第1章 联邦学习基础

本章将回顾联邦学习的基本概念，包括其提出的背景、定义和分类等相关知识，使读者对联邦学习有初步的认识。

1.1 联邦学习概述

1.1.1 联邦学习背景

随着算法不断被创新、训练数据不断被收集、硬件算力不断被增强，机器学习技术，特别是深度学习（Deep Learning，DL）技术在人工智能（Artificial Intelligence，AI）应用领域取得了巨大成功。例如：在图像识别领域，通过卷积网络实现的视觉算法在识别正确率上早已超越人类；在自然语言处理领域，Google 在 2018 年提出的 BERT 算法 [1]，刷新了自然语言处理的 11 项纪录；在推荐系统领域，YouTube、FaceBook、Netflix 等科技公司正在使用智能的推荐引擎，通过分析用户的历史数据，为用户推荐个性化的内容和商品，帮助提升用户的黏性和留存率。

在过去很长的一段时间，数据的价值主要体现在作为一种燃料，为人工智能模型提供大量的训练样本数据，帮助提升模型的效果。但随着移动互联网的快速发展，数据的规模变得越来越庞大与复杂，数据的存在价值已经不局限于作为训练数据而存在，而是以一种资产的形式服务于各家企业，并预期给企业带来经济收益。这种经济收益可以体现在两方面：一方面是数据作用于产品或者业务，从而间接帮助提高产品的收益，比如各运营商或者社交网络服务商都拥有丰富的用户数据，基于用户的行为数据、位置信息等数据，为每个用户构建完善的用户画像，帮助企业深入了解用户行为偏好和需求；另一方面，数据可以直接与企业收益相关，比如各金融机构有用户的历史逾期数据，一个有效的对逾期用户的识别模型，能够大大降低金融机构的贷款风险，减少公司的潜在经济损失。数据的资产属性也催生了一种新的商品交易模式，即大数据交易。

正是因为数据具有资产的属性，使得从政府、企业乃至个人都对数据越来越重视。但由于相互之间的竞争和隐私保护法规，各方的数据很难共享，导致数据呈现出割裂的状态，进而影响了对数据极度依赖的人工智能的发展。

为此，人们开始寻求一种方法，它能够不必将所有数据集中到一个中心存储点就能够训练机器学习模型。一种可行的方法是每个拥有数据源的机构利用自身的数据单独训练一个模型，之后各机构的模型彼此之间进行交互，最终通过模型聚合得到一个全局模型。为了确保用户隐私和数据安全，各机构间交换模型信息的过程将会被精心设计，使得没有机构能够猜测得到其他机构的隐私数据内容。同时，在

构建全局模型时，其效果与数据源被整合在一起进行集中式训练的效果几乎一致，这便是联邦学习（Federated Learning，FL）提出的动机和核心思想。

1.1.2 联邦学习定义与分类

联邦学习是通过隐私保护技术融合多方数据信息，协同构建全局模型的一种分布式训练范式。在模型训练过程中，模型相关的信息（如模型参数、模型结构、参数梯度等）能够在各参与方之间进行交换（可以通过明文、数据加密、添加噪声等方式），但本地训练数据不会离开本地。这一交换不会暴露本地的用户数据，极大地缓解了数据泄露的风险，训练好的联邦学习模型可以在各数据参与方之间进行共享和部署使用。

随着联邦学习研究的不断深入，已经有越来越多的传统机器学习算法开始支持在联邦学习框架上运行，本节对目前常用的机器学习算法在联邦学习上的实现进行简短小结。

横向联邦学习如图 1.1 所示，在文献 [2] 中首次被提出，常用于跨设备端（Cross-Device）场景，是当前研究最多的联邦学习类型。目前，线性模型（如线性回归、逻辑回归等）、提升树模型 GBDT [3]、递归神经网络 [4]，卷积神经网络 [5]、个性化推荐中的横向矩阵分解等都已经在横向联邦上实现。事实上，通常情况下使用梯度下降等最优化算法迭代优化的机器学习模型基本都能通过横向联邦学习框架进行训练。

