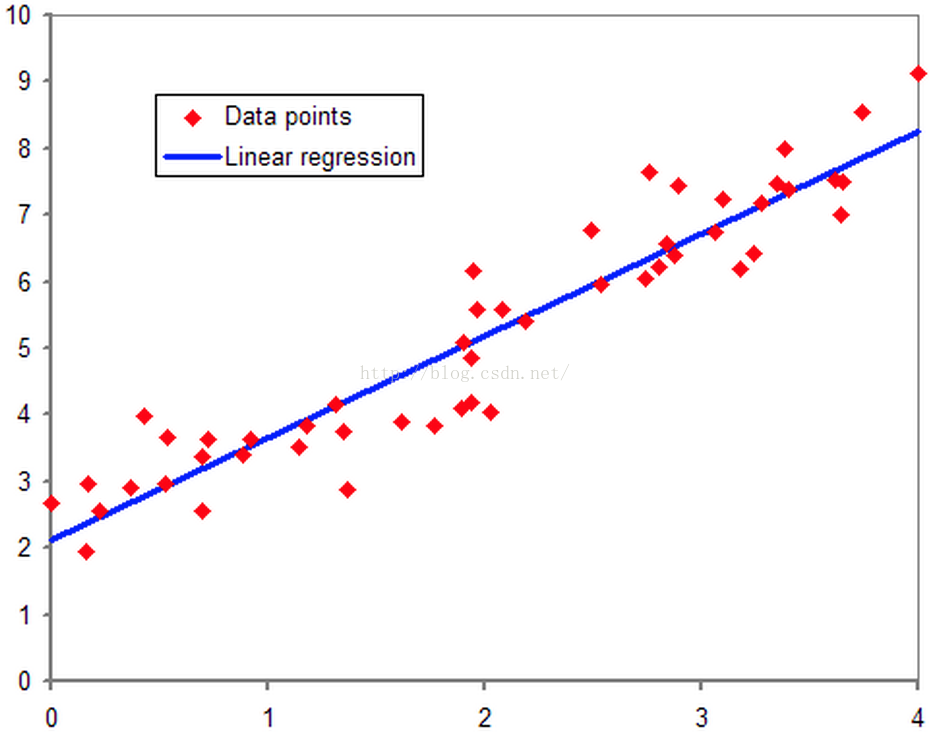
Logistics回归

**1. 综述**

logistics回归是应用非常广泛的一个分类机器学习算法，它将数据拟合到一个logit函数(或者叫做logistic函数)中，从而能够完成对事件发生的概率进行预测。Logistics回归是离散选择法模型之一，它是最早的离散模型，也是目前应用最广泛的模型。广泛应用于社会学、生物统计学、临床、数量心理学、计量经济学、市场营销等统计实证分析

**2. logistics回归的由来**

我们都学过线性回归，这是一种对多维空间中存在的样本点，用特征的线性组合去拟合空间中点的分布和轨迹的方法。如下图：



线性回归虽然可以对连续值结果进行预测，但是现实生活中更常见的问题是分类问题。最简单的情况是：是与否的二分类问题。比如说医生需要判断病人是否生病，银行要判断一个人的信用程度是否达到可以给他发信用卡的程度等等。

遇到这类问题的时候我们最直接的想法是，既然能够用线性回归预测出连续值结果，那根据结果设定一个阈值限定一下是不是就可以对这类问题进行求解呢？对于数据非常标准的时候也可以解决，但是现实中我们需要用来学习的数据并不都是那么精准的。

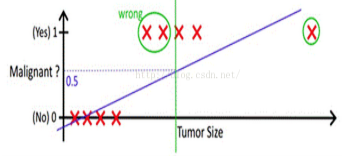
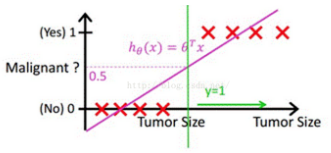
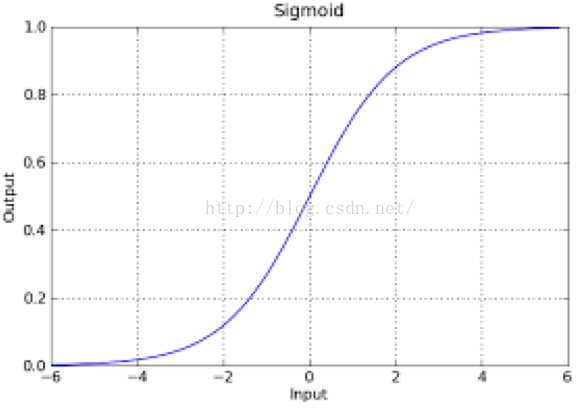


