联邦学习笔记

目录

[1. 联邦学习概述 2](#_Toc42077813)

[1.1 目标 2](#_Toc42077814)

[1.2 3个研究方向：通信效率、隐私、安全 2](#_Toc42077815)

[1.3 联邦学习分类：纵向联邦学习，横向联邦学习，联邦迁移学习 3](#_Toc42077816)

[1.4 通信效率 3](#_Toc42077817)

[1.4.1 分布式学习——平行梯度下降 3](#_Toc42077818)

[1.4.2 联邦学习——联邦均值算法 3](#_Toc42077819)

[2. 纵向联邦学习系统框架 4](#_Toc42077820)

[3. 纵向联邦学习示例 5](#_Toc42077821)

[3.1 加密样本对齐——隐私保护下的样本ID匹配（RSA+Hash） 6](#_Toc42077822)

[3.2 加密模型训练过程（线性回归+同态加密） 7](#_Toc42077823)

[3.3 安全性分析 8](#_Toc42077824)

[4. 联邦学习中的加密算法 9](#_Toc42077825)

[4.1 同态加密 9](#_Toc42077826)

[4.1.1 分类 9](#_Toc42077827)

[4.1.2 同态加密在机器学习上的应用 9](#_Toc42077828)

[4.2 安全多方计算 10](#_Toc42077829)

[4.3 差分隐私 10](#_Toc42077830)

[4.4 秘密共享 11](#_Toc42077831)

[4.5 对比 11](#_Toc42077832)

[4.6 应用 11](#_Toc42077833)

[4.6.1 SecureML 11](#_Toc42077834)

[5. Ficus工具使用 12](#_Toc42077835)

[5.0 Ficus使用反馈 12](#_Toc42077836)

[5.1 数据集说明 13](#_Toc42077837)

[5.2 示例test01——onehot, 特征分箱，特征选择，逻辑回归 14](#_Toc42077838)

[5.3 示例test04——SecureBoost 16](#_Toc42077839)

[5.4 HeteroSecureBoost训练效果 16](#_Toc42077840)

[6. 联合建模过程 17](#_Toc42077841)

[6.1 数据集说明 17](#_Toc42077842)

[6.2 本地与联合建模效果对比 18](#_Toc42077843)

[6.2.1 HeteroLR 18](#_Toc42077844)

[6.2.2 HeteroSecureBoost 20](#_Toc42077845)

[6.3 HeteroLR参数选择 26](#_Toc42077846)

[6.4 HeteroSecureBoost参数选择 28](#_Toc42077847)

[7. 联邦学习知识点分享 32](#_Toc42077848)

[7.1 20200408 联邦学习基础知识点 32](#_Toc42077849)

[7.2 20200409 清华创业团队联邦学习交流 33](#_Toc42077850)

[7.3 20200423 中原银行&京东 联邦模盒 33](#_Toc42077851)

[7.5 20200423 联邦学习技术与应用课程 33](#_Toc42077852)

[7.6 20200507 HeteroSecureBoost讲解与Ficus建模训练 34](#_Toc42077853)

[8. HeteroSecureBoost原理 34](#_Toc42077854)

[8.1 SecureBoost框架 34](#_Toc42077855)

[8.2 非联邦设定中的树集成模型XGBoost 35](#_Toc42077856)

[8.3 联邦学习中的SecureBoost过程 36](#_Toc42077857)

[8.4 使用学习的模型对新实例进行分类 40](#_Toc42077858)

[9. kuai操作超白特征表建模 41](#_Toc42077859)

[9.1 银河工具学习 41](#_Toc42077860)

[9.2 建模 41](#_Toc42077861)

[10. 联邦模盒建模记录 44](#_Toc42077862)

[10.0 联邦模盒使用的注意事项 44](#_Toc42077863)

[10.1使用方法 44](#_Toc42077864)

[10.2 建模过程 46](#_Toc42077865)

[10.3 SecureBoost参数选择 48](#_Toc42077866)

# 1. 联邦学习概述

## 1.1 目标

保障在大数据交换时的信息安全、保护终端数据和个人隐私，保证合法合规的前提下，在多参与方或多计算节点之间开展高效率的机器学习。

联邦学习的优势在于数据隔离，数据不会泄露到外部，满足用户隐私保护和数据安全的需求；能够保证模型质量无损，不会出现负迁移，保证联邦模型比割裂的独立模型效果好；参与者地位对等，能够实现公平合作；能够保证参与各方在保持独立性的情况下，进行信息与模型参数的加密交换，并同时获得成长。

## 1.2 3个研究方向：通信效率、隐私、安全

通信：Federate Averaging Algorithm vs. Parallel gradient descent

FedAvg通过牺牲计算，换取通信效率

（重点）**隐私**：联邦学习不是传原始数据，但是参数中包含了一些数据特征，例如最小二乘回归，传递参数同时也造成了信息泄露。可以根据梯度，反向推导出部分数据特征

现有的方法：加噪声（精度降低），随机矩阵变换

安全（鲁棒性）：拜占庭将军问题（攻击 防御）

## 1.3 联邦学习分类：纵向联邦学习，横向联邦学习，联邦迁移学习

纵向联邦学习：用户重叠多而特征重叠少，例如某地的银行和电商

横向联邦学习：用户重叠少而特征重叠多，例如两个银行

联邦迁移学习：用户和特征都重叠少，例如中国的银行和美国的电商

## 1.4 通信效率

### 1.4.1 分布式学习——平行梯度下降

**i-th worker:**

1. receiving model parameter w from server
2. using w and local data to compute gradient
3. sending to server

**Server:**

1. receiving gradients ,…, from all the m workers
2. computing
3. updating model parameter:

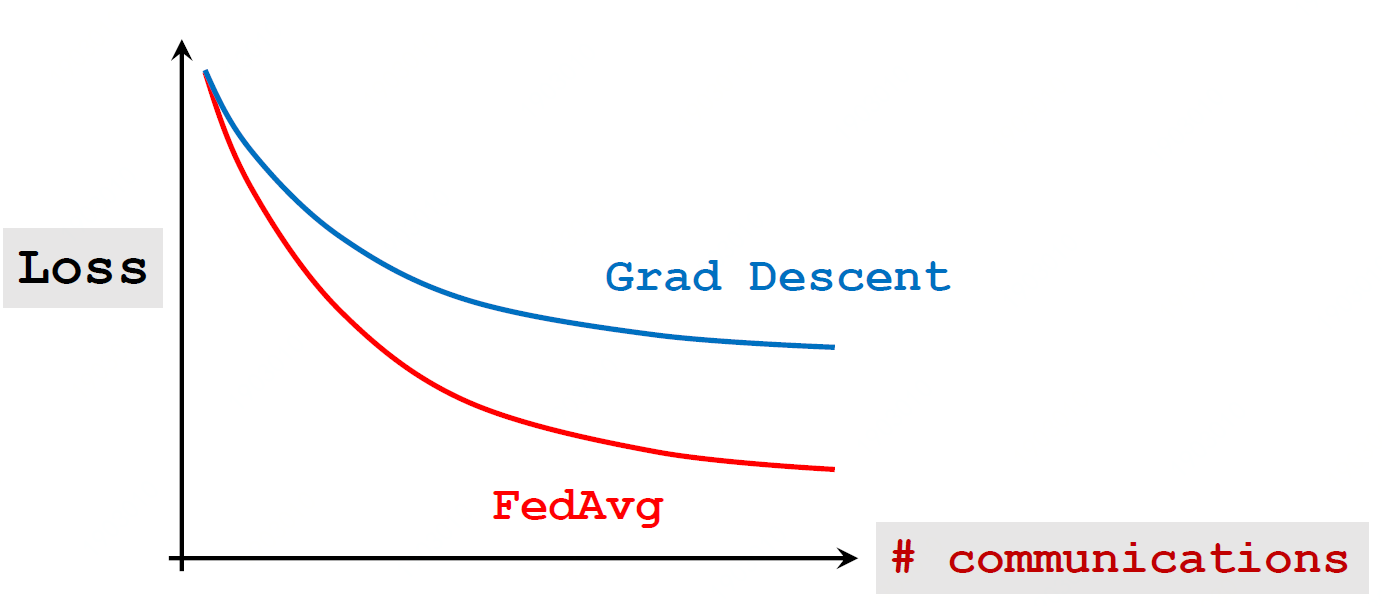
### 1.4.2 联邦学习——联邦均值算法

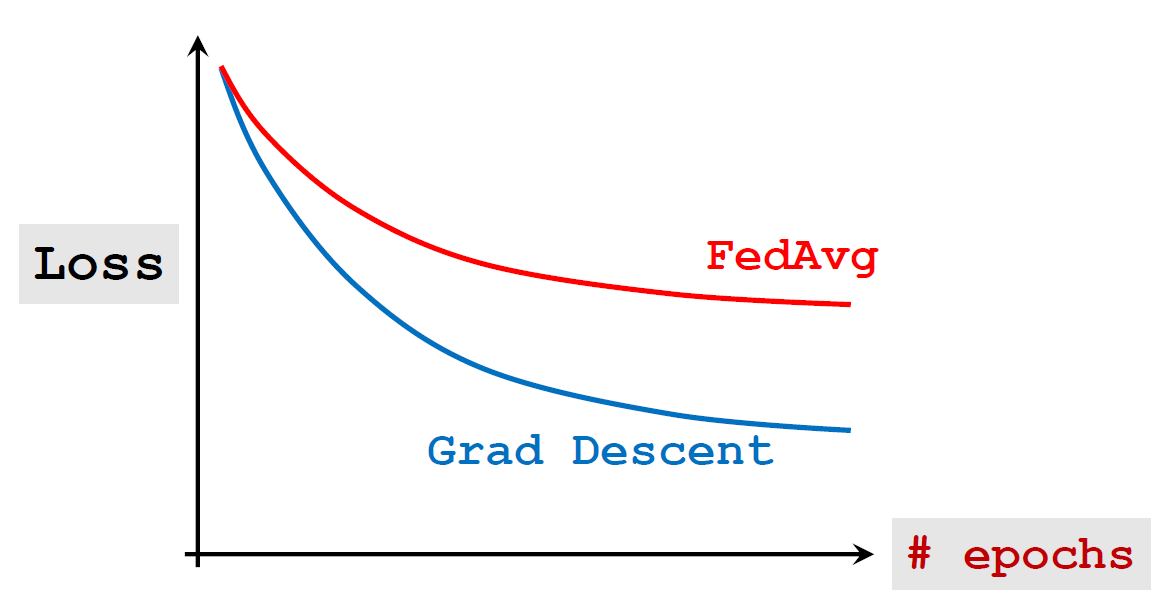
**i-th worker:**

1. receiving model parameter w from server
2. repeating:
   1. using w and local data to compute gradient g
   2. local updating:
3. sending to server