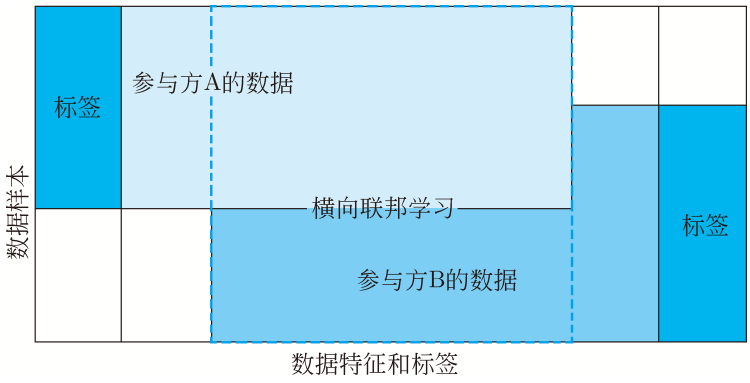

图 1.1 横向联邦学习示意 [6]

纵向联邦学习如图 1.2 所示，在文献 [6] 中正式提出，常用于跨机构（Cross-Silo）场景。目前，线性模型（如线性回归、逻辑回归等）、提升树模型 SecureBoost [7]、神经网络、个性化推荐中的纵向矩阵分解、纵向因子分解机等都已经在纵向联邦上实现。

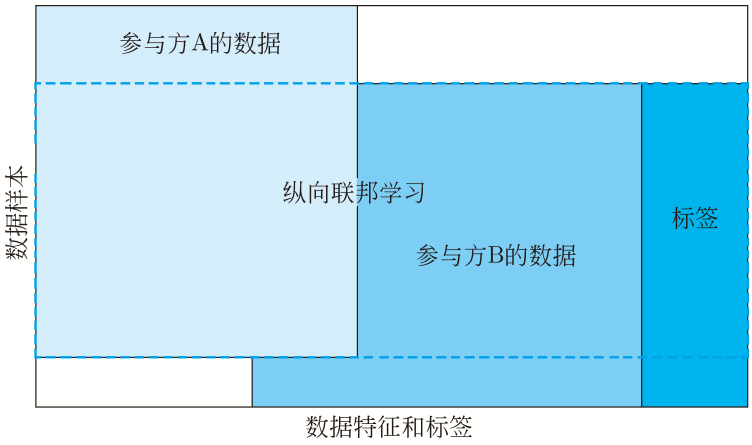

图 1.2　纵向联邦学习示意 [6]

联邦迁移学习（Federated Transfer Learning，FTL）如图 1.3 所示，它是将联邦学习与迁移学习相结合的一项新技术，其目的是在保护数据隐私的前提下，强调即使在异构特征分布的多方场景下，也能够协同并提升模型性能。文献 [8] 提出了一种安全的联邦迁移学习框架，包括基于同态加密（HE）和 secret sharing 的实现；文献 [9] 在 Google Cloud 上用 FATE（Federated AI Technology Enabler）对联邦迁移学习的性能进行了实验分析，并提出了可以提高性能的优化方案。总体来说，FTL 相对前面两种类型，当前的研究还比较少，也是今后联邦学习的重点研究方向。

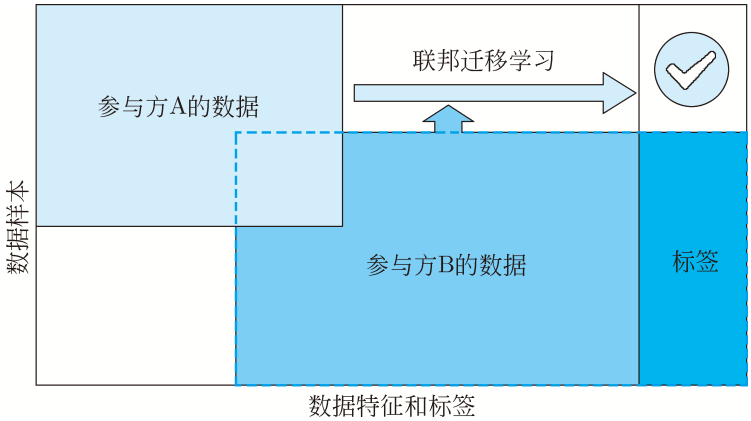

图 1.3　联邦迁移学习示意 [6]

