**基于神经网络的logistic回归模型**

董叶豪

指导老师：柯宗武

**摘要：**logistic回归是典型的二值分类器，用来处理二分类问题。人工神经网络是由多个神经元构成的，而logistic回归可视为只有一个神经元的单层神经网络。本文将从神经网络的构建角度来阐述logistic回归。包括：（1）数据预处理，即中心化和标准化特征数据；（2）神经网络的前向传播，即从输入特征数据到求代价函数；神经网络的反向传播，即从代价函数求对权重参数w和偏置参数b的偏导数；（3）使用梯度下降算法来优化模型参数；（4）计算预测值和预测精确度

**关键字：**logistic回归 神经网络

**0引言：**

logistic回归一般适用于简单的线性二分类问题，通常的研究的问题是在某些因素下某个结果是否发生，比如通过图片的所有像素点来预测图片是否是一只猫图片。神经网络是由多个神经元链接而成，正是多个神经元相互合作，神经网络才能解决复杂的非线性分类等问题。而logistic回归就是一个神经元，因此理解logistic回归对理解神经网络具有重要意义。

**1工作原理**

**1.1神经元模型**

如图1-1所示为一个神经元，它用于分析结果y与一些影响因素之间关系。将每个影响因素记为一个特征，假设有m组训练样本，每组样本中都有n个特征以及一个结果y，则特征矩阵X为维,结果向量Y是m维行向量。

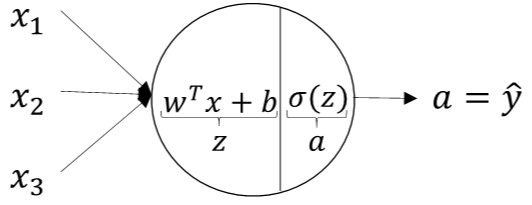


图 1-1 一个神经元

**1.2 预处理数据**

输入的特征x本来可能分布很不均匀，需要对其进行预处理，也就是归一化。经过归一化后的数据范围压缩在（0,1）之间或者（-1,1）之间，以免模型参数被分布范围较大或较小的数据支配。归一化的方法一般有两种：

(1)min-max标准化。

该方法是对原始数据的线性变换，使结果值映射到（0,1）之间。转换函数如下

**(1)**

其中max是输入特征的最大值，min是输入特征的最小值。该方法的缺点是当引入新数据时，max和min的值可能发生变化，需要重新定义。

(2)Z-score标准化方法

该方法需要先计算出原始数据的均值（mean）和标准差（standard deviation），**经过**处理后的数据均值为0，标准差为1，符合标准正态分布。转化函数为：

(2)

其中是每组训练样本特征的平均值，是每组训练样本特征的标准差

**1.3前向传播**

一个神经元包含多个特征的线性组合和激活函数，其中激活函数一般为非线性函数。每一个特征对应有一个权值，则对于一个训练样本，多个特征的线性组合表示为

**(3)**

其中b是一个标量，是对应的偏置（biase）。

对于m个训练样本，有

**(4)**

其中X是（n,m）维矩阵，w是n维行向量，Z，b是m维行向量。

对于二值分类问题，激活函数一般去sigmoid函数，其函数表达式为

如图1-3是sigmoid函数图像，其值域取值范围为（0,1），刚好适用于结果是[0 ,1]的二分类问题。从图像可以看出当很大时，sigmoid函数的斜率很小，会降低模型训练的速度，因此需要归一化输入特征，使|Z|接近于0，这也是需要预处理数据的原因之一。

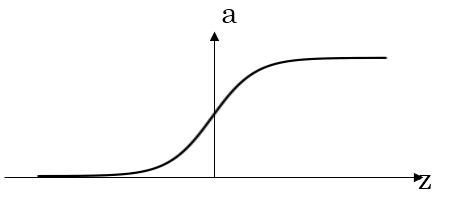


图1-3 sigmoid函数

**1.4初始化模型参数**

对于单个神经元来说，可以将权值向量w和偏置向量b初始化都为零向量。但是对于多个神经元的神经网络，对权值向量w的初始化一般采用标准正态分布，而不能将w初始化为零向量。因为若把w初始化为零向量，则每个神经元的训练结果一样，不管增加再多的神经元，训练的效果都一样。

**1.5计算代价函数(Cost Function)**

对m组训练样本的训练是相互独立的，对于其中的一组训练样本，)，若结果=1时，,则结果=0的概率。又训练m组数据是相互独立的事件，则那么

**(5)**

为了让预测的概率越准确，则概率P(y|x)应尽可能大。采用最大释然估计，将等式（5）两边取对数，有

**(6)**

为了求（5）式的最大值，则定义代价函数

**(7)**

**1.6反向传播**

代价函数J是关于和b的复合函数，由复合函数的链式法则，对于一组训练样本有

**(8)**

**(9)**

**由（9）可得**

**(10)**

其中j=1,2,…n。那么对n维列向量w的偏导数可表示为

**(11)**

其中x是n维列向量，是一组训练样本的输入特征

**(12)**

对于m组训练样本，每组训练样本之间相互独立，则

**(13)**

**(14)**

**其中i=1,2,…,m。由（13）（14）可得**

**(15)**

**(16)**

其中A是激活函数的输出向量，A是m维行向量。

**，是对m组训练样本每组求得的，取平均值，表达式如下**

**(17)**

**(18)**

**1.7更新参数**

使用公式(13)来更新模型参数，每训练一次神经网络，w,b都会朝着全局最优点移动，直到训练次数到达，或者梯度，近似为零、参数不再变化为止。如图1-7所示（为了画图方便，w取一维）。

**(19)**

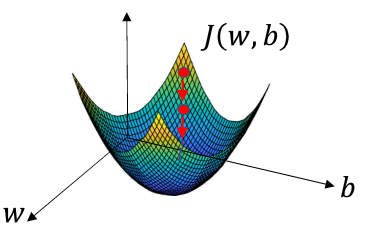


图1-7梯度下降

**1.8预测值与预测精确度**

由1.5节可知，是事件发生的概率，取阈值为0.5。当时，事件发生，预测值y\_prediction为1；当,事件不发生，预测值y\_prediction为0

预测精确度是把预测值y\_prediction与真实结果y作比较，计算方法如下

**(20)**

**2结语**

通过大量样本数据来训练神经网络，以得到最优模型参数。通过计算出来的模型参数，就能预测出在一定输入条件下的结果。Logistic回归可以看成是一个神经元，它一般用来解决二值分类问题，测试精确度为70%。然而多层神经网络的计算精度可以达到80%。因此基于单个神经元的logistics回归的计算能力是由局限性的，需要使用更为复杂的多层神经网络。

**参考文献：**

[1]周志华. 机器学习[M].北京：清华大学出版社，2016年:97-99

[2]韩立群. 人工神经网络理论、设计及应用（第二版）[M].北京：化学工业出版社，2007：24-27

# [3] Nielsen,M. Neural Networks and Deep Learning[M].Online,2016,1-20