**Server:**

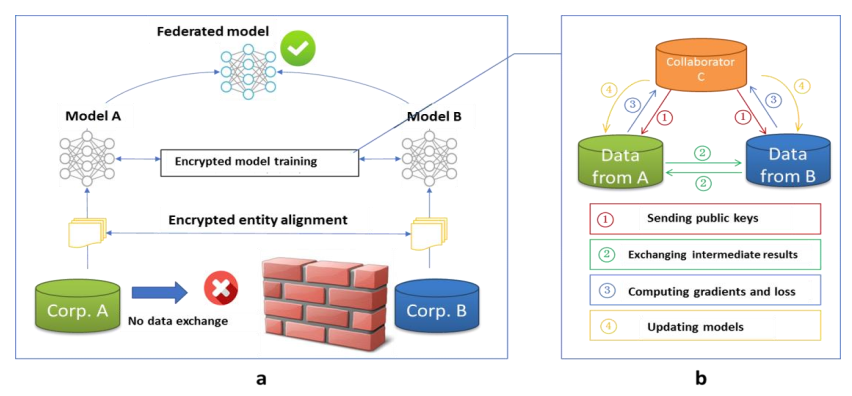
1. receiving ,…, from all the m workers
2. updating model parameter:





结论：联邦学习牺牲计算，换取通信效率

# 2. 纵向联邦学习系统框架



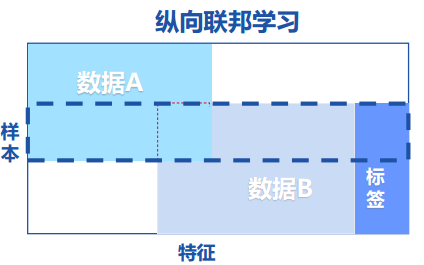
第一部分：**加密样本对齐**。由于两家企业的用户群体并非完全重合，系统利用基于加密的用户样本对齐技术，在 A 和 B 不公开各自数据的前提下确认双方的共有用户，并且不暴露不互相重叠的用户，以便联合这些用户的特征进行建模。

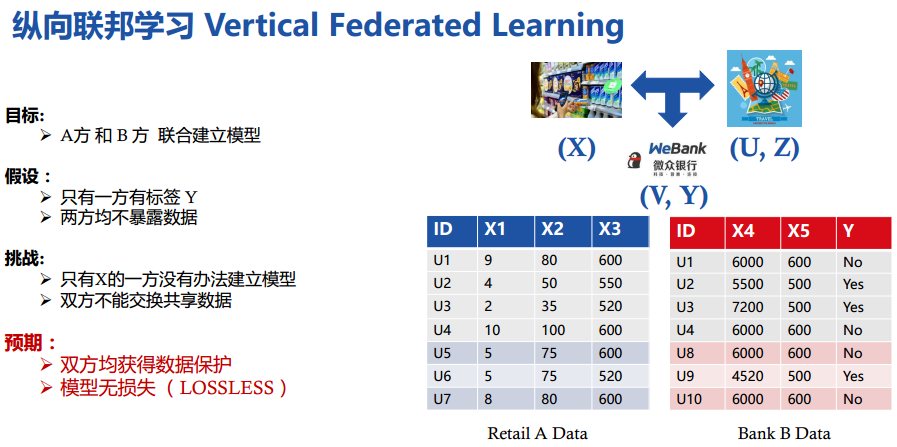
第二部分：**加密模型训练**。在确定共有用户群体后，就可以利用这些数据训练机器学习模型。为了保证训练过程中数据的保密性，需要借助第三方协作者 C 进行加密训练。以线性回归模型为例，训练过程可分为以下 4 步（如图b所示）：

1. 协作者 C 把公钥分发给 A 和 B，用以对训练过程中需要交换的数据进行加密。
2. A 和 B 之间以加密形式交互用于计算梯度的中间结果。
3. A 和 B 分别基于加密的梯度值进行计算，同时 B 根据其标签数据计算损失，并把结果汇总给 C。C 通过汇总结果计算总梯度值并将其解密。
4. C 将解密后的梯度分别回传给 A 和 B，A 和 B 根据梯度更新各自模型的参数。

      迭代上述步骤直至损失函数收敛，这样就完成了整个训练过程。在样本对齐及模型训练过程中，A 和 B 各自的数据均保留在本地，且训练中的数据交互也不会导致数据隐私泄露。因此，双方在联邦学习的帮助下得以实现合作训练模型。

# 3. 纵向联邦学习示例

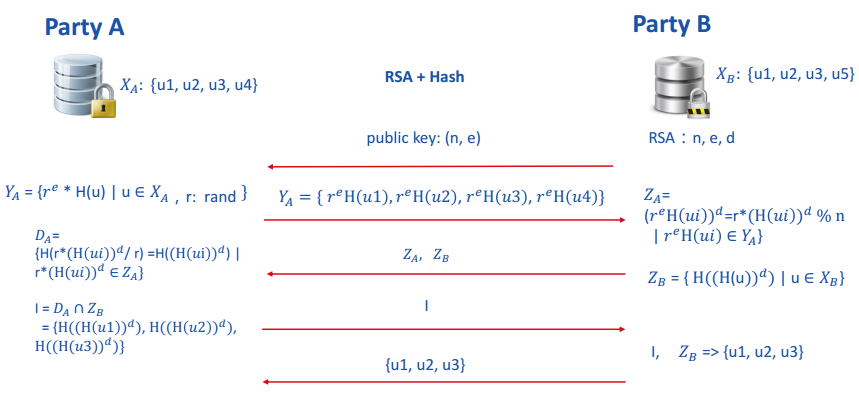




A方和B方拥有较多相同用户，但双方需要在不暴露各自原始数据的情况下联合训练模型，其中A含有特征，B含有特征和分类标签Y。

## 3.1 加密样本对齐——隐私保护下的样本ID匹配（RSA+Hash）

首先，需要确定A方和B方的共同用户ID.使用隐私保护下的样本ID匹配，如下图所示



A方包含用户，B方包含用户，目标得到共同用户ID:.这里使用RSA+Hash加密的方法：

1. B方RSA加密算法产生公共模数n，公钥e，私钥d。将公钥(n, e)传给A方
2. A方用户ID数据进行RSA+Hash加密：
3. B方解密A方传来的数据：

并对自身用户ID数据RSA+Hash解密后再哈希：

将和传给A方

1. A方对进行RSA+Hash解密：

计算交集：

将交集I传给B方

1. B方进行Hash还原得到共同用户ID: 并传给A方

## 3.2 加密模型训练过程（线性回归+同态加密）

设置一个可信的C方，提供加密解密的相关参数。

训练过程使用线性回归和同态加密方法，假设学习率，正则化参数，数据集，，初始化模型参数，则目标函数：

其中，令，则加密后的损失函数：

表示同态加密，同时记以下3个符号：

则加密后的损失函数记作：

同时，令，计算加密后的梯度：

训练过程如下表所示

|  |  |  |  |
| --- | --- | --- | --- |
| 步骤 | A方 | B方 | C方 |
| 1 | 初始化 | 初始化 | 创建密钥对，把公钥传给A方和B方 |
| 2 | 计算并传给B方 | 计算，，，，，并把传给A方，把传给C方 |  |
| 3 | 初始化，计算并传给C方 | 初始化，计算并传给C方 | 解密得到L，将传给A方，将传给B方 |
| 4 | 更新，重复234步操作 | 更新，重复234步操作 |  |

## 3.3 安全性分析

* 第三方C的安全性
* C方所有的学习对象都是对加密处理后的梯度，保证了加密后矩阵的随机性和隐私
* 这里假设第三方C是可信赖的，但也有数据泄露的风险，所以可以不用第三方，直接A和B交换加密数据，以效率会降低来换取更高的安全性
* A B双方之间的安全性
  + A方每一步学习梯度的过程，不足以得到B方的任何信息
  + 无法求解未知数个数多于方程数的方程
* 半诚实的对手威胁模型的设定保证安全性
* 半诚实的设定中，对手至多破坏2个数据客户端中的1个，余下数据还可以继续学习

# 4. 联邦学习中的加密算法

## 4.1 同态加密

同态加密（homomorphic encryption, HE）能够在不解密的情况下，对密文数据进行计算。无需破坏敏感源数据，同时可以对数据进行处理。使得其他人可以对加密数据进行处理，但是处理过程不会泄露任何原始内容。同时，拥有密钥的用户对处理过的数据进行解密后，得到的正好是处理后的结果。

### 4.1.1 分类

加法同态加密：

RSA算法

乘法同态加密：

Paillier算法

全同态加密：同时满足加法同态和乘法同态

Gentry算法

### 4.1.2 同态加密在机器学习上的应用

* 多项式近似

在的泰勒展开：

* 加密计算

逻辑回归的损失函数近似为

加密损失函数：

## 4.2 安全多方计算

安全多方计算（Secure Multi-Party Computation: MPC）解决一组互不信任的参与方之间保护隐私的协同计算问题，确保输入的独立性、计算的正确性、去中心化等特征，同时不泄露各种输入值给参与计算的其他成员。

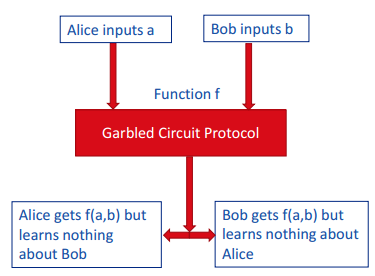
优点：

* 保证信息层面的数据安全
* 零知识证明
* 需要多方参与

缺点：

* 大量信息传输
* 降低数据安全可以提高效率

混淆电路（Yao’s Garbled Circuit Protocol）



## 4.3 差分隐私

差分隐私（differential privacy）采用了一种随机机制，使得当改变输入中的单个样本之后，输出的分布不会有太大的改变。或者说，对于差别只有一条记录的两个数据集，查询他们获得相同的输入的概率非常接近。

他要解决的问题是为了保护这样一种场景：早上查询医院流感人数700，朋友Alice早上去医院看病，下午查询医院流感人数为701。那么就可以推测Alice很大可能去医院看了流感。差分隐私方法：公布流感的人数 = 准确人数 + 噪声

在这些方法中，差分隐私由于其更强的加密保护、更简洁的算法、更小的系统开销，所以得到更广泛的应用。

## 4.4 秘密共享

秘密共享（secret sharing）中(k,n)门限秘密共享表示把秘密信息分成n份无意义的子秘密，只有拥有至少k份子秘密才能恢复秘密信息。

例如，(4,4)门限秘密共享，将秘密信息转成01字符串s，随机生成3个与s长度相同的01字符串s1, s2, s3，将s与s1,s2,s3异或，得到s4，即，然后将4个秘密信息s1,s2,s3,s4分配给4个人，只有4个人的时候才能解得

