Chapter06 作业

初始参数

参数设置：

```
{'num_thread': 4, 'num_leaves': 31, 'metric': 'binary',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
error_rate mean:
0.08132000000000002
```

num_round=1000，训练数据集比较小，预测1000轮应该足够了。learning_rate=0.1，按默认的先设为0.1。

后续调参在此基础上进行调整。

调整num_leaves

num_leaves：这是用于控制树模型的复杂度的主要参数。理论上设置num_leaves = 2^max_depth获得与树深度相同的叶子数，但实际上叶子数比树的深度更深，不受约束容易导致过拟合，所以通常设置num_leaves<2^max_depth。

当num_leaves设为[12, 31, 62, 81, 127, 256]时的输出结果如下：

参数设置：

```
{'num_thread': 4, 'num_leaves': 12, 'metric': 'binary',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 85
The minimum is attained in round 73
The minimum is attained in round 64
The minimum is attained in round 65
The minimum is attained in round 63
mean of error_rate :
0.08038000000000003
```

```
参数设置：
{'num_thread': 4, 'num_leaves': 31, 'metric': 'binary',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 70
The minimum is attained in round 51
The minimum is attained in round 58
The minimum is attained in round 51
The minimum is attained in round 48
mean of error_rate :
0.08132000000000002
参数设置：
{'num_thread': 4, 'num_leaves': 62, 'metric': 'binary',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 87
The minimum is attained in round 46
The minimum is attained in round 53
The minimum is attained in round 45
The minimum is attained in round 52
mean of error_rate :
0.08268
参数设置：
{'num_thread': 4, 'num_leaves': 81, 'metric': 'binary',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 72
The minimum is attained in round 36
The minimum is attained in round 61
The minimum is attained in round 39
The minimum is attained in round 46
mean of error_rate :
0.0826
参数设置：
{'num_thread': 4, 'num_leaves': 127, 'metric': 'binary',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 47
The minimum is attained in round 42
The minimum is attained in round 40
The minimum is attained in round 38
The minimum is attained in round 39
```

```
mean of error_rate :
0.08324000000000002
参数设置:
{'num_thread': 4, 'num_leaves': 256, 'metric': 'binary',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 49
The minimum is attained in round 34
The minimum is attained in round 42
The minimum is attained in round 34
The minimum is attained in round 35
mean of error_rate :
0.08427999999999998
```

当num_leaves=256时，误差率最好，后续设定num_leaves=256进行其他参数的调整。

调整metric

不同的统计方式，对统计结果会产生影响，所以对度量方法进行调整。

选取以下统计方式进行尝试：

l1

l1_loss最小绝对值偏差（LAD），把目标值与估计值的绝对值差的总和最小化。

l2

l2_loss最小平方误差（LSE），把目标值与估计值的差值的平方和最小化。

l1损失函数与l2损失函数的区别：

1. l1具有鲁棒性，l2不是非常的鲁棒。
2. l1是不稳定解，l2是稳定解。
3. l1可能存在多个解，l2总是一个解。

rmse

root square loss均方根误差，是真实因变量和预测因变量的差的平方根。

如果发现真实和预测的差异，可能会得到正值和负值。如果对该差异求和，该差异为零，这是无用的。所以做平方根。

如果存在连续因变量，使用rmse进行计算。

quantile

分位数回归是一种回归分析。在传统回归中，通过构建回归模型由自变量求出因变量的条件期望；而在分位数回归中，通过构建回归模型由自变量求出因变量的条件分位数。使用最小二乘法来寻找样本的中位数。

mape

mean_absolute_percentage_error平均绝对百分比误差，一种预测准确性的预测方法。通常用作回归问题和模型评估的损失函数，因为它在相对误差方面的解释非常直观。

