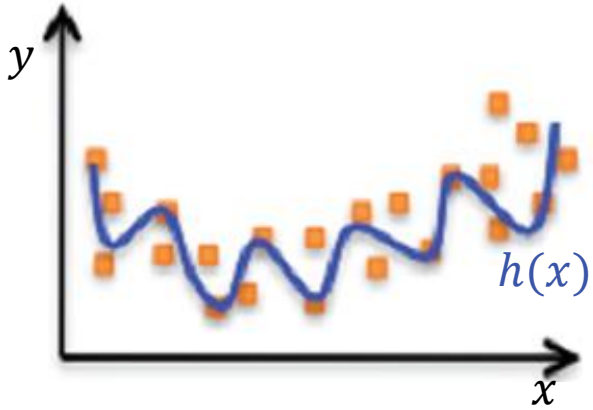

y
x
h(x)
y
x
h(x)
y
x
h(x)

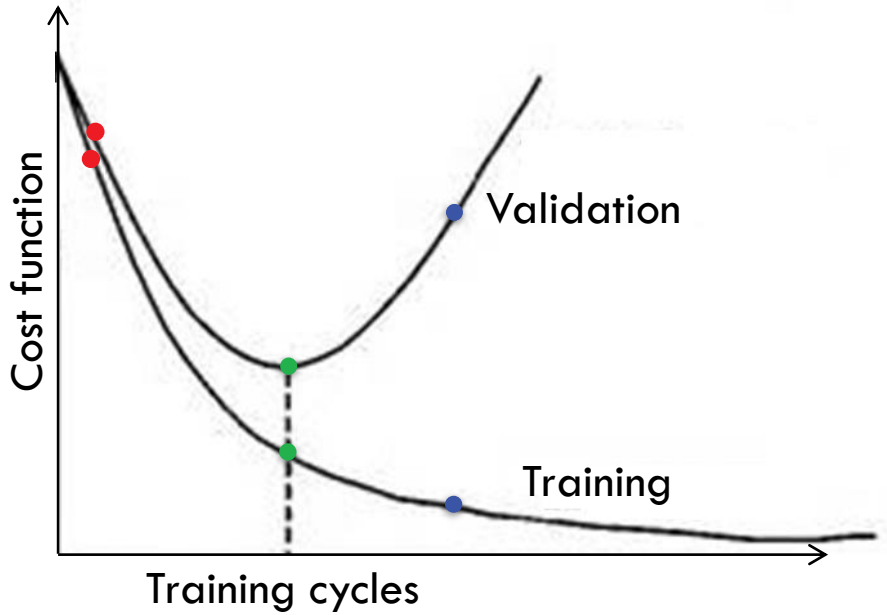

Custo
Validação
Treinamento
Ciclos de treinamento

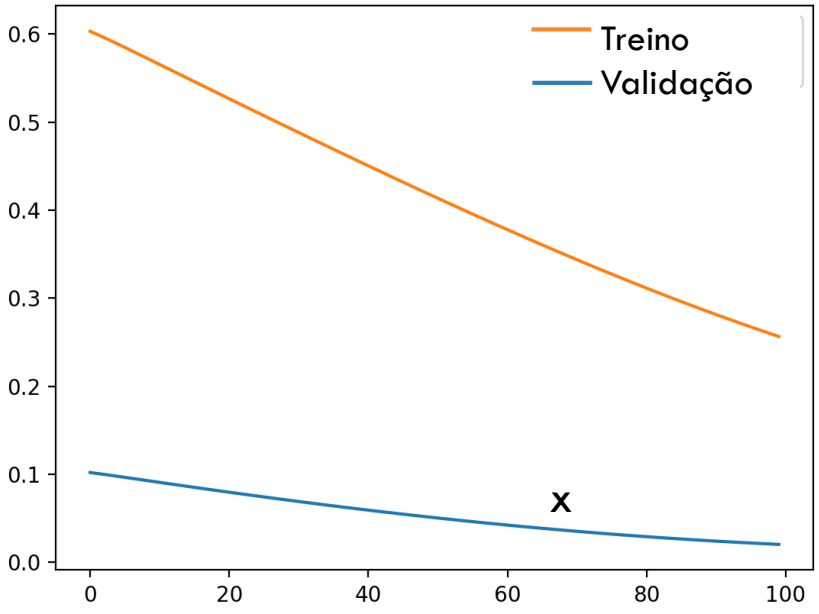

underfitting
overfitting
Erro total
Complexidade ótima do modelo
Variância
Viés²
Erro
Complexidade do modelo

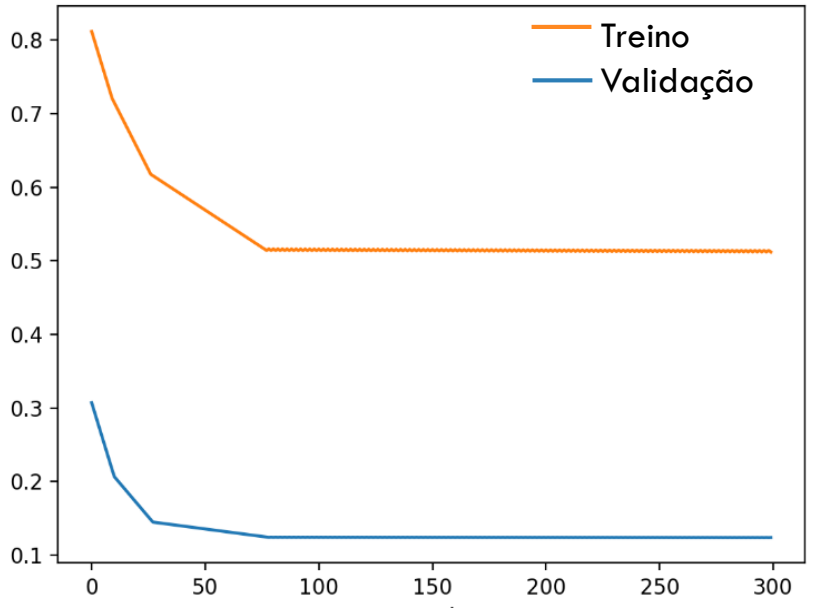

y
x
h(x)

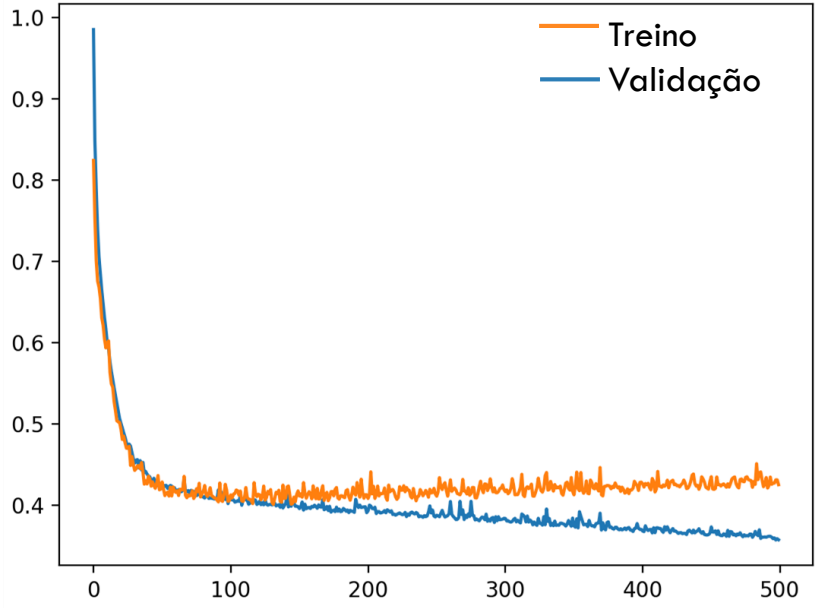

y
x
h(x)