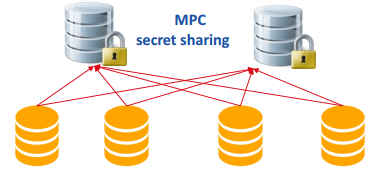
## 4.5 对比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 同态加密 | 秘密共享 | 混淆电路 | 差分隐私 |
| 计算成本 | 低 | 高 | 高 | 低 |
| 传输成本 | 高 | 低 | 低 | 高 |
| 需要第三方 | 不需要 | 需要 | 不需要 | 不需要 |
| 安全性 | 计算安全 | 信息理论安全 | 计算安全 | 计算安全 |

## 4.6 应用

### 4.6.1 SecureML

SecureML: MPC + secret sharing



[MZ17]

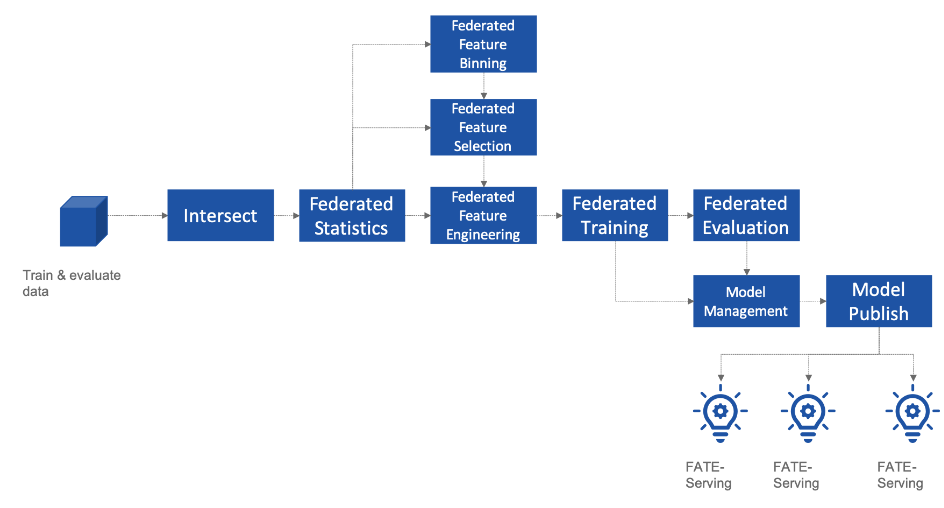
在线：MPC（混淆电路） + secret sharing 输出最后的模型

离线：服务在线阶段的乘法运算（同态加密，不经意传输），用客户端辅助三元组

特点：计算安全，但开销大

# 5. Ficus工具使用

FATE-Flow联邦学习Pipeline



## 5.0 Ficus使用反馈

1. 无法自主新建领域集

状态：已解决

2. 新建领域集时上传文件大小限制在10MB之内，但实际用到的数据量级很大（10w\*1000， 10w\*100）

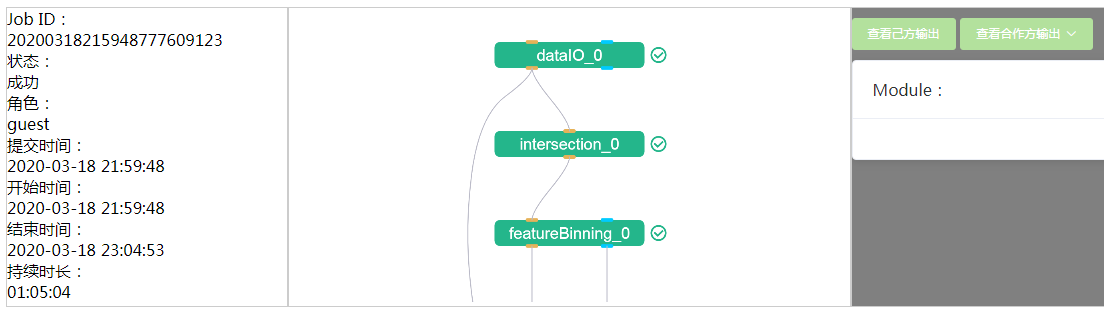
状态：已反馈

解决方法：后台上传大文件

3. Evaluation模块显示需要重复刷新几次

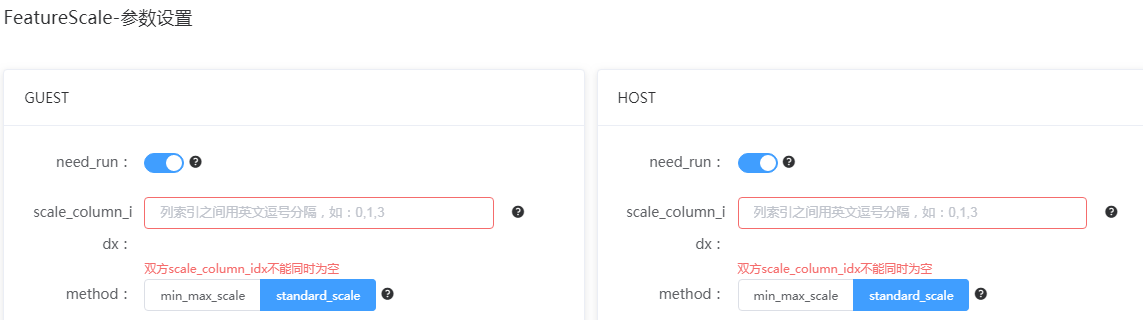
状态：已反馈

4. 作业查看时显示不全



5. ID对齐只处理数值

6. FeatureScale对所有特征缩放无法选择



7. 增加预测及结果输出模块

状态：已反馈（近期上线）

8. 会增加类似网格调参的功能吗？

状态：已反馈

解决方法：短期不会，改用spark计算任务会快很多

9. 如何证明数据没有出库

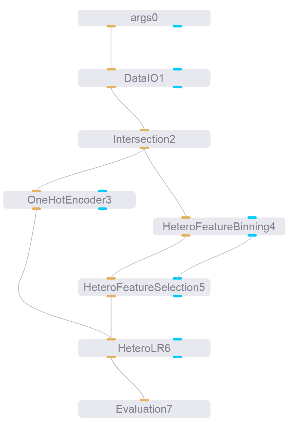
10. 连接数据库的方式

## 5.1 数据集说明

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 说明 | 我方（含y） | 其他参与方（不含Y） |
| 0 |  | 乳腺癌Y | 乳腺癌X |
| 数据规模 | (500+)\*3 | (500+)\*31 |
| 特征 | id, x0, y | id, x0~29 |
| ID对齐后数据量（重合率） | 501(0.935) |  |
| 1 |  | data\_small\_1 | data\_small\_2 |
| 数据规模 | 100000\*24 | 35000\*38 |
| 特征 | id, a0~5, a107~122, y | id, b0~20, a107~122 |
| ID对齐后数据量（重合率） | 7107(0.07107) |  |
| 2 |  | data12 | data22 |
| 数据规模 | 100000\*125 | 35000\*364 |
| 特征 | id, a0~122, y | id, b0~346, a107~122 |
| ID对齐后数据量（重合率） | 7107(0.07107) |  |
| 3 |  | data32\_10w\_part2 | data42\_10w\_part2 |
|  | 数据规模 | 3w\*125 | 3w\*553 |
|  | ID对齐后数据量（重合率） | 2.26w(0.749) |  |
| 4 |  | data32\_10w\_part3 | data42\_20w\_part3 |
|  | 数据规模 | 1w\*125 | 1w\*553 |
|  | ID对齐后数据量（重合率） | 2486(0.2472) |  |

## 5.2 示例test01——onehot, 特征分箱，特征选择，逻辑回归

登录http://ficus-panel-pro.jdfmgt.com/，创建模型test01，使用乳腺癌数据集（数据集0），分host和guest两端联合建模训练



args0：原始数据

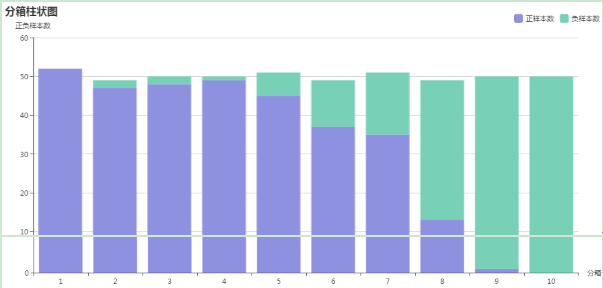
|  |  |  |
| --- | --- | --- |
| 角色 | 参与方 | 数据集（领域集） |
| guest | 京东数科银河 | 乳腺癌Y |
| host | 京东数科数资 | 乳腺癌X |

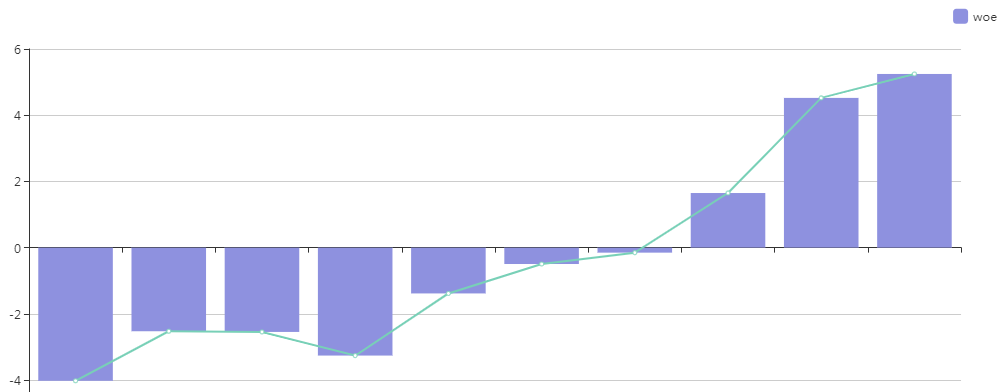
DataIO1：读取数据

Intersection2：ID对齐，输出双方ID相同的数据

OneHotEncoder3：对ID对齐后的数据进行onehot编码

HeteroFeatureBinning4：等频分箱，分成10组。以下是host中的分箱柱状图：

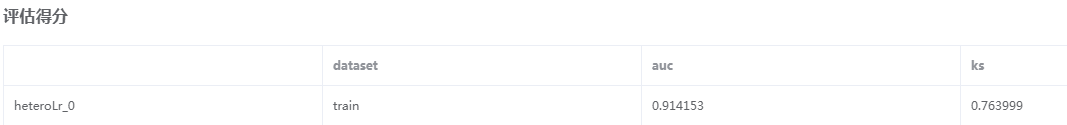




HeteroFeatureSelection5：特征选择

HeteroLR6：对数据进行HeteroLR异步逻辑回归训练

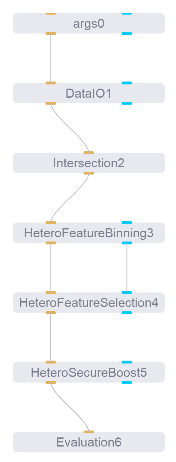
Evaluation7：评估训练结果如下图所示，auc为0.914，ks为0.764