图1 图2

如上图所示，图1阈值作用下可以很好地预测数据，但是图2中在阈值作用下还是有数据不能被回归模型描述出来，那么这个阈值就是失效的。这只是一个很简单的例子，现实中得到的数据会比这个复杂许多倍，因此线性回归模型就不能用了，实现不了我们想要的分类效果。当线性回归的结果输出是一个连续值，而值的范围是无法限定的，人们寻求把这个结果值映射为可以帮助我们判断的结果，因此逻辑回归就诞生了。为了能更好的做回归，人们假设输出结果是 (0,1) 的一个概率值，这个问题就很清楚了。于是找到了数学上的一个简单函数了， sigmoid函数，这个函数的特性表示如下：

sigmoid函数的图如下：



从函数图上可以看出，函数y=g(z)在z=0的时候取值为1/2，而随着z逐渐变小，函数值趋于0，z逐渐变大的同时函数值逐渐趋于1，而这正是一个概率的范围。

**3. logistics回归模型**

**LR分类器**

Logistic回归为概率型非线性回归模型，是研究二分类观察结果y与一些影响因素之间关系的一种多变量分析方法。通常的问题是，研究某些因素条件下某个结果是否发生。

我们先从LR分类器说起。LR分类器，即Logistic Regression Classifier。在分类情形下，经过学习后的LR分类器是一组权值 ，当测试样本的数据输入时，这组权值与测试数据按照线性加和得到：

 （1）

这里是每个样本的n个特征。

之后按照sigmoid函数的形式求出

 （2）

由于sigmoid函数的定义域为，值域为，因此最基本的LR分类器适合对两类目标进行分类。所以Logistic回归最关键的问题就是研究如何求得这组权值。这个问题是用极大似然估计来做的。

**回归模型**

考虑具有n个独立变量的向量，设条件慨率为根据观测量相对于某事件x发生的概率。那么Logistic回归模型可以表示为：

 (3)

这里称为Logistic函数。其中。

那么在x条件下y不发生的概率为:

 (4)

所以发生与不发生的概率之比为：

 （5）

这个比值称为事件的发生比，简记为odds。对odds取对数得到：

 (6)

可以看出Logistic回归都是围绕一个Logistic函数来展开的。

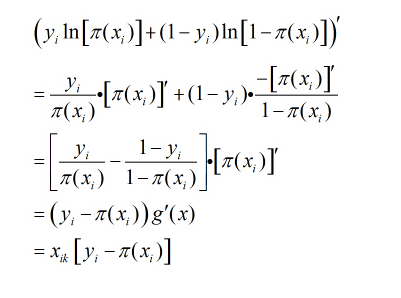
接下来就讲如何用极大似然估计求分类器的参数：假设有m个观测样本，观测值分别为，设为给定条件下得到的概率，同样地，的概率为，所以得到一个观测值的概率为。

因为各个观测样本之间相互独立，那么它们的联合分布为各边缘分布的乘积。得到似然函数为

 (7)

然后我们的目标是求出使这一似然函数的值最大的参数估计，最大似然估计就是求出参数，使得取得最大值，对函数取对数得到:

 (8)

继续对这个分别求偏导，得到个方程，比如现在对参数求偏导，由于：

得到：

 （10）

这样的方程一共有n+1个，所以现在的问题转化为解这n+1个方程形成的方程组。上述方程比较复杂，我们用牛顿迭代方法求解。

**4. logistics回归实现**

下面是对一组x和y进行回归的python的例子，行数=100偏置=15方差=5，迭代次数为100000次

得到的图如下所示：

