Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Fundamentos Matemáticos para Regressão Logística em Classificação Binária

Martin B Lobe[1], Luiz-Rafael Santos[2]
Universidade Federal de Santa Catarina, Blumenau, SC

A Regressão Logística possui seu início no século XIX, e serviu como um método estatístico para lidar com a crescente quantidade de informações do século. Com a modernização do século XX, a regressão logística ganhou diversas aplicações, nos campos médicos, de ciências sociais e de aprendizado de máquina, sendo esse último utilizando muito o método para classificação binária de tarefas. Isso é feito utilizando a função logística (ou sigmóide) para transformar a entrada de dados em um valor de probabilidade entre 0 e 1:

$$h_\theta(x) = \sigma(\theta^\intercal x) = \frac{1}{1+e^{-\theta^\intercal x}}, \quad \text{do qual} \quad \sigma(z) = \frac{1}{1+e^{-z}}. \tag{1}$$

Dado um contexto, o resultado obtido a partir da entrada da função (1) resulta na probabilidade de um evento ocorrer, sendo 1 representando certeza e 0 caso contrário. Note-se que ao derivarmos (1), a mesma é escrita como:

$$\begin{aligned} \sigma'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} = \frac{1}{(1+e^{-z})^2}(e^{-z}) = \\ &= \frac{1}{1+e^{-z}} \cdot (1 - \frac{1}{1+e^{-z}}) = \sigma(z)\cdot(1-\sigma(z)). \end{aligned} \tag{2}$$

Nosso modelo de regressão logística pode ser utilizado no contexto de um modelo de aprendizado de máquina: queremos minimizar a função de perda (*loss function*), que será responsável por verificar se ajustamos os parâmetros com probabilidade máxima. Isso é equivalente a maximizar a verossimilhança, isto é, para o conjunto de dados $\mathcal{X} = \{(x^{(i)}, y^{(i)}) \mid x^{(i)} \in \mathbb{R}^n, y^{(i)} \in \{0,1\}\}$, queremos maximizar:

$$L(\theta) = \prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)}; \theta), \tag{3}$$

Na função (3), m é o número de dados observados, enquanto $P(y^{(i)} \mid x^{(i)}; \theta)$ é a probabilidade de observar $y^{(i)}$ dado $x^{(i)}$ e os parâmetros θ. No caso da regressão logística, essa probabilidade é modelada pela função sigmoide h_θ

$$P(y^{(i)} \mid x^{(i)}; \theta) = \left(\sigma(\theta^\intercal x^{(i)})\right)^{y^{(i)}} \cdot \left(1 - \sigma(\theta^\intercal x^{(i)})\right)^{1-y^{(i)}}. \tag{4}$$

Assim, o problema de otimização consiste em encontrar os parâmetros θ que maximizam a verossimilhança.

[1]Estudante da Licenciatura em Matemática. E-mail:martinblobe@gmail.com
[2]Departamento de Matemática, UFSC/Blumenau-SC. E-mail: l.r.santos@ufsc.br

Normalmente, trabalhamos com o logaritmo da verossimilhança, que transforma o produtório em um somatório, mantendo o mesmo máximo, simplificando os cálculos e evitando problemas numéricos associados a produtos de probabilidades muito pequenas, isto é,

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{m} \log \sigma(y^{(i)} \cdot \theta^{\intercal} x^{(i)}). \tag{5}$$

A mudança para a função (5) não altera a localização do máximo, pois o logaritmo é uma função monotonicamente crescente, mas simplifica os cálculos e melhora a estabilidade numérica. Além disso, a log-verossimilhança está diretamente relacionada à função de perda utilizada no treinamento do modelo.

O problema é convexo e diferenciável, portanto é possível utilizar variantes do método do **gradiente descendente**. Em notação vetorial, a atualização dos parâmetros é dada por:

$$\theta^{k+1} \coloneqq \theta^k + \alpha_k \nabla \ell(\theta^k), \tag{6}$$

Na equação (6) α_k é o tamanho do passo (ou taxa de aprendizagem) e $\nabla \ell(\theta^k)$ é o gradiente da função de verossimilhança em relação aos parâmetros θ^k. Observe que o sinal é positivo, pois estamos maximizando a função, e não minimizando.

Note que o número de observações pode ser grande, de modo uma estratégia é selecionar aleatoriamente um único exemplo de treinamento $(x_i, y_i) \in \mathcal{X}$, o que nos dá a *regra do gradiente estocástico*:

$$\theta_j \coloneqq \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}. \tag{7}$$

Na equação (7) temos que: α é a taxa de aprendizado, $y^{(i)}$ é o rótulo verdadeiro, $h_\theta(x^{(i)}$ é a previsão do modelo para a entrada $x^{(i)}$, e $x_j^{(i)}$ é a j-ésima feature da i-ésima observação. Além disso assumimos conhecidos os rótulos $y^{(i)} \in \{0, 1\}$. Para outras codificações (como $y^{(i)} \in \{-1, 1\}$), a forma exata da atualização requer ajustes na derivação, mas preserva a estrutura fundamental do método.

Neste estudo, exploramos os fundamentos matemáticos da regressão logística aplicados ao problema de classificação, começando com sua formulação probabilística até a aplicação de métodos de otimização utilizando métodos do tipo gradiente descendente, em particular o gradiente estocástico (SGD). As notas de aula de Ng e Ma [3] e o livro de Deisenroth, Faisal e Ong [1] serviram como guias para este trabalho. Fizemos também uma implementação dos algoritmos na linguagem Julia. Além disso, destacou-se a importância dos fundamentos de álgebra linear e otimização, conforme abordado no livro de Friedlander [2]. Vimos que as técnicas de otimização e álgebra linear servem como uma ferramenta poderosa para a implementação e o sucesso de metéodos de aprendizagem de máquina.

Referências

[1] M. P. Deisenroth, A. A. Faisal e C. S. Ong. **Mathematics for Machine Learning**. Cambridge ; New York, NY: Cambridge University Press, 2020. 1 p. ISBN: 978-1-108-67993-0.

[2] A. Friedlander. **Elementos de Programação Não Linear**. Unicamp, jan. de 1994. ISBN: 85-268-0304-2.

[3] A. Ng e T. Ma. **CS229 Lecture Notes**. `https://cs229.stanford.edu/main_notes.pdf`. Disponível online. Jun. de 2023.