LinkeFL系统

- 数据接入
  - 系统自带数据集，开箱即用（可继续扩展丰富）
  - CSV 数据（提供路径即可加载）
  - 数据库（如 MySQL）
  - 数据仓库（如 HIVE）
- 隐私保护技术支撑
  - 密码学机制
    - Paillier 加法同态
    - RSA 公钥加密
    - ECC 椭圆曲线加密
  - 噪声机制
    - epsilon-LDP
    - (epsilon, delta)-DP
  - 多方安全计算
    - 算术秘密分享
    - 安全聚合机制
- 通信支撑
  - Python Native Socket（点对点通信）
  - gRPC（点对点通信）
  - NVIDIA NCCL（集合通信）
  - MPI（集合通信）
- 配置支撑
  - 专业用户
    - 在源 Python 文件中提供配置参数
  - 中级用户
    - 提供配置文件路径一键加载
      - YAML配置
      - JSON配置
  - 普通用户
    - web前端（如文本框、下拉列表）
- 特征工程
  - 标准化
    - standardization
    - normalization
  - 转换
    - 添加intercept列
    - label parser
    - one-hot
  - 特征重要性
    - xgboost
    - shap
  - EDA分析
  - 相关系数
  - woe&iv
- 联邦建模
  - 横向联邦
    - CPU 训练（轮询各个 client）
    - 单机单卡（轮训各个 client）
    - 单机多卡（一卡一个 client）
    - 多机多卡（分布式训练）
    - 嵌入式设备（树莓派4B，NVIDIA Jetson）
  - 纵向联邦
    - 隐私集合求交
      - RSA 盲签名
      - ECDH
      - KKRT, CM20
    - 纵向线性模型
      - 明文训练（无缝切换到密文）
      - 基于 Paillier 线性回归
      - 基于 Paillier 逻辑回归（还有进一步优化空间）
      - 基于 Paillier 和 SS 的混合模型
      - 带可信第三方协议
    - 纵向树模型
      - 明文训练（无缝切换到密文）
      - 密文训练（还有 bug 待修复）
    - 纵向神经网络模型
      - 明文训练（无缝切换到密文）
      - 密态训练
  - 拆分学习
    - 明文训练
    - 模型表征保护
    - 模型梯度保护
  - Pipeline
    - 分割离线阶段和在线阶段
    - 数据加载→求交→特征工程→建模 流程一键启动
- 代码文档
  - 重要类和函数编写rst注释，用sphinx自动构建API文档
- 安全隐私
  - 隐私保护的non-iid策略（封装算法加入横向联邦）
  - 纵向LR标签推断
    - 标签推断算法
    - 加噪保护机制（集成到训练算法中）
    - 混合保护机制（集成到训练算法中）
  - 样本交换标签推断
    - 标签推断算法
    - 保护机制（集成到训练算法中）
  - splitNN数据重构
    - 数据重构算法
    - 模型重构算法
    - 保护机制（集成到训练算法中）
  - 纵向联邦推理攻击
    - 攻击算法
    - 检测算法
- 通用模型表征学习
  - 先实现方法后面集成进训练算法
- 激励机制设计
  - money reward
  - 可解释性分析
  - 机制本身设计
- 贡献度评估
  - 特征（封装成算法后加入特征工程）
  - 用户（封装算法）
- 数据质量评估
  - 任务无关型数据质量评估
  - 重要样本筛选
  - 训练中样本筛选（基于样本grad、loss、bound等）封装成算法接口
  - 训练后model debugging
  - 纵向联邦特征筛选（算法封装后加入特征工程）
- 工程优化
  - 离线签名
  - 并行化签名
  - 快速 Paillier
  - 控制浮点数加密精度
  - 在加载数据前提前完成数据转换