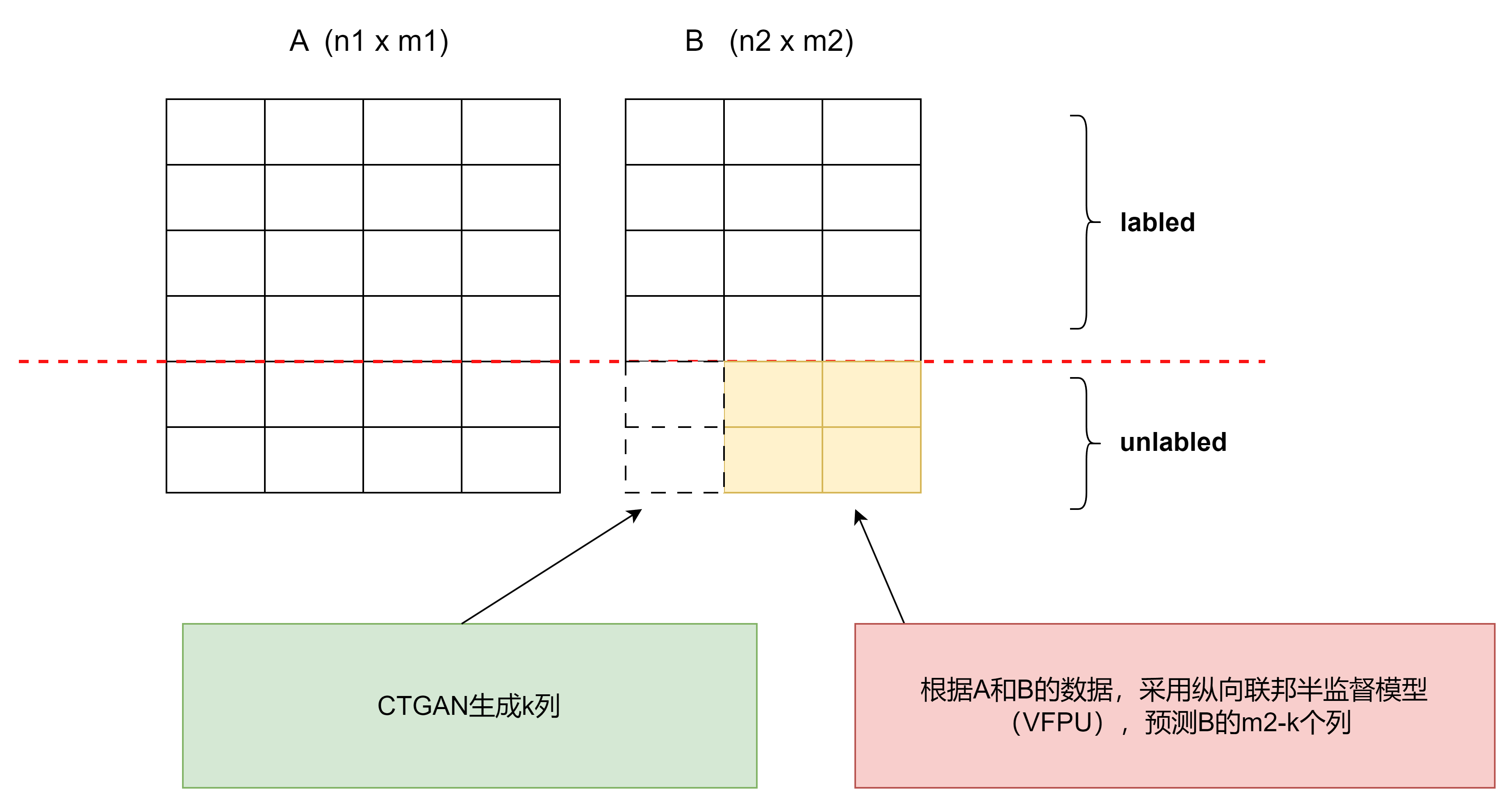
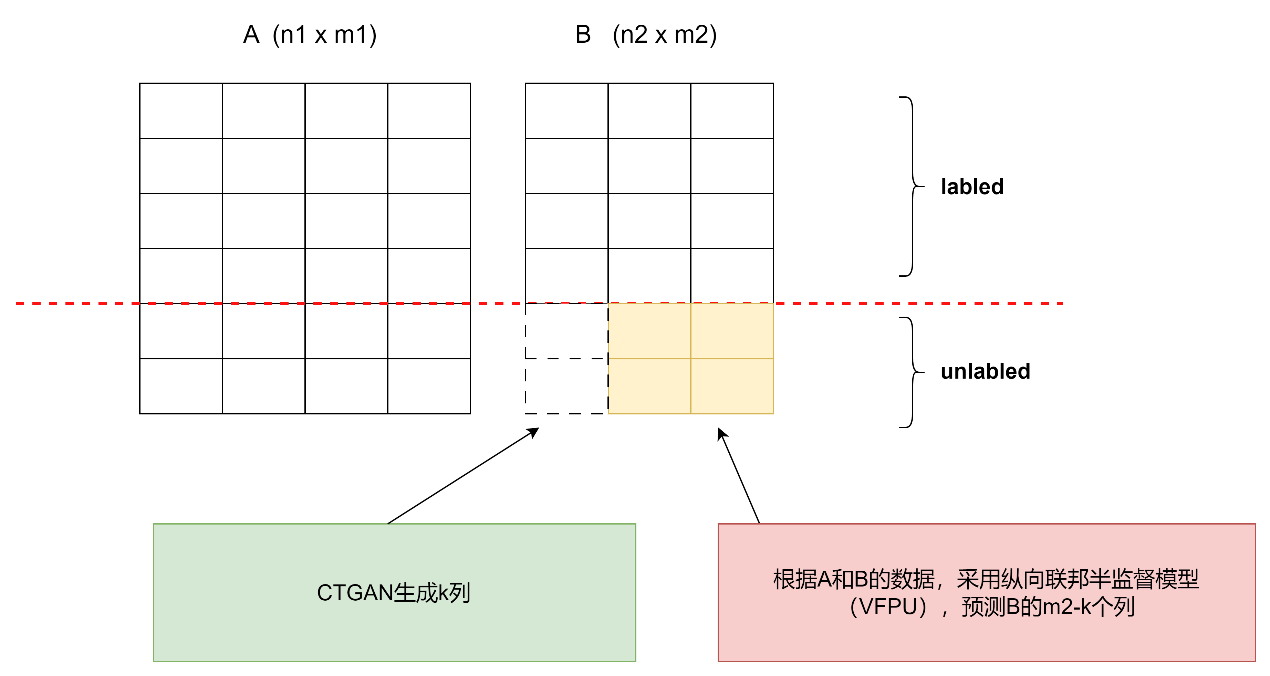
问题描述





