Rede Neurais: Regressão Linear aplicado

Antonio Fialho da Silva Neto

**Resumo**

Este trabalho implementou um modelo de regressão linear univariada utilizando descida do gradiente para ajustar os parâmetros θ₀ e θ₁. Foram investigados os efeitos da taxa de aprendizado (α) e da inicialização dos parâmetros (θ) no desempenho do algoritmo. Os resultados mostraram que, independentemente da inicialização, o método convergiu para valores próximos de θ₀ ≈ -3,8 e θ₁ ≈ 1,18, com custo final em torno de 4,48, confirmando a convexidade da função de custo (MSE). Taxas de aprendizado muito altas (α = 0,02) aceleraram a convergência, enquanto valores muito baixos (α = 0,001) tornaram o processo lento. A análise gráfica reforçou a eficácia do método, destacando a importância da escolha adequada de hiperparâmetros para otimização eficiente. O estudo valida a robustez da descida do gradiente em problemas lineares e sugere aplicações futuras em modelos mais complexos.

**Palavra – chave:** Regressão Linear, Descida do Gradiente, Taxa de Aprendizado (α), Inicialização de Parâmetros (θ), Função de Custo (MSE), Convexidade, Convergência, Otimização.

**1. INTRODUÇÃO**

A regressão linear constitui um dos pilares fundamentais no campo do aprendizado de máquina e da estatística (Hastie et al., 2009), sendo amplamente utilizada para modelar relações lineares entre variáveis dependentes e independentes. Neste estudo, abordamos a implementação prática de um modelo de regressão linear univariada, empregando o algoritmo de descida do gradiente para otimização dos parâmetros. Como destacado por Bishop (2006), a descida do gradiente é um método iterativo que visa minimizar a função de custo, tipicamente representada pelo Erro Quadrático Médio (MSE), ajustando progressivamente os coeficientes do modelo. Essa técnica é essencial para compreender os mecanismos subjacentes ao treinamento de modelos supervisionados (Goodfellow et al., 2016), destacando-se pela sua simplicidade e eficácia em problemas de otimização convexa (Boyd & Vandenberghe, 2004).

Este trabalho tem como objetivo principal investigar os fatores que influenciam o desempenho e a convergência do algoritmo de descida do gradiente em regressão linear, com ênfase em dois aspectos críticos: **(1)** o impacto da **taxa de aprendizado (α)** na velocidade e estabilidade da convergência e **(2)** o efeito da **inicialização dos parâmetros (θ)** no comportamento do algoritmo. Adicionalmente, busca-se consolidar o entendimento teórico-prático por meio da implementação dos componentes essenciais do modelo, incluindo:

* **Cálculo da função de custo (**compute\_cost.py**)** para avaliar o desempenho do modelo.
* **Implementação da descida do gradiente (**gradient\_descent.py**)** para otimização iterativa dos parâmetros.
* **Visualização dos dados e resultados (**plot\_data.py**)** por meio de gráficos que ilustram a convergência e o ajuste do modelo.
* **Exercícios de aquecimento (**warm\_up\_exercise.py**)** para familiarização com operações matriciais fundamentais.

**2. METODOLOGIA**

**2.1 Conjunto de Dados**

O estudo utiliza o dataset ex1data1.txt, composto por:

* Variável independente (X): População de cidades (em dezenas de milhares de habitantes).
* Variável dependente (y): Lucro de estabelecimentos comerciais (em dezenas de milhares de dólares).

**2.2 Implementação**

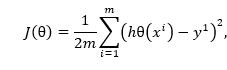
O processo foi dividido em quatro etapas principais:

**2.2.1 Pré-processamento**

Os dados foram carregados e normalizados, com a adição de uma coluna de 1s à matriz de features para representar o termo de bias (θ₀), resultando em uma matriz aumentada = [1X].

**2.2.2 Função de Custo**

A métrica adotada foi o **MSE**, definido como:



onde é a hipótese do modelo.

