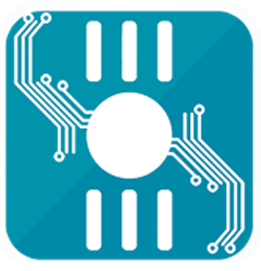
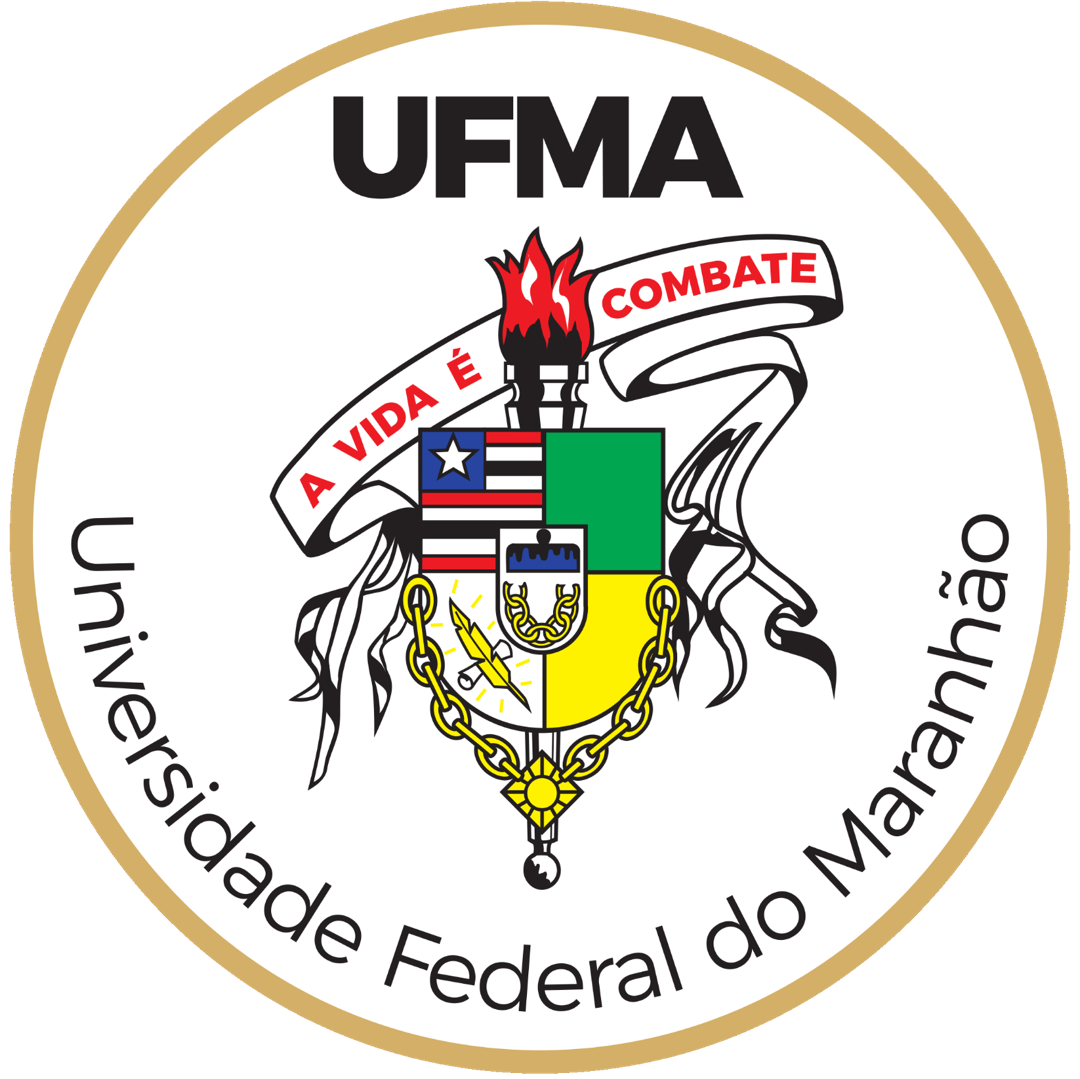
**Universidade Federal do Maranhão - UFMA Centro de Ciências Exatas e Tecnologia - CCET**

