

# Inteligência Artificial

Raoni F. S. Teixeira

## Aula 8 - Regressão Logística

### 1 Introdução

Nesta etapa do curso, estudaremos o problema de classificação. A classificação é semelhante à regressão, mas, em vez de prever valores contínuos, o objetivo é prever valores discretos para  $y$ . Vamos começar com um problema de classificação binária, em que  $y$  assume apenas dois valores: 0 e 1.

Por exemplo, se estivermos desenvolvendo um classificador de *spam* para e-mails,  $x$  representa características do e-mail e  $y$  é 1 se o e-mail for e 0 se não for. O valor 0 é chamado de classe negativa, e 1, de classe positiva, também podendo ser denotados por “-” e “+”. Dado  $x$ , o valor correspondente de  $y$  é chamado de rótulo do exemplo de treinamento.

### 2 Regressão Logística

Poderíamos tentar resolver o problema de classificação ignorando que  $y$  é um valor discreto e usar a regressão linear para prevê-lo. Porém, essa abordagem falha, pois não faz sentido que o modelo  $h_{\theta}(x)$  produza valores maiores que 1 ou menores que 0, já que  $y$  só pode ser 0 ou 1.

Para contornar esse problema, definimos a hipótese  $h_{\theta}(x)$  usando a função sigmoide  $g(z)$ :

$$h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}, \quad (1)$$

em que  $\theta \in \mathbb{R}^{n+1}$  é o vetor de parâmetros do modelo, e  $x \in \mathbb{R}^{n+1}$  é a entrada aumentada com o termo de viés  $x_0 = 1$ .

A função

$$g(z) = \frac{1}{1 + e^{-z}}$$

é uma curva em forma de “S” definida para todos os valores reais de  $z$ , como mostra a Figura 1.

A partir do gráfico, podemos observar que  $g(z)$  tende a 1 quando  $z \rightarrow \infty$  e a 0 quando  $z \rightarrow -\infty$ . Além disso,  $g(z)$  — e, consequentemente,  $h_{\theta}(x)$  — está sempre entre 0 e 1.

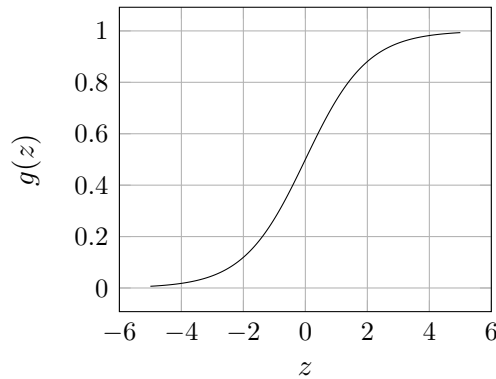


Figura 1: Gráfico da função sigmoide  $g(z)$ .

Essas propriedades nos permitem interpretar  $h_\theta(x)$  como a probabilidade de  $y = 1$  para uma entrada  $x$ :

$$P(y = 1|x; \theta) = h_\theta(x),$$

$$P(y = 0|x; \theta) = 1 - h_\theta(x).$$

De forma mais compacta, essa probabilidade pode ser expressa em única equação:

$$P(y|x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}. \quad (2)$$

### 3 Estimativa dos parâmetros

Considerando um conjunto de treinamento  $T = \{(x^{(1)}, y^{(1)}), \dots, (x^{(M)}, y^{(M)})\}$ , com  $M$  amostras independentes, os parâmetros do modelo  $h_\theta(x)$  podem ser estimados maximizando a função de verossimilhança  $L(\theta)$ :

$$\begin{aligned} L(\theta) &= P(T|\theta) \\ &= \prod_{i=1}^M P(y^{(i)}|x^{(i)}; \theta) \\ &= \prod_{i=1}^M (h_\theta(x^{(i)}))^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}. \end{aligned}$$

O objetivo é encontrar os parâmetros  $\theta$  que maximizem a probabilidade de observar os dados de treinamento  $T$ .

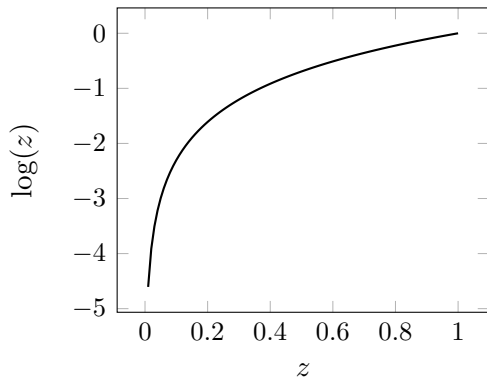
Para simplificar o cálculo, aplicamos o logaritmo à função de verossimilhança, transformando o produto em soma:

$$\begin{aligned}\ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^M (y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))) .\end{aligned}\tag{3}$$

