Exercício 1 — Regressão logística e regressão softmax

Objetivos de aprendizagem

- Demonstrar uma equivalência entre duas formulações de um mesmo modelo.
- Ficar à vontade manipulando expressões com softmax e função logística.

A regressão logística binária é um caso particular da regressão logística multiclasse, também chamada de *softmax regression*. A função softmax é a análoga multiclasse da função logística, transformando os scores $\boldsymbol{\theta}_k^\top \mathbf{x}$ em valores no intervalo $[0, 1]$ que somam 1:

$$\pi_k(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}_k^\top \mathbf{x})}{\sum_{j=1}^{g} \exp(\boldsymbol{\theta}_j^\top \mathbf{x})}, \qquad k \in \{1, \ldots, g\}.$$

Mostre que a regressão logística e a regressão softmax são equivalentes quando $g = 2$. (Dica: escreva π_1 e π_2 explicitamente, use o fato de que $\pi_1 + \pi_2 = 1$ e manipule a expressão até reconhecer a forma logística padrão $\frac{1}{1+\exp(-z)}$.)

Exercício 2 — Hiperplanos em 2D

Objetivos de aprendizagem

- Entender que hiperplanos dividem o espaço de features com uma fronteira linear.
- Ganhar intuição sobre o papel dos coeficientes em um hiperplano.

Classificadores lineares como a regressão logística aprendem uma fronteira de decisão que é um *hiperplano* linear. Hiperplanos são definidos por equações do tipo $\boldsymbol{\theta}^\top \mathbf{x} = b$, com coeficientes $\boldsymbol{\theta}$ e escalar $b \in \mathbb{R}$.

Em duas dimensões, um hiperplano é apenas uma reta. Considere a equação $\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$ e desenhe, no mesmo par de eixos (x_1, x_2), os hiperplanos correspondentes às três configurações abaixo. Em seguida, em até dois parágrafos, explique a diferença entre as parametrizações — o que mudar cada coeficiente provoca geometricamente?

(a) $\theta_0 = 0,\ \theta_1 = \theta_2 = 1.$
(b) $\theta_0 = 1,\ \theta_1 = \theta_2 = 1.$
(c) $\theta_0 = 0,\ \theta_1 = 1,\ \theta_2 = 2.$

Exercício 3 — Fronteiras de decisão e limiares na regressão logística

Objetivos de aprendizagem

- Entender que a regressão logística produz uma fronteira de decisão linear.
- Ganhar intuição sobre como a parametrização e o limiar afetam as probabilidades previstas.

Na regressão logística (caso binário), estimamos a probabilidade $\pi(\mathbf{x} \mid \boldsymbol{\theta}) = p(y = 1 \mid \mathbf{x}, \boldsymbol{\theta})$. Para decidir a classe de uma observação, definimos um limiar $\alpha \in (0, 1)$ e colocamos $\hat{y} = 1$ se, e somente se, $\pi(\mathbf{x} \mid \boldsymbol{\theta}) \geq \alpha$.

(a) Mostre que a fronteira de decisão de um classificador logístico é um hiperplano linear. *Dica:* parta de $\pi(\mathbf{x} \mid \boldsymbol{\theta}) = \alpha$ e isole $\boldsymbol{\theta}^\top \mathbf{x}$ como função de α.

(b) As quatro figuras abaixo mostram a função logística $\pi(\mathbf{x} \mid \boldsymbol{\theta})$ para um problema de classificação binária com duas features, sob diferentes valores de $\boldsymbol{\theta}^\top = (\theta_1, \theta_2)$ (gráficos 1 a 3) e diferentes limiares α (gráfico 4, com três cortes). O que você pode deduzir sobre os valores de θ_1, θ_2 e α em cada caso? Quais as implicações para a classificação?