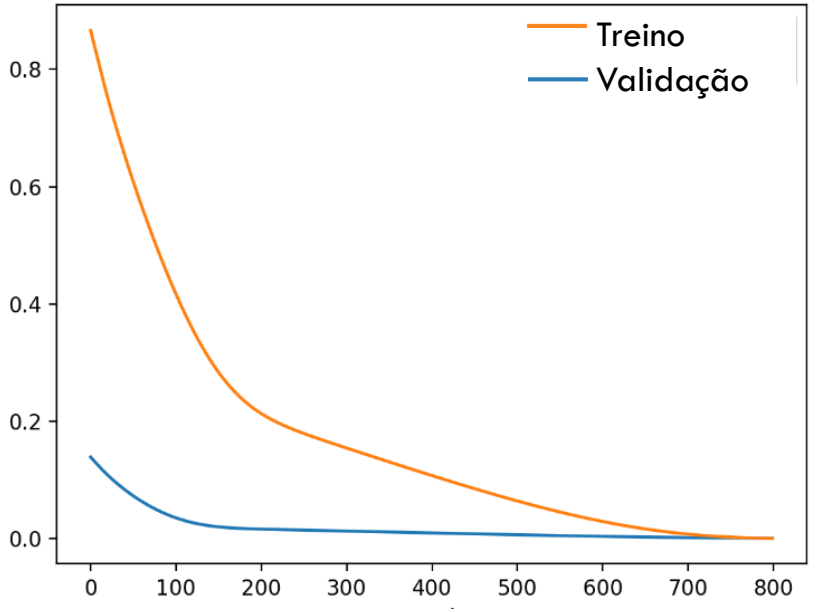

y
x
h(x)


Cost function
Validation
Training
Training cycles

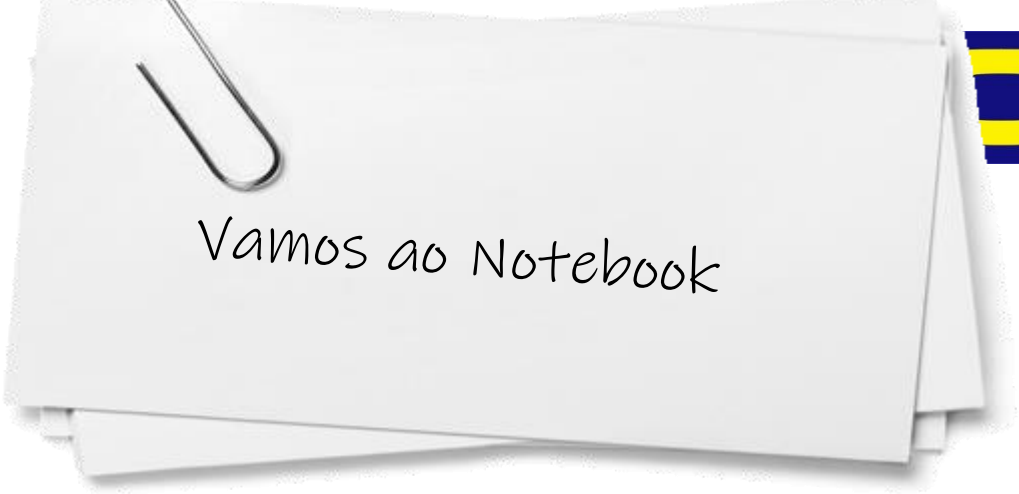

Treino
Validação
x
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
20
40
60
80
100


Treino
Validação
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
50
100
150
200
250
300


Treino
Validação
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
100
200
300
400
500


Treino
Validação
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
600
700
800

EXEMPLO



Com os dados abaixo,

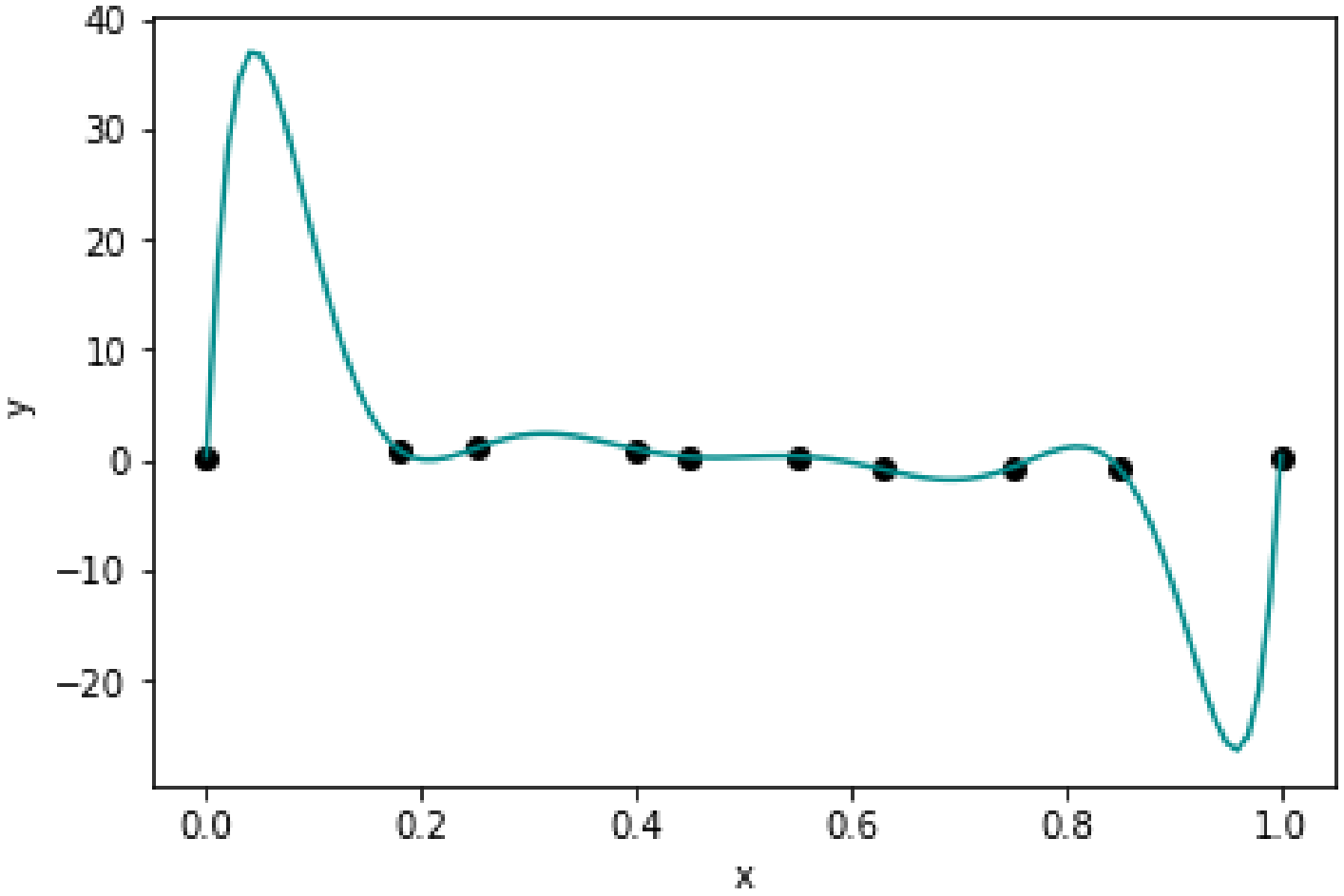
$$x = [0., 0.18, 0.25, 0.4, 0.45, 0.55, 0.63, 0.75, 0.85, 1.]$$

$$y = [0.3, 0.8, 1., 0.95, 0.25, 0.3, -0.9, -0.7, -0.8, 0.35]$$

modifique a ordem de aproximação de 0 a 9 e responda:

a) o que aconteceu com os valores dos parâmetros a medida que o grau do polinômio de interpolação aumentou?

b) porque isso aconteceu?