1.1.3　联邦学习发展与现状

2016 年，联邦学习的概念被首次提出后 [2]，研究者在隐私、公平性、知识迁移、

贡献度评估、个性化等问题上进行了大量深入探索。2019 年，杨强等在“Federated Learning: Concept and applications”[6] 中第一次正式形式化定义“联邦学习”和横向联邦、纵向联邦、联邦迁移等范式。2020 年，首部系统性联邦学习技术图书《联邦学习》[10] 出版，填补了这一领域的空白，其涵盖了理论知识总结、实践案例分析和未来技术展望的完整链，为研究者和从业者提供了全面的理论与实践指导。目前，联邦学习在多个行业已经获得了广泛的应用，包括医疗健康、金融等数据敏感行业。

在医疗领域，联邦学习应用场景有多中心数据整合与分析 [11]、疾病模型预测与诊断 [12]、个性化医疗方案制定 [13] 和医疗影像分析 [14,15]。在医疗行业，数据分布在不同的医疗机构，如医院、诊所和研究机构。由于数据隐私和安全问题，这些机构很难直接共享患者数据。联邦学习通过在本地设备上进行训练，仅共享模型参数，避免了数据的集中化，保护了患者隐私。例如，不同医院可以合作训练一个疾病预测模型，而无须将各自的患者数据上传到中央服务器。联邦学习被用于训练复杂的疾病预测模型，如癌症筛查、心脏病预测和糖尿病管理等。通过整合多中心的数据，联邦学习能够获得更全面和多样化的训练数据，提高模型的泛化能力。例如，在癌症预测中，多家医院可以联合训练一个模型，提高其对不同患者群体的预测准确性。联邦学习可以帮助实现个性化医疗。通过分析大量分布式的数据，模型可以学习到不同患者的个体差异，从而提供个性化的治疗方案。例如，在药物推荐系统中，联邦学习可以结合不同患者的病历数据，推荐最合适的治疗方案和药物剂量。医疗影像分析是联邦学习的重要应用领域之一。通过联合多个医疗机构的影像数据，联邦学习可以训练出高精度的图像识别模型，用于疾病检测和诊断。例如，在肺炎检测中，多个医院的 X 光片可以用于训练一个联合模型，提高其对肺炎的检测准确性。

在金融领域，联邦学习的应用主要体现在反欺诈检测、用户信用评估、个性化金融服务和金融市场的分析与预测中。金融机构利用联邦学习技术，结合多家银行的交易数据，共同训练反欺诈检测模型，提升检测效率和准确性 [6]。欺诈检测模型需要大量的交易数据，而这些数据通常分散在不同的银行和支付机构。通过联邦学习，各机构在本地训练模型，然后共享模型参数，而不是原始数据。这种方式不仅提高了模型的检测能力，还保护了用户的隐私。研究表明，联邦学习模型在检测信用卡欺诈交易方面的表现优于单一机构训练的模型。通过联邦学习技术，金融机构可以在不共享用户数据的情况下，共同训练信用评分模型，改善信用评估的准确性 [16]。信用评分模型通常需要大量的用户数据，包括信用记录、贷款历史等。各金融机构可以利用联邦学习技术，在本地对数据进行处理和训练，并共享模型参数。这种方法不仅提高了模型的泛化能力，还减少了数据泄露的风险。实验结果显

示，联邦学习模型在信用评估中的表现显著优于传统的单一机构模型。金融机构通过联邦学习技术，结合多方数据，为用户提供个性化的金融产品和服务。个性化金融服务需要大量的用户行为数据和交易记录。通过联邦学习，金融机构可以在本地处理数据，并通过共享模型参数的方式，联合其他机构的数据进行训练，开发出更精确的个性化服务模型。金融市场分析需要大量的历史数据和实时数据，通过联邦学习，可以结合多家金融机构的数据，提升市场分析与预测的准确性。金融市场预测模型需要结合多种数据源，包括股票价格、交易量、新闻等。通过联邦学习，多个金融机构可以在本地处理这些数据，并通过共享模型参数进行联合建模。

