**作业-02 LightGBM模型调参**

陈恬，2021-7-8

根据第六章所讲述的调参方式，请使用 LightGBM 进行调参：

目标变量：loan\_status

要求：

1. 请在word当中记录详细的调参过程，包括参数和结果；

2. 请解释选择该调参步骤的原因；

3. 请报告在测试集上的准确率；

4. （可选）尝试不同的模型的组合，检查是否可以提升。

**结论：当num\_leaves=8, num\_round=2000, learning\_rate=0.005时，测试集上的准确率为0.91774（错误率为：0.08226），表现较优。结果记录可见自动调参第1组结果。**

**数据说明：**

在训练集和测试集中，发现'continuous\_annual\_inc\_joint', 'continuous\_dti\_joint', 'continuous\_mths\_since\_last\_delinq', 'continuous\_mths\_since\_last\_major\_derog', 'continuous\_mths\_since\_last\_record‘特征的缺失率比较大，虽然Lightgbm可以自动处理缺失值，但发现直接丢弃，得到的误差率更小。这几个指标含义从业务的角度来说感觉对最终的结果预测还是有一定影响，但是直接做补0处理不太合适，暂时没有想到合适的填补缺失值策略，本次调参过程先采取了舍弃处理模式，后续会继续尝试一些符合实际的填补方式来进行调参。

**手动调参过程：**

以下调参过程暂时只输出了验证集5折交叉验证的均值误差率，并根据验证集的误差率进行调参。

**Test0: 初始参数**

Params: {'num\_thread': 4, 'num\_leaves': 50, 'metric': 'binary', 'objective': 'binary', 'num\_round': 1000, 'learning\_rate': 0.04, 'feature\_fraction': 0.8, 'bagging\_fraction': 0.8}

The minimum is attained in round 155

Finished loading model, total used 1000 iterations

The minimum is attained in round 115

Finished loading model, total used 1000 iterations

The minimum is attained in round 146

Finished loading model, total used 1000 iterations

The minimum is attained in round 115

Finished loading model, total used 1000 iterations

The minimum is attained in round 129

Finished loading model, total used 1000 iterations

**average error rate: 0.08376000000000001**

文本

描述已自动生成

分析：学习率过大，收敛过快，可将学习率调小

**Test1:降低learning\_rate至0.02**

Params: {'num\_thread': 4, 'num\_leaves': 50, 'metric': 'binary', 'objective': 'binary', 'num\_round': 1000, 'learning\_rate': 0.02, 'feature\_fraction': 0.8, 'bagging\_fraction': 0.8}

The minimum is attained in round 354

Finished loading model, total used 1000 iterations

The minimum is attained in round 248

Finished loading model, total used 1000 iterations

The minimum is attained in round 297

Finished loading model, total used 1000 iterations

The minimum is attained in round 235

Finished loading model, total used 1000 iterations

The minimum is attained in round 247

Finished loading model, total used 1000 iterations

error\_rate: [0.07440000000000002, 0.08220000000000005, 0.08689999999999998, 0.08509999999999995, 0.0827]

**average error rate: 0.08226**

文本

描述已自动生成

分析：相对于上一步误差率**0.08376,** 本次验证集误差率为**0.08226，有所下降**，说明调整方向是对的。从收敛的round来看，速度依然有点快，可以继续降低learning\_rate.

**Test2:降低learning\_rate至0.015**

Params:{'num\_thread': 4, 'num\_leaves': 50, 'metric': 'binary', 'objective': 'binary', 'num\_round': 1000, 'learning\_rate': 0.015, 'feature\_fraction': 0.8, 'bagging\_fraction': 0.8}

The minimum is attained in round 473

Finished loading model, total used 1000 iterations

The minimum is attained in round 359

Finished loading model, total used 1000 iterations

The minimum is attained in round 380

Finished loading model, total used 1000 iterations

The minimum is attained in round 306

Finished loading model, total used 1000 iterations

The minimum is attained in round 328

Finished loading model, total used 1000 iterations

error\_rate: [0.07389999999999997, 0.08209999999999995, 0.08530000000000004, 0.08550000000000002, 0.08330000000000004]

**average error rate: 0.08202000000000001**

文本

描述已自动生成

分析：相对于上一步验证集误差率**0.08226,** 本次验证集误差率为**0.08202，也有所下降**。从收敛的round来看，速度依然有点快，可继续降低learning\_rate.

