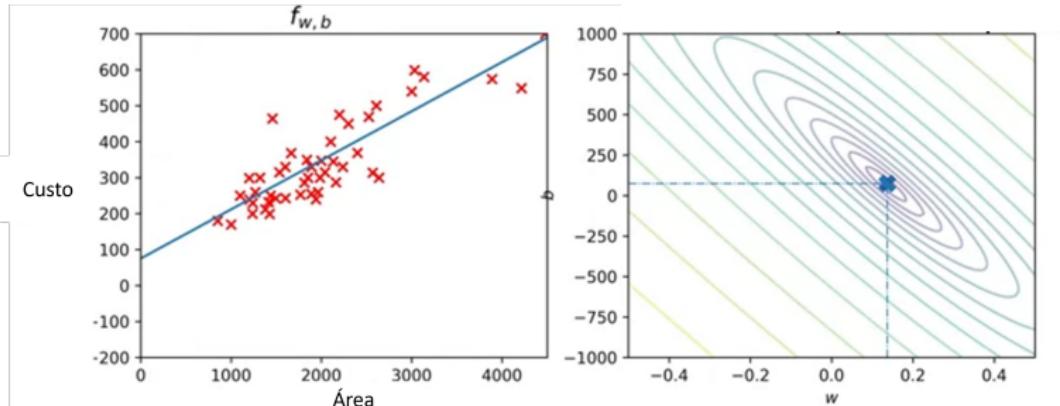


Minimizando a função custo pelo gradiente



Onde estamos e para onde vamos?

Na aula anterior, vimos que minimizando a função custo $J(w, b)$ nós conseguimos obter um modelo que minimiza a soma dos erros quadráticos $\left(f_{w,b}(x^{(i)}) - y^{(i)}\right)^2$, onde $f_{w,b}(x^{(i)})$ denota a previsão feita pelo modelo.



Pergunta:

Seria possível programar um algoritmo que busca automaticamente os parâmetros do modelo w, b que minimizam a função custo?

Resposta:

Sim. Nós faremos isso agora usando o **Método do Gradiente**.

Definição informal:

O **Método do Gradiente** consiste numa forma sistemática de busca por parâmetros (valores numéricos) que minimizam uma dada função.

Definição informal:

O **Método do Gradiente** consiste numa forma sistemática de busca por parâmetros (valores numéricos) que minimizam uma dada função.

No nosso caso...

No nosso caso, aplicaremos o **Método do Gradiente** para buscar sistematicamente os valores de w e b que minimizam a função custo

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)^2$$

Observações importantes:

O Método do Gradiente:

- pode ser usado para minimizar diversos tipos de funções (não precisa ser necessariamente uma função custo), e não se limita apenas a dois parâmetros.
- é vastamente utilizado na área de Aprendizado de Máquina, desde algoritmos mais simples (como o nosso, por enquanto), até algoritmos altamente avançados e sofisticados, como redes neurais profundas, etc.

No nosso caso...

Temos a função custo

$$J(w, b)$$

e queremos

$$\min_{w,b} J(w, b)$$

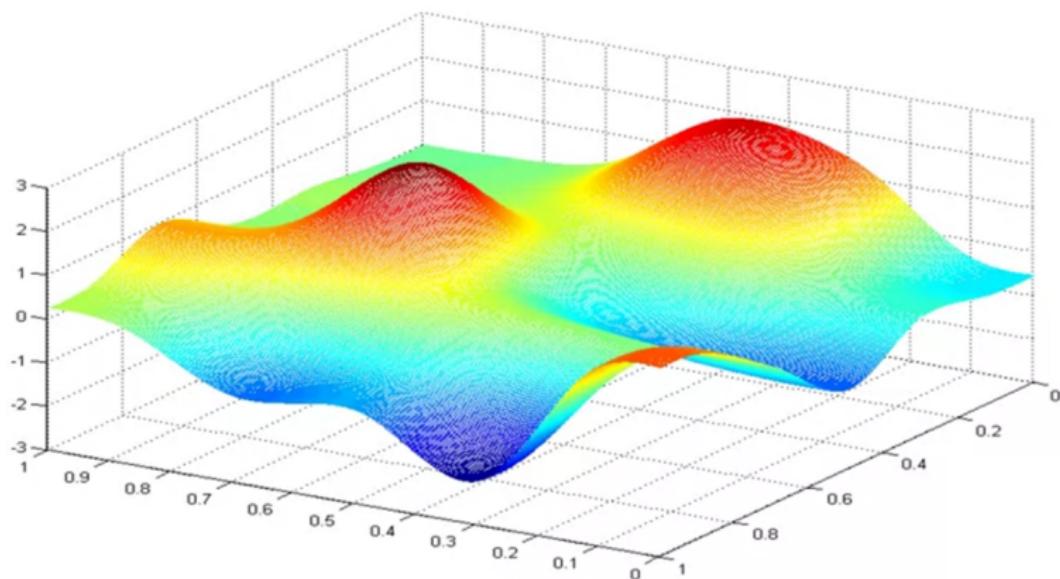
Observação

O Método do Gradiente:

- pressupõe um palpite inicial para os parâmetros. Ou seja, teremos que iniciar o algoritmo com valores para w , b . Por exemplo, podemos começar com $w = b = 0$.
- altera os valores de w , b com o objetivo de reduzir $J(w, b)$ até que estejamos próximos de um valor mínimo

