Escalonamento de características e escolha do α

Onde estamos e para onde vamos

Nas aulas anteriores, mostramos como o Método do Gradiente pode ser usado para resolver diversos tipos de problemas

Nesta aula, aprenderemos como aprimorar esse método por meio do **Escalonamento de Características** e de uma escolha apropriada para a **taxa de aprendizado** α.

Escalonamento de Características

Vamos começar com a técnica de **Escalonamento de Características**. Tal técnica faz com que o Método do Gradiente torne-se significativamente mais rápido.

Relação entre características e valores dos parâmetros

Suponha que você deseja obter um modelo que seja capaz de estimar o preço de uma casa:

$$\widehat{\text{preço}} = w_1 x_1 + w_2 x_2 + b$$

onde

- x_1 denota o tamanho em **feet**^{2} (um valor tipicamente entre 300 e 2000)
- x_2 denota o número de quartos (um valor tipicamente entre 0 e 5)

Relação entre características e valores dos parâmetros

Suponha que você deseja obter um modelo que seja capaz de estimar o preço de uma casa:

$$\widehat{\text{preço}} = w_1 x_1 + w_2 x_2 + b$$

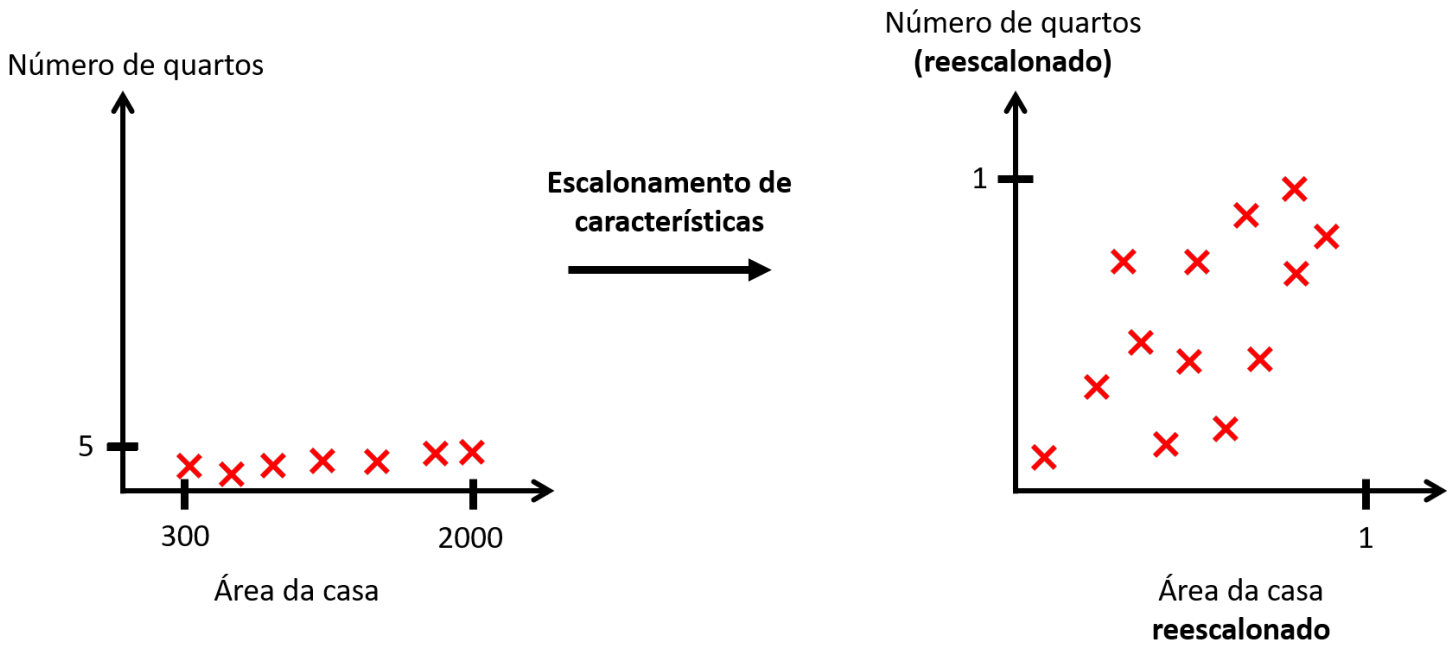
onde

- x_1 denota o tamanho em **feet**^{2} (um valor tipicamente entre 300 e 2000)
- x_2 denota o número de quartos (um valor tipicamente entre 0 e 5)

