原 Logistic Regression逻辑回归的简单解释

2016年08月26日 11:45:15 阿拉丁吃米粉 阅读数：15078

Logistic Regression也叫Logit Regression，在机器学习中属于参数估计的模型。逻辑回归与普通线性回归（Linear Regression）有很大的关系。在应
有所区别：

- 普通线性回归主要用于连续变量的**预测**，即，线性回归的输出y的取值范围是整个实数区间（$y \in R$）
- 逻辑回归用于离散变量的**分类**，即，它的输出y的取值范围是一个离散的集合，主要用于类的判别，而且其输出值y表示属于某一类的概率

一个单独的逻辑回归函数只能判别两个类，这里用0和1表示. 逻辑回归的结果会给出一个概率p，表示属于类别1的概率。既然是两类问题，那么属于类
然就是$1-p$。有没有发现一点二项分布（Binomial Distribution）的影子?

逻辑回归应用广泛，而且因为给出的结果是一个概率，比单纯的“是”或“不是”包含更多的信息，因此大受人们喜爱（误）。我们之前参加Kaggle广
测竞赛时使用的就是逻辑回归。因为用户要么点了广告，要么没点，我们给出一个概率，就可以判断用户的点击广告的可能性。这个预测看起来很简单
型很简单的，难的地方在于features的分析，选取，综合等，也就是常说的pre-processing。

文章内容

很多文章介绍逻辑回归时会直接给出一个叫sigmoid的函数，该函数的值域范围是$(0,1)$，刚好是概率的取值范围（也不完全是，因为是开区间）。本文
前一点点，从引入sigmoid函数之前介绍一下Logistic Regression。文章只做简单介绍（真的很简单），不涉及参数估计的内容。

Odds与Logit函数

逻辑回归的输入是一个线性组合，与线性回归一样，但输出变成了概率。而且逻辑回归用于预测两类问题，类似一个伯努利试验。假设在一个伯努利试
概率是p，失败的概率是$1-p$，我们设逻辑回归的输出是成功的概率p，那么需要一个函数将逻辑回归的输入（一个线性组合）与p联系起来。下面介
它的名字叫Logit.

我们定义：

$$Odds = \frac{p}{1-p}$$

上式很直观，表示成功的概率是失败概率的多少倍，中文叫做**发生比**。

在赌博中，发生比大概描述了赢的概率是输的概率的几倍。

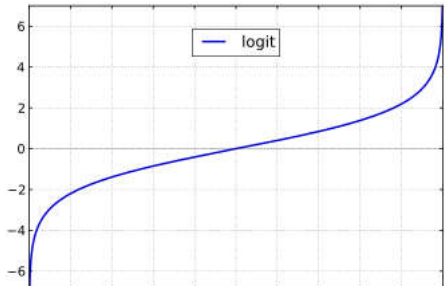
对Odds取自然对数：

$$\ln(Odds) = \ln(\frac{p}{1-p}) = \ln(p) - \ln(1-p)$$

上式即为logit函数的定义，参数为p，记为：

$$logit(p) = \ln(Odds)$$

$logit(p)$的图像如下所示，可以看到它的定义域是$[0,1]$，值域是R。



但我们要的是定义域是R，值域是$[0,1]$。于是我们求(3)式的反函数，并将参数p用另一个参数α表示，有：

$$logit^{-1}(\alpha) = \frac{1}{1+e^{-\alpha}} = \frac{e^{\alpha}}{1+e^{\alpha}}$$

上式中α可以取全体实数，而该函数的值域变成了$(0,1)$，这正是我们想要的效果。$logit(p)$的反函数$logit^{-1}(\alpha)$的名称就是我们常常听到的**sigmoid**状像字母S。

sigmoid由sigma和后缀-oid合成而来。sigma即希腊文第十八个字母σ，通常用来指代S，后缀-oid表示『像……的东西』，因此sigmoid函数实际上是命名，表示一个像S型的函数。

输入与输出

在(4)式中，输入的参数α可以是任何数，也可以将其作为一个**线性组合**输入。例如，另

$$\alpha = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

则(4)式的sigmoid函数可以写成：

$$sigmoid(\alpha) = logit^{-1}(\alpha) = \frac{e^{\theta_0+\theta_1 x_1+\theta_2 x_2}}{1+e^{\theta_0+\theta_1 x_1+\theta_2 x_2}}$$

上式就是逻辑回归的一般用法。注意到它的输入还是一个线性组合，跟线性回归的输入是一样的，只不过计算的时候比线性回归多了一层函数，因此这有文章说**逻辑回归的本质还是线性回归**，也会看到有一些文章说**在特征到结果的映射中多加了一层函数映射**，这个函数映射就是sigmoid。

(5)式是计算概率p的表达式，这个表达式也可以从$logit$函数来推导。**因为$logit(p)$与一个线性组合是等价的**（也再一次说明逻辑回归的本质还是线性回

令$logit$函数等于一个线性组合（这是可以的，因为$logit$函数的定义域和值域与一个线性组合的定义域和值域是一样的），即：

$$logit(p) = \ln(\frac{p}{1-p}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

对上式两边取自然底数，有：

$$\frac{p}{1-p} = e^{\theta_0+\theta_1 x_1+\theta_2 x_2}$$

$$\Rightarrow p = \frac{e^{\theta_0+\theta_1 x_1+\theta_2 x_2}}{1+e^{\theta_0+\theta_1 x_1+\theta_2 x_2}}$$

通常会将上式写成

$$\hat{p} = \frac{e^{\theta_0+\theta_1 x_1+\theta_2 x_2}}{1+e^{\theta_0+\theta_1 x_1+\theta_2 x_2}}$$

$\hat{p}$表示p的估计值。

上式就是(5)式。这样求概率p便变成了参数估计问题：估计参数θ，使得$\hat{p}$最接近p。

虽然逻辑回归通常用于两个类的判别问题，但是将多个逻辑回归函数组合起来就可以解决多类判别的问题。

Refrence

Youtube上有一个关于Logistic Regression的视频的入门级系列介绍，本文就是根据这个系列的介绍写的。想对Logistic Regression有快速的了解可以列视频（可惜要翻墙，QQ）
https://www.youtube.com/watch?v=zAULhNrnuL4