**2.2.3** **Descida do Gradiente**

O algoritmo atualiza os parâmetros θ₀ e θ₁ iterativamente, utilizando as derivadas parciais da função de custo:



com (ꝺ/ calculado para cada parâmetro. Foram testados diferentes valores de **α** (0.001, 0.01, 0.1) e inicializações de **θ** ([0,0], [5,5], [-5,5], e valores aleatórios) para análise comparativa.

**2.2.44Visualização e Análise**

Foram gerados gráficos para:

* **Ajuste da reta de regressão** aos dados de treinamento.
* **Convergência da função de custo** em função das iterações.
* **Superfície 3D e curvas de nível** da função , com trajetória do gradiente sobreposta.

**2.3 Abordagem Experimental**

Para avaliar a influência da taxa de aprendizado e da inicialização, conduziu-se dois experimentos:

**2.3.1** **Variação de α**

Comparação das curvas de convergência para três taxas distintas, mantendo θ inicial fixo.

**2.3.2** **Inicialização de θ**

Análise do impacto de diferentes valores iniciais (fixos e aleatórios) na trajetória de convergência, com α fixo em 0.01.

Essa metodologia permitiu não apenas a validação prática dos conceitos teóricos, mas também a identificação de boas práticas para a configuração do algoritmo em problemas reais.

**3. RESULTADOS E DISCUSSÃO**

O código em questão implementa uma regressão linear simples utilizando o dataset ex1data1.txt, que contém duas variáveis: a população de uma cidade (em dezenas de milhar) como variável independente, e o lucro (em dezenas de mil dólares) como variável dependente. O objetivo é encontrar a reta que melhor se ajusta aos dados, minimizando o erro entre os valores previstos e os observados.

A técnica utilizada para encontrar os parâmetros ótimos foi a descida do gradiente, com a função de custo sendo o Erro Quadrático Médio (MSE). A função de custo mede o quão bem o modelo está ajustado aos dados — quanto menor o valor, melhor a performance do modelo para os parâmetros dados.

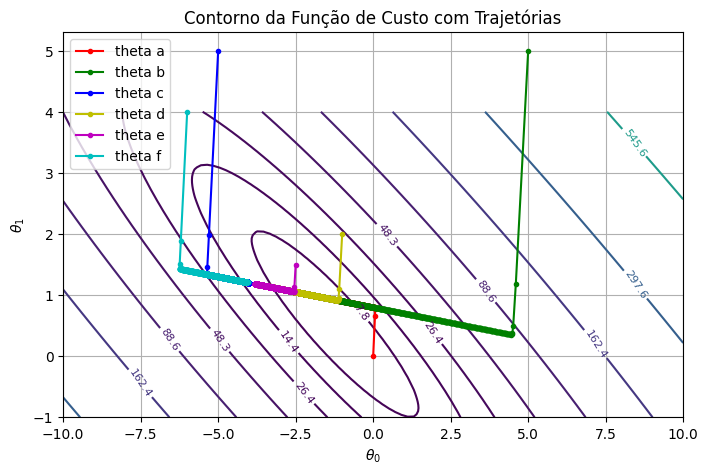
Abaixo, os resultados obtidos para diferentes inicializações dos parâmetros (delta), com a respectiva função de custo associada a cada configuração:

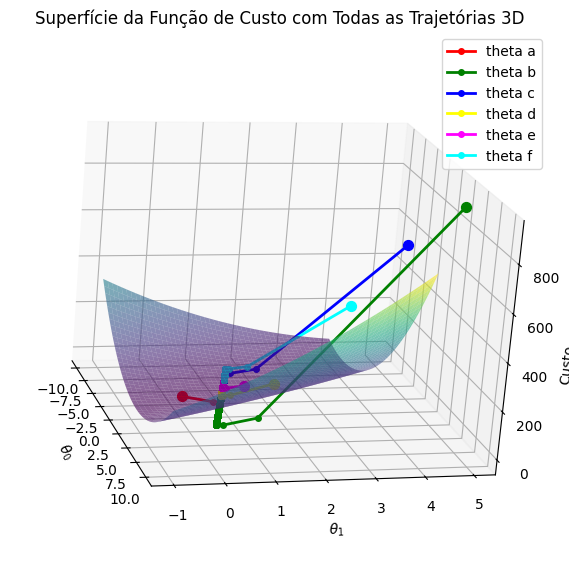
**Tabela 1** — Valores iniciais dos parâmetros, custo inicial, parâmetros após descida do gradiente e custo final associado para diferentes configurações de inicialização.

