文档

张苏月

1. 自定义函数说明
   1. standardization\_set(arr,n)

功能：0-1标准化

输入：数据集

输出：标准化后的数据集

* 1. normalization(x)

功能：0-1标准化一个list

输入：一个list

输出：标准化后的list

* 1. sigmoid(h)

功能：sigmoid函数的实现

目的说明：将计算出的直线数值归一化，转化为对应的概率，在此为是否为1类的概率。

参数：一个数

返回值：该数对应的sigmoid函数值

* 1. gradual\_decent(variable,classLabel,K, maxCycle=5000, alpha=0.01)

功能：逻辑回归函数的梯度下降。

流程说明：将输入都转化为矩阵，建立（属性值+1\*1）大小的θ矩阵，在最大迭代次数范围内进行整体样本和θ之间的计算，用每次计算出来的导数梯度下降，即，, m为样本总数

参数：训练集的属性值数组x(样本总行数,属性数+1)，类别值数组y(样本总行数,1)，指定的最大迭代次数(默认5000)，指定的学习率(默认0.01)

返回值：计算出的满足最大迭代次数或收敛的theta值。

* 1. logistic\_regression(theta,X)

功能：根据选择的系数建立逻辑回归函数进行分类

流程说明：根据和x计算全体样本判断为1类的概率，将大于0.5的分为1类，小于0.5的分为0类，将结果写入pre列表中并返回为（样本数\*1）的数组。

参数：回归方程系数向量theta(属性值+1,1),参与预测的属性向量x(样本总行数,属性数+1)

返回值：逻辑回归预测类别(样本总行数,1)

* 1. accuracy\_rate(predict,true\_value)

功能：计算模型在训练集上的正确率

参数：预测类别向量theta，真实类别向量y

返回值：该模型正确率

* 1. set\_stack(sets)

功能：训练集属性最前加上一列1方便计算截距

参数：训练集

返回值：更改后的训练集

1. 流程说明
2. 数据处理。使用numpy引入数据集和测试集，并将数据集0-1标准化处理。
3. 可视化处理。利用pandas库的loc索引，将0类和1类单独出来，通过matplotlib库进行可视化制作散点图，绿色为1类，蓝色为0类。
4. 分别实现sigmoid函数、逻辑回归梯度下降函数、逻辑回归预测结果函数、处理数据集和计算正确率的函数。
5. 利用写好的函数训练theta值，500000次迭代，学习率0.01
6. 打乱训练集，从中抽取600个数据进行验证，计算正确率。
7. 用读取好的验证集集和theta值进行测试计算正确率，再将输入test集进行预测，结果转化为DataFrame形式作为submit\_test.csv输出并显示。
8. 遇到的问题及解决办法

|  |  |
| --- | --- |
| 问题 | 解决方法 |
| 由于矩阵乘法顺序不当而报错 | 用np.shape()查看矩阵形状然后进行合适的调整 |
| 忘记theta为0时exp值应该为1，导致部分除数变为0，theta变为nan | 将计算概率矩阵初始值设为1。后来加上了最后一列theta，删除了这个初始值矩阵 |
| 写错了代价函数而导致theta并没有向代价函数减小的方向收敛，进而预测正确率低，迭代次数变大后还会导致除数为inf，theta全变为nan | 更正gradual\_decent函数中的公式 |
| 由于test集未0-1标准化而train样本标准化了，正确率低 | 将test集也0-1标准化 |