**Test3:降低learning\_rate至0.01**

Params:{'num\_thread': 4, 'num\_leaves': 50, 'metric': 'binary', 'objective': 'binary', 'num\_round': 1000, 'learning\_rate': 0.01, 'feature\_fraction': 0.8, 'bagging\_fraction': 0.8}

The minimum is attained in round 673

Finished loading model, total used 1000 iterations

The minimum is attained in round 493

Finished loading model, total used 1000 iterations

The minimum is attained in round 566

Finished loading model, total used 1000 iterations

The minimum is attained in round 524

Finished loading model, total used 1000 iterations

The minimum is attained in round 494

Finished loading model, total used 1000 iterations

error\_rate: [0.07420000000000004, 0.08199999999999996, 0.08589999999999998, 0.08379999999999999, 0.0817]

**average error rate: 0.08152**

文本

描述已自动生成

分析：相对于上一步验证集误差率**0.08202,** 本次验证集误差率为**0.08152，也有所下降**。从收敛的round来看，速度依然有点快，可继续降低learning\_rate.

**Test4:降低learning\_rate至0.008**

Params: {'num\_thread': 4, 'num\_leaves': 50, 'metric': 'binary', 'objective': 'binary', 'num\_round': 1000, 'learning\_rate': 0.008, 'feature\_fraction': 0.8, 'bagging\_fraction': 0.8}

The minimum is attained in round 858

Finished loading model, total used 1000 iterations

The minimum is attained in round 633

Finished loading model, total used 1000 iterations

The minimum is attained in round 722

Finished loading model, total used 1000 iterations

The minimum is attained in round 641

Finished loading model, total used 1000 iterations

The minimum is attained in round 614

Finished loading model, total used 1000 iterations

error\_rate: [0.07289999999999996, 0.0827, 0.08450000000000002, 0.08409999999999995, 0.08140000000000003]

**average error rate: 0.08112**

文本

描述已自动生成

分析：相对于上一步验证集误差率**0.08152,** 本次验证集误差率为**0.08112，下降了一点点**。从收敛的round来看，速度适中。在num\_leaves=50, num\_round=1000时，可考虑以将learning\_rate 设置在0.005~0.008左右。

**Test5:尝试降低树的深度，将num\_leaves降低至12，学习率和num\_round保持不变**

Params: {'num\_thread': 4, 'num\_leaves': 12, 'metric': 'binary', 'objective': 'binary', 'num\_round': 1000, 'learning\_rate': 0.008, 'feature\_fraction': 0.8, 'bagging\_fraction': 0.8}

The minimum is attained in round 995

Finished loading model, total used 1000 iterations

The minimum is attained in round 998

Finished loading model, total used 1000 iterations

The minimum is attained in round 955

Finished loading model, total used 1000 iterations

The minimum is attained in round 1001

Finished loading model, total used 1000 iterations

The minimum is attained in round 874

Finished loading model, total used 1000 iterations

error\_rate: [0.07150000000000001, 0.08089999999999997, 0.08389999999999997, 0.08409999999999995, 0.08020000000000005]

average error rate: 0.08012

文本

描述已自动生成

分析：相对于上一步验证集误差率**0.08112,** 本次验证集误差率为0.08012**，下降了0.1%**。从收敛的round来看，收敛较慢，可能round数不够，可调大round数。

**Test6: 在树的低深度条件下，将num\_round增加至4000**

Params: {'num\_thread': 4, 'num\_leaves': 12, 'metric': 'binary', 'objective': 'binary', 'num\_round': 4000, 'learning\_rate': 0.08, 'feature\_fraction': 0.8, 'bagging\_fraction': 0.8}

The minimum is attained in round 1003

Finished loading model, total used 4000 iterations

The minimum is attained in round 1556

Finished loading model, total used 4000 iterations

The minimum is attained in round 1020

Finished loading model, total used 4000 iterations

The minimum is attained in round 1576

Finished loading model, total used 4000 iterations

The minimum is attained in round 874

Finished loading model, total used 4000 iterations

error\_rate: [0.07250000000000001, 0.0826, 0.08620000000000005, 0.08240000000000003, 0.08140000000000003]

**average error rate: 0.08102000000000002**

文本

描述已自动生成

分析：相对于上一步验证集误差率0.08012**,** 本次验证集误差率同为0.08012**，**变化不大。说明将树的深度降低到小于50，是可以降低误差率的。但为避免过拟合情况，还可尝试一次较大树的深度。