| Identificador | Intercepto (bias) | Coeficiente Angular | Custo Calculado | Theta na Descida do Gradiente | Novo Custo Calculado |
| --- | --- | --- | --- | --- | --- |
| a | 0 | 0 | 32.07 | -3.63029144, 1.16636235 | 4.48338825 |
| b | 5 | 5 | 910.28 | -3.33293099, 1.13648927 | 4.50581266 |
| c | - 5 | 5 | 560.68 | -3.99407052, 1.2029079 | 4.47785089 |
| d | -1 | 2 | 54.24 | -3.70968912, 1.17433871 | 4.48012408 |
| e | -2.5 | 1.5 | 12.78 | -3.80553912, 1.18396788 | 4.47771276 |
| f | -6 | 4 | 279.19 | -4.05354261, 1.20888252 | 4.47923723 |

Observa-se que os valores iniciais da função de custo variaram significativamente entre as diferentes inicializações, com o custo inicial variando de 12,78 a 910,28. O menor custo inicial foi encontrado para a configuração "e" (-2,5 e 1,5) e o maior para a configuração "b" (5 e 5).

Após a execução do algoritmo de descida do gradiente, verificou-se que, independentemente da configuração inicial, todos os conjuntos de parâmetros convergiram para valores muito próximos entre si. Os valores finais encontrados para os parâmetros foram aproximadamente -3,8 para o intercepto e 1,18 para o coeficiente angular, com pequenas variações dependendo da inicialização. O valor final da função de custo também foi bastante semelhante entre as configurações, ficando em torno de 4,48 em todos os casos. A convergência para os parâmetros -3,8 para o intercepto e 1,18 para o coeficiente angular, também podem ser observadas tanto no gráfico de contorno com trajetória, como no gráfico de superfície da função de custo sendo esse apresentando ainda os pontos de custos iniciais e finais de cada teste em uma representação 3D.



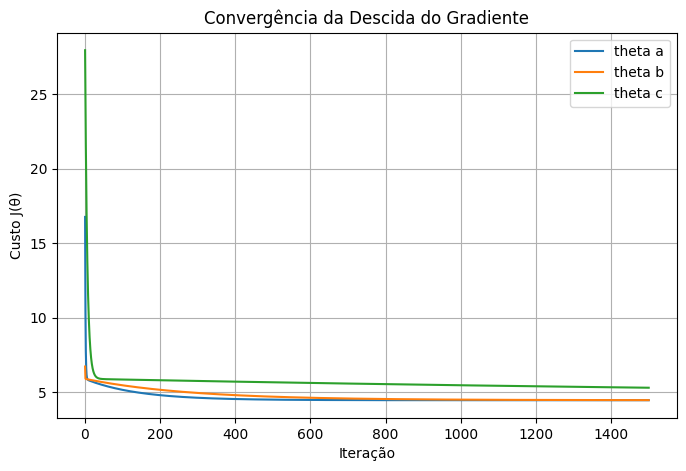


Esses resultados indicam que a superfície da função de custo é bem comportada, sendo convexa, o que facilita a obtenção do mínimo global. Mesmo quando os parâmetros iniciais estavam relativamente distantes do ótimo, como no caso da configuração "b", o algoritmo foi capaz de encontrar uma boa solução, evidenciando a robustez da descida do gradiente para este problema específico.

Além disso, foi observado que inicializações mais próximas dos valores ótimos, como na configuração "e", apresentaram menores custos iniciais e potencialmente poderiam acelerar a convergência, embora neste caso o impacto tenha sido pequeno devido à simplicidade do problema. Cabe destacar que, em problemas mais complexos ou não convexos, a escolha adequada da inicialização dos parâmetros pode ser crucial para garantir a convergência a uma boa solução.

