Regressão Logística

Mineração de Dados

Ronaldo C. Prati[1]

[1]Universidade Federal do ABC (UFABC), ronaldo.prati@ufabc.edu.br

Introdução

Introdução

- Na análise exploratória de dados, vimos que a associação entre uma variável numérica e categórica é normalmente analisada considerando as estatísticas descritivas para cada valor categórico
- E se codificarmos a variável categórica como uma variável numérica?

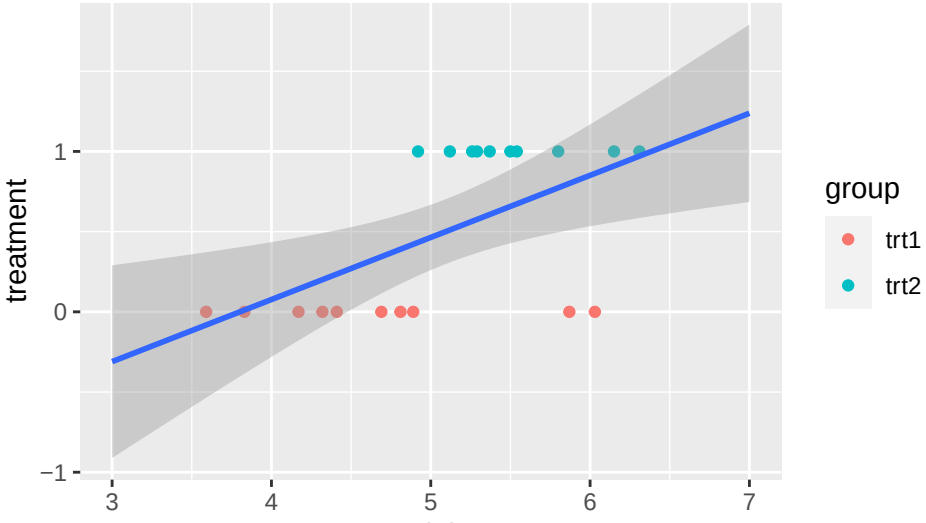
Base de dados

Considere a sequinte base de dados:

weight	group	weight	group
4.81	trt1	6.31	trt2
4.17	trt1	5.12	trt2
4.41	trt1	5.54	trt2
3.59	trt1	5.50	trt2
5.87	trt1	5.37	trt2
3.83	trt1	5.29	trt2
6.03	trt1	4.92	trt2
4.89	trt1	6.15	trt2
4.32	trt1	5.80	trt2
4.69	trt1	5.26	trt2

Introdução

- Se codificarmos o tretamento `trt1` como 0, e `trt2` como 1, podemos analisar a relação linear entre essas variáveis:



Introdução

- Essa codificação tem algumas limitações:
  - Temos valores discrepantes (menores que zero e maiores que 1).
  - Não há uma interpretação clara sobre os coeficientes e a variável resposta.

Transformação

- $y = b_0 + b_1 x$: se você mudar x em 1, espere uma mudança em y de b_1
- $\log(y) = b_0 + b_1 x$: se você mudar o x em 1, espere uma mudança em y de $100 * b_1$ por cento
- $y = b_0 + b_1 \log(x)$: se você mudar x em 1%, expere uma mudança em y de $b_1/100$.
- $\log(Y) = b_0 + b_1 \log(x)$: se você mudar x em 1%, espere uma mudança em y em b_1 %

Log Odds

- Seja p a probabilididade do exemplo ser da classe positiva. Na regressão logistica, queremos modelar

$$\log\left(\frac{p}{1-p}\right) = b_1 x + b_0$$

que com um pouco de algebra, pode ser reescrito como:

$$p = \frac{1}{1 + e^{-(b_1 x + b_0)}}$$

$\log(\frac{p}{1-p})$ é chamada de logarítmo razão de chance (*log odds ratio*).