Observações:

- Para que ambos os termos $w_1 x_1$ e $w_2 x_2$ tenham impacto significativo no cálculo do preço da casa, note que w_1 tenderá a ser pequeno em comparação com w_2, já que x_1 é um valor tipicamente maior que x_2.
- Isso significa que, enquanto o Método do Gradiente estiver buscando valores apropriados para w_1 e w_2, o valor da função custo J será **muito mais sensível** a um incremento unitário em w_1 em comparação com um incremento unitário em w_2.
- Se o Método do Gradiente "erra" um pouco na escolha de w_1, a precisão do modelo deteriora consideravelmente.
- Por outro lado, se escalonarmos os dados antes de aplicar o método do gradiente, criando duas **características escalonadas** entre 0 e 1, por exemplo, o Método do Gradiente torna-se mais rápido para encontrar valores apropriados para w_1 e w_2.

Escalonamento de Características



Observação

O escalonamento das características desempenha um papel fundamental no aumento de velocidade de convergência do método do gradiente, especialmente quando as características do problema possuem valores com ordens de grandeza diversas.

OPÇÃO 1: Dividindo pelo máximo

Se $300 \leq x_1 \leq 2000$, podemos escalonar x_1 da seguinte maneira:

$$x_{1,escalonado} = \frac{x_1}{2000}$$

Assim, teremos $0.15 \leq x_{1,escalonado} \leq 1$

Similarmente, se $0 \leq x_2 \leq 5$, podemos escalonar x_2 da seguinte maneira:

$$x_{2,escalonado} = \frac{x_2}{5}$$

Assim, teremos $0 \leq x_{2,escalonado} \leq 1$

OPÇÃO 2: Normalização pela média

Se $300 \leq x_1 \leq 2000$, podemos escalonar x_1 da seguinte maneira:

$$x_{1,escalonado} = \frac{x_1 - \mu_1}{2000 - 300} \quad \rightarrow \quad \text{onde } \mu_1 \text{ é a média de } x_1$$

Supondo $\mu_1 = 600$ feet^2, teremos $-0.18 \leq x_{1,escalonado} \leq 0.82$

Similarmente, se $0 \leq x_2 \leq 5$, podemos escalonar x_2 da seguinte maneira:

$$x_{2,escalonado} = \frac{x_2 - \mu_2}{5 - 0} \quad \rightarrow \quad \text{onde } \mu_2 \text{ é a média de } x_2$$

Supondo $\mu_2 = 2.3$ quartos, teremos $-0.46 \leq x_{2,escalonado} \leq 0.54$

Observação:

- Ao subtrair a média de uma sequência de números, a sequência resultante acaba ficando com média 0.

OPÇÃO 3: Normalização Z-score $\rightarrow$ Também chamada de **Padronização**

Se $300 \leq x_1 \leq 2000$, podemos escalonar x_1 da seguinte maneira:

$$x_{1,escalonado} = \frac{x_1 - \mu_1}{\sigma_1} \quad \rightarrow \quad \text{onde } \sigma_1 \text{ é o desvio padrão de } x_1$$

Supondo $\mu_1 = 600$ feet^2 e $\sigma_1 = 450$, teremos $-0.67 \leq x_{1,escalonado} \leq 3.1$

Similarmente, se $0 \leq x_2 \leq 5$, podemos escalonar x_2 da seguinte maneira:

$$x_{2,escalonado} = \frac{x_2 - \mu_2}{\sigma_2} \quad \rightarrow \quad \text{onde } \sigma_2 \text{ é o desvio padrão de } x_2$$

Supondo $\mu_2 = 2.3$ quartos e $\sigma_2 = 1.4$, teremos $-1.6 \leq x_{2,escalonado} \leq 1.9$

Observação:

- Ao dividir uma sequência de números pelo seu desvio padrão, a sequência resultante acaba ficando com desvio unitário.

Exemplos de escalonamento

Intervalos aceitáveis onde não é necessário reescalonar:

- $-1 < x_j < 1$
- $-3 < x_j < 3$
- $-0.3 < x_j < 0.3$
- $0 < x_j < 3$
- $-2 < x_j < 0.5$

Intervalos não aceitáveis onde pode ser importante reescalonar:

- $-100 < x_j < 100$
- $-0.001 < x_j < 0.001$
- $98.6 < x_j < 105$

Observação:

Reescalonar quase sempre irá melhorar o desempenho do Método do Gradiente. Raramente irá prejudicar.