40
30
20
10
0
−10
−20
y
0.0
0.2
0.4
0.6
0.8
1.0
x

	$n = 0$	$n = 1$	$n = 4$	$n = 9$
ω_0	0.155	−1.31	0.27	0.30
ω_1		0.82	5.41	2083.08
ω_2			−7.77	39882.57
ω_3			−15.21	315869.13
ω_4			17.67	−1355675.15
ω_5				3466475.49
ω_6				−5433998.03
ω_7				5120734.79
ω_8				−2661308.98
ω_9				585702.29

OVERFITTING

Detectar overfitting é útil, mas não resolve o problema!

Se seu modelo está ajustando demais os dados de treinamento, faz sentido executar ações que **reduzam a flexibilidade do modelo.**

Felizmente, existem várias técnicas eficientes que podem ser implementadas. As soluções mais populares e eficientes para sobreajuste são:

1. Treinamento com mais dados
2. Remoção de features
3. Parada antecipada
4. Validação cruzada (cross-validation)
5. Aumente a regularização

UNDERFITTING

O fraco desempenho nos dados de treinamento pode ser porque o modelo é muito simples para descrever bem o destino. O desempenho pode ser aprimorado **aumentando a flexibilidade do modelo**. Para aumentar a flexibilidade do modelo, tente o seguinte:

1. Adição de mais termos do polinômio ao seu conjunto de dados
2. Diminuição da quantidade de regularização usada
3. Divisão ruim entre treino/validação
4. Aumento de dados (data augmentation)

REGULARIZAÇÃO

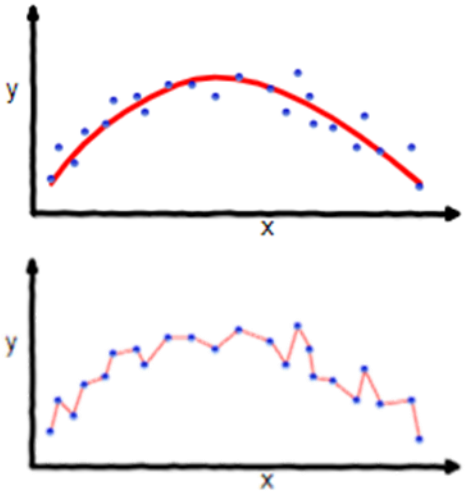
A palavra regularizar significa tornar as coisas regulares ou aceitáveis. É exatamente que faremos.

Regularizações são técnicas usadas para reduzir o erro, ajustando uma função adequadamente no conjunto de treinamento fornecido e evitando o ajuste excessivo.

$$h(\boldsymbol{x}) = \hat{y}(\boldsymbol{x}) = \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_{n-1} x_{n-1} + \omega_n x_n = \boldsymbol{\omega}^T \boldsymbol{x}$$

$$J(\boldsymbol{\omega}) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left[h\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right]^2 + 100|\omega_{n-1}| + 100|\omega_n|\right]$$

$$\omega_{n-1} \approx 0$$
$$\omega_n \approx 0$$



MODELO DE REGULARIZAÇÃO *ELASTIC NET*

$$J(\boldsymbol{\omega}) = \frac{1}{2m}\left\{\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - y^{(i)}\right]^2 + \lambda\left[\alpha\sum_{j=1}^{n}|\omega_j| + (1-\alpha)\sum_{j=1}^{n}|\omega_j|^2\right]\right\}$$

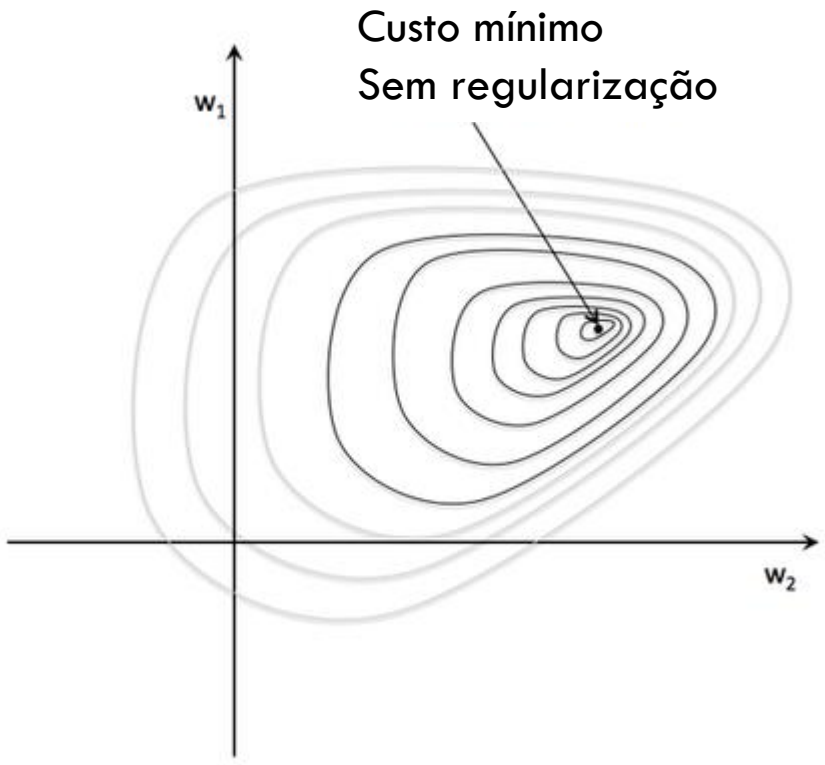
$\alpha = 1$ *regularização Lasso ou L1*