Observação importante

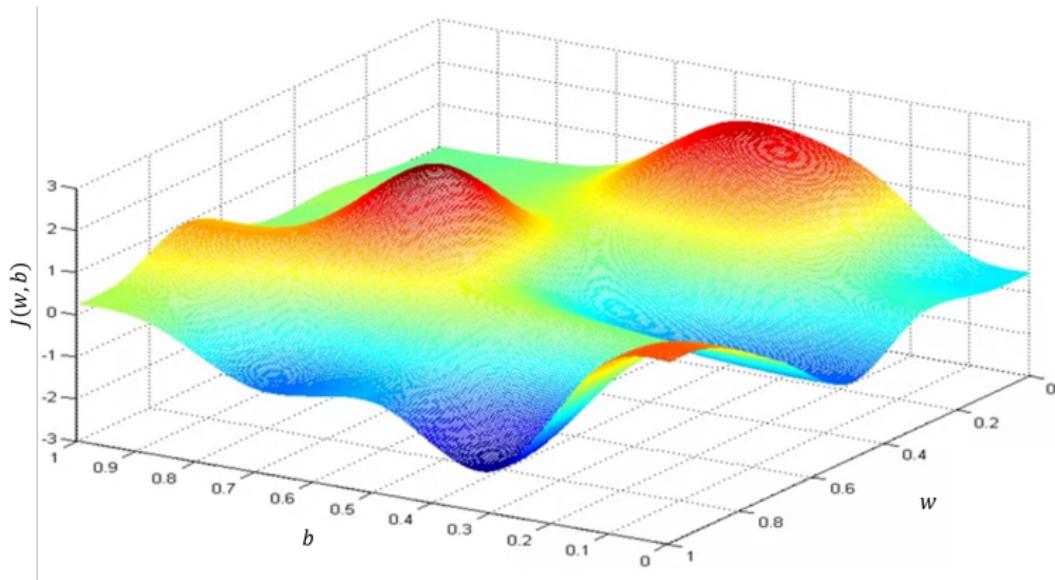
Funções mais complexas podem ter mais do que 1 mínimo.



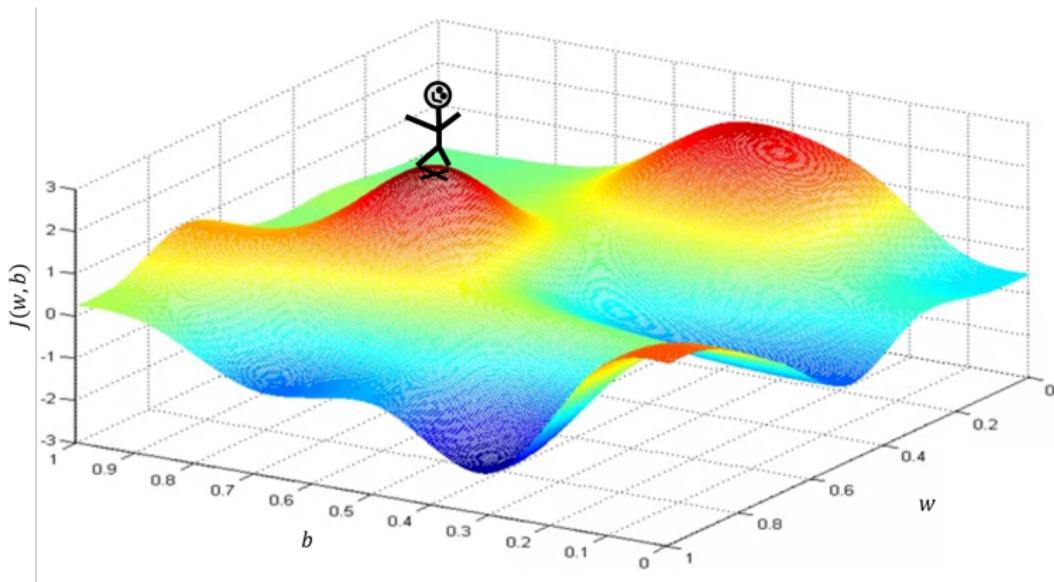
Observação

Esse tipo de função mais complexa é obtida quando treinamos modelos mais complexos, como redes neurais, por exemplo.

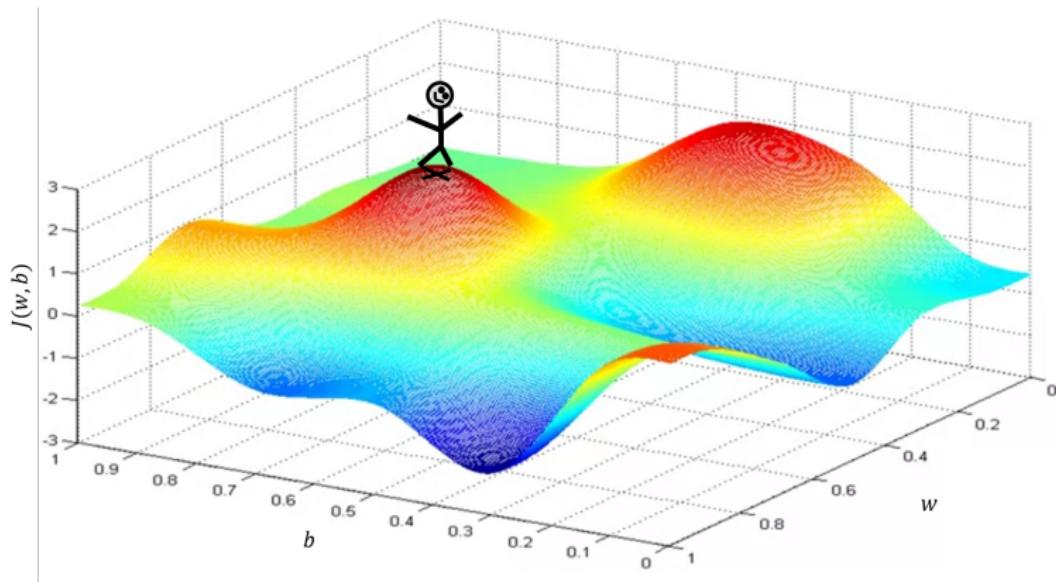
Suponha agora que os eixos são w , b e $J(w, b)$.



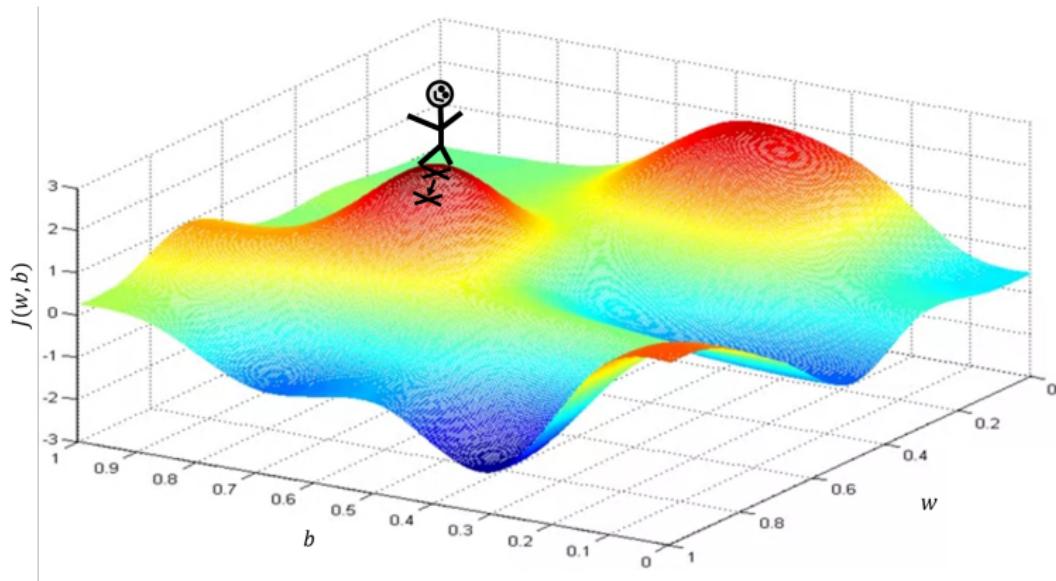
Você está numa posição específica dessa montanha e deseja chegar no seu ponto mais baixo.