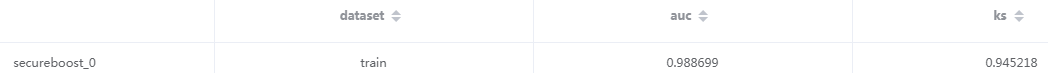
将guest和host两端的数据合并，即传统的机器学习方法，用逻辑回归对乳腺癌数据分类，训练得到的模型auc为0.917，与test01联邦学习训练得到的分类精度基本相同，可以证明Ficus工具在联邦学习中既保证模型的分类效果，又保证了用户的数据安全，不会造成隐私泄露。

## 5.3 示例test04——SecureBoost

对示例test01中的乳腺癌数据使用HeteroSecureBoost进行联合建模



训练结果如下图，auc为0.988，ks为0.945



## 5.4 HeteroSecureBoost训练效果

对于联合建模的分类效果，可查看第6章

Ficus优化spark后，对于6.4节的实验0，模型的训练速度提升了：

|  |  |  |
| --- | --- | --- |
| 实验 | 优化前用时 | 优化后用时 |
| 0 | 2h | 28min |

# 6. 联合建模过程

## 6.1 数据集说明

100000\*125 35000\*364 intersection(id):7000

300000\*126 300000\*554 intersection(phone):179494

小样本

|  |  |  |  |
| --- | --- | --- | --- |
|  | GUEST | HOST | intersection |
| 领域集名称 | data12 | data22 |  |
| 数据规模 | 100000\*24 | 35000\*38 | 7117 |
| 特征 | id, a0~5, a107~122, y | id, b0~20, a107~122 |  |

大样本

|  |  |  |  |
| --- | --- | --- | --- |
|  | GUEST（含y） | HOST（不含y） | intersection |
| 领域集 | data32 | data42 |  |
| 0 | 298919\*123 | 298821\*552 | 179494 |
| 1 | 5w | 10w | 9989 |
| 2 | 5w | 15w | 14988 |
| 3 | 5w | 20w | 20111 |
| 4 | 5w | 25w | 25188 |
| 5 | 5w | 30w | 30051 |
| 6 | 10w | 20w | 40221 |
| 7 | 10w | 30w | 60032 |
| 8 | 20w | 20w | 80501 |
| 9 | 20w | 30w | 120156 |

## 6.2 本地与联合建模效果对比

### 6.2.1 HeteroLR

实验一

在小样本集上，单方（本地）对ID重叠样本（7k）和测试样本（9w）使用LogisticRegression分类，结果：

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 测试 |
| accuracy | 0.7640 | 0.7518 |
| auc | 0.5262 | 0.5231 |

本地对全量数据（10w）分为训练集（7w）和测试集（3w）使用LogisticRegression分类，结果：

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 测试 |
| accuracy | 0.7572 | 0.7592 |
| auc | 0.5281 | 0.5268 |

实验二

选择特征较少（9个特征）的样本（训练集3w，验证集1w），使用LogisticRegression分类，结果如下：

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 测试 |
| accuracy | 0.7456 | 0.7495 |
| auc | 0.5 | 0.4998 |

以上面这一方为基础（作为GUEST方），增加特征（作为HOST方，10个特征），数据规模如下：

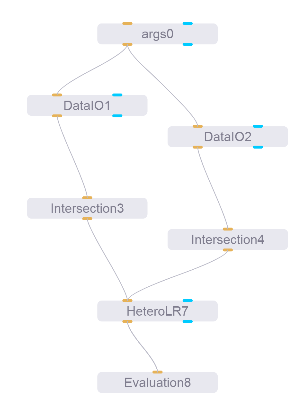
|  |  |  |  |
| --- | --- | --- | --- |
|  | GUEST | HOST | intersection |
| 训练集规模 | 3w\*9 | 3w\*10 | 2.6w |
| 验证集规模 | 1w | 1w | 0.25w |

联合建模实验结果如下：

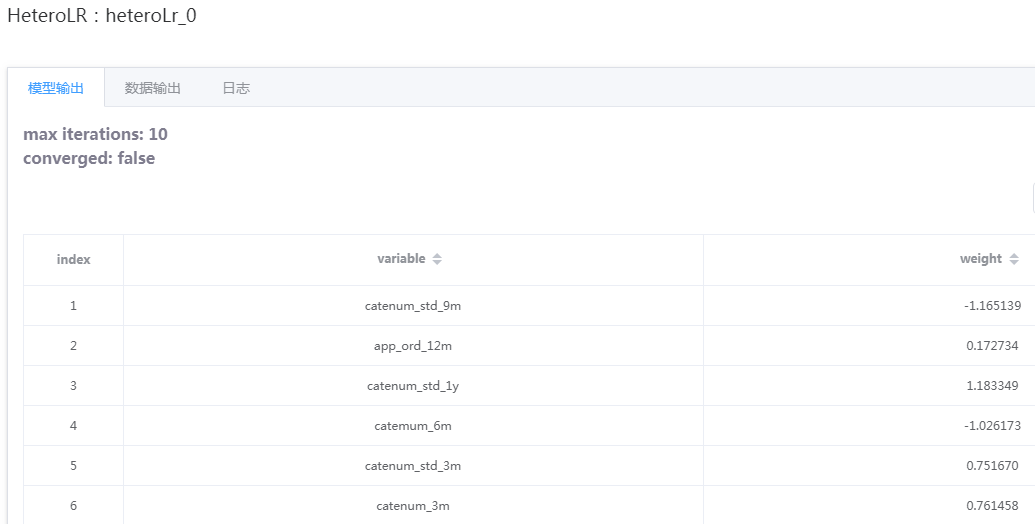
|  |  |  |
| --- | --- | --- |
|  | 训练集 | 测试 |
| accuracy | 0.7463 | 0.7595 |
| auc | 0.5181 | 0.5214 |

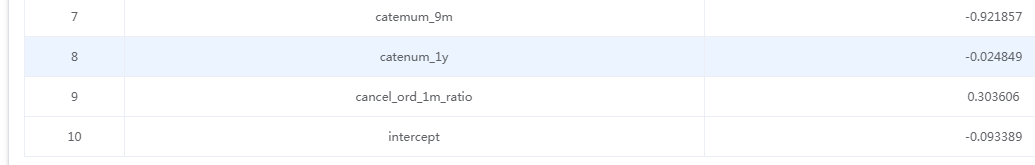
从实验结果可以看出，HeteroLR可以保持联合建模的效果，没有较大提升的原因可能是缺少数据预处理，以及线性模型的效果有限

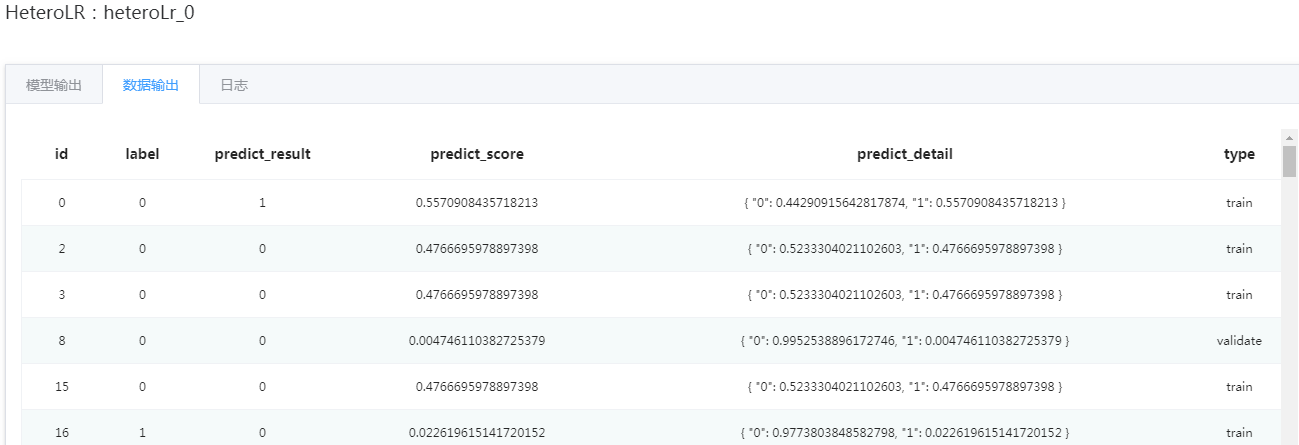
下面是HeteroLR的模型和数据输出：



其中，HeteroLR的模型输出和数据输出如下：







### 6.2.2 HeteroSecureBoost

实验一

在小样本集上，单方（本地）对ID重叠样本（7k）和测试样本（9w）使用xgboost分类，结果：

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 测试 |
| accuracy | 0.7837 | 0.7709 |
| auc | 0.5673 | 0.5617 |

本地对全量数据（10w）分为训练集（7w）和测试集（3w）使用xgboost分类，结果：

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 测试 |
| accuracy | 0.7764 | 0.7774 |
| auc | 0.5729 | 0.5709 |

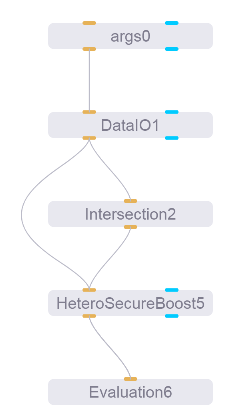
联合建模时，GUEST方（10w）和HOST方（3.5w）对ID重叠样本（7k）和测试样本（10w）使用HeteroSecureBoost分类（特征选择用xgboost），结果：

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 测试 |
| accuracy | 0.7775 | 0.8351 |
| auc | 0.6820 | 0.8408 |

实验二

改变领域集，分别在6.1节提出的大样本集（intersection规模：1w, 1.5w, 2w, 2.5w, 3w, 4w, 6w, 8w, 12w, 18w）上使用HeteroSecureBoost分类