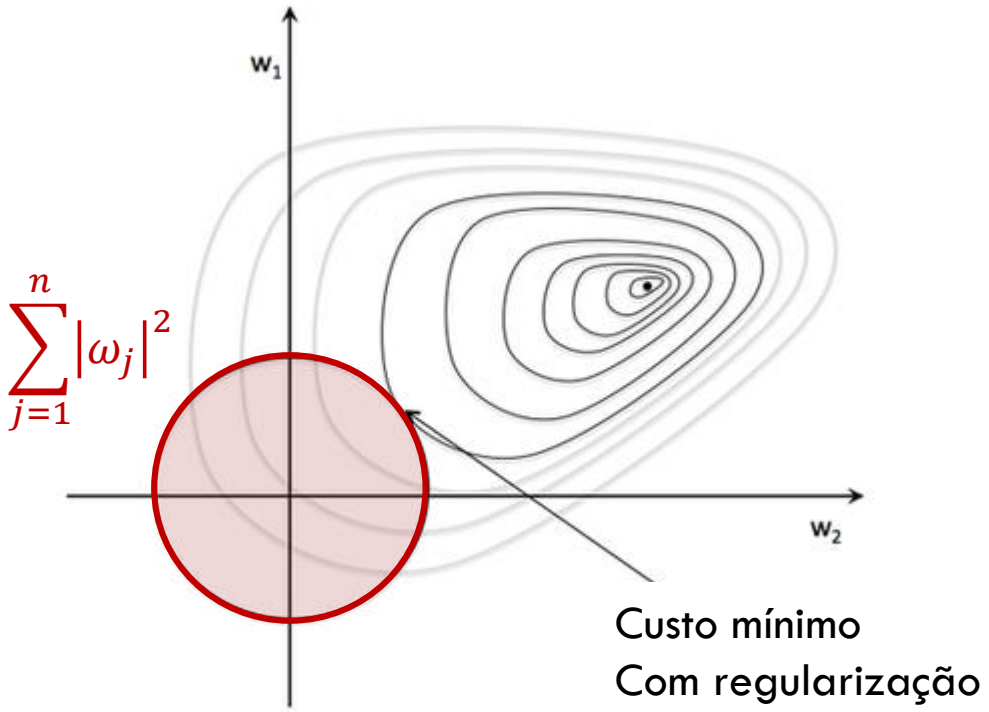
$\alpha = 0$ *regularização Ridge ou L2*

λ *é o parâmetro de regularização*, que controla o equilíbrio entre **regularizar** e **treinar.**

