对数几率回归实验报告

PB20010429 侯相龙

2022 年 10 月 16 日

1 实验内容

对借贷数据进行对数几率回归，比较评估模型。

2 实验设备和环境

- **实验设备：**

设备名称：LAPTOP-9J92NDCJ

处理器：Intel(R) Core(TM) i5-1035G1 CPU @ 1.00GHz 1.19 GHz

RAM：16.0 GB

- **实验环境：**

PyCharm 2021.3.2

3 实验方法与步骤

3.1 处理空列

用均值代替空值

3.2 类别特征编码

对于两类的问题，我们用 0、1 编码；对于多类（存在序关系）的问题，将 $[0,1]$ 等距划分取端点值编码

3.3 数据集划分

- **标准化：**采用去均值和方差标准化方法，将数据转化成均值为 0，方差为 1 的分布，以提高梯度下降法的速度
- **随机分层取样：**根据 Loan_Status 的分布进行分层采样，保持分布的一致性。
- **数据集划分：**首先划分为训练集、验证集、测试集，比例 $0.7:0.15:0.15$ 。再对各集合进行 X 属性和 y（Loan_Status）划分。

3.4 模型训练

Step1: 根据训练集对模型进行训练，作损失函数图并得到拟合系数。

Step2: 在验证集上训练比较，分类阈值更新为准确率最高的分类阈值。

Step3: 输入测试集 X，预测分类结果 $\hat{y}$，与测试集的真实分类结果 y 比较，进行准确率评估。

3.5 调参比较：

分别调整分类阈值、学习率、最大迭代次数等参数，对建立的模型进行比较。

4 实验结果与分析

本实验采用留出法进行了多次测试，因为随机取样样本不同，每次实验结果会有差异. 下面列举两组实验的结果。

4.1 实验结果

- 训练次数-损失函数曲线：

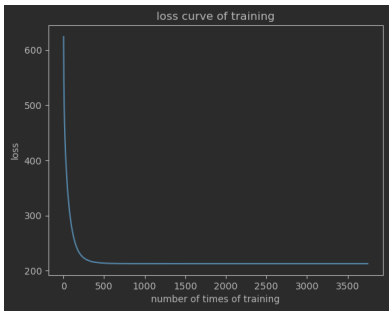

图 1: Test1

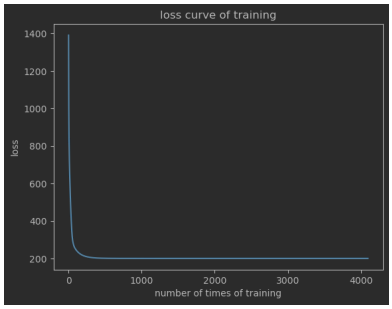

图 2: Test2

- 不同分类阈值下的正确率：

	默认阈值 0.5	验证得到的阈值 0.55
正确率	85.87%	85.87%

表 1: Test1

	默认阈值 0.5	验证得到的阈值 0.61
正确率	85.87%	86.96%

表 2: Test2

- **不同学习率的分类正确率：**

学习率	1e-1	1e-2	1e-3	1e-4
正确率	84.78%	84.78%	85.87%	85.87%

Test 1

学习率	1e-1	1e-2	1e-3	1e-4
正确率	79.35%	78.44%	86.96%	86.96%

Test 2

- **不同最大迭代次数的分类正确率：**

最大迭代次数	1e2	1e3	1e4	1e5
正确率	80.43%	85.87%	85.87%	85.87%

Test 1

最大迭代次数	1e2	1e3	1e4	1e5
正确率	80.43%	85.87%	86.96%	86.96%

Test 2

- **测试的最高精确度：**

经过多次试验，精确度最高为 86.96%

4.2 结果分析

(1) 就本实验数据而言，在某些情况下，在验证集上进行验证得出分类阈值确实能够增加分类正确率，但是与用默认参数 0.5 进行分类相差不大。但这对于处理某些特定问题（提高查全率），这是必要的。

(2) 在保证足够大迭代次数多情形下，在学习率比较高时，损失函数曲线会出现剧烈振动导致无法降低损失函数的值, 分类准确率低；学习率比较低时，迭代次数的增加会导致学习时间增长。

(3) 随着最大迭代次数多增加，分类正确率增加。