**Coordenação do Curso de Engenharia da Computação - CCEC Fundamentos de Redes Neurais**

**TÓPICOS EM ENGENHARIA DA COMPUTAÇÃO II - FUNDAMENTOS DE REDES NEURAIS – REGRESSÃO LINEAR**

Prof. Dr. Thales Levi Azevedo Valente Discente: Matheus Costa Alves

# 1 INTRODUÇÃO

A regressão linear é uma técnica estatística amplamente utilizada em tarefas de modelagem preditiva, cuja principal finalidade é estabelecer uma relação funcional entre variáveis independentes e uma variável dependente contínua. Seu caráter interpretável, simplicidade algorítmica e baixo custo computacional tornam-na uma escolha fundamental tanto em aplicações práticas quanto no ensino de fundamentos de aprendizado de máquina.

No contexto desta atividade, o algoritmo de descida do gradiente é empregado para otimizar os parâmetros do modelo, minimizando a função de custo associada ao erro quadrático médio entre as previsões e os valores reais. Para que esse processo ocorra de maneira eficiente, dois fatores exercem influência crítica: a taxa de aprendizado (α) e a inicialização dos parâmetros (θ). Enquanto a taxa de aprendizado regula o tamanho dos passos dados na direção do gradiente, a escolha dos valores iniciais dos parâmetros pode alterar significativamente o caminho de convergência e até mesmo impedir a obtenção de um mínimo global.

Este trabalho visa explorar o impacto desses dois fatores sobre a aprendizagem do modelo de regressão linear, por meio de experimentos sistemáticos. A partir da análise dos resultados obtidos por meio de visualizações gráficas, pretende-se não apenas consolidar os conhecimentos teóricos, mas também compreender suas implicações práticas no comportamento do algoritmo.

# 2 FUNDAMENTAÇÃO TEÓRICA

A regressão linear é um dos modelos estatísticos mais antigos e fundamentais da aprendizagem de máquina supervisionada (James et al., 2013). Seu objetivo principal é modelar a relação entre uma variável dependente y e uma ou mais variáveis independentes x, por meio de uma equação linear (James et al., 2013). No caso mais simples, denominado regressão linear simples, essa relação é descrita pela seguinte fórmula:

|  |  |
| --- | --- |
|  | (1) |

Nesta equação, representa o valor predito pelo modelo, x é a variável explicativa, é o intercepto (ou termo independente), e é o coeficiente angular da reta, responsável por quantificar o efeito de x sobre y. O modelo pode ser generalizado para múltiplas variáveis independentes, formando o que se chama de regressão linear múltipla (James et al., 2013). A equação é então expressa como:

|  |  |
| --- | --- |
|  | (2) |

Ou, de maneira compacta e vetorial:

|  |  |
| --- | --- |
|  | (3) |

Onde:

* é o vetor de atributos, geralmente incluindo o termo constante (intercepto),
* é o vetor de coeficientes do modelo,
* representa o produto escalar entre os vetores.

A suposição central da regressão linear é que existe uma relação aproximadamente linear entre as variáveis independentes e a variável dependente. Ou seja, espera-se que a variação em y possa ser explicada como uma combinação linear dos (Weisberg, 2014). Essa simplicidade matemática, somada à sua eficiência computacional e interpretabilidade, torna a regressão linear um modelo amplamente utilizado, especialmente como ponto de partida na análise de dados.

## 2.1 Função de Custo

Para que o modelo de regressão linear seja efetivamente útil na predição de valores, é necessário determinar os coeficientes que minimizam o erro entre os valores preditos e os valores reais y. Esse processo é orientado por uma função de custo, que quantifica a discrepância entre o modelo e os dados observados.

A função de custo mais comum utilizada na regressão linear é o Erro Quadrático Médio (Mean Squared Error – MSE), dada por:

|  |  |
| --- | --- |
|  | (4) |

Onde:

* n representa o número de amostras,
* ​ é a predição do modelo para a i-ésima amostra,
* ​ é o valor real correspondente,
* é o vetor de características da i-ésima amostra,
* é o vetor de parâmetros do modelo.

Essa função impõe penalidade quadrática aos erros cometidos pelo modelo, tornando a minimização mais sensível a outliers. No entanto, essa característica também contribui para uma superfície de custo suave e convexa, que garante a existência de um mínimo global.