同时，联邦学习被部署在边缘侧，也有了很多应用场景，主要体现在物联网（IoT）、智能家居、移动设备和自动驾驶等领域。通过在边缘设备上进行本地模型训练，联邦学习不仅提高了数据隐私和安全性，还降低了数据传输的延迟和带宽需求。在智能工厂中，成千上万的传感器和设备需要实时数据处理和决策。联邦学习可以在这些设备上本地训练模型，并通过共享模型参数提高整体系统的智能化水平 [17]。一个具体的应用是通过联邦学习优化设备的维护和故障检测模型，多个工厂的设备可以联合训练模型，而不需要共享敏感的生产数据。在智能家居的场景中，智能家居设备如智能音箱和语音助手需要处理大量用户的语音数据。通过联邦学习，这些设备可以在本地训练语音识别模型，保护用户隐私的同时，不断提高识别准确性 [18]。多个家庭的智能音箱可以联合训练语音识别模型，提升设备对不同口音和语速的适应性。移动设备上的应用，如新闻推荐和广告推送，需要个性化的数据处理。联邦学习可以在本地训练推荐模型，根据用户行为和偏好进行个性化推荐，同时保护用户的隐私 [19]。手机上的新闻应用可以通过联邦学习技术，在本地训练个性化推荐模型，并与其他用户的设备共享模型参数，提升推荐效果。自动驾驶车辆需要处理大量的传感器数据进行环境感知和决策。联邦学习可以在各个车辆上本地训练模型，提升整体系统的安全性和决策能力 [20]。多个自动驾驶汽车可以通过联邦学习技术，联合训练环境感知和驾驶决策模型，提高对复杂交通状况的处理能力。同时，可穿戴设备如智能手表和健身追踪器收集大量用户健康数据。通过联邦学习，这些设备可以在本地训练健康监测模型，提供个性化的健康建议，同时保护用户的隐私 [21]。

尽管联邦学习在数据隐私保护方面显示出了巨大的潜力，但它仍面临诸多挑战，如通信开销、异构性和安全问题等。未来的研究方向包括更高效的通信协议、更强的模型鲁棒性以及更完善的安全机制，以推动联邦学习在更广泛的实际应用中落地。联邦学习过程中，各设备需要频繁与中央服务器进行通信以传输模型参数，这会导致较高的通信开销，特别是在带宽受限的环境中，研究更高效的通信协议和压缩技术，如梯度压缩、模型剪枝和量化技术，以减少通信量。联邦学习模型在面

对异构数据（不同设备上的数据分布和质量不同）时，可能会导致模型的泛化能力下降，开发鲁棒的联邦学习算法，如个性化联邦学习、异构数据处理方法，以提高模型在不同设备上的性能一致性。尽管联邦学习在一定程度上保护了数据隐私，但仍然存在模型逆向工程和恶意攻击的风险（如差分攻击和中毒攻击），引入更强的隐私保护技术（如差分隐私（DP）、同态加密、联邦验证和防御机制），以提高系统的安全性和隐私保护水平。随着设备数量的增加，联邦学习系统需要处理更多的数据和更复杂的模型训练任务，面临扩展性和计算资源的挑战，研究分层联邦学习架构、动态设备调度和资源管理策略，以提高系统的可扩展性和高效性。不同应用场景对联邦学习算法的需求不同，需要算法具有更高的灵活性和适应性，以满足各种实际需求，开发通用且灵活的联邦学习框架，支持多种算法和模型的集成与调整，适应不同的数据分布和应用场景。联邦学习的广泛应用需要统一的标准和规范，以确保不同系统和平台之间的互操作性和协同工作，推动联邦学习的标准化进程，制定相关技术规范和标准，促进产业界和学术界的协同发展。

1.2 系统模型与威胁模型

1.2.1 联邦学习系统结构

联邦学习系统主要包含客户端、中央服务器、通信机制、模型聚合算法、隐私保护技术、系统管理与监控多个关键组件。

客户端是拥有数据并参与模型训练的边缘设备或节点，如手机、传感器、智能家居设备等。客户端在本地处理和存储数据，进行数据预处理和特征提取，使用本地数据训练模型，并生成本地模型更新（如模型权重或梯度）。客户端对模型更新进行加密或应用隐私保护技术，如差分隐私，以保护数据安全。