下图表示HeteroSecureBoost的工作流程：



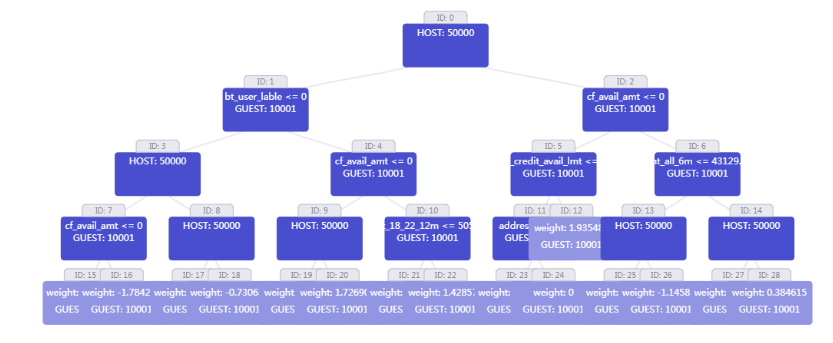
1. 数据读取（DataIO）

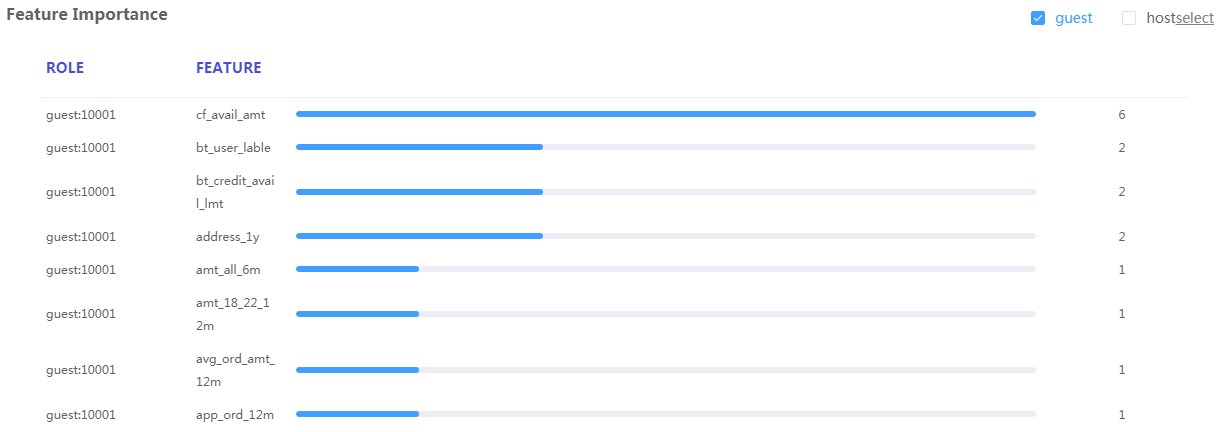
2. ID对齐（Intersection）



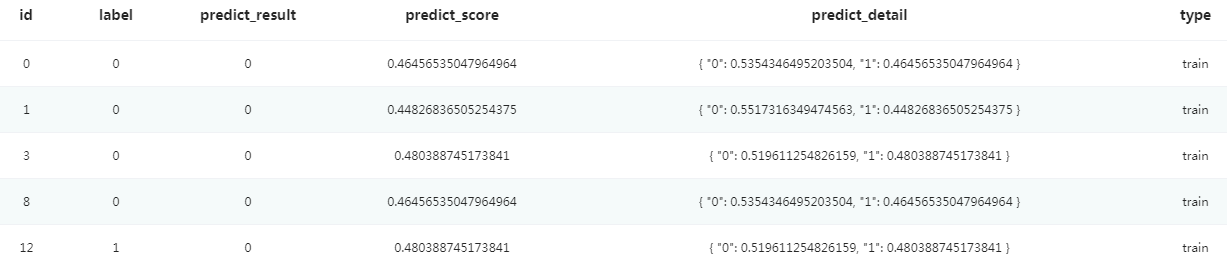
3. 模型训练：训练集是对齐样本（Intersection），测试集是读入的原始数据（DataIO）

树模型模型输出：





数据输出：



4. 效果评估

在不同量级的样本上训练的分类结果：

|  |  |  |
| --- | --- | --- |
| 领域集 | accuracy | auc |
| 1 | 0.7776 | 0.6935 |
| 2 | 0.7755 | 0.6880 |
| 3 | 0.7745 | 0.6884 |
| 4 | 0.7754 | 0.6895 |
| 5 | 0.7729 | 0.6878 |

从分类效果来看，GUEST方与HOST方联合建模的分类准确度比单方本地建模更高，这是在保证数据安全的情况下，增加训练样本的特征数，提高了GUEST方的分类精度，又避免了双方的隐私泄露问题。

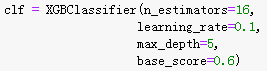
实验三

选择交叉样本较多的数据集来做训练和验证

单方xgboost分类：

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 验证集 |
| 样本量 | 3w | 1w |
| 特征数 | 123 |  |
| accuracy | 0.7772 | 0.5693 |
| auc | 0.5709 | 0.6123 |

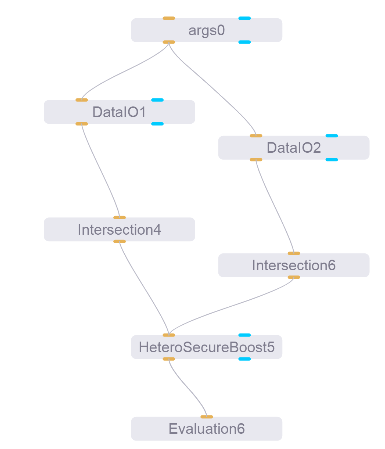
调参后，选择合适的参数，训练结果如下：



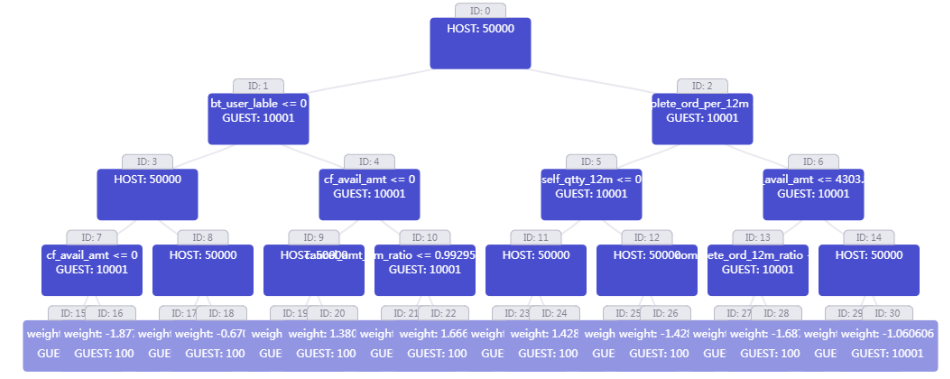
|  |  |  |
| --- | --- | --- |
|  | 训练集 | 验证集 |
| accuracy | 0.7767 | 0.7788 |
| auc | 0.5705 | 0.5710 |

对以上数据（GUEST方）增加特征（HOST方），进行联合建模

|  |  |  |  |
| --- | --- | --- | --- |
|  | GUEST | HOST | Intersection |
| 训练集数据量 | 3w | 3w | 2.26w |
| 验证集数据量 | 1w | 1w | 0.25w |



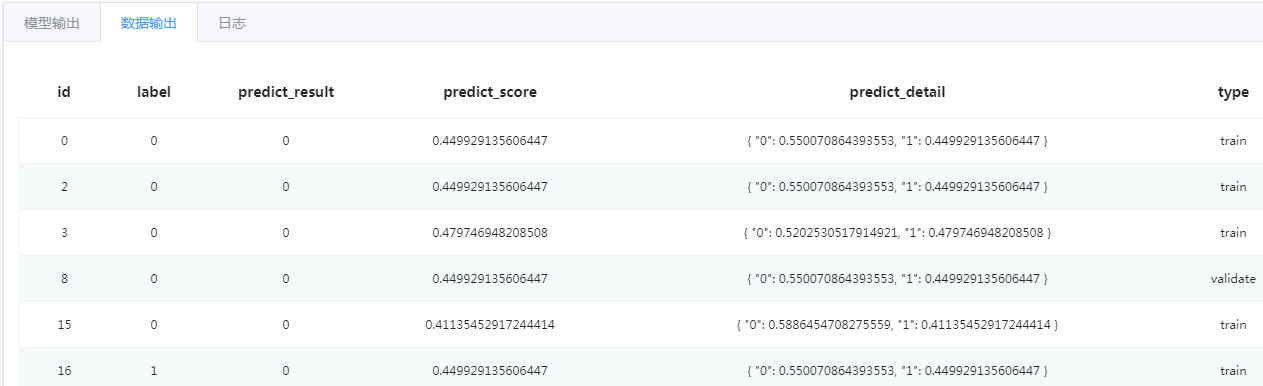
HeteroSecureBoost的模型输出和数据输出





从模型的结构来看，单方建模训练后树模型特征的重要性排序是cf\_avail\_amt, bt\_user\_label, complete\_ord\_per\_12m, cancel\_amt\_3m\_ratio, 双方联合建模后训练出来的特征按重要性排序是cf\_avail\_amt, ord\_per\_12m, bt\_user\_label, complete\_ord\_per\_12m, jdmall\_jdmordr\_f0002261, mother\_baby\_amount\_6m\_ratio, top1\_qtty\_amount\_3m, cancel\_amt\_3m\_ratio, 字段对应的中文含义如下，这也符合之前的建模逻辑：

|  |  |  |
| --- | --- | --- |
| GUEST/HOST | 特征 | 含义 |
| GUEST | cf\_avail\_amt | 白条、金条可用额度 |
| HOST | ord\_per\_12m | 最近12个月订单平均金额 |
| GUEST | bt\_user\_label | 白条账户类别 |
| GUEST | complete\_ord\_per\_12m | 最近12个月完成订单平均金额 |
| GUEST | jdmail\_jdmordr\_f0002261 | 商城末单距今天数 |
| HOST | mother\_baby\_amount\_6m\_ratio | 最近6个月购买母婴金额占最近6个月总完成金额比 |
| HOST | top1\_qtty\_amount\_3m | 最近3个月总购买件数Top1品类实付总金额 |
| GUEST | cancel\_amt\_3m\_ratio | 最近3个月取消金额占最近3个月总完成金额比 |



在训练集和验证集上的分类效果：

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 验证集 |
| accuracy | 0.7726 | 0.78 |
| auc | 0.6815 | 0.6747 |

可见，联合建模提高了训练效果，保证了数据安全性。

## 6.3 HeteroLR参数选择

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 参数名 | 含义 | 取值 |
| 0 | method | 惩罚函数 | L1/L2 |
| 1 | early\_stop | 收敛判断方法 | diff/abs/weight\_dif |
| 2 | tol | 收敛精度 |  |
| 3 | alpha | 正则化系数 |  |
| 4 | optimizer | 优化算法 | sgd/rmsprop/adam/adagrad/nesterov\_momentum\_sgd/sqn |
| 5 | batch\_size | 批尺寸 |  |
| 6 | learning\_rate | 学习率 |  |
| 7 | max\_iter | 最大迭代次数 |  |
| 8 | decay | 学习率的衰减率 |  |
| 9 | decay\_sqrt | 学习率开平方根 |  |
| 10 | multi\_class | 多分类策略 |  |
| 11 | validation\_freqs | 验证频率 |  |
| 12 | init\_method | 初始化方法 | random\_uniform/random\_normal/ones/zeros/const |
| 13 | init\_const | 初始化固定值 |  |
| 14 | fit\_intercept | 是否初始化截距 |  |
| 15 | init\_random\_seed | 初始化随机种子 |  |
| 16 | threshold | 分类阈值 |  |
| 17 | need\_cv | 是否交叉验证 |  |
| 18 | n\_splits | 交叉重数 |  |
| 19 | shuffle | 是否随机洗牌 |  |
| 20 | cv\_random\_seed | 交叉验证随机种子 |  |
| 21 | update\_interval\_l | 更新Hession矩阵迭代次数 |  |
| 22 | memory\_m | 曲率块大小 |  |
| 23 | sample\_size | 更新Hession矩阵样本数 |  |
| 24 | sqn\_random\_seed | SQN随机种子 |  |

