逻辑斯特回归

日期： 2019-3-18

作者： lunyang liu

1. 背景介绍

Logistic回归由统计学家David Cox在1958年提出。在统计学中logistic模型是一种广泛使用的统计模型，它使用logistic函数来对二元取值的因变量建模。在回归分析中，logistic（或者logit）回归估计logistic模型的参数，它是二项式回归的一种形式 。 从数学上来说，二元logistic模型的因变量有两种取值，例如通过/失败，赢/输，活/死等，这类取值通常由指示变量表示，将它们的值分别记作：0或1。在logistic模型中，标记为“1”的值的**对数几率（几率的对数）是一个或多个自变量（“预测值”）的线性组合**;自变量可以是二元变量（两类，由指示变量编码）或连续变量（任何实数）。标记为“1”的值的相应概率可以在0和1之间变化。将log-odds转换为概率的函数是逻辑函数，由此得名。这一点很像支持向量机和神经网络中的连接函数或者激活函数。

2. Logistic模型

2.1 logistic 分布

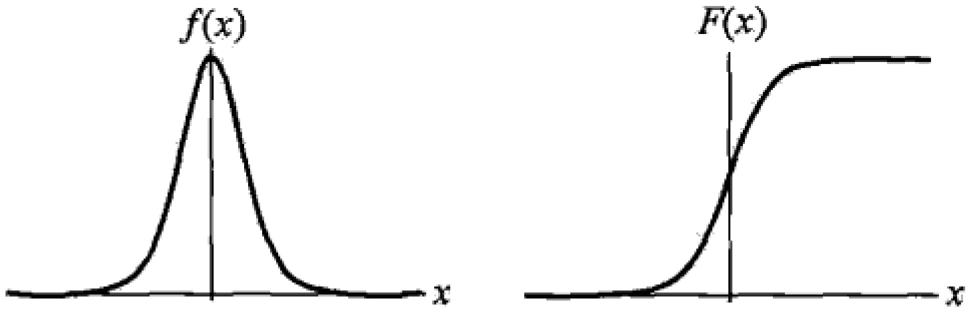
设X是连续随机变量，X服从logistic分布，即X具有如下分布函数和密度函数：

$$F(x) = P(X \leq x) = \frac{1}{1 + e^{-(x-\mu)/\gamma}} \tag{1}$$

$$f(x) = F'(x) = \frac{e^{-(x-\mu)/\gamma}}{\gamma(1 + e^{-(x-\mu)/\gamma})^2} \tag{2}$$

式中，μ为位置参数，γ 为形状参数。

logistic密度函数($f(x)$)及分布函数($F(x)$)如下图所示：



分布函数属于logistic函数，图形是一个S型曲线(sigmoid curve)，该曲线以点$(\mu, \frac{1}{2})$为中心对称，即满足：

$$F(-x + \mu) - 1/2 = -F(x - \mu) + 1/2 \tag{3}$$

曲线在中心附近增长较快，在两端增长较慢，形状参数γ越小，曲线在中心附近增长越快。

2.2 logistic分类模型

从线性回归模型出发转到logistic模型想法是很直接的。在线性回归中，输出和特征之间的关系我们使用线性等式进行建模：

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)} \tag{4}$$

式子(4)的输出是没有限定取值范围的，对于我们的想要的分类任务而言，期望的是输出值范围为$[0,1]$，也就是概率取值，结合前面提到的式子(1)，如果我们将(4)的输出作为自变量喂给logistic函数，那么其值被强制落于$[0,1]$，即：

$$P(y^{(i)} = 1) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))} \tag{5}$$

因此，logistic模型可以定义如下，二项logistic回归模型是如下的条件概率分布：

$$\begin{aligned} P(Y = 1|x) &= \frac{\exp(\beta . x + b)}{1 + \exp(\beta . x + b)} \\ P(Y = 0|x) &= \frac{\exp(\beta . x + b)}{1 + \exp(\beta . x + b)} \end{aligned} \tag{6}$$

这里，$x \in R^n$是输入，$Y \in \{0, 1\}$是输出，$\beta \in R^n$和$b \in R$是未知参数，β称作权重向量，b称为偏置，$\beta \cdot x$为β和x的内积。

对于给定的输入实例x，按照(6)，(7)可以求得$P(Y = 1|x)$和$P(Y = 0|x)$.Logistic回归比较两个条件概率值的大小，将实例x分到概率值较大的那一类.

