# 基于逻辑回归模型的糖尿病分析和预测

彭仁志

**摘要：**  
糖尿病是一组以高血糖为特征的代谢性疾病。. 高血糖则是由于胰岛素分泌缺陷或其生物作用受损，或两者兼有引起。长期存在的高血糖，导致各种组织，特别是眼、肾、心脏、血管、神经的慢性损害、功能障碍该数据集最初来自国家糖尿病、消化和肾脏疾病研究所。数据集的目的是根据数据集中包含的某些诊断测量值，诊断性地预测患者是否患有糖尿病。对从较大的数据库中选择这些实例。施加了若干限制。特别是，这里的所有患者都是至少21岁的皮马印第安人后裔女性文章利用逻辑回归方法在MATLAB环境下根据收集到的一些致病因素建立糖尿病预测模型，计算出其准确率为79%。本文结果表明，该模型在预测糖尿病风险的应用上具有一定价值。

**关键词：**

逻辑回归;糖尿病;梯度下降;预测;MATLAB2016a；

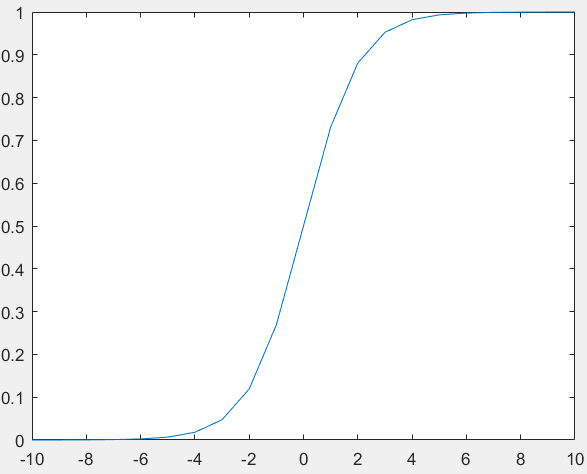
# 1、逻辑回归算法(Logistic Regression)

## （1）代价函数及其梯度(Cost function and gradient)

首先来看要拟合的函数，下面是我们之前拟合的函数的形式，很明显拟合结果为一个数值。



我们想要将一个范围(−∞,∞)的数值映射到(是/否)(1/0)我们非常自然的能够想到sigmoid函数，这是一个能将数值从(−∞,∞)映射到(0,1)的函数：g(z)=1/(1+e^(-z))



代价函数我们并没用平常的平方损失函数，而是用了如下的一个形式：



y=1时，我们考察的是hθ(x)与1的接近程度，越接近1，−log(hθ(x))就越趋近于0，y=0时，我们考察的是1−hθ(x)与1的接近程度，越接近1，−log(1−hθ(x))就越趋近于0，总而言之，这是一个越接近真实值，计算值越接近0，越远离真实值，计算值越接近正无穷的非常巧妙的函数。我们要求代价函数的极小值点，当然还是要让偏导等于0，这里我们对J(θ)求一下偏导得：



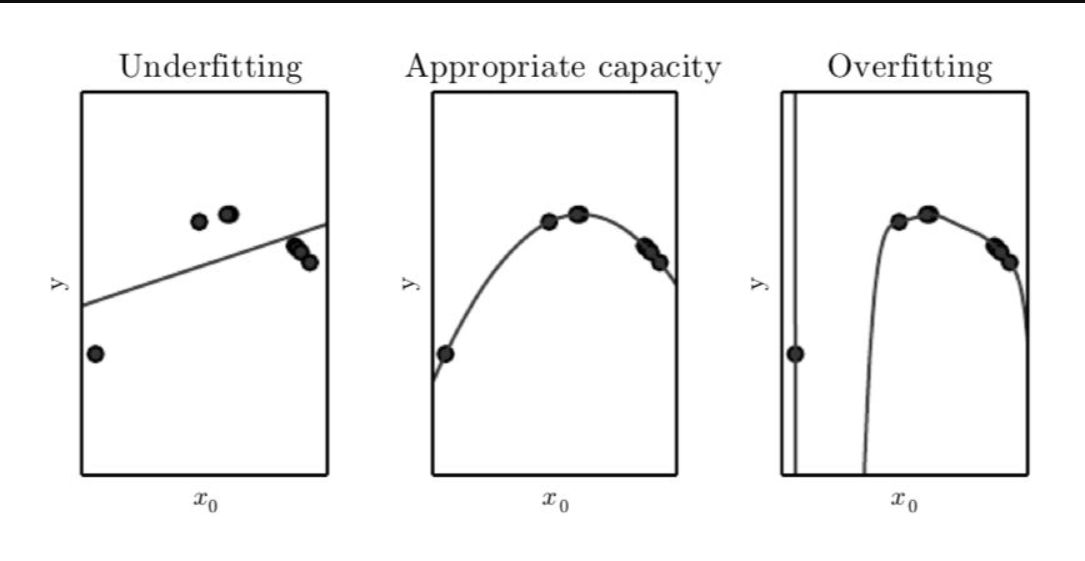
## （2）特征映射(Feature mapping)

