逻辑回归方法原理与应用

李卓冉
（河南省实验中学，河南　郑州　450000）

摘　要：本文重点讨论了逻辑回归方法的原理、模型和解决问题的一般步骤，并针对现有逻辑回归方法提出了一些改进。最后应用逻辑回归方法解决一个产品检测的实际问题，验证了逻辑回归方法对于非线性分类问题的有效性。

关键词：机器学习；逻辑回归；非线性；分类

DOI:10.19474/j.cnki.10-1156/f.001686

21 世纪的三大科学中的人工智能（Artificial Intelligence, AI）正如一颗彗星闪入了公众的视野，而在人工智能领域中能够较好体现智能的一个分支就是机器学习（Machine Learning, ML）。机器学习的发展并不漫长 从开始的符号机器学习到前些年来大热的统计机器学习，再到近几年深度学习的出现等。它的应用很广泛，例如大规模的数据挖掘（网页点击数据，治疗记录等）、自然语言处理、计算机视觉、推荐系统等。近年来机器学习出现了一些新的方向，例如深度学习和无终止学习等 [1]，掀起了世界各地对于人工智能学习的潮流。中国的众多高校也逐渐开设人工智能专业与机器学习学科，重视相关领域人才的培养。

本文详细讨论了逻辑回归的原理和解决问题的一般步骤，并利用逻辑回归方法解决实际非线性分类问题。针对逻辑回归方法的探索和研究，可以为 k 近邻、支持向量机等复杂机器学习方法和神经网络、增强学习等深度学习方法奠定基础。

一、逻辑回归方法

逻辑回归主要用来解决分类问题，最原始的逻辑回归用来解决二分类问题 [2]，在此基础上进行算法的修改，可以得到面向于解决多分类问题的逻辑回归方法。本文以二分类问题为例介绍逻辑回归的相关内容。逻辑回归的思想是：二分类问题的输出 y 并不是一定范围内连续的值，其输出只能为 0 或者 1 两种。为了将连续的输出转化为二值问题，可以通过非线性函数将连续值转换为离散的二值。

逻辑回归方法主要包括以下几个因素：假设函数、决策边界、代价函数以及参数优化方法。

（一）假设函数

假设函数（Hypothesis Representation）的构造方法以多变量线性回归问题为基础，其计算方法为综合考虑多个变量得到其线性组合。而对于二分类问题，假设函数的取值应该满足 $0 \leqslant h_\theta \leqslant 1$，因此采用非线性函数 Sigmoid 函数（Sigmoid Function）来将任意范围内的值规范化到 [0,1] 区间内。逻辑回归方法的假设函数计算如式 1–1 所示。

$$\begin{cases} g(z) = \dfrac{1}{1+e^{-z}} \\ h_\theta(x) = g\left(\theta^T x\right) \end{cases} \tag{1–1}$$

其中 x 为多维输入变量，θ 为多维输入变量对应的权值，h_θ（x）为输入变量为时模型给出的输出值，其范围在 [0,1] 区间内。

（二）决策边界

由于逻辑回归解决的是二分类问题，其输出结果必须为离散的二值。虽然经过 sigmoid 非线性函数可以将任意连续值规范化为 [0,1] 区间内的值，但是其仍然为连续值。此时假设函数的取值，可以进行这样的解读。逻辑回归方法面向的是一个二分类问题，即确定某个输入对应的输出是 0 或者是 1。当假设函数的输出为某个介于 0 到 1 之间的数字时，其数值大小代表着该输入对应输出是 1 的概率。例如 h_θ（x）=0.7，则意味着该输入变量对应的输出是 1 的概率为 0.7，是 0 的概率为 0.3。因此为了得到二分类问题的最终输出，可以通过设定决策边界来将连续值转化为离散的二值。例如对于某个问题，确定当假设函数的输出大于某个设定的边界时，即将该结果视为 1，否则视为 0，这样即可得到适用于二分类问题的离散二值输出。