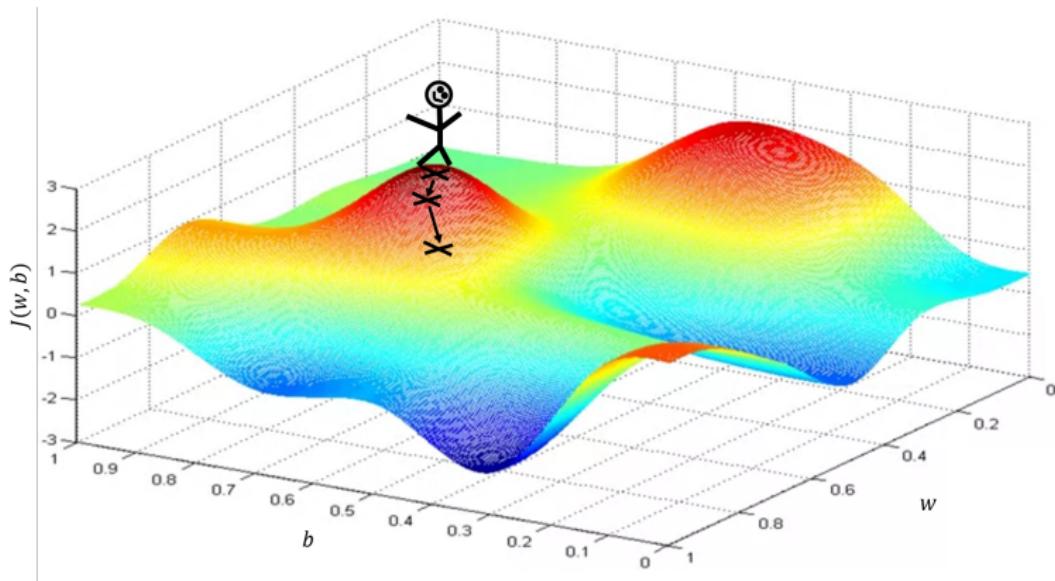
Você olha ao redor e tem que decidir para qual lado você dará o seu próximo passo. Para qual direção você iria?



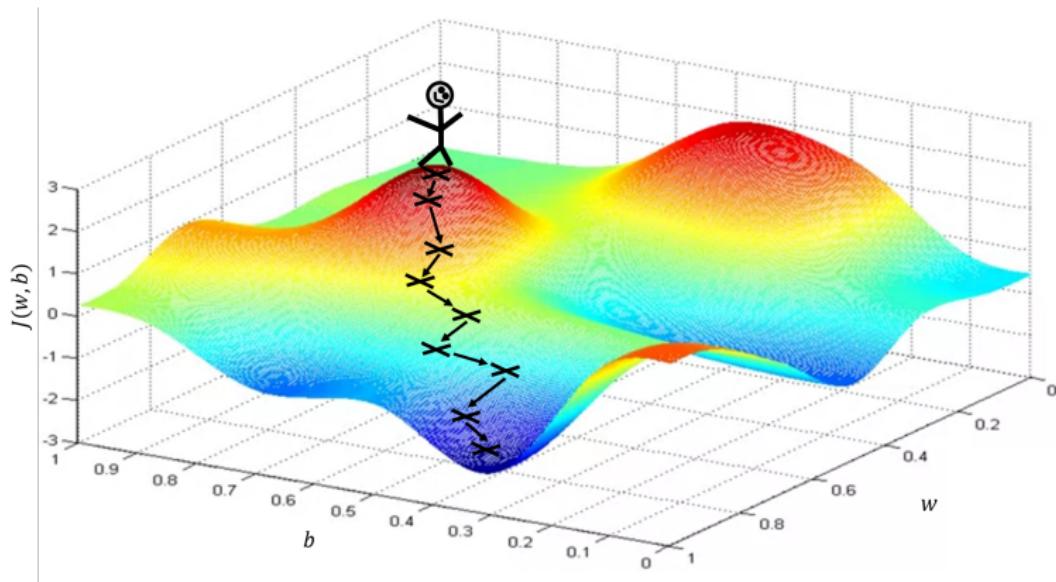
Após olhar ao redor, você decide avançar na direção indicada.



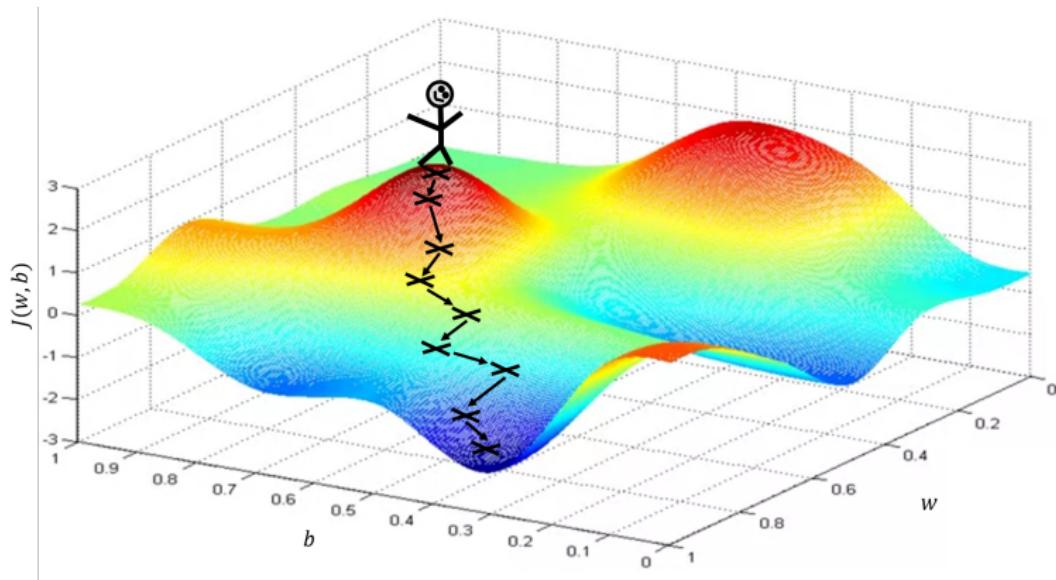
Agora você repete o processo. Você olha ao redor e decide seu próximo passo. Você então decide avançar mais um pouco.



