逻辑回归

分类函数

我们知道，对于逻辑回归有sigmoid函数：

$$h_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}}$$

$$g(z) = \frac{1}{1 + e^{-z}}$$

但为什么选择这个分类函数呢？

因为一般的函数是这样的：

$$y = 0, \sum_{i=1}^{n} \omega_i x \leq b$$
$$y = 1, \sum_{i=1}^{n} \omega_i x > b$$

我们可以看到，这是基于阈值的分类形式，会出现从0到1的突变，在数学上缺乏连续性。

而对于sigmoid函数来说有两个优势：

- 它的输入范围是−∞ → +∞，而输出刚好为（0，1），正好满足概率分布为（0，1）的要求。我们用概率去描述分类器，自然比单纯的某个阈值要方便很多
- 它是一个单调上升的函数，具有良好的连续性，不存在不连续点。

这个函数求导的时候会有有趣的现象：

$$\begin{aligned} g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= -\frac{1}{(1+e^{-z})^2} \cdot e^{-z} \cdot (-1) \\ &= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\ &= \frac{1}{(1+e^{-z})} \cdot \left(1 - \frac{1}{(1+e^{-z})}\right) \\ &= \left(\frac{1}{1+e^{-z}}\right) \cdot \left(\frac{1+e^{-z}-1}{1+e^{-z}}\right) \\ &= g(z)(1-g(z)) \end{aligned}$$

最大似然估计

那么，现在有逻辑回归模型了，咱们怎么去拟合一个合适的 θ 呢？我们之前已经看到了在一系列假设的前提下，最小二乘法回归可以通过最大似然估计来推出，那么接下来就给我们的这个分类模型做一系列的统计学假设，然后再用最大似然法来拟合参数。

首先我们可以假设：

$$P(y = 1|x;\theta) = h_\theta(x)$$
$$P(y = 0|x;\theta) = 1 - h_\theta(x)$$

这也可以简写为：

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

假设 m 个训练样本都是独立同分布的，所以可以按如下的方式来写参数的似然函数：

$$
\begin{aligned}
L(\theta) &= p(\vec{y}|X;\theta) \\
&= \prod_{i=1}^{m} p\left(y^{(i)}|x^{(i)};\theta\right) \\
&= \prod_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right)\right)^{y^{(i)}} \left(1-h_\theta\left(x^{(i)}\right)\right)^{1-y^{(i)}}
\end{aligned}
$$

显然我们可以用取对数的方法求最大值：

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{m} y^{(i)} \log h\left(x^{(i)}\right) + \left(1-y^{(i)}\right) \log\left(1-h\left(x^{(i)}\right)\right)
\end{aligned}
$$

梯度上升

怎么让似然函数最大？就跟之前咱们在线性回归的时候用了求导数的方法类似，咱们这次就是用**梯度上升法（gradient ascent）**。还是写成向量的形式，然后进行更新，也就是$\theta:=\theta+\alpha\nabla_\theta l(\theta)$。(注意更新方程中用的是加号而不是减号，因为我们现在是在找一个函数的最大值，而不是找最小值了。)还是先从只有一组训练样本(x,y)来开始，然后求导数来推出随机梯度上升规则：

$$
\begin{aligned}
\frac{\partial}{\partial\theta_j} l(\theta) &= \left(y\frac{1}{g\left(\theta^T x\right)} - (1-y)\frac{1}{1-g\left(\theta^T x\right)}\right) \frac{\partial}{\partial\theta_j} g\left(\theta^T x\right) \\
&= \left(y\frac{1}{g\left(\theta^T x\right)} - (1-y)\frac{1}{1-g\left(\theta^T x\right)}\right) g\left(\theta^T x\right)\left(1-g\left(\theta^T x\right)\right) \frac{\partial}{\partial\theta_j}\theta^T x \\
&= \left(y\left(1-g\left(\theta^T x\right)\right) - (1-y) g\left(\theta^T x\right)\right) x_j \\
&= \left(y - h_\theta(x)\right) x_j
\end{aligned}
$$

所以我们就得到了随机梯度上升的公式：

$$
\theta_j := \theta_j + \alpha\left(y^{(i)} - h_\theta\left(x^{(i)}\right)\right) x_j^{(i)}
$$

我们可以发现逻辑回归的随机梯度上升和线性回归的梯度下降十分相似，但这里要注意的是，两个回归的基础原函数不一样，逻辑回归是sigmoid函数，而线性回归则是线性方程。