避免加入到labled的数据噪声影响：

1选择置信度、排名k个参数进行训练

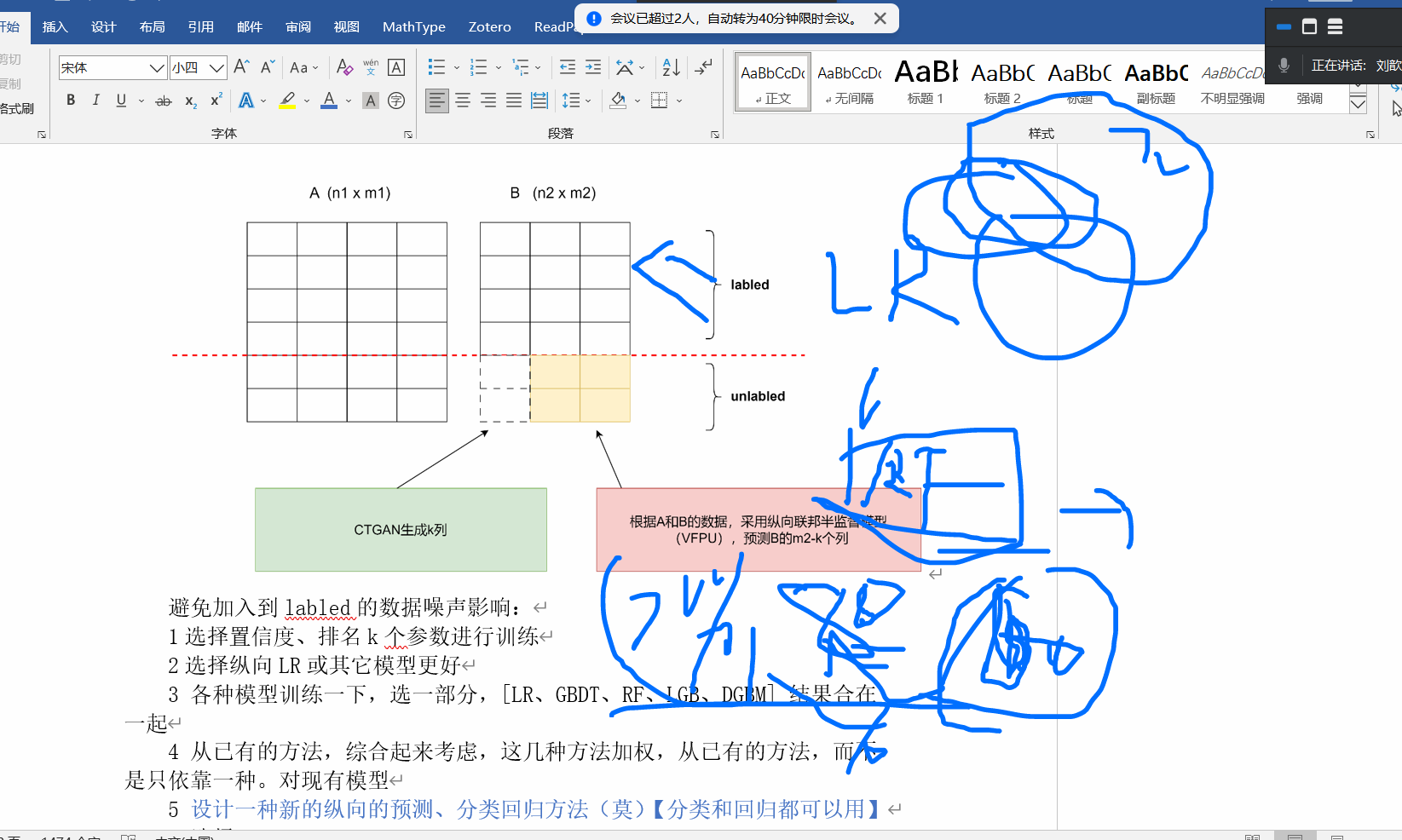
2选择纵向LR或其它模型更好

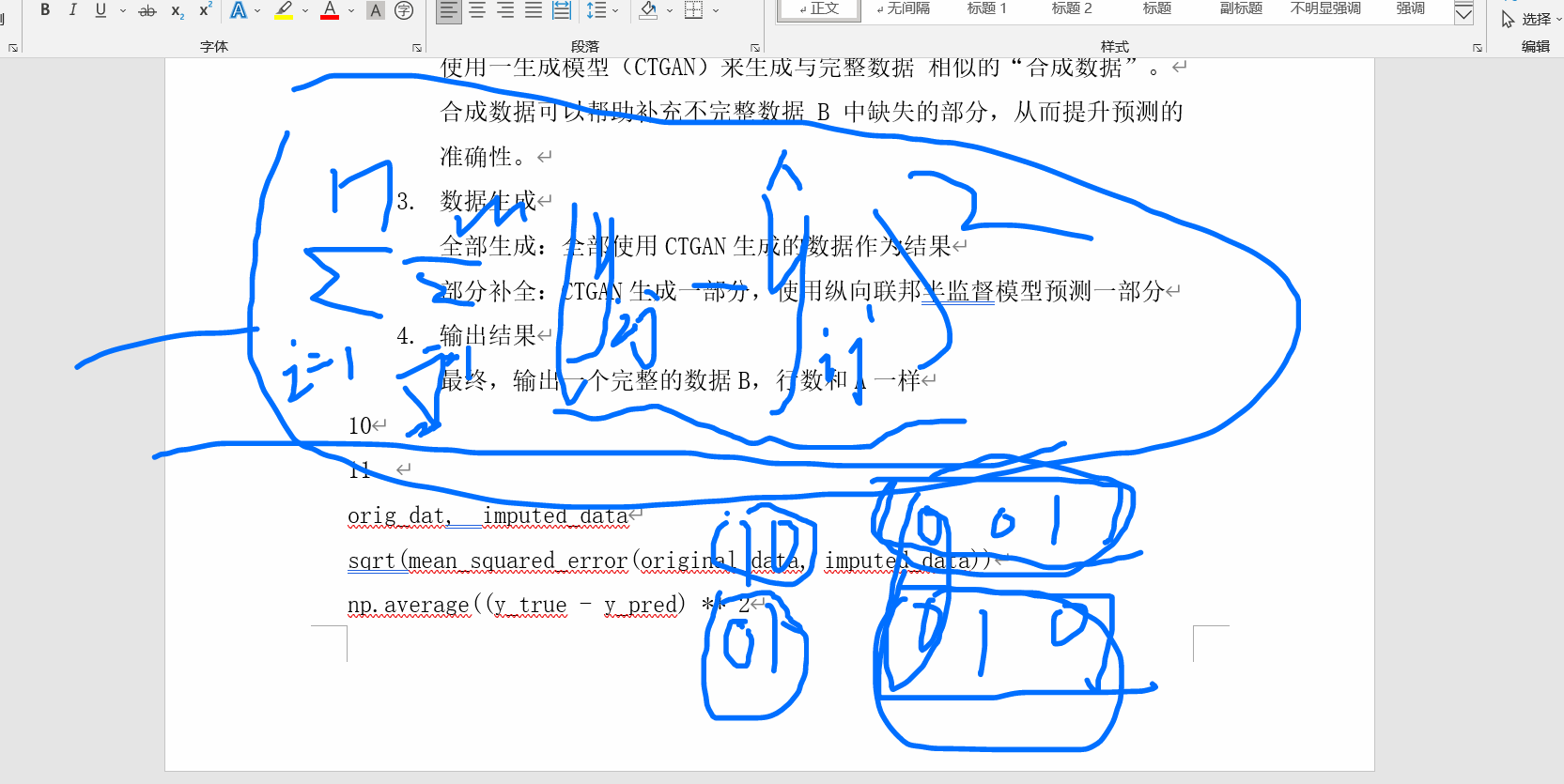
3 各种模型训练一下，选一部分，[LR、GBDT、RF、LGB、DGBM] 结果合在一起，每个模型结果取交集、取权重，加权平均，可以。

4 从已有的方法，综合起来考虑，这几种方法加权，从已有的方法，而不是只依靠一种。对现有模型

5 设计一种新的纵向的预测、分类回归方法（莫）【分类和回归都可以用】

6 选择





VFPU\_GEN 的工作可以分为以下几个步骤：

1. 数据准备

需要两部分数据：

* 完整数据 A：包含所有字段且没有缺失值的数据。
* 不完整数据 B：部分行缺失的数据。

1. 生成合成数据

使用一生成模型（CTGAN）来生成与完整数据 相似的“合成数据”。

合成数据可以帮助补充不完整数据 B 中缺失的部分，从而提升预测的准确性。

1. 数据生成

全部生成：全部使用CTGAN生成的数据作为结果

部分补全：CTGAN生成一部分，使用纵向联邦半监督模型预测一部分

1. 输出结果

最终，输出一个完整的数据B，行数和A一样

10

11

对于一个属性的预测，如何计算，gan中算损失的，交叉熵（KL）

对于数值型用rmse

数值列和属性列分开算在加起来

orig\_dat, imputed\_data

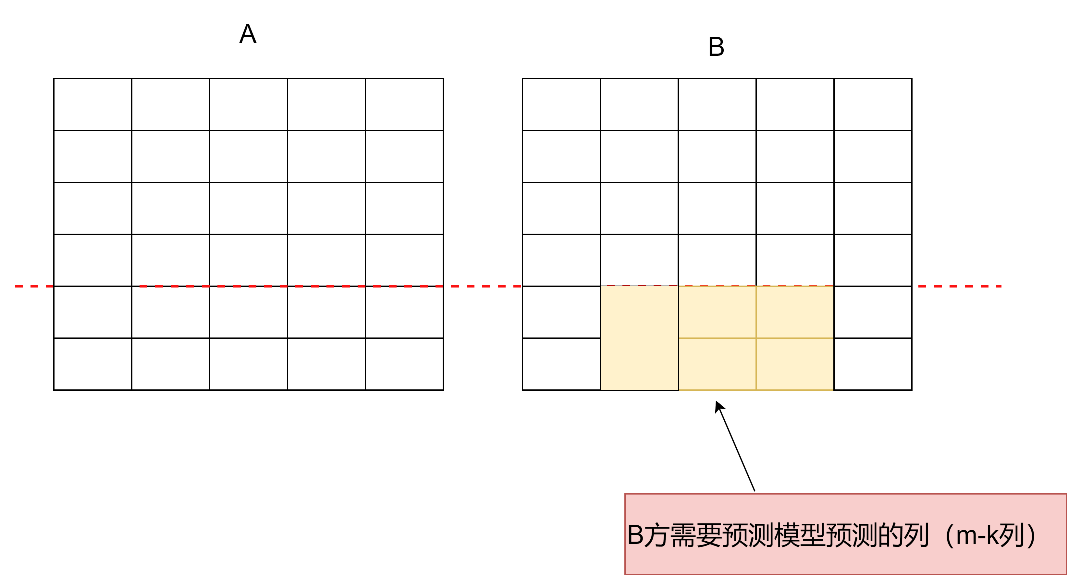
sqrt(mean\_squared\_error(original\_data, imputed\_data))

np.average((y\_true - y\_pred) \*\* 2

y\_ture [1 0 1 1]

y\_pred [0 1 0 1]

