

数据挖掘第二次项目

并行决策树集成

陈铭涛
16340024

May 21, 2019

1 CART 算法

CART 包含了分类决策树和回归决策树的算法，

2 Gradient Boosting

3 Random Forest

4 代码实现

出于内存和并行化实现的考虑，本次项目选择了使用 Rust 语言实现，Rust 的 RAII 机制使得临时资源可以及时地释放，提升内存利用率，由编译器提供的静态检查可以避免多线程时线程不安全的情况，降低 debug 难度。

程序实现中使用的第三方库如下：

1. rayon: 提供基于迭代器的便捷地编写并行代码的方法
2. rand: 提供随机数生成
3. csv: 提供对 csv 文件的读取
4. indicatif: 提供命令行进度条实现
5. ndarray: 提供类似 numpy 的多维数组的操作
6. num-traits: 提供数值类型上的一些实用方法，如最大最小值等
7. log: 程序日志
8. pretty_env_logger: 程序日志输出
9. num_cpus: 获取系统 CPU 核心数量

5 并行化表现

6 验证

以下程序测试均在一台搭载 6 核 12 线程 CPU，运行 macOS 系统的笔记本电脑上运行。

验证的标准为 R^2 ，其计算方法如下：

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i, \\ SS_{tot} &= \sum_i (y_i - \bar{y})^2, \\ SS_{res} &= \sum_i (y_i - f_i)^2, \\ R^2 &= 1 - \frac{SS_{res}}{SS_{tot}}\end{aligned}\tag{1}$$

其中 y_i 为第 i 个样本的实际观察值， f_i 为第 i 个样本的模型预测值。 R^2 的取值通常在 0 与 1 之间，越接近 1 代表预测值与真实值匹配程度越高。

使用 LightGBM 默认参数构建一个模型运行 3 折交叉验证进行对比：

```
In [6]: run_cross_validation(trains, labels, lgb.LGBMRegressor, params)

fit_time: [24.27433276 22.56340122 23.01793098]
Average fit_time: 23.285222
score_time: [10.66991425 9.20692468 8.52207994]
Average score_time: 9.466306
test_mse: [-0.33291203 -0.33487001 -0.34085721]
Average test_mse: -0.336213
train_mse: [-0.33630982 -0.33519583 -0.33240139]
Average train_mse: -0.334636
test_r2: [0.156102 0.15480654 0.15376245]
Average test_r2: 0.154890
train_r2: [0.15816958 0.15915648 0.15922097]
Average train_r2: 0.158849
```

Figure 1: LightGBM 交叉验证结果

对单棵决策树进行交叉验证获得的结果如下：

在训练集上获得的平均 R^2 分数为 0.14947528261278345

在验证集上获得的平均 R^2 分数为 0.143995380629807

训练时间平均为 682423 ms.

```

Finished release [optimized + debuginfo] target(s) in 0.84s
Running `target/release/cv`
INFO cv > Train data shape: [10000004, 13]
INFO cv > Label data shape: [1, 10000004]
INFO cv > Load time: 6919ms
INFO ensembles_rs::utils::cross_validate > train time: 712503, predict time: 295
INFO ensembles_rs::utils::cross_validate > train: 0.14879519294338295, validation: 0.144
71150292761092
INFO ensembles_rs::utils::cross_validate > train time: 658188, predict time: 254
INFO ensembles_rs::utils::cross_validate > train: 0.1489242194641912, validation: 0.1453
8125132877167
INFO ensembles_rs::utils::cross_validate > train time: 676580, predict time: 341
INFO ensembles_rs::utils::cross_validate > train: 0.15070643543077622, validation: 0.141
89338763303838
CrossValidateScore { train_time: [712503, 658188, 676580], predict_time: [295, 254, 341],
train_score: [0.14879519294338295, 0.1489242194641912, 0.15070643543077622], validation_sc
ore: [0.14471150292761092, 0.14538125132877167, 0.14189338763303838] }

```

Figure 2: 单棵决策树训练交叉验证结果

使用 Gradient Boosting 训练 150 步，设置单棵树最大生长至 2 层，进行交叉验证获得的结果如下：

在训练集上获得的平均 R^2 分数为 0.14595633826167811

在验证集上获得的平均 R^2 分数为 0.14329480642795986

训练时间平均为 223031 ms.

```

lr: 0.13
INFO ensembles_rs::boosting > lr 0.12 at step 149.
INFO ensembles_rs::boosting > Pred Score: 0.14566673131840036
lr: 0.12
INFO ensembles_rs::utils::cross_validate > train time: 217482, predict time: 767
INFO ensembles_rs::utils::cross_validate > train: 0.14566673131840036, validation: 0.145
5456685234573
CrossValidateScore { train_time: [220638, 230975, 217482], predict_time: [1013, 819, 767],
train_score: [0.147332346601316, 0.14486993686531802, 0.14566673131840036], validation_sc
ore: [0.1384225183769906, 0.14591623238343165, 0.1455456685234573] }
~/Desktop/junior-spring/mining/kaggle/assignment2/ensembles-rs master

```

Figure 3: GBDT 训练交叉验证结果

使用 Random Forest 训练，决策树数量为 150，不限制决策树生长深度，进行交叉验证获得的结果如下：

在训练集上获得的平均 R^2 分数为 0.1496125066152948

在验证集上获得的平均 R^2 分数为 0.14889900127096856

训练时间平均为 444804 ms.

```

INFO ensembles_rs::random_forest > score at step 146: 0.12003178060232811
INFO ensembles_rs::random_forest > score at step 147: 0.11941601849039751
INFO ensembles_rs::random_forest > score at step 148: 0.11914056387866856
INFO ensembles_rs::random_forest > score at step 149: 0.1054082041867459
INFO ensembles_rs::utils::cross_validate > train time: 442583, predict time: 6302
INFO ensembles_rs::utils::cross_validate > train: 0.1496643816056633, validation: 0.1491
2273384503627
CrossValidateScore { train_time: [446542, 445289, 442583], predict_time: [6959, 6392, 6302
], train_score: [0.14946317205633752, 0.14970996618388355, 0.1496643816056633], validation
_score: [0.14934094151491106, 0.14823332845295834, 0.14912273384503627] }
~/Desktop/junior-spring/mining/kaggle/assignment2/ensembles-rs master

```

Figure 4: 随机森林训练交叉验证结果

四个训练中使用取样工具查看内存占用值分别如下：

1. LightGBM: 5.2G

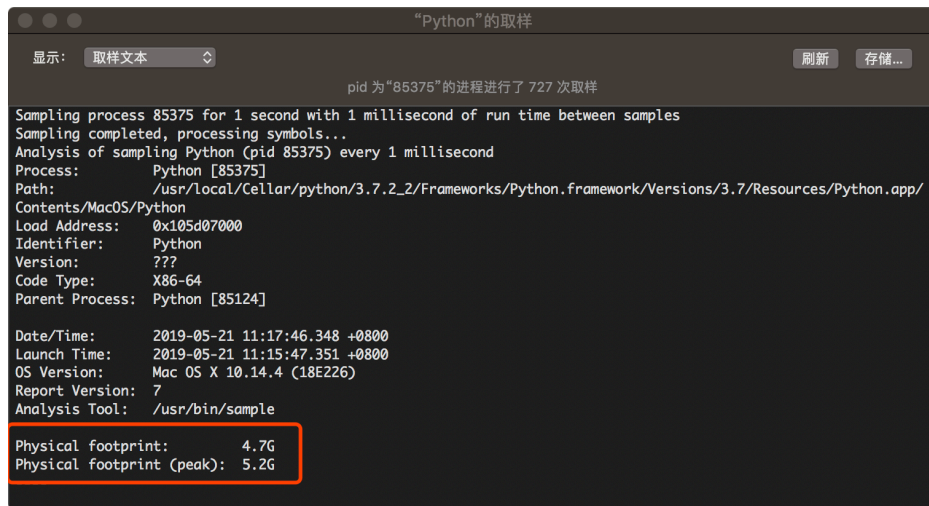
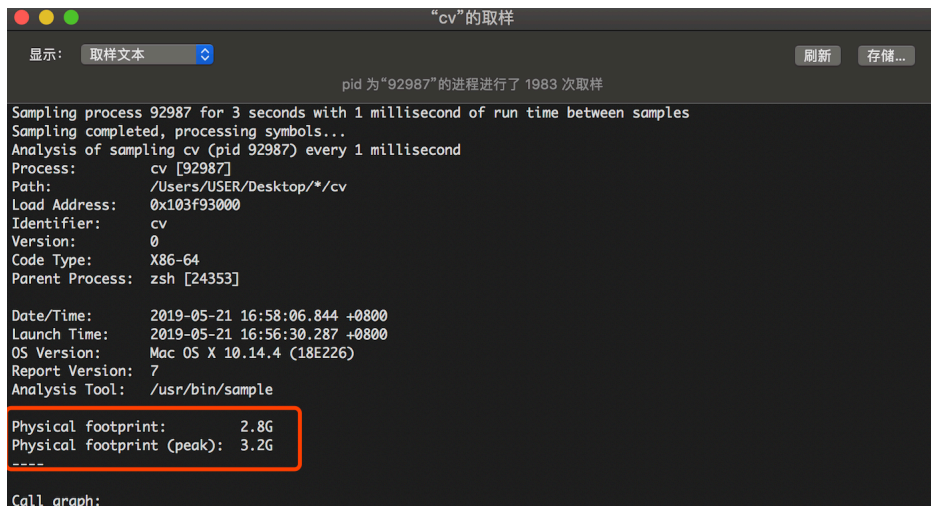


Figure 5: LightGBM 内存占用

2. 单决策树: 3.2G



```
显示: 取样文本 刷新 存储...

pid为"92987"的进程进行了 1983 次取样

Sampling process 92987 for 3 seconds with 1 millisecond of run time between samples
Sampling completed, processing symbols...
Analysis of sampling cv (pid 92987) every 1 millisecond
Process:      cv [92987]
Path:         /Users/USER/Desktop/*/cv
Load Address: 0x103f93000
Identifier:    cv
Version:      0
Code Type:    X86_64
Parent Process: zsh [24353]

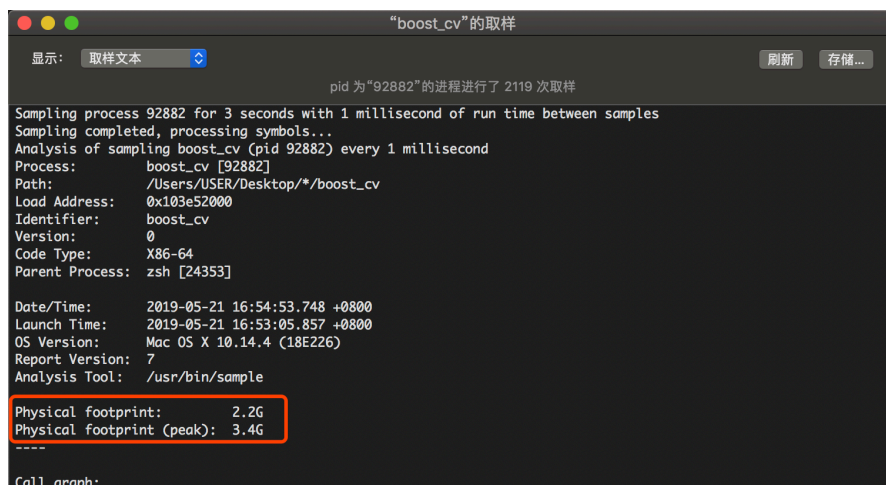
Date/Time:     2019-05-21 16:58:06.844 +0800
Launch Time:   2019-05-21 16:56:30.287 +0800
OS Version:    Mac OS X 10.14.4 (18E226)
Report Version: 7
Analysis Tool: /usr/bin/sample

Physical footprint: 2.8G
Physical footprint (peak): 3.2G
-----

Call graph:
```

Figure 6: 单棵决策树训练内存占用

3. Gradient Boosting: 3.4G



```
显示: 取样文本 刷新 存储...

pid为"92882"的进程进行了 2119 次取样

Sampling process 92882 for 3 seconds with 1 millisecond of run time between samples
Sampling completed, processing symbols...
Analysis of sampling boost_cv (pid 92882) every 1 millisecond
Process:      boost_cv [92882]
Path:         /Users/USER/Desktop/*/boost_cv
Load Address: 0x103e52000
Identifier:    boost_cv
Version:      0
Code Type:    X86_64
Parent Process: zsh [24353]

Date/Time:     2019-05-21 16:54:53.748 +0800
Launch Time:   2019-05-21 16:53:05.857 +0800
OS Version:    Mac OS X 10.14.4 (18E226)
Report Version: 7
Analysis Tool: /usr/bin/sample

Physical footprint: 2.2G
Physical footprint (peak): 3.4G
-----

Call graph:
```

Figure 7: Gradient Boosting 训练内存占用

4. Random Forest: 2.1G

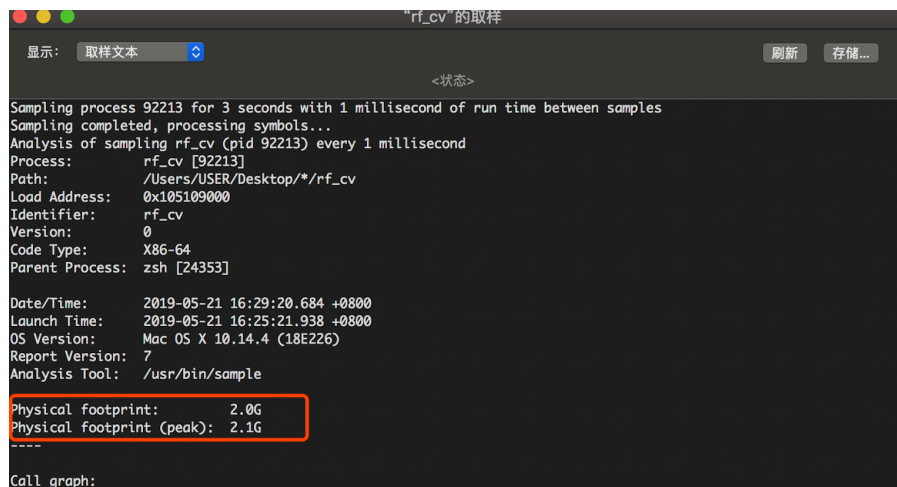


Figure 8: Random Forest 训练内存占用

7 Kaggle 分数

使用单棵决策树训练至 10 层后提交至 Kaggle 获得的分数为 0.16087:

Submission and Description	Public Score	Use for Final Score
decision-tree-10.csv 2 hours ago by Miguel Chan add submission details	0.16087	<input type="checkbox"/>

Figure 9: 单棵决策树分数

使用 Gradient Boosting, Learning Rate 固定为 0.25, 基学习器最大训练至 3 层, 训练步数为 400 时的分数为 0.16957:

GBDT-400-3-6-400-0.25-0.25.csv 17 hours ago by Miguel Chan add submission details	0.16957
---	---------

Figure 10: lr=0.25, GBDT