Aula 5 - Regressão Logística

João Florindo

Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas - Brasil
florindo@unicamp.br

Outline

- CLASSIFICAÇÃO: Saída discreta.
  - Email é spam ou não
  - Transação é fraudulenta ou não
  - Tumor benigno ou maligno
- Esses são exemplos de classificação **binária** (2 classes)

- CLASSIFICAÇÃO: Saída discreta.
  - Email é spam ou não
  - Transação é fraudulenta ou não
  - Tumor benigno ou maligno

- Esses são exemplos de classificação **binária** (2 classes)

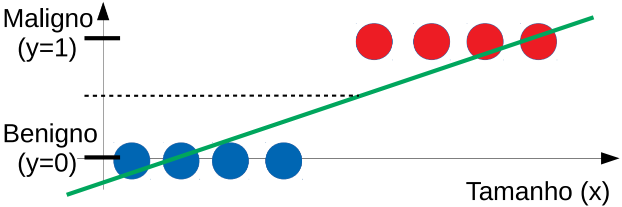
- CLASSIFICAÇÃO: Saída discreta.
  - Email é spam ou não
  - Transação é fraudulenta ou não
  - Tumor benigno ou maligno
- Esses são exemplos de classificação **binária** (2 classes)

- CLASSIFICAÇÃO: Saída discreta.
  - Email é spam ou não
  - Transação é fraudulenta ou não
  - Tumor benigno ou maligno

- Esses são exemplos de classificação **binária** (2 classes)

- CLASSIFICAÇÃO: Saída discreta.
    - Email é spam ou não
    - Transação é fraudulenta ou não
    - Tumor benigno ou maligno
- Esses são exemplos de classificação **binária** (2 classes)

Primeira ideia: usar regressão linear



PROBLEMAS:

- Um exemplo de treinamento a mais pode mudar completamente a reta
- Valores de $h_\theta(x)$ fora de $[0, 1]$

PROBLEMAS:

- Um exemplo de treinamento a mais pode mudar completamente a reta
- Valores de $h_\theta(x)$ fora de $[0, 1]$

Outline

SOLUÇÃO: **Regressão Logística**

Regressão Logística

Definir

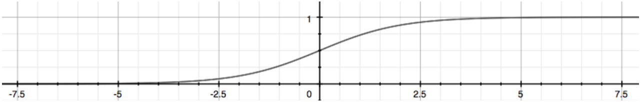
$$h_\theta(x) = g(\theta^T x),$$

em que $g(z)$ é a função sigmoide (logística):

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

ou seja:

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



- Note que $0 \leq h_\theta(x) \leq 1$.

Interpretação

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

Lê-se: "probabilidade de $y = 1$, dado x, parametrizado por θ".

Exemplo:

$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tamanho do tumor} \end{bmatrix}$$

e

$$h_\theta(x) = 0.7$$

$\Rightarrow$ Probabilidade de o tumor ser maligno ($y = 1$) é 70%.

- Como $y \in \{0, 1\}$:

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

Interpretação

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

Lê-se: "probabilidade de $y = 1$, dado x, parametrizado por θ".

Exemplo:

$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tamanho do tumor} \end{bmatrix}$$

e

$$h_\theta(x) = 0.7$$

$\Rightarrow$ Probabilidade de o tumor ser maligno ($y = 1$) é 70%.

- Como $y \in \{0, 1\}$:

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

Outline

Fronteira de Decisão

- Uma **fronteira de decisão** separa os exemplos com $y = 1$ dos com $y = 0$
- Definida por $h_\theta(x)$ e um limiar t: $y = 1$ se $h_\theta(x) \geq t$ e $y = 0$ caso contrário.
- Fronteira de decisão dada por $h_\theta(x) = t$.
- Na regressão logística, se por exemplo $t = 0.5$, temos $y = 1$ se $g(z) \geq 0.5$, ou ainda:

$$y = \begin{cases} 1 & \text{se} \quad \theta^T x \geq 0 \\ 0 & \text{se} \quad \theta^T x < 0. \end{cases}$$

Fronteira de Decisão

- Uma **fronteira de decisão** separa os exemplos com $y = 1$ dos com $y = 0$
- Definida por $h_\theta(x)$ e um limiar t: $y = 1$ se $h_\theta(x) \geq t$ e $y = 0$ caso contrário.
- Fronteira de decisão dada por $h_\theta(x) = t$.
- Na regressão logística, se por exemplo $t = 0.5$, temos $y = 1$ se $g(z) \geq 0.5$, ou ainda:

$$y = \begin{cases} 1 & \text{se} \quad \theta^T x \geq 0 \\ 0 & \text{se} \quad \theta^T x < 0. \end{cases}$$

