NO.2 逻辑回归

逻辑回归，虽然名为回归，其实是一个分类模型，而且是线性分类模型，属于判别式模型。这里从二分类为例子，来说明逻辑回归的推导。

对于一个二分类问题，$\{(\mathrm{x_i}, \mathrm{t_i}), i=1,\dots,n\}$,其中$\mathrm{t_i} \in \{\mathrm{C_1}, \mathrm{C_2}\}$。为了实现分类的目的，对于新的 x，从概率的角度，如果我们知道了两个类别关于 x 的条件概率，那么我们就可以根据这两个概率做出分类的决策。因此我们只需要求到其中某一个类别的概率即可（另一个类别的概率等于 1 减去此类别的概率）。于是，我们要求：

$$\begin{aligned} p(\mathcal{C}_1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)+p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)} \\ &= \frac{1}{1+\exp(-a)} = \sigma(a) \end{aligned}$$

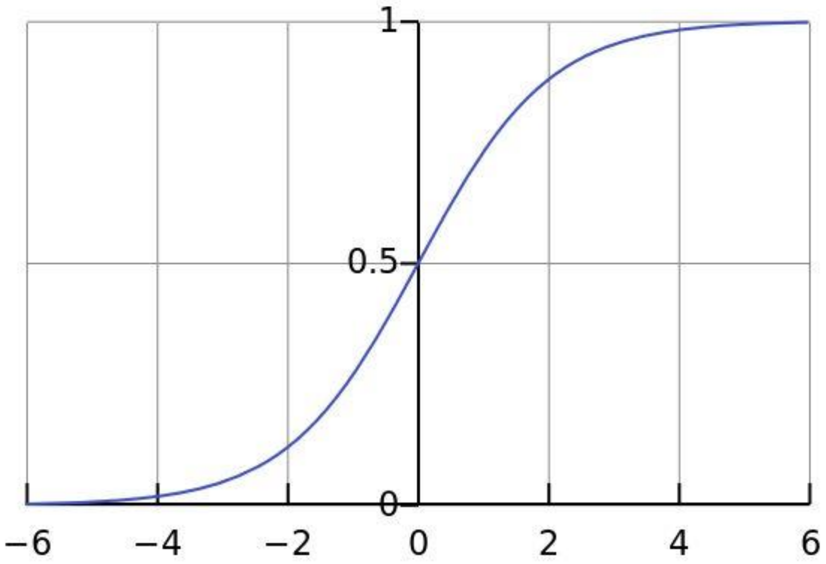
其中，

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

$\sigma(\mathrm{a})$是 logistic sigmoid 函数，定义为：

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

图形形状为：



可以看到，如果$\mathrm{a} = \ln\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)} = \ln\frac{p(C_1|x)}{p(C_2|x)} > 0$，即$\frac{p(C_1|x)}{p(C_2|x)} > 1$，那么$\sigma(\mathrm{a}) > 0.5$，因此我们可以以 0.5 为阈值，如果$\sigma(\mathrm{a}) > 0.5$，则判别为类别$\mathrm{C_1}$。这样的判别等价于如果

$\mathrm{p}(\mathrm{C_1}|x) > p(C_2|x)$，那么判别类别 $\mathrm{C_1}$。

一般情况下，我们会对输入向量做一个线性变换，变成特征向量$\mathrm{w^T}x$之后，再在完成上诉

步骤。逻辑回归的主要任务就是学习参数 w。
因此，逻辑回归的形式为

$$\mathrm{p}(\mathrm{C}_1|x) = y(x) = \sigma(w^T x)$$

注意：对于非线性问题，我们可以对输入向量做一个特征变换，得到特征向量$\phi(x)$,在特征空间上，分类问题变成线性可分的，逻辑回归的形式变为：

$$\mathrm{p}(\mathrm{C}_1|\phi) = y(\phi) = \sigma(w^T \phi)$$

最大似然法确定参数
在完成最大似然法的过程中，要用到对 logistic sigmoid 函数的导数：

$$\frac{d\sigma(a)}{da} = \sigma(1-\sigma)$$

对于一个数据集$\{(x_i, t_i), i = 1, \dots, n\}$，$\mathrm{t_i} \in \{0,1\}$，似然函数可以写成

$$\mathrm{p}(\mathbf{t}|\mathbf{w}) = \prod_{i=1}^{n} y_i^{t_i}\{1-y_i\}^{1-t_i}$$

其中$\mathrm{t} = (\mathrm{t}_1, \dots, t_n)$，$y_i = p(t_i = 1|x_i) = \sigma(\mathrm{w}^{\mathrm{T}} x_i)$
通过取似然函数的负对数，定义一个误差函数

$$\mathrm{E}(\mathbf{w}) = -\ln \mathrm{p}(\mathbf{t}|\mathbf{w}) = -\sum_{i=1}^{n}\{t_i \ln y_i + (1-t_i)\ln(1-y_i)\}$$

两侧关于 w 取误差函数的梯度，得到
$\nabla \mathrm{E}(\mathrm{w}) = \sum_{i=1}^{n}(y_i - t_i)x_i = \sum_{i=1}^{n}(\sigma(w^T x) - t_i)x_i$=$\mathrm{X}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{t})$
X 为设计矩阵。
没有解析解。
由于误差函数是凸函数（二次导大于 0），所以有唯一的最小值，因此采用牛顿-拉夫森迭代法：

$$\mathrm{w}^{\mathrm{new}} = w^{old} - H^{-1}\nabla E(w)$$

其中，H 为 Hessian 矩阵，其元素由$\mathrm{E}(\mathrm{w})$关于 w 的二阶导数构成。
于是有

$$\mathrm{w}^{\mathrm{new}} = w^{old} - (X^T R X)^{-1} X^T (\boldsymbol{y} - \boldsymbol{t})$$

其中 R 为对角阵，对角元素为$\mathrm{R_{nn}} = y_n(1-y_n)$

对于多分类问题，激活函数由 logistic sigmoid 函数变为 softmax 函数

$$\mathrm{p}(\mathrm{C_k}|x) = y_k(x) = \frac{exp(w_k^T x)}{\sum_j \exp(w_j^T x)}$$

其余步骤类似二分类情形。