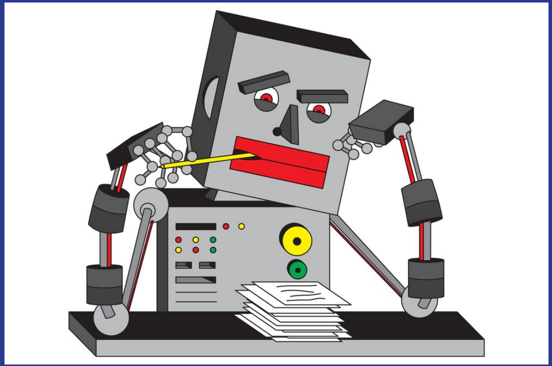
REGRESSÃO LOGÍSTICA

Prof. Eduardo Bezerra
(CEFET/RJ)
ebezerra@cefet-rj.br



Credits

- This presentation uses material from the following courses:
- CS229 – Machine Learning (prof. Andrew Ng)
- http://cs229.stanford.edu/
- CSE 4309 – Introdução to Machine Learning (prof. Vassilis Athitsos)
- http://vlm1.uta.edu/~athitsos/courses/cse4309_fall2019/

Visão Geral

- Introdução
  Hypothesis Representation
  Decision Boundary
  Cost Function
  Parameter learning

Introdução

Classificação Binária

5

- Tarefa de classificação onde o alvo (rótulo) pode assumir dois valores.
- Exemplos
- E-mail: spam/não spam?
- Transações online: fraudulenta/legítima?
- Tumor: maligno/benigno?

$$y \in \{0, 1\}$$

0: “Classe negativa” (e.g., non-spam message)

1: “Classe positiva” (e.g., message is spam)

Logistic regression

- Algoritmo usado para gerar a probabilidade de que um objeto esteja associado a uma determinada classe.
- e.g., approval/disapproval, victory/defeat, living/dead, healthy/sick, malignant/benign, spam/legitimate, etc..
- Pode ser estendido para modelar múltiplas classes.
- Por exemplo, determinar se uma imagem contém um gato, cachorro, leão etc.
- Cada objeto passado para o modelo de classificação resultante é mapeado para uma distribuição de probabilidade sobre as classes.

Notation

- $m \rightarrow$ Number of examples.
- $n \rightarrow$ number of features.
- $x^{(i)} \rightarrow$ i-th example.
- $y^{(i)} \rightarrow$ value for the *target* in the i-th example.
- $x_j \rightarrow$ j-th *feature* in the training set.
- $x_j^{(i)} \rightarrow$ value of the j-th *feature* in the i-th example.

NB: i and j start at 1, not at 0.

Conjunto de Treinamento

- Os dados usados para treinar um modelo de classificação, na sua forma mais simples, são tabulares.
- Por exemplo, estas são as primeiras 4 linhas do conjunto de dados Yeast*:

$$m = 1484 \qquad n = 8$$

0.5000	0.4600	0.6400	0.3600	0.5000	0	0.4900	0.2200	1
0.5300	0.5600	0.4900	0.4600	0.5000	0	0.5200	0.2200	1
0.5200	0.5300	0.5800	0.6900	0.5000	0	0.5000	0.2200	1
0.6700	0.6200	0.5400	0.4300	0.5000	0	0.5300	0.2200	1

* http://archive.ics.uci.edu/ml/datasets/yeast

Conjunto de Treinamento – tabular format

- Cada linha é um exemplo de treinamento.
- Todas as colunas, com exceção da última, representam a entrada, que é um vetor.
- A última coluna é o rótulo da classe.

$x^{(1)}$ $y^{(1)}$

0.5000	0.4600	0.6400	0.3600	0.5000	0	0.4900	0.2200	1
0.5300	0.5600	0.4900	0.4600	0.5000	0	0.5200	0.2200	1
0.5200	0.5300	0.5800	0.6900	0.5000	0	0.5000	0.2200	1
0.6700	0.6200	0.5400	0.4300	0.5000	0	0.5300	0.2200	1

Conjunto de Treinamento – tabular format

- Cada linha é um exemplo de treinamento.
- Todas as colunas, com exceção da última, representam a entrada, que é um vetor.
- A última coluna é o rótulo da classe.