Which of the following is a valid step used during feature scaling?

- ◯ Divide each value by the maximum value for that feature
- ◯ Multiply each value by the maximum value for that feature

Fonte: **Machine Learning Specialization**, *deeplearning.ai*, Stanford Online, Coursera.org.

Convergência do Método do Gradiente

Perguntas

- Ao rodar o Método do Gradiente, **como saber se ele já convergiu**?
- Qual a relação entre convergência e a taxa de aprendizado α?

Observação

É importante que consigamos olhar para uma implementação do Método do Gradiente e reconhecer se ela está rodando corretamente ou não.

Convergência do Método do Gradiente

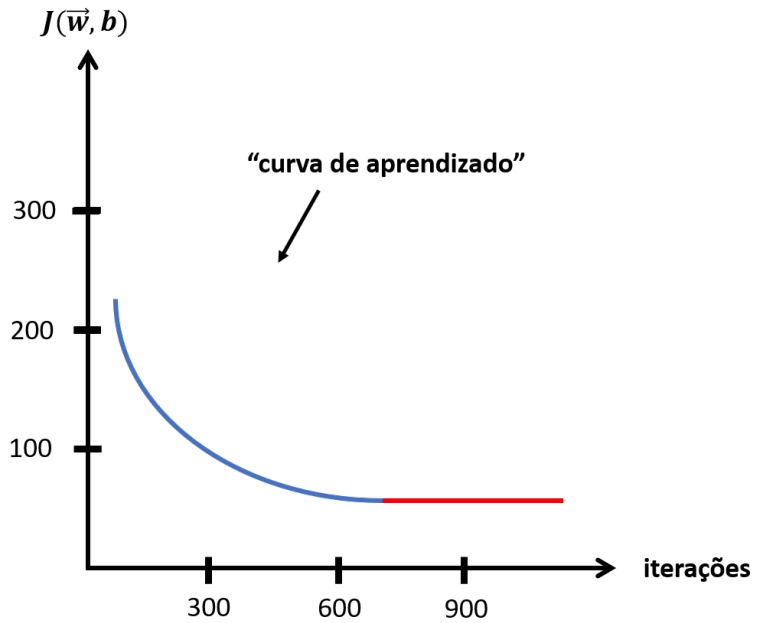
Apenas relembrando que o método do gradiente consiste em repetir até convergir:

$$w_j = w_j - \alpha \frac{d}{dw_j} J(\vec{w}, b)$$

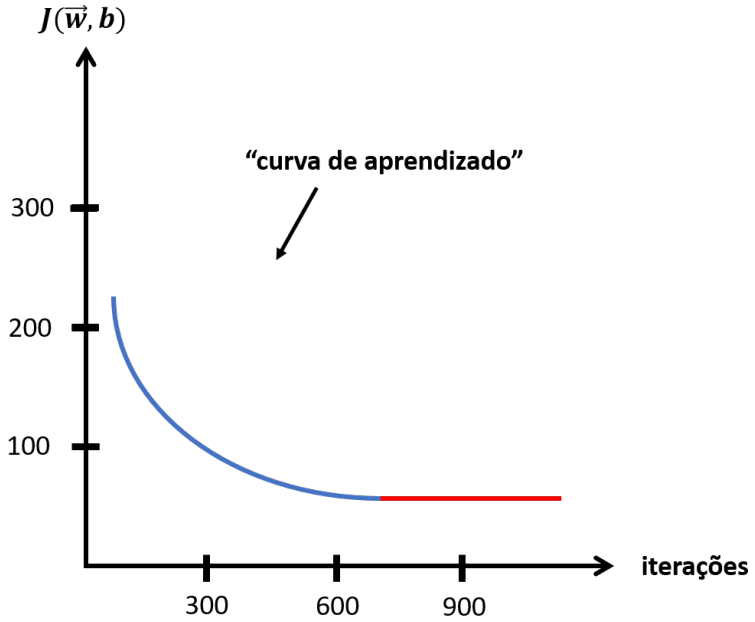
$$b = b - \alpha \frac{d}{db} J(\vec{w}, b)$$

Onde o objetivo dessas atualizações é encontrar os valores de $\vec{w}$ e b que minimizam $J(\vec{w}, b)$.

Portanto, é útil observarmos, ao longo das iterações:



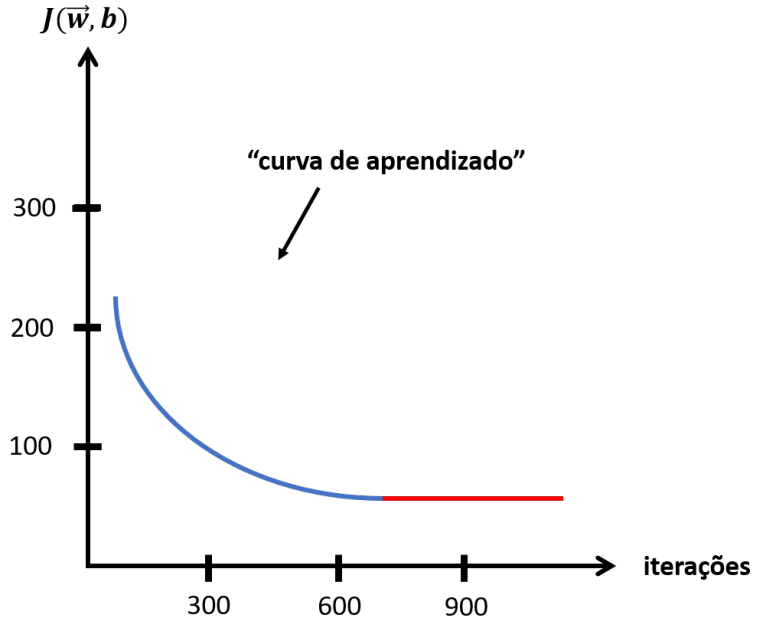
Convergência do Método do Gradiente



Observações:

- Após 300 iterações, o Método do Gradiente ainda está refinando significativamente os valores de w_j e b.
- Após 900 iterações, parece que o método já convergiu.
- O número de iterações que o Método do Gradiente leva para convergir pode variar bastante dependendo da aplicação (30 ou 100000).

Convergência do Método do Gradiente



Observações:

- Após 300 iterações, o Método do Gradiente ainda está refinando significativamente os valores de w_j e b.
- Após 900 iterações, parece que o método já convergiu.
- O número de iterações que o Método do Gradiente leva para convergir pode variar bastante dependendo da aplicação (30 ou 100000).

Observação final:

Após cada iteração, $J(\vec{w}, b)$ deve sempre decrescer. Se isso não ocorrer, então:

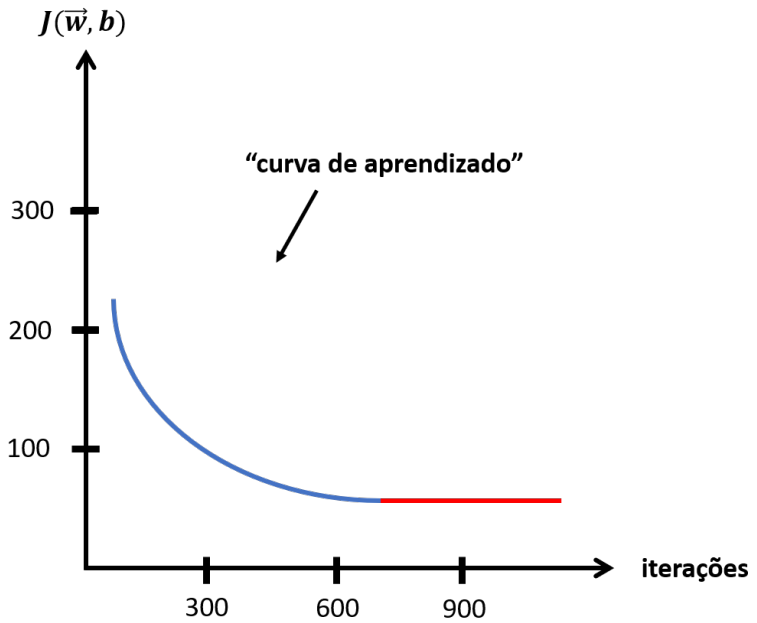
- Ou α não foi escolhido apropriadamente (geralmente α muito grande)
- Ou o Método do Gradiente não encontra-se implementado corretamente no código (*bug*)

Teste para detecção automática de convergência

Seja ε um valor pequeno, por exemplo, $\varepsilon = 10^{-3}$.
"*Se $J(\vec{w}, b)$ decresce menos que ε entre duas iterações consecutivas, então declarar convergência.*"

Observação

Encontrar um valor adequado para ε pode ser bastante desafiador. Portanto, na dúvida, olhe atentamente o gráfico!