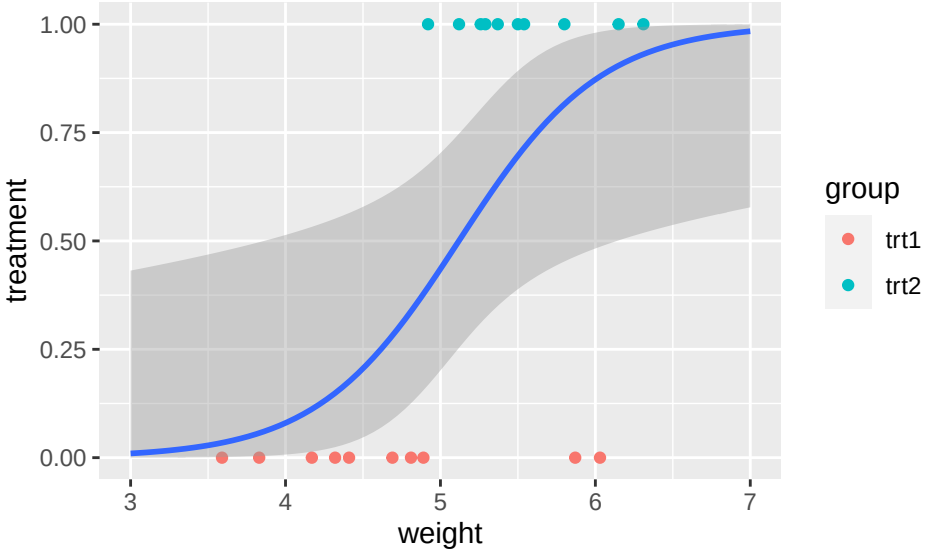
Derivação

$$
\begin{aligned}
p &= \frac{1}{1 + e^{-(b_1 x + b_0)}} \\
p \cdot \left(1 + e^{-(b_1 x + b_0)}\right) &= 1 \\
p + p e^{-(b_1 x + b_0)} &= 1 \\
p e^{-(b_1 x + b_0)} &= 1 - p \\
\frac{p}{e^{(b_1 x + b_0)}} &= 1 - p \\
\frac{p}{1 - p} &= e^{(b_1 x + b_0)} \\
\log\left(\frac{p}{1 - p}\right) &= b_1 x + b_0
\end{aligned}
$$

Regressão logística

- A regressão logística nos permite predizer a probabilide da classe, baseado em variáveis contínuas (apesar de categóricas também serem utilizadas)
- O modelo logístico binário pode ser usado quando a classe tem somente duas categorias
- O modelo logístico multinomial pode ser usado quando temos mais de duas classes
- Nos permite calcular como uma mudança em x afeta a chance de p

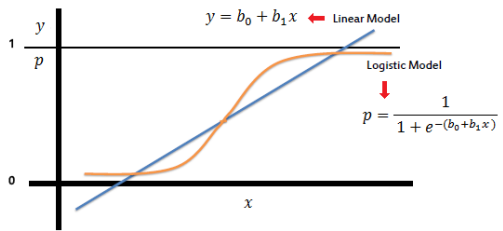
Função Logística



Função Logística

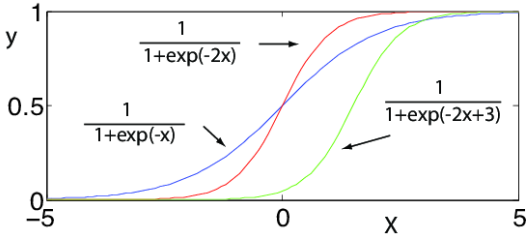
A função logística (também chamada de **função sigmoid**) ´ está no "coração" da regressão logística

$$f(x) = \frac{1}{1 + e^{-x}}$$



Regressão Logística

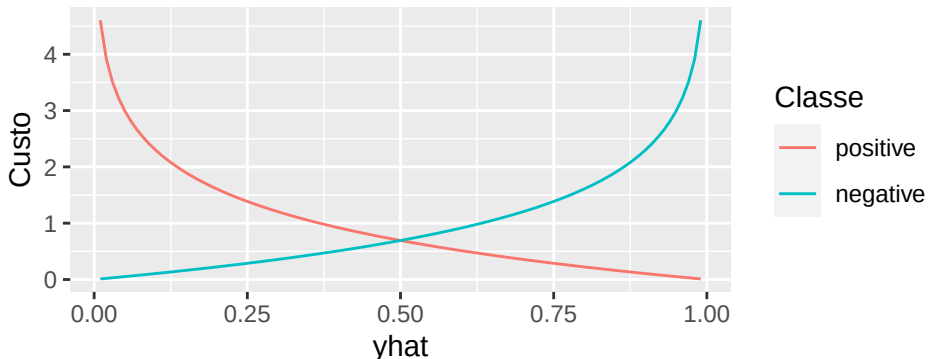
- Como encontrar b_0 e b_1 ?



Função de Custo

- Seja $\hat{y}$ o valor predito pelo modelo, para um certo valor de parâmetros b_0 e b_1.
- Vamos definir a função de erro:

$$Custo(y, \hat{y}) = \begin{cases} -\log(\hat{y}), \text{se } y = 1 \\ -\log(1 - \hat{y}), \text{se } y = 0 \end{cases}$$



Função de Custo

- Como $y = \{0, 1\}$, podemos reescrever como

$$Custo(y, \hat{y}) = -\left(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})\right)$$

- Custo médio

$$E(Custo(y, \hat{y})) = -\frac{1}{n} \sum_{i}^{n} \left(y_i \cdot \log(\hat{y}_i) + (1 - y) \cdot \log(1 - \hat{y}_i)\right)$$

Fronteira de decisão

- Para predizer a classe, utilizamos aquela com maior probabilidade, segundo o modelo
  - Isso é equivalente a predizer a classe 1 se $\hat{y} > 0.5$, e a classe 0 caso contrário
- Observando a função logística, temos que $\hat{y} = 0.5$ quando $b_0 + b_1 x = 0$
- A linha $b_0 + b_1 x = 0$ é a fronteira de decisão entre as classes

Regressão Logística

- Apesar de termos apresentado para uma variável, raciocínio similar se aplica a casos multivariados (teremos um vetor de coeficientes ao invés de um único)
- Existe diversas maneiras de minimizar o erro para o conjunto de treinamento
  - Máxima verossimilhança (pode ser computacionalmete computacionalmente caro)
  - Método de descida do gradiente
  - Métodos de otimização
- Poblemas com mais de duas classes podem ser tratados fazendo uma decomposição em vários problemas binários, e predizendo a classe mais provável dentre as possíveis combinações.