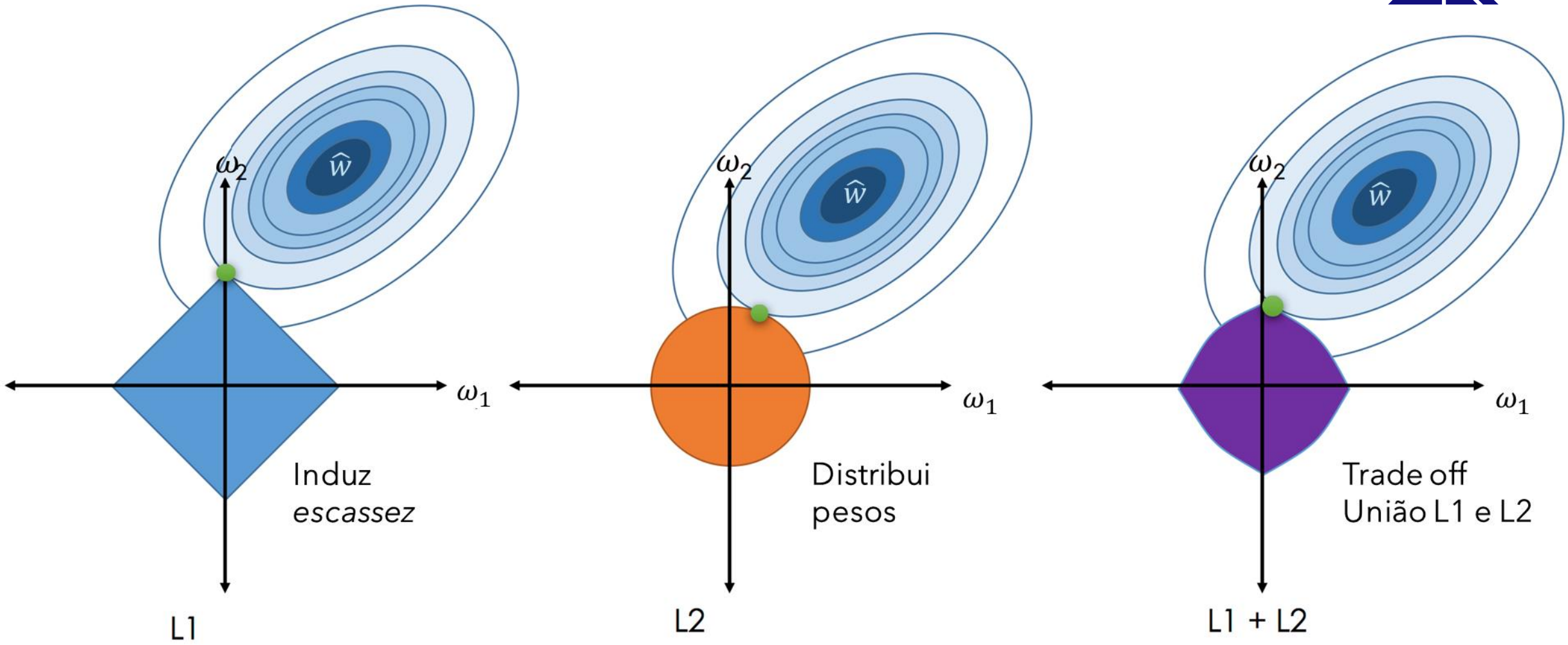
α, λ *são hiperparâmetros do modelo*

DEFINIÇÃO GEOMÉTRICA

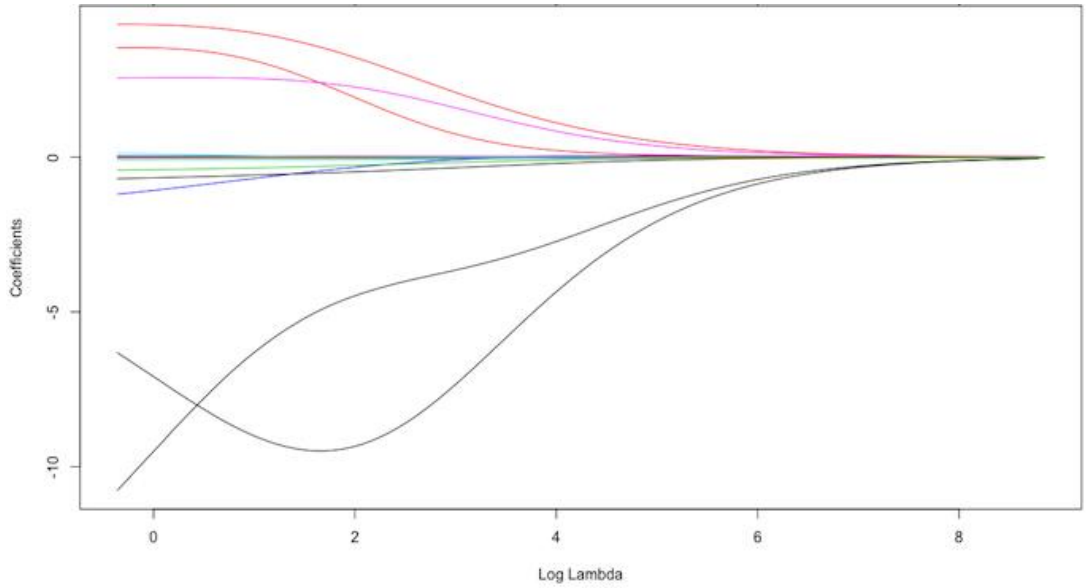


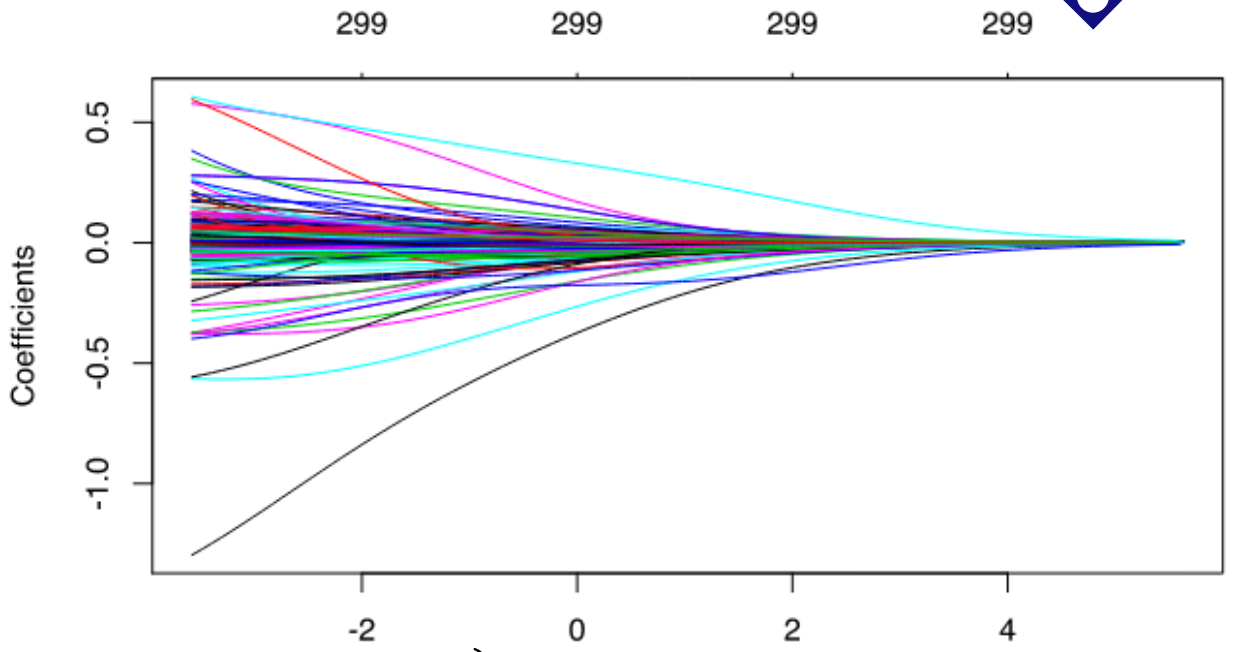


https://sebastianraschka.com/faq/docs/regularized-logistic-regression-performance.html


ω_2
$\widehat{w}$
ω_1
Induz
escassez
L1
ω_2
$\widehat{w}$
ω_1
Distribui
pesos
L2
ω_2
$\widehat{w}$
ω_1
Trade off
União L1 e L2
L1 + L2

RIDGE

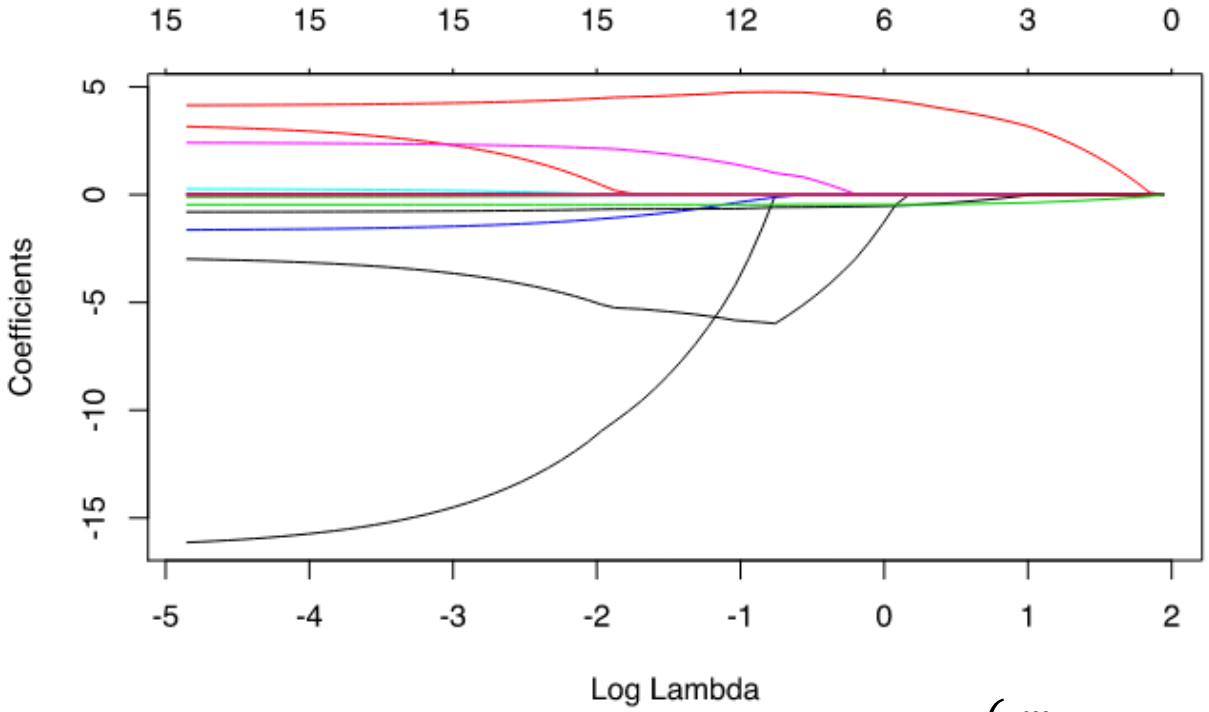


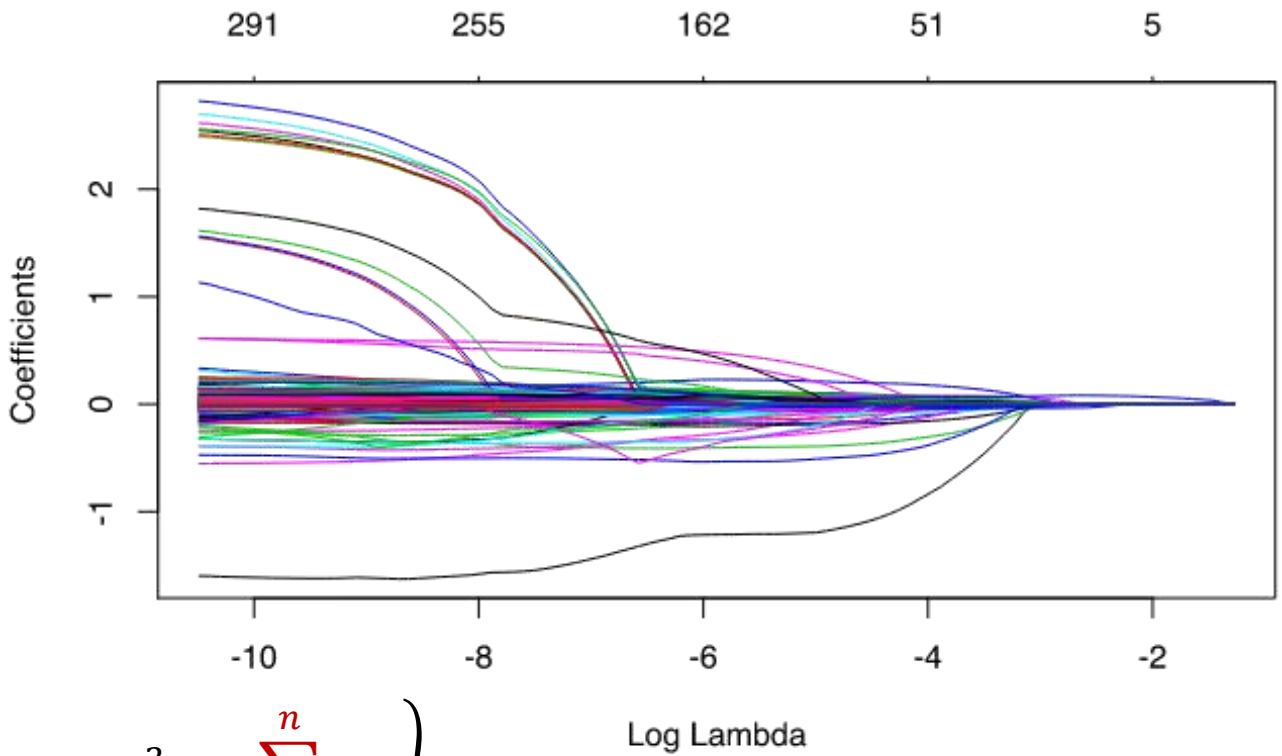


$$J(\boldsymbol{\omega}) = \frac{1}{2m}\left\{\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - y^{(i)}\right]^2 + \lambda \sum_{j=1}^{n}\left|\omega_j\right|^2\right\}$$

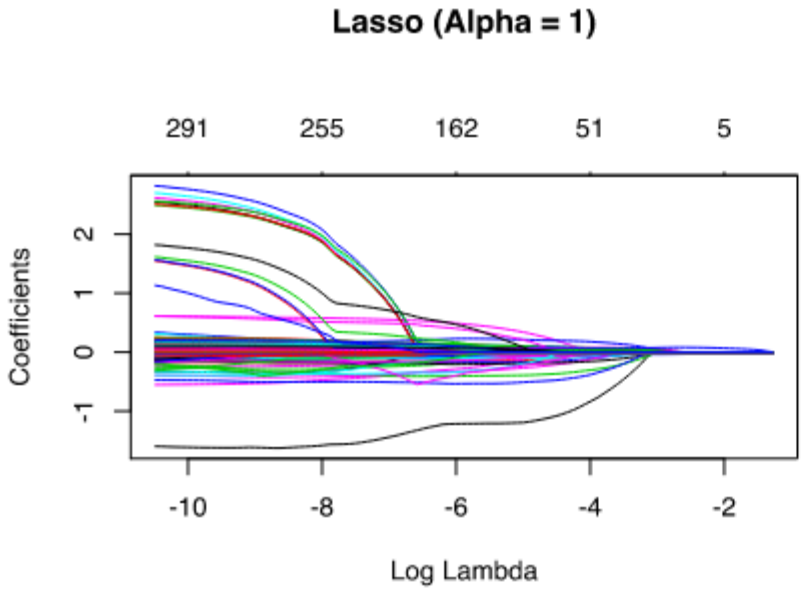
https://uc-r.github.io/regularized_regression

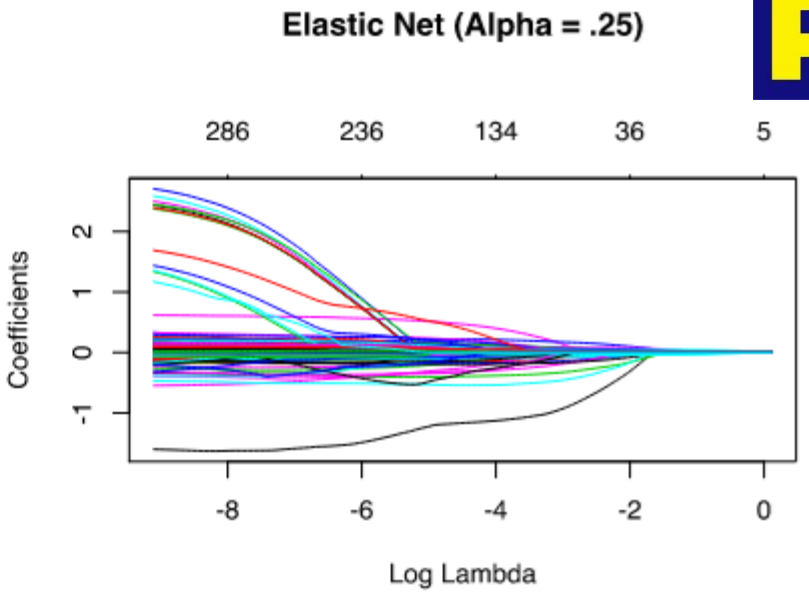
LASSO

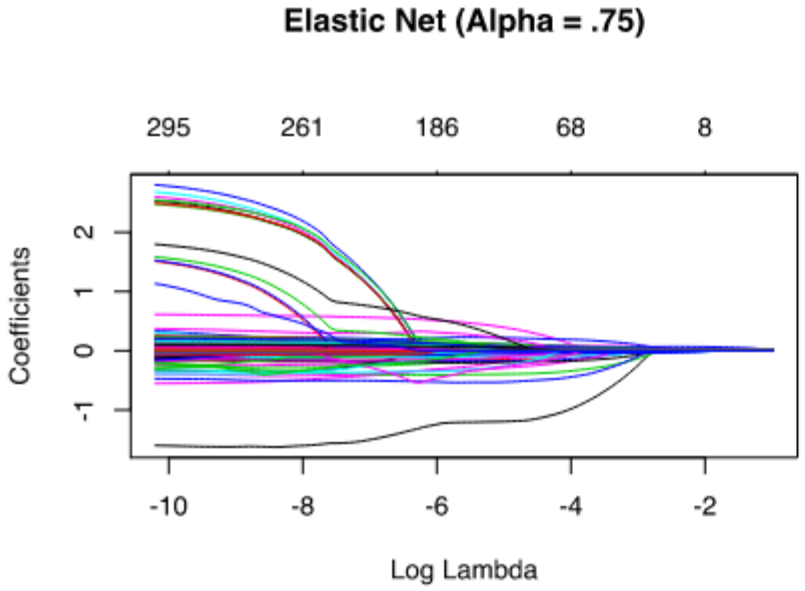


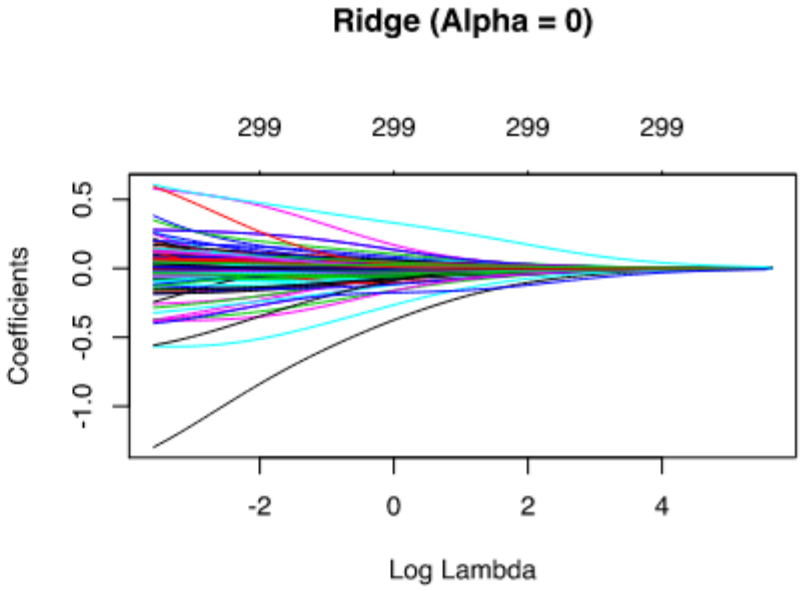


$$J(\boldsymbol{\omega}) = \frac{1}{2m}\left\{\sum_{i=1}^{m}\left[h\left(x^{(i)}\right) - y^{(i)}\right]^2 + \lambda\sum_{j=1}^{n}|\omega_j|\right\}$$


Lasso (Alpha = 1)
291
255
162
51
5
Coefficients
-1
0
1
2
-10
-8
-6
-4
-2
Log Lambda

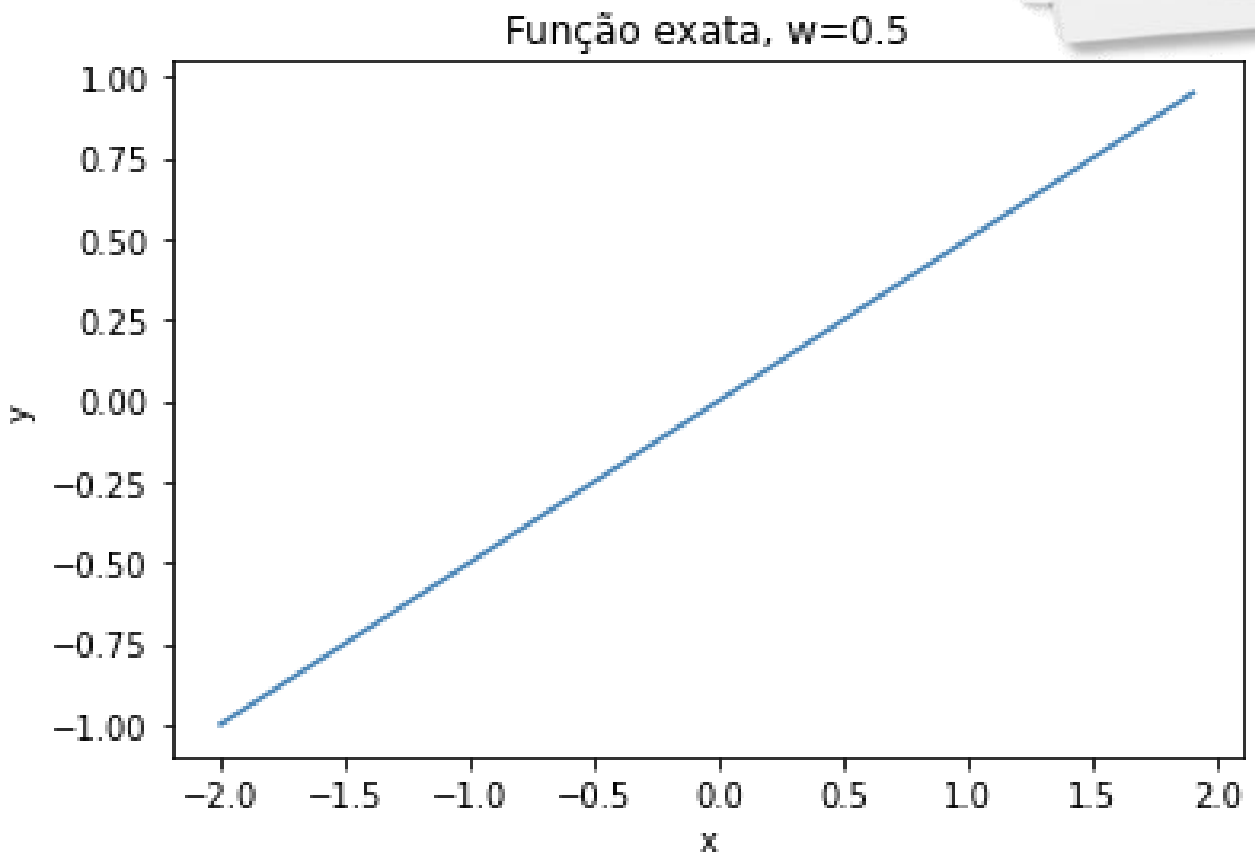

Elastic Net (Alpha = .25)
286
236
134
36
5
Coefficients
-1
0
1
2
-8
-6
-4
-2
0
Log Lambda


Elastic Net (Alpha = .75)
295
261
186
68
8
Coefficients
-1
0
1
2
-10
-8
-6
-4
-2
Log Lambda

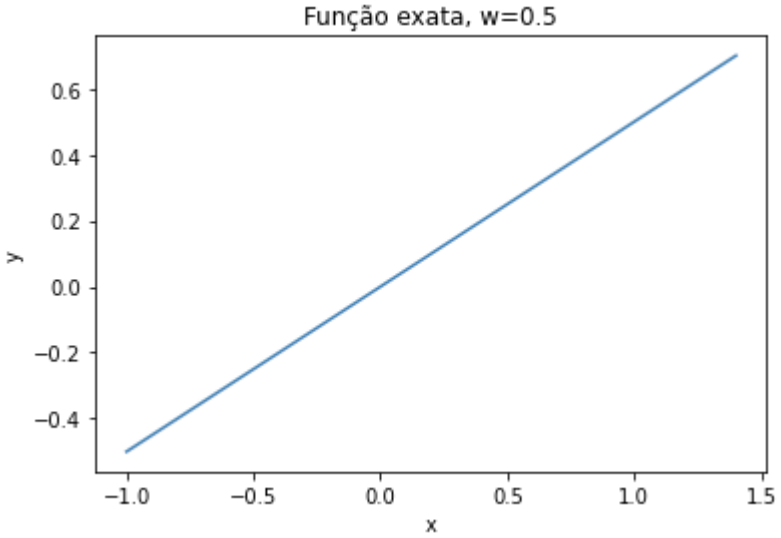

Ridge (Alpha = 0)
299
299
299
299
Coefficients
-1.0
-0.5
0.0
0.5
-2
0
2
4
Log Lambda