Você pode repetir esse processo iterativamente até que você chegue no ponto desejado.



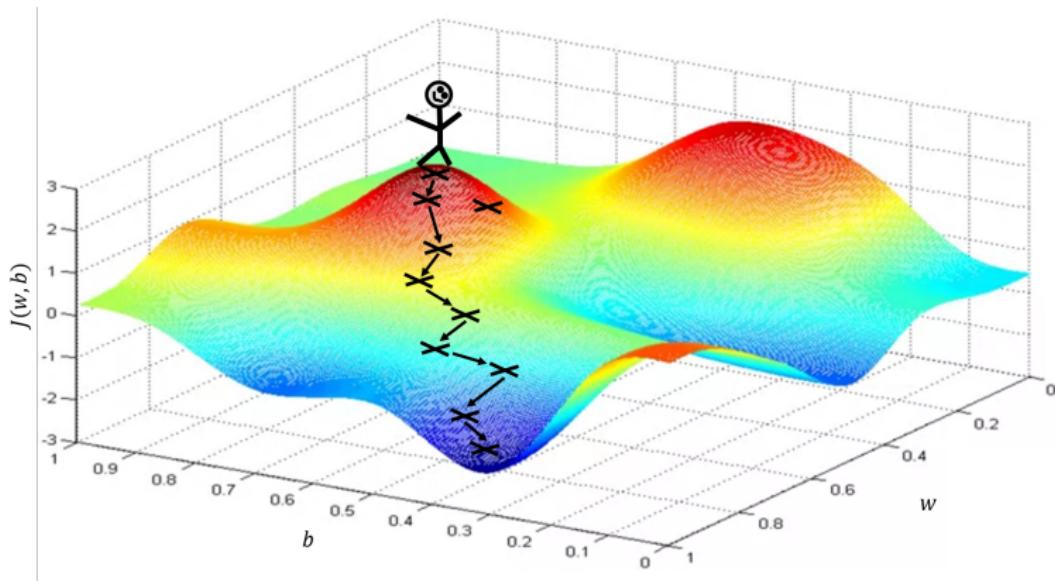
Você pode repetir esse processo iterativamente até que você chegue no ponto desejado.



Observação

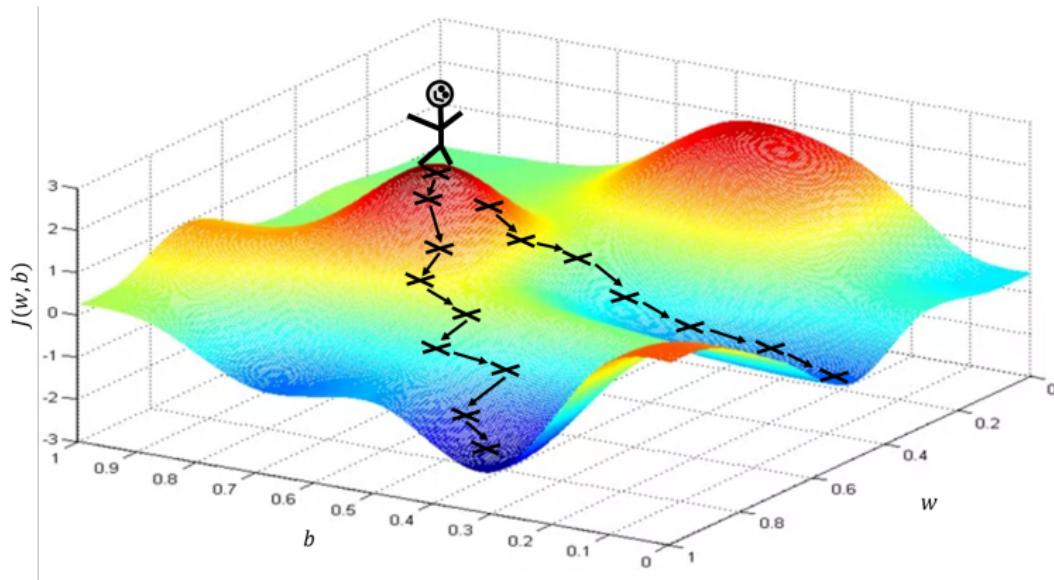
O Método do Gradiente faz justamente isso. Por isso, ele também é chamado de método do gradiente descendente, do inglês, *gradient descent*.

Lembre-se que você precisa escolher a posição inicial no Método do Gradiente. Vamos considerar agora que você escolheu uma posição inicial um pouco mais à direita.



Método do Gradiente

Acontece que, no Método do Gradiente, ainda que você escolha uma posição inicial apenas ligeiramente diferente, você pode acabar com parâmetros w e b bem diferentes...



Pergunta:

Qual solução é um ótimo local e qual é um ótimo global?

Pergunta:

Mas como implementamos na prática o Método do Gradiente?

Resposta:

Veremos isso agora

Em cada passo (iteração), devemos...

Atualizar w da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

Método do Gradiente: Como implementar

Em cada passo (iteração), devemos...

Atualizar w da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

- "Atualize seu parâmetro w a partir do seu valor atual ajustado por um pequeno valor dado pelo termo $\alpha \frac{d}{dw} J(w, b)$ "

Em cada passo (iteração), devemos...

Atualizar w da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

- "Atualize seu parâmetro w a partir do seu valor atual ajustado por um pequeno valor dado pelo termo $\alpha \frac{d}{dw} J(w, b)$ "
- o sinal = na expressão denota a operação de atribuição (não é igualdade) → (exemplo parecido: $i = i + 1$)

Em cada passo (iteração), devemos...

