

Aprendizagem de Máquina 1

Inteligência Artificial



AULA 01 – APRESENTAÇÃO DA DISCIPLINA E REGRESSÃO

Larissa Driemeier
Thiago Martins

CRONOGRAMA

Data	Professor	Assunto
07/05	Larissa	Definição de aprendizado de máquina. Aprendizado supervisionado e não supervisionado. Regressão linear. Regressão polinomial.
14/05	Thiago	Exercícios de acompanhamento. Nota 01.
21/05	Larissa	Régressão Logística
28/05	Thiago	Exercícios de acompanhamento. Nota 02.
04/06	Larissa	Máquinas de vetores de suporte
11/06	Thiago	Exercícios de acompanhamento. Nota 03.
18/06	Larissa	Aprendizado não supervisionado
25/06	Thiago	Exercícios de acompanhamento. Nota 04.
02/07	Larissa	Redução de similaridade: análise de componentes principais (PCA) e suas variações.
16/07	Larissa/Thiago	Exercícios “Melhores Momentos”. Nota 05.

AULA DE HOJE

- Introdução ao Aprendizado de Máquinas
- O que é regressão
- Regressão Linear
 - Simples
 - Múltipla
 - Polinomial



INTRODUÇÃO AO APRENDIZADO DE MÁQUINAS

Aprendizado supervisionado e
não supervisionado
Problemas de regressão e
classificação

O PROBLEMA DE APRENDIZADO...



1 ano	35000
2 anos	30800
3 anos	23800
...	

12% por ano???

Em termos de Aprendizagem de máquinas, o Coiote inventou a regressão – ele previu um valor baseado em dados históricos. Fazemos isso todo tempo!

MAS É TÃO SIMPLES ASSIM?

Ano de fabricação ?

Farol de milha, ar condicionado, bancos de couro...?

Direção elétrica...?

Está perto do Natal, dia das mães... ?

Help!!! O coiote precisa de um robô para fazer as contas por ele – alimentamos a máquina com todos os dados possíveis e pedimos que ela encontre todos os padrões ocultos relacionados ao preço.

Isto É aprendizado de máquinas!



NETFLIX: PREVENDO COMO UM ESPECTADOR AVALIARÁ UM FILME

Para uma **melhoria de previsão de 10%**, a Netflix deu um prêmio de US \$ 1 milhão!



<https://www.kaggle.com/netflix-inc/netflix-prize-data>



<https://www.slideshare.net/justinbasilico/making-netflix-machine-learning-algorithms-reliable>



<http://arek-paterek.com/predict.pdf>

ESSÊNCIA DO APRENDIZADO DE MÁQUINAS

- Existência de um padrão;
- Não é possível definir o problema matematicamente;

**O fato de haver um padrão e não haver uma relação matemática fechada é a razão pela qual usamos ML.
Razão pela qual usamos aprendizagem por dados.**
- Disponibilidade de dados.

CARTÃO PLATINUM

Modelo para prever se um indivíduo é aprovado para um cartão platinum ou não.

$$X = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}\}$$

$$\mathbf{x}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\}$$

$$Y = \{y^{(1)}, y^{(2)}, \dots, y^{(m)}\}$$

$$y^i = \begin{cases} 1 & - \text{bom cliente} \\ 0 & - \text{mau cliente} \end{cases}$$

$$P(y^j = 0,1 | \mathbf{x}^{(j)}) = ???$$

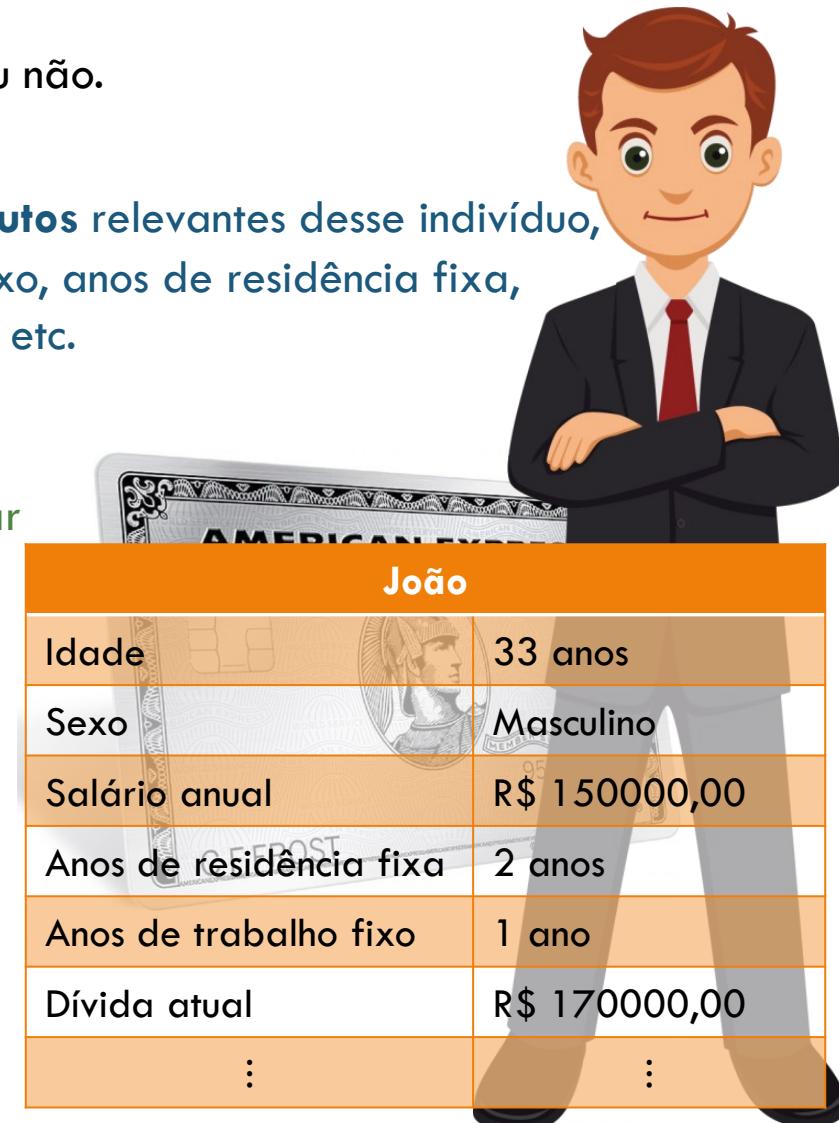
$$(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(m)}, y^{(m)})$$

Função alvo (target function) $f: X \rightarrow Y$

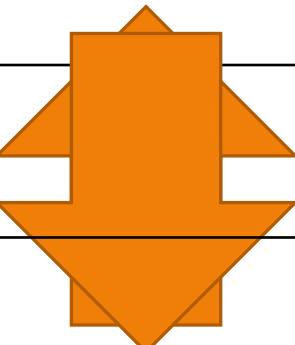
Histórico de dados (data set)

$\mathbf{x}^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)}\}$ seriam os **atributos** relevantes desse indivíduo, com informações pessoais como idade, sexo, anos de residência fixa, anos de emprego estável, salário, dívida, etc.

Com base nesses dados, queremos treinar um modelo capaz de prever se esse indivíduo deve ou não receber o cartão (binário).



	Número de filhos	Idade	Sexo	Histórico de trabalho	Débito
Cliente 01					...
Cliente 02					...
Cliente 100000					...



⋮

⋮



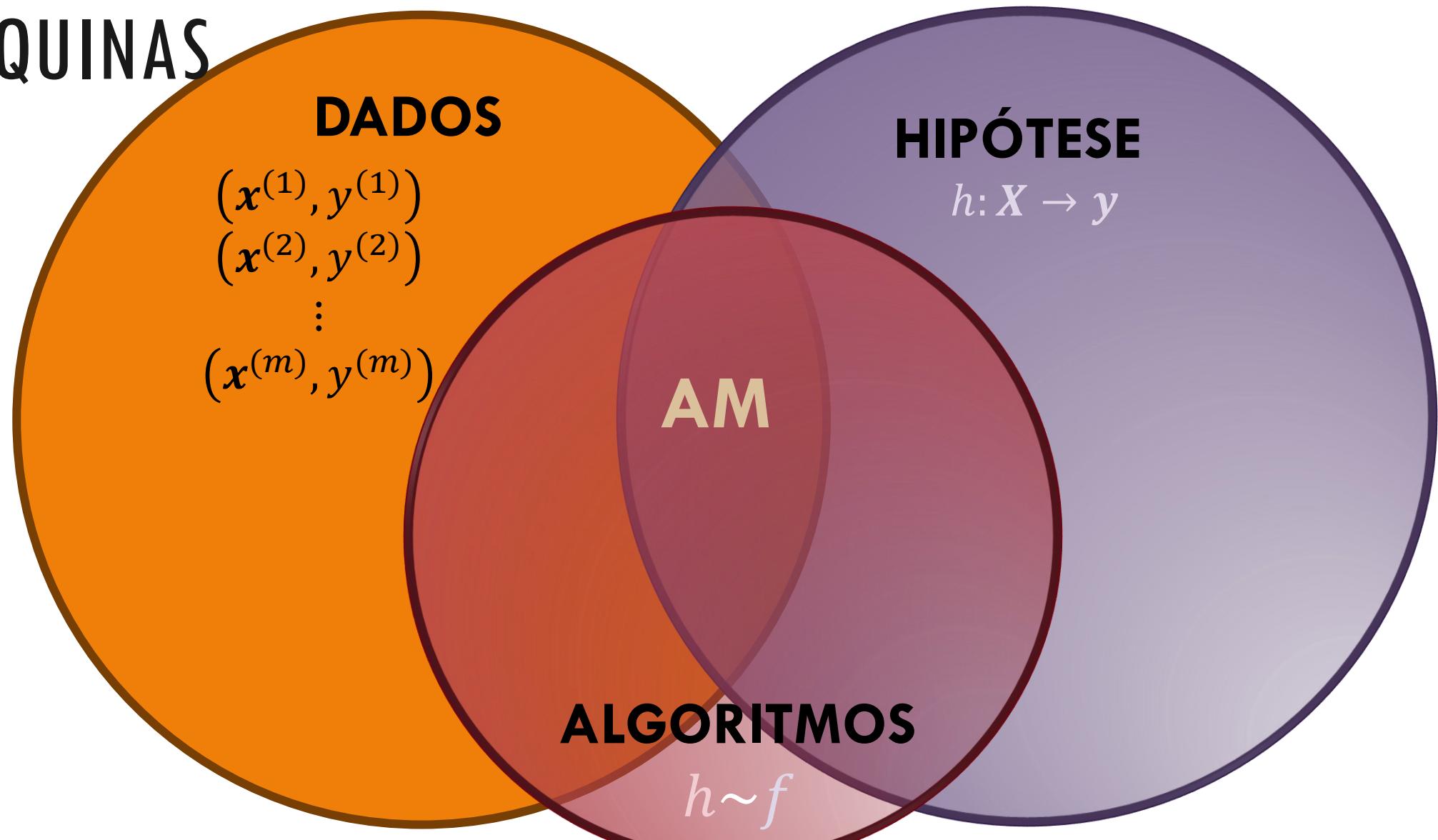
BOM CLIENTE



MAU CLIENTE



COMPONENTES DO APRENDIZADO DE MÁQUINAS



Função alvo

$$f: X \rightarrow Y$$

Dados de treinamento

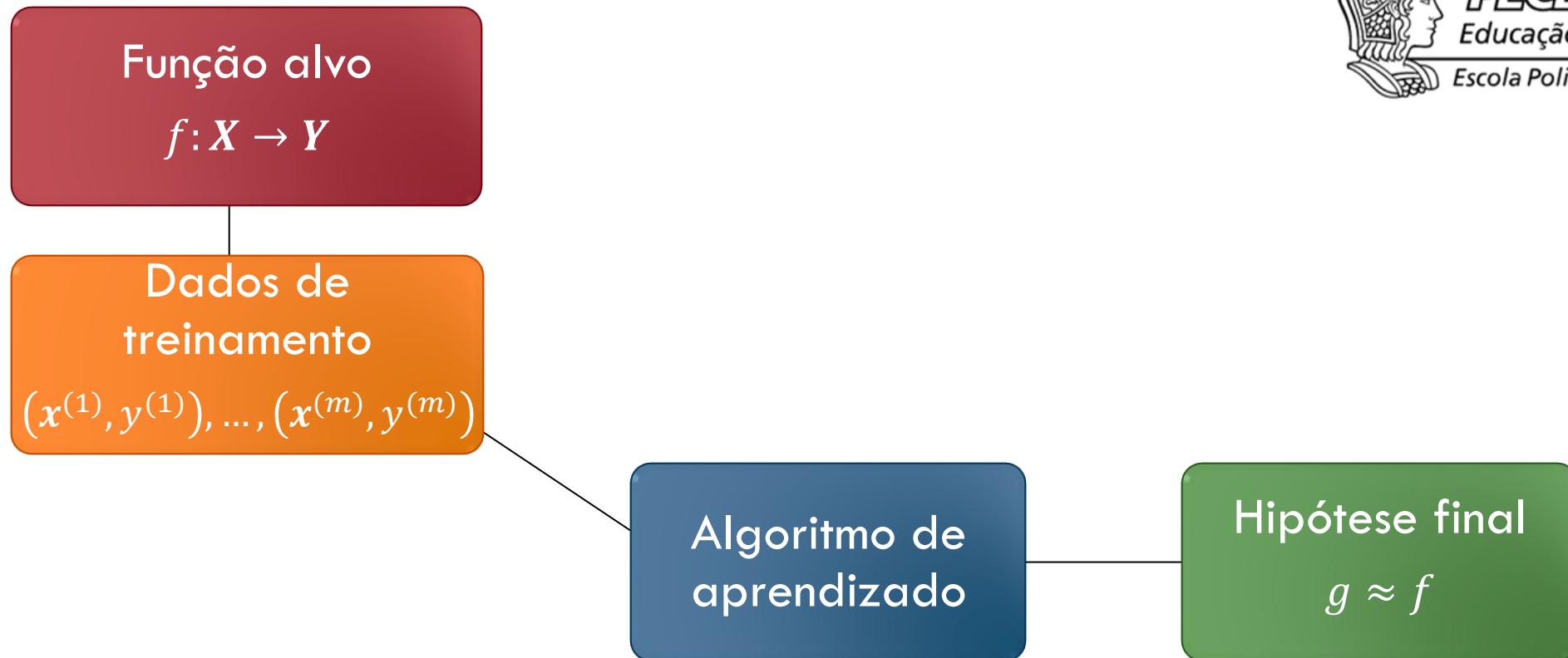
$$(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})$$

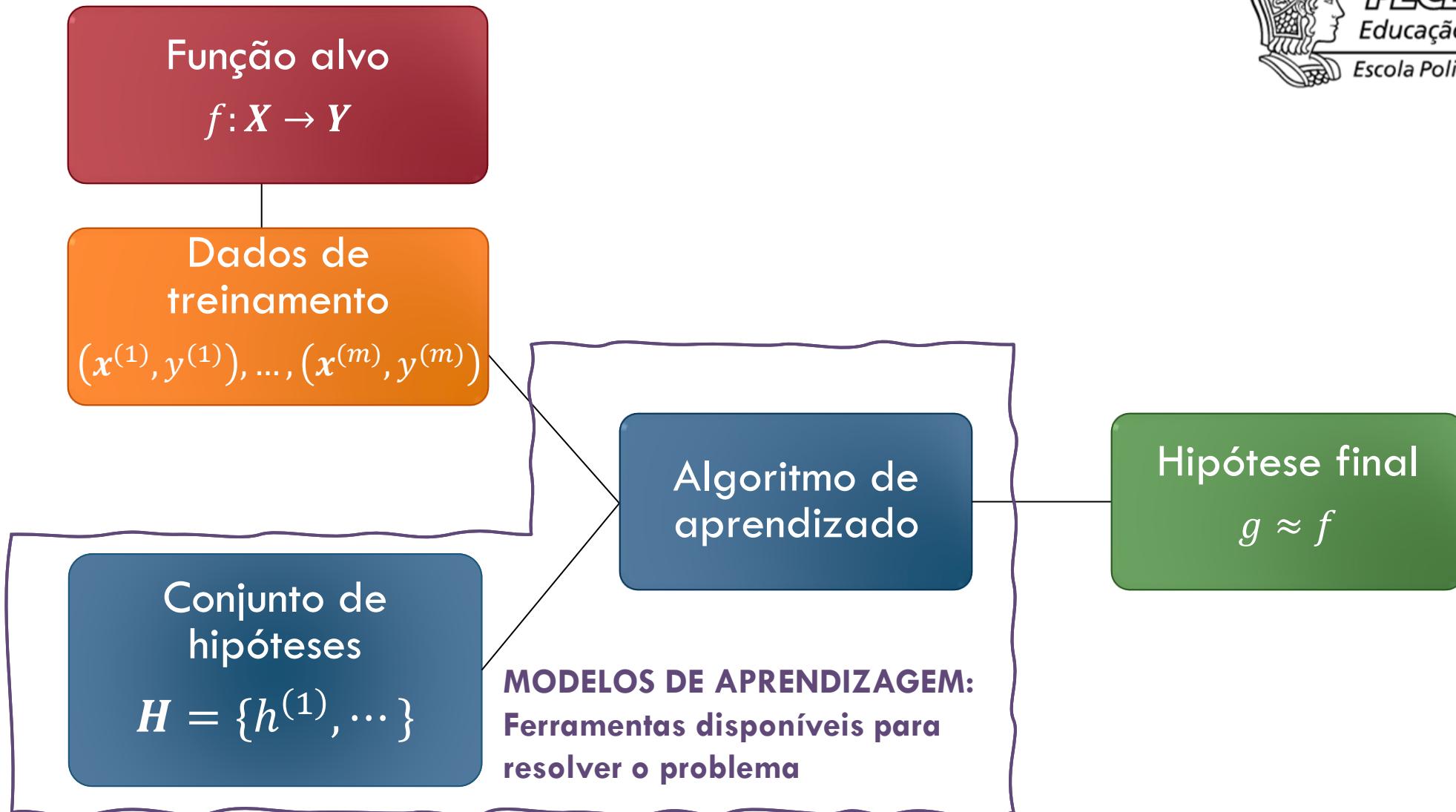
Vemos a função alvo apenas por meio de dados de treinamento. Os dados são nosso veículo para entender como é nossa função de destino.

A premissa de aprendizagem é usar um conjunto de observações para descobrir um processo subjacente...

Hipótese final

$$h \approx f$$





Aprendizado de Máquinas

Aprendizado supervisionado

Você supervisiona o processo de aprendizado, o que significa que os dados coletados são rotulados e, assim, você sabe qual entrada precisa ser mapeada para qual saída. Isso ajuda você a corrigir seu algoritmo se ele cometer um erro ao dar a resposta.

Problemas de classificação, regressão

Aprendizado não supervisionado

Os dados coletados aqui não têm rótulos e você não tem certeza sobre as saídas. Então, você modela seu algoritmo de forma que ele possa entender os padrões dos dados e gerar a resposta necessária. Você não interfere quando o algoritmo aprende.

Clusterização, classificação, detecção de anomalias, segmentação, reconhecimento de objetos...

Aprendizado por reforço

Não há dados nesse tipo de aprendizado, nem você ensina nada ao algoritmo. Você modela o algoritmo de forma que ele interaja com o ambiente e se o algoritmo fizer um bom trabalho, você o recompensa, senão pune o algoritmo. Com interações e aprendizado contínuos, o modelo aprende a fazer uma tarefa da melhor forma que pode.

Jogos, robótica, veículos autônomos

APRENDIZADO SUPERVISIONADO E NÃO SUPERVISIONADO

A maioria do aprendizado de máquina prático usa aprendizado supervisionado.

No aprendizado supervisionado você tem variáveis de entrada e variáveis de saída correspondentes, e aprende a função de mapeamento da entrada para a saída.



O aprendizado não supervisionado é uma técnica de aprendizado de máquina, na qual você não precisa *supervisionar* o modelo. Em vez disso, você permite que o modelo se vire para descobrir informações. Usamos dados não rotulados.

APRENDIZADO SUPERVISIONADO

O aprendizado supervisionado pode ser dividido com base no método de previsão usado.

Régressão: problemas de regressão observam dados contínuos. O problema clássico a ser resolvido com regressão é: *dado um valor x específico, qual é o valor esperado da variável y?*

- Regressão linear, polinomial
- Regressão Logística

Classificação: em problemas de classificação o algoritmo prevê um valor discreto, identificando os dados de entrada como o membro de uma classe ou grupo específico.

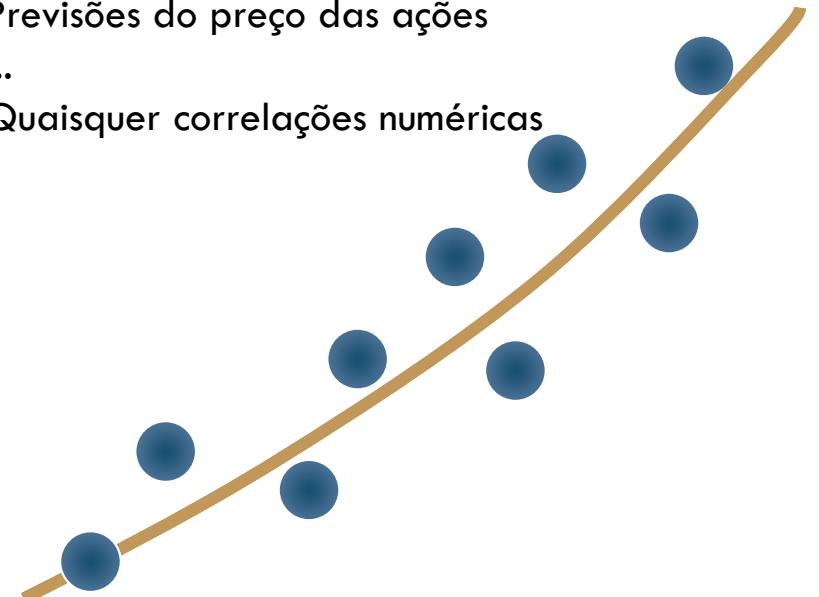
- Árvores de decisão (*Decision trees*)
- Algoritmo Bayesiano Ingênuo (*Naive Bayes Algorithm*)
- Redes neurais (*Neural networks*)
- K vizinhos mais próximos (*K-Nearest Neighbours*)
- Máquinas de vetores de suporte (SVM, do inglês *Support Vector Machines*)

CLASSIFICAÇÃO VS REGRESSÃO

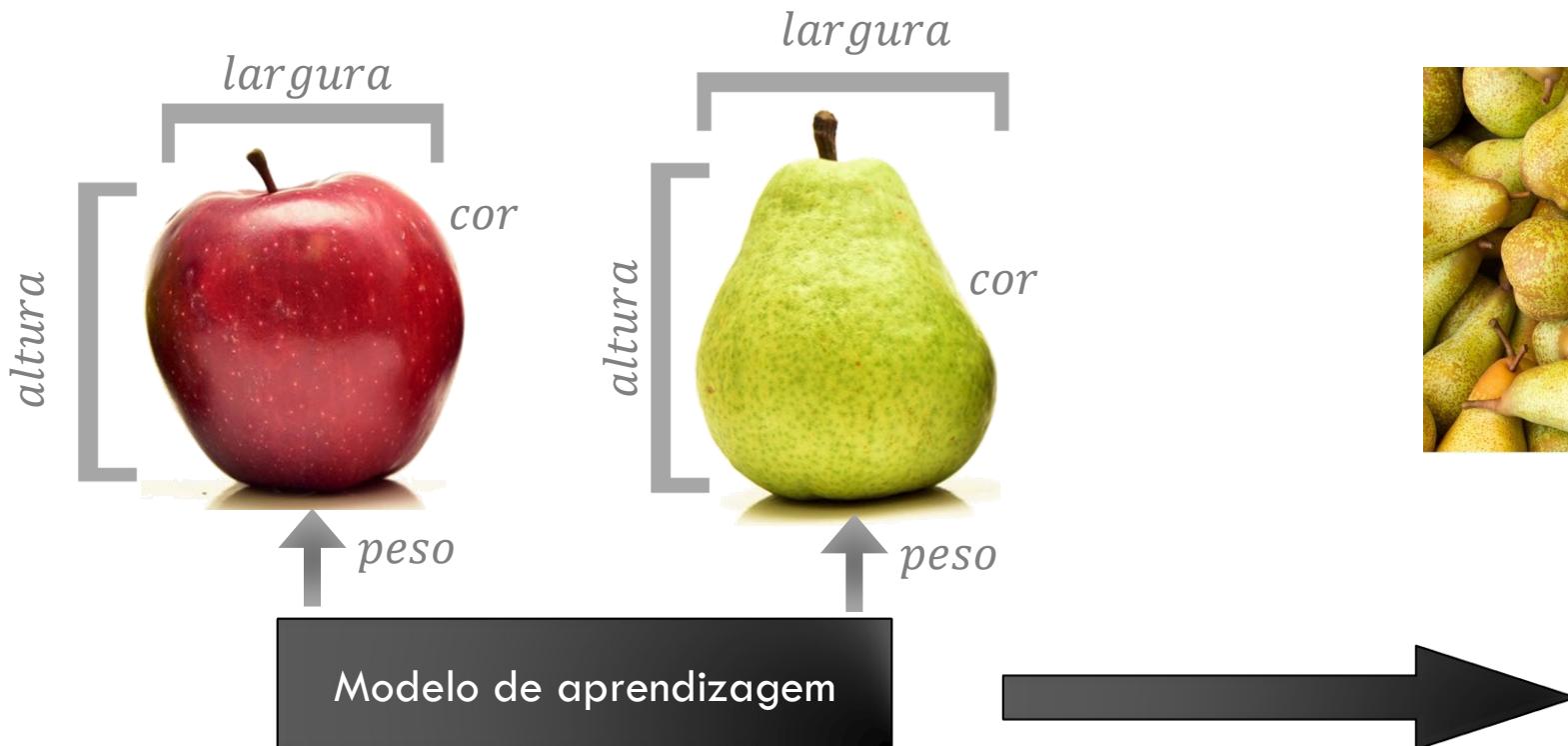


Regressão

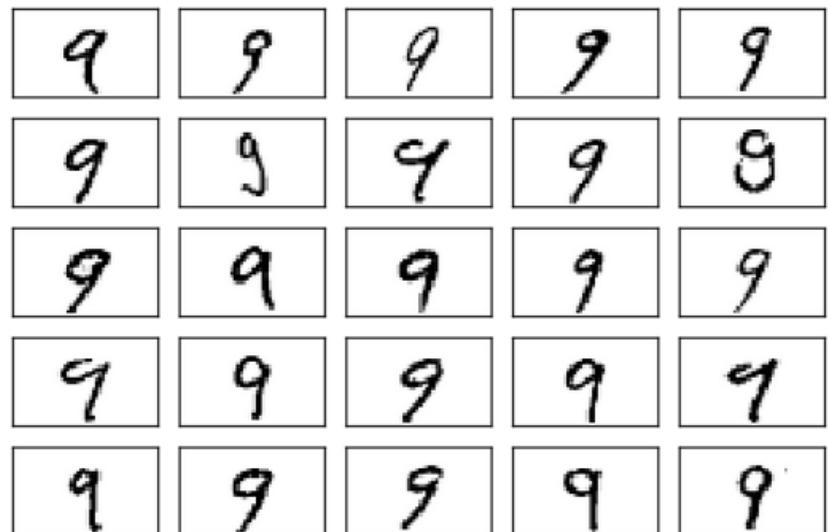
- Preço de imóvel, com base em metragem, localização e proximidade ao transporte público.
- Previsões do preço das ações
- ...
- Quaisquer correlações numéricas



EXEMPLOS DE CLASSIFICAÇÃO



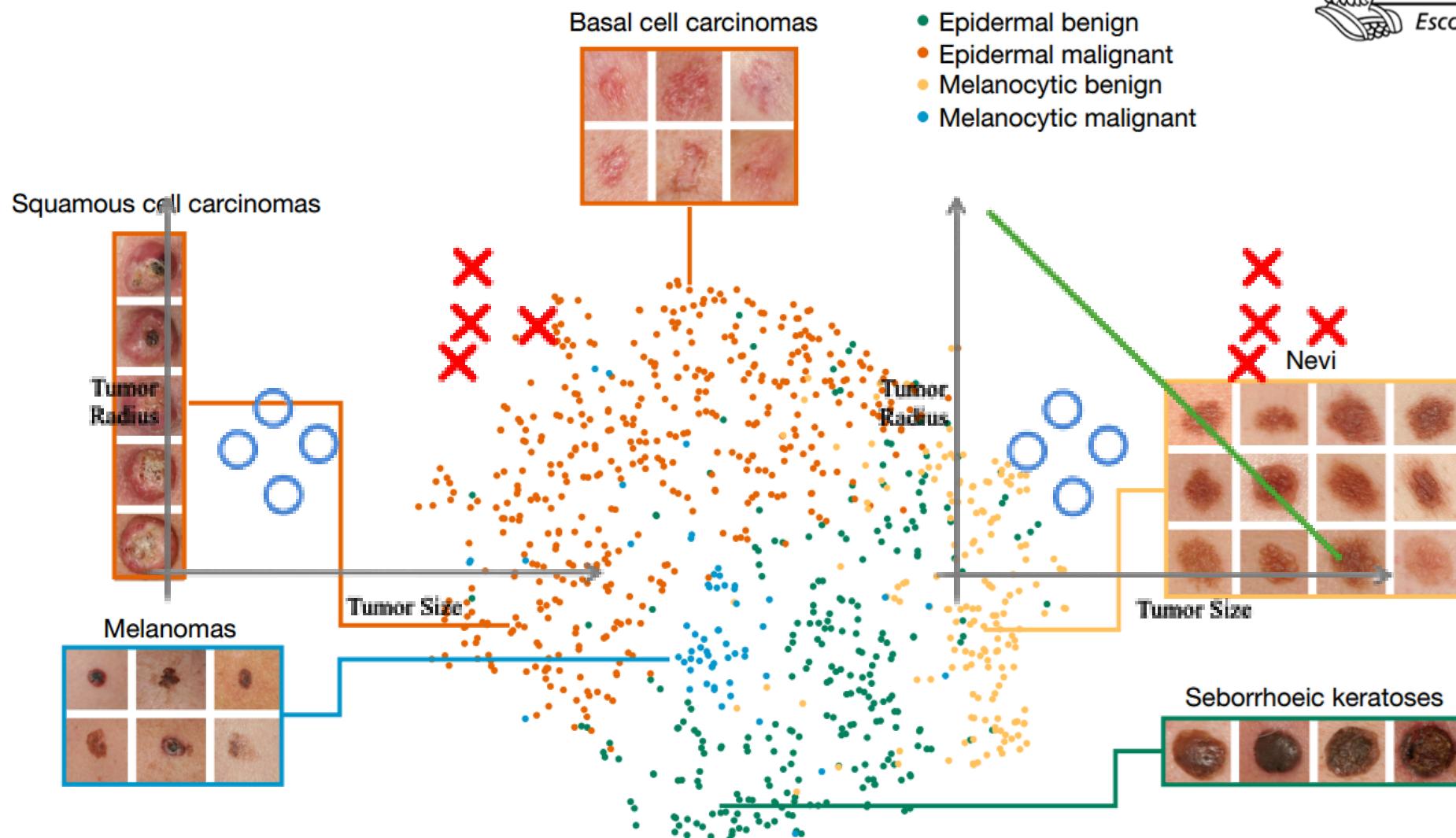
Identificando o código postal de dígitos manuscritos em um envelope