Minimizar o valor de significa encontrar os valores ótimos dos coeficientes que melhor ajustam a linha de regressão aos dados observados. Para isso, diferentes técnicas podem ser utilizadas, como métodos analíticos baseados em álgebra linear (via equações normais) ou métodos iterativos como o gradiente descendente, que será explorado na próxima seção.

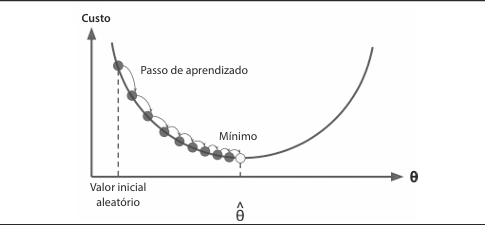
A escolha da função de custo adequada é essencial, pois influencia diretamente o comportamento do modelo durante o treinamento e sua capacidade de generalização. No contexto da regressão linear, o MSE é preferido por sua simplicidade matemática e propriedades analíticas bem definidas.

## 2.2 Gradiente Descendente

O gradiente descendente é um algoritmo de otimização amplamente utilizado no treinamento de modelos de regressão linear. Sua função é encontrar os valores ótimos dos parâmetros que minimizam a função de custo do modelo, geralmente o Erro Quadrático Médio (MSE). Em termos simples, trata-se de um processo iterativo que ajusta os coeficientes do modelo com base na inclinação da função de custo — o gradiente — na tentativa de descer em direção ao ponto mais baixo da curva, ou seja, o mínimo global (GÉRON, 2019).

Géron (2019) ilustra o funcionamento do gradiente descendente com a metáfora de estar perdido em uma montanha coberta por neblina: sem enxergar o vale, você sente o terreno sob seus pés e caminha sempre na direção da descida mais íngreme. De maneira análoga, o algoritmo ajusta os parâmetros em direção à menor inclinação local da função de custo. O processo continua até que o gradiente seja praticamente nulo — indicando que um mínimo foi alcançado.

Figura 01 – Gradiente Descendente



Fonte: GÉRON (2019, p. 120).

A atualização dos parâmetros é feita por meio da seguinte fórmula:

|  |  |
| --- | --- |
|  | (5) |

Em que:

* é a taxa de aprendizado (*learning rate*), um hiperparâmetro que determina o tamanho dos passos,
* é o vetor gradiente da função de custo com respeito aos parâmetros.

No caso da regressão linear com MSE, a derivada da função de custo é dada por:

|  |  |
| --- | --- |
|  | (6) |

Portanto, a equação de atualização se torna:

|  |  |
| --- | --- |
|  | (7) |

Essa abordagem, conhecida como gradiente descendente em lote (batch), utiliza todo o conjunto de dados para calcular o gradiente a cada iteração. Embora isso assegure uma trajetória suave até o mínimo, o método pode ser lento quando aplicado a grandes conjuntos de dados (GÉRON, 2019).

A escolha da taxa de aprendizado (*learning rate*) é um fator crítico no treinamento de modelos de aprendizado de máquina, especialmente em algoritmos baseados em gradiente descendente. Se a taxa de aprendizado for muito pequena, o algoritmo pode convergir de forma extremamente lenta, exigindo um número excessivo de iterações para alcançar um mínimo da função de custo, o que aumenta o tempo computacional e o custo de recursos. Por outro lado, se a taxa de aprendizado for muito alta, os passos dados na direção do gradiente podem ser grandes demais, fazendo com que o algoritmo oscile em torno do mínimo ou até mesmo divirja, ultrapassando soluções ótimas e impedindo a convergência (GÉRON, 2019).

Além disso, Géron destaca que o escalonamento das características é fundamental para o bom desempenho do algoritmo. Características em escalas muito diferentes podem distorcer a superfície da função de custo, transformando a "tigela convexa" em uma "tigela alongada", o que dificulta a convergência.

Por fim, uma vantagem importante é que, como a função de custo do MSE é convexa no caso da regressão linear, o gradiente descendente tem garantia de convergir para o mínimo global, desde que a taxa de aprendizado seja bem escolhida e o número de iterações seja suficiente.