Atualizar w da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

- "Atualize seu parâmetro w a partir do seu valor atual ajustado por um pequeno valor dado pelo termo $\alpha \frac{d}{dw} J(w, b)$ "
- o sinal = na expressão denota a operação de atribuição (não é igualdade) → (exemplo parecido: $i = i + 1$)
- símbolo α (letra grega "alfa") denota a chamada taxa de aprendizado → com ela, você consegue controlar o quanto você quer que o termo $\frac{d}{dw} J(w, b)$ impacte na atualização do parâmetro w

Em cada passo (iteração), devemos...

Atualizar w da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

- "Atualize seu parâmetro w a partir do seu valor atual ajustado por um pequeno valor dado pelo termo $\alpha \frac{d}{dw} J(w, b)$ "
- o sinal = na expressão denota a operação de atribuição (não é igualdade) → (exemplo parecido: $i = i + 1$)
- símbolo α (letra grega "alfa") denota a chamada taxa de aprendizado → com ela, você consegue controlar o quanto você quer que o termo $\frac{d}{dw} J(w, b)$ impacte na atualização do parâmetro w
- um valor elevado para α indica um processo de aprendizado agressivo, onde w será drasticamente atualizado pelo termo $\frac{d}{dw} J(w, b)$ (cuidado! você pode acabar se perdendo na sua busca pelo mínimo da função)

Em cada passo (iteração), devemos...

Atualizar w da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

- "Atualize seu parâmetro w a partir do seu valor atual ajustado por um pequeno valor dado pelo termo $\alpha \frac{d}{dw} J(w, b)$ "
- o sinal = na expressão denota a operação de atribuição (não é igualdade) → (exemplo parecido: $i = i + 1$)
- símbolo α (letra grega "alfa") denota a chamada taxa de aprendizado → com ela, você consegue controlar o quanto você quer que o termo $\frac{d}{dw} J(w, b)$ impacte na atualização do parâmetro w
- um valor elevado para α indica um processo de aprendizado agressivo, onde w será drasticamente atualizado pelo termo $\frac{d}{dw} J(w, b)$ (cuidado! você pode acabar se perdendo na sua busca pelo mínimo da função)
- um valor muito pequeno para α indica um processo de aprendizado lento, onde w será pouco afetado pelo termo $\frac{d}{dw} J(w, b)$

Em cada passo (iteração), devemos...

Atualizar w da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

- "Atualize seu parâmetro w a partir do seu valor atual ajustado por um pequeno valor dado pelo termo $\alpha \frac{d}{dw} J(w, b)$ "
- o sinal = na expressão denota a operação de atribuição (não é igualdade) → (exemplo parecido: $i = i + 1$)
- símbolo α (letra grega "alfa") denota a chamada taxa de aprendizado → com ela, você consegue controlar o quanto você quer que o termo $\frac{d}{dw} J(w, b)$ impacte na atualização do parâmetro w
- um valor elevado para α indica um processo de aprendizado agressivo, onde w será drasticamente atualizado pelo termo $\frac{d}{dw} J(w, b)$ (cuidado! você pode acabar se perdendo na sua busca pelo mínimo da função)
- um valor muito pequeno para α indica um processo de aprendizado lento, onde w será pouco afetado pelo termo $\frac{d}{dw} J(w, b)$
- $\frac{d}{dw} J(w, b)$ é a derivada da função $J(w, b)$ em relação ao parâmetro w → mostra a direção em que $J(w, b)$ mais cresce para uma pequena variação de w .

Método do Gradiente: Como implementar

Em cada passo (iteração), devemos...

Atualizar w da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

- "Atualize seu parâmetro w a partir do seu valor atual ajustado por um pequeno valor dado pelo termo $\alpha \frac{d}{dw} J(w, b)$ "
- o sinal = na expressão denota a operação de atribuição (não é igualdade) → (exemplo parecido: $i = i + 1$)
- símbolo α (letra grega "alfa") denota a chamada taxa de aprendizado → com ela, você consegue controlar o quanto você quer que o termo $\frac{d}{dw} J(w, b)$ impacte na atualização do parâmetro w
- um valor elevado para α indica um processo de aprendizado agressivo, onde w será drasticamente atualizado pelo termo $\frac{d}{dw} J(w, b)$ (cuidado! você pode acabar se perdendo na sua busca pelo mínimo da função)
- um valor muito pequeno para α indica um processo de aprendizado lento, onde w será pouco afetado pelo termo $\frac{d}{dw} J(w, b)$
- $\frac{d}{dw} J(w, b)$ é a derivada da função $J(w, b)$ em relação ao parâmetro w → mostra a direção em que $J(w, b)$ mais cresce para uma pequena variação de w .

Observação

Lembre-se que também temos o parâmetro b no nosso modelo, então também teremos que atualizá-lo a partir da expressão equivalente

$$b = b - \alpha \frac{d}{db} J(w, b)$$

Em cada passo (iteração), devemos...

Atualizar w e b da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

$$b = b - \alpha \frac{d}{db} J(w, b)$$

Importante:

- Devemos repetir esse processo até a **convergência**, ou seja, até que w e b deixem de atualizar (isso ocorre quando estamos próximos de um mínimo da função J)

Em cada passo (iteração), devemos...

Atualizar w e b da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

$$b = b - \alpha \frac{d}{db} J(w, b)$$

Forma **correta** de implementação (atualização simultânea):

$$tmp_w = w - \alpha \frac{d}{dw} J(w, b)$$

$$tmp_b = b - \alpha \frac{d}{db} J(w, b)$$

$$w = tmp_w$$

$$b = tmp_b$$

Em cada passo (iteração), devemos...

Atualizar w e b da seguinte maneira:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

$$b = b - \alpha \frac{d}{db} J(w, b)$$

Forma **INCORRETA** de implementação:

$$\text{tmp_}w = w - \alpha \frac{d}{dw} J(w, b)$$

$$w = \text{tmp_}w$$

$$\text{tmp_}b = b - \alpha \frac{d}{db} J(w, b)$$

$$b = \text{tmp_}b$$

Pergunta:

O Método do Gradiente é um algoritmo que visa encontrar os valores dos parâmetros w e b que minimizam a função custo J . O que a expressão abaixo faz? (assuma que α é pequeno)

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

- A) Checa-se w é igual a $w - \alpha \frac{d}{dw} J(w, b)$
- B) Atualiza w ligeiramente

Vimos até agora que o Método do Gradiente pode ser implementado da seguinte forma:

repetir até a convergência:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

$$b = b - \alpha \frac{d}{db} J(w, b)$$

onde α denota a taxa de aprendizado (geralmente um valor positivo $\alpha > 0$ pequeno).

