

Logistic regression

- [🔗 背景和基本内容](#)
 - [🔗 基本设定](#)
 - [🔗 回顾：线性回归 \(Linear Regression\)](#)
 - [🔗 Logistic Regression 的引入](#)
 - [🔗 Some definition](#)
 - [🔗 优势比 \(Odds\)](#)
 - [🔗 Logit 变换](#)
 - [🔗 针对响应变量的 Logit 变换](#)
 - [🔗 Logit 的逆变换 \(Sigmoid\)](#)
 - [🔗 Sigmoid 性质](#)
- [🔗 Logistic 回归模型与预测](#)
 - [🔗 观测与响应变量](#)
 - [🔗 新观测的预测](#)
- [🔗 Logistic 回归中 \$\beta\$ 的 MLE 与损失函数](#)
 - [🔗 极大似然估计 \(MLE\)](#)
 - [🔗 \$\beta\$ 的更深层含义：从线性回归到 Logistic 回归的损失函数](#)
 - [🔗 多元线性回归中 \$\beta\$ 的估计 \(LSE 与 MLE 的等价性\)](#)
 - [🔗 最小二乘估计 \(LSE\)](#)
 - [🔗 极大似然估计 \(MLE\)](#)
 - [🔗 Logistic 回归中 \$\beta\$ 的估计 \(损失函数的“距离”意义\)](#)
 - [🔗 损失函数定义](#)
 - [🔗 损失函数的“距离”含义](#)
 - [🔗 线性回归与 Logistic 回归的核心联系](#)
 - [🔗 损失函数 \(交叉熵\)](#)
- [🔗 Logistic 回归中 \$\beta\$ 的数值求解：牛顿-拉夫逊法 \(IRLS\)](#)
 - [🔗 对数似然的泰勒展开](#)

- [🔗](#) 梯度与海森矩阵的表达式
- [🔗](#) 迭代更新公式
- [🔗](#) 等价于加权最小二乘
- [🔗](#) 迭代加权最小二乘 (IRLS)
- [🔗](#) 停止条件

logistic regression 对率回归 (常称为“逻辑回归”)

背景和基本内容

基本设定

有 n 个独立观测 $\{x_i^T, y_i\}_{i=1}^n$, 其中:

- 设计矩阵 X (维度 $n \times (k+1)$) 定义为:

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix}$$

- y_i 是 **两类响应变量** (Type I/Type II), 取值为:

$$y_i = \begin{cases} 1 & \text{对应概率 } p_i \\ 0 & \text{对应概率 } 1 - p_i \end{cases}$$

回顾: 线性回归 (Linear Regression)

线性回归中, 响应变量 y_i 是 **连续无界变量** (取值 $(-\infty, \infty)$), 模型形式为:

$$y_i = x_i^T \beta + \varepsilon_i$$

满足:

- 误差项 $\varepsilon_i \sim N(0, \sigma^2)$, 即 $y_i \sim N(x_i^T \beta, \sigma^2)$;
- 条件期望为: $\mathbb{E}(y_i | x_i) = x_i^T \beta$.

Logistic Regression 的引入

当 y_i 是 **离散有界变量** (仅取 $\{0, 1\}$) 时不能直接用线性回归, 因为线性回归的期望无界, 而概率必须在 $[0, 1]$ 中。

对于离散响应变量:

$$\mathbb{E}(y_i | x_i) = P(y_i = 1 | x_i) = p_i$$

于是需要对 p_i 做 **变换映射**, 将其从 $[0, 1]$ 扩展到无界实数, 再建立线性关系, 这就是 Logistic Regression 的核心思想。

Some definition

优势比 (Odds)

事件 E 的优势比:

$$\text{Odds}(E) = \frac{P(E)}{1 - P(E)} = \frac{P(E)}{P(E^c)}$$

Logit 变换

对概率 $p \in (0, 1)$:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

将 $(0, 1)$ 映射到 $(-\infty, \infty)$ 。

针对响应变量的 Logit 变换

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \log(\text{Odds}(y_i = 1))$$

Logit 的逆变换 (Sigmoid)

若 $y = \log \frac{x}{1-x}$, 则:

指数化:

$$e^y = \frac{x}{1-x}$$

解得:

$$x = \frac{e^y}{1 + e^y}$$

这就是 Sigmoid 函数:

$$\sigma(y) = \frac{e^y}{1 + e^y}$$

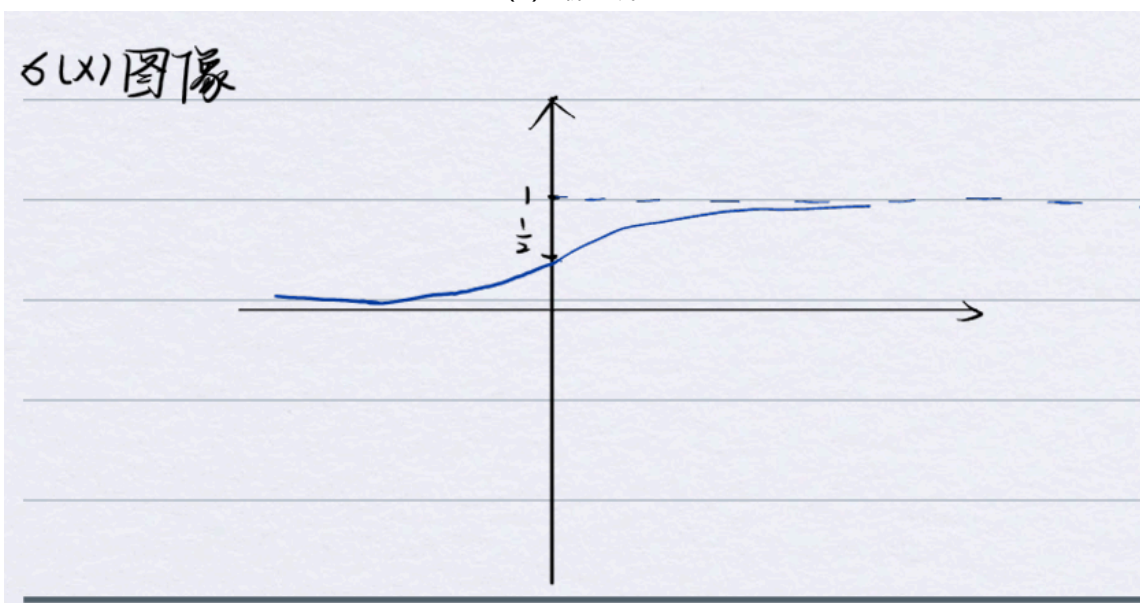
Sigmoid 性质

- (1) 对称性:

$$\sigma(-x) = 1 - \sigma(x)$$

-

(2) 渐近线:



$$x \rightarrow -\infty : \sigma(x) \rightarrow 0;$$

$$x \rightarrow +\infty : \sigma(x) \rightarrow 1.$$

- (3) 导数:

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Logistic 回归模型与预测

观测与响应变量

已知观测 $\{x_i^T, y_i\}_{i=1}^n$:

- x_i : $(k+1) \times 1$ 特征向量
- $y_i \in \{0, 1\}$, 其中:

$$P(y_i = 1 \mid x_i) = p_i$$

Logistic 模型核心关系

$$\text{logit}(p_i) = x_i^T \beta$$

条件期望显式表达式

利用 Sigmoid:

$$p_i = \mathbb{E}(y_i \mid x_i) = \sigma(x_i^T \beta) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}}$$

新观测的预测

对新 x_{n+1} :

1. 计算概率

$$\hat{p}_{n+1} = \sigma(x_{n+1}^T \beta)$$

2. 判别:

- $\hat{p}_{n+1} > 1/2 \Rightarrow y_{n+1} = 1$
- $\hat{p}_{n+1} < 1/2 \Rightarrow y_{n+1} = 0$

Logistic 回归中 β 的 MLE 与损失函数

极大似然估计 (MLE)

对单个样本：

$$P(y_i | x_i) = p_i^{y_i} (1 - p_i)^{1-y_i}$$

联合似然：

$$L(\beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$$

对数似然：

$$\ell(\beta) = \sum_{i=1}^n [y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