binary

二分类方法。

当metric参数设为['l1', 'l2', 'rmse', 'quantile', 'mape', 'binary']，其输出结果如下：

```
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'l1', 'objective':
'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 862
The minimum is attained in round 735
The minimum is attained in round 783
The minimum is attained in round 980
The minimum is attained in round 798
mean of error_rate :
0.08452
```

```
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'l2', 'objective':
'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 39
The minimum is attained in round 32
The minimum is attained in round 27
The minimum is attained in round 34
The minimum is attained in round 30
mean of error_rate :
0.084
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'rmse',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 39
The minimum is attained in round 32
The minimum is attained in round 27
The minimum is attained in round 34
The minimum is attained in round 30
mean of error_rate :
0.084
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 862
The minimum is attained in round 684
The minimum is attained in round 813
The minimum is attained in round 986
The minimum is attained in round 798
mean of error_rate :
0.08461999999999999
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'mape',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 862
The minimum is attained in round 735
The minimum is attained in round 783
The minimum is attained in round 980
The minimum is attained in round 798
```

```
mean of error_rate :
0.08452
参数设置:
{'num_thread': 4, 'num_leaves': 256, 'metric': 'binary',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 49
The minimum is attained in round 34
The minimum is attained in round 42
The minimum is attained in round 34
The minimum is attained in round 35
mean of error_rate :
0.08427999999999998
```

根据结果，当metric=quantile时，准确率最好。

调整learning_rate

较高的学习速率是因为可以让收敛更快，但准确度没有较低的好。通过降低学习速率来训练数据，看可不可以进一步优化分数。

当learning_rate设为[0.1, 0.08, 0.06, 0.05, 0.0008, 0.005, 0.003]，其输出结果如下：

```
参数设置:
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 862
The minimum is attained in round 684
The minimum is attained in round 813
The minimum is attained in round 986
The minimum is attained in round 798
mean of error_rate :
0.08461999999999999
参数设置:
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.08,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 935
```

```
The minimum is attained in round 980
The minimum is attained in round 802
The minimum is attained in round 1001
The minimum is attained in round 879
mean of error_rate :
0.08423999999999998
参数设置:
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.06,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 980
The minimum is attained in round 995
The minimum is attained in round 795
The minimum is attained in round 1000
The minimum is attained in round 959
mean of error_rate :
0.08463999999999998
参数设置:
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.05,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 992
The minimum is attained in round 1000
The minimum is attained in round 968
The minimum is attained in round 1001
The minimum is attained in round 989
mean of error_rate :
0.08458000000000002
参数设置:
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate':
0.0008, 'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
mean of error_rate :
0.09766000000000001
参数设置:
```

```
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.005,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
mean of error_rate :
0.08292000000000002
参数设置:
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate': 0.003,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
mean of error_rate :
0.08317999999999999
```

根据结果，当learning_rate=0.0008时，误差率结果最好。

```
0.09766000000000001
```

再次调整num_round&num_leaves

根据上面的输出结果，发现learning_rate调整后，最小收敛值的获取基本落在1000轮了，故将num_round设为2000（同时调小num_leaves，将其设为137）再次重跑learning_rate参数。

其输出结果如下：

```
参数设置:
{'num_thread': 4, 'num_leaves': 137, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate': 0.1,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 1662
The minimum is attained in round 1730
```

```
The minimum is attained in round 1491
The minimum is attained in round 1991
The minimum is attained in round 1241
mean of error_rate :
0.08492
参数设置:
{'num_thread': 4, 'num_leaves': 137, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate': 0.08,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 1965
The minimum is attained in round 1986
The minimum is attained in round 1211
The minimum is attained in round 1985
The minimum is attained in round 1363
mean of error_rate :
0.08494000000000002
参数设置:
{'num_thread': 4, 'num_leaves': 137, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate': 0.06,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 1996
The minimum is attained in round 1984
The minimum is attained in round 1735
The minimum is attained in round 2000
The minimum is attained in round 1909
mean of error_rate :
0.08435999999999999
参数设置:
{'num_thread': 4, 'num_leaves': 137, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate': 0.05,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 1995
The minimum is attained in round 2001
The minimum is attained in round 1986
The minimum is attained in round 1992
The minimum is attained in round 2000
mean of error_rate :
0.08416000000000001
参数设置:
{'num_thread': 4, 'num_leaves': 137, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate':
0.0008, 'feature_fraction': 1.0, 'bagging_fraction': 0.8}
```

```
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
mean of error_rate :
0.08368000000000002
参数设置:
{'num_thread': 4, 'num_leaves': 137, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate': 0.005,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
mean of error_rate :
0.08348
参数设置:
{'num_thread': 4, 'num_leaves': 137, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate': 0.003,
'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
mean of error_rate :
0.08326000000000003
```