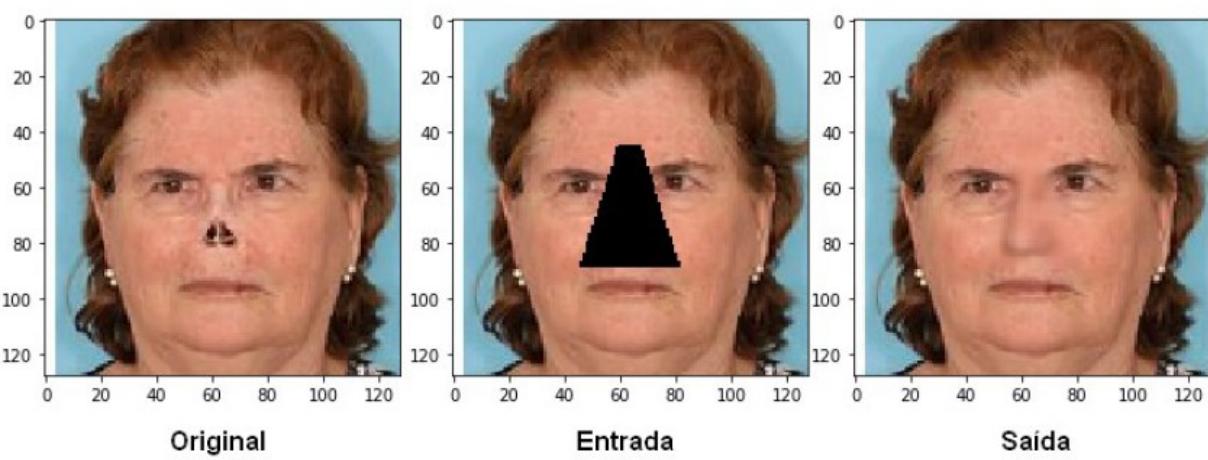
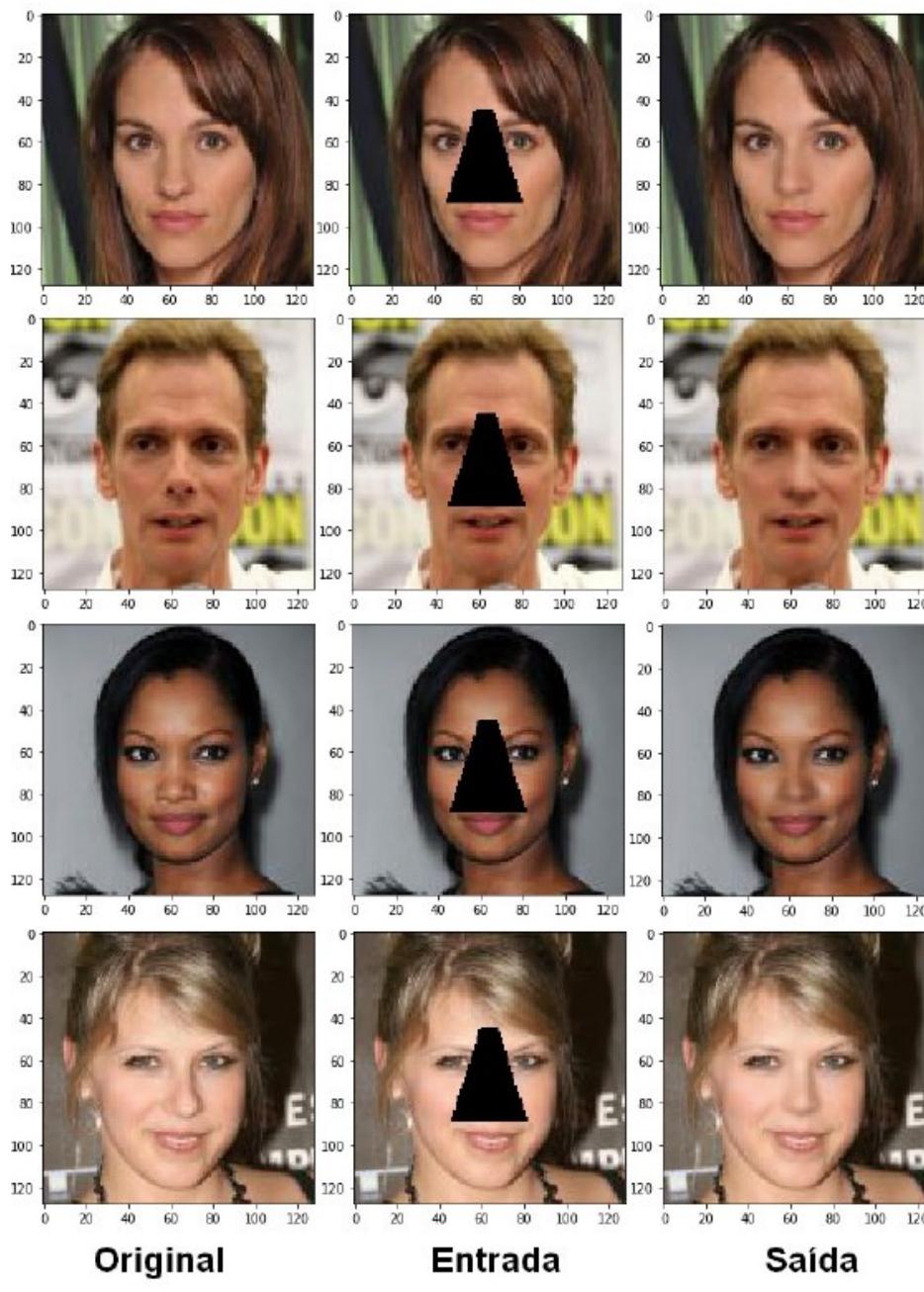
LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D.
Backpropagation applied to handwritten zip code recognition, *Neural Computation*, 1, 541-551
(1989).

80322-4129 80206
40004 14310
37878 05753
~~33502~~ 75316
35460 44209

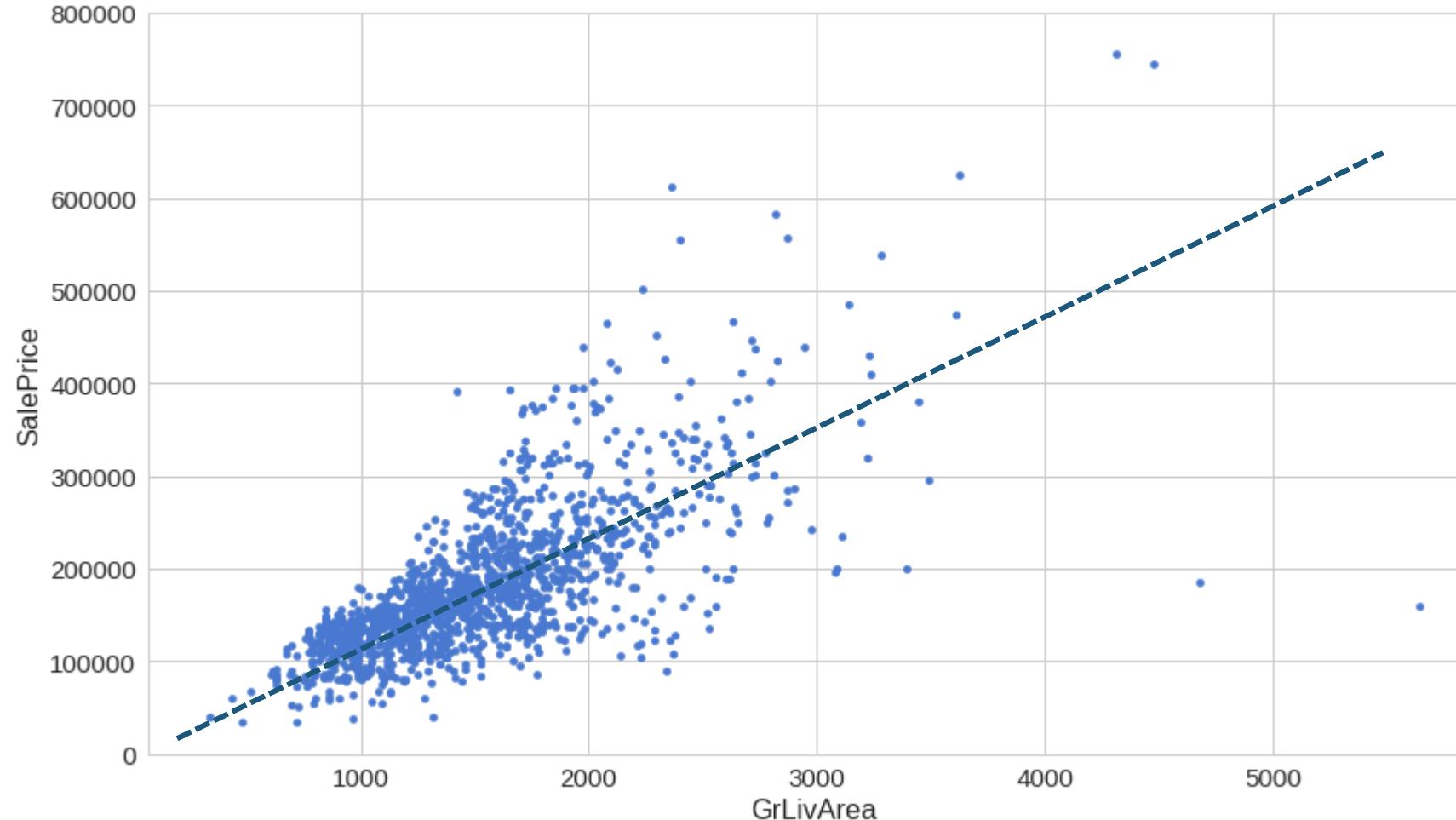
Determinando se um tumor é benígo com base em uma imagem médica

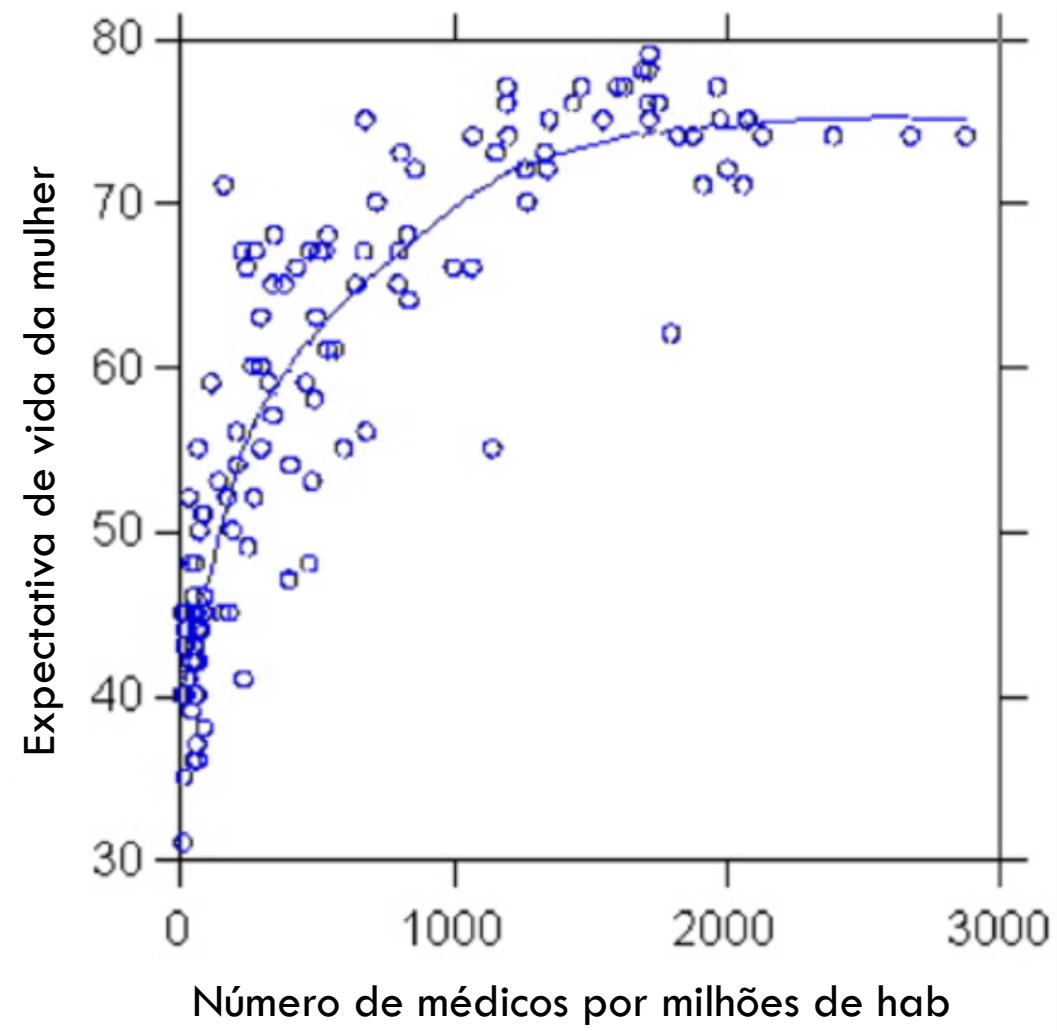


Fonte: <https://www.extremetech.com/extreme/243352-deep-learning-algorithm-diagnoses-skin-cancer-seasoned-dermatologists>



EXEMPLO REGRESSÃO





PEQUENA NOTA SOBRE OS “RÓTULOS” DOS DADOS

Os rótulos de treinamento geralmente são fornecidos através de um julgamento humano. Alguns pontos são importantes na análise da viabilidade de obtenção desses rótulos:

- quantidade de dados necessários;
- nível de conhecimento humano ou conhecimento especializado necessário para fornecer um rótulo preciso;
- complexidade da tarefa de rotulagem.

Existem plataformas *crowdsourcing*, como o Mechanical Turk, Crowd Flower, etc... Ainda há o processo chamado HITL (human-in-the-loop), quando a máquina ou o sistema de computador é incapaz de oferecer uma resposta a um problema, necessitando de uma intervenção humana. Rótulos *implícitos* também são bastante utilizados.

APRENDIZADO NÃO SUPERVISIONADO

O aprendizado não supervisionado é mais difícil quando comparado ao aprendizado supervisionado, porém:

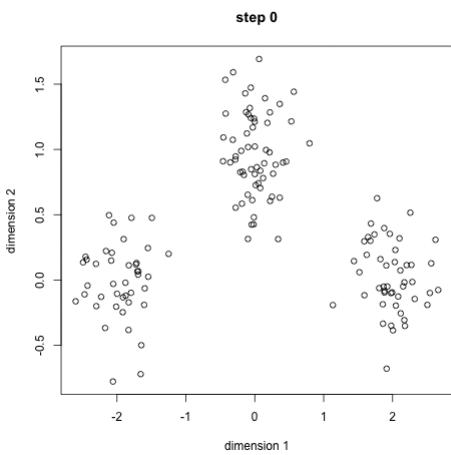
- Rotular grandes conjuntos de dados é muito caro e, portanto, podemos rotular apenas alguns exemplos manualmente;
- Pode haver casos em que não sabemos quantas nem quais classes os dados são divididos;
- Podemos usar a clusterização para obter algumas informações sobre a estrutura dos dados antes de projetar um classificador.

Em aprendizado não supervisionado, desejamos aprender a estrutura inerente de nossos dados sem usar rótulos fornecidos explicitamente. Alguns algoritmos comuns incluem agrupamentos (K-means), análise de componentes principais (PCA).

É somente capaz de agrupar dados semelhantes.

Clusterização

envolve o agrupamento de pontos de dados. Dados que estão no mesmo grupo devem ter características semelhantes, enquanto dados em grupos diferentes devem ter características显著mente diferentes.



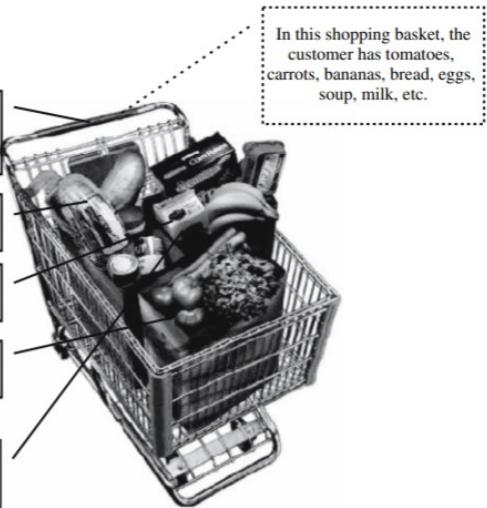
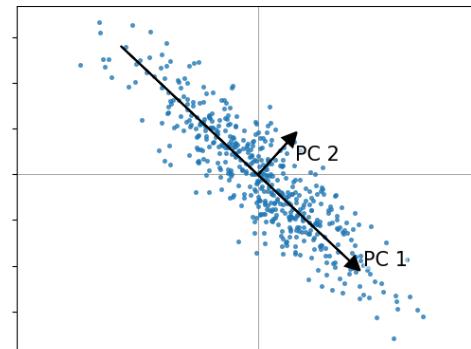
Não supervisionado

Associação

Identifica estruturas, associações (relacionamentos, dependências) em grandes conjuntos de itens de dados

Redução de dimensão

se refere aos métodos usados para representar dados usando menos recursos e entender onde está a maior variância dos dados.





Pesquise assuntos, locais e fontes



Página inicial

Para você

Segundo

Destaques Jornalísticos

Brasil

Mundo

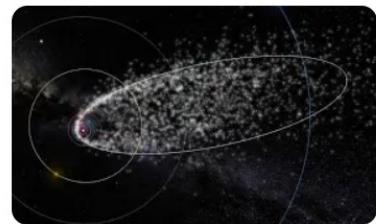
Local

Negócios

Ciência e tecnologia

Entretenime

17 horas atrás



Diário

Chuva de meteoros Eta Aquáridas poderá ser observada no Brasil neste...

Ontem

Superinteressante

Como observar a chuva de meteoros Eta Aquáridas este final de semana
Ontem

UOL

Como ver Eta Aquáridas, chuva de meteoros de fragmentos do cometa Halley
2 dias atrás

CNN Brasil

Chuva de meteoros Eta Aquáridas tem pico neste domingo (5)
Ontem

Cobertura completa

Salvar para ler depois

Compartilhar

Acessar UOL

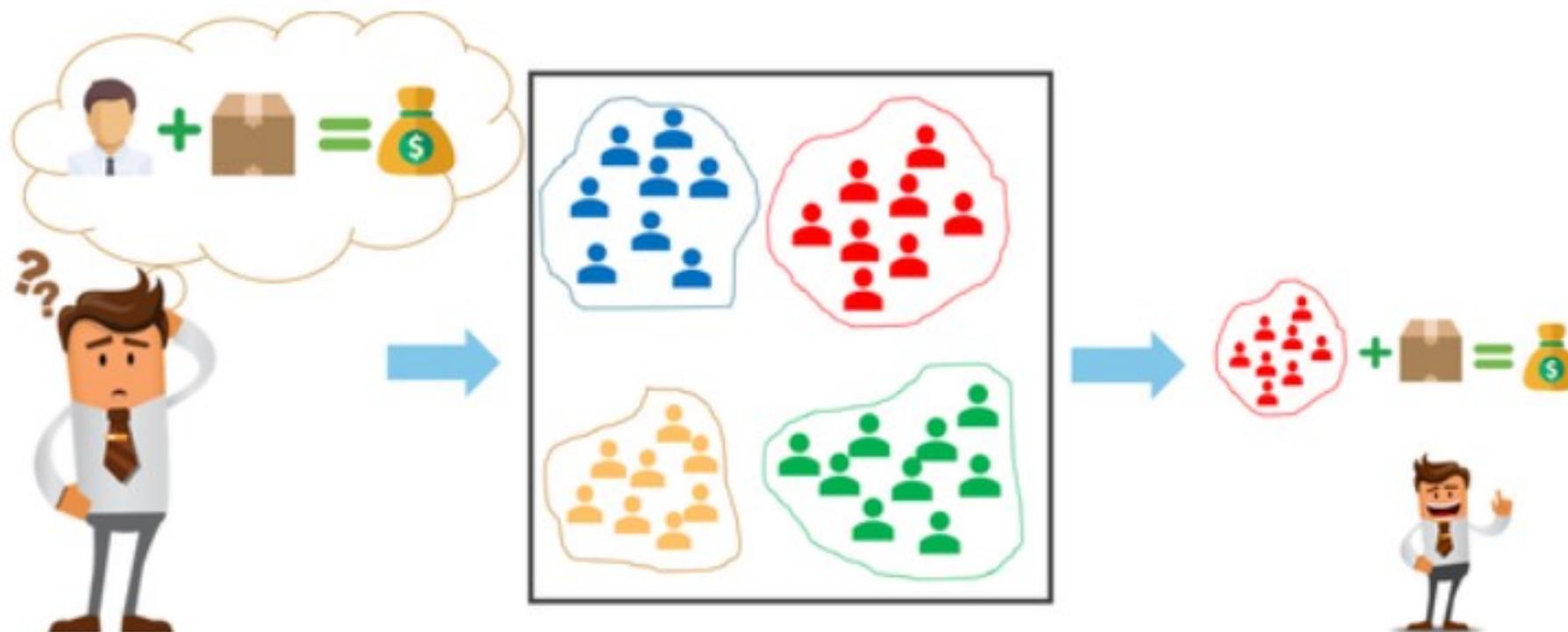
Ocultar todas as reportagens desta fonte: UOL

Outras reportagens como esta

Menos reportagens como esta

O que o Google Notícias faz: analisa dezenas de milhares ou centenas de milhares de novas histórias na Web e as agrupa em notícias coesas.

Segmentando clientes em grupos com preferências semelhantes



Trying to determine the appropriate audience for the product

Using clustering algorithms on the customer base

Selling the product to the targeted audience

<https://www.quora.com/What-is-clustering>

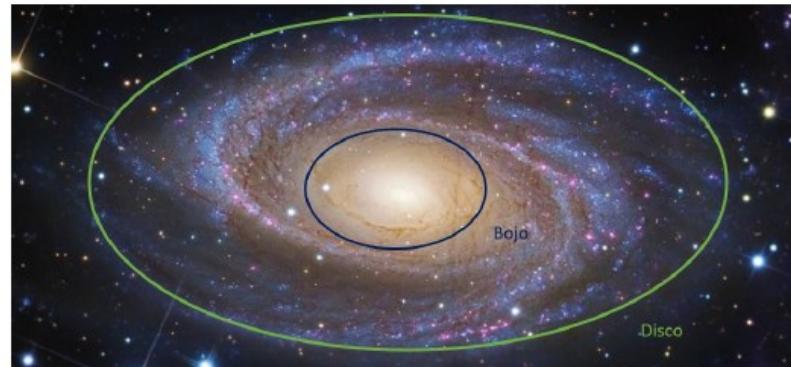


Figura 1 – Esquema básico de uma galáxia espiral. Fonte: adaptado de SOLLIMA et al. (2018).

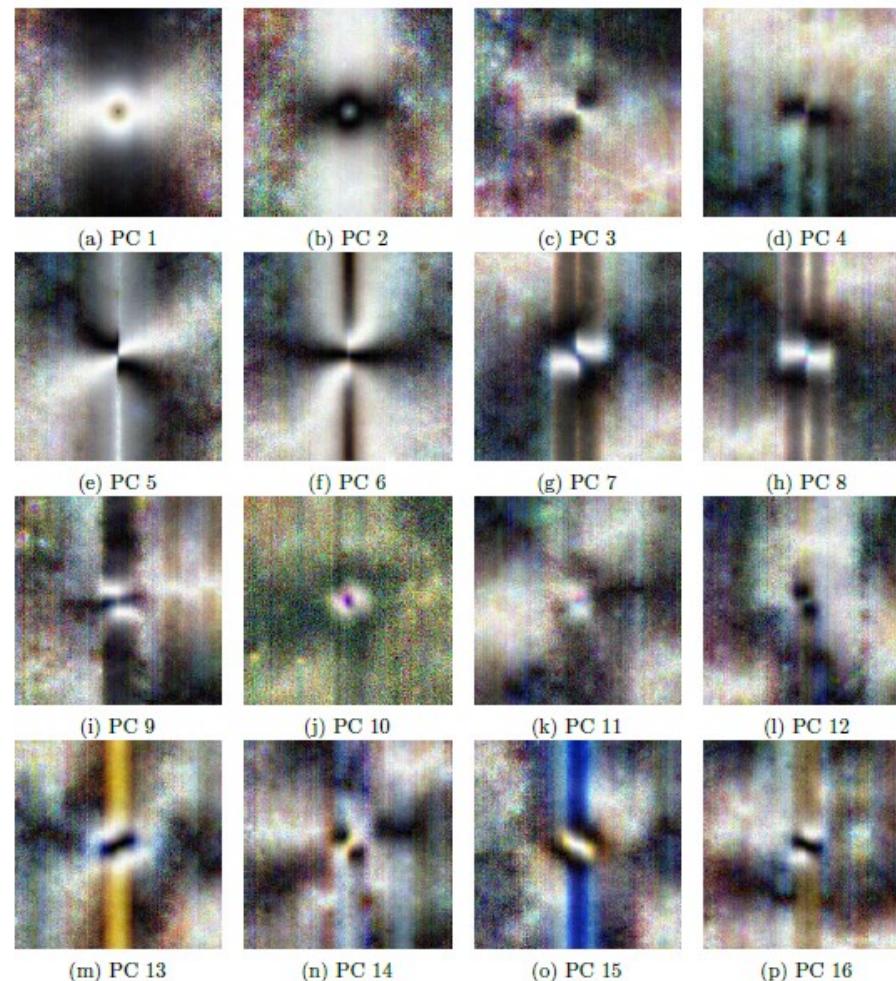


Figura 16 – Representação visual dos 16 primeiros componentes principais.

Sarah Pires Pérez Classificação Morfológica de Galáxias com Machine Learning Trabalho de conclusão do Curso de Engenharia Mecatrônica da Escola Politécnica da Universidade de São Paulo, 2018. Orientação: Thiago Martins.

SEPARAÇÃO CEGA DE SINAIS (BSS)

assume que processos físicos diferentes geram sinais não correlacionados

COCKTAIL PARTY



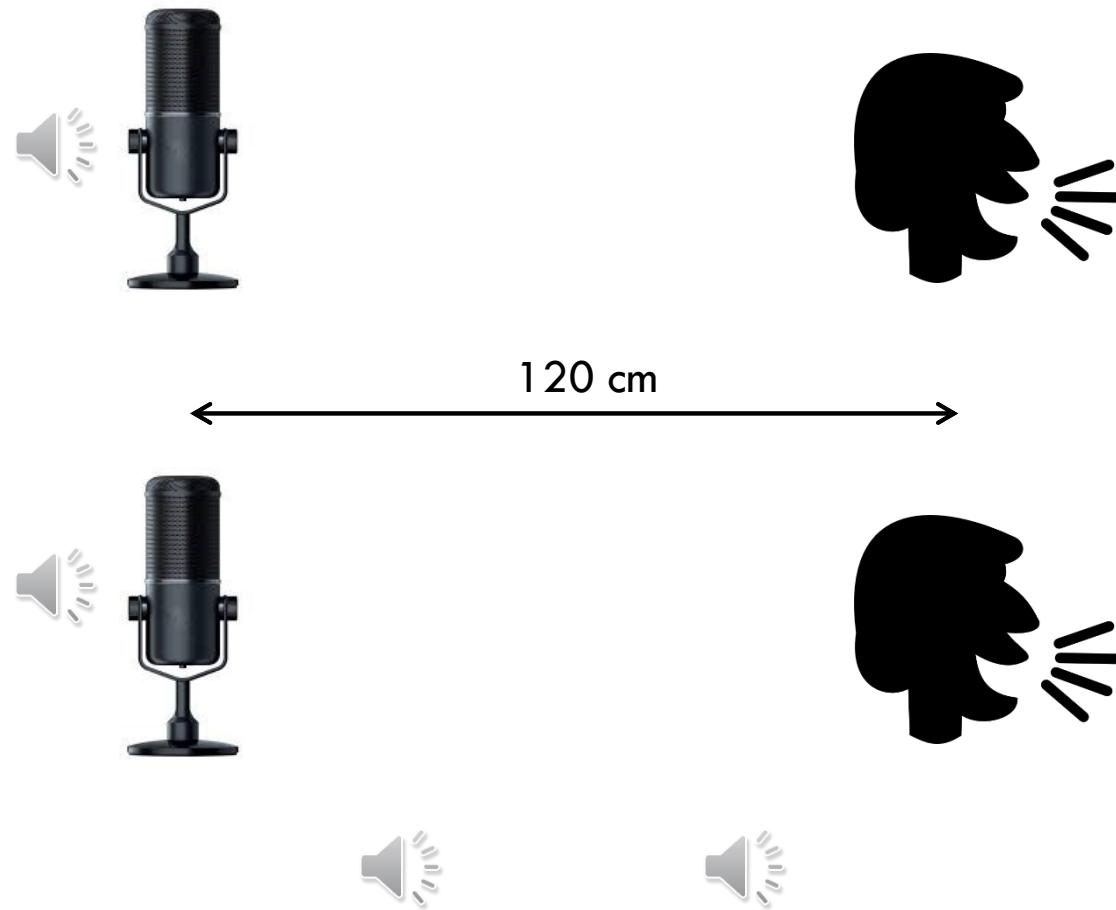
"Existem aspectos da voz de uma pessoa que são distintos do indivíduo, e concentramos nossa atenção nesses recursos para rastrear sua voz em uma sala barulhenta"

Psicólogo Frederic Theunissen, da Universidade da Califórnia, em Berkeley.

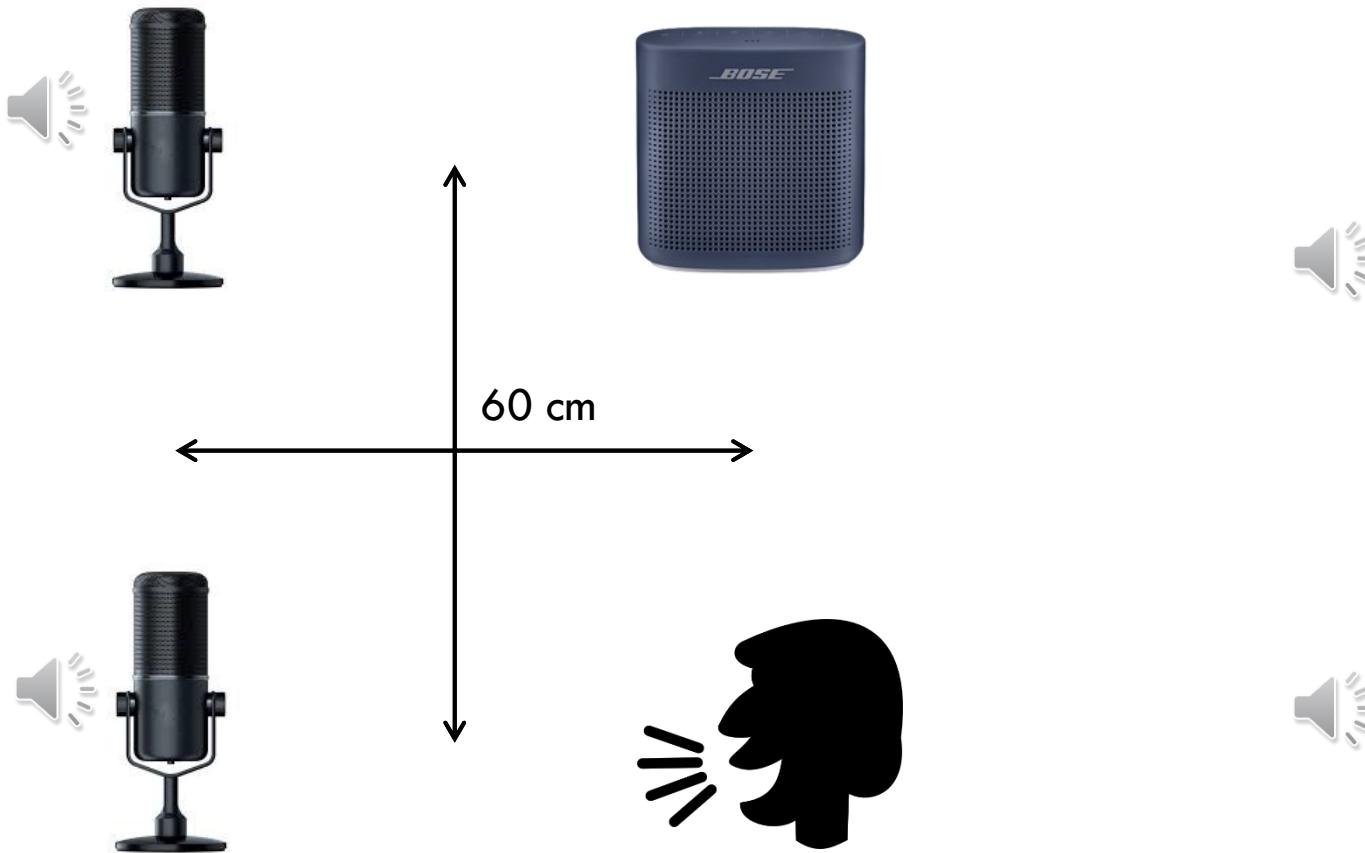
Por exemplo, os ouvintes focam no sotaque, tom e no timbre da voz do falante.