**Test7: 将树的深度调至127，num\_round设置为1000**

Params: {'num\_thread': 4, 'num\_leaves': 127, 'metric': 'binary', 'objective': 'binary', 'num\_round': 1000, 'learning\_rate': 0.008, 'feature\_fraction': 0.8, 'bagging\_fraction': 0.8}

The minimum is attained in round 796

Finished loading model, total used 1000 iterations

The minimum is attained in round 587

Finished loading model, total used 1000 iterations

The minimum is attained in round 637

Finished loading model, total used 1000 iterations

The minimum is attained in round 535

Finished loading model, total used 1000 iterations

The minimum is attained in round 515

Finished loading model, total used 1000 iterations

error\_rate: [0.07530000000000003, 0.08499999999999996, 0.08640000000000003, 0.08450000000000002, 0.0816]

average error rate: 0.08256000000000001

文本

描述已自动生成

分析：将树的深度提高到较大深度，相对于上一步验证集误差率0.08012**,** 本次验证集误差率为0.08256**，**误差率反倒提高了，说明在特征变量中，可能是存在几个与目标变量强相关的特征变量，树的深度不宜设置太深。

**Test8: 将树的深度调至32，num\_round设置为1000**

Params: {'num\_thread': 4, 'num\_leaves': 32, 'metric': 'binary', 'objective': 'binary', 'num\_round': 1000, 'learning\_rate': 0.008, 'feature\_fraction': 0.8, 'bagging\_fraction': 0.8}

The minimum is attained in round 949

Finished loading model, total used 1000 iterations

The minimum is attained in round 714

Finished loading model, total used 1000 iterations

The minimum is attained in round 729

Finished loading model, total used 1000 iterations

The minimum is attained in round 778

Finished loading model, total used 1000 iterations

The minimum is attained in round 612

Finished loading model, total used 1000 iterations

error\_rate: [0.07340000000000002, 0.08220000000000005, 0.0847, 0.08379999999999999, 0.08079999999999998]

average error rate: 0.08098000000000001

文本

描述已自动生成

分析：将树的深度降低到32，相对于深度为12时的误差率0.08012**,** 本次验证集误差率为0.0809**，**说明还可将树的深度降低。

**Test9: 将树的深度调至8，学习率降至0.005， num\_round设置为2000**

Params: {'num\_thread': 4, 'num\_leaves': 8, 'metric': 'binary', 'objective': 'binary', 'num\_round': 2000, 'learning\_rate': 0.005, 'feature\_fraction': 0.8, 'bagging\_fraction': 0.8}

The minimum is attained in round 1991

Finished loading model, total used 2000 iterations

The minimum is attained in round 1996

Finished loading model, total used 2000 iterations

The minimum is attained in round 1991

Finished loading model, total used 2000 iterations

The minimum is attained in round 2000

Finished loading model, total used 2000 iterations

The minimum is attained in round 1485

Finished loading model, total used 2000 iterations

error\_rate: [0.0716, 0.0807, 0.08330000000000004, 0.08379999999999999, 0.07950000000000002]

**average error rate: 0.07978**

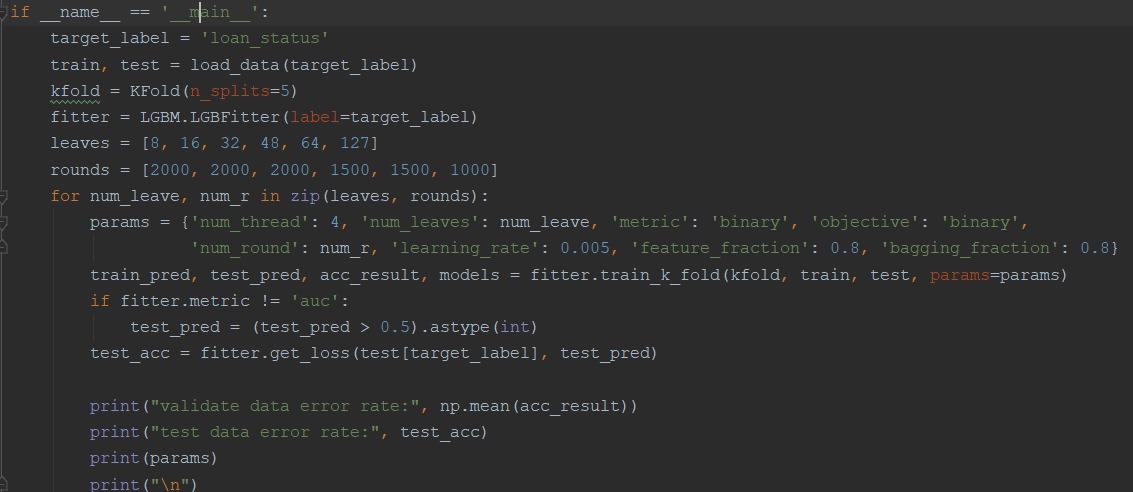
文本

描述已自动生成

分析：将树的深度降低到8，将num\_round设置为2000，learning\_rate调至到0.005，验证集上可将错误率降至0.07978 ，为目前验证集上较好的结果。