# 3 RESULTADOS E DISCUSSÃO

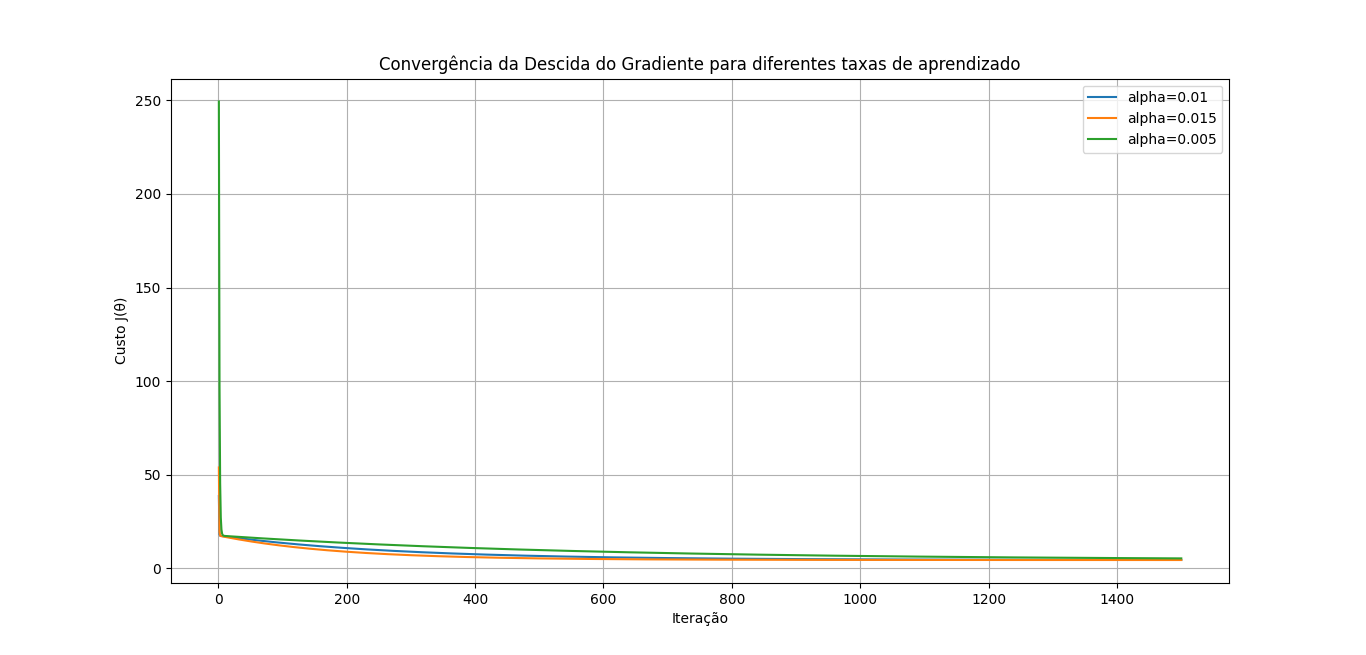
Neste capítulo, são apresentados e discutidos os resultados obtidos a partir dos experimentos com o modelo de regressão linear, incluindo a implementação do algoritmo de gradiente descendente. Serão analisados aspectos como a convergência da função de custo, o ajuste da reta aos dados, bem como a visualização da superfície e das curvas de contorno da função de custo.

Além disso, serão avaliados os impactos provocados por diferentes valores da taxa de aprendizado (η) e pela inicialização dos pesos (β). A discussão abordará como essas escolhas influenciam a estabilidade, a velocidade de convergência e a qualidade final do modelo treinado.