COCKTAIL PARTY

O algoritmo permite encontrar estrutura em um ambiente caótico. Isto é, identificar vozes e músicas individuais a partir de uma malha de sons em uma festa.



https://cnl.salk.edu/~tewon/Blind/blind_audio.html



```
from numpy import *
U, S, Vh = linalg.svd(dot((tile(sum(x*x,0),(x.shape[0],1))*x),x.T))
```

Detecção de anomalias:

Detecção de anomalias é a identificação de itens, eventos ou observações que não estão em conformidade com um padrão esperado ou outros itens em um conjunto de dados.



Processamento de linguagem natural: o processamento de linguagem natural é uma maneira de os computadores analisarem, entenderem e derivarem significado da linguagem humana de maneira inteligente e útil.

O PLN usa o aprendizado de máquina para aprender padrões analisando um conjunto de exemplos (coleção de artigos) e fazer uma inferência estática.

Eu nasci na Itália e, apesar de ter vivido em Portugal e Brazil a maior parte da minha vida, ainda falo fluentemente.

Espero que você possa ver que a resposta mais provável é italiano...

RESOLUÇÃO DE ANÁFORA

Aquele homem não podia levantar o filho nos braços porque **ele** estava muito fraco.

Aquele homem não conseguia levantar o filho nos braços porque **ele** estava muito pesado.

Este par de frases faz parte do WSC - Winograd Schema Challenge, uma competição criada como uma implementação alternativa ao teste de Turing.

Humanos: 96% de acertos; Máquinas: 56% (em 2015) e até 90% (BERT, 2019)

CHATGPT

- I'm sure that my map will show this building; it is very famous.
"It" in this sentence refers to
- In this sentence, "it" refers to the building that the speaker believes will be shown on the map because it is very famous.
- I'm sure that my map will show this building; it is very good.
"It" in this sentence refers to
- In this sentence, "it" refers to the map that the speaker is confident will display the building because the map is very good.
- This book introduced Shakespeare to Ovid; it was a major influence on his writing.
"His" in this sentence refers to
- In this sentence, "his" refers to Shakespeare's writing.
- This book introduced Shakespeare to Goethe; it was a major influence on his writing.
"His" in this sentence refers to
- In this sentence, "his" refers to Shakespeare's writing.
- Please consider the historical facts and correct yourself
- My apologies for the confusion in my previous response. Given the historical context, it is not possible for Shakespeare to be introduced to Goethe, as Shakespeare lived from 1564 to 1616, and Goethe lived from 1749 to 1832. The correct interpretation should be that the book introduced Shakespeare's works to Goethe, and "his" in this sentence would then refer to Goethe's writing.

CREATING FACES



It's Getting Harder to Spot a Deep Fake Video

Progressive Growing of GANs for Improved Quality,
Stability, and Variation, 2017.



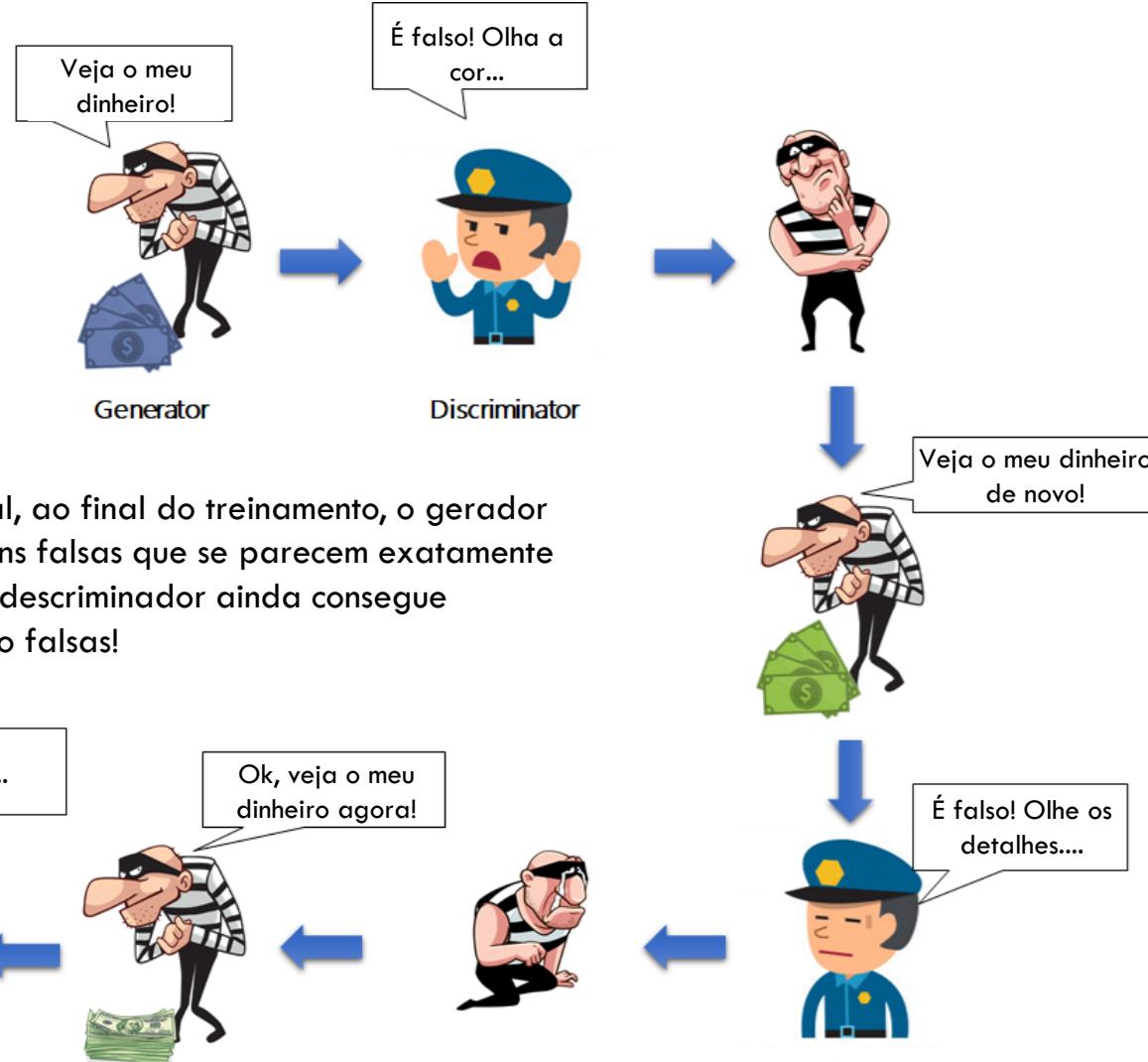
<https://www.thispersondoesnotexist.com/>

GANS – GENERATIVE ADVERSARIAL NETWORKS

O núcleo de uma GAN é composto de duas redes: uma geradora e outra descriminadora, que representam dois agentes competindo entre si. A rede geradora tenta enganar a descriminadora gerando imagens falsas que se parecem com as imagens reais, enquanto a descriminadora tenta detectar a imagem falsa. Uma aprende com a outra ...



Em uma situação ideal, ao final do treinamento, o gerador pode produzir imagens falsas que se parecem exatamente como as reais, mas o descriminador ainda consegue perceber que elas são falsas!



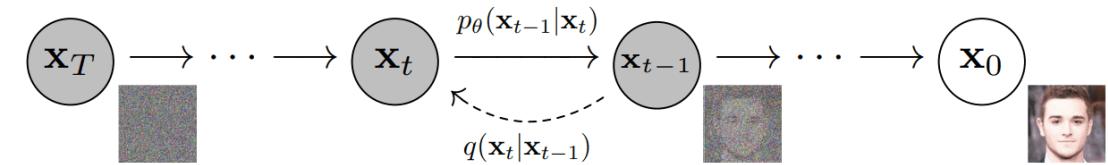
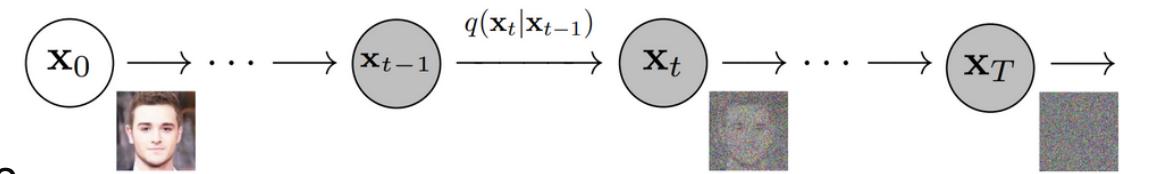
MÉTODOS DE DIFUSÃO

Começamos com um conjunto de dados existente, e a técnica é dividida em:

1. Difusão Direta: Este é o primeiro passo. Aqui, gradualmente "difundimos" ou "espalhamos" a informação nos dados existentes. Imagine isso como misturar um líquido em um copo de água. Isso "quebra" a estrutura original dos dados.

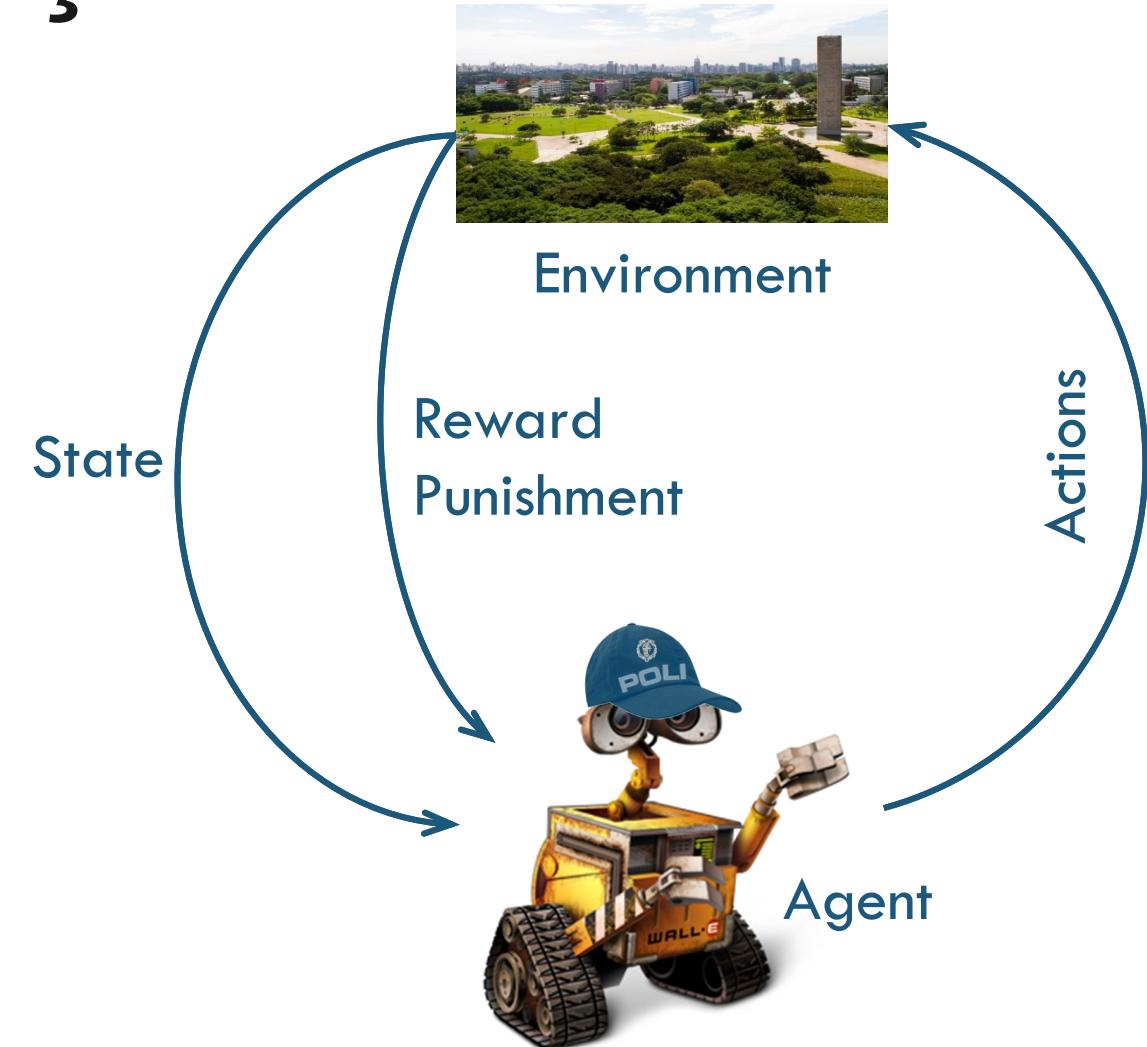
2. Aprendizado da Difusão Reversa: Em seguida, aprendemos a reverter o processo de difusão. Ou seja, tentamos descobrir como "desfazer" o que fizemos na etapa anterior, reconstruindo lentamente a estrutura original dos dados.

Uma vez que aprendemos a reverter o processo, agora temos um modelo que pode gerar novos dados semelhantes aos dados originais. É flexível porque pode gerar uma variedade de dados diferentes, mantendo as características importantes dos dados originais.



APRENDIZADO POR REFORÇO

O método de aprendizado por reforço modela a interação entre dois elementos - o ambiente e o agente de aprendizado.





“A IA é um risco fundamental para o futuro da civilização humana de uma maneira que não foram acidentes de carro, acidentes de avião, drogas defeituosas ou comida ruim. Eles eram prejudiciais para um conjunto de indivíduos na sociedade, mas não eram prejudiciais para todos os indivíduos.”

CEO da SpaceX, Elon Musk

“Sucesso na criação de IA efetiva, poderia ser o maior evento da história da nossa civilização. Ou o pior. Nós simplesmente não sabemos. Portanto, não podemos saber se seremos infinitamente ajudados pela IA, ou ignorados por ela e forrados de lado, ou concebivelmente destruídos por ela.”

“A menos que aprendamos como nos preparar e evitar os riscos potenciais, a IA poderia ser o pior acontecimento na história de nossa civilização. Traz perigos, como poderosas armas autônomas, ou novas maneiras para os poucos oprimirem muitos. Isso pode trazer grandes transtornos para nossa economia, [...]”

“Sou otimista e acredito que podemos criar IA para o bem do mundo. Que ela possa funcionar em harmonia conosco. Simplesmente precisamos estar cientes dos perigos, identificá-los, empregar as melhores práticas e gerenciamento possíveis, e prepare-se para suas consequências com bastante antecedência”, Hawking.



EPA... ESTAMOS PREVENDO O FUTURO NOVAMENTE???

Elon Musk



“Os humanos precisam se tornar ciborgues para se manterem relevantes em um futuro dominado pela inteligência artificial. Haverá menos e menos empregos que um robô não pode fazer melhor ”, 2017.

“[...] todas as nações desenvolvidas do mundo verão taxas de perda de empregos de até 47% nos próximos 25 anos,[...]”,
School of Business da Universidade da Pensilvânia, 2017.

Art Bilger



“Qualquer pessoa viva hoje que sobreviva até 2050 talvez nunca precise enfrentar a morte. Em 2050, será realmente apenas para os ricos e famosos. A maioria das pessoas com renda da classe média e com renda razoável da classe trabalhadora provavelmente pode pagar isso nos anos 2060.”, 2016.

Dr. Ian Pearson, engenheiro, futurista



“Os robôs podem superar os HUMANOS em apenas 15 anos e sentirão ‘emoções genuínas’ até 2028”, 2016.

intervenções neurológicas? Afinal, não seria a tristeza acumulada o que nos faz vergar sob o peso da velhice?

O paradoxo da imortalidade é que, sem a finitude, a vida perde seu significado. Atingir a imortalidade nunca poderia ser um fato, mas apenas um desejo. Embora a morte não faça parte da vida por estar além da experiência consciente, a finitude é um dos componentes essenciais da vida. Mais do que morrer, o importante é *saber* que morreremos. A brevidade da vida e a indefinição sobre quando ela terminará faz com que lutemos para torná-la o mais digna possível. Uma vida infinita ou extremamente longa é incompatível com a dignidade e com o desejo de aproveitá-la da melhor forma possível. O que é muito abundante acaba perdendo valor.



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 (se existir...) entre conjuntos de dados.

Em Aprendizado de Máquina, essa equação é obtida através de um *algoritmo de treinamento* utilizando os m dados $(x^{(i)}, y^{(i)}), i = 1, \dots, m$.

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**



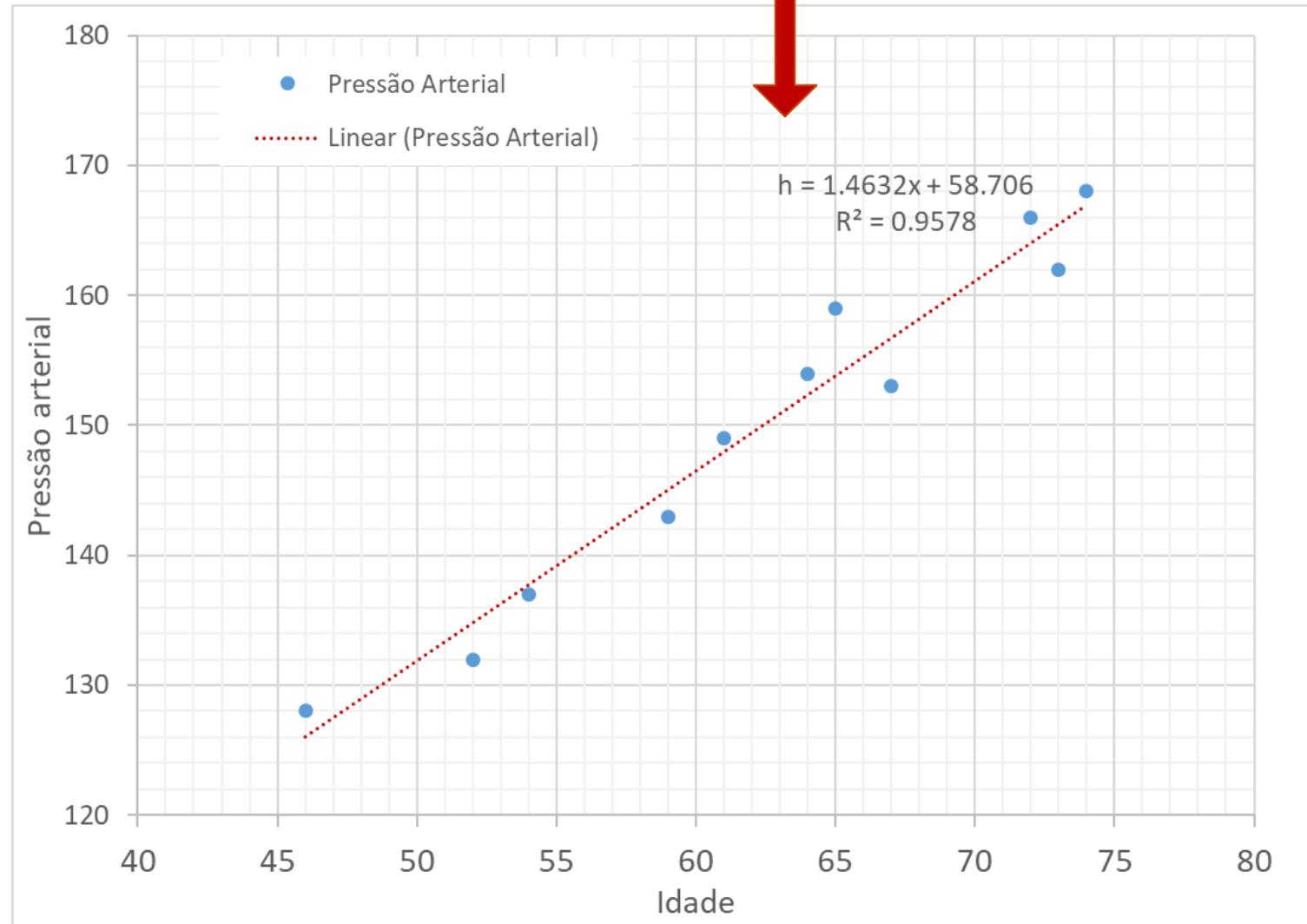
x: Variável
independente y: variável
dependente

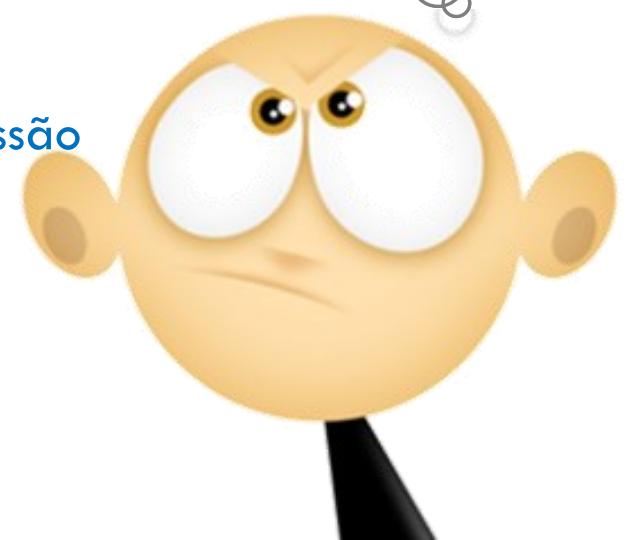
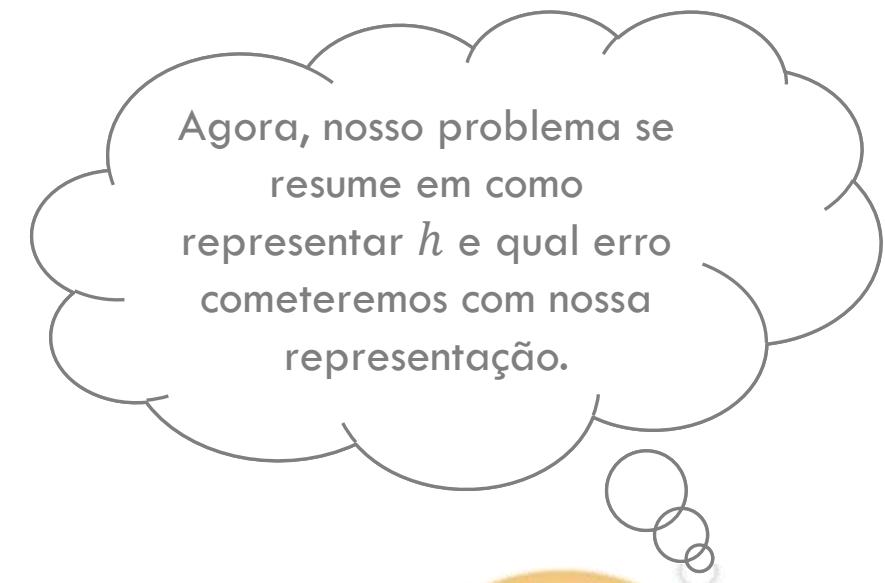
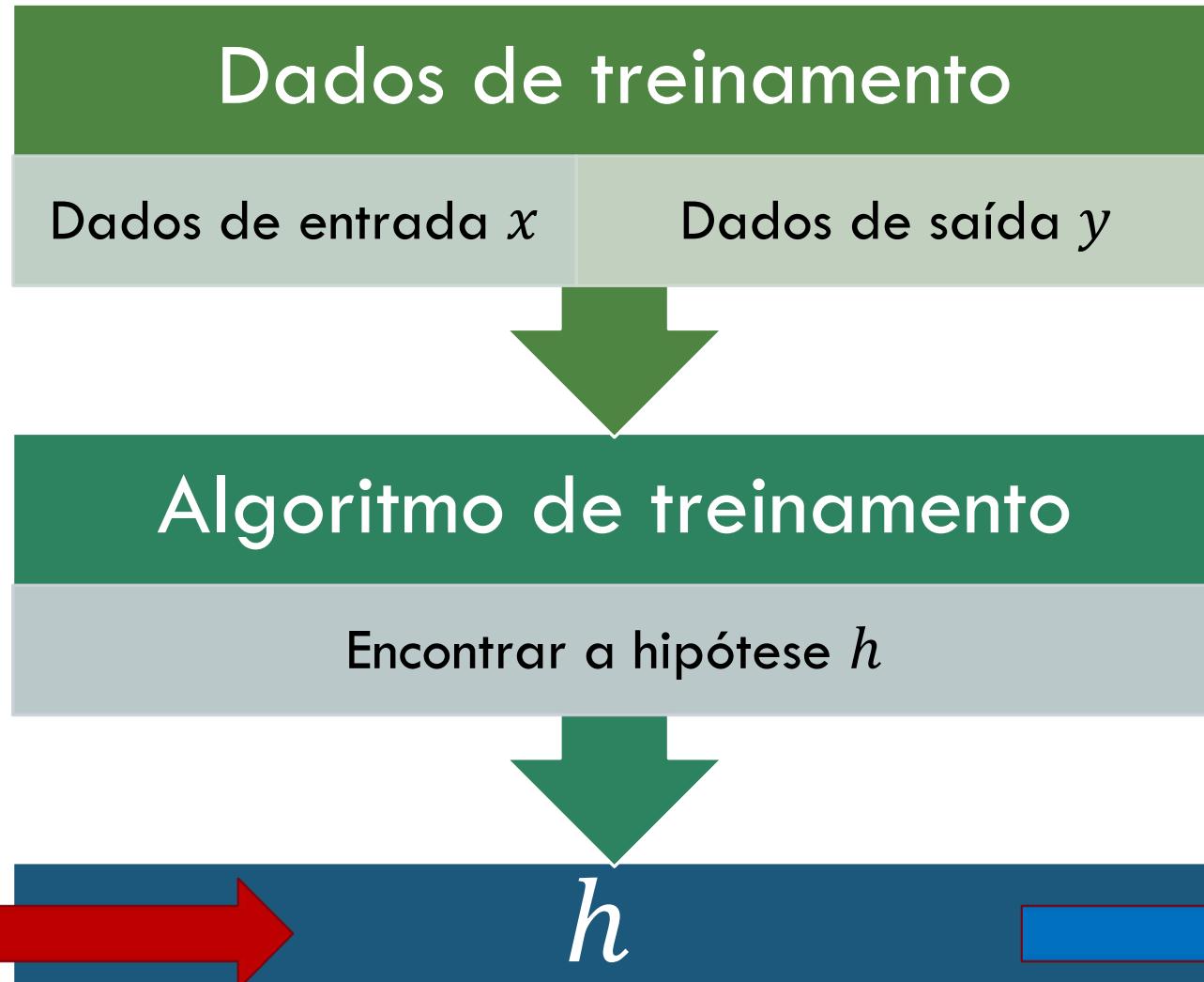
**Pressão
Arterial**

Idade	
52	132
59	143
67	153
73	162
64	154
74	168
54	137
61	149
65	159
46	128
72	166

Hipótese h mapeia x em y .