（三）代价函数

对于算法中的权值的不同取值，需要通过代价函数来衡量其对实际问题拟合效果的优劣。代价函数越小代表模型对于实际问题的适应度越好，模型参数的优化过程就是不断的变化权值的取值，使得代价函数不断减小。

通常回归问题都会选择误差的平方均值作为代价函数，而对于二分类逻辑回归问题，代价函数的选取规则稍有变化。逻辑回归的代价函数统计每一个训练样本的偏差值并取平均，而每个样本的偏差通过对数运算以更好的适应二分类问题，代价函数的具体计算规则如式 1–2 所示，式 1–3 为代价函数的向量化表达。

$$\begin{cases} J = \dfrac{1}{m}\sum_{i=1}^{m} \mathrm{Cost}\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) \\ \mathrm{Cost}\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) = -\log\left(h_\theta(x)\right), \text{if } y = 1 \\ \mathrm{Cost}\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) = -\log\left(h_\theta(x)\right), \text{if } y = 0 \end{cases} \tag{1–2}$$

$$J = \frac{1}{m}\left(-y^T \log(h) - (1-y)^T \log(1-h)\right) \tag{1–3}$$

其中，x 为多维输入变量，θ 为多维输入变量对应的权值，为训练样本总数，h 为样本对应的实际输出，为样本对应的计算输出。

（四）参数优化方法

参数的优化过程即为不断改变权值的取值使得代价函数不断降低，进而得到较好的符合实际应用的模型。逻辑回归问题的参数更新规则采用梯度下降算法 [3]，通过不断计算代价函数关于权值的梯度，并利用梯度的负方向为函数下降速度最快的方向这一准则更新权值，使得代价函数随着梯度的更新而不断下降。对应于式 1–2 所示的代价函数构造形式，可以得到其相对于各个权值的梯度，进而得到权值的更新规

则，如式 1–4 所示，式 1–5 为权值更新规则的向量化表达，其中 α 为学习率，其大小代表了参数更新的速度。

$$\theta_j = \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} \quad (1-4)$$

$$\theta = \theta - \frac{\alpha}{m}X^T(g(X\theta) - \vec{y}) \quad (1-5)$$

二、逻辑回归方法的改进

（一）逻辑回归的潜在问题

逻辑回归方法适用于分类问题，但是在处理一些问题时，为了使算法具有更强的拟合能力，我们会构造能尽量拟合所有数据的假设函数，此时假设函数表达式中变量较多。同时如果没有足够的数据集(训练集)去约束这个变量较多的模型，就会发生过拟合问题，即假设函数过度拟合已知数据，使得算法在训练集上的表现较好但是难以泛化，导致算法无法对未知数据进行很好的预测。

（二）逻辑回归的正则化

为了解决过拟合问题，要对上文的逻辑回归方法施加正则化改进，改进的目的是避免假设函数出现过度拟合训练数据的情况。可以通过修改算法的代价函数、假设函数的表达式以及更新权值的迭代方法来实现正则化：

1、正则化的代价函数：

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(h_\theta(x^{(i)})\right) + (1-y^{(i)})\log\left(1-h_\theta(x^{(i)})\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \quad (2-1)$$

正则化的代价函数与上文的代价函数形式类似，只是在结尾加了一项用于惩罚权值取值过大的正则项，其中是正则化权重，用来平衡预测误差的惩罚与参数正则的惩罚。

2、基于梯度下降算法的正则化权值更新规则

$$\theta_j = \theta_j - \alpha\left[\left(\frac{1}{m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)}\right) + \frac{\lambda}{m}\theta_j\right], j \in \{1,2\ldots n\} \quad (2-2)$$

通过对原始逻辑回归方法施加正则化改进，可以避免学习到绝对值很大的权值，算法收敛后模型的权值都维持在较小的范围内，使得泛化过程中算法不会因某个输入量而产生较大的影响，保证算法的稳定性。正则化方法不仅是逻辑回归方法中有效避免过拟合的方法，也为其他复杂机器学习以及深度学习解决过拟合问题提供了参考和改进思路 [4]。