为方便测试，选择只包含少数特征的训练集和验证集

|  |  |  |
| --- | --- | --- |
| 本地 | 训练集 | 验证集 |
| 样本量 | 3w | 1w |
| 特征数 | 8 |  |
| 具体特征 | 'cancel\_ord\_1m\_ratio', 'catenum\_std\_9m', 'app\_ord\_12m', 'catenum\_std\_1y', 'catenum\_3m', 'catenum\_1y', 'catemum\_9m', 'catemum\_6m', 'catenum\_std\_3m' |  |
| accuracy | 0.7466 | 0.7495 |
| auc | 0.5 | 0.4998 |

将以上特征作为GUEST方，增加HOST方

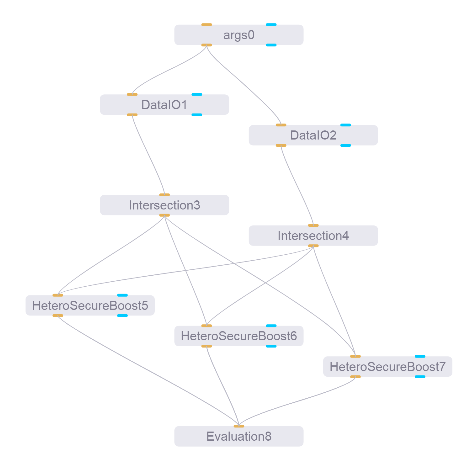
|  |  |  |
| --- | --- | --- |
|  | GUEST | HOST |
| 训练集样本量 | 3w | 3w |
| 训练集ID对齐后样本量 | 2.26w |  |
| 验证集样本量 | 1w | 1w |
| 验证集ID对齐后样本量 | 2486 |  |
| 特征数 | 8 | 10 |
| 具体特征 | 'cancel\_ord\_1m\_ratio', 'catenum\_std\_9m', 'app\_ord\_12m', 'catenum\_std\_1y', 'catenum\_3m', 'catenum\_1y', 'catemum\_9m', 'catemum\_6m', 'catenum\_std\_3m' | 'discount\_amt\_6m\_ratio', 'discount\_amt\_1m\_ratio', 'discount\_amt\_12m\_ratio', 'discount\_amt\_3m\_ratio', 'off\_ord\_6m\_ratio', 'off\_ord\_12m\_ratio', 'off\_ord\_3m\_ratio', 'prc50\_amount\_12m\_ratio', 'prc50\_amount\_6m\_ratio', 'top3\_qtty\_nsku\_3m' |

## 6.4 HeteroSecureBoost参数选择

以下是HeteroSecureBoost的可选参数

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 参数名 | 含义 | 取值 |
| 0 | task\_type | 任务类型 | 分类/回归 |
| 1 | learning\_rate | 学习率 |  |
| 2 | num\_trees | 最大树数 |  |
| 3 | subsample\_feature\_rate | 子采样特征 |  |
| 4 | n\_iter\_no\_change | 是否启用tol | 是/否 |
| 5 | tol | 收敛精度 |  |
| 6 | bin\_num | 分段数 |  |
| 7 | use\_missing | 是否使用缺失值 | 是/否 |
| 8 | zero\_as\_missing | 是否把0当成缺失值 | 是/否 |
| 9 | validation\_freqs | 验证频率 |  |
| 10 | threshold | 分类阈值 |  |
| 11 | need\_cv | 是否交叉验证 | 是/否 |
| 12 | n\_splits | 交叉重数 |  |
| 13 | shuffle | 是否随机洗牌 | 是/否 |
| 14 | cv\_random\_seed | 交叉验证随机种子 |  |
| 15 | criterion\_method | 特征选择标准 | xgboost |
| 16 | criterion\_params | 标准参数 |  |
| 17 | max\_depth | 决策树最大深度 |  |
| 18 | min\_sample\_split | 内部节点再划分所需最小样本数 |  |
| 19 | min\_impurity\_split | 最小不纯阈值 |  |
| 20 | min\_leaf\_node | 最小叶子节点数 |  |
| 21 | max\_split\_nodes | 最大分割点数 |  |
| 22 | feature\_importance\_type | 特征重要性计算类型 |  |
| 23 | objective | 目标函数 | cross\_entropy |
| 24 | params | 目标函数参数 |  |

模型是test35系列，点看查看HeteroSecureBoost的具体参数设置



我们以上一节的HeteroSecureBoost模型为baseline，相同的GUEST和HOST领域集，调整参数，进行以下实验，查看模型分类结果的accuracy和auc

|  |  |  |
| --- | --- | --- |
| 实验序号 | JobID | 参数设置 |
| 0 | 14501963755853 | learning\_rate=0.1,  num\_trees=2,  bin\_num=50,  max\_depth=5,  threshold=0.5  feature\_importance\_type=split |
| 1 | 10350721679455 | 同0, num\_trees=5 |
| 2 | 16092142209963 | 同0, bin\_num=70 |
| 3 | 12142152094664 | 同0, bin\_num=30 |
| 4 | 18453596500970 | 同0，  feature\_importance\_type=gain |
| 5 | 18463442349771 | 同0，  启用tol=0.0001 |
| 6 | 8392894046472 | 同0，  max\_depth=3 |
| 7 | 8441826313073 | 同0，  threshold=0.6 |
| 8 | 9070795814977 | 同0，  need\_cv=True, n\_splits=5 |

实验对应的分类结果：

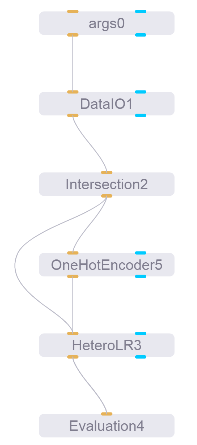
|  |  |  |
| --- | --- | --- |
| 实验序号 | accuracy（训练集/验证集） | auc（训练集/验证集） |
| 0 | 0.7726/0.78 | 0.6815/0.6746 |
| 1 | 0.7726/0.7796 | 0.7011/0.6916 |
| 2 | 0.7727/0.7796 | 0.6813/0.6733 |
| 3 | 0.7726/0.7804 | 0.6803/0.6713 |
| 4 | 0.7726/0.78 | 0.6813/0.6747 |
| 5 | 0.7726/0.78 | 0.6813/0.6747 |
| 6 | 0.7693/0.7784 | 0.6320/0.6268 |
| 7 | 0.7726/0.78 | 0.6813/0.6747 |
| 8 | 0.7725/0.7765 | 0.6848/0.6784 |