HIPÓTESE h

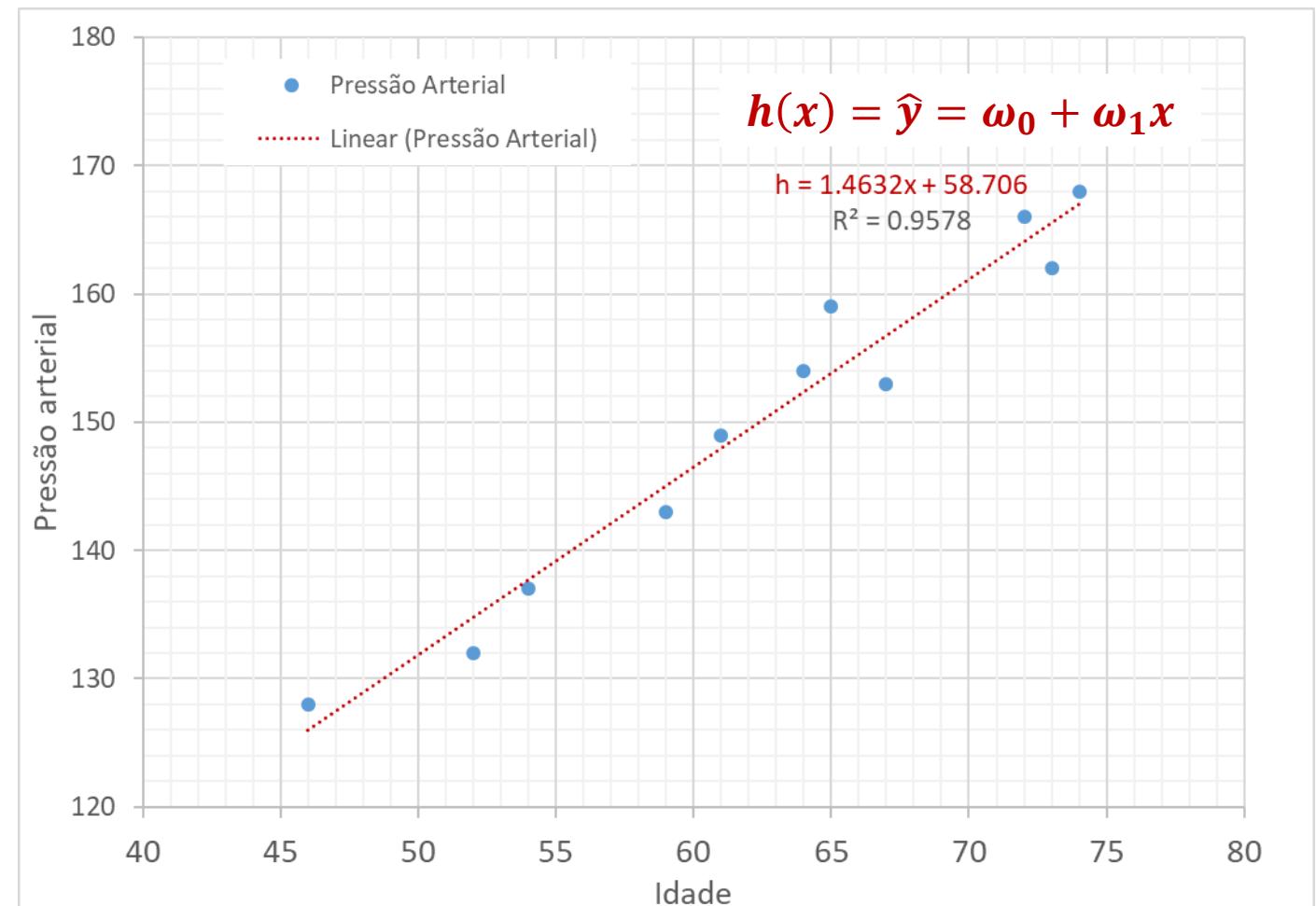


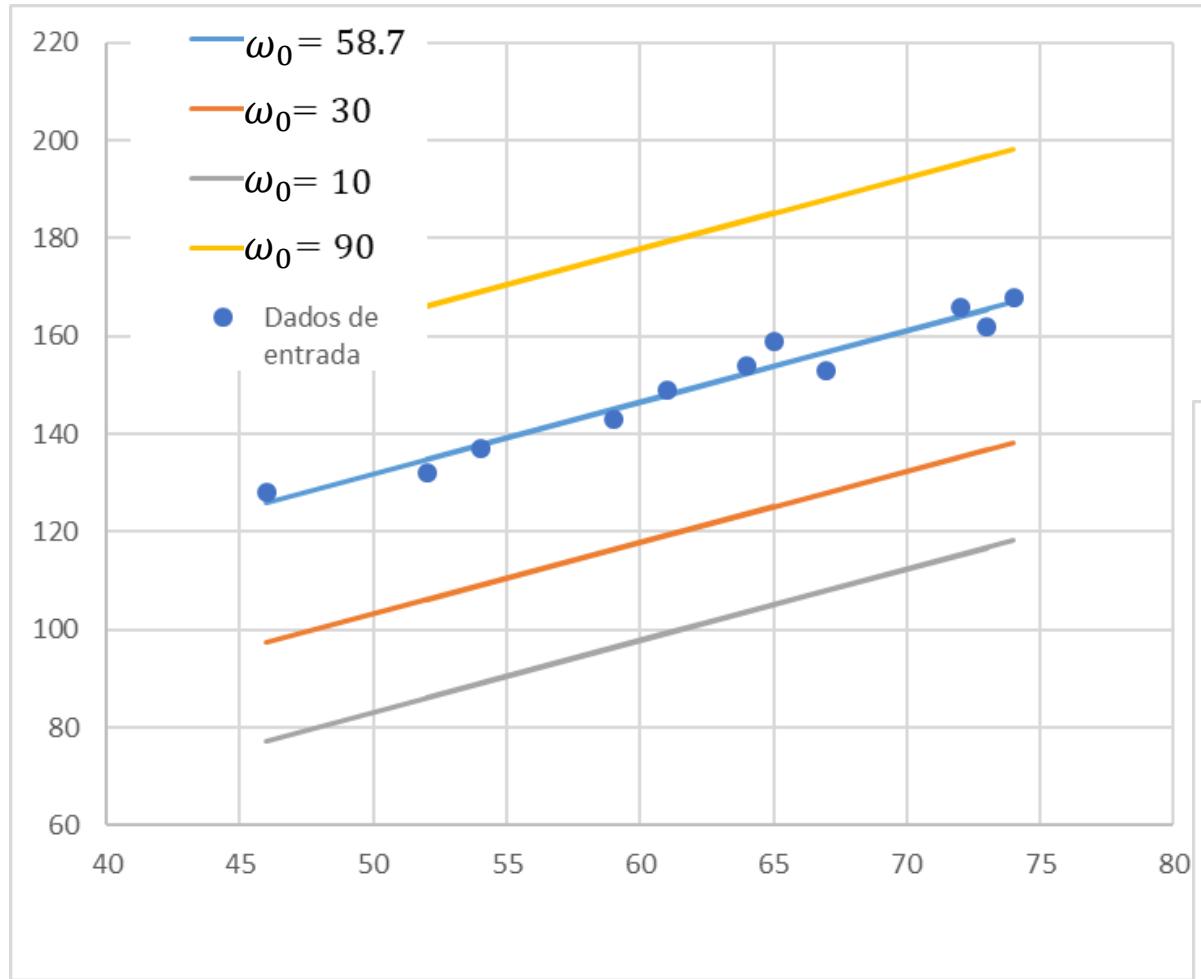


REGRESSÃO LINEAR SIMPLES

Tenho $m = 11$ dados de treinamento (\bullet),
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



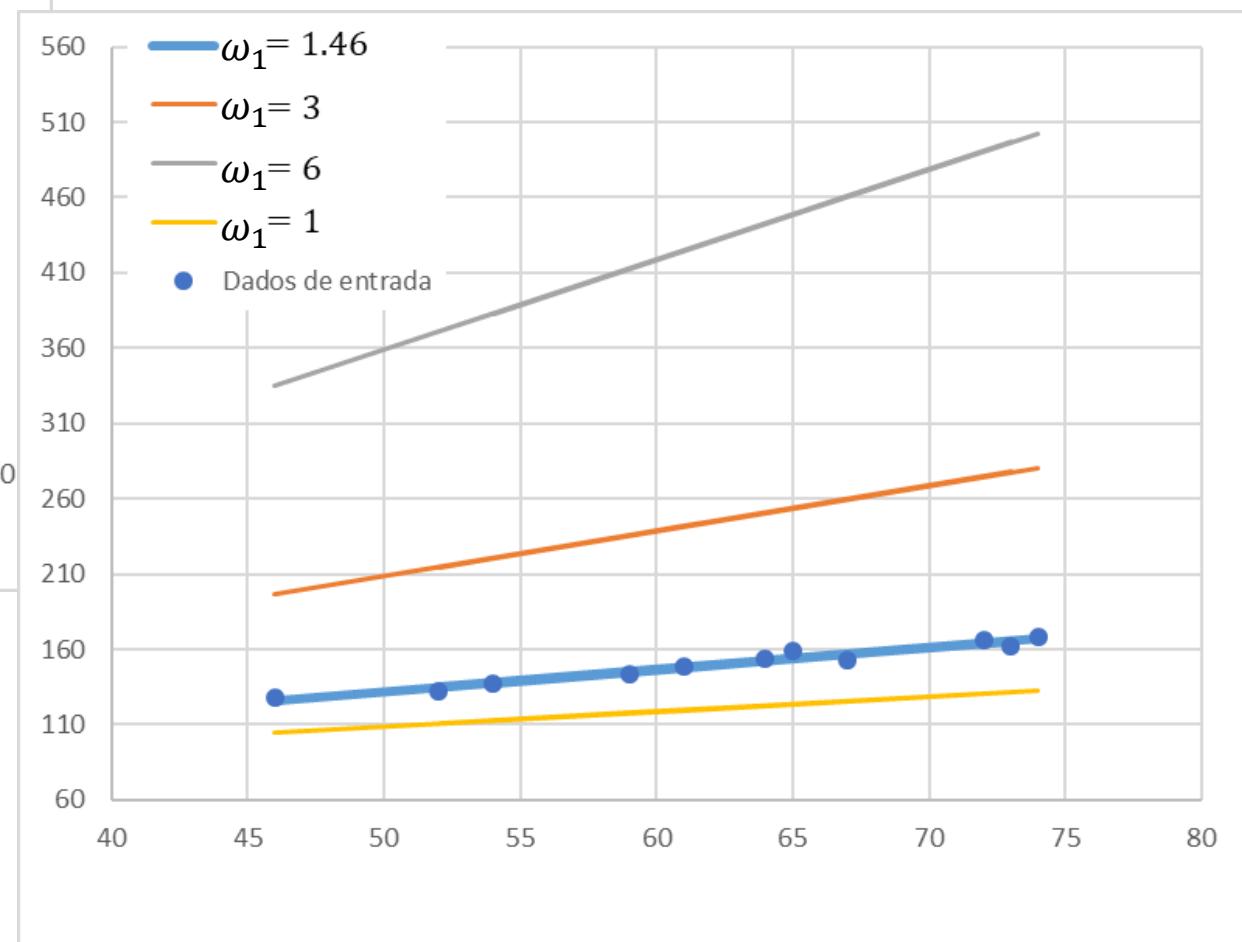


ω_0 e ω_1 serão definidos através dos meus m dados.

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???

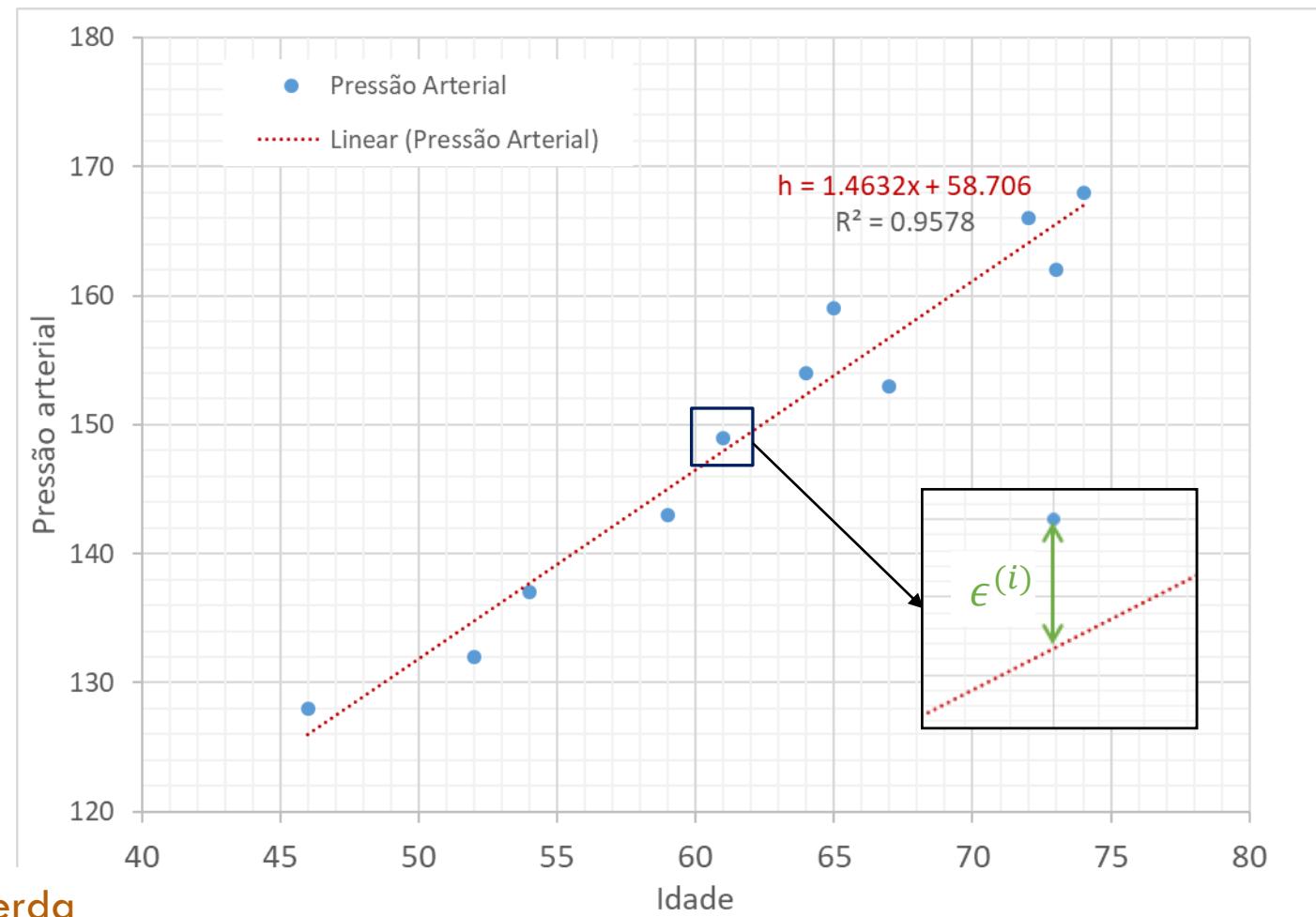
Define-se o erro no conjunto de dados i como a distância entre a resposta da minha hipótese, $\hat{y}^{(i)}$, e a resposta exata $y^{(i)}$,

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

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

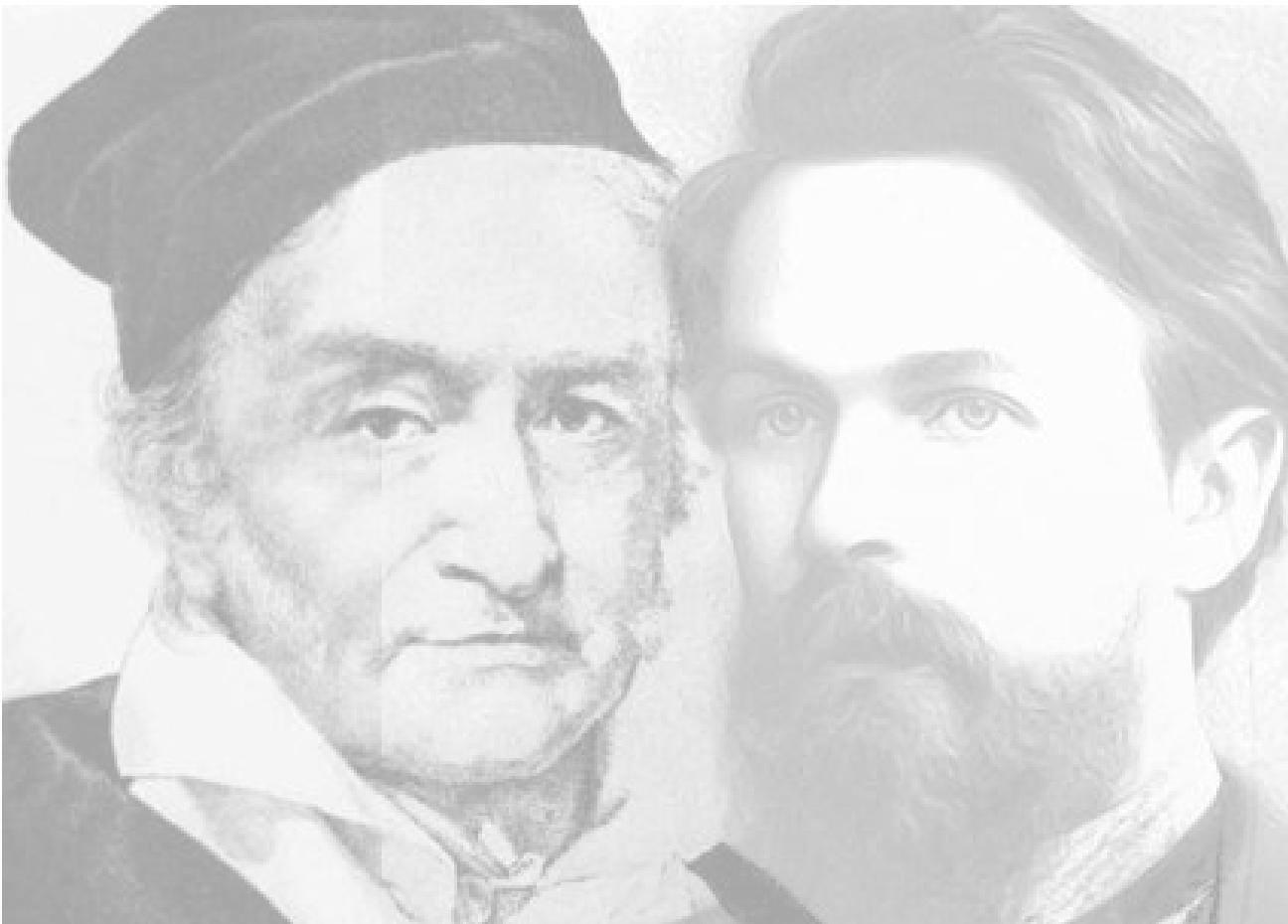
$$J(\omega_0, \omega_1) = \frac{1}{2m} \sum_{i=1}^m [\epsilon^{(i)}]^2$$

Função perda
(Loss function)



TEOREMA DE GAUSS MARKOV

O teorema de Gauss Markov nos diz que, se um determinado conjunto de suposições for atendido, a estimativa de mínimos quadrados ordinários para coeficientes de regressão fornecerá a melhor estimativa linear não viésada (BLUE, do inglês *best linear unbiased estimate*) possível.



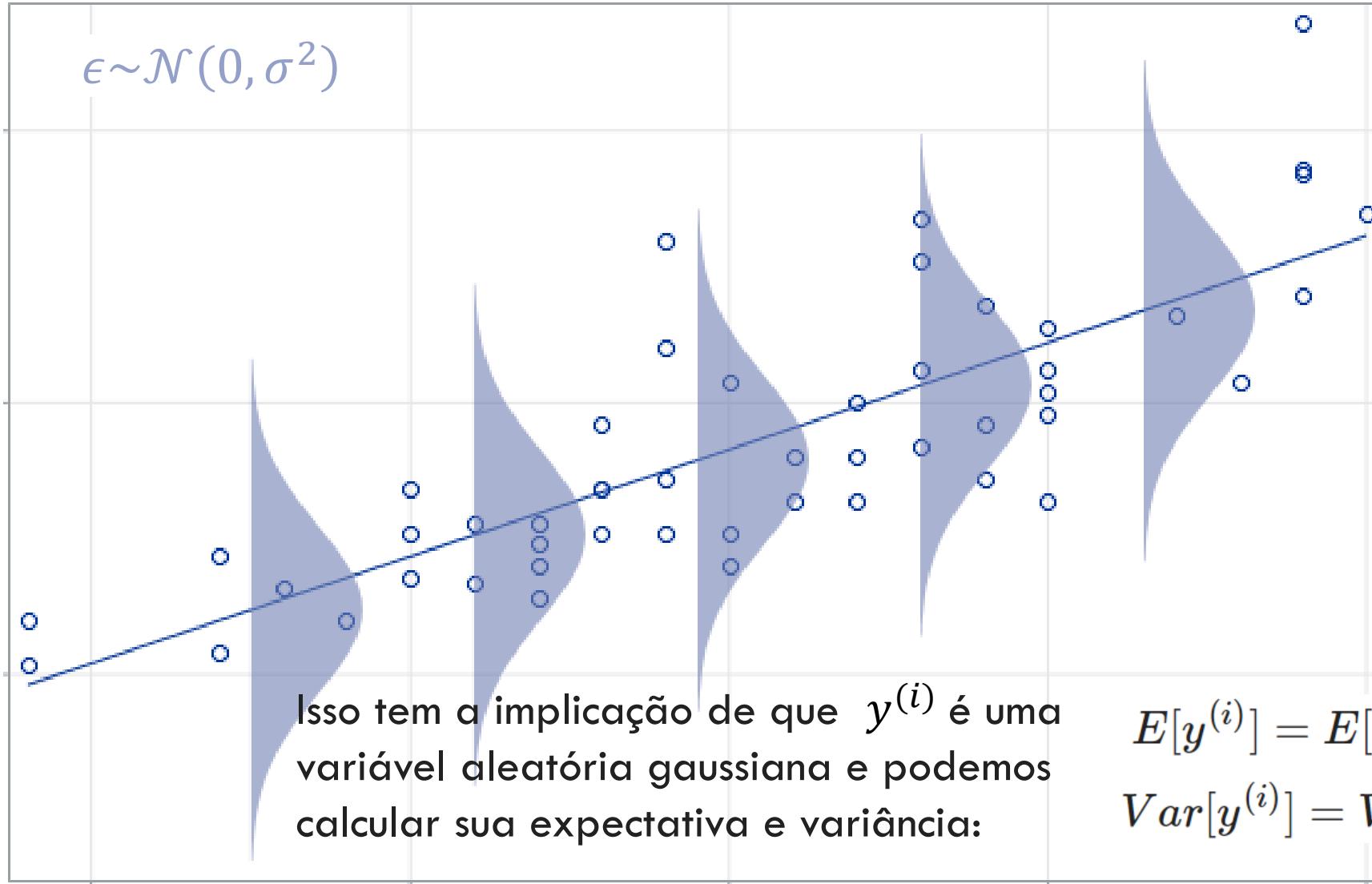
ALGEBRA E GAUSS MARKOV

Um modelo de regressão linear representado por,

e gerado pela estimativa dos mínimos quadrados ordinários é possivelmente BLUE se

- $\mathbb{E}(\epsilon_i) = 0, \forall i$
 - $Var(\epsilon_i) = \sigma^2 < \infty, \forall i$
 - $Cov(\epsilon_i, \epsilon_j) = 0, \forall i \neq j$

Essencialmente, nossos dados têm uma relação linear que
 é corrompida pelo ruído gaussiano aleatório que tem
 média zero e variância constante.



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

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

FUNÇÃO CUSTO

i	$x^{(i)}$	$y^{(i)}$	$\hat{y}^{(i)}$	$[\epsilon^{(i)}]^2$
1	52	132	134.79	7.80
2	59	143	145.04	4.14
3	67	153	156.74	13.99
4	73	162	165.52	12.39
5	64	154	152.35	2.72
6	74	168	166.98	1.03
7	54	137	137.72	0.52
8	61	149	147.96	1.08
9	65	159	153.81	26.89
10	46	128	126.01	3.95
11	72	166	164.06	3.78
$J =$				3.56

Se o objetivo é cometer o menor erro possível, então, o que queremos é ...

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

PROBLEMA DEFINIDO

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

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

$\boldsymbol{x} = \begin{bmatrix} x_0 = 1 \\ x_1 \end{bmatrix}$ $x_0 = 1, x_1$ arranjados em um vetor \boldsymbol{x}

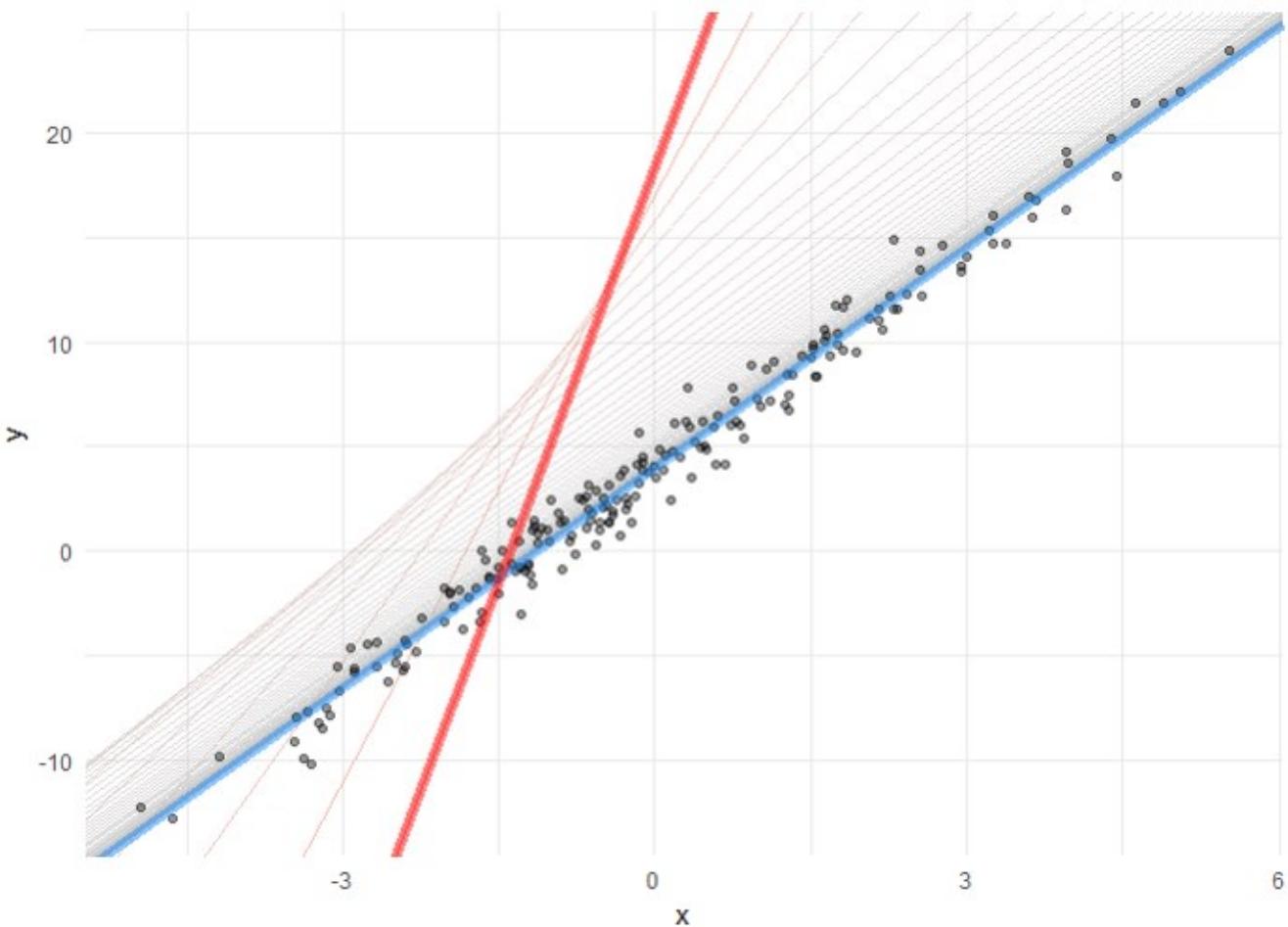
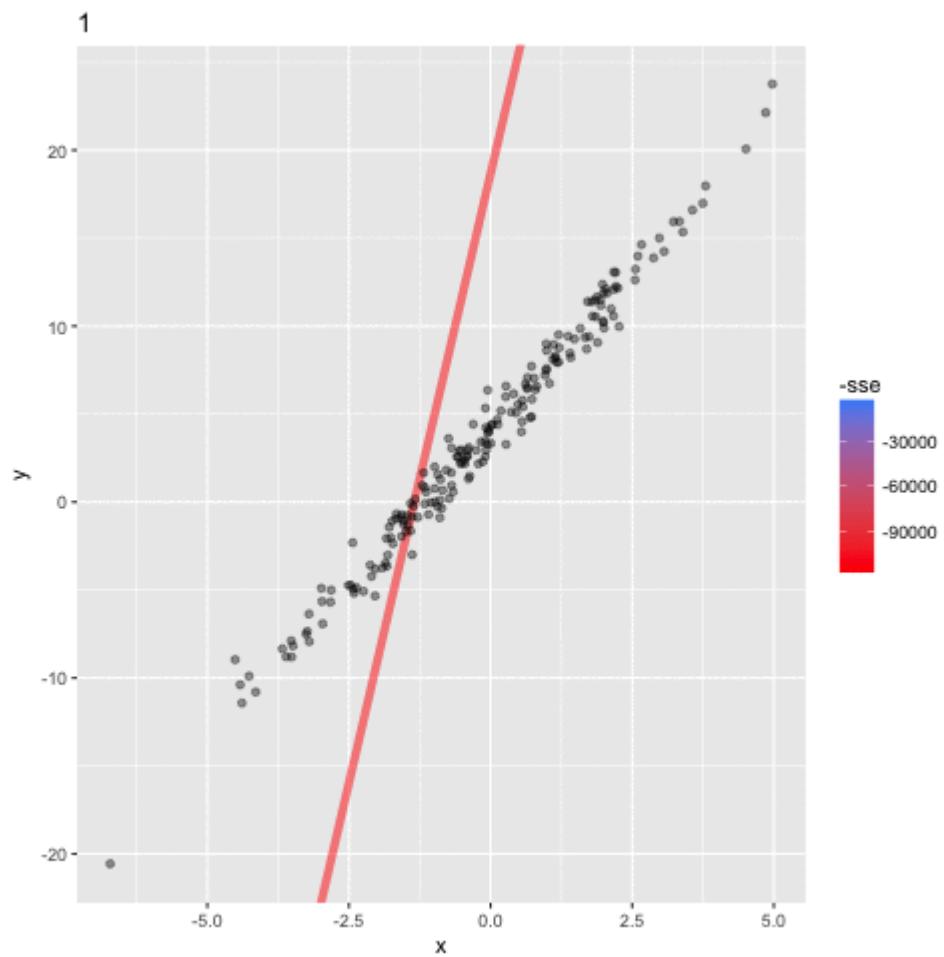
FUNÇÃO CUSTO: $J(\omega_0, \omega_1) = \frac{1}{2m} \sum_{i=1}^m [\epsilon^{(i)}]^2$, $\epsilon^{(i)} = \hat{y}^{(i)} - y^{(i)}$

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

“A problem well stated is a problem half solved.”

