Regressão e Regressão Logística

Carlos Diego Rodrigues

28 de outubro de 2021

Universidade Federal do Ceará

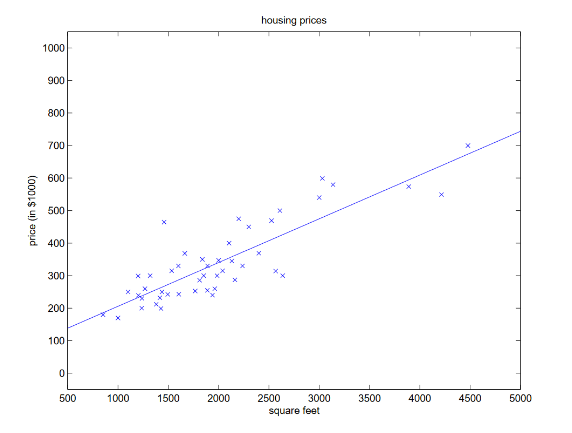
Métodos lineares

- Regressão: consiste em encontrar uma função que traduz as relações entre uma variável dependente (classe) e diversas variáveis independentes (atributos). Essa relação se traduz matematicamente como uma função e o tipo mais comum é a função **linear**.

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_1 x_1 + \cdots + \theta_n x_n = \sum_{j=0}^{n} \theta_j x_j = \theta^T x$$

- Método canônico na estatística
- Existem diversas formas de encontrar uma regressão linear, mas tipicamente utiliza-se o método de **mínimos quadrados**.

Exemplo regressão



Mínimos quadrados

- Neste método queremos minimizar o erro quadrático do modelo gerado em relação às observações presentes no conjunto de treinamento.

$$\min \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2 = 2 \min \frac{1}{2} (h_\theta(x^i) - y^i)^2 = J(\theta)$$

- Isto pode ser feito através de um processo de otimização iterativo que atualiza um vetor tentativa θ iterativamente até convergir para aquele que corresponde ao mínimo da expressão acima, na direção da derivada da função a ser otimizada.

$$\theta_j \leftarrow \theta_j - \alpha \frac{d}{d\theta_j} J(\theta)$$

- O fator α é chamado taxa de aprendizado do método.

Derivando e concluindo

- Derivando temos:

$$\begin{aligned}\frac{d}{d\theta_j}J(\theta) &= \frac{d}{d\theta_j}(h_\theta(x) - y)^2 \\ &= 2 \cdot \frac{1}{2}(h_\theta(x) - y) \cdot \frac{d}{d\theta_j}(h_\theta(x) - y) \\ &= (h_\theta(x) - y) \cdot \frac{d}{d\theta_j}(\sum_{k=1}^{n} \theta_k x_k - y) \\ &= (h_\theta(x) - y)x_j\end{aligned}$$

- Portanto podemos atualizar o valor de θ fazendo

$$\theta_j \leftarrow \theta_j + \alpha(y^i - h_\theta(x^i))x_j^i$$

- O algoritmo consiste portanto em repetir estas atualizações até que se obtenha a convergência desejada.

Algoritmo Descida do gradiente

Algorithm 1 Descida do gradiente

Entrada: X, Y: conjunto de atributos e classes

Saída: θ: um modelo de regressão

Encontre um valor inicial para θ
repeat
    for $i \leftarrow 1 \cdots m$ **do**
        for $j \leftarrow 1 \cdots n$ **do**
            $\theta_j \leftarrow \theta_j + \alpha(y^i - h_\theta(x^i))x_j^i$
until convergência
Retorna θ

Explicação probabilística

- Podemos pensar que a resposta do modelo de regressão se relaciona aos dados originais através da expressão:

$$y^i = \theta^T x^i + \epsilon^i$$

  onde os valores ϵ^i são independentes e normalmente distribuídos ($\epsilon^i \sim N(0, \sigma)$.
- Dessa forma, vamos considerar a distribuição de cada erro:

$$p(\epsilon^i) = \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{(\epsilon^i)^2}{2\sigma^2}\right)$$

