Logistic回归

2019-2-27

**一、概念**

Logistic回归模型是一种分类模型，一般用于二分类问题（当然也可以用于多分类），本质上是一种基于条件概率的判别模型，由条件概率P(Y|X)表示，这里X是随机变量，Y取0或1。

Logistic回归模型是如下条件概率分布：

 （1）

 （2）

对于任意输入样本x，我们通过求解P(Y=1|x)和P(Y=0|x)的结果，选择较大的概率值将x归为那一类，通常我们一般选择P(Y=1|x)的分布作为Logistic回归的模型。

另外，Logistic回归模型也可以通过sigmoid函数来获得。

sigmoid函数公式：

 （3）

为实现Logistic回归模型，在样本的每个特征 上都乘以一个回归系数 ，然后将加权和 带入sigmoid函数中，就得到一个在0~1之间的结果，结果大于0.5的数据被归为1类，否则归为0类，这样也就通过（3）式得到了（1）式。

**二、参数计算**

为了使得到Logistic回归分类模型尽可以准确，我们就需要求解出公式中的最优参数w。对于求解最优参数w的方法，可以采用极大似然估计法来估计模型：

首先我们将（1）（2）式重新写为：

 （4）

 （5）

这个时候似然函数为：

 （6）

对数似然函数为：

 （7）

满足L(w)最大的w值即为满足Logistic模型的最优参数。一般通过对L(w)求导等于零，得到的w值即为最优w值。但是由于实际求解过程中，w不容易直接解出，我们就需要通过间接的方法：比如梯度下降（上升）法、拟牛顿法等来求解。

这里介绍通过梯度上升法来求解的步骤。因为求解的是最大值，所以使用梯度上升法；当求解使L(w)最小的w时，我们就需要的使用梯度下降法。

根据公式（1）（2），上面（7）式可以进一步变形为：

 （8）

对L(w)求偏导数，可得：

 （9）

对于梯度上升法，通过迭代得方法来求解参数w，迭代公式如下：

  （10）

通过迭代的方法就可以得到的w即为所求。

**三、算法实现**

《机器学习实战》一书中提供了几个实际算法实现的例子与对应的数据集，这里进行复现。

梯度上升法和随机梯度上升法的具体代码见**Logistic.ipynb。**

从疝气病症状预测病马的死亡率的具体代码见**Logistic\_house.ipynb。**

**四、参考文献**

1. 《统计学习方法》李航
2. 《机器学习实战》Peter Harrington
3. Python3《机器学习实战》学习笔记（六）：Logistic回归基础篇之梯度上升算法（https://blog.csdn.net/c406495762/article/details/77723333）