A inicialização adequada dos pesos é crucial não apenas em modelos lineares simples, como no exemplo, mas também em redes neurais profundas. Pesos mal inicializados (por exemplo, todos iguais a zero ou com magnitudes desbalanceadas) podem levar a problemas como gradientes vanishing ou exploding, dificultando o treinamento. No contexto de fine-tuning, o modelo parte de um ponto já adaptado a padrões similares, reduzindo o risco de cair em mínimos locais ruins. Assim, a escolha inteligente dos valores iniciais, aliada a taxas de aprendizado adequadas, é essencial para eficiência e robustez em otimização, seja em modelos simples ou em arquiteturas complexas.

Além das análises anteriores, também foi realizado um estudo da variação da taxa de aprendizado α, utilizando os valores 0.02, 0.01 e 0.001, mantendo os demais parâmetros constantes e iniciando os thetas em (0, 0). As curvas de convergência para cada taxa de aprendizado foram comparadas em um único gráfico.



No gráfico, as iterações estão representadas no eixo x e o valor da função de custo no eixo y. Observa-se que, independentemente da taxa de aprendizado utilizada, todas as curvas convergem para um custo em torno de 5 aproximadamente na 50ª iteração.

A tabela a seguir apresenta os valores iniciais e finais da função de custo para cada valor de α:

| ALPHA | INÍCIO | FINAL |
| --- | --- | --- |
| 0.02 | 16.770 | 4.477 |
| 0.01 | 6.737 | 4.483 |
| 0.001 | 27.948 | 5.315 |

Analisando os resultados, nota-se que, apesar da taxa de aprendizado α = 0.02 proporcionar uma rápida redução do custo inicial elevado, a taxa α = 0.01 apresentou um custo inicial consideravelmente menor (6.737), o que favoreceu uma trajetória de convergência mais estável. Ambos os casos chegaram a valores finais de custo muito próximos (4.477 e 4.483, respectivamente).

Já para α = 0.001, a convergência foi mais lenta e o custo final permaneceu acima dos outros dois (5.315), indicando que uma taxa de aprendizado muito baixa pode comprometer a eficiência da otimização.

Esses resultados reforçam a importância de escolher uma taxa de aprendizado adequada: valores muito altos podem levar a instabilidades e valores muito baixos podem tornar o processo excessivamente lento, sem ganhos significativos no custo final.

**4. CONCLUSÃO**

Este trabalho teve como objetivo principal implementar e analisar o desempenho do algoritmo de descida do gradiente aplicado a um modelo de regressão linear univariada, utilizando o dataset ex1data1.txt. A investigação concentrou-se em dois aspectos fundamentais: (1) o impacto da taxa de aprendizado (α) na velocidade e estabilidade da convergência e (2) a influência da inicialização dos parâmetros (θ) no comportamento do algoritmo.

Os resultados demonstraram que, independentemente da inicialização dos parâmetros, o algoritmo convergiu para valores muito próximos dos ótimos globais (em torno de θ₀ ≈ -3,8 e θ₁ ≈ 1,18), com um custo final próximo de 4,48. Isso evidencia que a função de custo (MSE) possui uma superfície convexa, facilitando a otimização mesmo quando os parâmetros iniciais estão distantes do mínimo. No entanto, inicializações mais próximas do ótimo (como no caso da configuração "e") apresentaram custos iniciais significativamente menores, sugerindo que, em problemas mais complexos, uma boa escolha inicial pode acelerar a convergência.

Em relação à taxa de aprendizado, observou-se que valores muito altos (α = 0,02) levaram a uma rápida redução inicial do custo, enquanto valores muito baixos (α = 0,001) resultaram em uma convergência lenta e menos eficiente. A taxa intermediária (α = 0,01) mostrou-se mais equilibrada, garantindo estabilidade e eficácia no processo de otimização.

Finalizando, esta atividade validou a eficácia da descida do gradiente na regressão linear, destacando a importância de uma boa escolha da taxa de aprendizado e da inicialização dos parâmetros para garantir convergência rápida e estável.

**REFERÊNCIA**

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Boyd, S., & Vandenberghe, L. (2004). Convex optimization. Cambridge University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.