Fronteira de Decisão

- Uma **fronteira de decisão** separa os exemplos com $y = 1$ dos com $y = 0$
- Definida por $h_\theta(x)$ e um limiar t: $y = 1$ se $h_\theta(x) \geq t$ e $y = 0$ caso contrário.
- Fronteira de decisão dada por $h_\theta(x) = t$.
- Na regressão logística, se por exemplo $t = 0.5$, temos $y = 1$ se $g(z) \geq 0.5$, ou ainda:

$$y = \begin{cases} 1 & \text{se} \quad \theta^T x \geq 0 \\ 0 & \text{se} \quad \theta^T x < 0. \end{cases}$$

Fronteira de Decisão

- Uma **fronteira de decisão** separa os exemplos com $y = 1$ dos com $y = 0$
- Definida por $h_\theta(x)$ e um limiar t: $y = 1$ se $h_\theta(x) \geq t$ e $y = 0$ caso contrário.
- Fronteira de decisão dada por $h_\theta(x) = t$.
- Na regressão logística, se por exemplo $t = 0.5$, temos $y = 1$ se $g(z) \geq 0.5$, ou ainda:

$$y = \begin{cases} 1 & \text{se} \quad \theta^T x \geq 0 \\ 0 & \text{se} \quad \theta^T x < 0. \end{cases}$$

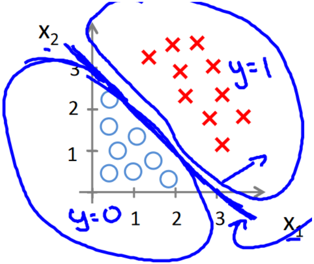
Exemplo

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

dado que

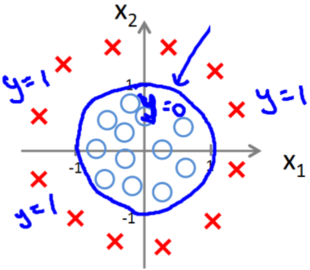
$$\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix},$$

$y = 1$ se $-3 + x_1 + x_2 \geq 0$.



Fronteira Não Linear

E se uma reta não separar os exemplos?



Fronteira Não Linear

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2),$$

com

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix},$$

de modo que $y = 1$ se $-1 + x_1^2 + x_2^2 \geq 1$.

Outline

- Função de custo:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right].$$

- Gradiente descendente:

Repita até convergir: $\{$

$$\theta_j := \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

$\}$ (atualização simultânea para todo θ_j)

- Função de custo:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(x^{(i)})) + (1 - y^{(i)})\log(1 - h_\theta(x^{(i)}))\right].$$

- Gradiente descendente:

Repita até convergir: $\{$

$$\theta_j := \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

$\}$ (atualização simultânea para todo θ_j)

- ATENÇÃO 1: O gradiente descendente é idêntico ao da regressão linear, MAS agora:

$$h_\theta(x) = \frac{1}{1 + \mathrm{e}^{-\theta^T x}}.$$

- ATENÇÃO 2: Normalizar os atributos.
- O gradiente pode ser vetorizado:

$$\theta := \theta - \frac{\alpha}{m} X^T (g(X\theta) - \mathbf{y}).$$

- ATENÇÃO 1: O gradiente descendente é idêntico ao da regressão linear, MAS agora:

$$h_\theta(x) = \frac{1}{1 + \mathrm{e}^{-\theta^T x}}.$$

- ATENÇÃO 2: Normalizar os atributos.
- O gradiente pode ser vetorizado:

$$\theta := \theta - \frac{\alpha}{m} X^T (g(X\theta) - \mathbf{y}).$$

- ATENÇÃO 1: O gradiente descendente é idêntico ao da regressão linear, MAS agora:

$$h_\theta(x) = \frac{1}{1 + \mathrm{e}^{-\theta^T x}}.$$

- ATENÇÃO 2: Normalizar os atributos.
- O gradiente pode ser vetorizado:

$$\theta := \theta - \frac{\alpha}{m} X^T (g(X\theta) - \mathbf{y}).$$

Otimização Avançada

- Algoritmos mais avançados podem ser usados para minimizar $J(\theta)$. Exemplos:
  - Gradiente conjugado
  - BFGS
  - L-BFGS
  - $\vdots$

Vantagens	Desvantagens
Não precisa escolher α manualmente	Consumo de memória
Normalmente usa menos iterações	Cada iteração é mais complexa

- Técnicas adaptativas para α são também populares:
  - RMSProp
  - Adam

Otimização Avançada

- Algoritmos mais avançados podem ser usados para minimizar $J(\theta)$. Exemplos:
  - Gradiente conjugado
  - BFGS
  - L-BFGS
  - ⋮

Vantagens	Desvantagens
Não precisa escolher α manualmente	Consumo de memória
Normalmente usa menos iterações	Cada iteração é mais complexa

- Técnicas adaptativas para α são também populares:
  - RMSProp
  - Adam

Otimização Avançada

- Algoritmos mais avançados podem ser usados para minimizar $J(\theta)$. Exemplos:
  - Gradiente conjugado
  - BFGS
  - L-BFGS
  - $\vdots$