中央服务器负责协调和管理联邦学习过程，聚合来自客户端的模型更新。中央服务器初始化全局模型并分发给客户端，确保所有客户端从相同的起点开始训练。中央服务器接收来自各客户端的模型更新，使用聚合算法（如 FedAvg（Federated Averaging））计算全局模型的更新。中央服务器应用隐私保护机制，确保在聚合过程中不泄露任何单个客户端的私人信息。

通信机制用于在客户端与中央服务器之间传输模型参数和更新，确保数据传输的高效和安全。通信机制主要包含通信协议和带宽优化。通信协议定义客户端和中央服务器之间的通信方式，包括数据传输格式和加密方式。关于带宽优化的问题，使用模型压缩、梯度剪枝和差分隐私等技术，减少通信量和带宽需求，提高通信效率。

模型聚合算法用于中央服务器端，将来自不同客户端的本地模型更新聚合成全

局模型。常见的聚合算法包括 FedAvg，它通过加权平均方式聚合各客户端的模型更新。模型聚合算法通过设计不同的模型聚合方式，处理异构数据和不平衡数据，确保聚合过程对噪声和异常数据具有鲁棒性。

隐私保护技术确保联邦学习过程中不泄露任何客户端的私人数据。常见的隐私保护技术可以分为差分隐私、同态加密和安全多方计算（Secure Multi-Party Computation，SMPC）等几类。其中：差分隐私是在模型更新中添加噪声，防止敏感信息泄露；同态加密允许对加密数据进行计算，保护数据隐私；安全多方计算在多个参与方之间安全地计算函数，确保数据隐私。

系统管理与监控确保联邦学习过程的稳定运行和性能优化。其负责管理和调度客户端的训练任务，确保资源的有效利用，并监控系统性能和训练过程，检测异常情况并进行调整。同时记录系统运行日志，进行分析和优化。

1.2.2 联邦学习威胁模型

1. 攻击者设定

攻击者对联邦学习系统发动不同攻击时有不同的攻击目标，同时也需要不同的背景知识和能力，因此本节从攻击者目标、攻击者能力以及攻击者知识三个维度对安全攻击的威胁模型（Threat Model）进行分析。

1）攻击者目标

攻击者目标是降低联邦学习全局模型的性能（如准确率、$F1$ 分数等），根据其具体目标可细分为非定向攻击和定向攻击两类。其中：非定向攻击是影响模型对任意输入数据的推理；定向攻击只降低模型对特定标签的输入数据的推理准确率，而不影响或轻度影响其他标签数据的性能。以自动驾驶应用的交通标志识别模型为例，非定向攻击是使模型无法识别所有交通标志；定向攻击可以使模型将停车标志识别为限速标志，而不影响其他标志的识别。

2）攻击者能力

攻击者能力是指攻击者对联邦学习系统的角色和数据所拥有的操作权限。在现有的安全研究工作中，攻击者能力从高到低依次包括控制服务器、控制多个参与方、控制单个参与方和控制参与方训练数据。其中，控制服务器和控制参与方是指攻击者可以随意访问修改服务器或参与方的模型和数据，干扰其执行的操作；控制训练数据是指攻击者可以读取、插入或修改参与方的训练数据集。攻击要求的能力越低，在实际应用中越容易实施。

3）攻击者知识

攻击者知识是指攻击者对目标联邦学习系统所拥有的背景知识，具体包括服务

器采用的聚合算法、每轮迭代中所有参与方上传的模型更新、参与方训练数据集的数据分布等。攻击所需知识越少，在实际应用中越容易实施。

2. 攻击方式

1）隐私推理攻击和中毒攻击

隐私推理攻击一般不会改变目标模型，而是收集有关模型的特征来导致隐私和鲁棒性问题。推理攻击一般分为四种：第一种是成员推理攻击；第二种是属性推理攻击，攻击者试图获取其他用户的私有数据的属性；第三种是训练输入和标签推断攻击，这种攻击方式可以确定 FL 模型类的标签和客户端的训练输入，往往更具有破坏性；第四种是基于生成对抗网络（GAN）的推理攻击，这种情况下可以生成对抗网络来执行强大的攻击。

中毒攻击发生在联邦学习的训练阶段，可分为数据中毒和模型中毒两种方式。数据中毒主要通过添加噪声或者翻转标签来改变训练数据集，模型中毒通过操作模型更新导致全局模型偏离正常模型。

2）后门攻击和拜占庭攻击

后门攻击是指攻击者在模型训练过程中通过某种方式对模型植入后门，当后门没有被激活时，被攻击的模型与正常模型无异；当后门被激活时，模型的输出变成攻击者事先指定好的标签来达到恶意攻击的目的。拜占庭攻击旨在阻止全局模型收敛。

1.3 联邦学习系统目标

联邦学习作为一个分布式机器学习系统，其目标不仅局限于模型性能的优化，还包括隐私、安全、效率和公平多个重要的目标。这些目标共同构成了联邦学习的核心价值，使其成为在数据隐私保护和分布式计算中不可或缺的技术。联邦学习“没有免费的午餐”定理 [22] 证实了联邦学习系统不可能同时达到效能、效率、隐私的最优，而是获得一个满足不同偏好的帕累托最优前沿。研究者也在持续努力寻找推进帕累托前沿的方案。

1.3.1 隐私目标

保护隐私是联邦学习的核心目标之一。在传统的集中式机器学习方法中，数据需要被集中到中央服务器进行训练和推理，这样做可能导致数据泄露的严重风险。在联邦学习系统中，数据始终保留在联邦参与方内部，通过使用差分隐私、安全多方计算和同态加密等方式，客户端和中央服务器共享模型梯度或参数等信息，从而达到保护隐私的目标。

在横向和纵向联邦学习场景中，客户端通常与中央服务器共享模型参数或者模型梯度。然而，共享模型参数等隐私数据的间接信息并不能提供隐私安全的保证，已经有一系列工作证明了其中的安全漏洞。在模型逆向攻击 [23] 中，攻击者利用访问的模型参数或梯度信息逆向推测训练数据的某些特征或重建训练数据；在深度泄露攻击 [24] 中，攻击者使用接收到的梯度信息以及初始的模型参数反向优化一个随机生成的输入数据，使其产生的梯度与共享的梯度尽可能匹配，从而逐步恢复出原始的训练数据；在生成对抗攻击 [25] 中，攻击者利用生成式对抗网络的生成能力构建一个生成模型来模拟训练数据的分布，攻击者通过这种方式可以生成与真实训练数据高度相似的数据样本，从而间接推测出训练数据的某些特征或分布。面对联邦学习共享模型参数或模型梯度隐私保护不完备的问题，我们需要设计一系列方案，达到隐私保护目标。

在大模型参数的联邦迁移场景中，对于隐私保护又提出了新的要求。在推理场景中，大模型由于体量巨大和版权问题，往往难以进行本地部署，需要部署在云端执行推理任务。客户端需要上传提示词至中央服务器进行模型推理，从而造成隐私泄露，这就需要设计隐私保护算法，保护提示词隐私，包括提示词的概率分布和上传的辅助文件。在利用大模型辅助本地训练的场景中，中央服务器部署性能强大的基础模型，本地部署轻量化的特定功能的小模型，针对大小模型之间共享信息造成的隐私泄露问题，也对联邦学习的隐私保护提出了新的目标。

1.3.2 安全目标

在联邦学习中，除了隐私安全问题，还涉及模型的安全。攻击者试图在不被发现的情况下，在联合训练的模型中植入恶意行为或功能。这种攻击可以在多个场景中发生，且对联邦学习系统的安全性和可靠性构成严重威胁。这种攻击称为后门攻击。

具体而言，攻击者在本地数据集中插入一些带有特定触发模式的样本，这些样本与目标标签不一致。例如，在图像分类任务中，可以在部分训练图像中添加特定的噪声或图案，然后将这些图像的标签更改为攻击者想要的标签。或者，攻击者在本地模型训练阶段通过修改损失函数或梯度来引入后门，使得更新后的模型在特定输入下表现异常。攻击者通过在本地数据或模型更新中植入后门，尽可能保持模型在正常任务上的性能不变，从而不被检测到。后门通常只在特定条件下（如特定输入模式）被触发，而在其他情况下模型表现正常，这使得攻击更加隐蔽和难以发现。攻击者可以是参与联邦学习的多个节点中的一个或几个，利用联邦学习系统的分布式特性来实施攻击。