对于乳腺癌数据使用LogisticRegression在本地直接建模，找到合适的正则化系数C

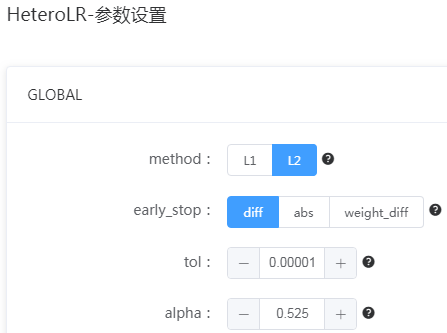


得到L2正则化和坐标轴下降法的最佳正则化系数为0.525

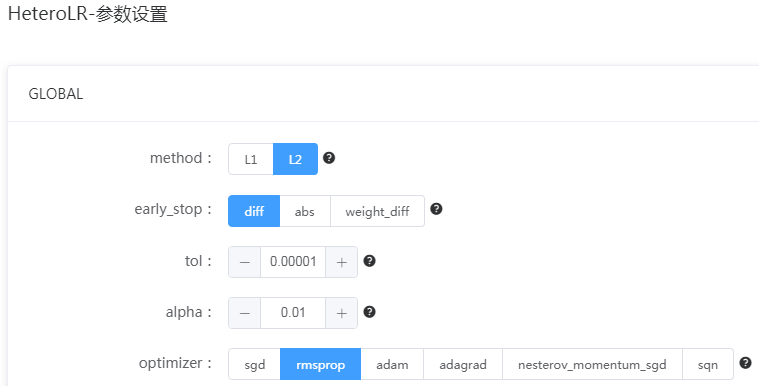
在Ficus工具中，对于模型test20



将正则化系数alpha置为0.525，模型在训练集上的auc从0.514提高为0.552



改变优化方法，从sgd换成rmsprop，训练集的auc从0.52变为0.55



# 7. 联邦学习知识点分享

## 7.1 20200408 联邦学习基础知识点

参考：ZRobot联邦学习建模手册0408.pptx

涉及内容：

1. 联邦学习的目标、研究方向、分类

2. 加密算法，主要是同态加密

3. 纵向联邦学习的过程：加密样本对齐，加密模型训练

4. 原理：线性回归+同态加密

5. 安全性分析

讨论：

1. 数据出库问题

2. 对数据加密：列名脱敏，原始数据缩放

3. 对ID重叠的样本建模，模型适用于具有所有特征的样本

4. 代码开源

## 7.2 20200409 清华创业团队联邦学习交流

技术不太成熟

## 7.3 20200423 中原银行&京东 联邦模盒

会议重点：

1. 训练模盒，推理模盒

分离：硬件要求不同；训练在生产环境，隐私级别高，数据量大，计算资源多，带宽高

2. 开放47个特征

消费画像：金额，次数

金融画像：白条

3. 加密安全

线性回归和同台加密的模型训练过程

4. 模盒优势

阿里云的硬件，传数据密文，私钥在云端；模盒传加密后的梯度

5. 产品演示

6. 中原业务落地场景：反欺诈、信贷、营销

## 7.5 20200423 联邦学习技术与应用课程

主要内容：

1. 联邦模盒（目前联邦学习的发展概况）

2. 联邦学习中的密码学

## 7.6 20200507 HeteroSecureBoost讲解与Ficus建模训练

主要内容：

1. SecureBoost框架及公式推导

2. Ficus建模训练过程，介绍每个模块的使用方法及作用

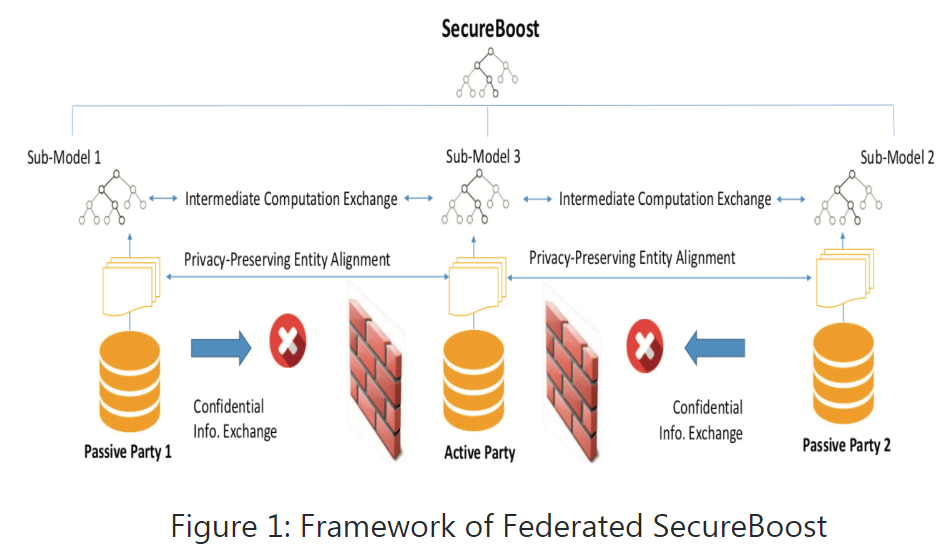
# 8. HeteroSecureBoost原理

参考：https://github.com/FederatedAI/FATE/blob/master/federatedml/tree/README.md

SecureBoost A Lossless Federated Learning Framework

## 8.1 SecureBoost框架

下图是SecureBoost的框架



Active Party（主动方）：我们将活动方定义为同时拥有数据矩阵和类别标签的数据提供者。 由于班级标签信息对于监督学习必不可少，因此必须有一个活跃的参与者可以访问标签y。 活跃的一方自然会在联合学习中担当主导服务器的角色。

Passive Party（被动方）：我们将只有一个数据矩阵的数据提供者定义为被动方。 被动方在联合学习环境中扮演客户的角色。 他们还需要建立一个模型来预测类别标签y，以用于其预测目的。 因此，他们必须与主动方合作以建立他们的模型，以便使用自己的功能为未来的用户预测y。

SecureBoost分为两步：

（1）它在隐私约束下对齐数据。

（2）它可以协作学习共享的梯度树提升模型，同时将所有训练数据在多个私人团体中保密。

数据对齐后，（2）中联邦学习的一般步骤：

（1）每个客户端都从服务器下载当前的全局模型。

（2）每个客户端根据其本地数据计算更新的模型和当前的全局模型，该模型驻留在活动方中。

（3）每个客户端都将加密后的模型更新发送回服务器。

（4）服务器汇总这些模型更新并构建改进的全局模型。

## 8.2 非联邦设定中的树集成模型XGBoost

给定数据集（n个样本，d个特征），XGBoost用K个回归树预测输出：

为学习回归树模型，采用贪心的方法在第t次迭代增加树，最小化损失函数：

其中，，，

在第t次迭代中构造回归树时，它从深度为0的树开始，并为树的每个叶节点添加一个拆分，直到达到最大深度。使用以下方程式确定最佳分割：

其中，和是在拆分后的左右子树节点。将分数最高的分割区选作最佳分割区。

当获得最佳树形结构时，叶j的最佳权重可通过以下公式计算：

其中，是叶j的实例空间

从上面的公式可以得出：

（1）候选分裂的评估和计算叶的最佳权重仅依赖于和。

（2）计算和需要类标签

例如，当我们将逻辑损失作为损失函数，此时的和如下：

因此，一旦我们获得，就能通过和得到类标签

## 8.3 联邦学习中的SecureBoost过程

二分类问题的损失函数：

最佳分割需要满足：

其中，和是在拆分后的左右子树节点

利用加法同态加密，最佳分割可以通过以下方式找到：

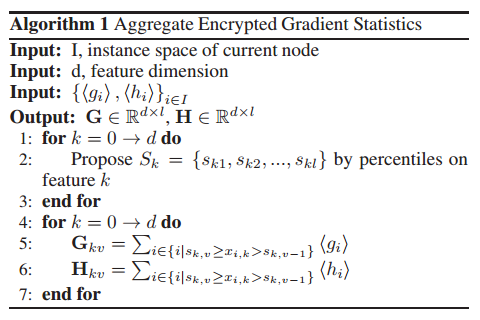
（1）每个被动方为本地所有可能的拆分计算和

（2）被动方将值发送回主动方

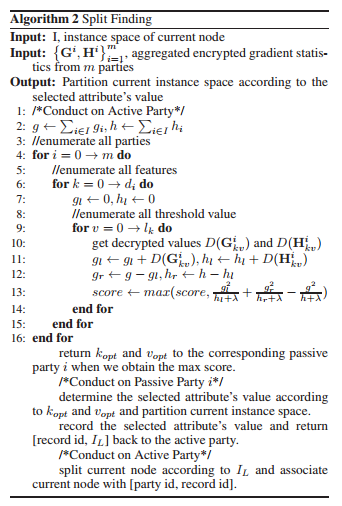
（3）在从所有被动方收集值之后，主动方对和进行解密，并计算全局最优分割

在这种情况下，主动方和每个被动方之间的通信成本对于单个拆分为. 此处，ct表示密文的大小，n表示与要拆分的节点关联的实例数，d表示被动方保留的特征数。

该解决方案效率不高，因为它要求将和转移给所有可能的拆分候选对象。为了构建通信成本较低的树，我们利用（Chen and Guestrin 2016）提出的近似框架，其详细计算：算法1

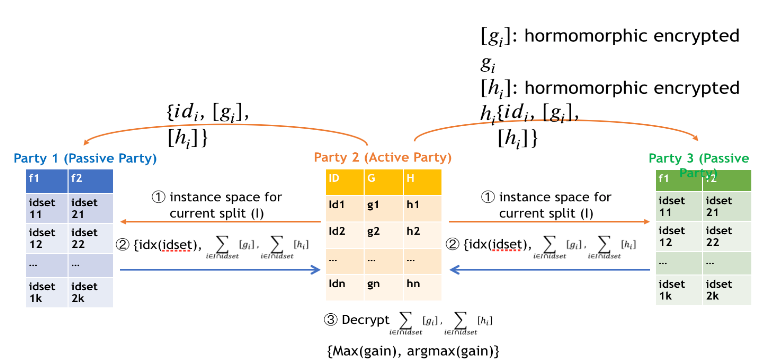


对于每个被动方，它不直接计算和，而是将特征映射到存储桶中，然后基于存储桶聚合加密的梯度统计信息。这样，主动方只需要从所有被动方收集汇总的加密梯度统计信息。然后，如算法2中所述确定全局最佳拆分：



在这种情况下，构建回归树的通信成本可以降低到，其中q表示一个存储桶中的实例数。显然，我们有（1 / q）<<1。因此，我们确实可以降低通信成本。

主动方获得全局最优拆分（）后，将k和v返回给相应的被动方i。被动方i根据k和v确定所选属性的值。然后，它根据所选属性的值对当前实例空间进行划分。此外，它会在本地构建一个查找表，以记录所选属性的值。然后，它返回记录的索引和分割后左节点的实例空间（）返回给主动方。主动方根据接收到的实例空间拆分当前节点，并将当前节点与关联，直到停止标准或达到最大深度。所有叶子节点都存储在主动方内部。



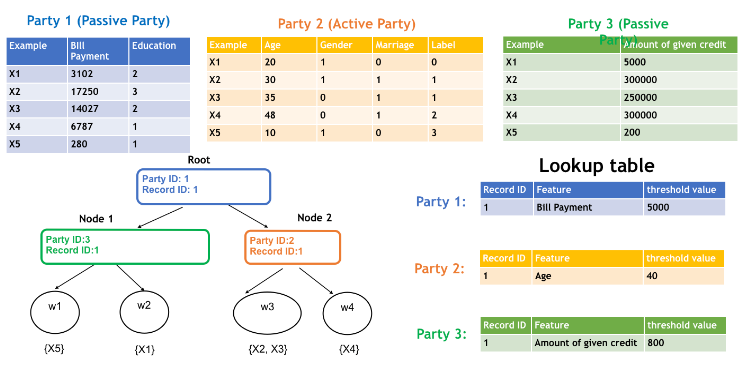
加密训练过程：

（1）主动方会将加密的[[gradient]]和[[hessian]]发送给被动方

（2）被动方使用[[gradient]]和[[hessian]]计算加密的特征直方图，然后对（feature，split\_bin\_val）进行编码并构造一个（feature，split\_bin\_val）查找表，将（feature，split\_bin\_val）的编码值与特征直方图一起发送给主动方