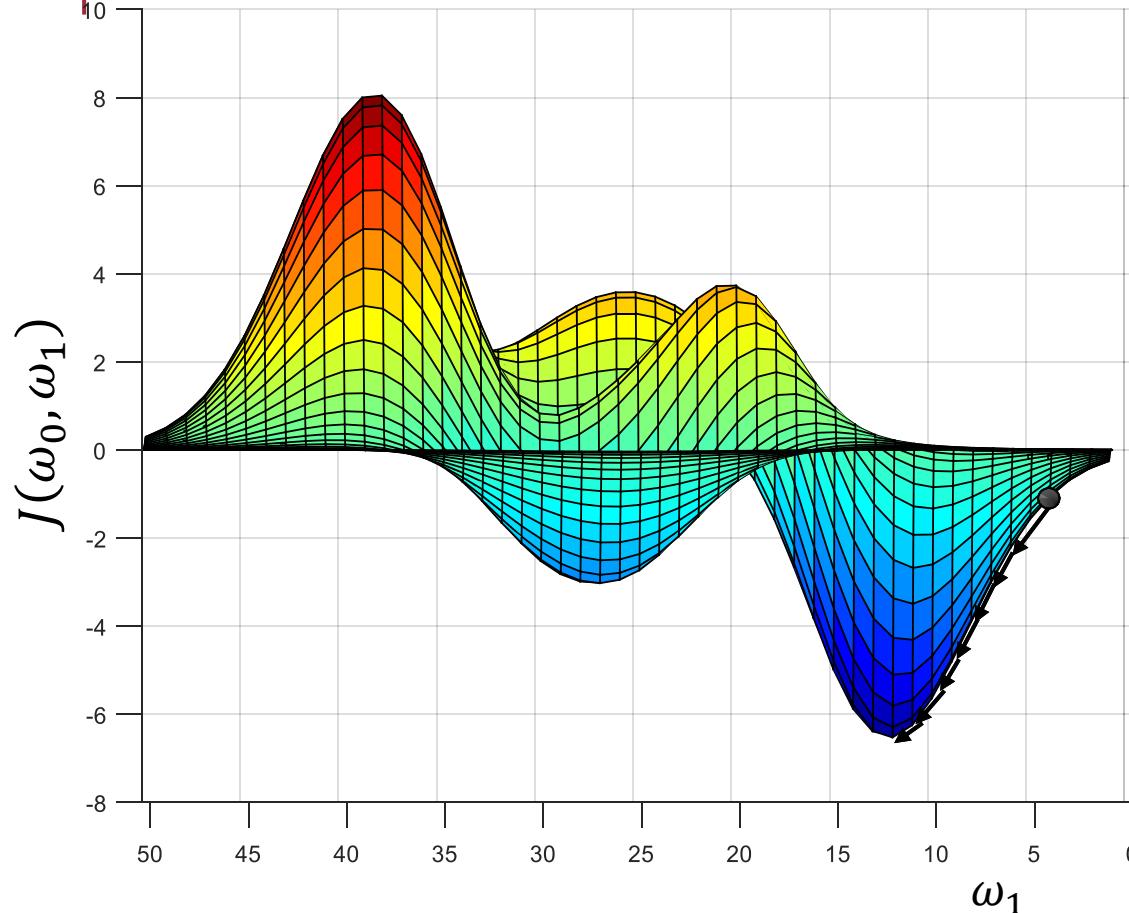
Charles F. Kettering

INTUIÇÃO DE NOSSO PROBLEMA DE REGRESSÃO



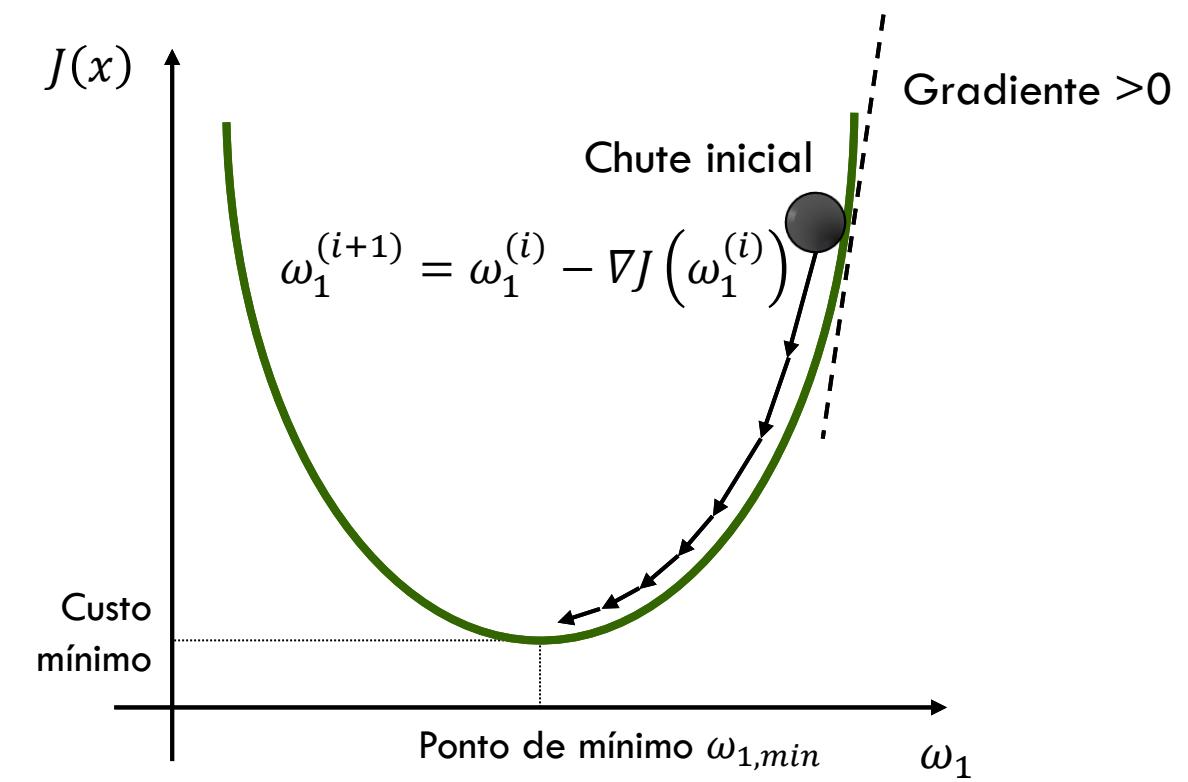
Fonte: <https://towardsdatascience.com/machine-learning-fundamentals-via-linear-regression-41a5d11f5220>

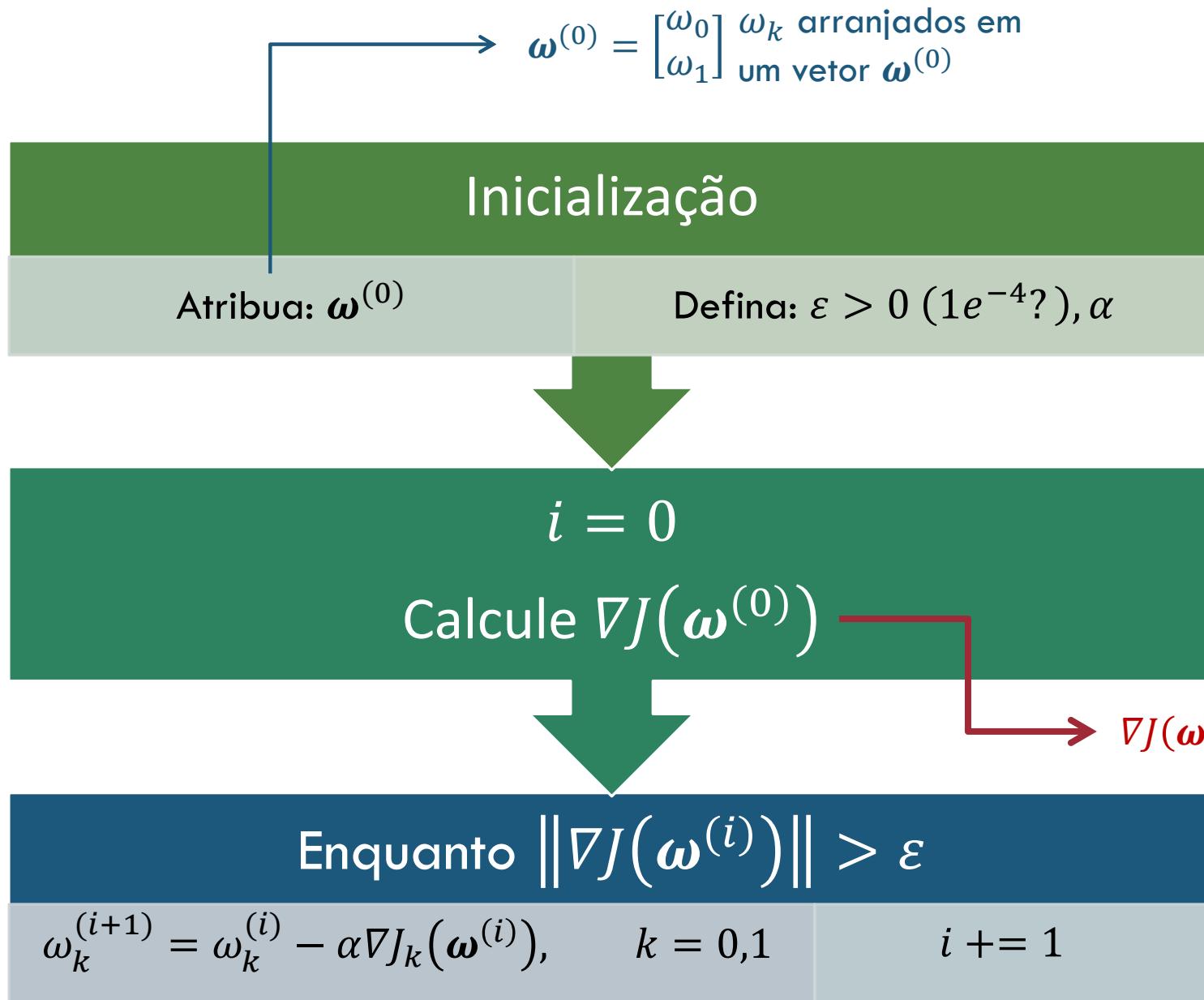
GRADIENTE DESCENDENTE



$$\omega_1^{(i+1)} = \omega_1^{(i)} - \alpha \nabla J(\omega_1^{(i)})$$

Supondo $\omega_0 = 0$ para ilustrar de forma clara o algoritmo





ALGORITMO

repetir até convergência{

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

}

Update simultâneo:

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

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

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

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

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

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

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

ALGORITMO DE GRADIENTE DESCENDENTE

Repita até convergência {

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

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

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

}

Update ω_0, ω_1 simultaneamente!

OU...

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

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

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

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

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

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

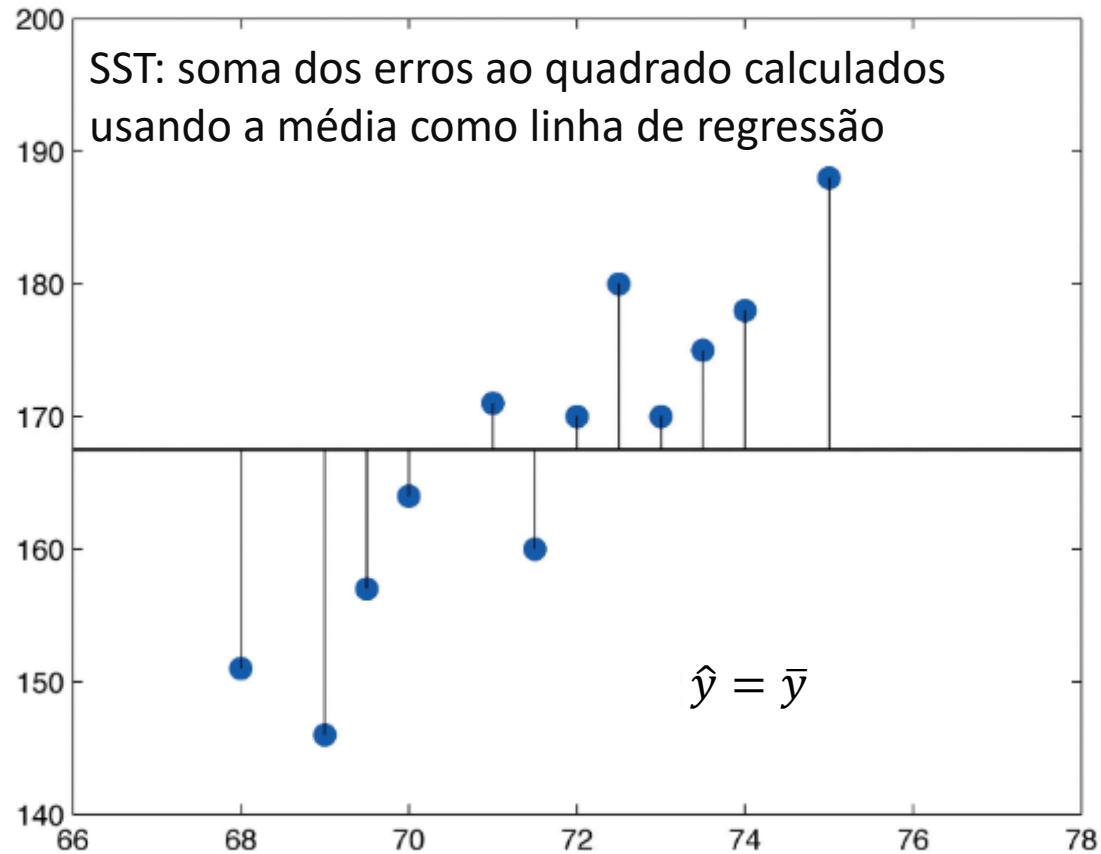
$$s_{xy} = \sum_i (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$$

$$w_1 = \frac{s_{xy}}{s_{xx}} \qquad \qquad \sum_i [\epsilon^{(i)}]^2 = s_{yy} \left(1 - \frac{s_{xy}^2}{s_{xx}s_{yy}} \right)$$

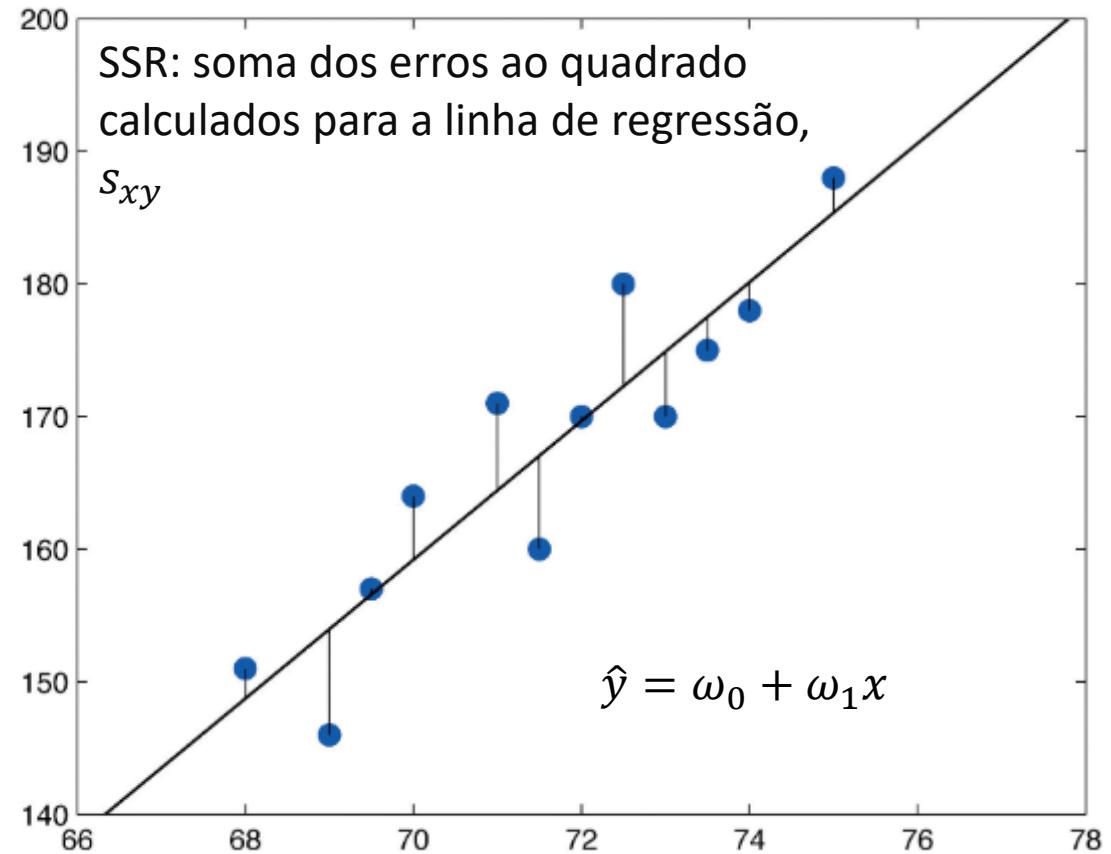
$$w_0 = \bar{y} - w_1 \bar{x}$$

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

R^2



$$\hat{y} = \bar{y}$$



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

Uma medida de quão útil é usar a equação de regressão para a previsão de y é o quanto menor SSR é do que SSE.

R^2

Vamos ao Notebook

R^2 (R-quadrado) é uma medida estatística de quão próximos os dados estão da linha de regressão ajustada. Também é conhecido como coeficiente de determinação.

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}}, \quad \text{Var(regressao)} = \sum_{i=1}^m (y_i - \text{regressao})^2$$

$$= 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = \frac{s_{xy}^2}{s_{xx} s_{yy}}$$

$$\text{Var(media)} = \sum_{i=1}^m (y_i - \text{media})^2$$

$$R^2 = \frac{\text{Var(media)} - \text{Var(regressao)}}{\text{Var(media)}}$$

COMO FICA NOSSO PROBLEMA DE IDADE VS PRESSÃO ARTERIAL?

i	x Idade (variável de entrada)	y 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

$$m = 11$$

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

$$\bar{x} = \frac{52+59+67+73+64+74+54+61+65+46+72}{11} = \frac{687}{11} = 62,45$$

$$\bar{y} = \frac{132+143+153+162+154+168+137+149+159+128+166}{11} = 150,09$$

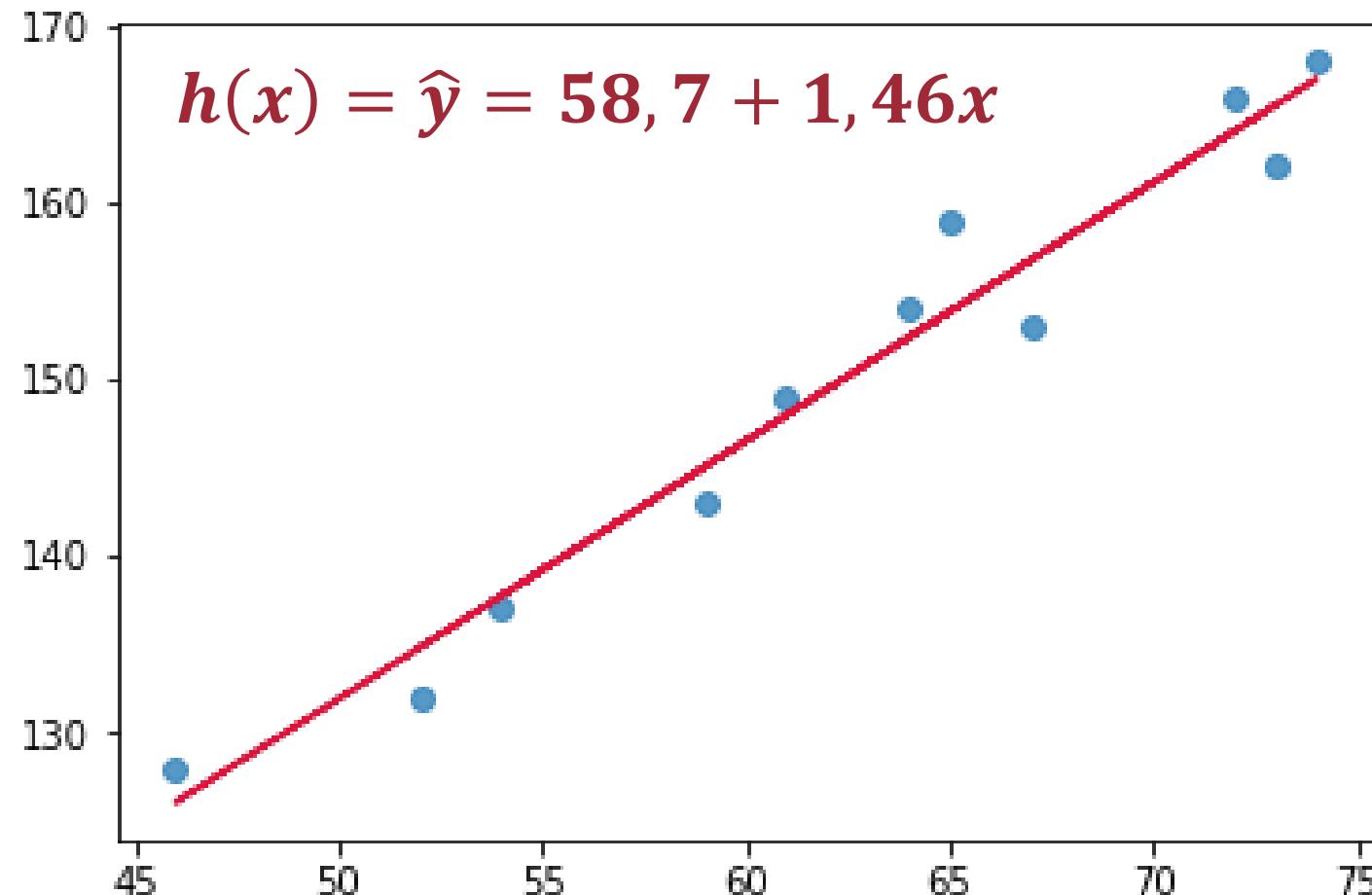
$$s_{xx} = \sum_{i=1}^m (x_i - \bar{x})^2 = \sum_{i=1}^{11} (x_i - 62,45)^2 = 830,73$$

$$s_{xy} = \sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{11} (x_i - 62,45)(y_i - 150,09) = 1215,54$$

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

$$\omega_1 = \frac{s_{xy}}{s_{xx}} = \frac{1215,54}{830,73} = 1,46$$

$$\omega_0 = \bar{y} - \omega_1 \bar{x} = 150,09 - 62,45 \times 1,46 = 58,7$$



VAMOS ao Notebook

$$\begin{aligned}\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_0} &= \frac{1}{m} \sum_{i=1}^m [h(x^{(i)}) - y^{(i)}] = \frac{1}{11} \sum_{i=1}^{11} \left[[0 \quad 0,5] \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix} - y^{(i)} \right] \\ &= \frac{1}{11} \left\{ \left[[0 \quad 0,5] \begin{bmatrix} 1 \\ 52 \end{bmatrix} - 132 \right] + \dots + \left[[0 \quad 0,5] \begin{bmatrix} 1 \\ 72 \end{bmatrix} - 166 \right] \right\} = -118,86\end{aligned}$$

$$\begin{aligned}\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_1} &= \frac{1}{m} \sum_{i=1}^m [h(x^{(i)}) - y^{(i)}] x^{(i)} = \frac{1}{11} \sum_{i=1}^{11} \left[[0 \quad 0,5] \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix} - y^{(i)} \right] x^{(i)} \\ &= \frac{1}{11} \left\{ \left[[0 \quad 0,5] \begin{bmatrix} 1 \\ 52 \end{bmatrix} - 132 \right] \times 52 + \dots + \left[[0 \quad 0,5] \begin{bmatrix} 1 \\ 72 \end{bmatrix} - 166 \right] \times 72 \right\} \\ &= -7496,32\end{aligned}$$

$$\omega_0^{(1)} = \omega_0^{(0)} - \alpha \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_0} = 0 - 0,0005 \times (-118,86) = 0,059$$

$$\omega_1^{(1)} = \omega_1^{(0)} - \alpha \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_1} = 0,5 - 0,0005 \times (-7496,32) = 4,248$$

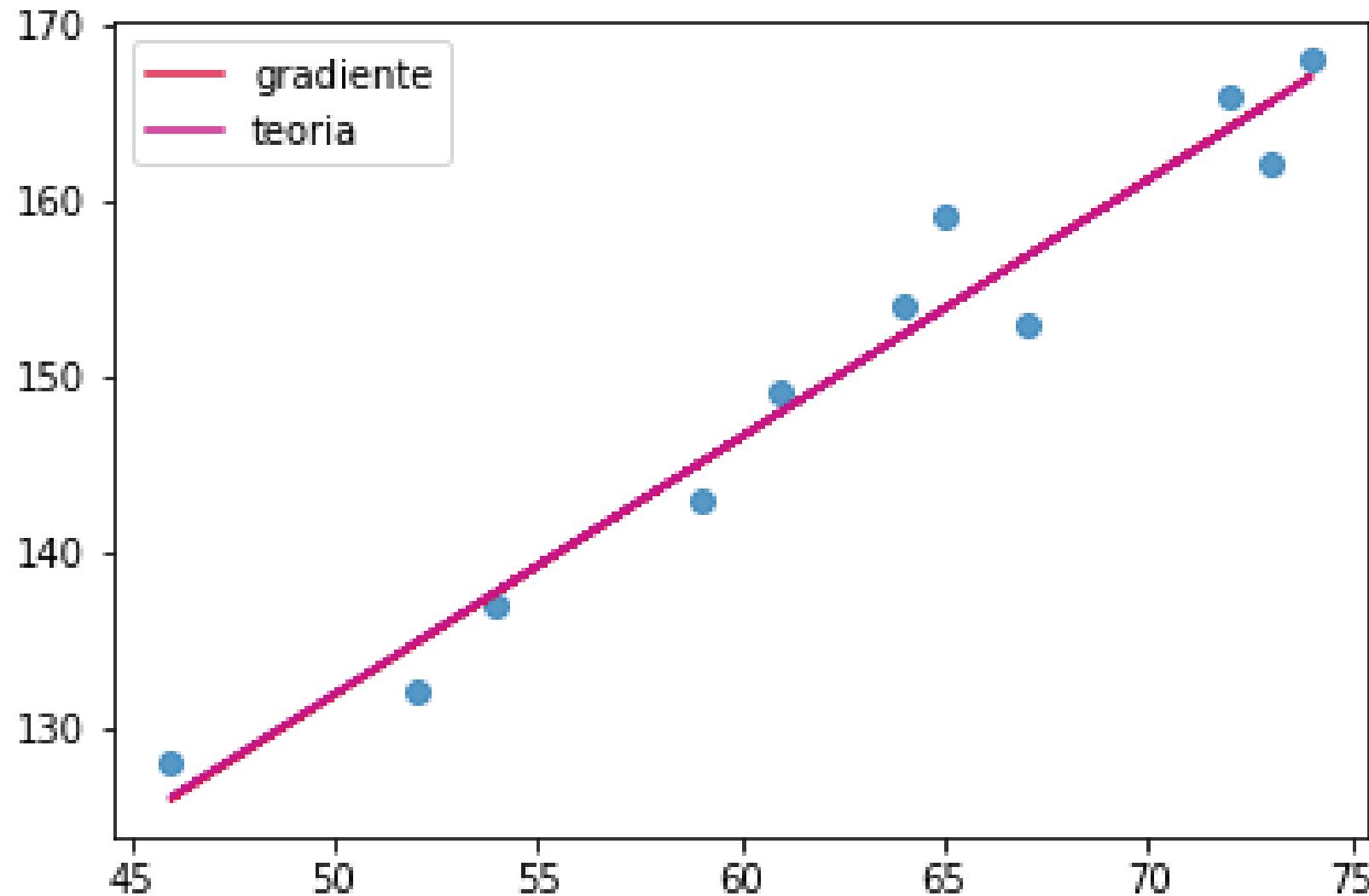
$$\omega_0^{(0)} = 0, \quad \omega_1^{(0)} = 0,5$$

$$\alpha = 0,0005$$

$$h = \hat{y} = 0x_0 + 0,5x$$

$$h = \hat{y} = [0 \quad 0,5] \begin{bmatrix} 1 \\ x \end{bmatrix}$$

<i>i</i>	<i>x</i> Idade	<i>y</i> Pressão Arterial
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





REGRESSÃO LINEAR MÚLTIPLA

A regressão multilinear trata da relação de uma variável dependente com *múltiplas* variáveis independentes.

PROBLEMA DEFINIDO

HIPÓTESE: $h(\mathbf{x}) = \hat{y}(\mathbf{x}) = \omega_0x_0 + \omega_1x_1 + \omega_2x_2 + \cdots + \omega_nx_n = \boldsymbol{\omega}^T\mathbf{x}$

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

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

$$\mathbf{x} = \begin{bmatrix} x_0 = 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{bmatrix}$$

x_k arranjados em um vetor \mathbf{x}

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

<i>m</i>	<i>Idade</i>	<i>Peso (Kg)</i>	<i>Pressão Arterial</i>
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

$$X = \begin{pmatrix} - & x^{(1)T} & - \\ - & x^{(2)T} & - \\ - & \vdots & - \\ - & x^{(m)T} & - \end{pmatrix} x^{(i)} = \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{pmatrix}$$

m: linhas de $X \rightarrow$ número de dados de entrada (11)

n: colunas de $X \rightarrow$ número de features (01: idade, 02:peso)

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

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

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



RESPOSTA ANALÍTICA

Dada uma função diferenciável, é teoricamente possível encontrar seu mínimo analiticamente: sabe-se que o mínimo de uma função é um ponto em que a derivada é 0, então tudo que você precisa fazer é encontrar todos os pontos onde a derivada vai para 0 e verificar para qual desses pontos, a função tem o menor valor.

$$\omega = (X^T X)^{-1} X^T y$$

Veja que $m \geq n$ e, portanto, $(X^T X)^{-1} X^T$ é a *pseudo inversa* de X e $(X^T X)$ é uma matriz positiva definida.

Lembrar das recomendações do Thiago:
Não inverta diretamente a matriz,
nunca!

ALGORITMO ANALÍTICO PARA REGRESSÃO LINEAR

- Construir a matriz X e o vetor y a partir dos dados

$$X = (\mathbf{x}^{(1)}, \mathbf{x}^{(i)}, \dots, \mathbf{x}^{(m)})$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ \omega_n \end{bmatrix}$$

- Computar a pseudo inversa $X^+ = (X^T X)^{-1} X^T$
- Retornar $\boldsymbol{\omega} = X^+ \mathbf{y}$

ALGORITMO DE GRADIENTE DESCENDENTE

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

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

Defina α e $i = 0$

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

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

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

$$i += 1$$

$\nabla J(\omega)$

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

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

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

EXEMPLO

$$m = 6, n = 2$$

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

Vamos ao Notebook

$$h(\mathbf{x}) = \hat{y}(\mathbf{x}) = \omega_0 x_0 + \omega_1 x_1 + \omega_2 x_2$$

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

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

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

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

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

$$\omega_k^{(i+1)} = \omega_k^{(i)} - \alpha \nabla J_k(\boldsymbol{\omega}^{(i)})$$

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

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

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

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

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

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

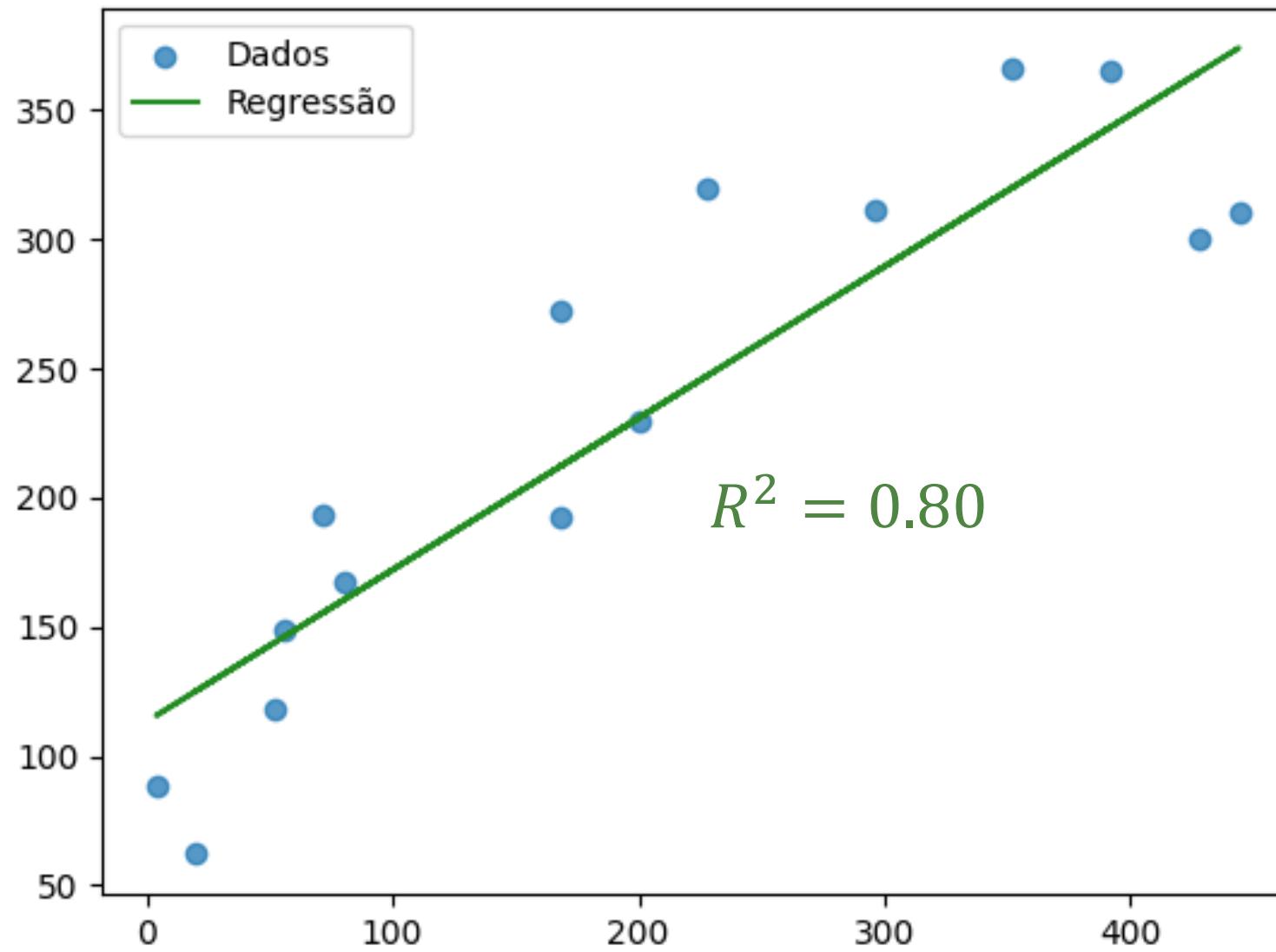
$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_1} = \frac{1}{m} \sum_{i=1}^m [h(x^{(i)}) - y^{(i)}] x_1^{(i)} = \frac{1}{6} \left\{ \left[[0 \quad -0.017 \quad -0.048] \begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} - 2 \right] 4 + \left[[0 \quad -0.017 \quad -0.048] \begin{bmatrix} 1 \\ 2 \\ 8 \end{bmatrix}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