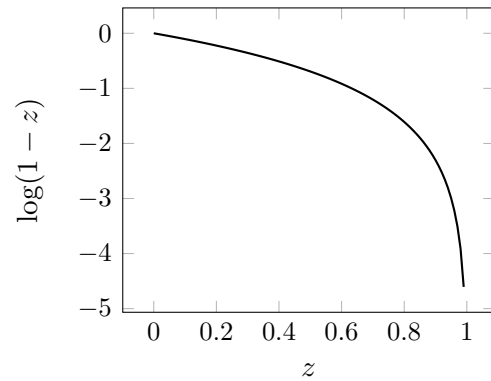
De forma mais detalhada, a log-verossimilhança pode ser separada em dois termos associados a  $y = 1$  e  $y = 0$ :

$$\ell(\theta) = \sum_{i=1}^M \left( y^{(i)} \underbrace{\log(h_{\theta}(x^{(i)}))}_{y=1} + (1 - y^{(i)}) \underbrace{\log(1 - h_{\theta}(x^{(i)}))}_{y=0} \right) .$$

A Figura 2 apresenta os gráficos das funções  $\log(h_{\theta}(x))$  e  $\log(1 - h_{\theta}(x))$ , que correspondem aos dois termos. Observe que  $\log(h_{\theta}(x))$ , associado a  $y = 1$ , cresce à medida que  $h_{\theta}(x)$  se aproxima de 1, enquanto  $\log(1 - h_{\theta}(x))$  cresce à medida que  $h_{\theta}(x)$  se aproxima de 0. Esse comportamento mostra que quanto maior o valor de  $\ell(\theta)$  mais preciso será o modelo —  $h_{\theta}(x)$  será o mais próximo de  $y$ .



(a) Termo associado a  $y = 1$



(b) Termo associado a  $y = 0$

Figura 2: Gráficos das funções  $\log(z)$  e  $\log(1 - z)$  associadas a  $y = 1$  e  $y = 0$ , respectivamente.

Para maximizar  $\ell(\theta)$ , usamos o método de subida de gradiente. Assim como na regressão linear, as atualizações dos parâmetros são realizadas iterativamente. Em notação vetorial, as atualizações são dadas por:

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} \ell(\theta),$$

em que  $\alpha$  é a taxa de aprendizado, e o sinal positivo indica que estamos maximizando  $\ell(\theta)$ , em vez de minimizá-la.

A derivada de  $\ell(\theta)$  para um único exemplo de treinamento  $(x, y)$  é:

$$\begin{aligned}\frac{\partial}{\partial \theta_j} \ell(\theta) &= \left( y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\ &= \left( y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)} \right) g(\theta^T x)(1 - g(\theta^T x)) \frac{\partial}{\partial \theta_j} \theta^T x \\ &= (y(1 - g(\theta^T x)) - (1 - y)g(\theta^T x))x_j \\ &= (y - h_\theta(x))x_j.\end{aligned}$$

Assim, a regra de atualização de  $\theta$  pela subida do gradiente é:

$$\theta_j \leftarrow \theta_j + \alpha \sum_{i=1}^M (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)}. \quad (4)$$

A primeira vista, essa regra de atualização pode parecer idêntica à da regressão linear, mas não é, pois  $h_\theta(x^{(i)})$  agora é igual a  $g(\theta^T x^{(i)})$ .

## 4 Classificação

Como  $h_\theta(x)$  prevê a probabilidade de  $y = 1$ , definimos um classificador binário como uma função  $\gamma : \mathbb{R} \rightarrow \{0, 1\}$  que determina a classe de uma amostra verificando se essa probabilidade excede 0.5:

$$\gamma(x) = \begin{cases} 1, & \text{se } h_\theta(x) \geq 0.5, \\ 0, & \text{caso contrário.} \end{cases} \quad (5)$$

Essa regra define uma superfície de decisão, que separa o espaço de entrada em duas regiões correspondentes às classes  $y = 1$  e  $y = 0$ . A superfície de decisão é descrita pela equação:

$$h_\theta(x) = 0.5, \quad (6)$$

que, substituindo a definição de  $h_\theta(x)$  (Equação 1), resulta em:

$$\frac{1}{1 + e^{-\theta^T x}} = 0.5.$$

E simplificando obtemos:

$$\theta^T x = 0.$$

Assim, a superfície de decisão é o conjunto de pontos  $x$  que satisfazem  $\theta^T x = 0$ . Essa equação é:

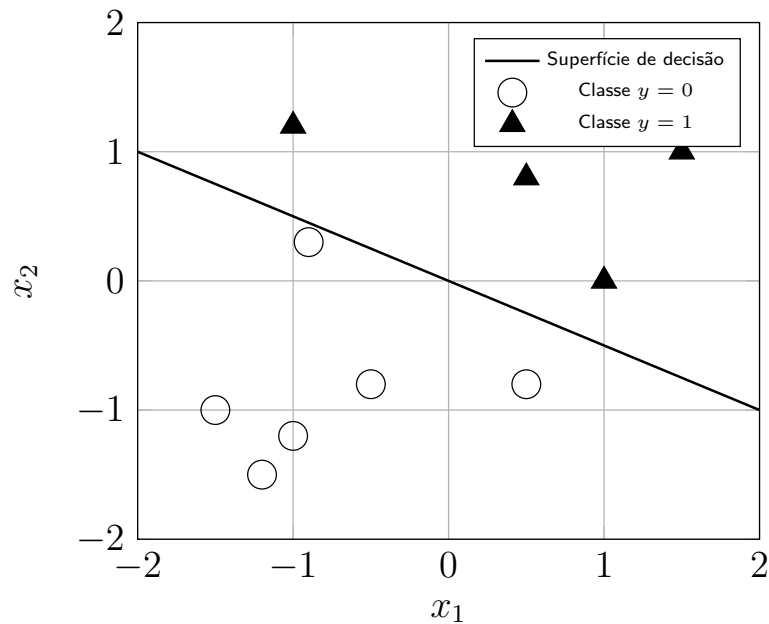


Figura 3: Exemplo de uma superfície de decisão em duas dimensões.

- um ponto (*threshold*) em uma dimensão.
- uma reta em duas dimensões,
- um plano em três dimensões e
- um hiperplano em dimensões superiores.

A Figura 3 apresenta um exemplo de superfície de decisão em duas dimensões. Os círculos correspondem à classe  $y = 0$ , e os triângulos, à classe  $y = 1$ . A reta é a superfície de decisão ( $\theta^T x = 0$ ). A posição e a orientação da superfície de decisão dependem dos parâmetros  $\theta$ . A reta da Figura 3 é dada por  $\frac{x_1}{2} - x_2 = 0$ . Se  $\theta^T x \geq 0$ , o ponto  $x$  é classificado como  $y = 1$  — triângulos da Figura 3. Caso contrário ( $\theta^T x < 0$ ), é classificado como  $y = 0$  — círculos da Figura 3.

## 5 Classificação Não Linear

Tal como na aula anterior, podemos utilizar um mapeamento não linear  $\phi$  para definir superfícies mais complexas. Por exemplo, o mapeamento

$$\phi(x_1, x_2) = \begin{bmatrix} 1 \\ x_1^2 \\ x_2^2 \end{bmatrix} \in \mathbb{R}^3. \quad (7)$$

gera a superfície de decisão circular da Figura 4. A superfície de decisão é definida pela equação  $x_1^2 + x_2^2 - 1.5 = 0$ . Em outras palavras, o círculo da Figura 4 é um hiperplano (linear) no espaço  $\mathbb{R}^3$  criado por  $\phi$ .

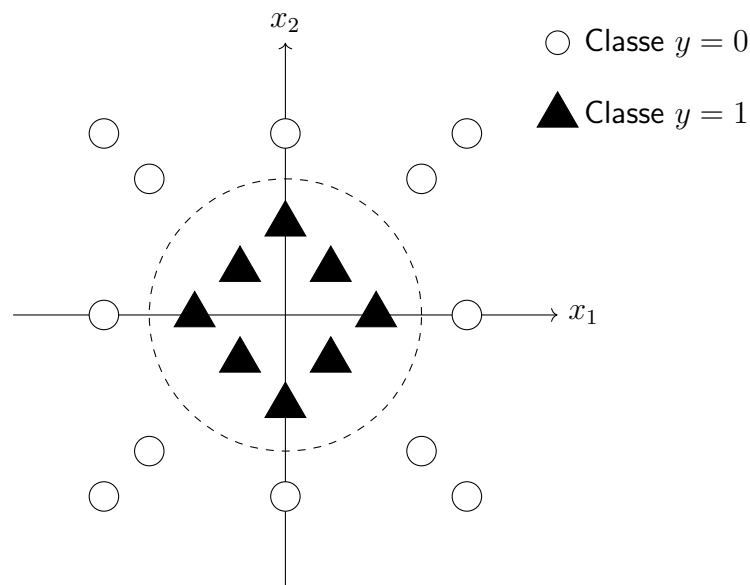


Figura 4: Exemplo de classificação não linear com fronteira circular.

É necessário tomar o mesmo cuidado da aula anterior com a dimensionalidade, com a escala dos dados mapeados e empregar regularização.

Para regularizar, transformamos o problema de maximização da Equação 3 em um de minimização multiplicando-o por  $-1$  e adicionamos o termo de regularização:

$$J(\theta) = \underbrace{-\sum_{i=1}^M y^{(i)} \log(h_{\theta}(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\theta}(x^{(i)}))}_{\text{negativo do log-verossimilhança}} + \underbrace{\lambda \sum_{j=1}^n |\theta_j|}_{\text{regularização}}. \quad (8)$$

## 6 Exercícios

1. Qual a diferença entre a regressão linear e a regressão logística e como a função sigmoide ajuda a resolver o problema de classificação binária?
2. Como a superfície de decisão é definida na regressão logística e qual a importância de mapeamentos não lineares para a classificação não linear?