Vantagens	Desvantagens
Não precisa escolher α manualmente	Consumo de memória
Normalmente usa menos iterações	Cada iteração é mais complexa

- Técnicas adaptativas para α são também populares:
  - RMSProp
  - Adam

Otimização Avançada

- Algoritmos mais avançados podem ser usados para minimizar $J(\theta)$. Exemplos:
  - Gradiente conjugado
  - BFGS
  - L-BFGS
  - $\vdots$

Vantagens	Desvantagens
Não precisa escolher α manualmente	Consumo de memória
Normalmente usa menos iterações	Cada iteração é mais complexa

- Técnicas adaptativas para α são também populares:
  - RMSProp
  - Adam

Otimização Avançada

- Algoritmos mais avançados podem ser usados para minimizar $J(\theta)$. Exemplos:
  - Gradiente conjugado
  - BFGS
  - L-BFGS
  - $\vdots$

Vantagens	Desvantagens
Não precisa escolher α manualmente	Consumo de memória
Normalmente usa menos iterações	Cada iteração é mais complexa

- Técnicas adaptativas para α são também populares:
  - RMSProp
  - Adam

Otimização Avançada

- Algoritmos mais avançados podem ser usados para minimizar $J(\theta)$. Exemplos:
  - Gradiente conjugado
  - BFGS
  - L-BFGS
  - $\vdots$

Vantagens	Desvantagens
Não precisa escolher α manualmente	Consumo de memória
Normalmente usa menos iterações	Cada iteração é mais complexa

- Técnicas adaptativas para α são também populares:
  - RMSProp
  - Adam

Otimização Avançada

- Algoritmos mais avançados podem ser usados para minimizar $J(\theta)$. Exemplos:
  - Gradiente conjugado
  - BFGS
  - L-BFGS
  - $\vdots$

Vantagens	Desvantagens
Não precisa escolher α manualmente	Consumo de memória
Normalmente usa menos iterações	Cada iteração é mais complexa

- Técnicas adaptativas para α são também populares:
  - RMSProp
  - Adam

Otimização Avançada

- Algoritmos mais avançados podem ser usados para minimizar $J(\theta)$. Exemplos:
  - Gradiente conjugado
  - BFGS
  - L-BFGS
  - $\vdots$

Vantagens	Desvantagens
Não precisa escolher α manualmente	Consumo de memória
Normalmente usa menos iterações	Cada iteração é mais complexa

- Técnicas adaptativas para α são também populares:
  - RMSProp
  - Adam

Outline

Exemplos:

- Emails em pastas: trabalho, amigos, família, lazer
- Exame médico: saudável, resfriado, gripe
- Clima: ensolarado, nublado, chuvoso

Normalmente as classes são numeradas - ex.: $y = 1, 2, 3, \cdots$.

Exemplos:

- Emails em pastas: trabalho, amigos, família, lazer
- Exame médico: saudável, resfriado, gripe
- Clima: ensolarado, nublado, chuvoso

Normalmente as classes são numeradas - ex.: $y = 1, 2, 3, \cdots$.

Exemplos:

- Emails em pastas: trabalho, amigos, família, lazer
- Exame médico: saudável, resfriado, gripe
- Clima: ensolarado, nublado, chuvoso

Normalmente as classes são numeradas - ex.: $y = 1, 2, 3, \cdots$.

Exemplos:

- Emails em pastas: trabalho, amigos, família, lazer
- Exame médico: saudável, resfriado, gripe
- Clima: ensolarado, nublado, chuvoso

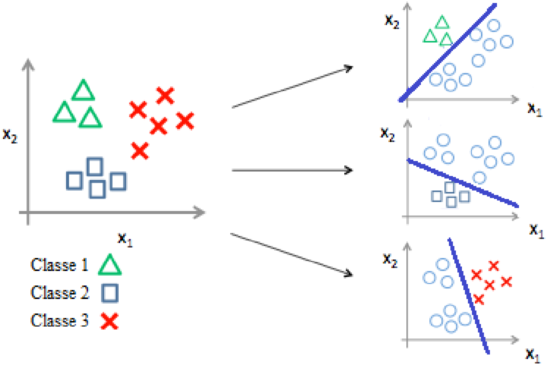
Normalmente as classes são numeradas - ex.: $y = 1, 2, 3, \cdots$.

Um-*vs*-Todos

Um-*vs*-Todos

- Para k classes temos k classificadores binários.
- No i-ésimo classificador $h_\theta^{(i)}(x)$, a i-ésima classe é "positiva" e todas as demais são "negativas".

Um-*vs*-Todos



$$h_\theta^{(i)}(x) = P(y = i|x;\theta), \qquad i = 1, 2, 3.$$

Classe predita:

$$\underset{i}{\operatorname{argmax}}\, h_\theta^{(i)}(x).$$