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

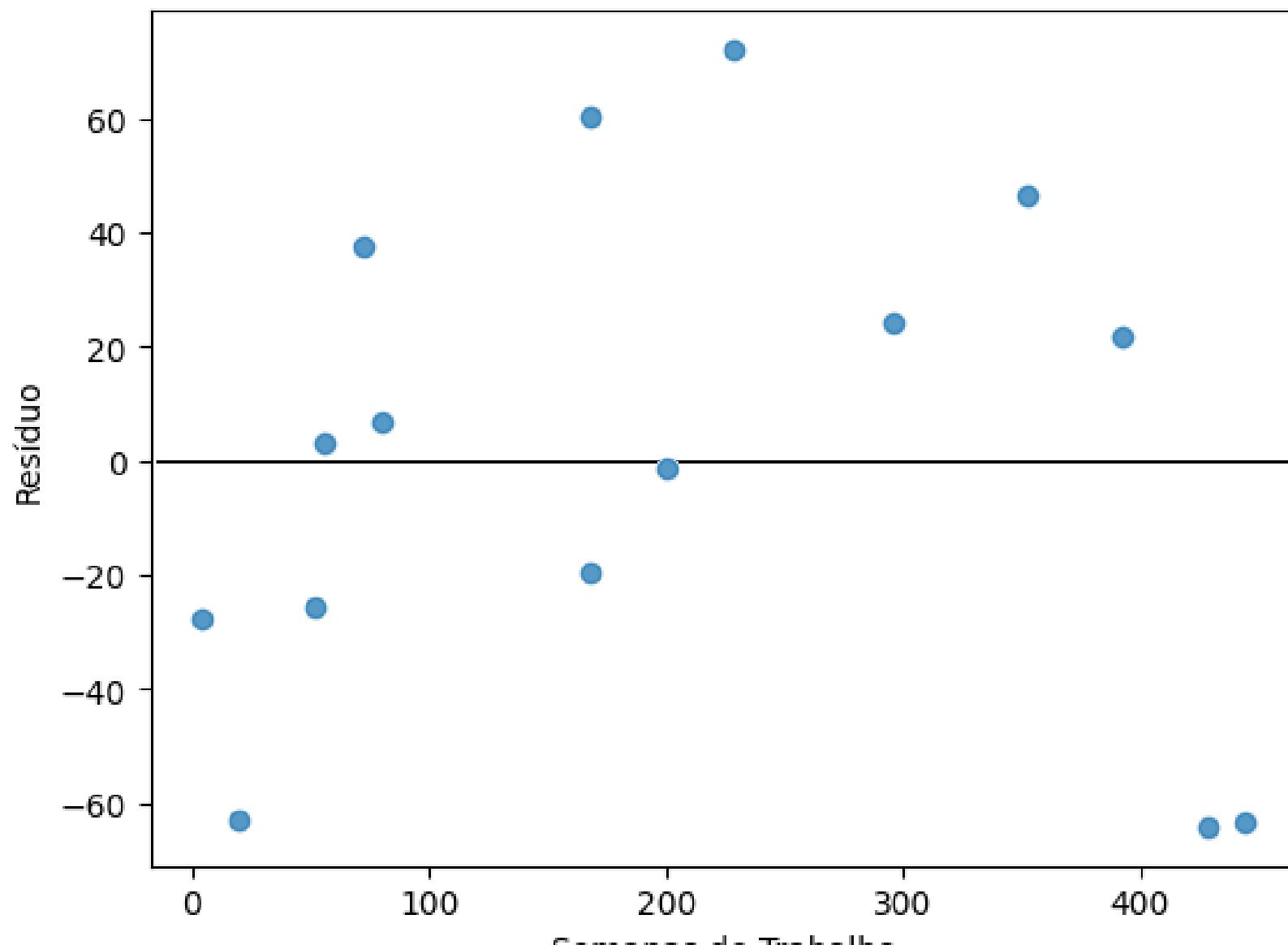
$$\frac{\partial J(\boldsymbol{\omega})}{\partial \omega_2} = \frac{1}{m} \sum_{i=1}^m [h(x^{(i)}) - y^{(i)}] x_2^{(i)} = \frac{1}{6} \left\{ \left[[0 \quad -0.017 \quad -0.048] \begin{bmatrix} 1 \\ 4 \\ 1 \end{bmatrix} - 2 \right] 1 + \left[[0 \quad -0.017 \quad -0.048] \begin{bmatrix} 1 \\ 2 \\ 8 \end{bmatrix}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

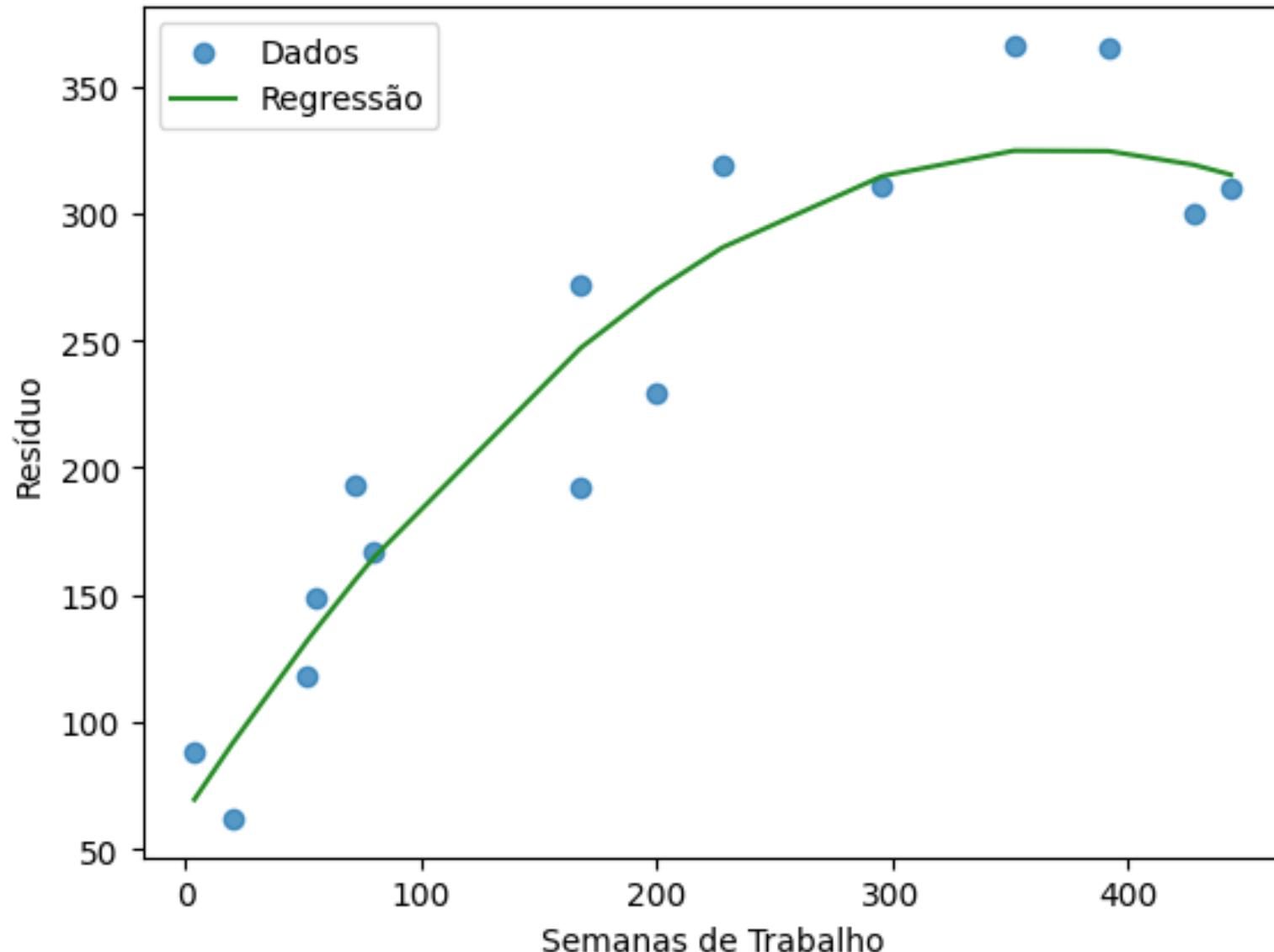
$$\omega_2^{(1)} = \omega_2^{(0)} - \alpha \frac{\partial J(\boldsymbol{\omega})}{\partial \omega_2}$$





168	272
428	300
296	311
392	365
80	167
56	149
352	366
444	310
168	192
200	229
4	88
52	118
20	62
228	319
72	193

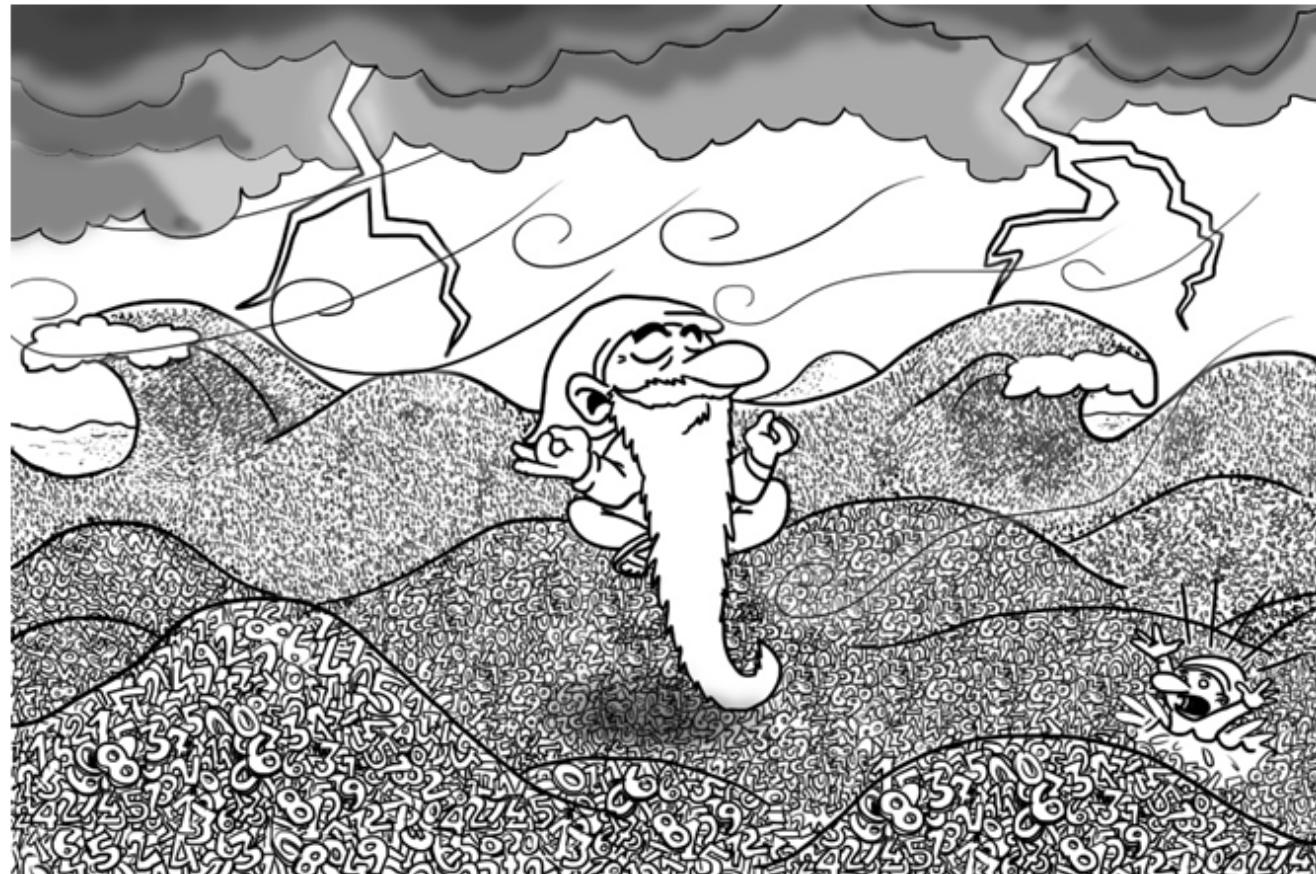


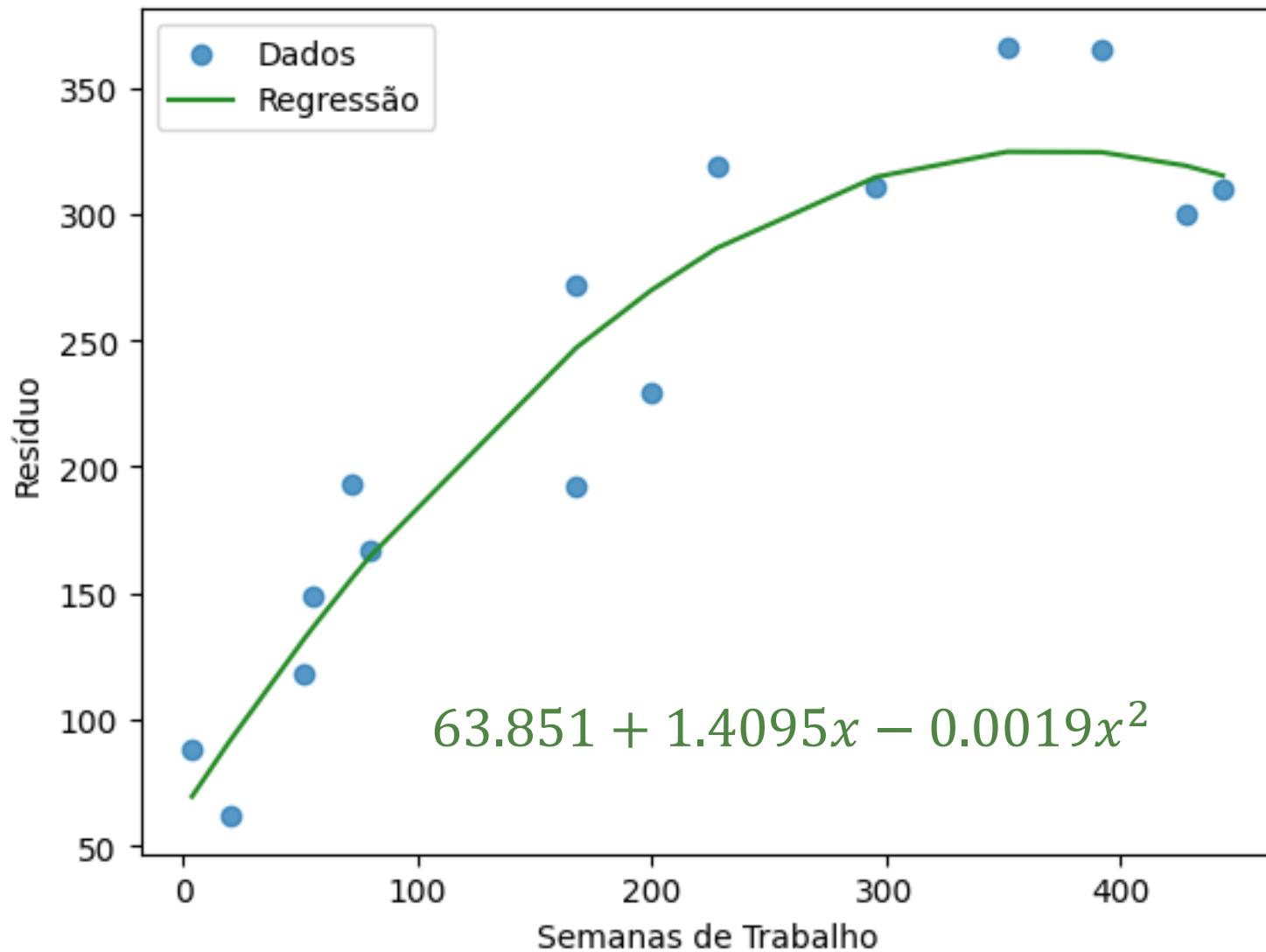


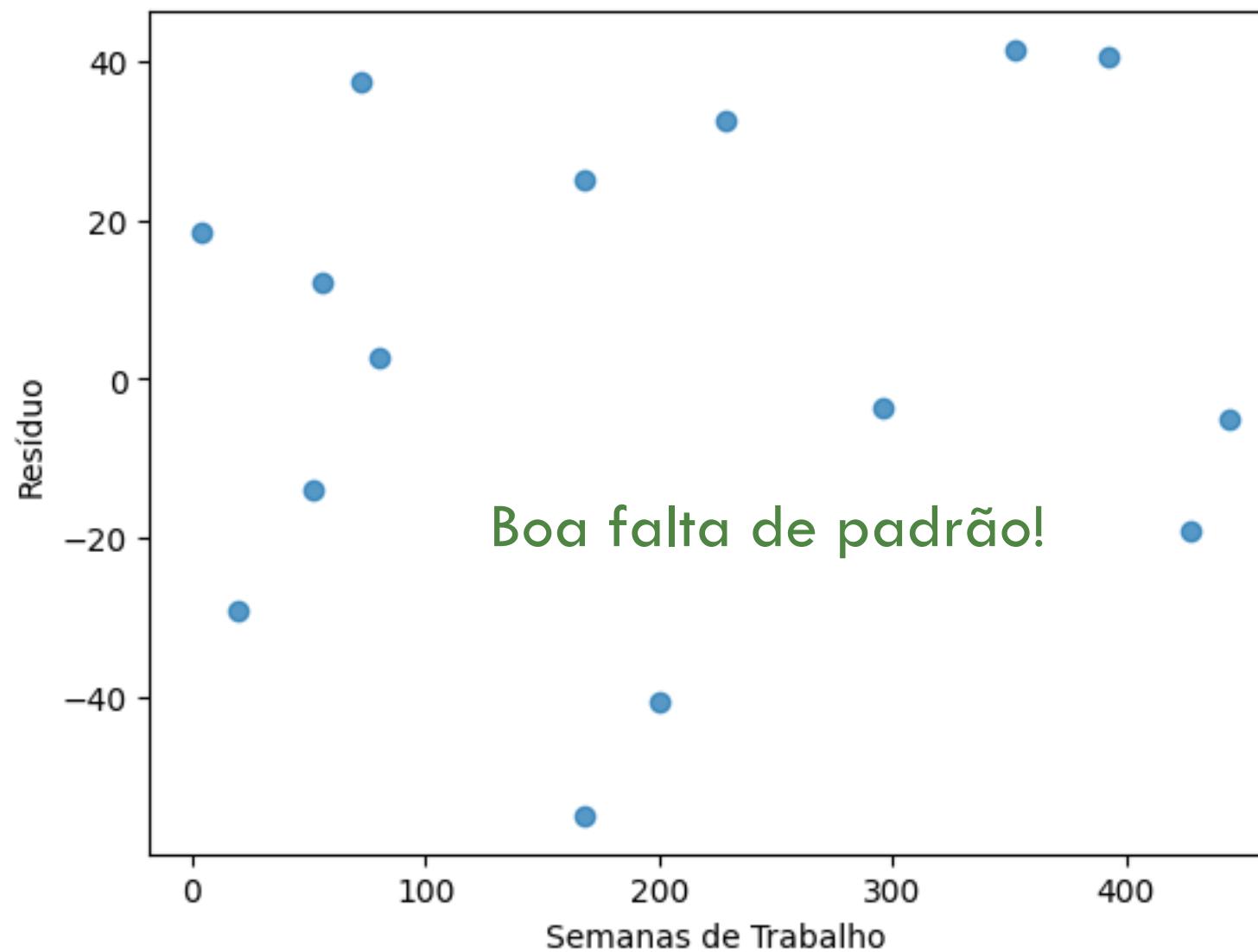
MODELO NÃO LINEAR

- Parece que o modelo linear não é melhor escolha para modelar os dados originais...
- Erros muito altos!
- A plotagem do resíduo possui uma curvatura definida – isso indica que o modelo pode ser não linear.
- Modelos polinomiais adicionam variáveis independentes extra que são a potência das variáveis originais: x, x^2, x^3, \dots
- Termos extra devem ser vistos com cuidado. Porque?

Mineração de dados,
regra 1. Torturar os dados até que
confessem... e se você os torturar o
suficiente, pode fazer com que
confessem qualquer coisa.

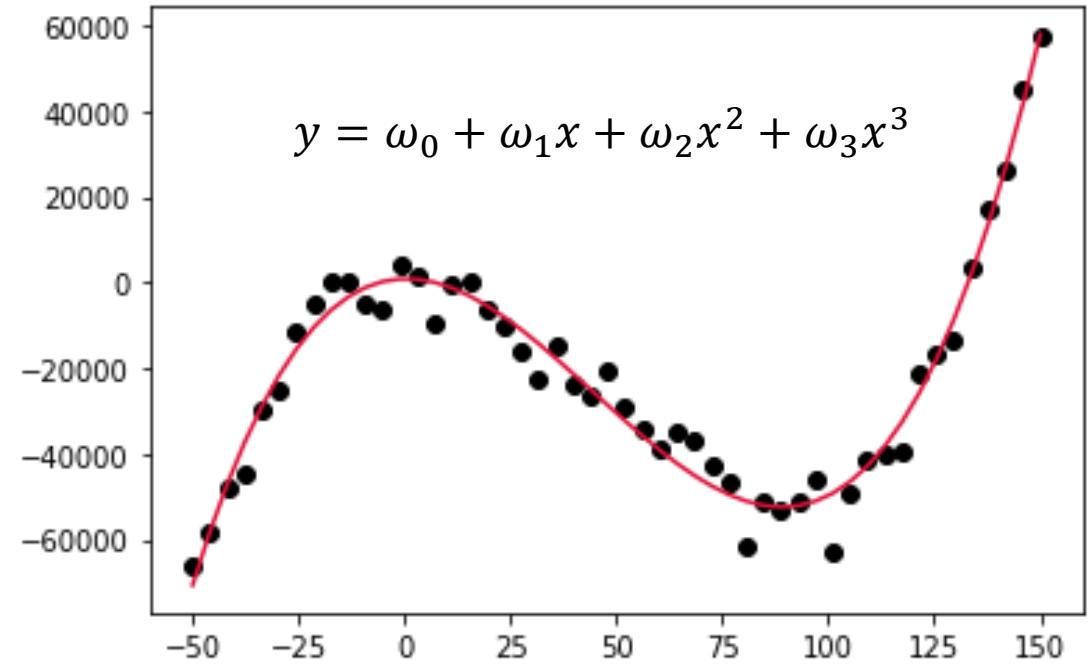
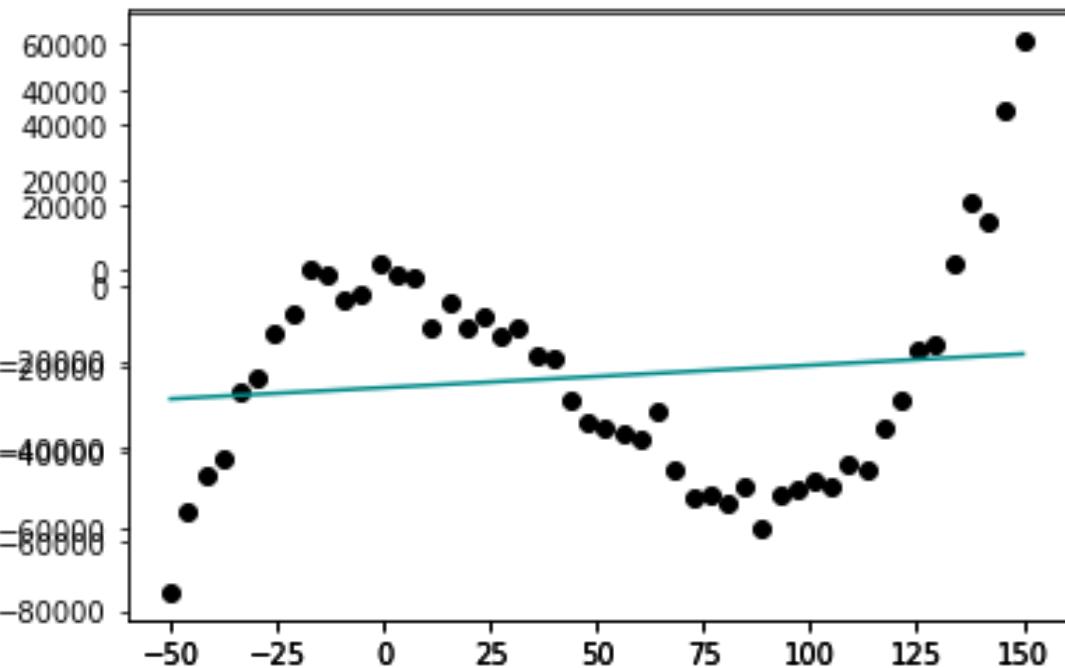






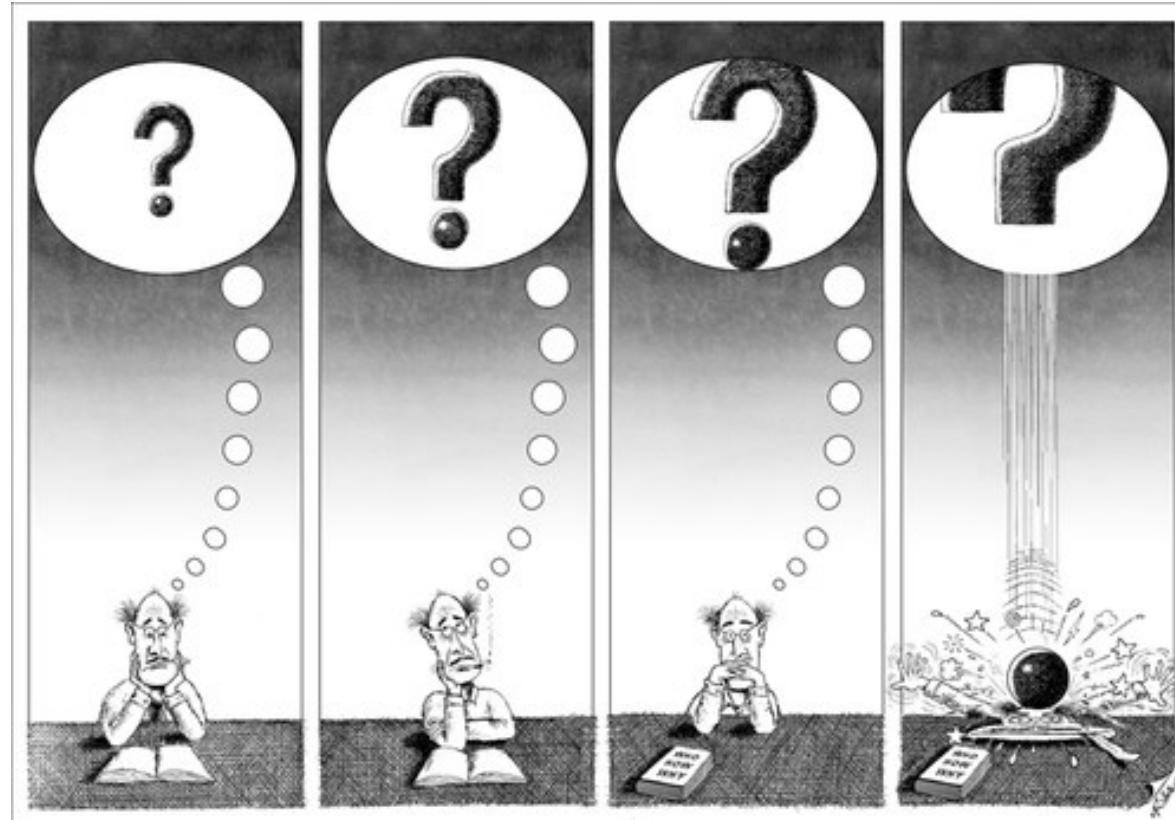
REGRESSÃO POLINOMIAL

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



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 variável dependente correspondente de y .

Regressão linear polinomial ???



Regressão Linear Simples

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

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

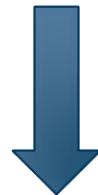
Regressão Linear Múltipla

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

Regressão Linear Polinomial

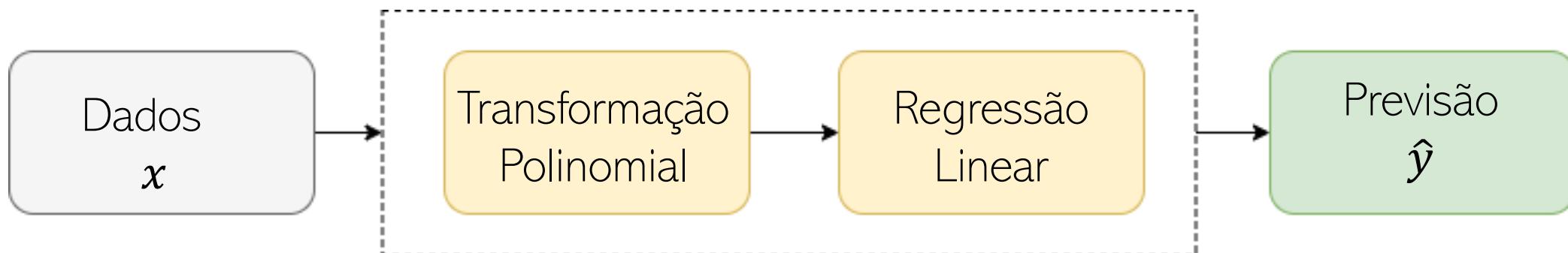
$$\hat{y} = \omega_0 + \omega_1 x + \omega_2 x^2 + \omega_3 x^3 + \cdots + \omega_n x^n$$

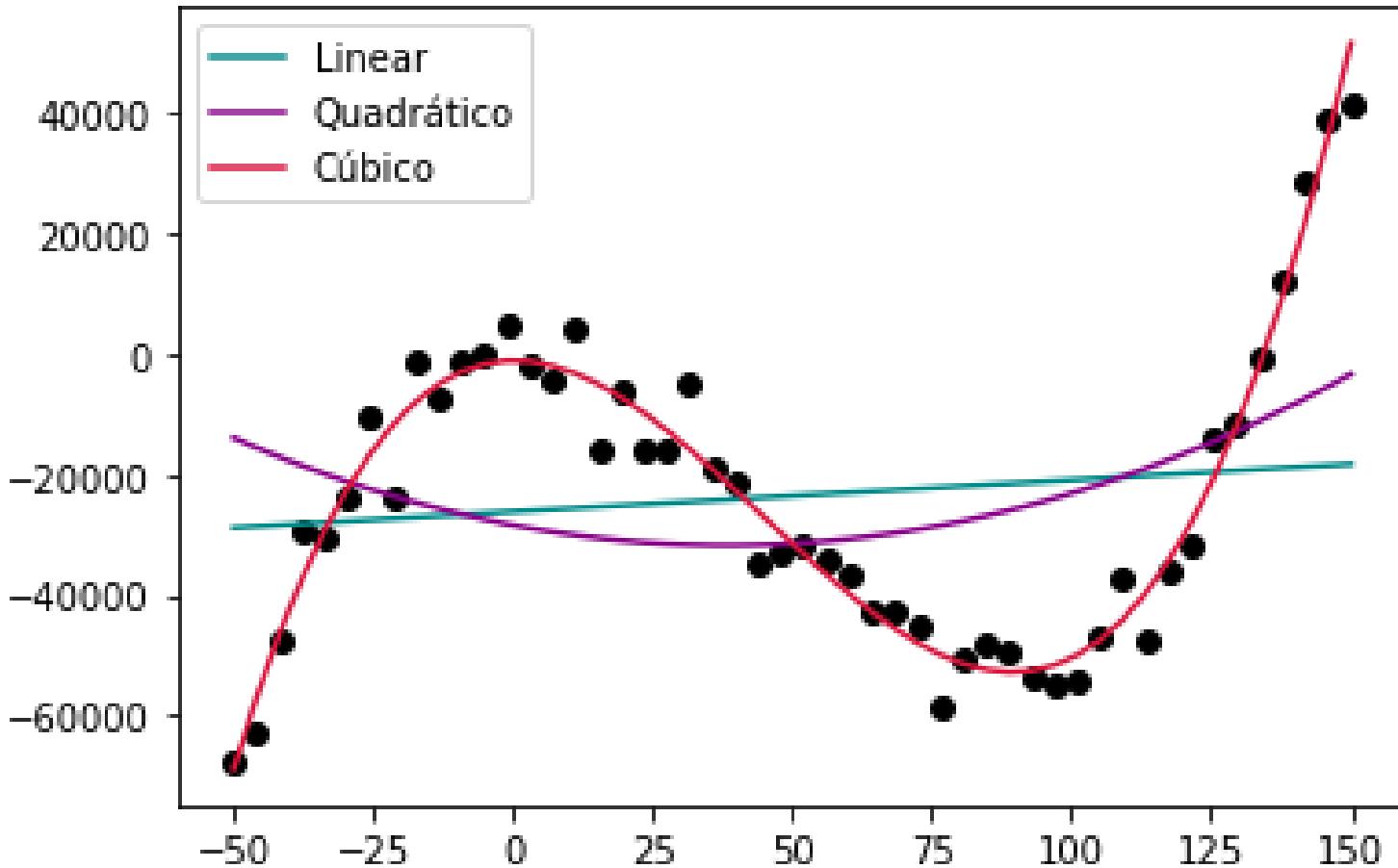
$$h(x) = \omega_0 + \omega_1 x + \omega_2 x^2 + \omega_3 x^3 + \cdots + \omega_n x^n + \varepsilon$$



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

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: $h(\mathbf{x}) = \hat{y}(\mathbf{x}) = \omega_0x_0 + \omega_1x_1 + \omega_2x_2 + \cdots + \omega_nx_n = \boldsymbol{\omega}^T\mathbf{x}$

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

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

$$\mathbf{x} = \begin{bmatrix} x_0 = 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{bmatrix}$$

x_k arranjados em um vetor \mathbf{x}

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

O MÉTODO DOS MÍNIMOS QUADRADOS NO AJUSTE DE UM MODELO POLINOMIAL

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

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

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

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

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

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

FINALMENTE...