	0.5000	0.4600	0.6400	0.3600	0.5000	0	0.4900	0.2200	1	
$x^{(2)}$	0.5300	0.5600	0.4900	0.4600	0.5000	0	0.5200	0.2200	1	$y^{(2)}$
	0.5200	0.5300	0.5800	0.6900	0.5000	0	0.5000	0.2200	1	
	0.6700	0.6200	0.5400	0.4300	0.5000	0	0.5300	0.2200	1	

Conjunto de Treinamento – tabular format

- Cada coluna (exceto a última) corresponde a uma característica.
- No conjunto de dados Yeast, há 8 características.
- Diferentes conjuntos de dados possuem quantidades variadas de características.

x_1	x_2						x_8	y
0.5000	0.4600	0.6400	0.3600	0.5000	0	0.4900	0.2200	1
0.5300	0.5600	0.4900	0.4600	0.5000	0	0.5200	0.2200	1
0.5200	0.5300	0.5800	0.6900	0.5000	0	0.5000	0.2200	1
0.6700	0.6200	0.5400	0.4300	0.5000	0	0.5300	0.2200	1

Conjunto de Treinamento – tabular format

- Qual é o valor da característica 4 do terceiro exemplo de treinamento?

0.5000	0.4600	0.6400	0.3600	0.5000	0	0.4900	0.2200	1
0.5300	0.5600	0.4900	0.4600	0.5000	0	0.5200	0.2200	1
0.5200	0.5300	0.5800	0.6900	0.5000	0	0.5000	0.2200	1
0.6700	0.6200	0.5400	0.4300	0.5000	0	0.5300	0.2200	1

Conjunto de Treinamento – tabular format

- What is the value of feature 4 of the third training example?

$$x_4^{(3)} = 0.6900$$

				x_4						
	0.5000	0.4600	0.6400	0.3600	0.5000	0	0.4900	0.2200	1	
	0.5300	0.5600	0.4900	0.4600	0.5000	0	0.5200	0.2200	1	
$x^{(3)}$	0.5200	0.5300	0.5800	**0.6900**	0.5000	0	0.5000	0.2200	1	$y^{(3)}$
	0.6700	0.6200	0.5400	0.4300	0.5000	0	0.5300	0.2200	1	

Hypothesis Representation

Aqui, estudamos como representamos hipóteses na Regressão Logística..

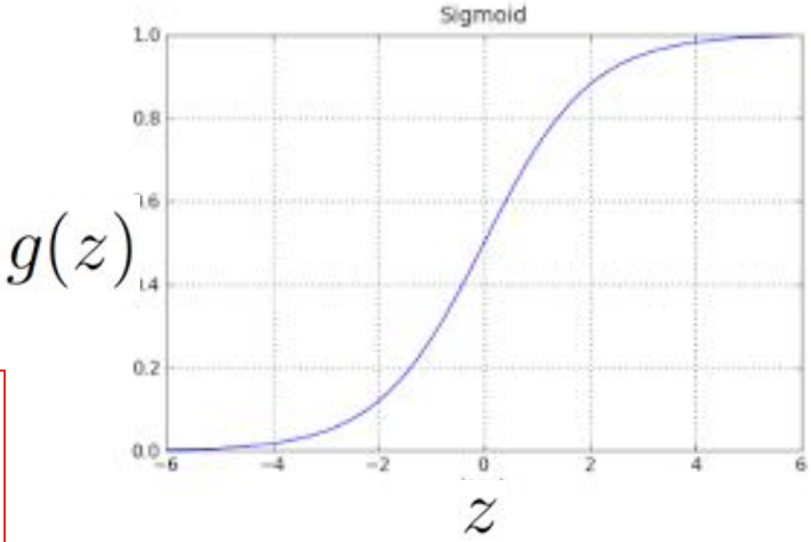
Hypothesis Representation

- Queremos uma representação tal que:

$$0 \leq h_\theta(x) \leq 1$$

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

sigmoid function
(aka logistic function)



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

Interpretation

$$\boxed{h_\theta(x)\text{: estimativa da probabilidade de que } y = 1 \text{ para a entrada } x}$$

- Example: $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$ $\quad h_\theta(x) = 0.7$
  - Há 70% de chance de que o tumor seja maligno

$$h_\theta(x) = \Pr(y = 1 \mid x; \theta) \qquad \Pr(y = 0 \mid x; \theta) = 1 - \Pr(y = 1 \mid x; \theta)$$

Decision Boundary

Aqui, estudamos o conceito de fronteira de decisão, que resulta da aplicação do algoritmo de regressão logística.

