逻辑回归（对数几率回归）

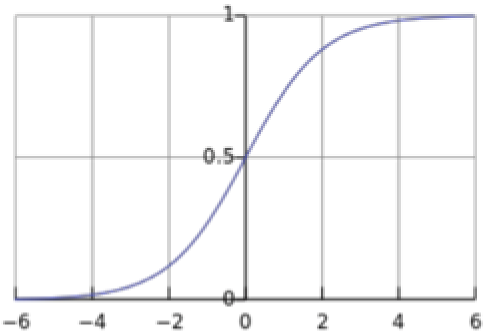
一、 逻辑回归

逻辑回归模型：$\log\left(\frac{p}{1-p}\right) = w_0 + \sum_{j=1}^{d} w_j x_j + u = \widetilde{W}^T \tilde{X}$ (1)

通过(1)式反解可以得到：$p = \frac{1}{1+e^{-\widetilde{W}^T \tilde{X}}}$ (2)

而（2）式是一种 sigmoid 函数，它将 x 值转化为一个接近 0 或 1 的 y 值：

$\varphi(x) = \frac{1}{1+e^{-x}}$ 其图像大致为



若将 p 视为样本 x 作为正例的可能性，则 1-p 是其反例的可能性，两者的比值$\frac{p}{1-p}$称为“几率”，反映了 p 作为正例的相对可能性。所以(1)式其实是在用线性回归模型的预测结果去逼近真实标记的对数几率。

由于 sigmoid 函数的取值在[0,1]之间，所以可以将其视为类 1 的后验概率估计 p(y=1|x)，所以我们把 sigmoid 函数计算得到的值大于等于 0.5 的归为类别 1，小于 0.5 的归为类别 0。

二、　最大似然估计

逻辑回归的第一个假设是：假设数据服从伯努利分布。第二个假设为假设模型的输出值是样本为正例的概率。所以整个模型可以描述为：

$$h_\theta(x;\theta) = p - \frac{1}{e^{-\theta^T x}} \quad \theta = (w;b)$$

若模仿线性回归的做法（利用误差平方和作为代价函数），则会发现得到得的结果是一个非凸函数，不利于求解，所以选取最大似然作为 cost function.

因为逻辑回归是一种训练学习，所以$h_\theta(x;\theta)$可以视为类 1 的后验概率，则：

$$h_\theta(x;\theta) = p(p=1|x;\theta) \quad = \varphi(w^T + b) = \varphi(z) = \frac{1}{1+ e^{-\theta^T x}}$$

$$p(p=0|x;w) = 1 - \varphi(z)$$

合并之后得到：

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

则极大似然估计函数为：

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

取对数之后：

$$L(\tilde{\mathbf{W}}) = \sum_{i=1}^{n}\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right)$$

$$= \sum_{i=1}^{n}\left(y_i \log\frac{p_i}{1-p_i} + \log(1-p_i)\right)$$

$$= \sum_{i=1}^{n}\left(y_i \tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{X}}_i - \log(1+e^{\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{X}}_i})\right)$$

$$\frac{\partial L(\tilde{\mathbf{W}})}{\partial \tilde{\mathbf{W}}} = \sum_{i=1}^{n}\left[\left(y_i - \frac{e^{\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{X}}_i}}{1+e^{\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{X}}_i}}\right)\tilde{\mathbf{X}}_i\right]$$

需要求得使$L(\widetilde{W})$最大的$\widetilde{W}$，在其前面加一个负号就变为最小化负对数似然函数：

$$C(\tilde{\mathbf{W}}) = -L(\tilde{\mathbf{W}}) = -\sum_{i=1}^{n}\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right)$$

所以，如果样本的类别为 1/0，估计值越接近 1/0 付出的代价越小，反之越大。

三、 梯度回归

因为梯度的负方向就是代价函数下降最快的方向，所以可以通过迭代求梯度更新权重来求解。

重复此公式：

$$\begin{aligned}\tilde{\mathbf{W}}_{new}^{(j)} &= \tilde{\mathbf{W}}^{(j)} - \eta \frac{\partial C(\tilde{\mathbf{W}})}{\partial \tilde{\mathbf{W}}^{(j)}} \\ &= \tilde{\mathbf{W}}^{(j)} - \eta \sum_{i=1}^{n}\left[\left(\frac{e^{\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{x}}_i}}{1+e^{\tilde{\mathbf{W}}^{\mathrm{T}}\tilde{\mathbf{x}}_i}} - y_i\right)\tilde{\mathbf{X}}_i^{(j)}\right]\end{aligned}$$

直到$\widetilde{W}$收敛