（3）主动方收到被动方的特征直方图后，将其解密并获得最佳收益，如果特征属于被动方，则将编码（feature，split\_bin\_val）发送回被动方。

各方继续进行拆分查找过程，直到完成构建树为止。各方仅知道拆分特征所属的树节点的详细拆分信息。下图显示了单个决策树的最终结构。



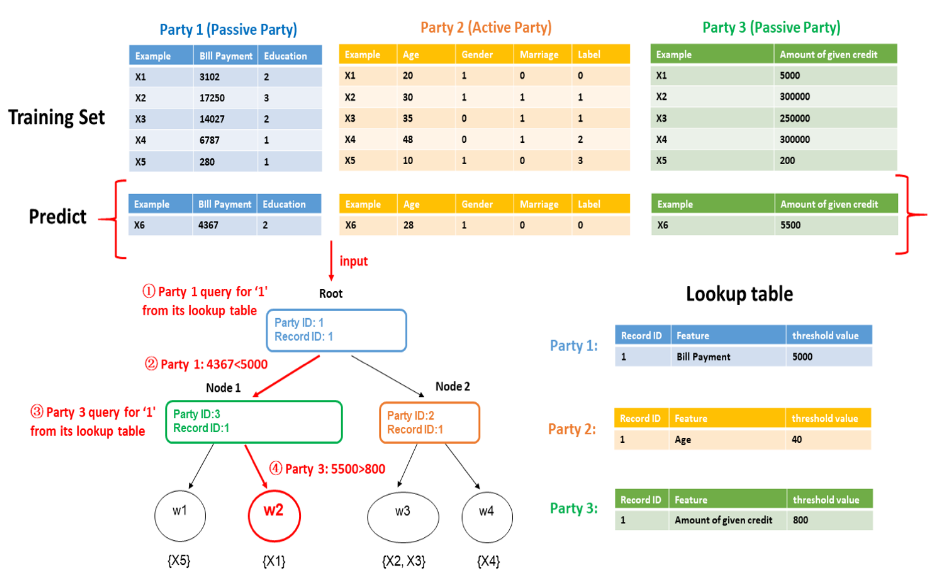
## 8.4 使用学习的模型对新实例进行分类

主动方判断当前树节点是否属于它。

（1）如果当前树属于主动方，则它可以使用其（功能，split\_bin\_val）查找表来决定向左/右节点移动

（2）否则主动方将节点ID发送给指定的被动方，被动方查看它的查找表，然后将当前节点转到哪个分支发送回活动方。

直到达到叶子节点，该过程停止。下图显示了联合推理过程。



如图所示，对于已经训练好的树模型，预测样本的类别为0

联邦模盒

用户的消费、金融画像：dmrc\_model\_t07\_chaobai\_feats\_s\_m

# 9. kuai操作超白特征表建模

## 9.1 银河工具学习

1. 权限申请

2. 元数据

3. 数据开发平台IDE（SQL取数）

4. kuai机器学习平台

jupyter调用5k数据：



## 9.2 建模

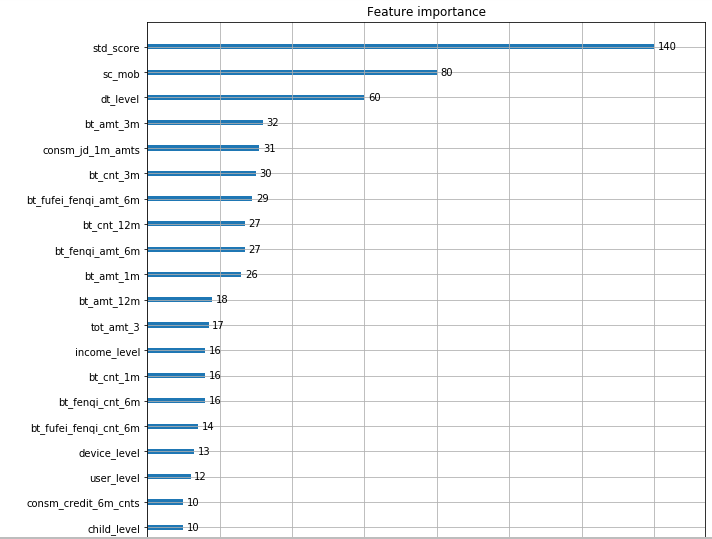
数据

|  |  |
| --- | --- |
| 数据来源 | 超白特征表（dmr\_c.dmrc\_model\_t07\_chaobai\_feats\_s\_m）  业务表现表（dmr\_zr.dmrzr\_umodel\_001\_bt\_jt\_jqpt\_biz\_inds\_target\_sampl\_s\_m）  以pin和user\_pin连接，dt选择2019-05-31和2019-05 |
| 特征数 | 47 |
| 样本数 | 训练集150000，验证集50000 |

建模效果

|  |  |
| --- | --- |
| 好坏样本比例 | 4:1 |
| xgboost-AUC | 0.7568（训练集），0.7572（验证集） |
| xgboost-KS | 0.5137（训练集），0.5143（验证集） |
| importance特征（前10） | std\_score 京东信用分-std  sc\_mob 京东商城帐龄等级  dt\_leve 多头等级  bt\_amt\_3m 近3个月白条累计交易金额  consm\_jd\_1m\_amts 京东商城近1个月消费总金额  bt\_cnt\_3m 近3个月白条累计交易笔数  bt\_fufei\_fenqi\_amt\_6m 白条账户近6个月付费分期金额  bt\_cnt\_12m 近12个月白条累计交易笔数  bt\_fenqi\_amt\_6m 白条账户近6个月分期金额  bt\_amt\_1m 近1个月白条累计交易金额 |
| 对比 | 第6章使用底层特征建模，auc分别为0.57（本地单方xgboost）、0.68（ficus联邦学习HeteroSecureBoost） |

importance:



iv:



|  |  |  |
| --- | --- | --- |
| 序号 | 按importance降序排列 | iv值 |
| 0 | std\_score | nan |
| 1 | sc\_mob | inf |
| 2 | dt\_level | 0.607 |
| 3 | bt\_amt\_3m | 0.399 |
| 4 | consm\_jd\_1m\_amts | 0.199 |
| 5 | bt\_cnt\_3m | nan |
| 6 | bt\_fufei\_fenqi\_amt\_6m | nan |
| 7 | bt\_cnt\_12m | 0.387 |
| 8 | bt\_fenqi\_amt\_6m | inf |
| 9 | bt\_amt\_1m | 0.388 |

# 10. 联邦模盒建模记录

## 10.0 联邦模盒使用的注意事项

1. 联邦模盒可以自己上传数据，大小没有限制（ficus限制在10MB以内）

2. 联邦模盒的可选参数比ficus少

3. 选择特征时，注意不要选y，y必须是int型

## 10.1使用方法

联邦学习需要至少两方参与模型训练，联邦模盒的两方分别是：

发起方（业务方，含有标签y）：<http://172.25.66.113:8081/federation_mbox/#/home>

处理方（京东平台）：<http://172.25.66.112:8081/federation_mbox/#/home>

账号：zhangxiaoyuan6

密码：eljIDA3j

联邦模盒的使用方法

1. 发起方和处理方新建数据集（新建表结构-确认上传本地文件-共享数据集）



2. 发起方在联邦学习项目中新建任务



选择SecureBoost/LogisticRegression算法，配置参数、模型验证方法、选择特征及标签所在的列名、ID对齐方法（可选RSA/不加密）







3. 处理方确认训练任务，进行训练

## 10.2 建模过程

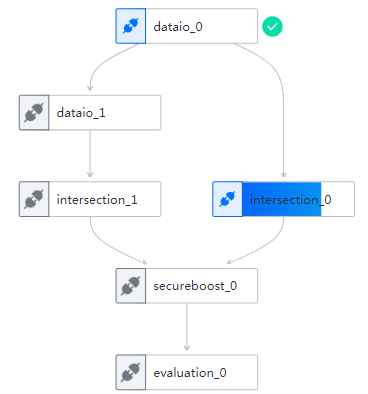
使用数据集：

|  |  |  |
| --- | --- | --- |
|  | 训练集 | 验证集 |
| 特征数 | 合作方：123  京东平台：552 | |
| 样本数 | 3w | 1w |
| ID对齐率 | 0.75 | 0.25 |

可选参数（4个）设置为：

|  |  |
| --- | --- |
| 学习率 | 0.1 |
| 树深度 | 2 |
| 树个数 | 2 |
| 样本采样比例 | 0.8 |

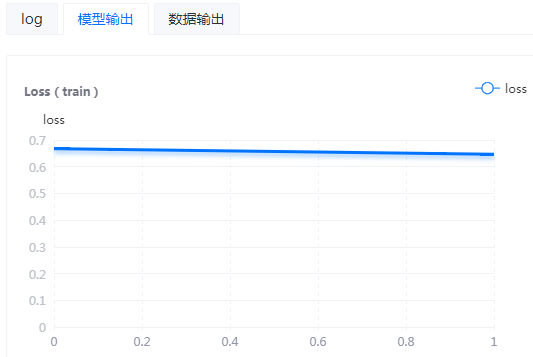
SecureBoost流程图如下

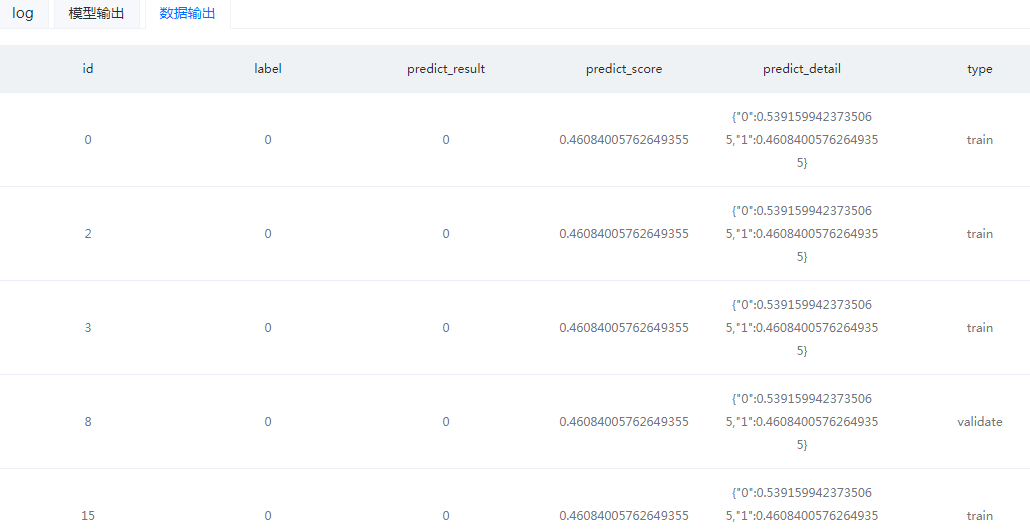


训练成功后，evaluation输出模型效果：



SecureBoost的模型输出是训练集的loss图像，数据输出是模型预测值与实际值的比较





## 10.3 SecureBoost参数选择

训练任务：

http://172.25.66.113:8081/federation\_mbox/#/taskList?platformId=10002&projectSeq=000000366683235&projectName=zhangxiaoyuan6&platformList=%5B%7B%22partyId%22%3A10002%7D,%7B%22partyId%22%3A10001%7D%5D

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 实验序号 | 作业编号 | 参数 | auc（训练集） | ks（训练集） |
| 0 | 000000366683235-20200529111631 | 学习率=0.1，  树深度=2  树个数=2  样本采样比例=0.8 | 0.5888 | 0.1762 |
| 1 | 000000366683235-20200529115215 | 同0，  树深度=3 | 0.6328 | 0.1755 |
| 2 | 000000366683235-20200529121942 | 同0，  树深度=3，  树个数=3 | 0.6319 | 0.1752 |
| 3 | 000000366683235-20200529142846 | 同0，  学习率=0.05  树深度=3 | 0.6388 | 0.1938 |
| 4 | 000000366683235-20200529152601 | 同0，  学习率=0.01，  树深度=3 | 0.6313 | 0.1755 |
| 5 | 000000366683235-20200529165639 | 同0，  学习率=0.05，  树深度=3，  样本采集比例=1 | 0.6313 | 0.1755 |
| 6 | 000000366683235-20200603113342 | 同0，  学习率=0.05，  树深度=3，  树个数=3 | 0.6328 | 0.1755 |

目前来看，表现较好的参数：

学习率=0.05，

树深度=3

树个数=2

样本采样比例=0.8

后续工作：

将超白特征表的47个特征作为处理方（京东平台）特征，尝试联邦学习建模