代入 $p_i = \sigma(x_i^T \beta)$ 后可写为：

$$\ell(\beta) = \sum_{i=1}^n [y_i x_i^T \beta + \log(1 - \sigma(x_i^T \beta))]$$

MLE：

$$\hat{\beta} = \arg \max_{\beta} \ell(\beta)$$

一阶导数（得似然方程）：

$$\sum_{i=1}^n [y_i - \sigma(x_i^T \beta)] x_i = 0$$

此方程无解析解，需要数值方法求解。

β 的更深层含义：从线性回归到 Logistic 回归的损失函数

多元线性回归中 β 的估计 (LSE 与 MLE 的等价性)

考虑多元线性回归模型：

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$$

最小二乘估计 (LSE)

线性回归中， β 的最小二乘估计是最小化 **残差的 L_2 距离**：

$$\hat{\beta}_{\text{LSE}} = \arg \min_{\beta} \|Y - X\beta\|_2^2$$

其中

$\|Y - X\beta\|_2^2 = \sum_{i=1}^n (y_i - x_i^T \beta)^2$ ，
代表观测值与拟合值的“欧氏距离平方和”。

极大似然估计 (MLE)

由于 $\varepsilon \sim N(0, \sigma^2 I)$ ， Y 的似然函数为：

$$L(\beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \|Y - X\beta\|_2^2 \right\}$$

最大化似然函数等价于最小化指数部分的残差平方和，因此：

$$\hat{\beta}_{\text{MLE}} = \arg \max_{\beta} L(\beta) = \arg \min_{\beta} \|Y - X\beta\|_2^2 = \hat{\beta}_{\text{LSE}}$$

此时，线性回归的 LSE 与 MLE 是等价的， β 的估计本质是最小化 **欧氏距离**。

Logistic 回归中 β 的估计 (损失函数的“距离”意义)

Logistic 回归中，响应变量 $y_i \in \{0, 1\}$ ，无法直接用欧氏距离度量误差，因此使用 **交叉熵损失函数**。

损失函数定义

Logistic 回归中, β 的估计是最小化 **平均交叉熵损失**:

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n f(y_i)$$

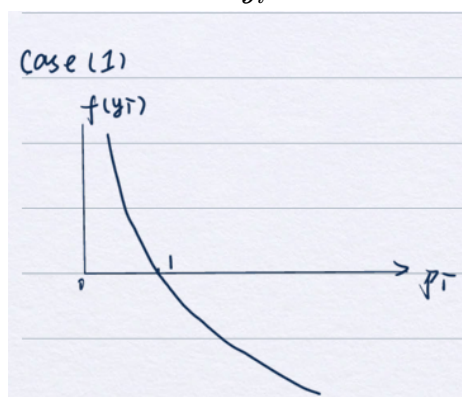
其中单个观测的损失函数为:

$$f(y_i) = -[y_i \log p_i + (1 - y_i) \log(1 - p_i)], \quad p_i = \sigma(x_i^T \beta)$$

损失函数的“距离”含义

交叉熵虽然不是欧氏距离, 但可看作 **概率空间的距离**, 衡量“预测概率与真实标签的匹配程度”。

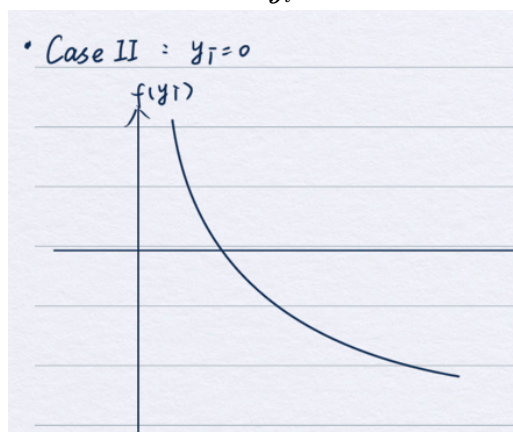
- **Case 1: $y_i = 1$**



损失为 $f(y_i) = -\log p_i$:

- 若 $p_i = 1$ (预测完全正确), 损失 = 0;
- 若 $p_i \rightarrow 0$ (预测完全错误), 损失 $\rightarrow +\infty$ 。

- **Case 2: $y_i = 0$**



损失为 $f(y_i) = -\log(1 - p_i)$:

- 若 $p_i = 0$ (预测完全正确), 损失 = 0;
- 若 $p_i \rightarrow 1$ (预测完全错误), 损失 $\rightarrow +\infty$ 。