我们认为y的取值不光与X1，X2...X8有关，还和其次方项及交叉项有关，所以将X变量进行扩充为X\_exp，该过程就是将原本8项,扩充为包含常数项在内的44项，变成X1,X2

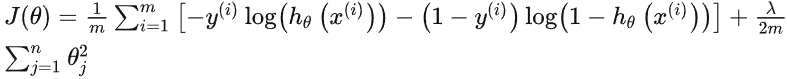
..X8,X1^2,X1\*X2..X8^2。

## （3）正则化逻辑回归(Regularized logistic regression)

这里再说明一下正则化的作用，我们要对逻辑回归进行正则化首先是要防止其过拟合和欠拟合(主要是过拟合)，就是防止图中左图及右图的情况：



什么时候容易出现左图情况？就是拟合所用的参数数量比较少，或者说大部分系数数值比较小的情况，而出现右图这种情况一般是参数数量比较多，或者说部分没那么重要的系数数值比较大的情况，因此我们想要拟合结果合理，我们有两种途径，一是调整参数数量，另一个便是将系数数值限定在一定范围内，这里正则化就是用到的第二种方法，我们将代价函数改写为：

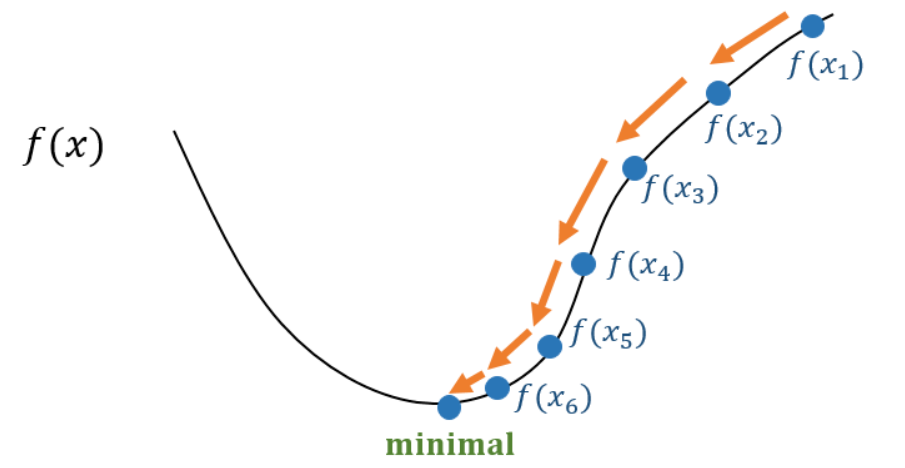


这样就能起到限制θ大小的作用，需要注意的是λ过大容易让θ过小从而出现欠拟合，λ过小甚至为0容易让θ过大从而出现过拟合，因此调节合适的λ大小非常重要。相应的偏导也变为：



# 2、梯度下降法(gradient descent)

既然已经找到了偏导，那么怎么找到它的最小值呢，我们这里采用的是梯度下降法，在机器学习的核心内容就是把数据喂给一个人工设计的模型，然后让模型自动的“学习”，从而优化模型自身的各种参数，最终使得在某一组参数下该模型能够最佳的匹配该学习任务。那么这个“学习”的过程就是机器学习算法的关键。梯度下降法就是实现该“学习”过程的一种最常见的方式，尤其是在深度学习(神经网络)模型中，BP反向传播方法的核心就是对每层的权重参数不断使用梯度下降来进行优化。梯度下降法(gradient descent)是一种常用的一阶(first-order)优化方法，是求解无约束优化问题最简单、最经典的方法之一。我们来考虑一个无约束优化问题minxf(x), 其中f(x)为连续可微函数，如果我们能够构造一个序列x0,x1,x2,...，并能够满足：f(xt+1)<f(xt),t=0,1,2,...，如图



# 3、数据概括

注意：在导入数据集diabetes.csv之前去除了第一行，形成新的数据集diabetes1.csv，本实验使用的是只有数据的diabetes1.csv。数据集属性信息：