- Portanto:

$$p(y^i|x^i;\theta) = \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{(y^i - \theta^T x^i)^2}{2\sigma^2}\right)$$

Minimizando a verossimilhança

- Definindo a verossimilhança:

$$L(\theta) = \prod_{i=1}^{m} p(y^i|x^i;\theta)$$

- No entanto o produtório é uma função difícil de trabalhar. Podemos tomar, contudo, o logaritmo desta função, cujo máximo é dado pelo mesmo argumento.
- Assim temos:

$$\begin{aligned} l(\theta) &= log\ L(\theta) \\ &= mlog\frac{1}{\sqrt{2\pi\sigma}} - \frac{1}{\sigma^2} \cdot \frac{1}{2}\sum_{i=1}^{m}(y^i - \theta^T x^i)^2 \end{aligned}$$

- A conclusão é que minimizar o erro sob estas hipóteses é equivalente a encontrar o melhor θ com o método de mínimos quadrados.

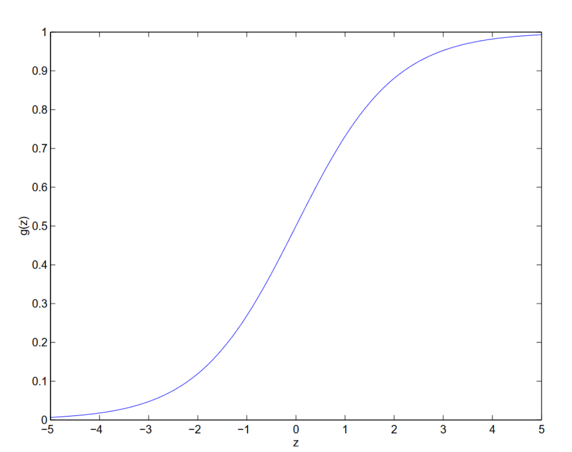
Regressão logística

- Se ignorarmos o fato que y é uma variável binária poderíamos usar a regressão para os problemas de classificação.
- No entanto facilmente encontraríamos exemplos onde esse método funciona muito, pois os valores intermediários são difíceis de distinguir para uma função linear.
- Isto pode ser corrigido se utilizarmos uma função logística.

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

onde a função g é chamada função logística ou sigmoid.

Exemplo regressão



Derivando a função logística

- Para aplicarmos uma metodologia similar à anterior seria interessante conhecer a derivada da função g:

$$\begin{aligned} g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}e^{-z} \\ &= \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right) \\ &= g(z)(1-g(z)) \end{aligned}$$

- Vamos utilizar a mesma ideia da regressão linear para derivar um estimador interessante θ para a verossimilhança do modelo logístico.

Verossimilhança do modelo logístico

- Vamos supor que:

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

- Isto pode ser transformado em uma função de distribuição:

$$p(y|x;\theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}$$

- Definindo a função de verossimilhança:

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} p(y^i|x^i;\theta) \\ &= \prod_{i=1}^{m} (h_\theta(x^i))^{y^i} (1 - h_\theta(x^i))^{1-y^i} \end{aligned}$$

Otimizando o logaritmo

- Definindo o logaritmo da verossimilhança

$$
\begin{aligned}
l(\theta) &= logL(\theta) \\
&= \sum_{i=1}^{m} y^i logh_\theta(x^i)) + (1-y^i)log(1-h_\theta(x^i))
\end{aligned}
$$

- Derivando a função $l(\theta)$:

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

Descida do gradiente para Regressão logística

- Podemos otimizar o valor de $l(\theta)$ a partir do valor de sua derivada, fazendo

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} l(\theta)$$

- Encontramos portanto que a minimização da verossimilhança (e portanto a definição de um modelo) para a regressão logística pode ser obtida de forma semelhante àquela obtida na regressão linear, através de sucessivas atualizações do modelo θ:

$$\theta_j \leftarrow \theta_j + \alpha(y^i - h_{\theta}(x^i))x_j$$

- Esta função é muito semelhante, porém não idêntica àquela encontrada para a regressão pois as funções $h_{\theta}(x)$ são distintas.