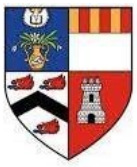

UNIVERSITY OF ABERDEEN

阿伯丁大学

自然与计算科学学院 计算科学系

2024 - 2025

编程作业 - 单独评估（无团队合作）

标题JC4001 - 分布式系统	注：本作业占课程总分的 30%。

学习成果

成功完成这部分内容后，学生将证明自己能够

- 了解分布式系统中联合学习（FL）的原理，以及它与集中式机器学习的区别。
- 利用 MNIST 数据集，在分布式系统中实施基本的联合学习，进行图像分类。
- 模拟分布式系统中的联合学习环境，其中多个客户端独立训练模型，服务器对其进行汇总。
- 探索模型聚合的效果，并与集中培训进行比较。
- 评估 FL 模型在不同条件下的性能，如非 IID 数据分布和客户数量变化。

剽窃和串通信息：源代码和您的报告可能会被提交进行抄袭检查。有关避免抄袭的更多信息，请参阅 *MyAberdeen* 上的幻灯片，然后再开始评估工作。使用大型语言模型（如 ChatGPT）编写代码或报告也可能被视为抄袭。此外，与其他学生一起提交类似的作业也会被视为串通行为。还请阅读以下由大学提供的信息：

https://www.abdn.ac.uk/sls/online-resources/avoiding-plagiarism/

导言

在本作业中，您的任务是在分布式系统中构建一个联合学习（FL）算法。FL 是一种训练机器学习模型的分布式方法，旨在通过在没有集中数据集的情况下训练学习模型来保证本地数据的私密性。如图 1 所示，FL 结构应包括两个部分。第一部分是用于模型聚合的边缘服务器。第二部分应包括多个设备，每个设备都有一个本地数据集，用于更新本地模型。然后，每个设备将更新后的本地模型传送到边缘服务器进行本地模型聚合。

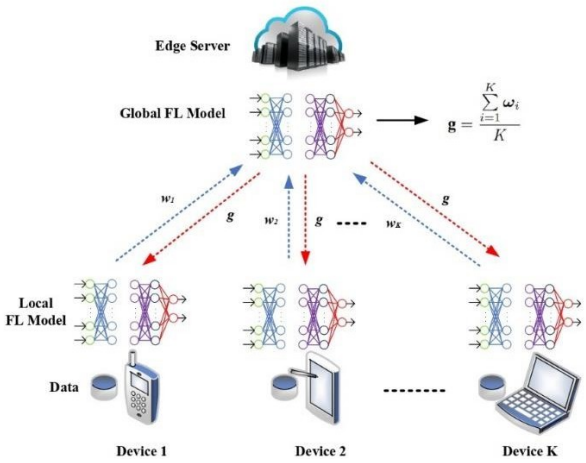

图 1：FL 结构示意图FL 结构示意图。

一般指导和要求

您的作业代码和报告必须符合以下要求，并包含各部分概述的必要内容。您**必须**提供一份书面报告以及相应的代码，其中包含所有不同的章节/子任务，对所进行的过程进行全面的批判和反思。

本作业可以在您自己的设备上使用 Python/PyCharm 完成。如果您在自己的设备上完成作业，请务必定期将文件移动到 MyAberdeen，以便我们运行应用程序并对其进行标记。

请注意，确保代码在 Python/PyCharm 上运行是您的责任。默认情况下，您的代码应通过直接点击 "运行 "按钮来运行。如果您的实现使用了其他命令来启动代码，则必须在报告中提及。

提交指南。完成作业后，请将所有文件压缩到一个压缩文件中，然后在 MyAberdeen 中提

交（内容 -> 作业提交 -> 查看说明 -> 提交（将文件拖放到此处））。

第 1 部分：了解联合学习 [5 分]

1. 阅读研究论文：您应该阅读一篇关于联合学习的基础性论文，例如 McMahan 等人（2017 年）撰写的《从分散数据中高效学习深度网络》（Communication-Efficient Learning of Deep Networks from Decentralized Data）。
2. 总结任务：写一份 500 字的摘要，解释联合学习的关键要素（客户端-服务器架构、数据隐私以及非 IID 数据等挑战）。**[5 分]**

第 2 部分：集中学习基线[15 分]

1. **实施集中训练**：您应该使用集中式方法实施一个简单的神经网络，对 MNIST 数据集中的数字进行分类。这将作为一个基准。
    - 输入：MNIST 数据集。**[5分]**
    - 模型具有多个隐藏层的基本神经网络。**[5 分]**
    - 任务：训练模型并评估其准确性。**[5 分]**

第 3 部分：联合学习的实施 [30 分]

1. **模拟客户端**：将 MNIST 数据集拆分成几个分区，以表示本地存储在不同客户端的数据。实现一个模拟客户端的 Python 类，每个客户端保存一个数据子集。**[10 分]**
    - 任务：执行一个函数，以 IID（独立且同分布）和非 IID 两种方式分割数据。
2. **在客户端上进行模型训练**：修改集中式神经网络代码，使每个客户端都能使用本地数据独立训练模型。**[5 分]**
3. **服务器端聚合**：实施一个简单的参数服务器，聚合客户端发送的模型更新。使用联

合平均（FedAvg）算法：[**10 分**]

- 每个客户端在对本地数据进行训练后，都会向服务器发送其模型参数。
- 服务器汇总这些参数（根据每个客户端的样本数量加权）并更新全局模型。

4. **通信轮**：实现一个循环，客户机训练其本地模型，服务器在多轮通信中汇总这些模型。[**5 分**]

第 4 部分：实验和分析 [20 分]

1. **实验 1 - 客户数量的影响：[10分]**
    - 改变客户端数量（如 5、10、20），评估最终联合模型的准确性。
    - 绘制每种情况下的训练精度和通信回合损失图。
2. **实验 2 - 非 IID 数据：[10分]**
    - 修改客户端的数据分布，模拟非 IID 情景（客户端的数据子集有偏差或倾斜）。
    - 比较联合学习模型在客户端拥有 IID 数据和非 IID 数据时的性能。绘制两种情况下的准确率和通信回合损失图。

第 5 部分：与集中学习的性能比较 [5 分]

- 比较联合学习模式（IID 和非 IID）与集中学习基线在以下方面的差异：
    - 最终精度
    - 收敛所需的历时/通信轮数

项目报告的要求和评分标准 [25 分]

您应该撰写一份报告。您的报告应描述分布式系统中联合学习的整体设计，以及联合学习编程过程中面临的挑战。

报告的评分标准如下：

- 结构和完整性（涵盖所有方面）**[5 分]**。
- 清晰易读（语言通俗易懂）**[5 分]**。
- 设计说明 **[5 分]**。
- 讨论的挑战 **[5 分]**。
- 参考资料来源 **[5 分]**。

提交材料

您应在 MyAberdeen 中使用 "作业提交 "链接提交代码和报告。**截止日期为 2024 年 12 月 22 日。请勿迟于截止日期。**

联系方式

如有任何问题或说明，请联系课程教师：刘晓楠博士 (xiaonan.liu@abdn.ac.uk)。