**Logistic 回归**

**一、背景**

Logistic回归来源于线性回归，线性回归是用一个线性函数去拟合一些点，如果我们想进行分类的话，那就只需要找一个单调可微的函数，将分类任务的真实标记和回归模型的预测值联系起来。

对于二分类任务，其输出标记是0和1，而回归模型的预测值z = wTx + b是实数，于是我们需要一个函数将实数转换为0/1值。最简单的方法，就是使用阶跃函数，大于0判为1，小于0判为0。但是单位阶跃函数不连续，我们希望有一个近似单位阶跃的函数，而且单调可微，对数几率函数就是这样一个常用的替代函数，即Logistic函数。

**Logistic回归**与**多重线性回归**实际上有很多相同之处，最大的区别就在于它们的**因变量不同**，其他的基本都差不多。正是因为如此，这两种回归可以归于同一个家族，即广义线性模型（generalizedlinear model）。

这一家族中的模型形式基本上都差不多，**不同的就是因变量不同**。

* 如果是连续的，就是**多重线性回归**；
* 如果是二项分布，就是**Logistic回归**；
* 如果是Poisson分布，就是**Poisson回归**；
* 如果是负二项分布，就是**负二项回归**。

Logistic回归的因变量可以是二分类的，也可以是多分类的，但是二分类的更为常用，也更加容易解释。所以实际中最常用的就是二分类的Logistic回归。

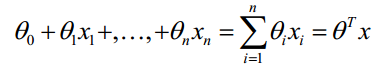
**二、模型**

**1、构造预测函数h**

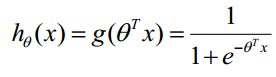
Logistic回归虽然名字里带“回归”，但是它实际上是一种分类方法，主要用于两分类问题（即输出只有两种，分别代表两个类别），**所以利用了Logistic函数（或称为Sigmoid函数）**，函数形式为：

https://img-blog.csdn.net/20140716153623240

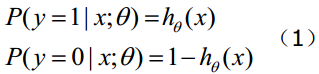
对于线性边界的情况，边界形式如下：



构造预测函数为：

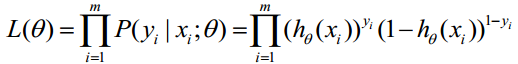


函数https://img-blog.csdn.net/20140716153711397的值有特殊的含义，它表示结果取1的概率，因此对于输入x分类结果为类别1和类别0的概率分别为：



**2、构造似然函数**

取似然函数如下：



对数似然函数：

http://img.my.csdn.net/uploads/201407/16/1405496926_1023.png

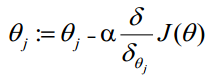
最大似然估计就是求使http://img.my.csdn.net/uploads/201407/16/1405496927_9629.png取最大值时的参数θ，这里实际上可以使用梯度上升法来求解最大值，但是一般都取负数然后用梯度下降法估计http://img.my.csdn.net/uploads/201407/16/1405496928_3363.png的最小值来求解：

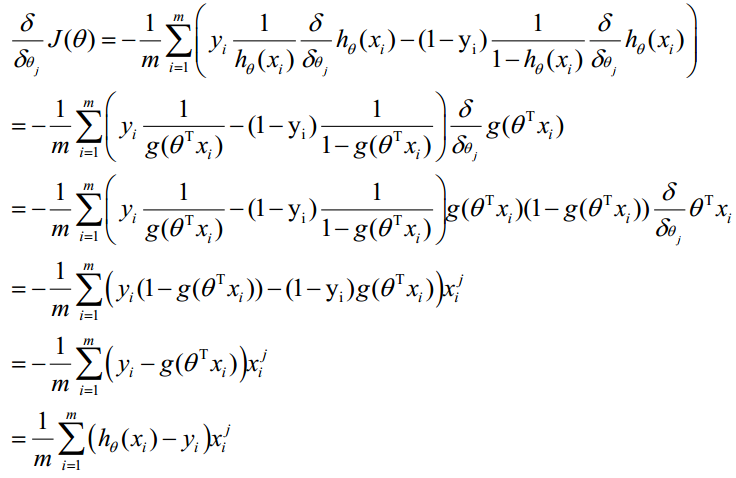
http://img.my.csdn.net/uploads/201407/16/1405496967_1305.png

所以实际上可以把J看成Logistic回归模型的损失函数。一般来说，在实际应用中，logistic会配合正则化使用，特别是L1正则化。

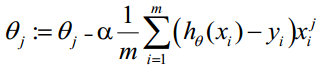
**3、梯度下降法求的最小值**

θ更新过程：





θ更新过程可以写成：



**4、求解过程**

除了可以用上述的梯度下降法来求解，一般还可以用牛顿法、拟牛顿法等来进行求解。

**三、应用**

**Logistic回归的主要用途**：

* 寻找危险因素：寻找某一疾病的危险因素等；
* 预测：根据模型，预测在不同的自变量情况下，发生某病或某种情况的概率有多大；
* 判别：实际上跟预测有些类似，也是根据模型，判断某人属于某病或属于某种情况的概率有多大，也就是看一下这个人有多大的可能性是属于某病。

Logistic回归主要在流行病学中应用较多，比较常用的情形是探索某疾病的危险因素，根据危险因素预测某疾病发生的概率，等等。例如，想探讨胃癌发生的危险因素，可以选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群肯定有不同的体征和生活方式等。这里的因变量就是是否胃癌，即是或否，自变量就可以包括很多了，例如年龄、性别、饮食习惯、幽门螺杆菌感染等。自变量既可以是连续的，也可以是分类的。