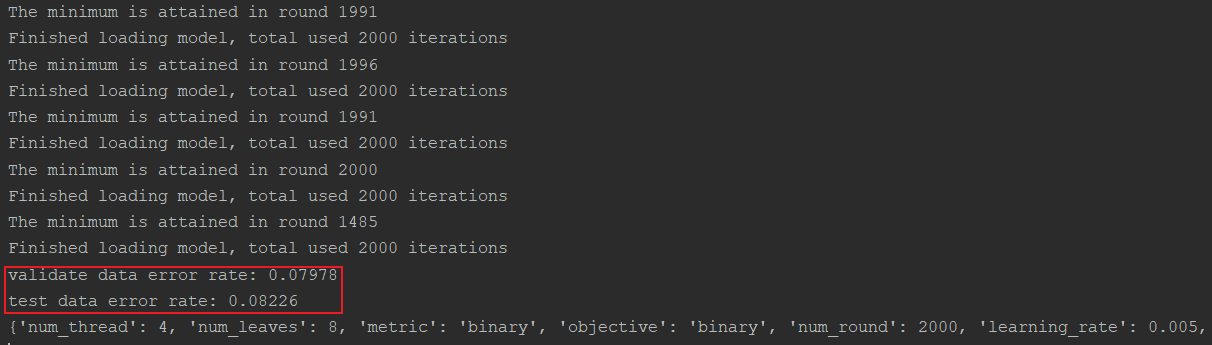
**自动调参过程：同时输出验证集和测试集上的误差率**

根据手动调参过程，将num\_leaves和num\_round设置为两组变量，主要观测这样组合过程中是否会存在一些过拟合的状况。组合调参代码如截图所示：

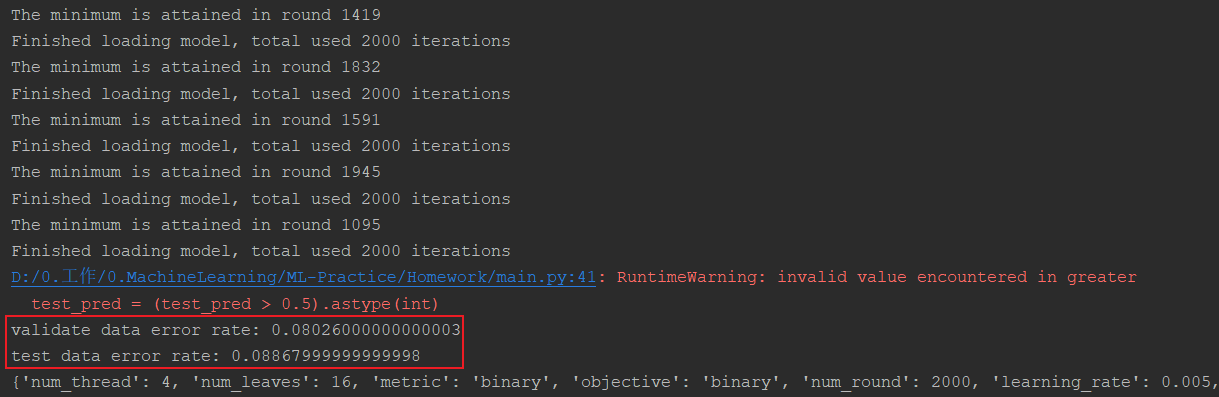


**结论：从验证集的误差率看，随着树的深度的增加，误差率单调递增的，但测试集上会有些波动；整体结果为在num\_leaves=8; num\_round=2000; learning\_rate=0.005表现最优。**

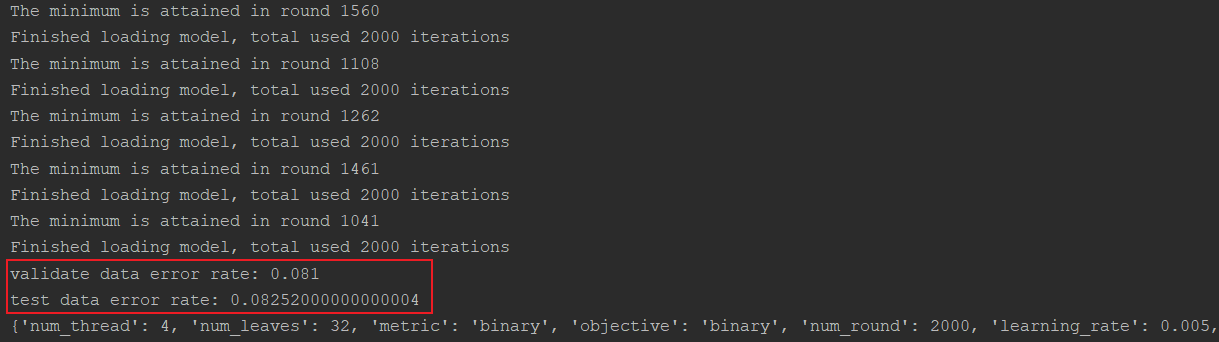