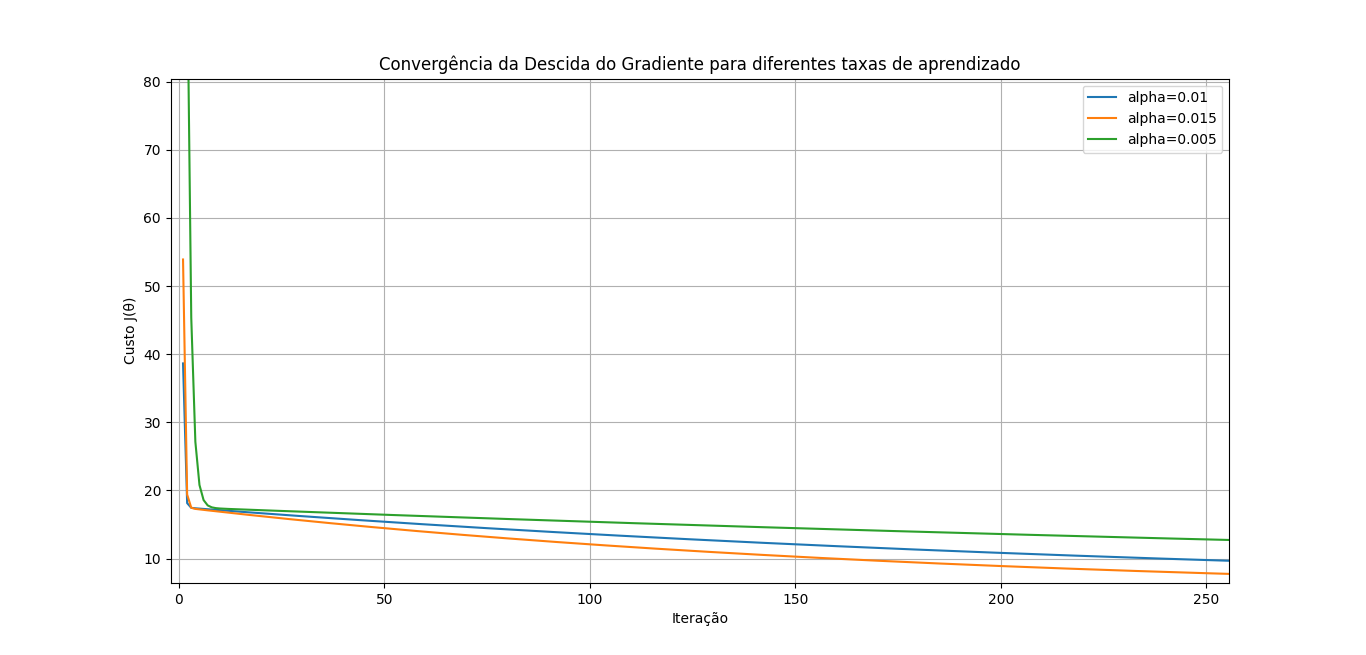
## 3.1 Impacto do Valor de α na Convergência

Ao testar diferentes valores de α (learning rate), observa-se que alguns valores fazem o modelo convergir mais rápido e de forma mais estável, enquanto valores pequenos demoram mais a convergir. Por outro lado, valores muito grandes podem fazer o gradiente explodir, impedindo o algoritmo de encontrar o mínimo. A Figura 1 e 2 apresentam a comparação da taxa de convergência para os valores de α definidos em .

Figura 01 – Curvas de convergência para diferentes valores de α.