（三）应用逻辑回归方法解决实际分类问题

应用正则化的逻辑回归方法解决生产线产品检测的非线性分类问题。已知某厂篮球生产线产品检测环节，篮球的合格与否主要取决于篮球的半径与篮球的质量两个指标，记待检篮球的半径为变量 x_1，质量为变量 x_2。某厂随机挑选 118 个待检篮球产品，检验其是否合格并记录每个产品的半径与质量。将这两个变量做零均值与归一化处理后，得到图 3–1a 的统计结果。根据式(2–1)建立逻辑回归模型，并根据式(2–2)进行模型参数的更新，进而得到能将合格产品与不合格产品区分开来的边界曲线。正则化系数值较大时，边界简单但是误分情况相对较多；较小时能将大部分样本正确划分边界，但是边界曲线的形状也更为复杂。图 3–2 为 λ =0.1 时的分类边界。

从该产品生产线产品检测问题结果来看，逻辑回归方法能够找到比较合理的分类边界，将合格产品与不合格产品较大程度的分离开来，证明了该方法对于非线性分类问题的有效性与合理性。

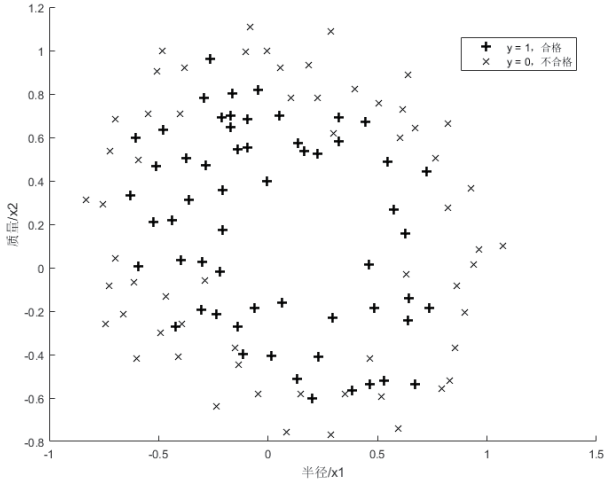

图 3–1　某厂 118 个篮球产品检测结果与变量统计图

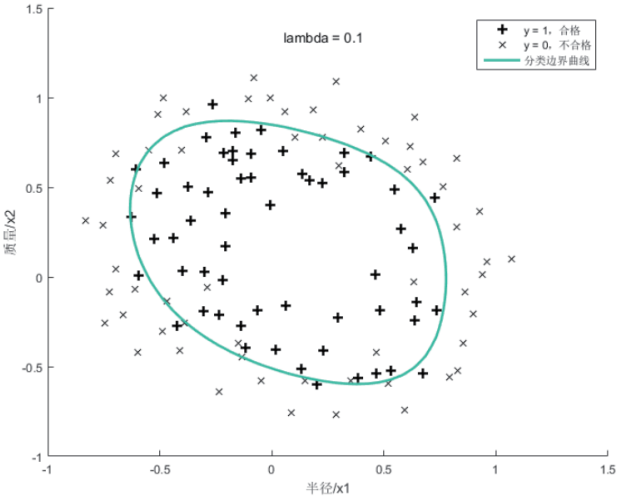

图 3–2　λ =0.1 时的分类边界

参考文献：

[1] 周志华 . 机器学习及其应用 2007[M]. 清华大学出版社 ,2007.

[2] 许冲 , 徐锡伟 . 逻辑回归模型在玉树地震滑坡危险性评价中的应用与检验 [J]. 工程地质学报 ,2012,20(3):326–333.

[3] 刘颖超 , 张纪元 . 梯度下降法 [J]. 南京理工大学学报自然科学版 ,1993 (2) :12–16.

[4] 朱劲夫 , 刘明哲 , 赵成强 , 等 . 正则化在逻辑回归与神经网络中的应用研究 [J]. 信息技术 , 2016 (7) :1–5.

作者简介：

李卓冉，男，汉族，籍贯河南省永城市，就读于河南省实验中学，研究方向是机器学习。