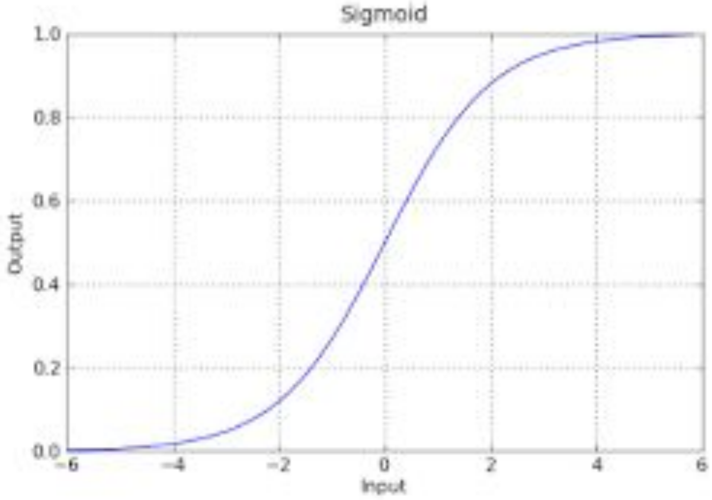
Decision boundary

18

- Note that
  - $g(z) \geq 0.5$ if $z \geq 0$
  - $g(z) < 0.5$ if $z < 0$

$$h_\theta(x) = g(\theta^T x) = \Pr(y = 1 \mid x; \theta)$$

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

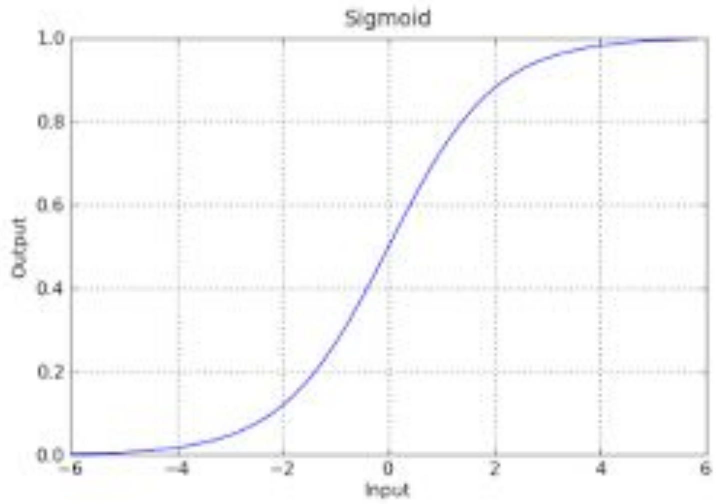


Decision boundary

- Note that
  - $g(z) \geq 0.5$ if $z \geq 0$
  - $g(z) < 0.5$ if $z < 0$
- Therefore
  - predict $y = 1$ if $\theta^T x \geq 0$
  - predict $y = 0$ if $\theta^T x < 0$

$$h_\theta(x) = g(\theta^T x) = \Pr(y = 1 \mid x; \theta)$$

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



Decision boundary

- A fronteira de decisão na regressão logística é o conjunto de todos os pontos $\boldsymbol{x}$ que satisfazem a expressão abaixo.

$$\Pr(\mathrm{y} = 1 \mid x) = \Pr(y = 0 \mid x) = 0.5$$

Decision boundary - example

- Suponha um conjunto de dados com duas características. Então:

$$\boldsymbol{x} = (x_1, x_2)$$
$$\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)$$

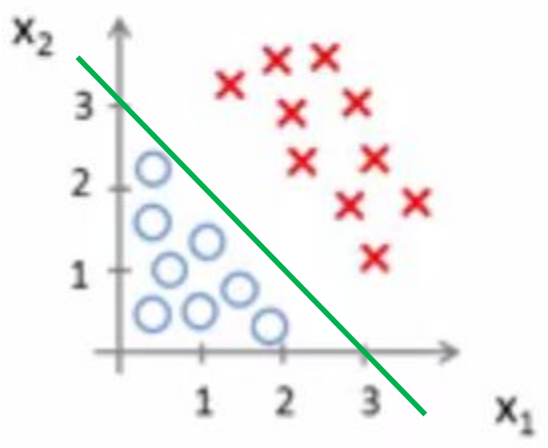
- Nesse caso, a fronteira de decisão forma uma linha dada pela seguinte equação:

$$x_2 = -\frac{\theta_1}{\theta_2}x_1 - \frac{\theta_0}{\theta_2}$$

$$\frac{1}{1+e^{-\boldsymbol{\theta}^t \boldsymbol{x}_+}} = \frac{1}{2}$$
$$\Rightarrow \boldsymbol{\theta}^t \boldsymbol{x}_+ = 0$$
$$\Rightarrow \theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d = 0$$

datascience.stackexchange.com/questions/49573/how-to-plot-logistic-regression-decision-boundary