因为最好的误差率没有比之前的更好，故后续只将num_round=2000继续调整。

```
0.08494000000000002
```

调整feature_fraction

feature_fraction参数来进行特征的子抽样。这个参数可以用来防止过拟合及提高训练速度。如果这个参数设置<1.0，比如0.8，那么在训练树之前，模型会选择80%的特征子抽样。

将该参数设为[1.0, 0.8, 0.7, 0.5, 0.4, 0.3]进行调参，其输出结果如下：

```
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate':
0.0008, 'feature_fraction': 1.0, 'bagging_fraction': 0.8}
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
mean of error_rate :
0.08471999999999999
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate':
0.0008, 'feature_fraction': 0.8, 'bagging_fraction': 0.8}
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
mean of error_rate :
0.08514000000000002
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate':
0.0008, 'feature_fraction': 0.7, 'bagging_fraction': 0.8}
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
mean of error_rate :
0.08536000000000002
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate':
0.0008, 'feature_fraction': 0.5, 'bagging_fraction': 0.8}
The minimum is attained in round 2001
The minimum is attained in round 2001
```

```
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
mean of error_rate :
0.08628
参数设置:
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate':
0.0008, 'feature_fraction': 0.4, 'bagging_fraction': 0.8}
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
mean of error_rate :
0.08922000000000001
参数设置:
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 2000, 'learning_rate':
0.0008, 'feature_fraction': 0.3, 'bagging_fraction': 0.8}
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
The minimum is attained in round 2001
mean of error_rate :
0.09803999999999999
```

通过观察，当feature_fraction越来越小时，误差率结果越来越好。

```
0.09803999999999999
```

考虑到特征抽样不应过低，后续将feature_fraction=0.5进行调参。

调整bagging_fraction

bagging_fraction相当于subsample样本采样，类似于feature_fraction参数，可以使bagging更快的运行，同时也可以降拟合。

bagging_freq参数默认0，表示bagging的频率，0意味着没有使用bagging，k意味着每k轮迭代进行一次bagging。这里将该参数设为50，bagging_freq=50。

将bagging_fraction设为[1.0, 0.8, 0.75, 0.5, 0.4, 0.3]时，其输出结果如下：

```
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate':
0.0008, 'feature_fraction': 0.5, 'bagging_fraction': 1.0,
'bagging_freq': 50}
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
mean of error_rate :
0.12663999999999997
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate':
0.0008, 'feature_fraction': 0.5, 'bagging_fraction': 0.8,
'bagging_freq': 50}
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
mean of error_rate :
0.12728
参数设置：
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate':
0.0008, 'feature_fraction': 0.5, 'bagging_fraction': 0.75,
'bagging_freq': 50}
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
mean of error_rate :
0.12735999999999997
```

参数设置：
```
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate':
0.0008, 'feature_fraction': 0.5, 'bagging_fraction': 0.5,
'bagging_freq': 50}
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
mean of error_rate :
0.12892
```
参数设置：
```
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate':
0.0008, 'feature_fraction': 0.5, 'bagging_fraction': 0.4,
'bagging_freq': 50}
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
mean of error_rate :
0.12908000000000003
```
参数设置：
```
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate':
0.0008, 'feature_fraction': 0.5, 'bagging_fraction': 0.3,
'bagging_freq': 50}
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
mean of error_rate :
0.12998
```

最终版本参数的输出结果

参数设置：

```
{'num_thread': 4, 'num_leaves': 256, 'metric': 'quantile',
'objective': 'binary', 'num_round': 1000, 'learning_rate':
0.0008, 'feature_fraction': 0.5, 'bagging_fraction': 0.8}
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
The minimum is attained in round 1001
mean of error_rate :
0.12663999999999997
```

准确率为：

```
0.12663999999999997
```