Pergunta:

E se quisermos encontrar o mínimo de uma função f que possui apenas 1 parâmetro? Por exemplo, $J(w)$?

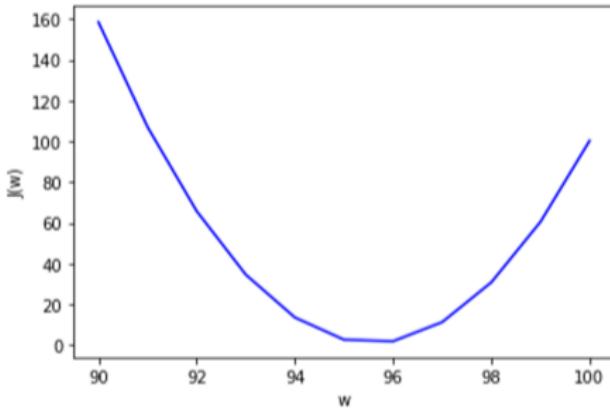
Resposta:

Nesse caso, precisamos fazer apenas

$$w = w - \alpha \frac{d}{dw} J(w)$$

Exemplo

Na atividade de programação da última aula, obtivemos a seguinte função $J(w)$ quando consideramos $b = 0$:



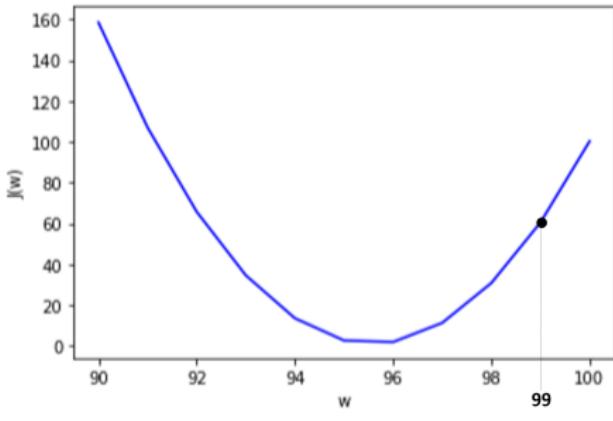
Pergunta:

Como ficaria a aplicação do Método do Gradiente para esse caso? Ou seja, como ficaria a atualização

$$w = w - \alpha \frac{d}{dw} J(w) ?$$

Exemplo

Considerando uma inicialização em $w = 99$, temos:

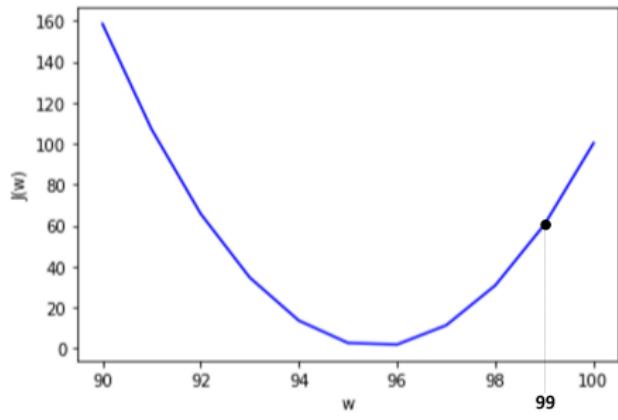


Pergunta:

Sabendo que a derivada é a inclinação da reta que tangencia o ponto $w = 99$, teremos $\frac{d}{dw} J(w) > 0$ ou $\frac{d}{dw} J(w) < 0$?

Exemplo

Considerando uma inicialização em $w = 99$, temos:



Observação:

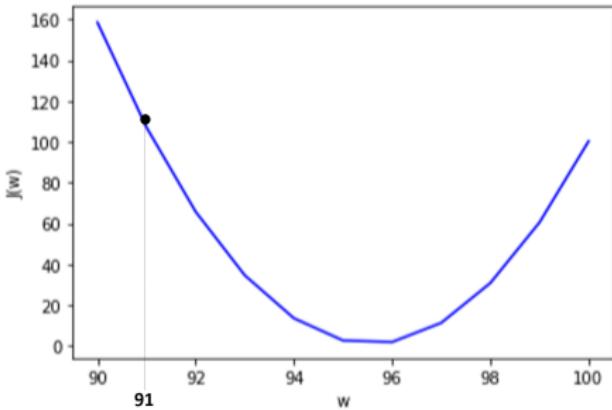
Como $\frac{d}{dw} J(w) > 0$ para $w = 99$, note que teremos

$$w = w - \alpha \text{ (valor positivo)}$$

Ou seja, o valor atualizado para w será **menor** que 99, já que $\alpha > 0$. Para um valor apropriado para α , estaremos caminhando em direção ao mínimo da função $J(w)$.

Exemplo

Supondo agora uma inicialização em $w = 91$, temos:

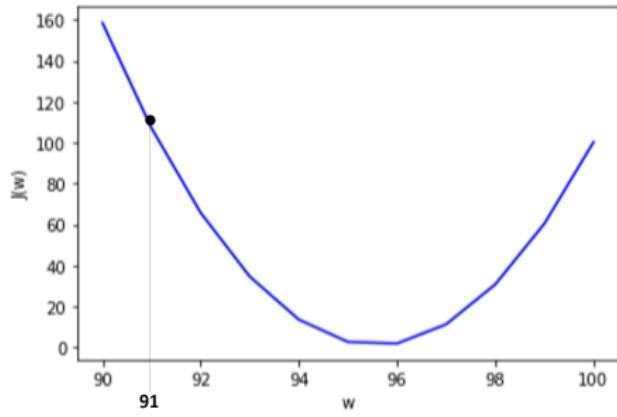


Pergunta:

Sabendo que a derivada é a inclinação da reta que tangencia o ponto $w = 91$, teremos $\frac{d}{dw} J(w) > 0$ ou $\frac{d}{dw} J(w) < 0$?

