

# Trivial Big Data

wu

2022 年 11 月 5 日

## 目录

1 一元线性回归	1
2 多元线性回归	3
3 逻辑回归	4

## 1 一元线性回归

方差

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

标准差

$$S_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

协方差

$$Cov = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

相关系数

$$\rho = \frac{Cov}{S_x S_y}$$

一元线性回归：找一条直线，拟合数据点

$$y = \beta_0 + \beta_1 x$$

最小二乘法

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n ((\beta_0 + \beta_1 x_i) - y_i)^2$$

$$f(\beta_0, \beta_1) = \sum_{i=1}^n ((\beta_0 + \beta_1 x_i) - y_i)^2$$

$$\frac{\partial f}{\partial \beta_0} = 2 \sum_{i=1}^n (\beta_0 + \beta_1 x_i - y_i) = 0$$

$$\Rightarrow n\beta_0 + \beta_1 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0$$

$$\Rightarrow \beta_0 + \beta_1 \bar{x} - \bar{y} = 0$$

$$\frac{\partial f}{\partial \beta_1} = 2 \sum_{i=1}^n x_i (\beta_0 + \beta_1 x_i - y_i) = 0$$

$$\sum_{i=1}^n x_i (\beta_0 + \beta_1 x_i - y_i) = 0$$

注意到

$$\sum_{i=1}^n \bar{x} (\beta_0 + \beta_1 x_i - y_i) = \bar{x} (n\beta_0 + \beta_1 \sum_{i=1}^n x_i - \sum_{i=1}^n y_i) = 0$$

所以

$$\sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 x_i - y_i) = 0$$

而

$$\sum_{i=1}^n (x_i - \bar{x}) (\beta_0 + \beta_1 \bar{x} - \bar{y}) = (\beta_0 + \beta_1 \bar{x} - \bar{y}) \sum_{i=1}^n (x_i - \bar{x}) = 0$$

所以

$$\sum_{i=1}^n (x_i - \bar{x}) (\beta_1 (x_i - \bar{x}) - (y_i - \bar{y})) = 0$$

$$\begin{aligned}\beta_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{Cov}{S_x^2} \\ &= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}} = \rho \frac{S_y}{S_x}\end{aligned}$$

因此

$$y = \bar{y} - \beta_1 \bar{x} + \beta_1 x \Rightarrow y - \bar{y} = \beta_1 (x - \bar{x}) = \rho \frac{S_y}{S_x} (x - \bar{x})$$

## 2 多元线性回归

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

最小二乘法

$$\min_{\beta_0, \dots, \beta_m} \sum_{i=1}^n ((\beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}) - y_i)^2$$

令

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

最小二乘形式

$$\min_{\beta} \|X\beta - y\|^2$$

$$g(\beta) = \langle w, \beta \rangle = w^T \beta = \sum_{i=0}^m w_i \beta_i$$

$$\nabla g = \begin{pmatrix} \frac{\partial g}{\partial \beta_0} \\ \frac{\partial g}{\partial \beta_1} \\ \vdots \\ \frac{\partial g}{\partial \beta_m} \end{pmatrix} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{pmatrix} = w$$

假设  $A = A^T$

$$h(\beta) = \langle A\beta, \beta \rangle = \beta^T A \beta = \sum_{i,j} a_{ij} \beta_i \beta_j$$

定义  $p(u, v) = \langle Au, v \rangle = \langle Av, u \rangle$  令

$$u(\beta) = \beta, v(\beta) = \beta \Rightarrow h(\beta) = p(u(\beta), v(\beta))$$

$$\nabla h = \frac{\partial p}{\partial u} \frac{\partial u}{\partial \beta} + \frac{\partial p}{\partial v} \frac{\partial v}{\partial \beta} = Av(\beta) + Au(\beta) = 2A\beta$$

$$\begin{aligned} f(\beta) &= (X\beta - y)^T (X\beta - y) \\ &= (\beta^T X^T - y^T)(X\beta - y) \\ &= \beta^T X^T X \beta - \beta^T X^T y - y^T X \beta + y^T y \end{aligned}$$

$$\nabla_{\beta} f = 2X^T X \beta - X^T y - X^T y = 2(X^T X \beta - X^T y) = 0$$

因此

$$\beta = (X^T X)^{-1} X^T y$$

为了增加鲁棒性，通常会最小化如下目标函数

$$\|X\beta - y\|^2 + \lambda \|\beta\|^2 \quad (\lambda > 0)$$

此时

$$\beta = (X^T X + \lambda I)^{-1} X^T y$$

### 3 逻辑回归

分类问题建模：

$$f : \mathbb{R}^m \rightarrow \{0, 1\}$$

$$f(x) = \sigma(\beta^T x)$$

$\sigma$  的一种取法:

$$\sigma(z) = \begin{cases} 1 & z \geq 0 \\ 0 & z < 0 \end{cases}$$

问题:  $\sigma$  在 0 点处间断

光滑化: 逻辑函数 (Logistic Function)

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$
$$f(x) = \sigma(\beta^T x) = \frac{1}{1 + e^{-z}}$$

求解如下优化问题:

$$\min_{\beta} \sum_{i=1}^n \left( \frac{1}{1 + e^{-\beta^T x_i}} - y_i \right)^2$$

对于输入  $x$ , 当  $f(x) \geq 0.5$  时预测 1, 当  $f(x) < 0.5$  时预测  $y = 0$ 。

如何求  $\beta$ : 梯度下降法

$$\min_{\beta} C(\beta)$$

$$\beta_{m+1} = \beta_m - \lambda \nabla C(\beta_m)$$

$$C(x) \approx C(x') + \nabla C(x')(x - x')$$

$$\begin{aligned} C(\beta_{m+1}) &= C(\beta_m - \lambda \nabla C(\beta_m)) \\ &\approx C(\beta_m) - \lambda \|\nabla C(\beta_m)\|^2 \\ &\leq C(\beta_m) \end{aligned}$$

混淆矩阵

		Actual class	
		positive class	negative class
Predicted class	positive class	True Positive(TP)	False Positive(FP)
	negative class	False Negative(FN)	True Negative(TN)