1. Visualização padrão



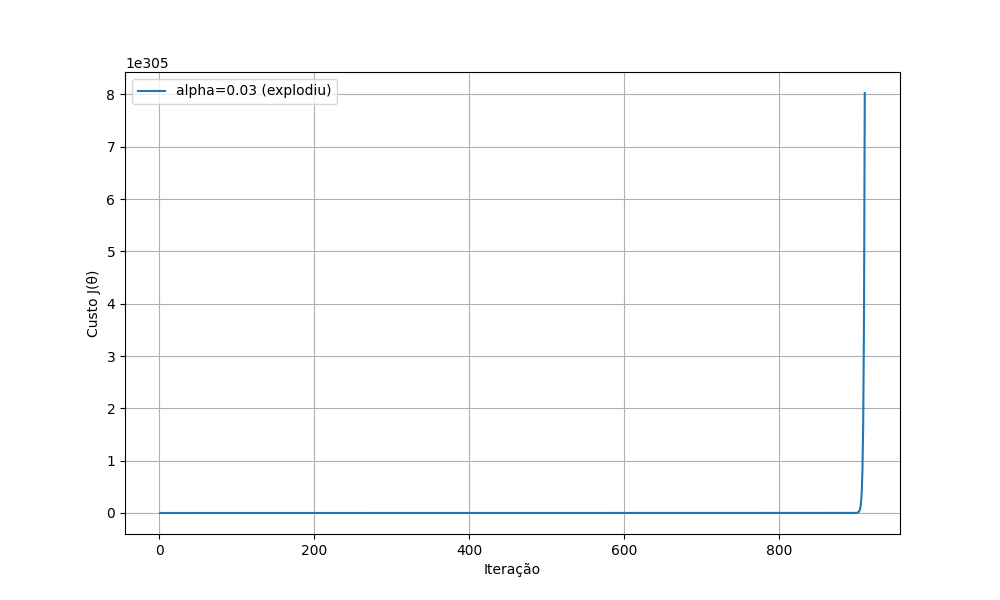
1. Visualização amplificada

Fonte: Autor

Na Figura 1, vemos que a menor taxa (α = 0,005) apresenta a convergência mais lenta, iniciando e terminando com o maior valor de erro. As taxas α=0,01 e α = 0,015 são mais rápidas na convergência, com α = 0,01 reduzindo o erro de maneira suave, enquanto α = 0,015 cai um pouco mais rapidamente, embora comece com um erro inicial maior. Ao atingir cerca de 1000 iterações, essas duas taxas praticamente se sobrepõem, estabilizando-se em um valor de erro próximo de 7.

Esses resultados são consistentes com a teoria de aprendizado de máquina sobre o impacto da taxa de aprendizado. Valores menores de α tendem a provocar uma convergência mais lenta, especialmente se o erro inicial for elevado. Por outro lado, valores muito grandes podem levar à explosão do gradiente, comprometendo a convergência. Esse comportamento será ilustrado na Figura 2 a seguir.

Figura 02 – Curva de convergência com explosão de gradiente.



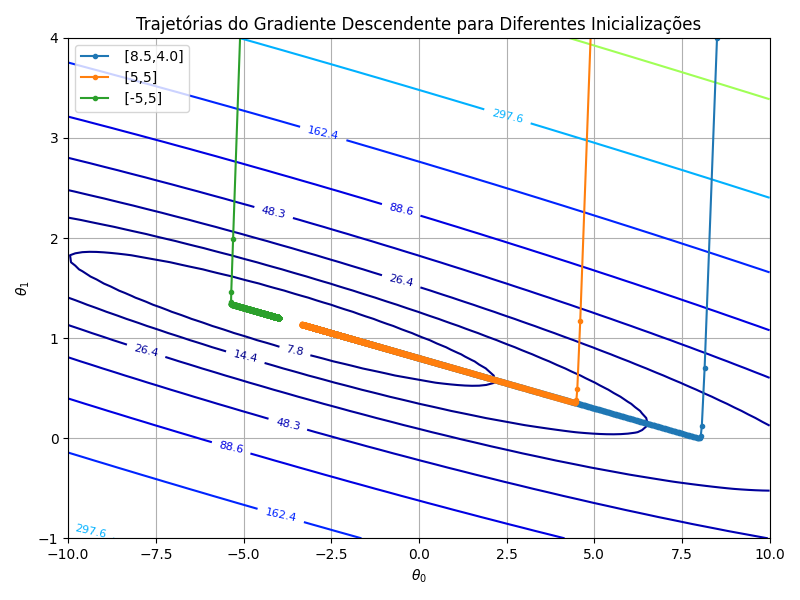
Fonte: Autor

Na Figura 2 observamos que custo permanece em 0 por um pouco mais de 900 iterações, após isso, ele dispara abruptamente pra um valor extremamente auto (1e305), indicando um alta instabilidade causada pelo auto valor de learning rate (0.03) para este problema, que faz com que os passos do algoritmo se tornem grandes demais, levando a uma divergência ao invés de uma convergência.

## 3.2 Impacto do Valor das Inicializações na Convergência

Os resultados dos experimentos com diferentes valores de inicialização evidenciaram que escolhas mais adequadas para θ podem acelerar o processo de convergência do algoritmo. Inicializações mais distantes do mínimo global exigem movimentos maiores e podem levar mais tempo para estabilizar. Podemos interpretar os valores iniciais dos pesos como as coordenadas iniciais a partir das quais o algoritmo inicia sua busca pelo objetivo, o mínimo global. As Figura 3 e 4 ilustram esse comportamento para diferentes valores de inicialização de θ.