Decision boundary - example



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

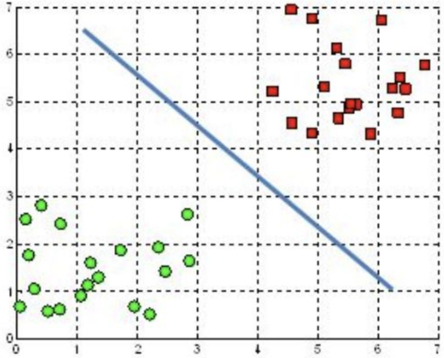
Suponha que $\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$

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

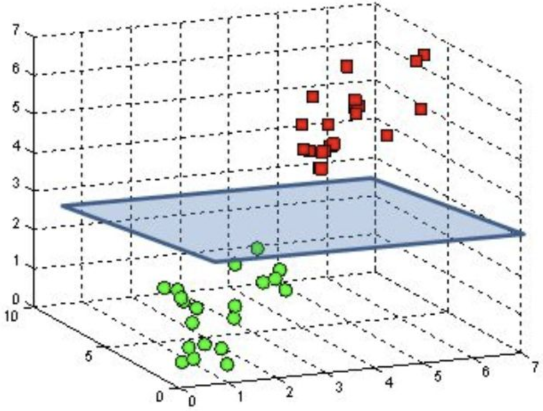
Decision boundary

- Em geral, a fronteira de decisão criada por um modelo de regressão logística é um hiperplano que separa da melhor forma possível as duas classes de exemplos:

A hyperplane in $\mathbb{R}^2$ is a line



A hyperplane in $\mathbb{R}^3$ is a plane



Cost Function

Aqui, apresentamos a função de custo utilizada na regressão logística.

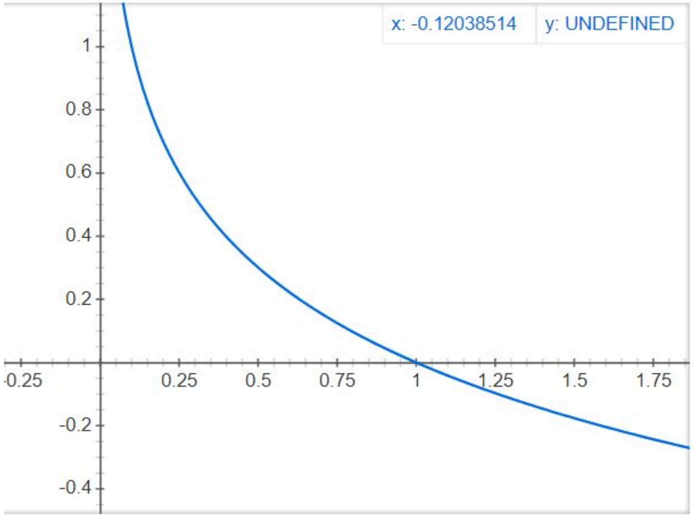
Parameter selection (optimization)

- Given:
  - Training dataset $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$
  - Each example: $x \in [x_0, x_1, \dots, x_n]^T$, with $x_0 = 1$ and $y \in \{0,1\}$.
  - General form of $h_\theta(x)$: $h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$
- How to select the parameter vector θ?

Quiz AM-2.1

- When x is restricted to the range [0,1], what is the range of values for $y = -\log(x)$?

- A) [-∞, +∞]
- B) [0, +∞]
- C) [-∞, 0]
- D) [+∞, 0]

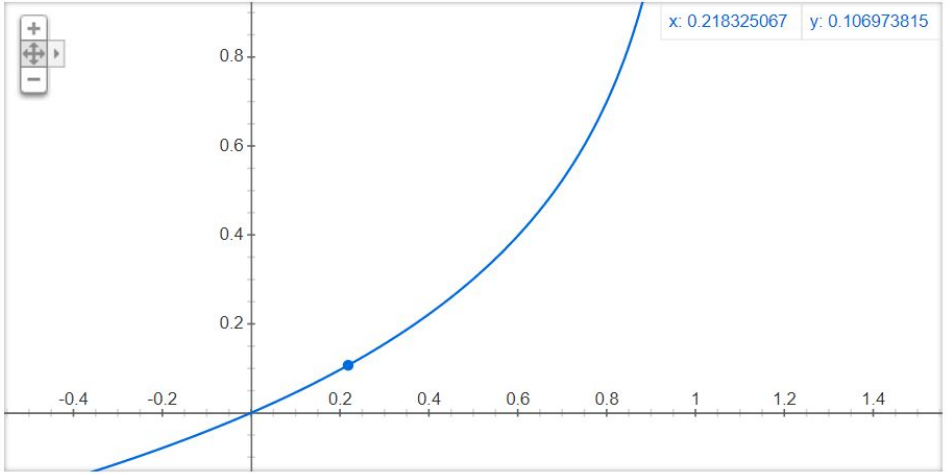


Quiz AM-2.2

27

- When x is restricted to the range [0,1], what is the range of values for $y = -\log(1-x)$?

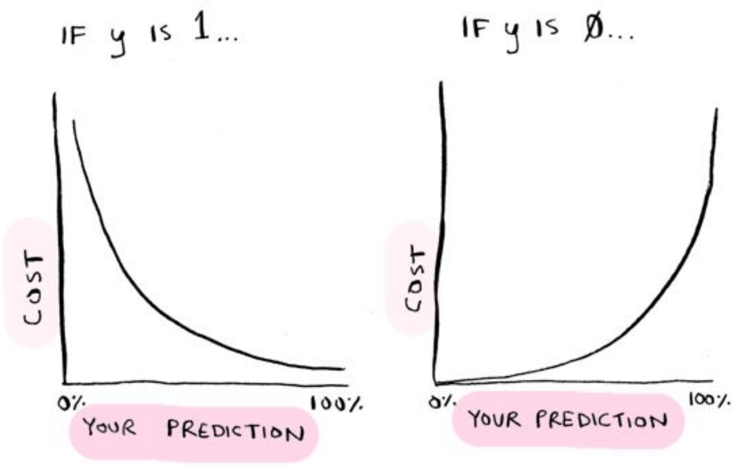
  - A) [-∞, +∞]
  - B) [0, +∞]
  - C) [-∞, 0]
  - D) [0,1]



Cost Function – Logistic Regression

- Cost function for logistic regression:

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



Cost Function – Intuition

- Observe que o custo é zero quando a previsão e a hipótese coincidem:

$$\boxed{(h_\theta = 1) \text{ e } (y = 1) \implies \text{Custo} = 0} \quad \boxed{(h_\theta = 0) \text{ e } (y = 0) \implies \text{Custo} = 0}$$

- No entanto, se a previsão e a hipótese divergirem:

$$\boxed{\begin{array}{l} (h_\theta \to 0) \text{ e } (y = 1) \implies \text{Custo} \to \infty \\ (h_\theta \to 1) \text{ e } (y = 0) \implies \text{Custo} \to \infty \end{array}}$$

- Portanto, o algoritmo é penalizado quando faz uma previsão incorreta.

Simplifying the cost function

- $$\boxed{\mathrm{Custo}(h_\theta(x), y) = -y \ \log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))}$$

- Note that
  - if $y = 1$, the second term is equal to zero.
  - if $y = 0$, the first term is equal to zero.

Simplifying the cost function

- $$\boxed{\text{Custo}(h_\theta(x), y) = -y\ \log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))}$$

- Note that
  - if $y = 1$, the second term is equal to zero.
  - if $y = 0$, the first term is equal to zero.
- Then we can rewrite the cost function:

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

Optimization

Aqui, estudamos como minimizar a função de custo da Regressão Logística utilizando o algoritmo de Descida do Gradiente.

Minimization with Gradient Descent

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

- To minimize $J(\theta)$, GD does the following:

$$\begin{aligned}
&Repetir\ \{\\
&\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)\\
&\}
\end{aligned}$$

Minimization with Gradient Descent

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

- It is possible to prove that:

$$\frac{\partial}{\partial\theta_j}J(\theta) = \sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}$$

Minimization with Gradient Descent

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

- So, to minimize $J(\theta)$, we do:

$$\begin{aligned} &Repetir \ \{ \\ &\quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \\ &\} \end{aligned}$$

$$\begin{aligned} &Repetir \ \{ \\ &\quad \theta_j := \theta_j - \alpha \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)} \\ &\} \end{aligned}$$

Final remarks

- The discussion of the following topics that we made in the context of linear regression also applies to logistic regression:
    - Gradient debugging
    - Learning rate value
    - Feature scaling