Exemplo

Considerando uma inicialização em $w = 91$, temos:

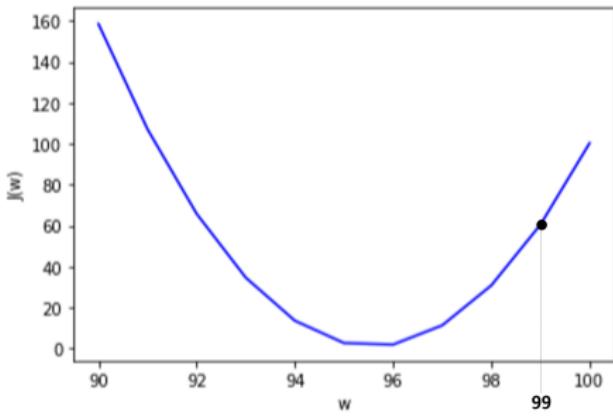


Observação:

Como $\frac{d}{dw} J(w) < 0$ para $w = 91$, note que teremos

$$w = w - \alpha (\text{valor negativo})$$

Ou seja, o valor atualizado para w será **maior** que 91, já que $\alpha > 0$. Para um valor apropriado para α , estaremos caminhando em direção ao mínimo da função $J(w)$.



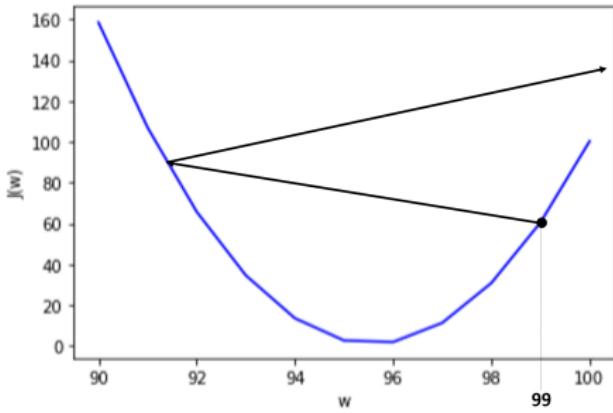
Pergunta:

O que acontece se tivermos um valor excessivamente grande para a taxa de aprendizado α , por exemplo, $\alpha = 1000000$?

Apenas relembrando que

$$w = w - \alpha \frac{d}{dw} J(w)$$

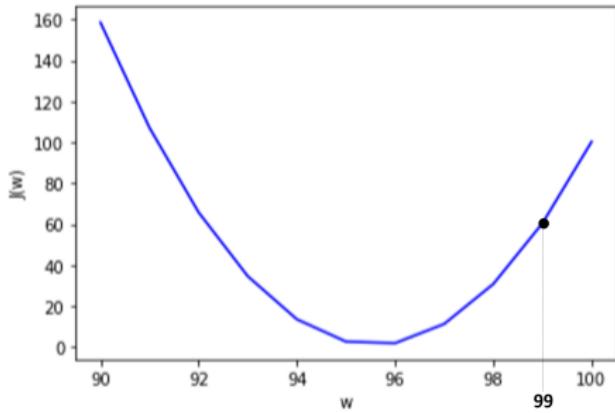
Exemplo



Resposta:

O Método do Gradiente irá divergir, e o valor de w que minimiza $J(w)$ nunca será encontrado.

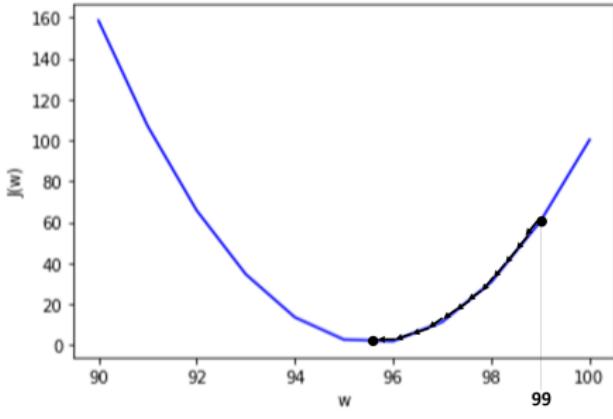
Exemplo



Pergunta:

E se tivermos um valor excessivamente pequeno para a taxa de aprendizado α , por exemplo, $\alpha = 0.000001$?

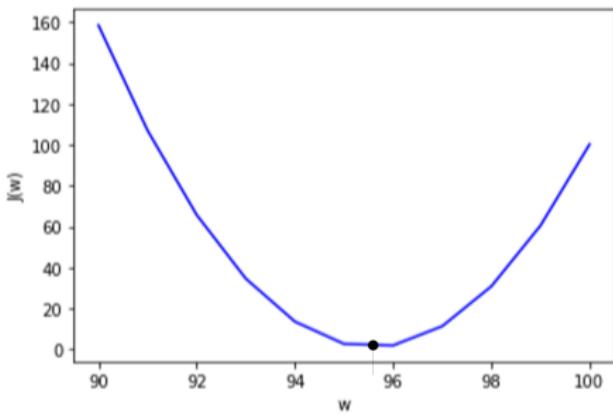
Exemplo



Resposta:

O Método do Gradiente será lento, e demorará muitas iterações até convergir para o mínimo.

Exemplo



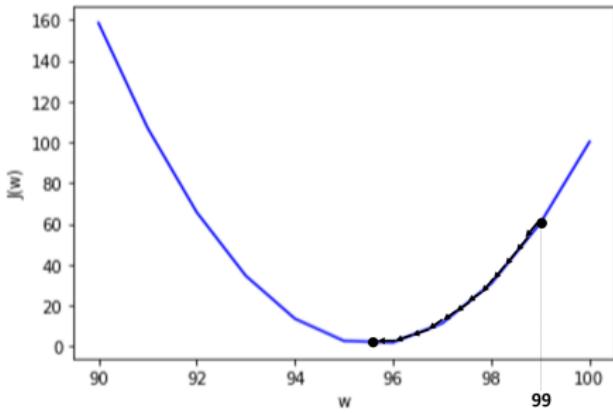
Pergunta:

O que acontece se inicializarmos w já no ponto de mínimo da função $J(w)$?

Resposta:

O método permanecerá no mínimo, já que, no mínimo, temos $\frac{d}{dw} J(w) = 0$.