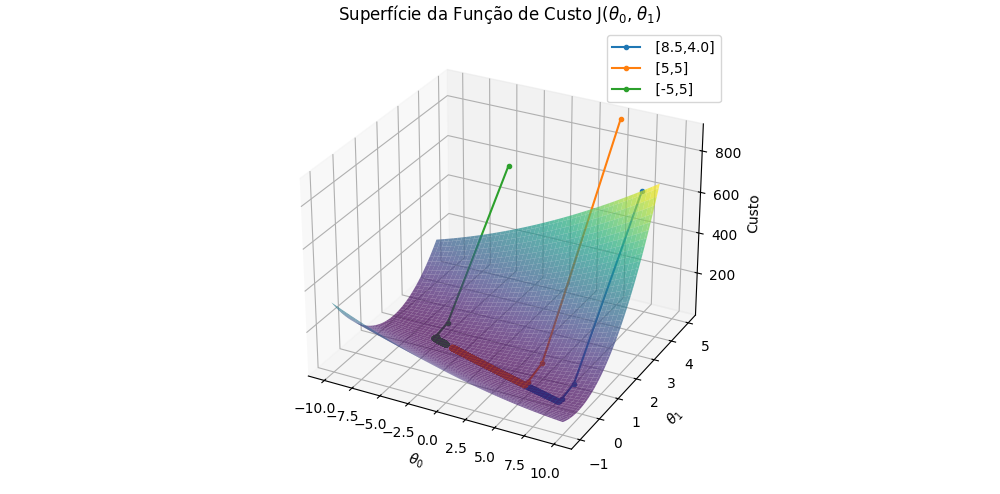
Figura 3 – Gráfico de contorno 2d



Fonte: Autor

Como pode-se observar, algumas trajetórias estão mais distantes do ponto ótimo do que outras. Para os valores [−5,5], a trajetória ficou mais "alinhada" ao objetivo, muito por conta de estar mais próximo do valor de no mínimo global, necessitando de menos ajustes. Em contrapartida, para [5,5] a trajetória ficou mais distante e precisou de mais ajustes no seu valor de ​. A figura 4 mostra o gráfico em 3d para uma melhor visualização.

Figura 04 – Superfície da função de custo 3d



Fonte: Autor

Esse gráfico ilustra de forma mais clara o que foi discutido anteriormente. Podemos observar que, de fato, os valores [−5,5] podem ser considerados os melhores dentre os três para a inicialização, pois apresentam uma trajetória mais curta, evidenciando a importância do valor correto de θ. Em projetos mais complexos, a escolha adequada de θ, juntamente com a taxa de aprendizado (*learning rate*), será fundamental para garantir a convergência da função de custo ao seu mínimo local.

Essa observação se conecta diretamente ao conceito de *fine-tuning*, onde, a partir de uma boa inicialização, buscamos ajustes ainda mais precisos dos parâmetros do modelo para otimizar seu desempenho. O *fine-tuning* torna-se crucial principalmente em cenários mais desafiadores, nos quais pequenas diferenças nos valores iniciais de θ e nas taxas de aprendizado podem impactar significativamente a qualidade da solução final.

# 4 CONCLUSÃO

Este trabalho abordou a importância e o impacto da escolha da taxa de aprendizado (α) e da inicialização dos pesos (θ) no comportamento do algoritmo de regressão linear. Foi demonstrado que diferentes configurações para α e θ influenciam diretamente a eficiência do processo de otimização, afetando tanto a velocidade quanto a estabilidade da convergência da função de custo J(θ). Experimentos mostraram que taxas de aprendizado inadequadas podem comprometer a redução eficaz do erro, enquanto escolhas apropriadas de inicialização dos pesos podem acelerar o processo de convergência para o mínimo global.

Referências

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning: with applications in R**. Nova York: Springer, 2013.

WEISBERG, S. **Applied linear regression**. 4. ed. Nova York: Wiley, 2014.

GÉRON, Aurélien. Mãos à obra: aprendizado de máquina com Scikit-Learn, Keras e TensorFlow: conceitos, ferramentas e técnicas para construir sistemas inteligentes. 2. ed. Rio de Janeiro: Alta Books, 2019.