会议纪要

1 使用何种指标来验证生成的结果

生成后的数据使用的均方根误差（Root Mean Square Error, RMSE）指标作为衡量生成数据的好坏，是用来衡量两个矩阵之间差异的常用指标。假设有两个矩阵 $\mathbf{orig}$ 和 $\mathbf{imputed}$，它们的维度相同，均为 $m \times n$。RMSE 的计算公式可以用 如下：

$$\text{RMSE} = \sqrt{\frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\mathbf{orig}_{ij} - \mathbf{imputed}_{ij} \right)^2}$$

其中：

- $\mathbf{orig}_{ij}$ 表示原始矩阵 $\mathbf{orig}$ 中第 i 行第 j 列的元素。
- $\mathbf{imputed}_{ij}$ 表示生成矩阵 $\mathbf{imputed}$ 中第 i 行第 j 列的元素。
- m 和 n 分别是矩阵的行数和列数。

直接用 RMSE 指标计算可能会有问题，原因在于类别列（经过 one-hot 编码）和数值列的性质不同。

1. 类别列和数值列的尺度不同

- 数值列：通常是连续的数值，可能有较大的范围（例如 0 到 100）。
- 类别列：经过 one-hot 编码后，值通常是 0 或 1，表示类别的存在与否。

如果直接用 RMSE 计算，会导致数值列的误差对 RMSE 的贡献远大于类别列的误差，因为数值列的范围通常更大。这样，类别列的误差可能被数值列的误差掩盖，导致 RMSE 对类别列的预测质量不敏感。

2. 类别列的特殊性质

- 对于类别列，one-hot 编码后的值（0 或 1）实际上表示的是类别的离散性，而不是连续性。直接用 RMSE 计算类别列的误差，可能无法很好地反映类别预测的质量。
- 例如，如果一个类别的真实值是 `[0, 1, 0]`，而预测值是 `[0.2, 0.7, 0.1]`，虽然 RMSE 可以计算误差，但它并不能很好地衡量预测的类别是否正确。

3. 类别列和数值列的权重问题

- 如果矩阵中类别列和数值列的数量不均衡（例如类别列占了很大比例），RMSE 的计算会受到类别列的主导，反之亦然。
- 这种不均衡会导致 RMSE 对某些列类型的误差更加敏感，而对其他列类型的误差不敏感。

解决方法

为了更合理地衡量误差，可以考虑以下方法：

1. 分开计算误差

- 对类别列和数值列分别计算误差：
  - 对数值列使用 RMSE。
  - 对类别列使用分类指标（如准确率、F1 分数、交叉熵损失等），或者对 one-hot 编码的列使用类似于 RMSE 的指标（如均方误差）。

- 最后根据实际需求，将两部分误差加权组合。

2. 归一化数值列

- 在计算 RMSE 之前，将数值列归一化到与类别列相同的范围（例如 [0, 1]）。这样可以减少数值列和类别列之间的尺度差异对 RMSE 的影响。

3. 加权 RMSE

- 为类别列和数值列的误差赋予不同的权重，确保它们对总误差的贡献更加平衡。例如：

$$\text{Weighted RMSE} = \sqrt{\frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} w_j \cdot \left(\mathbf{orig}_{ij} - \mathbf{imputed}_{ij}\right)^2}$$

  其中 w_j 是第 j 列的权重，可以根据列的类型（类别列或数值列）进行调整。

4. 使用其他指标

- 对于类别列，可以直接使用分类指标（如准确率、F1 分数、交叉熵等）来评估预测质量，而不是用 RMSE。
- 对于整个矩阵，可以结合多种指标（如 RMSE + 分类指标）来综合评估。

2 如何避免噪声数据对已标注数据集的影响

解决方案讨论:

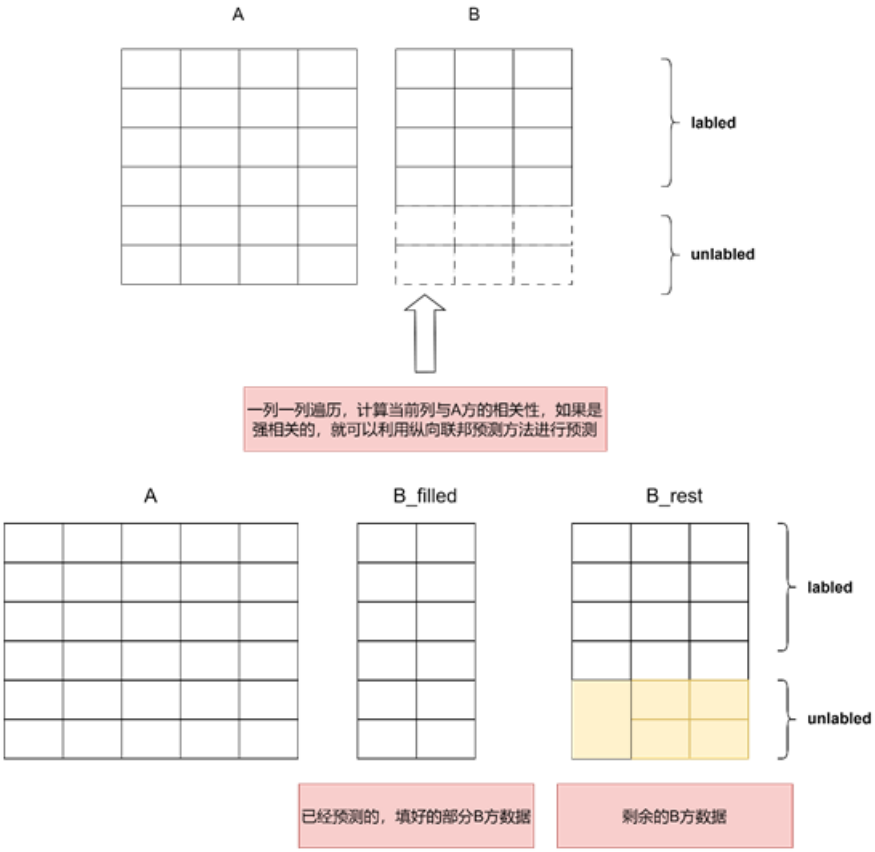
- 超参数设置:
  - 强调了设置合适的超参数的重要性，如迭代次数(max_iter)和每轮选择的样本比例(k)。
- 置信度选择:
  - 讨论了选择合适的置信度指标，以确保高置信度样本的预测结果更为准确。
- 基学习器选择:
  - 提出了从纵向逻辑回归、纵向线性回归、纵向GBDT、纵向RF、纵向LGBM中选择最优基分类器的建议。
- 多模型组合策略:
  - 探讨了使用多种模型进行训练，并通过取交集、加权平均等方法组合结果的可能性。
  - 建议对不同模型的结果进行验证，以确定最佳组合策略。
- 新基学习器设计:
  - 提议设计一种新的基学习器，适用于分类和回归任务。
  - 鼓励师弟在此方向上进行深入研究和探索。

后续步骤:

- 进一步研究和优化超参数设置和置信度选择。
- 实验不同基学习器和多模型组合策略的效果。
- 开展新基学习器的设计和验证工作。

3 基于相关性来选择预测那些列

在填补B方进行数据时，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，就没必要基于A方和B方已填数据进行纵向联邦半监督预测。这个过程当中，先基于A这边，对B中与A强相观的列进行预测，剩下相关性不是很强的列，进行填补。



整个选取预测列和生成列的过程的伪代码如下:

1. **输入**: $\mathbf{X}^A$, $\mathbf{X}^B$, 相关性阈值 δ
2. **对 B 方各列** $j = 1, \ldots, d_B$:
   a) 计算与 A 方数据的相关性指标 Corr_j。
   b) 若 $\mathrm{Corr}_j \geq \delta$（强相关）：
   - 使用纵向联邦半监督学习在 A 方的特征空间进行训练；
   - 预测生成 $\mathbf{x}_j^{B_{\text{filled}}}$ 并使行数扩展到 N_A。
     c) 若 $\mathrm{Corr}_j < \delta$（弱相关）：
   - 使用生成模型对缺失行/缺失值进行补全，得到 $\mathbf{x}_j^{B_{\text{gen}}}$。
3. **输出**:
   - $\mathbf{X}^A$;
   - 拼接各列后得到的 $\mathbf{X}^{B_{\text{filled}}}$，其行数与 $\mathbf{X}^A$ 相同;
   - 其余部分 $\mathbf{X}^{B_{\text{rest}}}$ （若需要保留）。

严升

分类任务流程：将无标记样本进行分类，对于每一条样本都有其属于某一类的概率，从该类中选取TopK个样本，将这部分样本加入训练集。剩余部分重复迭代，迭代结束条件由自己决定。

回归任务流程：对于预测出来的数据，进行标准化计算（（预测数据-预测的所有数据的方差）/预测的所有数据的方差）。所得到的值越小则可信度越高，从中选取TopK个样本，将这部分样本加入训练集。剩余部分重复迭代，迭代结束条件由自己决定。

迭代终止问题：是否存在一种通用的可接受的方式确定迭代次数。

数据填补问题：对于某些列的填补如何选择填补方式，例如该需填补列与已有的一些列具有相关性，那么就可以通过这些列进行填补，另外如何寻找数据中列的相关性仍需解决。

关于自己后面的研究方向

1. 最终目标找到一种能够从未标记样本中，通过纵向半监督的方法选择尽量正确的填补样本，避免噪声过多的进入训练中的模型影响效果。可以通过已有方法相结合去实现或者自己构建未有模型。
   深入纵向联邦学习，明白其中原理。从学术论文阅读和项目实战入手。最好是能够通过这些学习能够自己建立模型完成上述目标。

2. 短期安排：
   1.收集基于半监督的正负样本区分并去做实验对比。
   2.看吕师兄关于选择正样本的论文并复现论文。

吴艳

从无标签数据中选择k%个数据加入标签数据中时，需要避免噪声的影响，首先考虑现有的模型LR，GBDT，RF，LGB，DGBM中某一种方案或者是综合使用，找到一种尽量最优的方法。其次考虑设计一种符合要求的新方案。