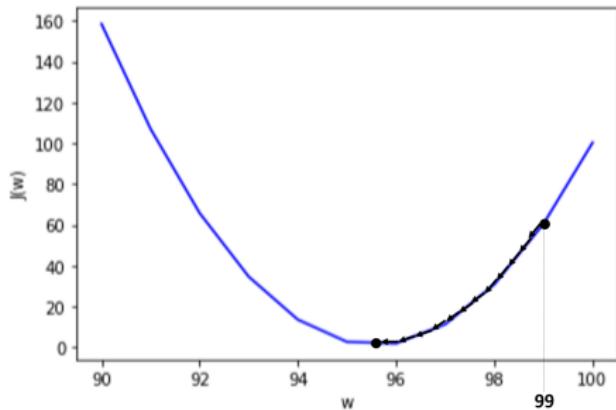
Exemplo



Observação:

Note que $\frac{d}{dw} J(w)$ diminui (em termos do seu módulo) à medida com que nos aproximamos do mínimo da função. Isso significa que, utilizando um valor pequeno e fixo para α , passos cada vez menores em direção ao mínimo são dados ao longo das iterações.

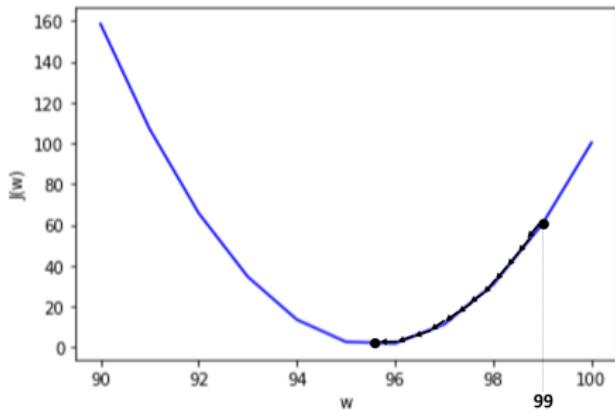
Exemplo



Pergunta:

Como fazemos então para selecionarmos um valor adequado para α ?

Exemplo



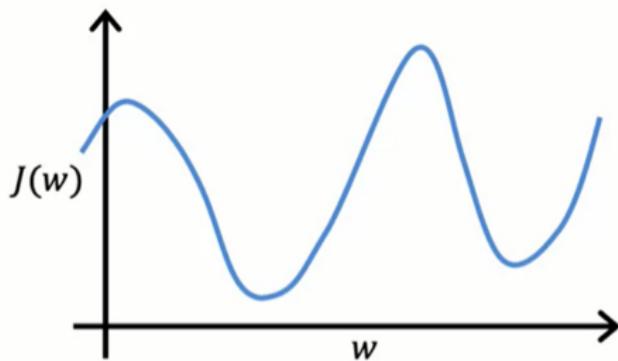
Pergunta:

Como fazemos então para selecionarmos um valor adequado para α ?

Resposta:

Depende muito de cada problema. Dica: teste diferentes valores!

Exemplo



Pergunta:

Caso estejamos minimizando uma função $J(w)$ que possui diversos mínimos locais, é possível que o Método do Gradiente não converja para o mínimo global?

Resposta:

Sim. Note que o mínimo para o qual o método irá convergir depende do ponto de inicialização.

Voltando para o nosso problema de regressão linear

Queremos utilizar o **Método do Gradiente** para encontrar os parâmetros w, b do modelo $f_{w,b}(x) = wx + b$ que minimizam a função custo

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)^2$$

Em outras palavras, queremos usar o Método do Gradiente para encontrar a **reta que melhor representa nossos dados**.

Para fazermos isso, basta escolhermos valores iniciais para w, b e...

repetir até a convergência:

$$w = w - \alpha \frac{d}{dw} J(w, b)$$

$$b = b - \alpha \frac{d}{db} J(w, b)$$

Para essa função custo, é possível mostrar que

$$\frac{d}{dw} J(w, b) = \frac{1}{m} \sum_{i=1}^m \left(f_{w,b}(x^{(i)}) - y^{(i)} \right) x^{(i)}$$

$$\frac{d}{db} J(w, b) = \frac{1}{m} \sum_{i=1}^m \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)$$

Como deduzir as expressões acima?

Ou seja, encontraremos os parâmetros w, b do modelo $f_{w,b}(x) = wx + b$ que minimizam a função custo

$$J(w, b) = \frac{1}{2m} \sum_{i=1}^m \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)^2$$

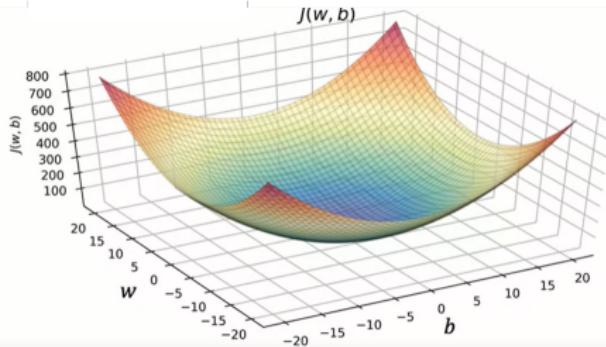
escolhendo valores iniciais para w, b e então repetindo até a convergência as seguintes atualizações:

$$w = w - \alpha \frac{1}{m} \sum_{i=1}^m \left(f_{w,b}(x^{(i)}) - y^{(i)} \right) x^{(i)}$$

$$b = b - \alpha \frac{1}{m} \sum_{i=1}^m \left(f_{w,b}(x^{(i)}) - y^{(i)} \right)$$

Observação importante:

Em problemas de **regressão linear**, a função quadrática J é **convexa**, ou seja, ela não possui outros mínimos locais além do próprio mínimo global.



Conhecendo esse resultado teórico, sabemos que o Método do Gradiente nos levará, obrigatoriamente, para o mínimo global da função.

De olho no código!

Vamos agora ver como implementar na prática o **Método do Gradiente**

Acesse o Python Notebook usando o QR code ou o link abaixo:



https://colab.research.google.com/github/xaximppv2/master/blob/main/codigo_aula5_metodo_do_gradiente.ipynb

Parte 1

Rode todo o código. Responda às questões nele contidas e complete-o, se necessário.

Parte 2

Insira no código da Parte 1 o conjunto de medições que você já criou anteriormente para um resistor de 50Ω , faça as adaptações necessárias e verifique os resultados.