纵向联邦半监督预测方法（VF\_TwoStep）概述

L << U

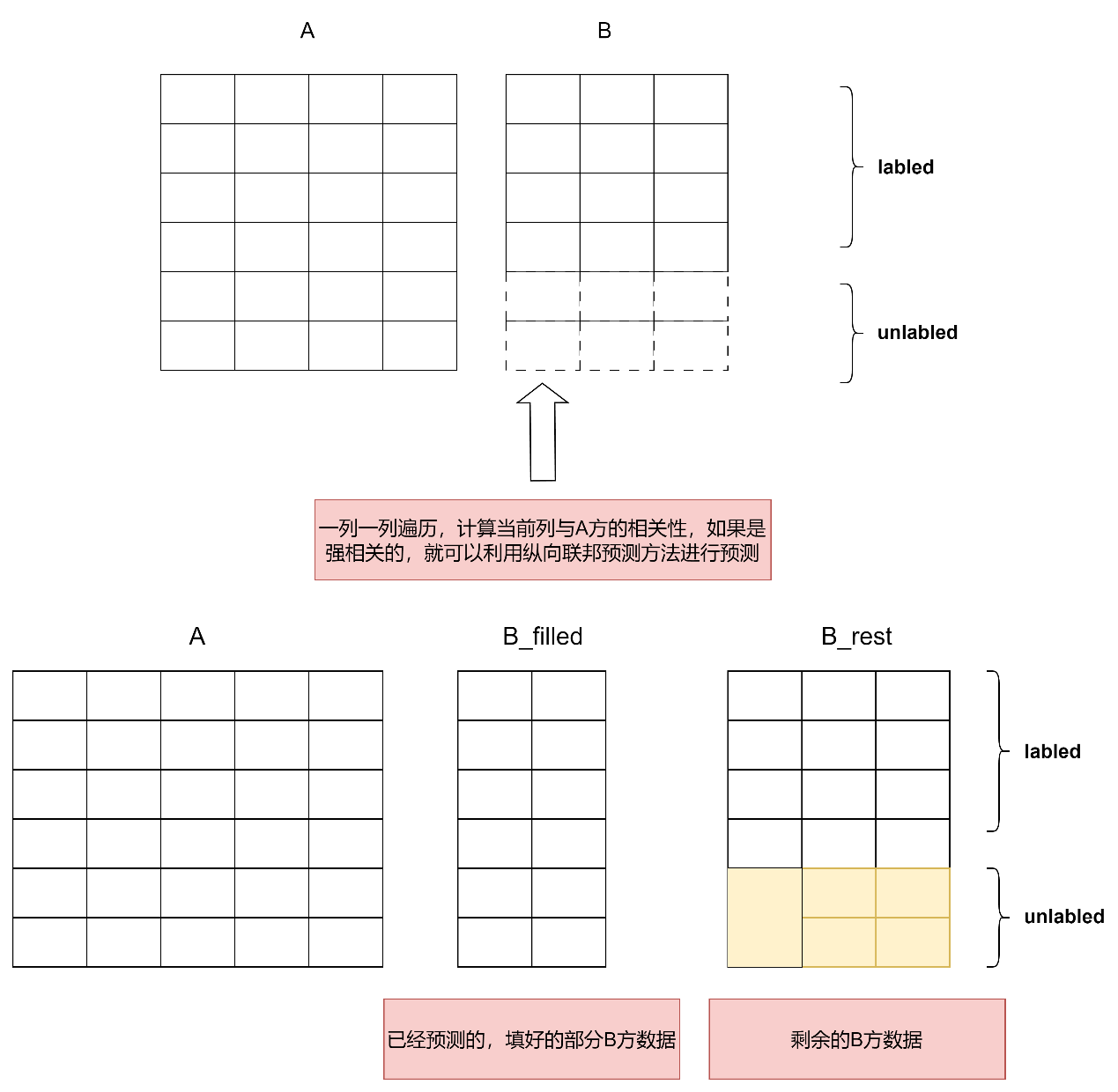
L: 只有正样本

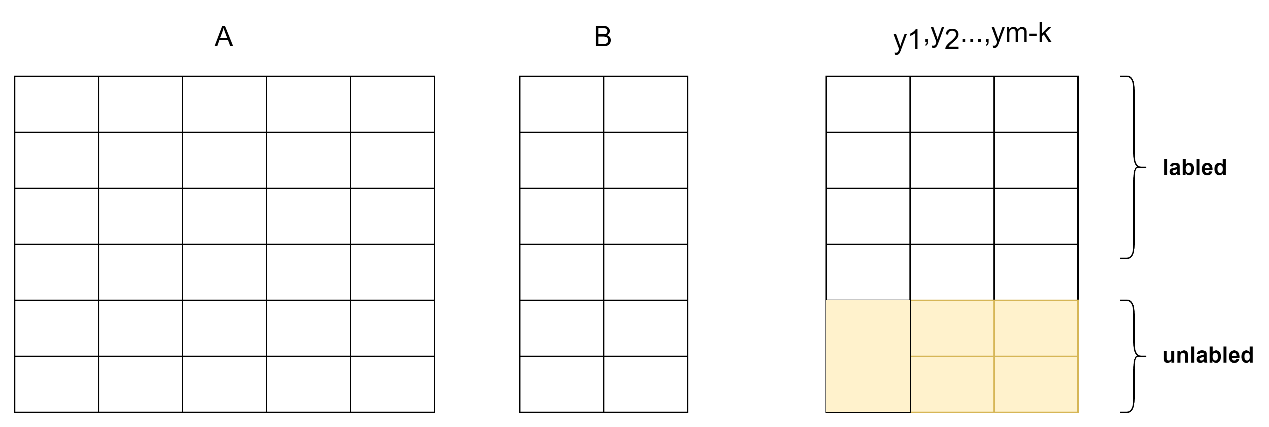
L: 不仅有正样本，也有负样本

初始：A提供数据，B提供标签

B提供的数据在增多，然后A和B一起填

不是所有的列都是预测，如果当前列与其它数据相关性很低，这个过程当中，先基于A这边，B这边与A强相观的列进行预测，这些列中，相关性，如果没有任何关系，就，如果有强相，就用预测方法，





总体思路

给定一组带标签的样本集 和一组未标注的样本集 ，目标是充分利用 和 中的信息来训练模型，对每一个未标记数据进行标注。

VF\_TwoStep 的整体流程如下：

1. 初始化：将已标注数据 用于训练初始模型（纵向联邦分类器或纵向联邦回归器），并对未标注数据 进行预测。
2. 置信度估计：对未标注样本的预测结果打分，以衡量其被模型正确预测的确信程度。对于分类问题，常用最大后验概率 作为置信度；对于回归问题，使用模型的预测值计算标准分数（Z分数，衡量一个数据点与其分布的平均值之间的偏离程度的标准化值）。作为置信度。
3. 选取高置信度样本：在每次迭代中，从未标注数据集中筛选出置信度大于某一阈值或排名前 的样本加入到带标签集合 中，以丰富真实或“准真实”的标注信息。
4. 重复迭代：将新增的带伪标签样本加入训练，更新模型参数，并在下一轮迭代中继续对剩余未标注数据进行筛选，直到达到最大迭代次数或没有足够的高置信度样本为止。
5. 最终预测：若仍有剩余未标注数据，则使用最终训练得到的模型进行推断或预测。