2.3 进一步理解

首先我们了解下**几率**。一个事件的几率(odd)是指该事件发生的几率与不发生的概率的比值。 如果事件发生的概率是p，那么该事件的几率是$\frac{p}{1-p}$，该事件的对数几率为（log-odd）或logit函数是：

$$logit(p) = log\frac{p}{1 - p} \tag{7}$$

那么，神奇的一幕来了，对于logistic模型而言，我们将(6), (7)带入(8)，那么得到：

$$log\frac{P(Y = 1|x)}{1 - P(Y = 1|x)} = \beta \cdot x \tag{8}$$

这说明，在logistic回归模型中，输出$Y = 1$的对数几率是输入x的线性函数.或者说，输出$Y = 1$的对数几率是由输入x的线性函数表示的模型，即logistic回归模型。

通过对数几率函数，还能增强logistic回归模型的可解释性。我们知道，logistic回归模型的输出在$[0, 1]$，因此，对权重参数的解释并不能和线性回归模型一样。权重对概率的影响并不是线性的。权重加和通过logistic函数转换为概率了，然而，通过logit函数，线性关系又出来了，见公式(9). 可是，这有如何，看起来没什么用啊？不然.下面我们通过一系列的变化，你就能看出一个特征x_i变化一个单元如何影响输出。首先，对公式(9)两边取指数变换：

$$\frac{P(y = 1)}{1 - P(y = 1)} = odds = exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) \tag{9}$$

接下来我们对其中的一个特征加1，然后对两个模型作几率的比值：

$$\frac{odds_{x_j+1}}{odds} = \frac{exp(\beta_0 + \beta_1 x_1 + ... + \beta_j(x_j+1) + ... + \beta_p x_p)}{exp(\beta_0 + \beta_1 x_1 ... + \beta_p x_p)}$$

约简一下：

$$\frac{odds_{x_j+1}}{odds} = \frac{exp(\beta_j x_j + 1)}{exp(\beta_j x_j)} = exp(\beta_j)$$

竟然简单至此，最后只剩一个特征的**权重的指数幂**！ 如果一个特征变化一个单位，那么这个变化前后的**几率**比值会是$exp(\beta_j)$. 举个例子，如果几率为2，也就是说$y=1$的概率是$y=0$的概率的2倍. 假设我们取一个权重为0.7的特征，增加其一个单位，那么新的几率变为$exp(0.7)*2$，结果大约为4.但是呢，实际上我们并不会这样去解释特征的变化，这样做只不过为了加强理解而已。

2.4 参数估计

考虑一个参数为β广义线性回归模型:

$$h_\beta(X) = \frac{1}{1+e^{-\beta^T X}} = P(Y=1|X;\beta) \tag{11}$$

因为$Y \in \{0,1\}$,所以：

$$P(Y=0|X;\beta) = 1 - h_\beta(X) \tag{12}$$

这样对于$P(y|X;\beta)$来说，通过幂技巧可以这样来表示：

$$P(y|X;\beta) = h_\beta(X)^y (1-h_\beta(X))^{1-y} \tag{13}$$

式子(14)就是我们的模型来预测Y取1还是0，我们假设所有的样本都是来自于伯努利分布的独立样本，那么所有样本的似然函数为：

$$\begin{aligned} L(\theta|x) &= P(Y|X;\beta) \\ &= \prod_i P(y_i|x_i;\beta) \\ &= \prod_i h_\beta(x_i)^{y_i}(1-h_\theta(x_i))^{(1-y_i)} \end{aligned}$$

对数似然函数为：

$$\begin{aligned} log(L(\beta|x)) &= \sum_i [y_i log(h_\beta(x_i)) + (1-y_i) log(1-h_\beta(x_i))] \\ &= \sum_i [y_i log\frac{h_\beta(x_i)}{1-h_\beta(x_i)} + log(1-h_\beta(x_i))] \\ &= \sum_i [y_i(\beta.x_i) - log(1+e^{\beta x_i})] \end{aligned}$$

求$log(L(\beta|x))$的最大值，得到参数β.一般还会除以样本总数，然后取负，求解$-\frac{log(\beta|x)}{N}$的最小值，两者其实是等价的。其中N为样本总数.

2.5 梯度求解公式推导

先根据公式(16)写出代价函数：

$$J(\beta) = -\frac{1}{N}\sum_{i}[y_i\beta x_i - log(1+e^{\beta x_i})] \tag{14}$$

下面要做的就是$J(\beta)$对β求偏导，其实最终也就是对$y_i\beta x_i$和$log(1+e^{\beta x_i})$求偏导：

$$\frac{\partial y_i\beta x_i}{\partial\beta} = y_i x_i$$
$$\frac{log(1+e^{\beta x_i})}{\partial\beta} = \frac{x_i e^{\beta x_i}}{1+e^{\beta x_i}} = x_i h_\beta(x_i)$$

整理可得代价函数的更新梯度为：

$$\frac{J(\beta)}{\partial\beta} = \frac{1}{N}\sum_{i}(h_\beta(x_i) - y_i)x_i \tag{15}$$

2.6 损失函数选择讨论

这里借用李宏毅老师的两张片子进行解释：

Logistic Regression + Square Error

为什么不能用平方差？

Step 1: $f_{w,b}(x) = \sigma\left(\sum_i w_i x_i + b\right)$

Step 2: Training data: $(x^n, \hat{y}^n)$, $\hat{y}^n$: 1 for class 1, 0 for class 2

$$L(f) = \frac{1}{2}\sum_n \left(f_{w,b}(x^n) - \hat{y}^n\right)^2$$

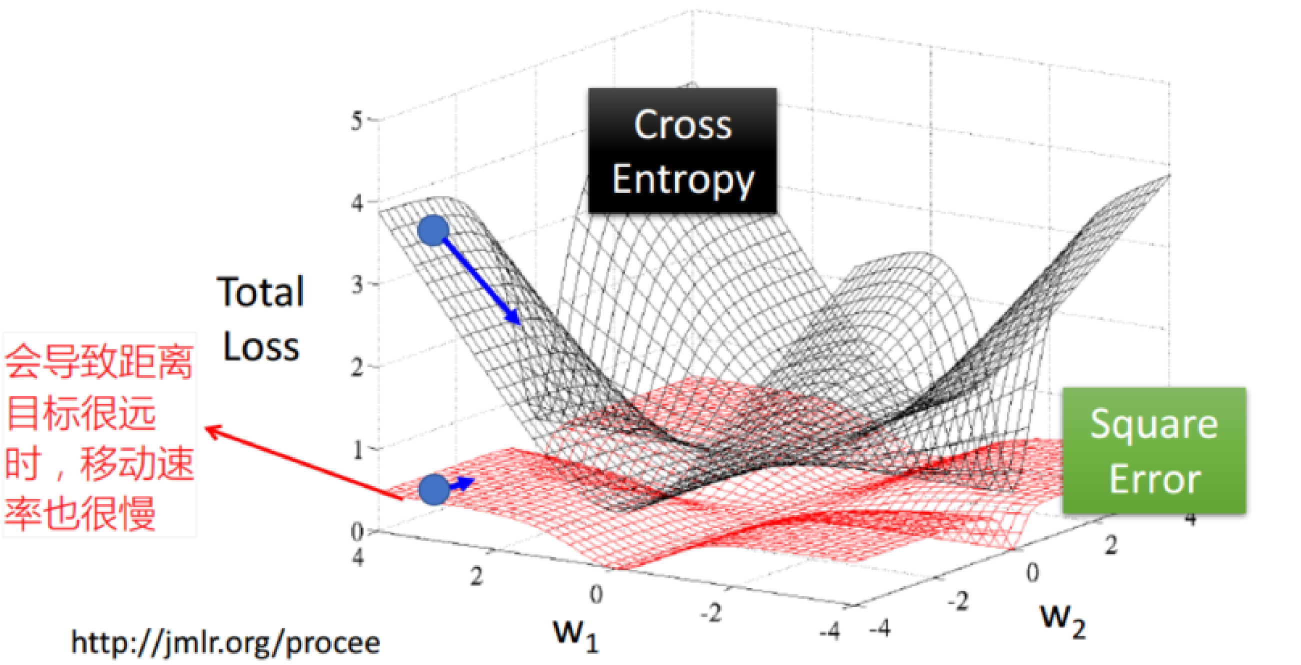
Step 3:

$$\frac{\partial \left(f_{w,b}(x) - \hat{y}\right)^2}{\partial w_i} = 2\left(f_{w,b}(x) - \hat{y}\right)\frac{\partial f_{w,b}(x)}{\partial z}\frac{\partial z}{\partial w_i}$$

$$= 2\left(f_{w,b}(x) - \hat{y}\right) f_{w,b}(x)\left(1 - f_{w,b}(x)\right) x_i$$

$\hat{y}^n = 0$ If $f_{w,b}(x^n) = 1$ (far from target) ➡ $\partial L / \partial w_i = 0$

如下图表示 If $f_{w,b}(x^n) = 0$ (close to target) ➡ $\partial L / \partial w_i = 0$



http://jmlr.org/proceedings/papers/v9/glorot10a/glorot10a.pdf

3. Logistic算法优缺点

线性回归模型的许多优缺点同业适用于logistic回归模型，logistic模型的一大缺点是可解释性差，对于权重的解释是基于乘法的，而不是可加的。另外它还可能受困于**完全可分**问题，如果有一个特征就能完美的区分两个类别，那么logistic模型可能就无法训练了，因为该特征的权重不会收敛，可能趋于无限大。但是真有这样**独秀**的特征也就不需要机器学习，使用专家经验就直接搞定了。如果出现这样的问题可以考虑添加惩罚项，或者定义一个权重的先验概率分布.

Logistic模型的优点是它不仅是一个分类模型，还能给出概率值，相对于其他只能给出分类结果的模型来说这简直是太棒了，因为对于一个样本来说0.99的概率与0.51的概率还是存在很大差别的.

Logistic 回归模型也可以从二元分类扩展到多类别分类，叫做**多项回归模型**(Multinomial Regression).