Gráfico (1)

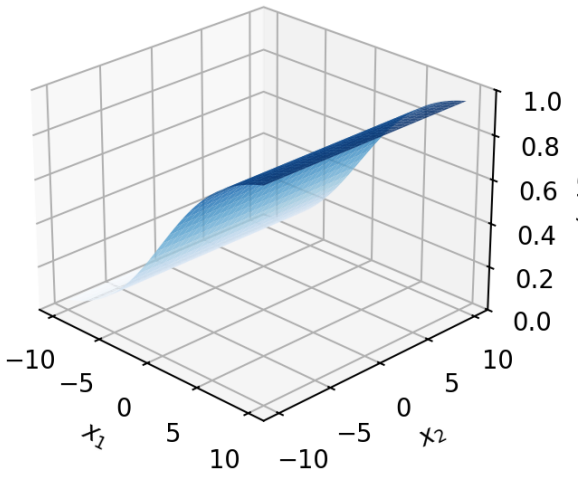


Gráfico (2)

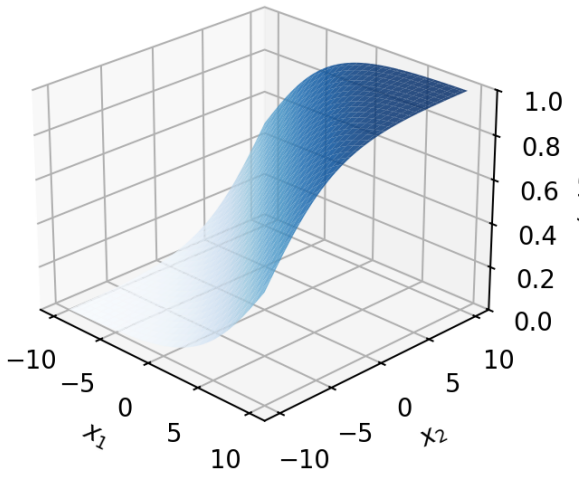


Gráfico (3)

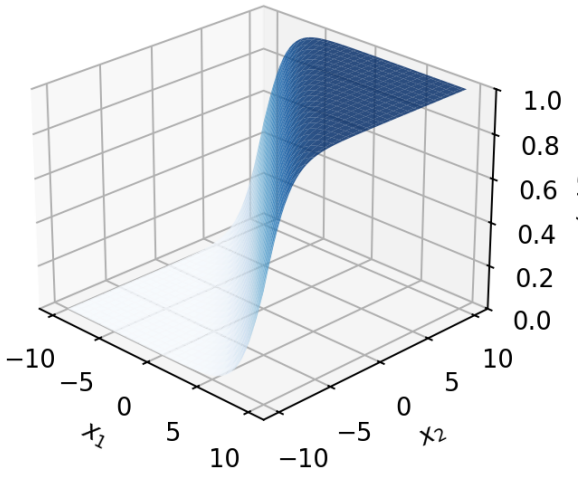
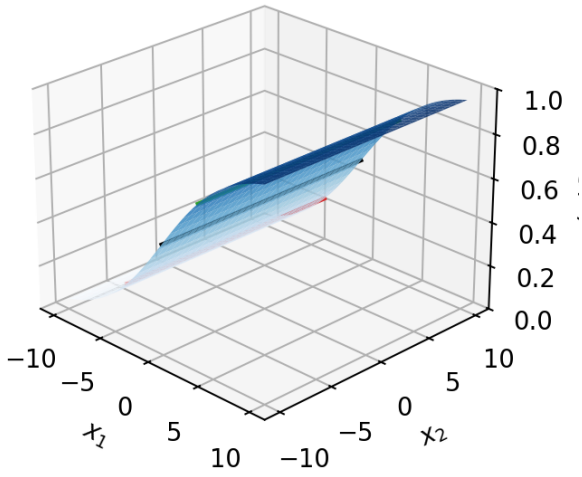


Gráfico (4)



(c) Derive a equação do hiperplano de decisão para $\alpha = 0{,}5$.

(d) Explique quando faz sentido escolher $\alpha = 0{,}5$ e quando seria preferível escolher um limiar diferente.