1. num\_leaves=8, num\_round=2000;



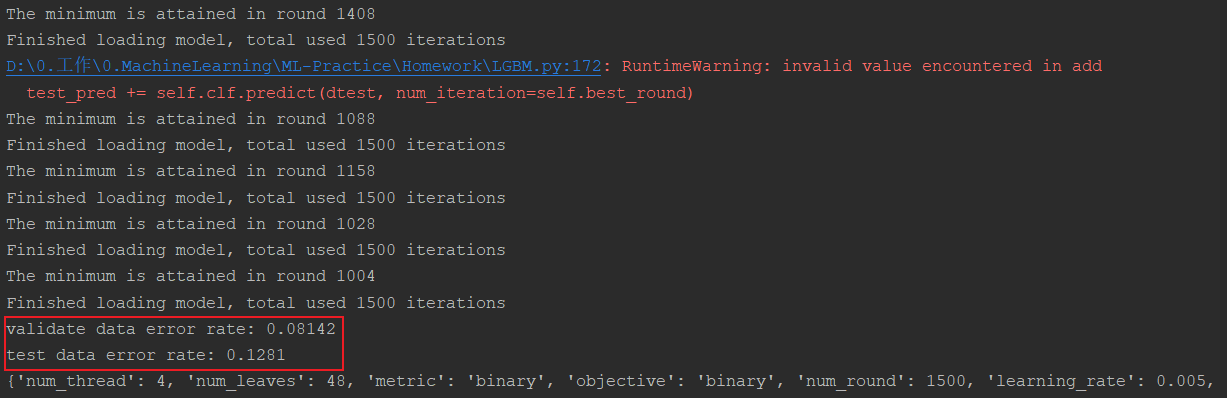
2) num\_leaves=16, num\_round=2000;



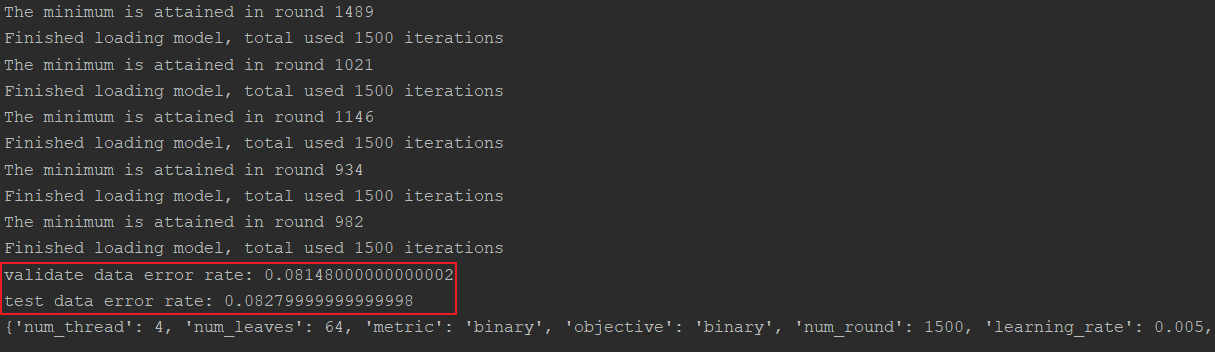
3) num\_leaves=32, num\_round=2000;



4) num\_leaves=48, num\_round=1500;



5) num\_leaves=64, num\_round=1500;



6) num\_leaves=127, num\_round=1500;