线性回归与 Logistic 回归的核心联系

两类模型中, β 的估计本质都是 **最小化“观测与拟合的距离”**, 但“距离”的定义适配了响应变量类型:

- 线性回归: 连续响应 \rightarrow **欧氏距离** (残差平方和)
- Logistic 回归: 二分类响应 \rightarrow **交叉熵损失** (概率空间距离)

这一“距离最小化”思想是参数估计的核心逻辑。

损失函数 (交叉熵)

单样本损失:

$$f(y_i) = -[y_i \log p_i + (1 - y_i) \log(1 - p_i)]$$

整体损失:

$$\text{Loss}(\beta) = \frac{1}{n} \sum_{i=1}^n f(y_i) = -\frac{1}{n} \ell(\beta)$$

MLE 等价于最小化 Loss。

Logistic 回归中 β 的数值求解: 牛顿-拉夫逊法 (IRLS)

对数似然:

$$\ell(\beta) = \sum_{i=1}^n [y_i x_i^T \beta + \log(1 - \sigma(x_i^T \beta))]$$

对数似然的泰勒展开

对对数似然函数 $\ell(\beta)$, 在当前迭代值 $\beta^{(t)}$ 处做 **二阶泰勒展开**, 并最大化该近似式以更新 β :

$$\hat{\beta}_{\text{MLE}} = \arg \max_{\beta} \left[\ell(\beta^{(t)}) + \ell'(\beta^{(t)})^T (\beta - \beta^{(t)}) + \frac{1}{2} (\beta - \beta^{(t)})^T H(\beta^{(t)}) (\beta - \beta^{(t)}) \right]$$

(记展开后的近似函数为 $f(\beta)$)

梯度与海森矩阵的表达式

- **梯度（一阶导数）：**

对数似然的梯度为 (Y 是响应向量, $P = [\sigma(x_1^T \beta), \dots, \sigma(x_n^T \beta)]^T = \mathbb{E}(Y|X)$) :

$$\ell'(\beta) = (Y^T - P^T)X$$

- **海森矩阵（二阶导数）：**

海森矩阵为 ($D = \text{diag}(\sigma(x_1^T \beta)(1 - \sigma(x_1^T \beta)), \dots, \sigma(x_n^T \beta)(1 - \sigma(x_n^T \beta)))$) :

$$H(\beta) = \sum_{i=1}^n -\sigma(x_i^T \beta) x_i x_i^T = -X^T D X$$

迭代更新公式

对泰勒展开式 $f(\beta)$ 求导并令导数为 0, 得到迭代公式:

$$f'(\beta) = X^T(Y - P) - X^T D X (\hat{\beta} - \beta^{(t)}) = 0$$

整理得下一次迭代的参数:

$$\hat{\beta}^{(t+1)} = (X^T D X)^{-1} X^T (Y - P) + \hat{\beta}^{(t)}$$

这一方法称为 **迭代加权最小二乘 (Iterative Reweighted Least Squares, IRLS)** 。

等价于加权最小二乘

令加权设计矩阵 $\tilde{X} = D^{1/2} X$ 、加权响应向量 $\tilde{Y} = D^{1/2}(X\beta^{(t)} + D^{-1}(Y - P))$, 则迭代公式等价于:

$$\hat{\beta}^{(t+1)} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y} = \arg \min_{\beta} \|\tilde{Y} - \tilde{X}\beta\|^2$$

即每次迭代都是对 **加权后的数据** 做最小二乘估计。

设

$$Y = [y_1, \dots, y_n]^T,$$

$$P = [p_1, \dots, p_n]^T,$$

$$X = [x_1, \dots, x_n]^T.$$

则：

$$\ell'(\beta) = X^T(Y - P)$$

迭代加权最小二乘 (IRLS)

令

$$\tilde{X} = (D^{(t)})^{1/2} X,$$

$$\tilde{Y} = (D^{(t)})^{1/2}(X\beta^{(t)} + (D^{(t)})^{-1}(Y - P^{(t)})), \text{ 则:}$$

$$\beta^{(t+1)} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}$$

这是一个加权最小二乘问题。

停止条件

当

$$\|\beta^{(t+1)} - \beta^{(t)}\| < 10^{-6}$$

或对数似然变化足够小时停止。