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

MÉTODO DO GRADIENTE DESCENDENTE

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

$$\boldsymbol{x} = \begin{bmatrix} x_0 = 1 \\ x \\ x^2 \end{bmatrix} \quad \boldsymbol{\omega} = \begin{bmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \end{bmatrix}$$

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

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

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

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

POR EXEMPLO....

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

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

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

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

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

$$= m\omega_0 + \omega_1 \sum_{i=1}^m x^{(i)} + \omega_2 \sum_{i=1}^m (x^{(i)})^2 - \sum_{i=1}^m y^{(i)}$$

$$= \omega_0 \sum_{i=1}^m x^{(i)} + \omega_1 \sum_{i=1}^m (x^{(i)})^2 + \omega_2 \sum_{i=1}^m (x^{(i)})^3 - \sum_{i=1}^m x^{(i)} y^{(i)}$$

$$= \omega_0 \sum_{i=1}^m (x^{(i)})^2 + \omega_1 \sum_{i=1}^m (x^{(i)})^3 + \omega_2 \sum_{i=1}^m (x^{(i)})^4 - \sum_{i=1}^m (x^{(i)})^2 y^{(i)}$$

ALGORITMO DE GRADIENTE DESCENDENTE

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

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

Defina α e $i = 0$

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

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

$$\omega_k^{(i+1)} = \omega_k^{(i)} - \alpha \nabla J_k(\omega^{(i)}), \quad k = 0, 1, \dots, n$$

$$i += 1$$

$\nabla J(\omega)$

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

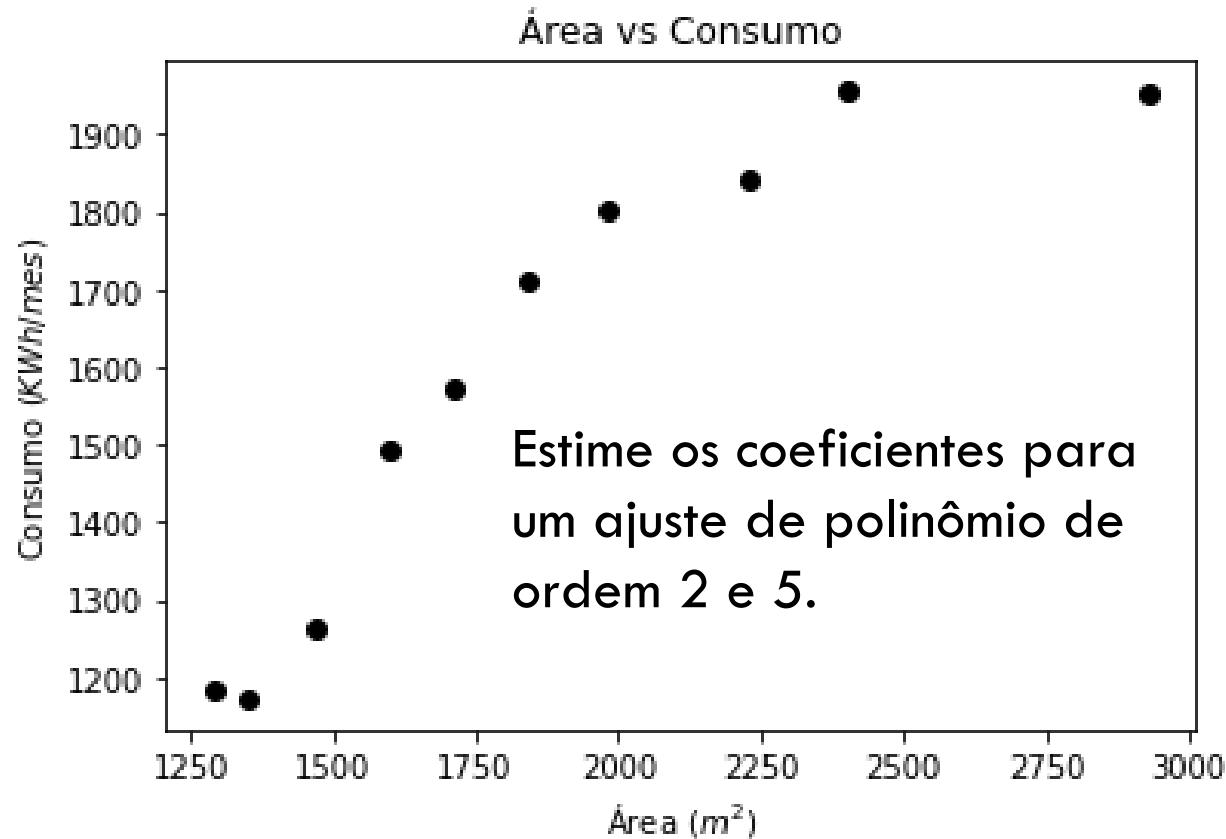
$$\frac{\partial J(\omega)}{\partial \omega_k} = \frac{1}{m} \sum_{i=1}^m [h(x^{(i)}) - y^{(i)}] (x_k^{(i)})^k$$

EXEMPLO

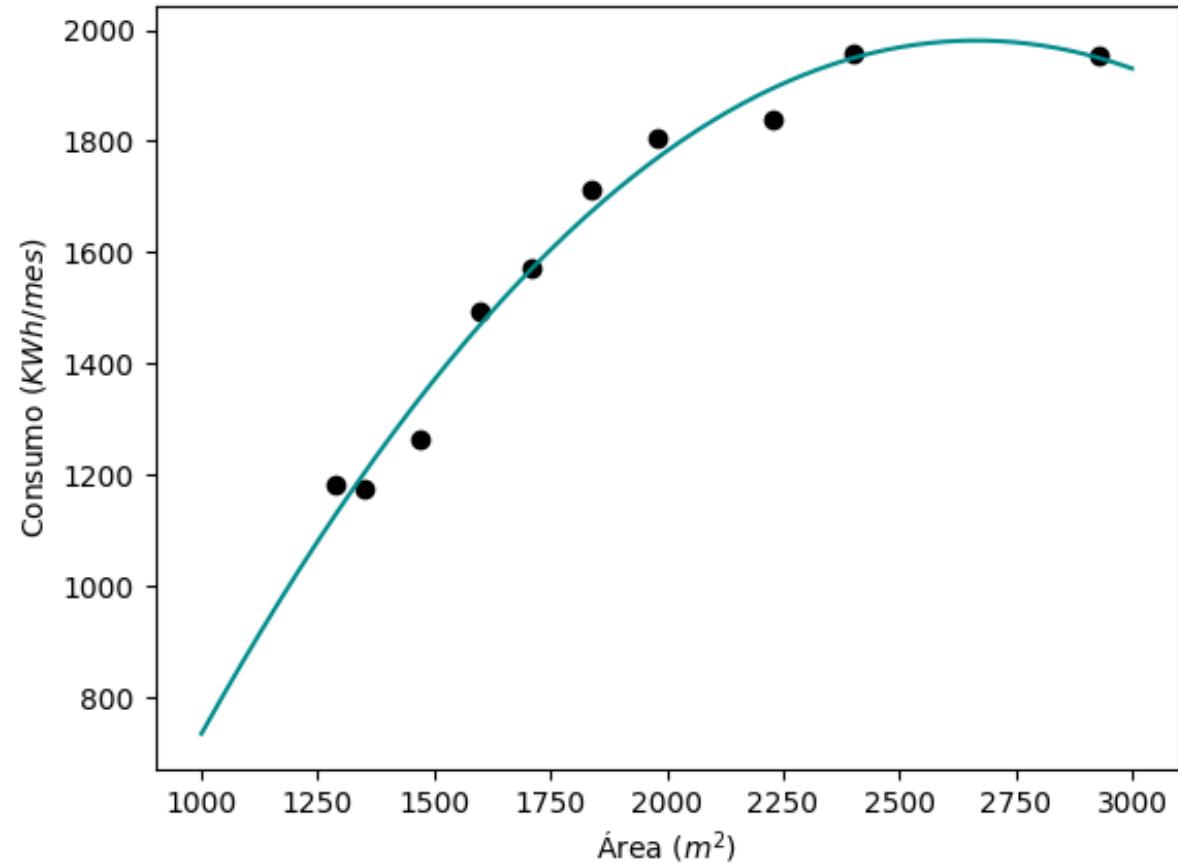
VAMOS ao Notebook

Os dados são os consumos de eletricidade, em quilowatts-hora por mês, de dez casas e suas respectivas áreas, em metros quadrados,

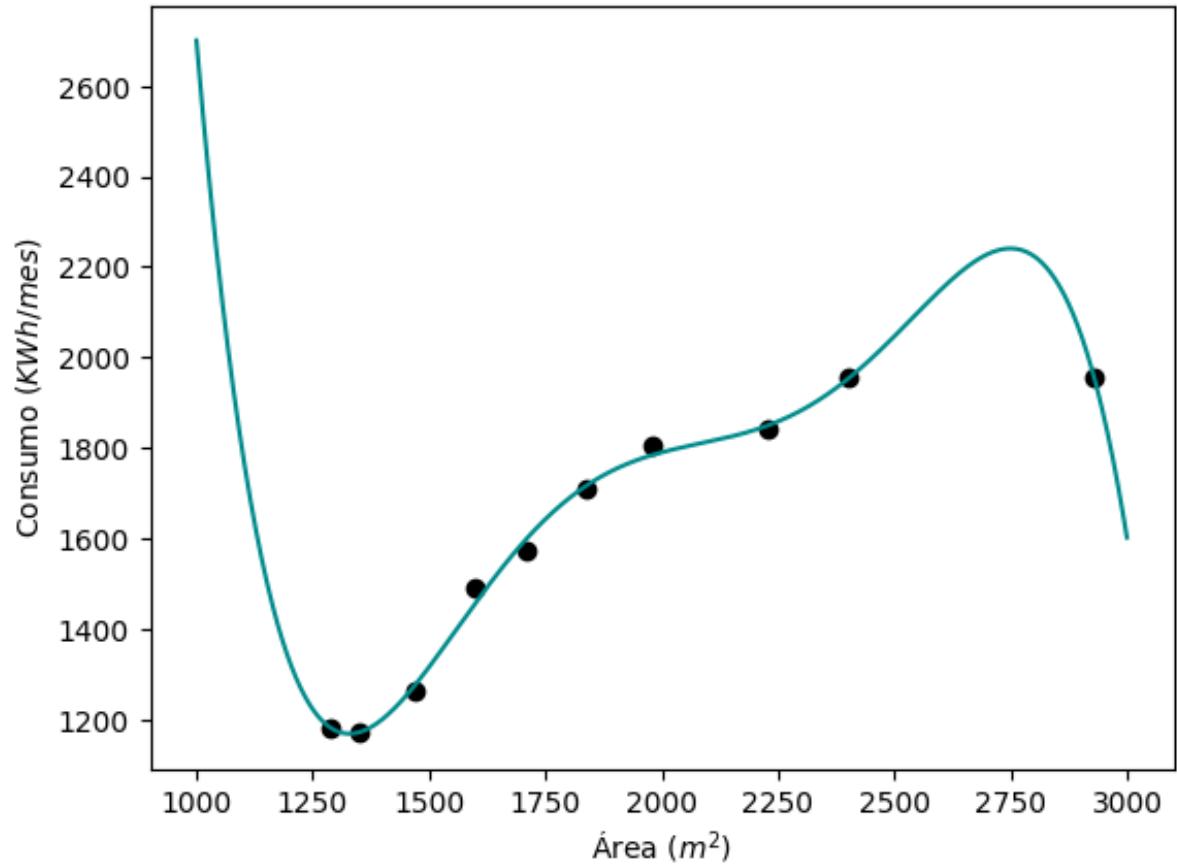
Área	KW Hrs/Mes
120	1182
135	1172
147	1264
160	1493
171	1571
184	1711
198	1804
223	1840
240	1956
293	1954



Área vs Consumo



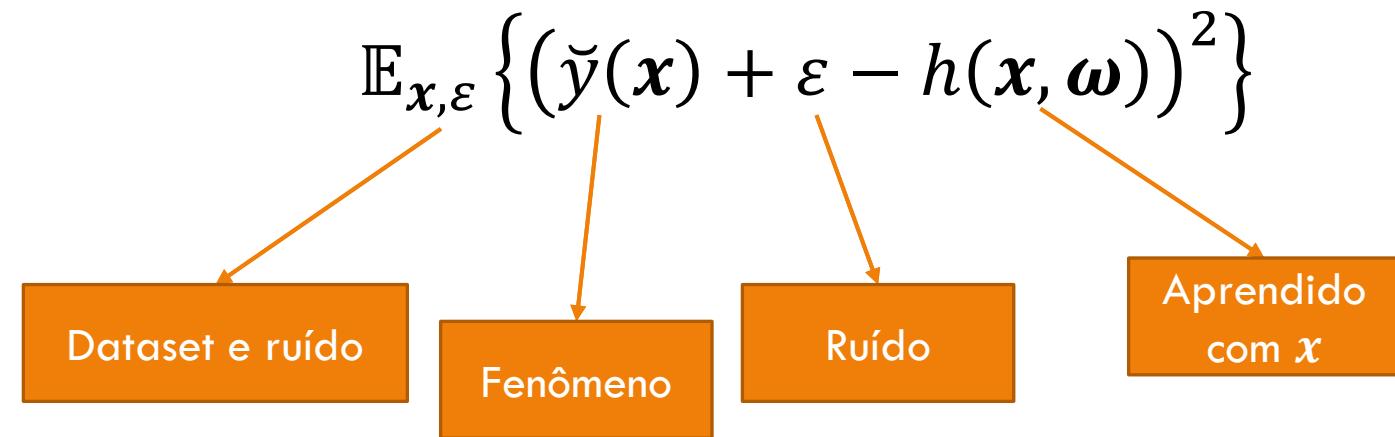
Área vs Consumo





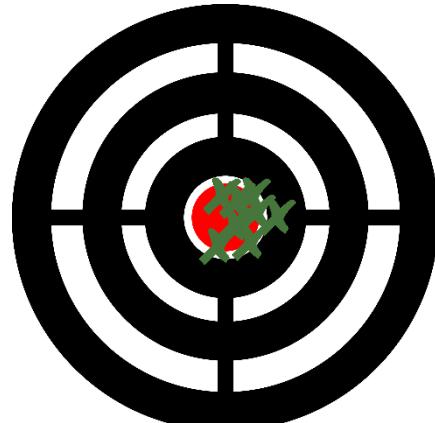
ERRO QUADRÁTICO MÉDIO

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.



VIÉS E VARIÂNCIA

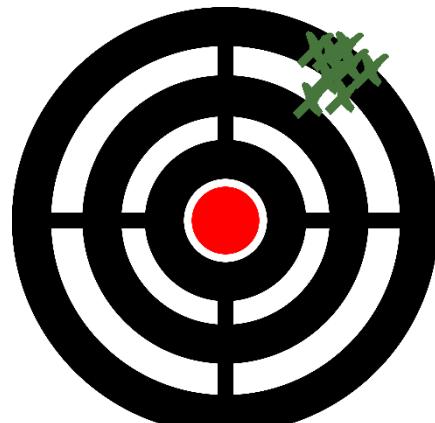
Baixa variância



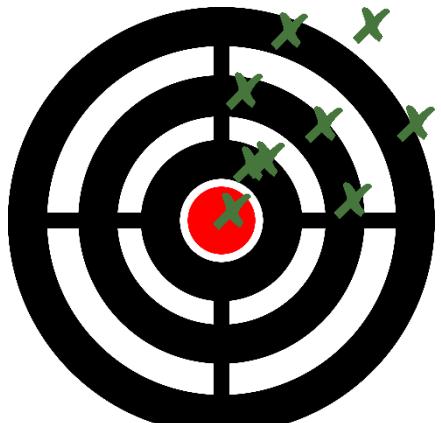
Alta variância



Baixo viés



Alto viés



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

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

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,

$$EQM = \varepsilon^2 + Var(\hat{y}) + (Bias(\hat{y}))^2$$

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.

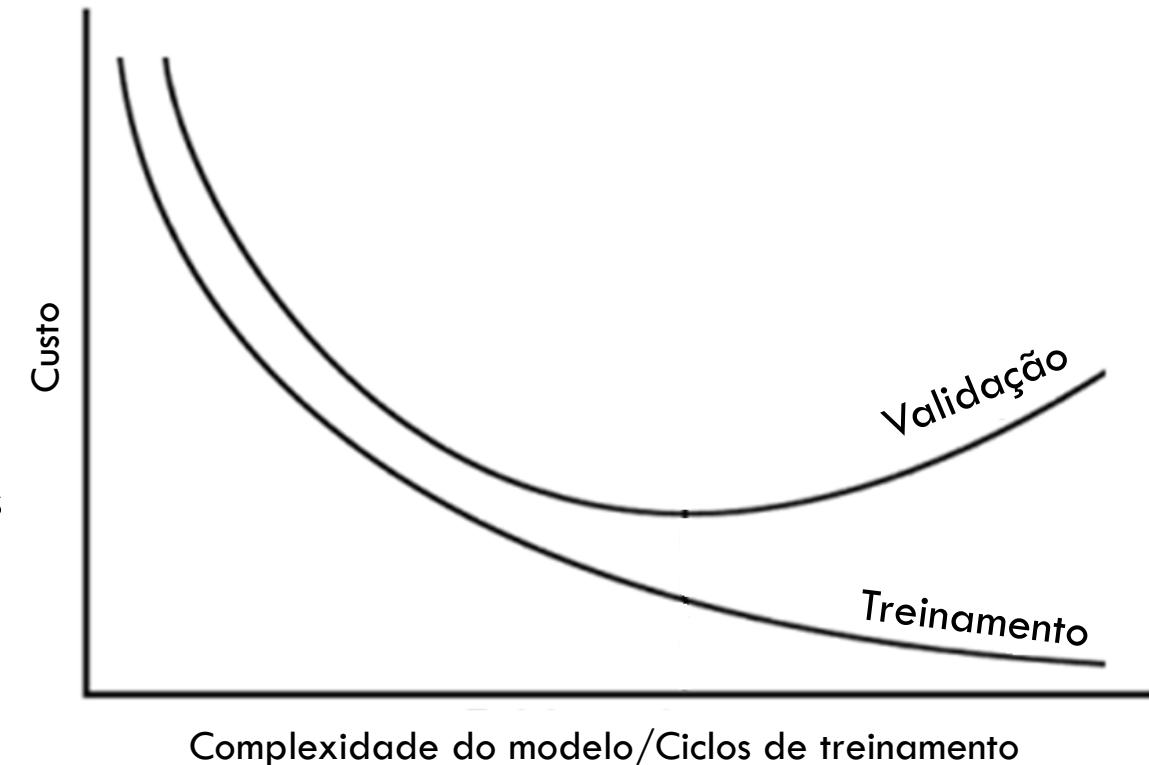
CURVA DE APRENDIZADO

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.

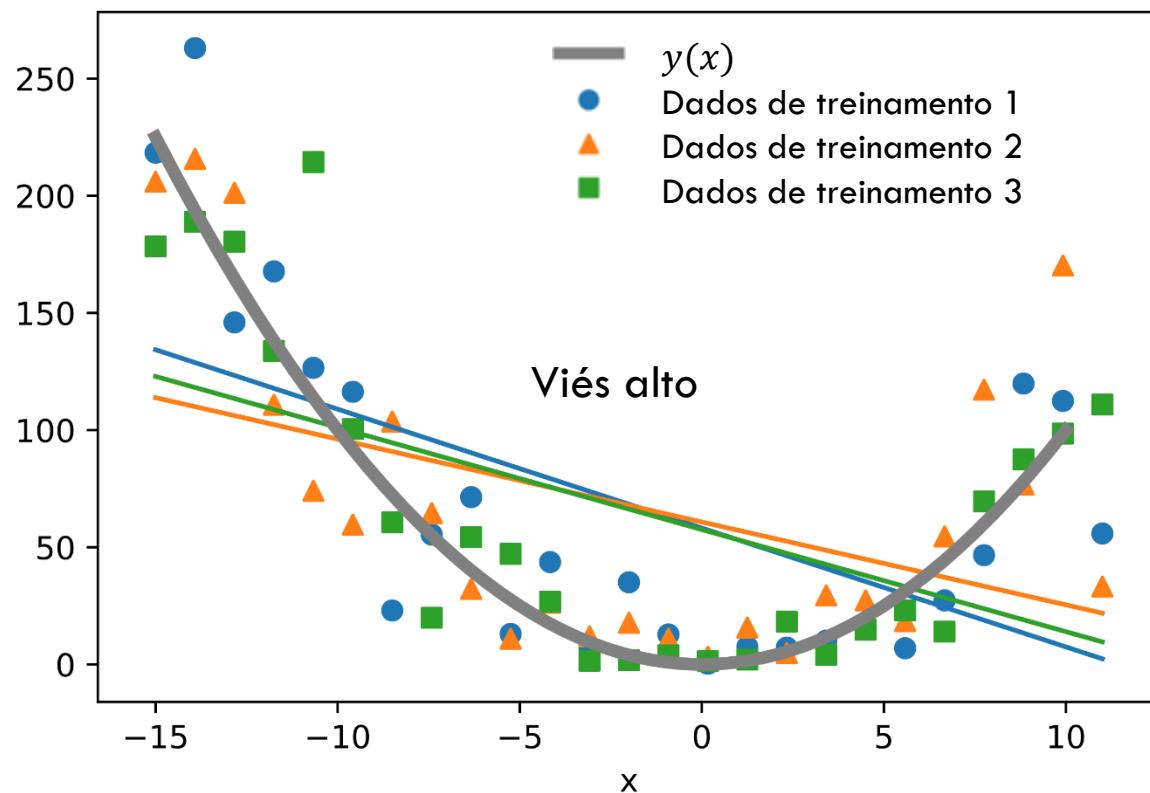
Curva de aprendizado de treinamento: Curva de aprendizado calculada a partir do conjunto de dados de treinamento que fornece uma ideia de quão bem o modelo está **aprendendo**.

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**.

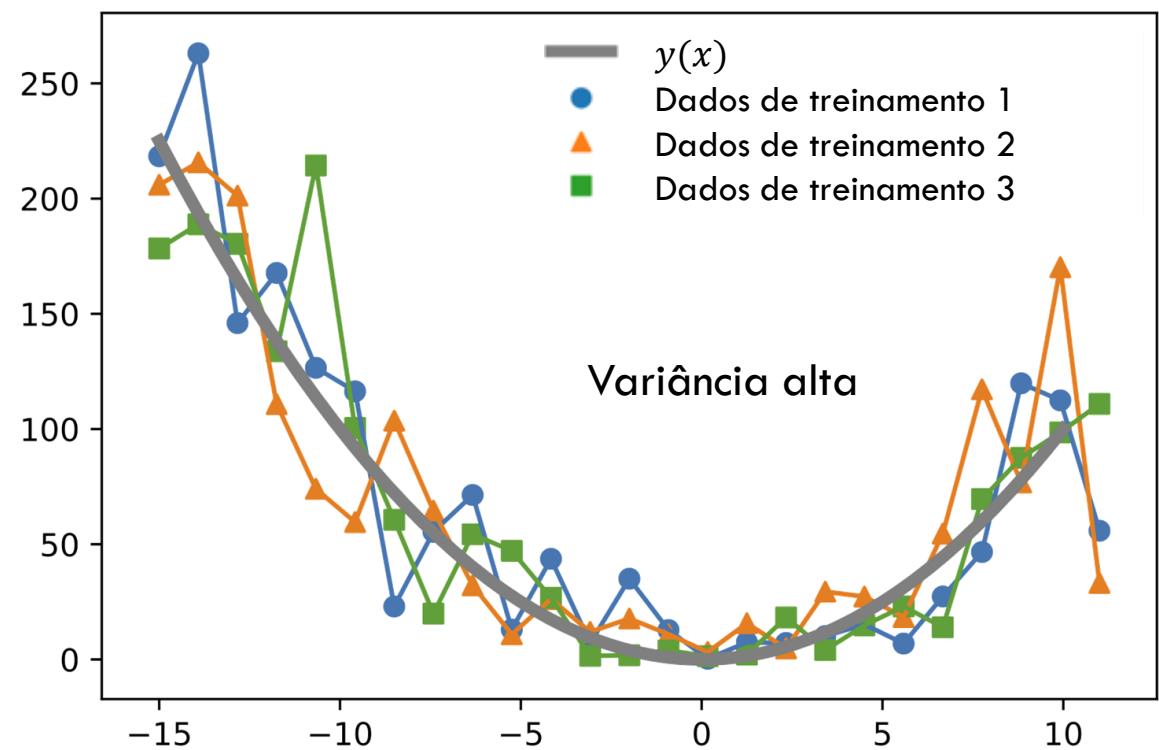
É comum criar as duas curvas de aprendizado para um modelo durante o treinamento.



EXEMPLO EM DADOS 2D

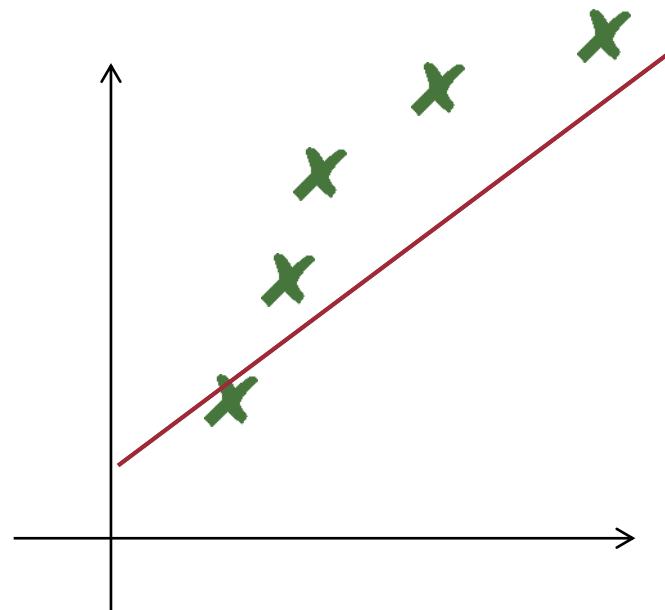


Viés alto

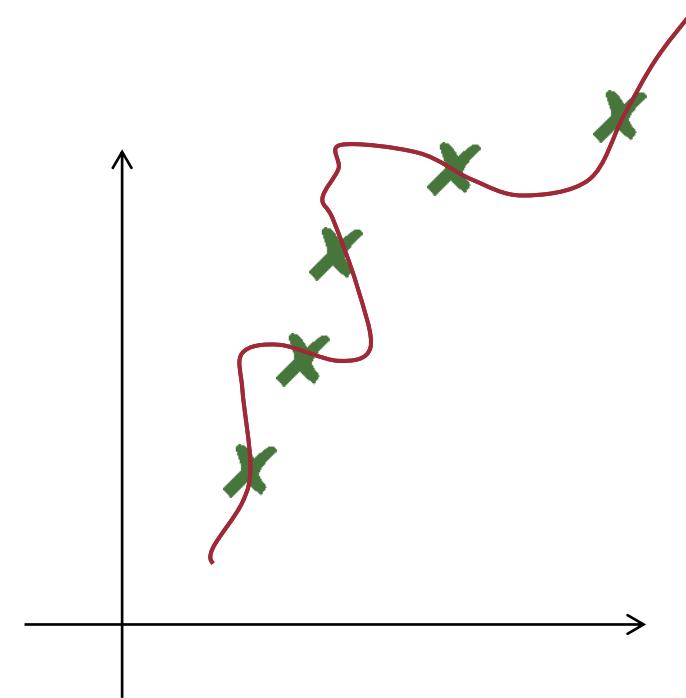


Variância alta

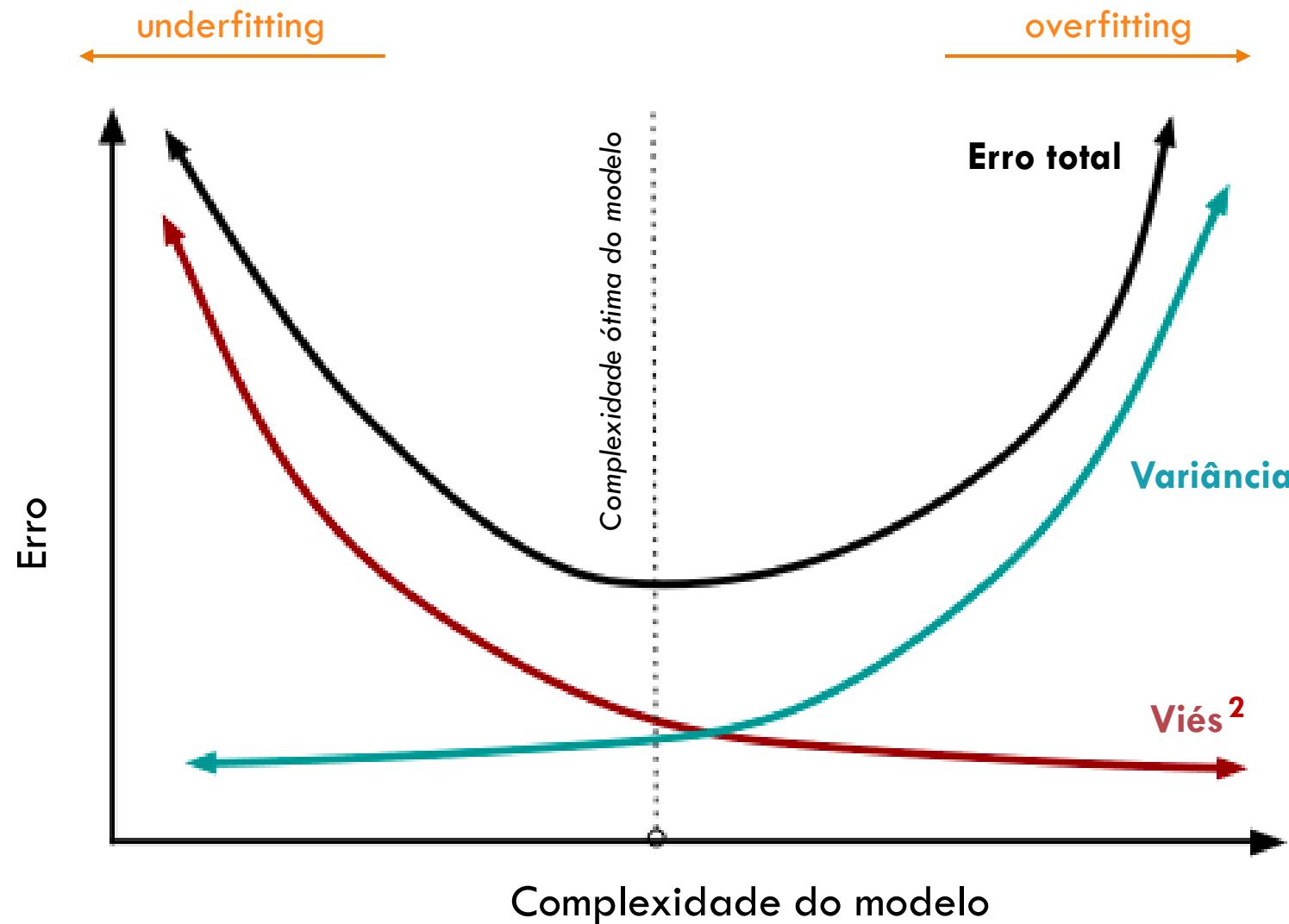
Existe um “trade-off”, ou seja, um balanço, entre viés e variância, de forma que, quando se aumenta a complexidade de um modelo, o quadrado de seu viés tem seu valor diminuído, enquanto a variância tem seu valor aumentado.