怀孕：表示怀孕次数

葡萄糖：表达血液中的葡萄糖水平

血压：表示血压测量

皮肤厚度：表示皮肤的厚度

胰岛素：表达血液中的胰岛素水平

BMI：表示体重指数

糖尿病谱系功能：表示糖尿病百分比

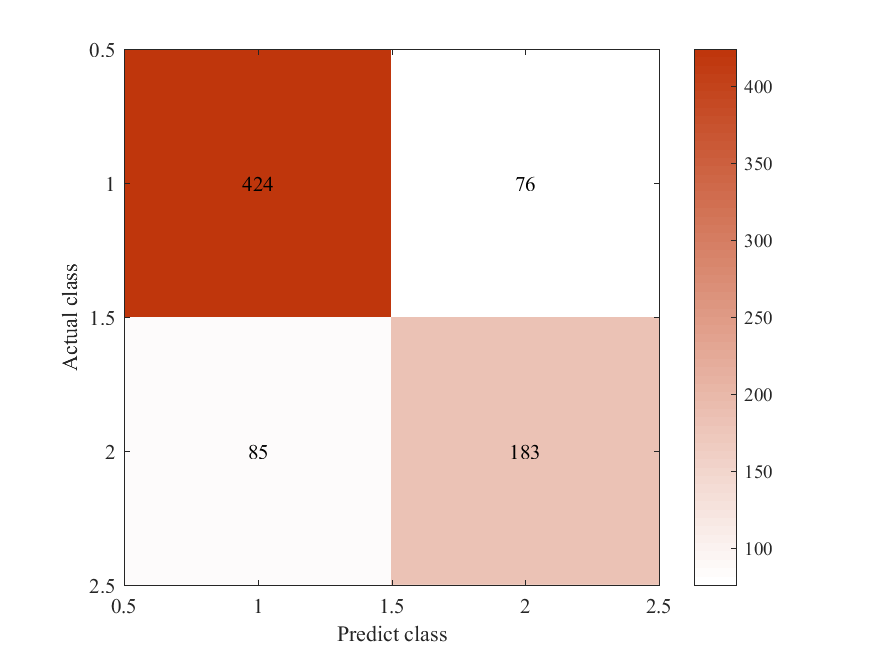
年龄：表示年龄

结果：为了表示最终结果 1 为“是”，0 为“否”

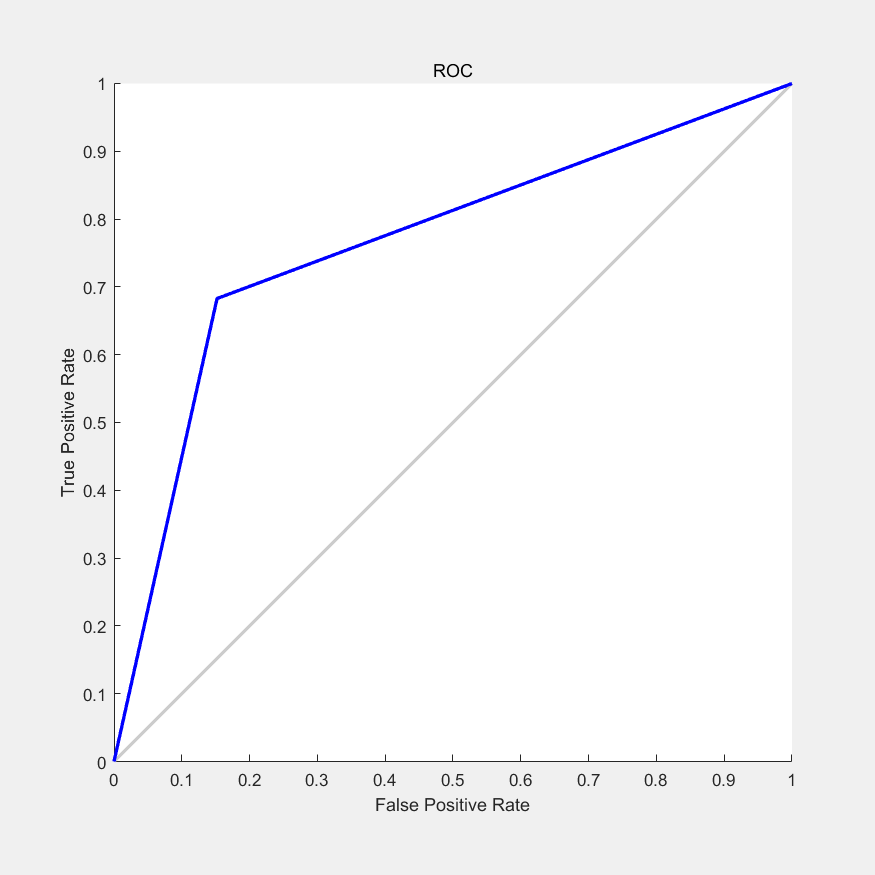
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 怀孕 | 葡萄糖 | 血压 | 皮肤厚度 | 胰岛素 | BMI | 糖尿病谱系功能 | 年龄 |
| 最大值 | 17 | 199 | 122 | 99 | 846 | 67.10 | 2.42 | 81 |
| 最小值 | 0 | 0 | 0 | 0 | 0 | 0 | 0.078 | 21 |
| 平均值 | 3.84 | 120.89 | 69.10 | 20.53 | 79.79 | 31.99 | 0.471 | 33.24 |
| 标准差 | 3.36 | 31.97 | 19.35 | 15.95 | 115.24 | 7.88 | 0.33 | 11.76 |
| 上四分位数 | 6 | 140.5 | 80 | 32 | 127.5 | 36.6 | 0.626 | 41 |
| 下四分位数 | 1 | 99 | 62 | 0 | 0 | 27.3 | 0.244 | 24 |

# 4、实验结果

混淆矩阵如图：



准确率为79.04%。



# 5、总结

逻辑回归也称作logistic回归分析，是一种广义的线性回归分析模型，属于机器学习中的监督学习。其推导过程与计算方式类似于回归的过程，但实际上主要是用来解决二分类问题（也可以解决多分类问题）。所以在本实验对糖尿病数据的分类中发挥着至关重要的作用，但也存在许多不够完善的地方。

数据来源：[the comparison of ml and dl and consideration of sampling... | Kaggle](https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset)