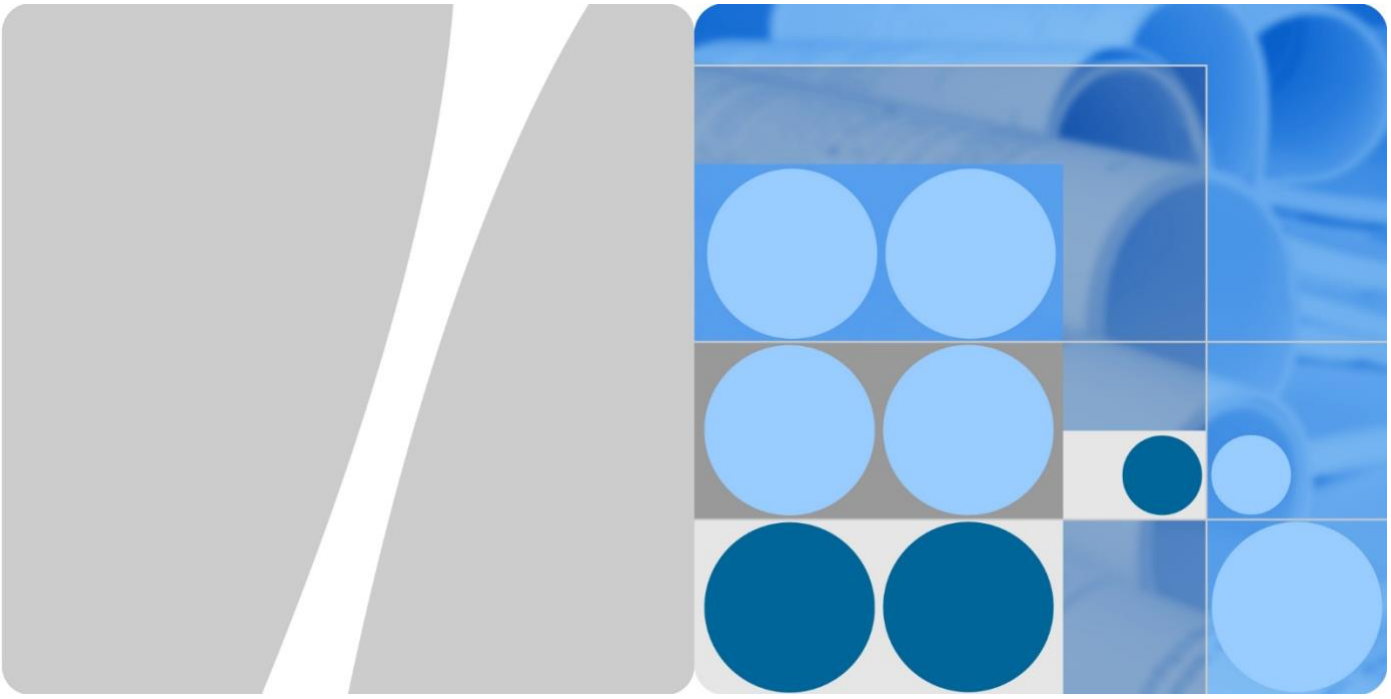

金融风控欺诈检测建模

热身赛题说明

发布日期　　2020-01-17

目 录

1 题目说明

机器学习是金融风控中使用到的核心技术之一。金融风控中，会结合各类特征，进行风险预估（常见的特征如：学历、性别、收入、负债情况、商品购买记录、历史逾期行为、人际社交等）。数据分析工程师会针对上述特征进行特征工程处理，再选用合适的机器学习模型进行数据建模工作。

在“大数据”的时代背景下，保证机器学习算法效果的同时，充分挖掘 IT 基础设施算力，提升算法计算性能，一方面有利于保护企业现有 IT 投资，另一方面能让数据分析师以更短的时间完成建模，从而可以选择出来更优化的业务模型。

在本次比赛中，我们准备了已经做好了特征工程处理的数据和对应的样例代码，您需要优化模型提升准确率和性能。

请您结合对机器学习算法的理解并结合鲲鹏处理器的特点对其进行优化，我们将会利用您的代码针对训练数据进行建模，并对测试数据进行预测。建模+预测要求在 15 分钟之内完成。模型准确率低于 70%不计分，高于 70%分为四档，最终执行时间最短者胜出。

- 准确率在[70%,80%)之间，其中最终执行时间为实际时间乘以 200%
- 准确率在[80%,90%)之间，其中最终执行时间为实际时间乘以 150%
- 准确率在[90%,95%)之间，其中最终执行时间为实际时间乘以 120%
- 准确率在>=95%，其中最终执行时间为实际时间乘以 100%

1.1 输入信息

输入分为三部分：

- train_data.txt 为已经做好特征工程处理的本地训练集文件。每一行为一条数据记录，以逗号分开。最后一列为类别（二分类），前面的列为特征值。
- test_data.txt 为需要预测的本地测试集文件。特征数和训练集一致。不含类别信息。
- 示例代码为准确率和性能待优化的参考代码，支持的语言分别为C++/Python/JAVA。

说明

answer.txt 为 test_data.txt 的二分类结果，用于练习的时候使用。

1.2 输出信息

输出信息为一个文件 result.txt，按行顺序放置测试集记录的预测结果，每一行代表一条训练数据的二分类结果。

1.3 限制条件

- 选手拿到的训练集和测试集并不是最终判题用的数据。
- 示例代码的算法实现为 LR（逻辑回归），选手可以将其改为其它的机器学习算法，但程序中定义的输入输出文件路径不能改。
- 不允许使用外部机器学习库。

1.4 排分规则

- 结果准确，最终执行时间（参考题目说明里的描述）最短者胜出。
- 如果最终执行时间一样，先提交的选手排名靠前。
- 选手成绩取个人多次提交里面的最好成绩。

1.5 其它说明

- 在解决大型机器学习问题时，一般会利用高级优化算法来加快梯度下降的计算过程。比如，除了梯度下降法，还可以用共轭梯度法、BFGS 或 L-BFGS 来加快运算速率。
- 可选基于鲲鹏 920 的特点（如：多核，NEON，Cache 大小）进行加速。
- C++语言的编译命令为：g++ -O3 main.cpp -o test -lpthread
- C++语言的运行命令为：./test
- Python 语言的运行命令为： python3 ./Main.py
- 只能使用 Python 标准库和 numpy，其中判题程序使用的是 numpy 1.17.2 版本。

numpy 1.17.2 在 Euler OS 上的安装方法为：

1. 安装 python3-devel

wget https://developer.huawei.com/ict/site-euleros/euleros/repo/yum/2.8/os/aarch64/updates/python3-devel-3.7.0-9.h15.eulerosv2r8.aarch64.rpm

rpm -ivh python3-devel-3.7.0-9.h15.eulerosv2r8.aarch64.rpm --nodeps

2. 修改 pip 的安装源

创建如下的文件~/.pip/pip.conf，并在文件中添加如下配置，保存并退出

[global]

index-url = https://repo.huaweicloud.com/repository/pypi/simple

trusted-host = repo.huaweicloud.com

timeout = 120

3. 安装 numpy 1.17.2

pip3 install numpy==1.17.2

- JAVA 语言的编译命令为：javac Main.java
- JAVA 语言的运行命令为：java Main
- JAVA 代码请使用 UTF-8 编码

2 运行环境

- 选手使用的练习资源：2U4G
- 判题系统使用的判题资源：4U8G
- 操作系统：Euler OS
- 服务器：TaiShan 服务器
- 芯片：鲲鹏 920