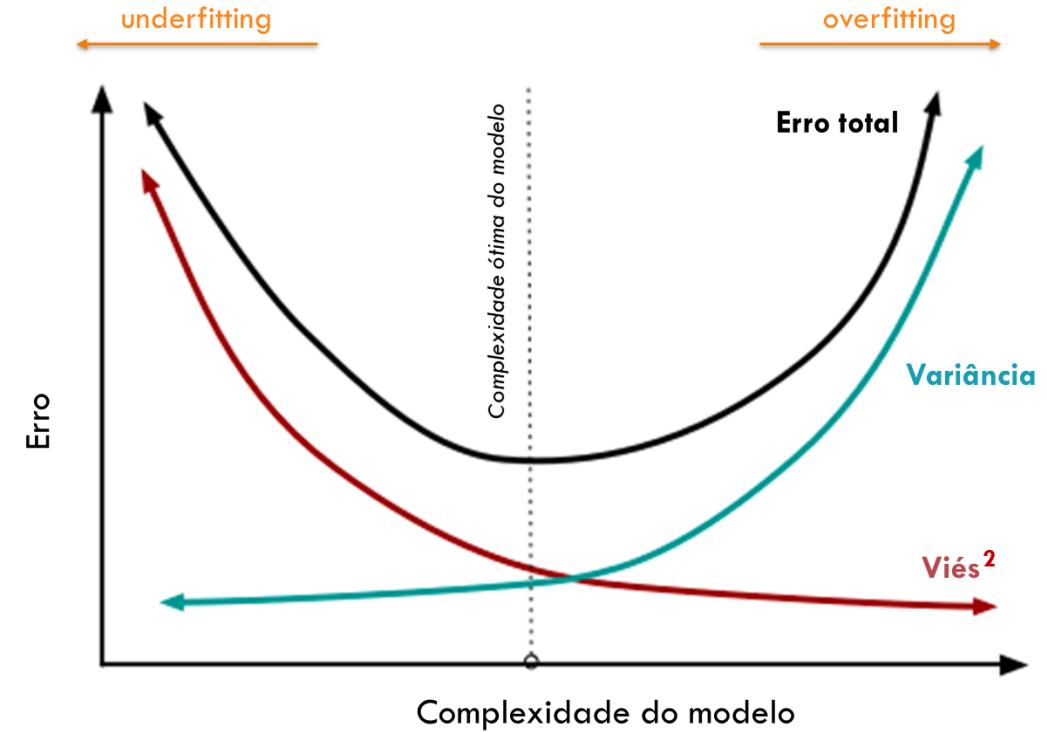
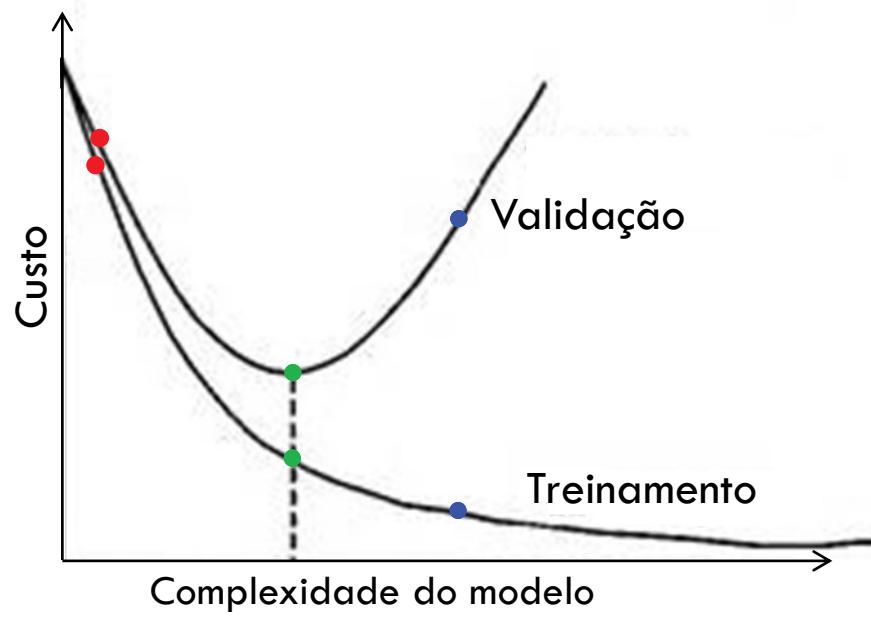
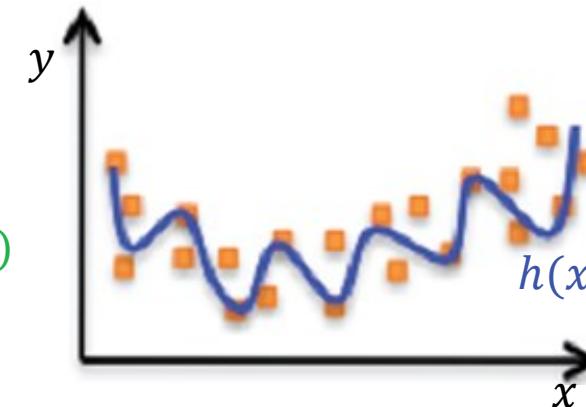
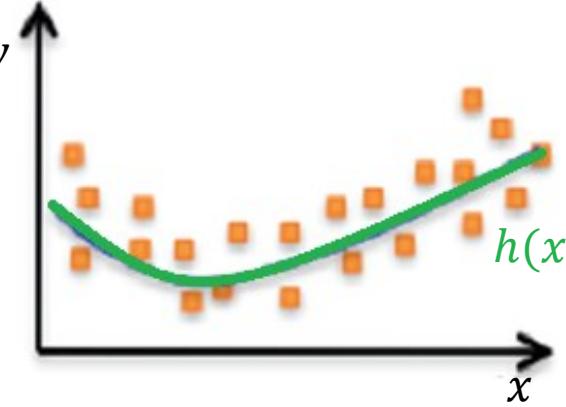
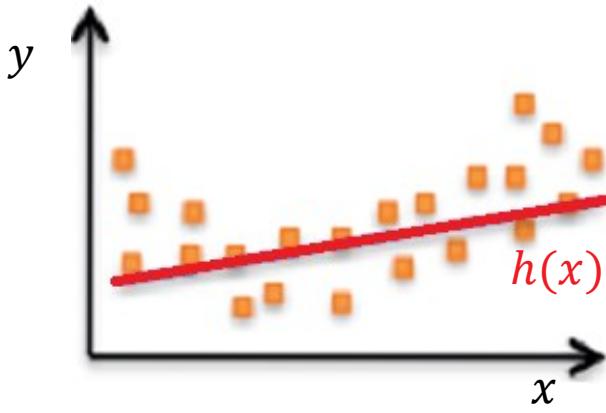


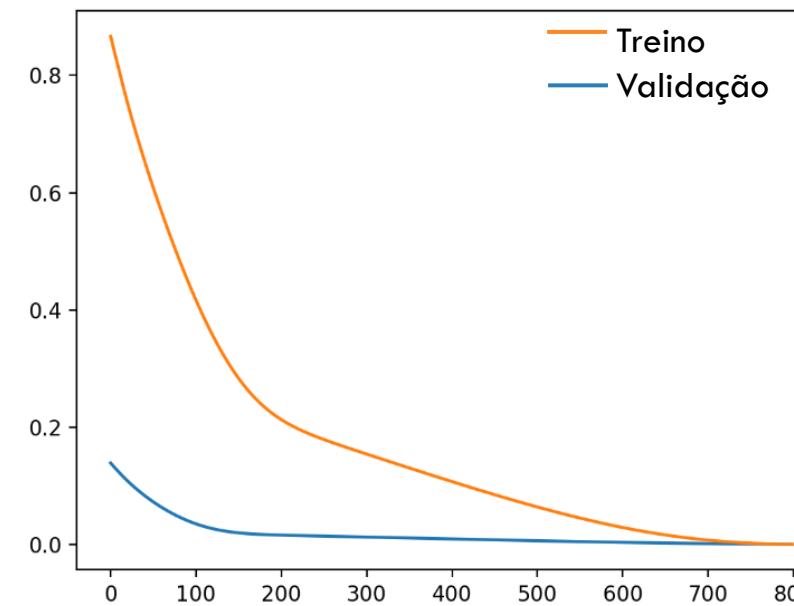
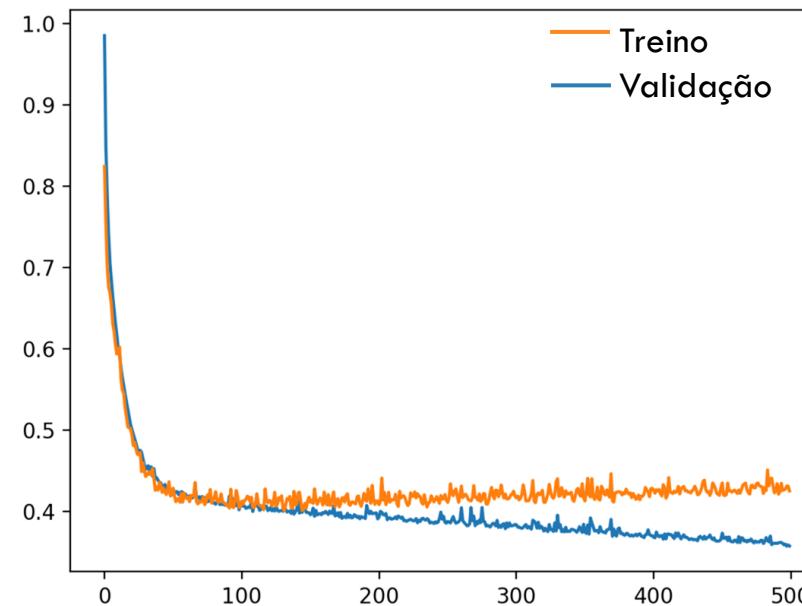
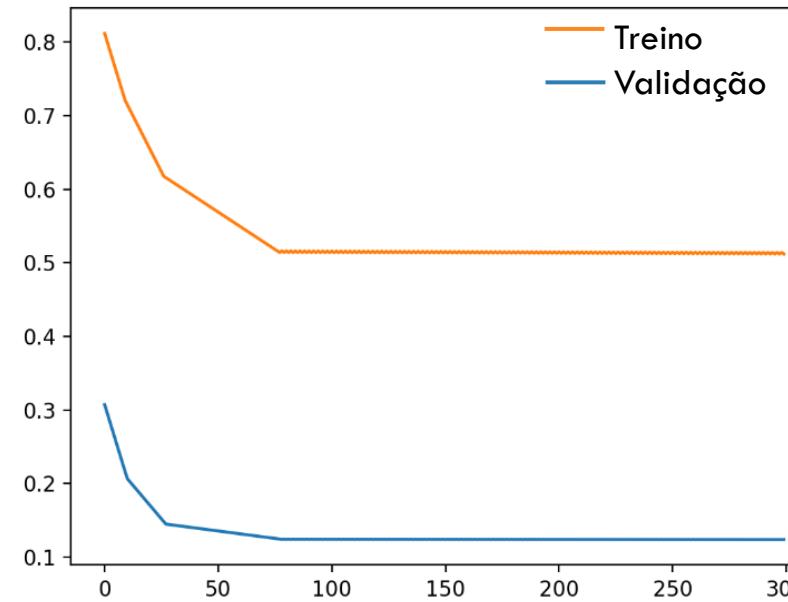
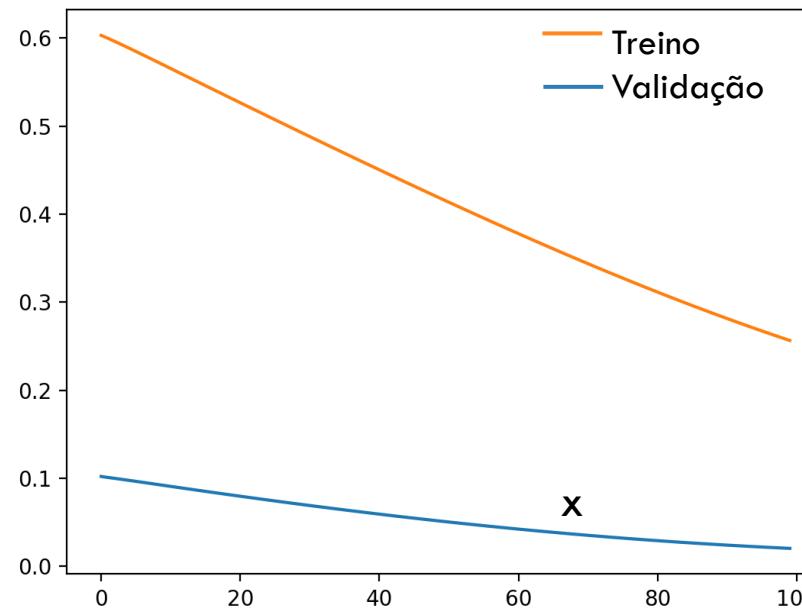
Alto viés (bias): incapacidade do modelo de capturar a verdadeira relação entre variáveis e o objeto a ser predito.
UNDERFITTING



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







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)

EXEMPLO

VAMOS ao Notebook

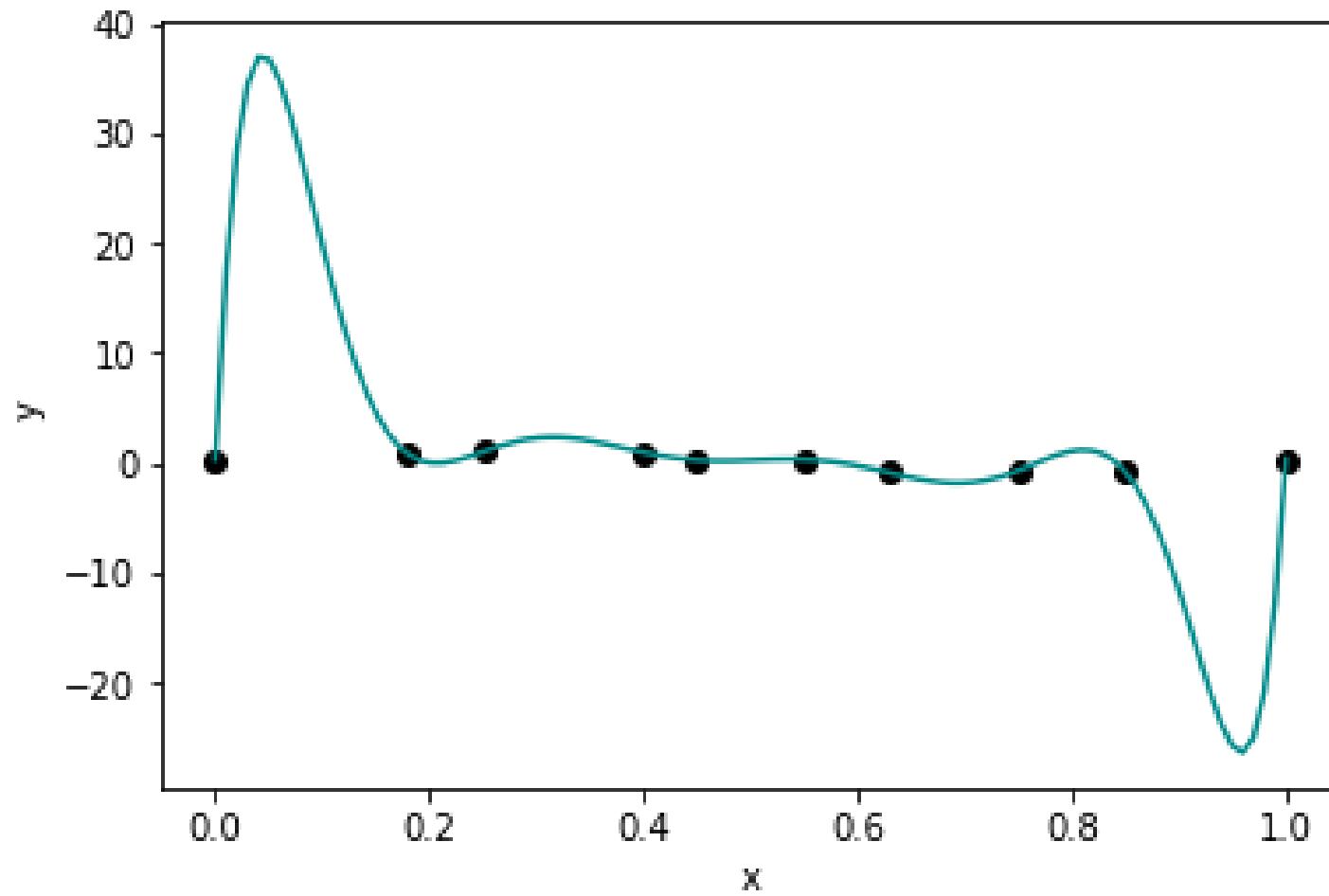
Com os dados abaixo,

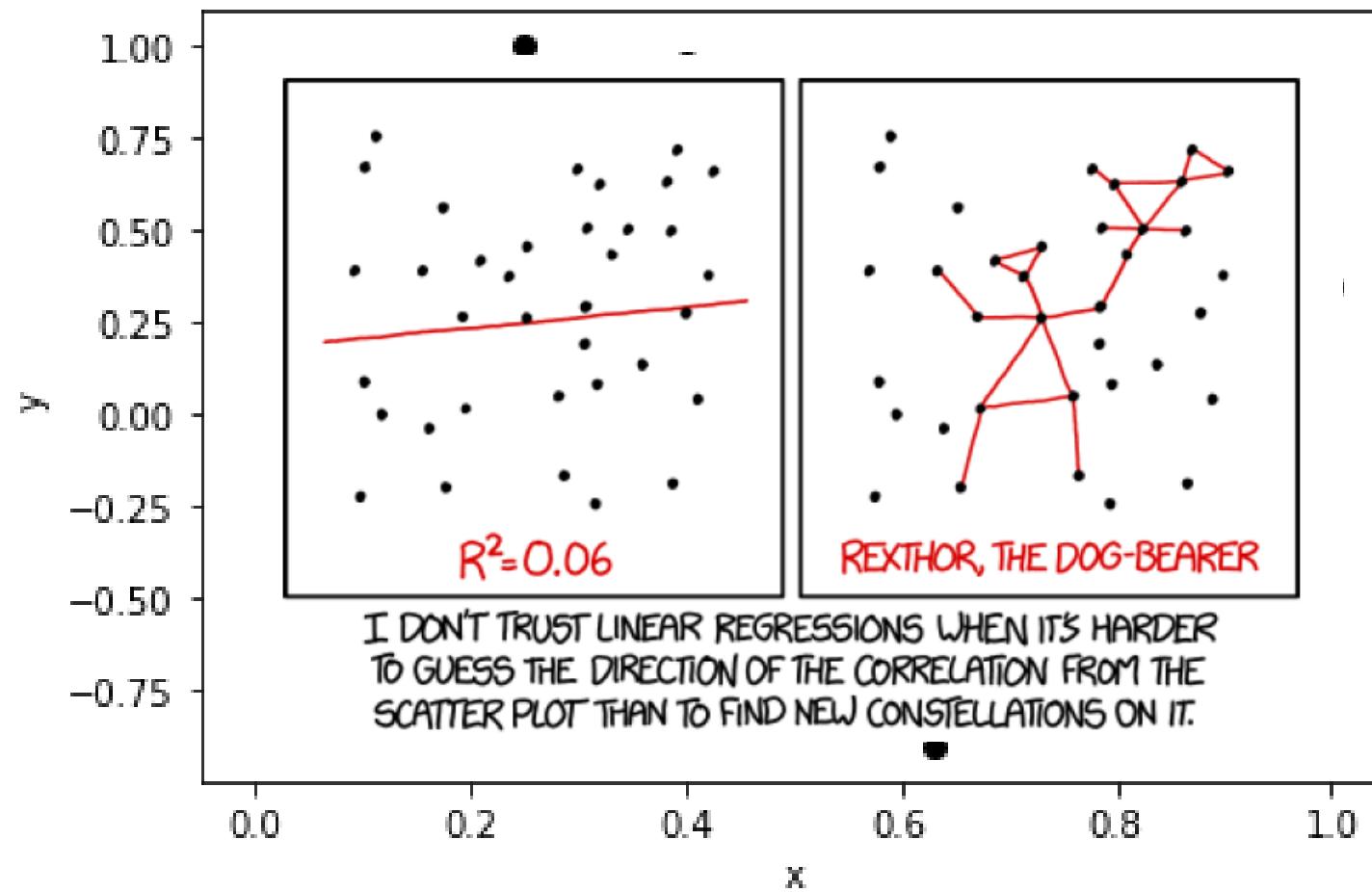
$$x = [0., 0.18, 0.25, 0.4, 0.45, 0.55, 0.63, 0.75, 0.85, 1.]$$

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

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?



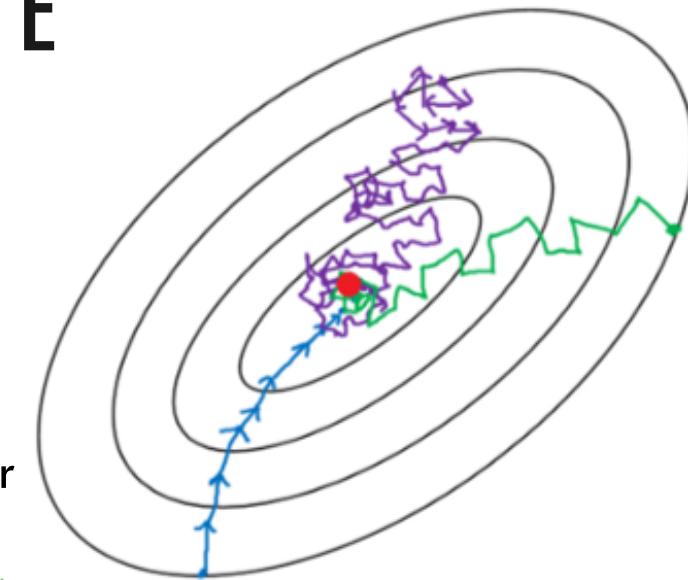


	$n = 0$	$n = 1$	$n = 4$	$n = 9$
ω_0	0.155	-1.31	0.27	0.30
ω_1		0.82	5.41	2083.08
ω_2			-7.77	39882.57
ω_3			-15.21	315869.13
ω_4			17.67	-1355675.15
ω_5				3466475.49
ω_6				-5433998.03
ω_7				5120734.79
ω_8				-2661308.98
ω_9				585702.29

VARIAÇÕES DO GRADIENTE DESCENDENTE

Gradiente descendente em lote (Batch Gradient Descent, BGD): calcula-se o **gradiente** usando **todo o dataset de treinamento** em cada iteração, para atualizar os parâmetros. Portanto, o que fizemos até agora...

Mas se o número de exemplos de treinamento for grande, então o gradiente descendente em lote é computacionalmente muito caro! Imagine se você tem 10000 dados, cada dado com 10 features, são 100mil valores para computar a cada iteração, em cada época...



Gradiente descendente em mini lotes (Mini-batch Gradient Descent, MBGD): Este é um tipo de gradiente de descida que funciona mais rápido. Calcula-se o **gradiente** usando **$b < m$ dados do dataset** em cada interação, para atualizar os parâmetros.

Gradiente descendente estocástico (Stochastic Gradient Descent, SGD): calcula-se o **gradiente** usando **$b = 1$ dados aleatórios de treinamento por iteração**, para atualizar os parâmetros. O SGD converge mais rapidamente para conjuntos de dados maiores. Porém, como no SGD usamos apenas um dado de cada vez, não podemos usar implementação vetorizada. Isso pode desacelerar os cálculos.

CADA ÉPOCA (EPOCH)

Gradiente descendente em lote (BGD)

Tomamos a média dos gradientes de **todos os exemplos de treinamento** e usamos esse gradiente médio para atualizar nossos parâmetros.

Gradiente descendente estocástico (SGD):

Tomamos **um exemplo de treinamento** para cálculo do gradiente e usamos esse gradiente médio para atualizar nossos parâmetros.

Gradiente descendente em mini lotes (MBGD)

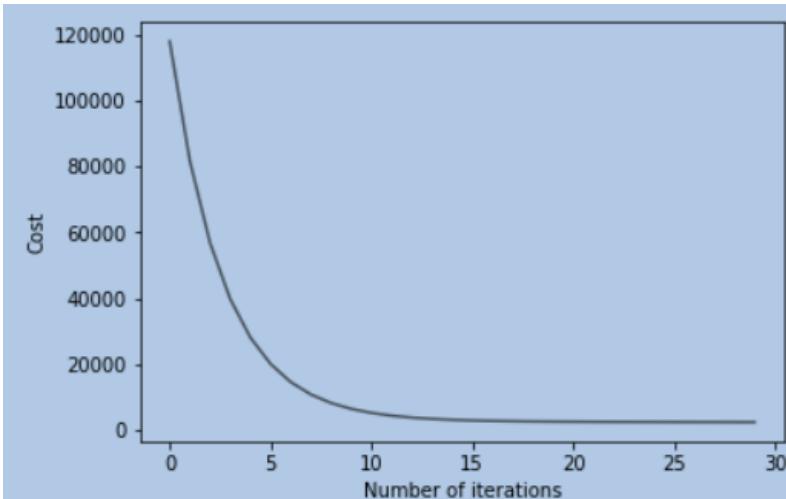
O mini lote tenta encontrar um equilíbrio entre **BGD** e **SGD**.

Para cada época:

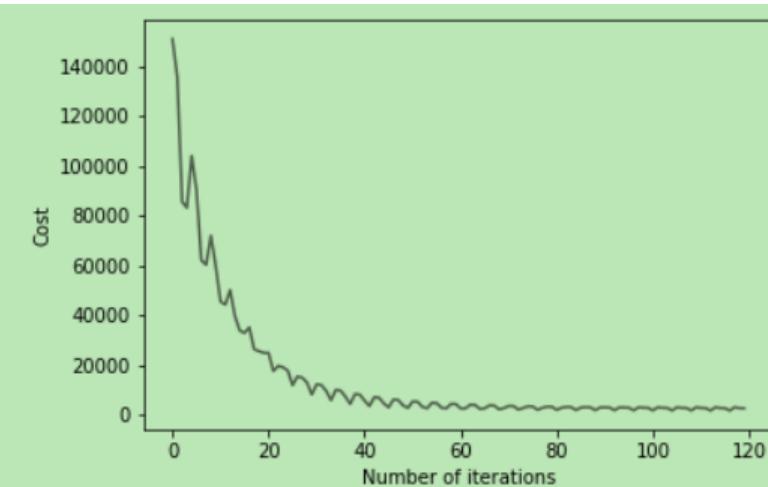
1. Use os dados de treinamento: **O LOTE, 1 MINI LOTE ou UM ÚNICO DADO**
2. Calcule o gradiente
3. Use o gradiente calculado na etapa anterior para atualizar os pesos
4. Repita as etapas 1 a 3 para todos os exemplos no conjunto de dados de treinamento

$$\begin{aligned}
 x &= [0.5 \quad 1.0 \quad 1.5 \quad 2.0 \quad 2.5 \quad 3.0 \quad 3.5 \quad 4.0 \quad 4.5 \quad 5.0 \quad 5.5 \quad 6.0 \quad 6.5 \quad 7.0]; \\
 y &= [173 \quad 284 \quad 310 \quad 439 \quad 485 \quad 520 \quad 605 \quad 699 \quad 728 \quad 735 \quad 790 \quad 801 \quad 840 \quad 850]
 \end{aligned}$$

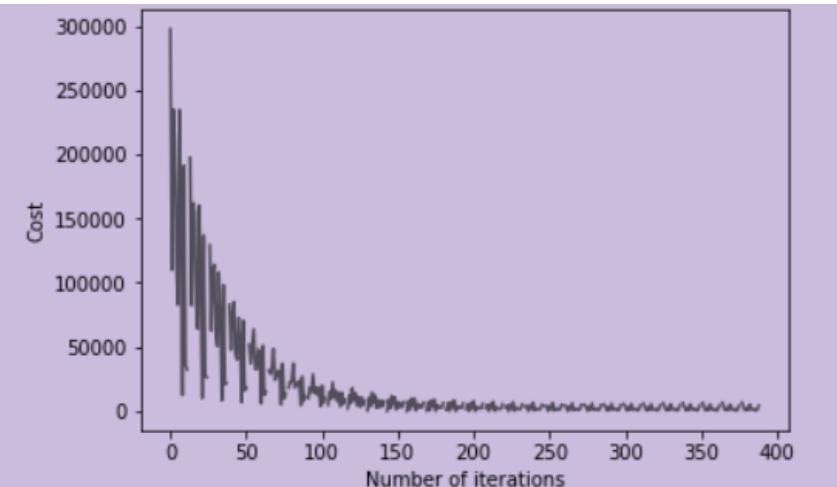
Batch Gradient Descent, BGD



Mini-batch Gradient Descent, MBGD

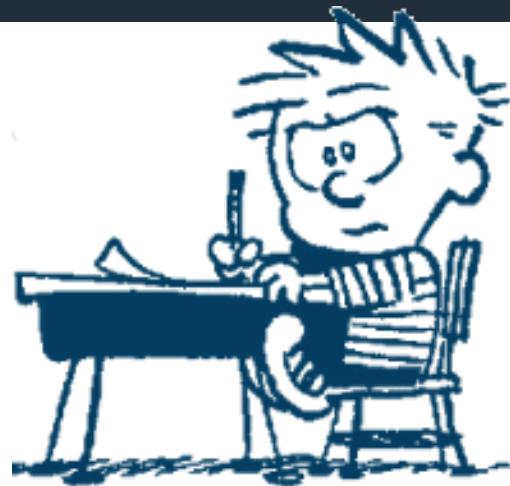


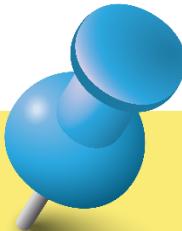
Stochastic Gradient Descent, SGD





REVISÃO





Lição de casa

Você tem até domingo,
23:59hs. Estude os slides da
aula de hoje e entre no Moodle e
faça o teste com 3 questões
múltipla escolha.

NEVER
GIVE UP



ACABOU...

Reveja a aula antes
de resolver os
exercícios.