Escolhendo α adequadamente

Escolhendo α adequadamente

Observações iniciais

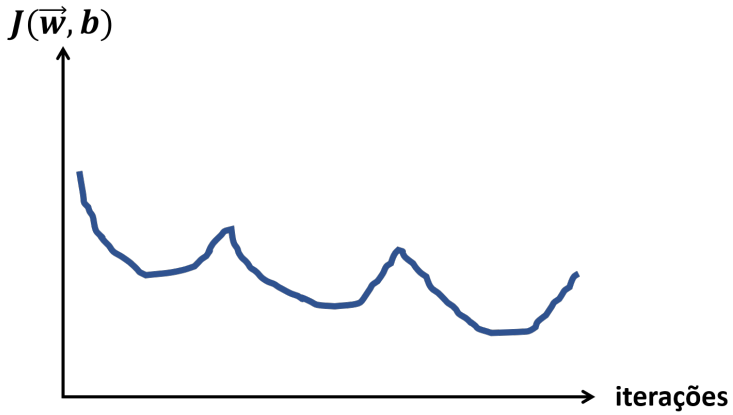
- Se α é muito pequeno, o aprendizado será lento
- Se α é muito grande, o método pode não convergir

Pergunta

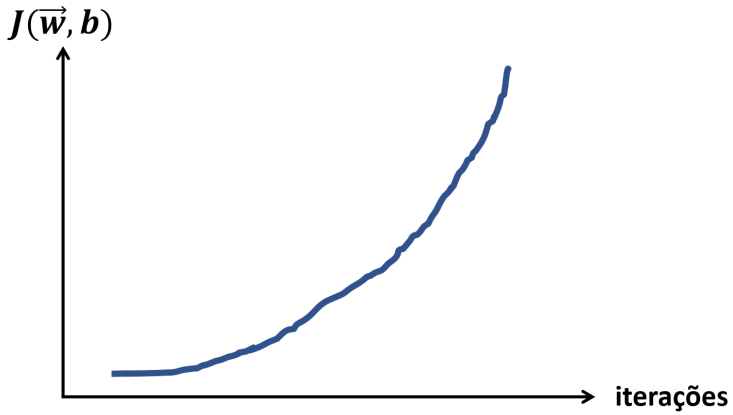
Como escolher um valor adequado para α?

Escolhendo α adequadamente

Supondo α muito grande, pode acontecer o seguinte:



Supondo um problema de código, por exemplo, $w_j = w_j + \alpha d_j$ pode ocorrer o seguinte:



Escolhendo α adequadamente

Dica para debugar o código:

Escolha um valor suficientemente pequeno para α (valor bem pequeno), e verifique se $J(\vec{w}, b)$ está sempre decrescendo iteração após iteração.

Se $J(\vec{w}, b)$ cresce em algum momento, então provavelmente tem-se um bug no código.

Importante:

Usar um valor bem pequeno para α consiste numa boa estratégia para debugar, entretanto fará com que o aprendizado do seu modelo seja lento.

Escolhendo α adequadamente

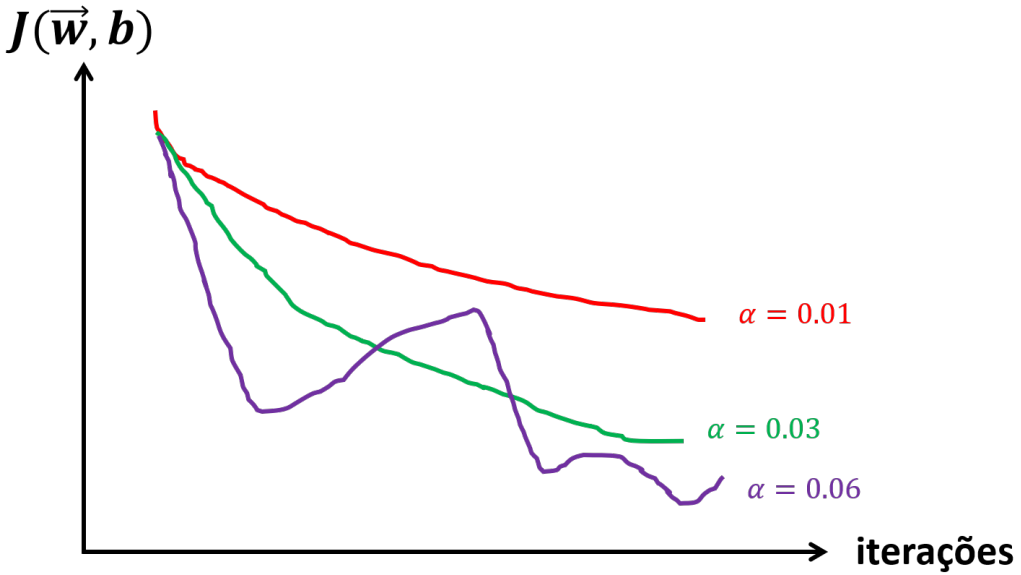
Um método eficiente:

Teste diferentes valores para α

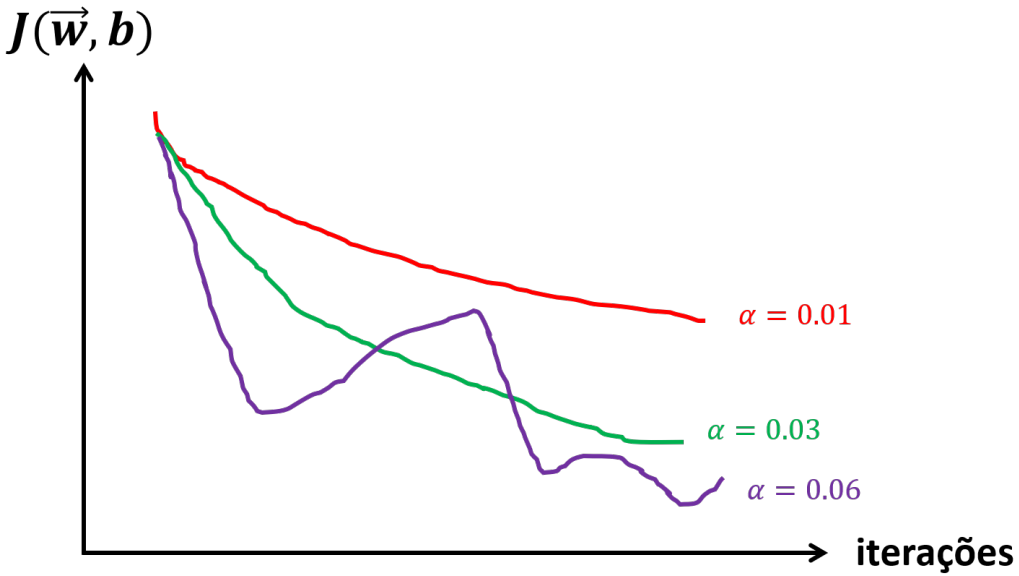
$$\ldots \quad 0.001 \quad 0.01 \quad 0.1 \quad 1 \quad \ldots$$

Para cada valor de α acima, rodar o método do gradiente por um certo número de iterações e:

- verificar para qual escolha de α a função $J(\vec{w}, b)$ decai rapidamente, porém mantendo consistência no seu decaimento.



Escolhendo α adequadamente



Observações finais (nível *hard*)

- A função custo leva em conta todos os parâmetros do modelo simultaneamente. Como todos os parâmetros são atualizados a partir do mesmo α, pode acontecer de um determinado α ser adequado para um parâmetro e não para outro.
- Em alguns raros casos, um determinado valor menor para α pode fazer a função custo decair mais rapidamente a cada iteração em comparação com um α ligeiramente maior. Isso pode acontecer justamente quando esse α não é um valor adequado para certos parâmetros do modelo (o parâmetro encontra-se saltando o seu valor ótimo, por exemplo), ainda que a função custo decaia como um todo a cada iteração.

De olho no código!

Vamos agora ver como realizar o **escalonamento de características e a escolha do** α.

Nome do arquivo que trabalharemos agora:

codigo - escalonamento de caracteristicas e escolha do alpha.ipynb

Atividade de aula

Parte 1

Rode todo o "codigo - escalonamento de caracteristicas e escolha do alpha.ipynb" sem fazer qualquer tipo de alteração. Certifique-se de que você o compreendeu.

Parte 2

1. Explique, com as suas próprias palavras, como a escolha do α afeta a convergência do Método do Gradiente.
2. Explique, com as suas próprias palavras, o que é o Escalonamento de Características e qual o seu impacto no Método do Gradiente.