对数几率回归

Logistic Regression 比较好的翻译是周志华《机器学习》的"**对数几率回归**"，亦被翻译为“**逻辑回归**”或“**逻辑斯蒂回归**”。

虽然名字是回归，但对数几率回归是一种常用于**分类**(classification)的算法，可以解决二分类或多分类问题，因为**对数几率回归**的输出是离散的，**线性回归**的输出才是连续的。

二分类问题

现假设二分类问题的输出标记为 $y \in \{0,1\}$，可以有下列`阶跃函数`（unit-step function）：

$$y = \begin{cases} 0, & z < 0 \\ 0.5, & z = 0 \\ 1, & z > 0 \end{cases} \tag{1}$$

单位阶跃函数不连续，我们希望得到其**单调可微**的替代函数，`对数几率函数`（logistic function）正是这样一个常用的替代函数。

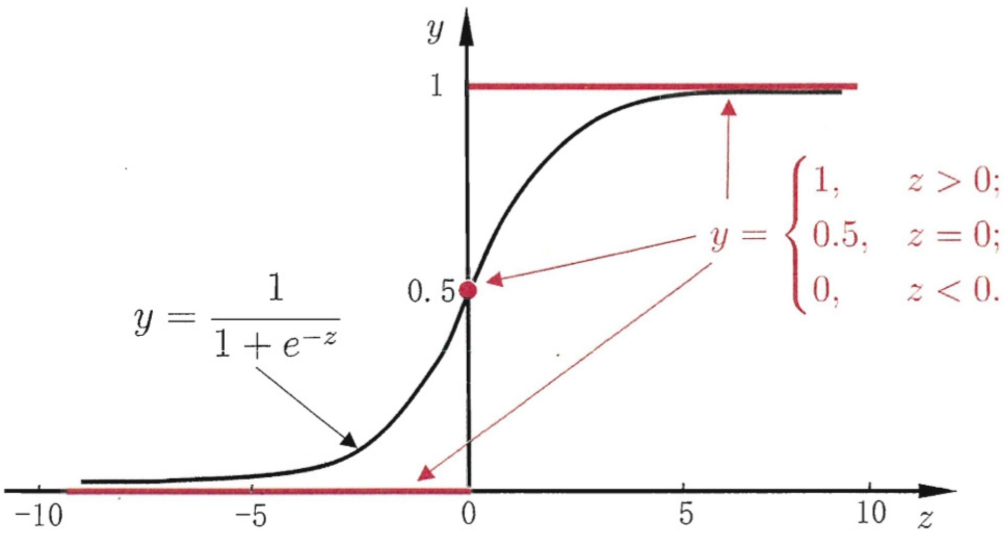

图 3.2 单位阶跃函数与对数几率函数

$$y=\frac{1}{1+e^{-z}} \tag{2}$$

其实这个函数就是 **Sigmoid 函数**，将 z 用 w^Tx+b 表示：

$$y=\frac{1}{1+e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}} \tag{3}$$

取对数，可转化为为：

$$\ln\frac{y}{1-y}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b \tag{4}$$

对数几率的名字怎么来的呢？若将 y 为视为样本 x 作为正例的相对可能性，则 $1-y$ 是其反例可能性，两者（正反例可能性）的比值即为**几率 odds**：$\frac{y}{1-y}$，又由于我们取了对数，因此 $\ln\frac{y}{1-y}$ 称为**对数几率 （log odds， 亦称为 logit）**

$$\ln\frac{y}{1-y}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b \tag{5}$$

这个式子实际上是用**线性回归模型** w^T+b 预测的结果来逼近真实标记的**对数几率**， 因此此模型也被称为 对数几率回归 Logistic Regression （或 Logit Regression）， **虽然名字是回归，但实际是一种分类学习方法。**

对数几率回归的优点：

1. 直接对分类可能性建模，无需实现假设数据分布，避免了假设分布不准确带来的问题
2. 不仅预测出类别，还得到近似概率预测
3. 对数几率函数是任意阶可导的凸函数，有很好的数学性质，可利用很多数值优化算法求最优解

通过极大似然估计w和b

还是上面那个公式

$$\ln \frac{y}{1-y}=\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b \tag{6}$$

我们把 y 视为类后验概率，上式可重写为：

$$\ln \frac{p(y=1 | \boldsymbol{x})}{p(y=0 | \boldsymbol{x})}=\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b \tag{7}$$

$$p(y=1 | \boldsymbol{x})=\frac{e^{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b}}{1+e^{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b}} \tag{8}$$

$$p(y=0 | \boldsymbol{x})=\frac{1}{1+e^{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b}} \tag{9}$$

我们通过极大似然法（maximum likelihood method） 估计 w 和 b，给定数据集 $\left\{\left(\boldsymbol{x}_{i}, y_{i}\right)\right\}_{i=1}^{m}$，对率回归模型最大化**对数似然** $\ell(w, b)$

$$\ell(\boldsymbol{w}, b)=\sum_{i=1}^{m} \ln p\left(y_{i} | \boldsymbol{x}_{i} ; \boldsymbol{w}, b\right) \tag{10}$$

为了方便，把 w 和 b 堆叠一起标记为 $\boldsymbol{\beta}=\boldsymbol{(w;b)}$，令 $\hat{\boldsymbol{x}}=(\boldsymbol{x} ; 1)$ 则 $\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b=\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}$.

对数似然项 $p\left(y_{i} | \boldsymbol{x}_{i} ; \boldsymbol{w}, b\right)$ 写成：

$$p\left(y_{i} | \boldsymbol{x}_{i} ; \boldsymbol{w}, b\right)=y_{i} p_{1}\left(\hat{\boldsymbol{x}}_{i} ; \boldsymbol{\beta}\right)+\left(1-y_{i}\right) p_{0}\left(\hat{\boldsymbol{x}}_{i} ; \boldsymbol{\beta}\right) \tag{11}$$

又由于

$$p(y=1 | \boldsymbol{x})=\frac{e^{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b}}{1+e^{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b}} \tag{12}$$

$$p(y=0 | \boldsymbol{x})=\frac{1}{1+e^{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}+b}} \tag{13}$$

最大化$\ell(\boldsymbol{w}, b)=\sum_{i=1}^{m} \ln p\left(y_{i} | \boldsymbol{x}_{i} ; \boldsymbol{w}, b\right)$ 等价于最小化$\ell(\boldsymbol{w}, b)=\sum_{i=1}^{m} -\ln p\left(y_{i} | \boldsymbol{x}_{i} ; \boldsymbol{w}, b\right)$，即负对数

由于 y_i 只能取 $0,1$ 值，因此 $\ln p_0, \ln p_1$ 只能有一个生效，$-\ln p_0=1+e^{\beta^T \hat{x}_i}$，而 $-\ln p_1=1+e^{\beta^T \hat{x}_i}+1+\beta^T \hat{x}_i$，那么应该最小化下列 $\ell(\beta)$

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^{m}\left(-y_{i} \boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_{i}+\ln \left(1+e^{\boldsymbol{\beta}^{\mathrm{T}} \hat{\boldsymbol{x}}_{i}}\right)\right) \tag{14}$$

$\boldsymbol{\beta}=(\boldsymbol{w};b)$，这是关于 β 的高阶可导连续凸函数，根据凸优化理论，可以通过经典数值优化算法：**梯度下降法 gradient descent method**、**牛顿法 Newton method** 得到最优解，于是得到

$$\boldsymbol{\beta}^*=\underset{\boldsymbol{\beta}}{\arg\min}\,\ell(\boldsymbol{\beta}) \tag{15}$$