具体描述

令带标签样本为：

其中 与 代表了可并行输入给分类模型或回归模型的两种特征表征。未标注样本为：

max\_iter

k

（超参数，需要进行实验）

排名取多少，去前面多少个

分了两个过程，这两个过程的k设置不一样是否会达到较高的准确率

现在使用纵向LR和纵向Linear，使用其它

假设当前迭代轮次下，任务被判定为分类，使用纵向联邦分类器 。对于任意未标注样本 ，得到预测概率分布：

xj [0.1, 0.1,0.7,0.1] 0.7

[0.6, 0.1,0.2,0.1] 0.6

…

m条向量，m个置信度

10%

则其置信度可定义为

其中 表示样本 被预测为类别 c 的概率。当 大于设定阈值 或位于排名前 k\%，则视为高置信度样本。

如果任务被判定为回归，使用纵向联邦回归 。对于未标注样本 ，可得预测值

x1 2.35

x2 3.15

x3 1.55

…

m个未标记样本

z\_j = (y\_j -u)/q

在回归场景中，由于 不再是概率，而是未标记样本 具体的预测值，置信度可以用这一批未标记样本预测的值来计算计算每个预测值的 Z 分数（标准化偏离程度），这个值越小，说明，偏离

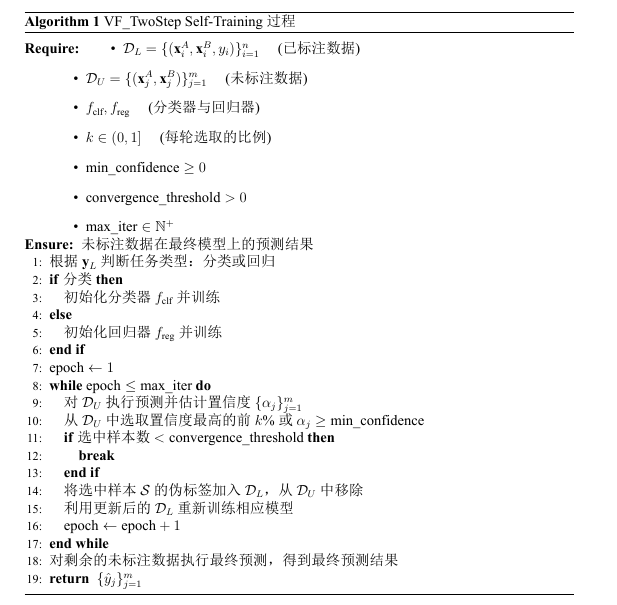
其中， 是所有预测值 的平均值， 是所有预测值的标准差。

计算所有预测值   的平均值 ：

计算所有预测值的标准差 ：

同样地，选出 较大的样本并视为高置信度。每轮迭代执行完后，若高置信度样本数量小于收敛阈值 ，则停止迭代；否则将新增样本并更新模型。

算法流程



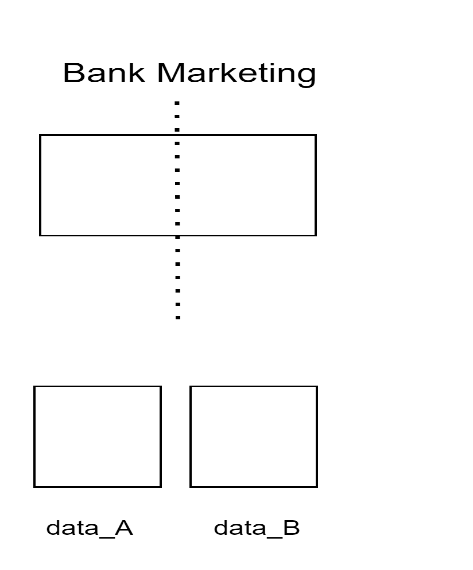
实验设置

数据集：Bank Marketing

1 数据预处理

预处理之后的数据集的大小为：（41188，63）

2 垂直划分数据，模拟两个参与方A和B



data\_A的大小 （41188，32）

data\_B的大小 （41188，31）

miss\_rate