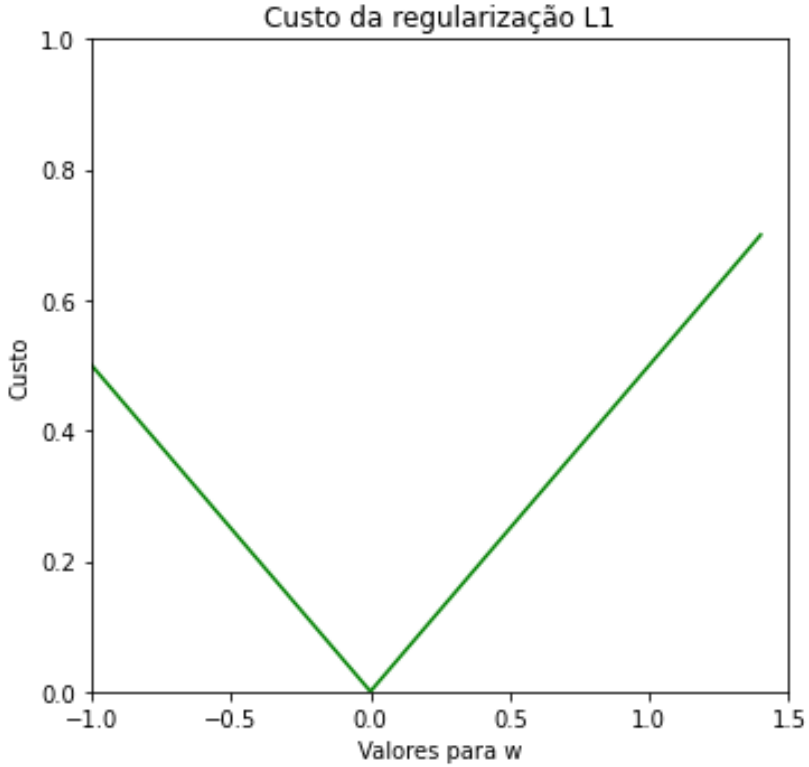
FUNÇÃO $y = \omega x$

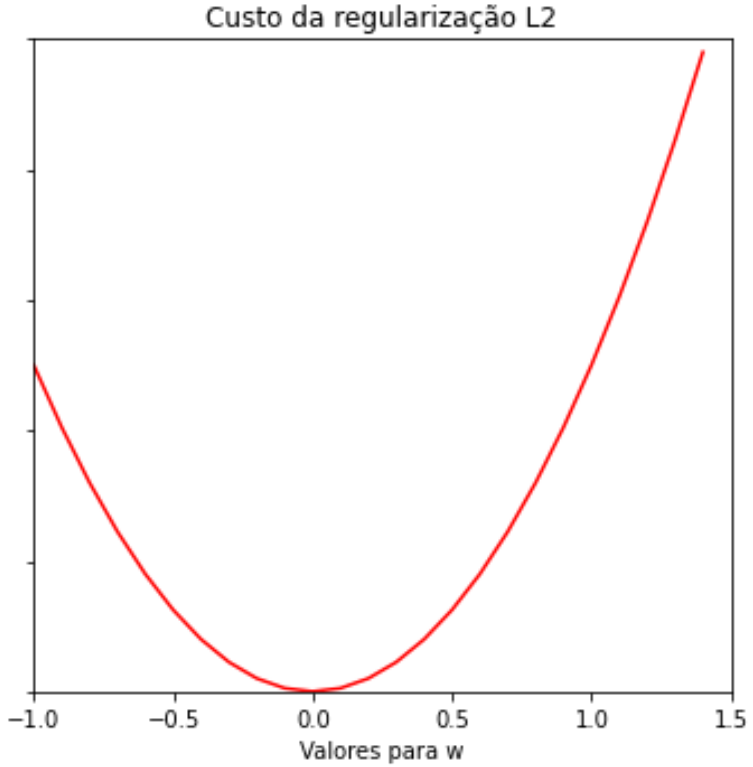
Vamos ao Notebook



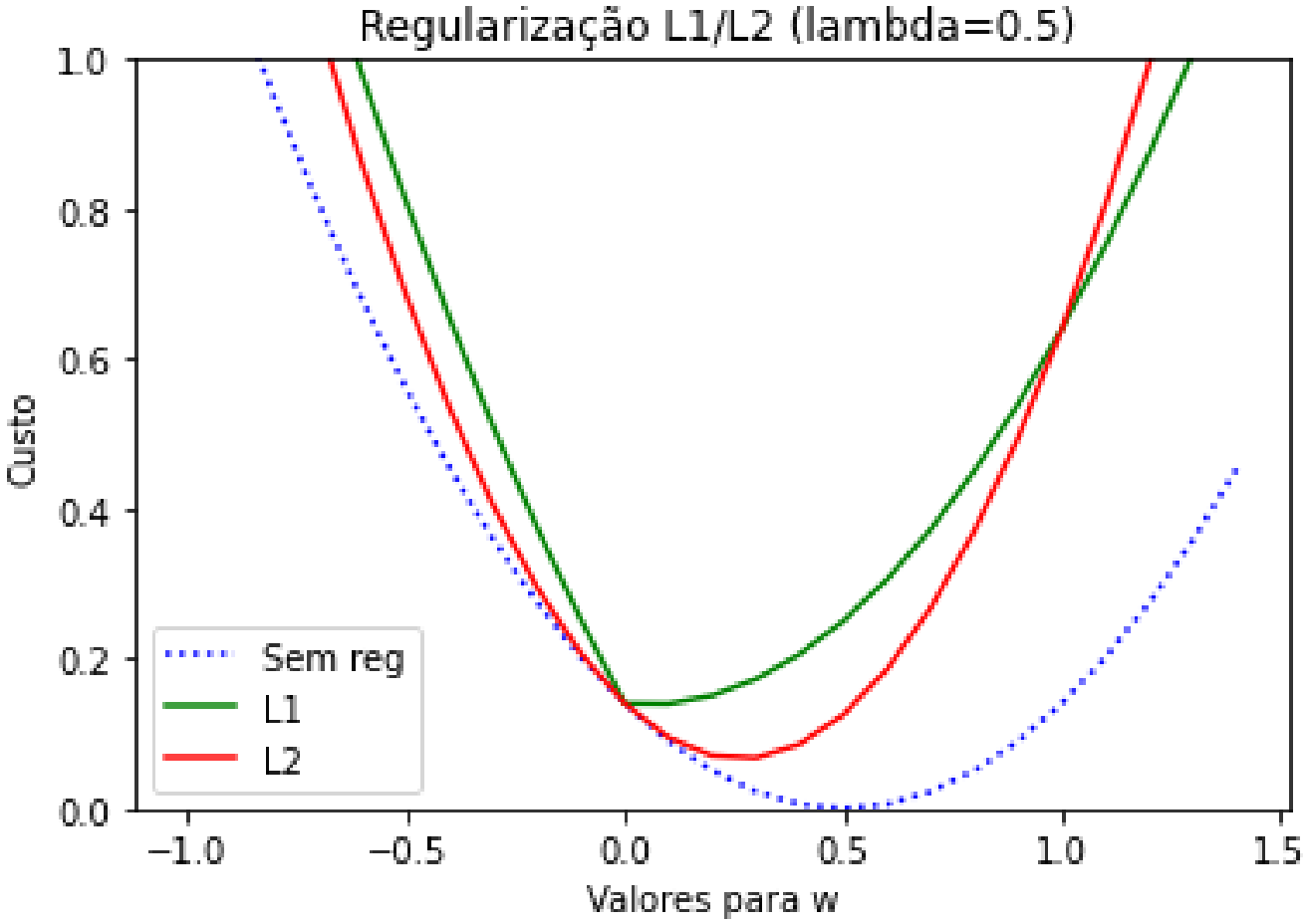
$$\hat{y} = \omega_0 + \omega_1 x = \omega x$$



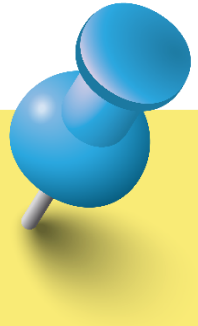




$$L = (\hat{y} - y)^2 = (\omega x - y)^2$$
$$L_1 = (\omega x - y)^2 + \lambda|\omega|$$
$$L_2 = (\omega x - y)^2 + \lambda\omega^2$$


Regularização L1/L2 (lambda=0.5)
Custo
Valores para w
Sem reg
L1
L2
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
1.5

REVISÃO



Lição de casa

Estude os slides da aula de hoje e refaça o Notebook. Complete a parte em azul e entregue no Moodle.

ENOUGH IS ENOUGH!

ACABOU...

Nossa próxima aula de ML será sobre problemas de classificação. Faremos exemplos usando redes neurais ou regressão logística