# 预处理和隐私求交

## 预处理

预处理算子可能会暴露隐私，需要评估

### DataFrame

仿照pandas的dataframe，是联邦表格数据的封装。支持三种切分模式：

水平切分：特征一致，样本不一样，HDataFrame

垂直切分：每一方有自己的特征，但是样本是对齐的（经过PSI）, VDataframe

混合切分：上两个切分混合

### FedNdarray

联邦ndarray的封装，每一方用numpy的ndarray表示

### 具体预处理操作

可以用DataFrame的API预处理，也可以用sf.preprocessing包的预处理组件进行处理

缺失值填充：

df.count()[“sepal width(cm)”]

df.fillna(value={“sepal width(cm)”:10})

df.count()[“sepal width(cm)”]

标准化：

from secretflow.preprocessing import MinMaxScaler

scaler = MinMaxScaler()

scaled\_sepal\_len = scaler.fit\_transform(df[“sepal width(cm)”])

print(scaled\_sepal\_len.max(), scaled\_sepal\_len.min())

会暴露最大值和最小值

## PSI

垂直拆分场景中的第一步，需要对齐

接口：

spu.psi\_csv()

data.vertical.read\_csv()

from secretflow.data.vertical import read\_csv as v\_read\_csv

vdf = v\_read\_csv(

{alice:a\_path, bob:b\_path},

spu,

keys,

drop\_keys,

psi\_protocol

)

vdf.columns

# 决策树模型、线性回归模型

## XGB决策树算法

import secretflow.ml.boost as boost

|  |  |  |  |
| --- | --- | --- | --- |
| 算法 | SS-XGB(MPC) | SecureBoost(FL) | 水平XGBoost(FL) |
| API | boost.ss\_xgb\_v.Xgb | boost.sgb\_v.Sgb | boost.homo\_boost.SFXgboost |
| 场景 | 垂直切分 | 垂直 | 水平 |
| 安全 | 安全，基于秘密共享 | 存在已知攻击 | 泄露中间信息 |
| 性能 | 通信量大，慢 | 计算量大，通常快 | —— |

## 回归和二分类

import secretflow.ml.linear as ln

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 算法 | SS-SGD | HESS-SGD | SS-GLM | 混合联邦LR |
| API | ln.  SSRegression | ln.  HESSLogisticRegression | ln.  SSGLM | ln.  FlLogisticRegressionMix |
| 场景 | 垂直切分 | 垂直 | 垂直 | 混合 |
| 安全 | 安全，基于秘密共享 | 安全，秘密分享+同态 | 安全，秘密共享 | 泄露信息 |
| 算法 | 线性/逻辑回归 | 逻辑回归 | 广义线性回归 | 逻辑回归 |
| 性能 | 通信量大，带宽好时更快 | 计算量大，网络不好时更快 | —— | —— |

# 神经网络算法

注意：这些算法都不是可证安全算法，会泄露中间信息

通过安全加固组件（安全聚合、差分隐私、稀疏化）可以加强安全性

两类：

|  |  |  |
| --- | --- | --- |
| 算法名称 | 水平联邦学习 | 垂直拆分学习 |
| API | secretflow.ml.nn.FLModel | secretflow.ml.nn.SLModel |

## 水平联邦学习

支持后端：Tensorflow、pytorch

可以用torchvision, tf.keras.applications等封装好的模型

FLModel是sf封装好的水平联邦训练器，提供一些选项：

数据加载器：DF、文件、图片等

安全聚合算法，有内置，如secure\_aggregator

联邦学习策略，也有内置，如fed\_avg\_w

## 垂直拆分

模型被拆分为几个部分，分别存在每隔参与方，建模时需要针对拆分学习架构设计模型结构

Sf拆分好的模型库：sf.ml.nn.applications

常见拆分类型：

1、两边都有特征，label在一方

2、label方没有特征

SLModel提供的选项：

数据加载器：DF、文件、图片等

通信优化算法，内置了一些稀疏化、量化压缩方法，如qkm=QuantizedKmeans()

垂直拆分策略，有内置，也可以自定义