逻辑回归模型

对于一个二元分类问题，我们可以使用最小二乘法的回归模型来做：

$$y = w_0 + \sum_{j=1}^{d} w_j x_j + u$$

$$= \tilde{\mathbf{W}}^{\mathrm{T}} \tilde{\mathbf{X}}$$

如果基于上述模型预测的y值来作为样本属于某个类别的概率，那么预测的值会超出0到1的范围。为了解决预测概率超出0到1范围的问题，我们可以使用logit变换来解决上述问题：

$$\log\left(\frac{p}{1-p}\right) = w_0 + \sum_{j=1}^{d} w_j x_j + u$$

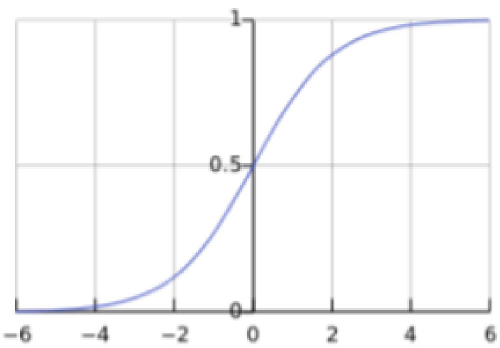
$$= \tilde{\mathbf{W}}^{\mathrm{T}} \tilde{\mathbf{X}}$$

上式实际上就是将y替换为$log(p/(1-p))$，其中

p是事件y发生的概率，$p/(1-p)$称为机率比或优势比。因此，只要通过求解上式的p，将它表示成y的表达式，就可以将回归模型转化为输出为0到1范围的逻辑回归模型。逻辑回归模型的定义如下：

$$p = \frac{1}{1 + e^{-w_0 - \sum_{j=1}^{d} w_j x_j}} = \frac{e^{w_0 + \sum_{j=1}^{d} w_j x_j}}{1 + e^{w_0 + \sum_{j=1}^{d} w_j x_j}}$$

$$= \frac{1}{1 + e^{-\tilde{\mathbf{W}}^{\mathrm{T}} \tilde{\mathbf{X}}}} = \frac{e^{\tilde{\mathbf{W}}^{\mathrm{T}} \tilde{\mathbf{X}}}}{1 + e^{\tilde{\mathbf{W}}^{\mathrm{T}} \tilde{\mathbf{X}}}}$$



- 如果 $w_0 + \sum_{j=1}^{d} w_j x_j = 0$，那么 $p = 0.5$
- 当 $w_0 + \sum_{j=1}^{d} w_j x_j$ 很大时，p 趋近于 1
- 当 $w_0 + \sum_{j=1}^{d} w_j x_j$ 很小时，p 趋近于 0

PLA ?

对于最小二乘法回归模型，使用了最小二乘的公式，直接得到了最终的模型。

对于逻辑回归模型，可以使用极大似然估计，配合梯度下降法，计算出最终模型。

那么，什么是极大似然估计？

首先，我们需要定义一个似然函数作为我们的目标函数，我们需要最大化这个目标函数。假设有n个样本，样本有类别1和0，p_i是样本属于类别1的概率，那么样本属于类别0的概率为 1 - p_i似然函数的定义为：

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

理解上述公式可以分两种情况来考虑，对于一个样本

1、假设该样本的类别为1，那么此时概率p_i 大于 1-p_i，上述公式只剩下$(p_i)_i^y$

2、假设该样本的类别为0，那么此时概率1-p_i 大于 p_i，上述公式只剩下$(1-p_i)^{(1-y_i)}$

那么上式怎样才能达到最大呢？显然，逻辑回归模型能决定预测样本属于哪个类，例如样本实际标签为1，那么预测的概率p_i的值为1；如果样本实际标签为0，那么预测的概率p_i的值为0。如果能准确预测，那么上述似然函数的值为1且为最大值。

我们实际处理过程中不好处理连乘符号，我们对它取对数，可得：

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

此时，我们的目标是最小化上式。

那么如何训练逻辑回归模型呢？

答案是梯度下降：

首先我们需要计算梯度向量，需要对代价函数求关于权值的梯度。

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

计算出梯度向量后，每次用梯度向量的反方向来更新权重，即重复下面的操作直至收敛：

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