HZ BOOKS
华章IT

智能系统与技术丛书

某大型金融集团科技公司联邦学习团队负责人撰写

两位院士及多位学术界和企业界专家联袂推荐

从基础、原理、实战、拓展4个维度系统讲解联邦学习

深入浅出联邦学习

原理与实践

王健宗 李泽远 何安珣 著

Dive into Federated Learning

Principle and Practice

机械工业出版社
China Machine Press

智能系统与技术丛书

深入浅出联邦学习

原理与实践

王健宗 李泽远 何安珣 著

Dive into Federated Learning

Principle and Practice

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

深入浅出联邦学习：原理与实践 / 王健宗，李泽远，何安珣著 . -- 北京：机械工业出版社，2021.5
（智能系统与技术丛书）
ISBN 978-7-111-67959-2

Ⅰ. ①深… Ⅱ. ①王… ②李… ③何… Ⅲ. ①机器学习 Ⅳ. ① TP181

中国版本图书馆 CIP 数据核字（2021）第 063209 号

深入浅出联邦学习：原理与实践

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：杨绣国 罗词亮 责任校对：殷 虹
印 刷：三河市宏达印刷有限公司 版 次：2021 年 5 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：12.5
书 号：ISBN 978-7-111-67959-2 定 价：79.00 元

客服电话：（010）88361066 88379833 68326294 投稿热线：（010）88379604
华章网站：www.hzbook.com 读者信箱：hzit@hzbook.com

华章图书

一本打开的书，一扇开启的门，

通向科学殿堂的阶梯，托起一流人才的基石。

PRAISE

赞　誉

进入数字时代以来，机构和个人不断产生海量数据，如何将大数据与人工智能技术完美结合是人们一直在探索的问题，联邦学习技术的诞生为之提供了一条新的解决思路。本书从扎实的理论基础出发，内容涵盖联邦学习的各种框架、实战案例、应用场景和前沿研究，这些是作者在联邦学习领域的耕耘成果，更是经验分享。想要学习联邦学习技术的读者一定不要错过这本书。

——郭嵩　加拿大工程院院士/香港理工大学电子计算学系教授/IEEE Fellow/长江学者讲座教授

数字经济时代的来临，正加速各行各业完成数字化转型与业务的降本增效。如果说数据是智能时代的石油，那么联邦学习无疑是极具潜力的“石油挖掘机”。本书从联邦学习的基本概念出发，深入浅出地讲解了其技术原理，并结合实例分析了联邦学习的应用。有兴趣了解联邦学习是什么、如何实践和应用的读者读之，必大有裨益。

——李晓林　同盾科技合伙人/知识联邦产学研联盟理事长

数据是具有战略价值的核心资产，相比传统的数据授权和数据传输模式，联邦学习的优势是既能满足隐私保护要求，又能实现商业合作的诉求。本书作者结合自己在联邦学习领域长期的沉淀和研究，系统介绍了联邦学习技术的基本知识。仔细研读这本书，读者可以详尽地了解和掌握联邦学习，一定会很有收获。

——谢长生　华中科技大学武汉光电国家研究中心教授/CCF 信息存储专业委员会常委

PREFACE

前　言

联邦学习到底是什么呢？

我们认为可以这样定义：它是在数据不出本地的前提下，由多个参与方联合、协作完成建模任务的分布式机器学习范式。据统计，2020 年产生的联邦学习相关论文超过 6000 篇，是之前所有相关论文的三倍多。作为大数据时代下人工智能发展不可或缺的核心技术，联邦学习已经成为当前学术界、产业界争相研究和应用的对象。

在绝大部分的行业中，数据是以孤岛的形式存在的，即数据在不同机构或部门中独立存储、分仓管理，难以流通和利用，而人工智能的发展又往往会涉及多个领域的数据。在过去，为了打破数据孤岛，数据需求方通常会收集来自不同机构的数据信息，并统一整合到中心数据集群后进行集中处理和应用。然而，由于数据隐私泄露和数据获取成本过高，这一方法变得越来越不可取。同时，在愈发重视数据隐私安全的全球性趋势下，社会各界逐渐提升了数据所有权、资产化的保护意识，各国也逐步出台新的法律法规来严格规范数据的管理和使用。例如，2018 年 5 月，欧盟实施《通用数据保护条例》(GDPR)来保护用户的个人隐私和数据安全，禁止数据在实体间转移、交换和交易。2020 年 10 月，我国公布《中华人民共和国个人信息保护法(草案)》，为个人信息保护提供了强有力的法律保障。在法律法规强监管的环境下，如何在确保数据隐私安全的前提下解决数据孤岛问题，已然成为人工智能发展的首要挑战。

联邦学习成为打破人工智能发展困境的“头雁”，其核心价值是在数据安全合规的前提下提升模型效果，实现降本增效。那么联邦学习是如何做到的呢？对于联邦模型的训练而言，模型可以基于各参与方的本地数据库进行训练，训练过程中的模型参数通过加密机制在各参与方间通信，数据无须出本地，既保证了数据隐私安全合规，又间接共享了数据资源，促进了数据生产要素的流通。对于联邦模型的推理而言，由多

个参与方联合共建的最优模型可以在密态基础上实现金融、医疗、政务等多个行业的赋能应用。

联邦学习能有效解决人工智能发展面临的数据隐私安全与孤岛问题，这为大数据与人工智能的健康发展和颠覆式变革奠定了基础，并为其在更复杂、更前沿、更尖端领域的应用落地创造了更多的机会和可能。

为什么要写本书

联邦学习技术一经提出，便引起了社会各界人士的广泛关注。联邦学习能够满足各方在不共享数据源的前提下进行数据联合训练的需求，帮助多方组织构建最优的机器学习模型。这一技术不仅能够推动互联网时代下海量数据的价值变现，还能助力人工智能的发展革新和应用落地。

目前，联邦学习的相关学习资源过于分散，相关图书屈指可数。为了更好地普及联邦学习知识，传递联邦学习价值，我们特写作本书，旨在系统全面地介绍联邦学习的来龙去脉，为有志于联邦学习理论研究和实践的读者提供指引和参考。希望本书能够给广大读者带来启示。

读者对象

大数据、人工智能相关产业的从业者和研究人员，包括但不局限于：

- 想要全面了解、探索联邦学习的读者；
- 想要上手实践联邦学习的读者。

本书主要内容

全书共 9 章，分为 4 部分。

第一部分　基础（第 1～2 章）

主要介绍了联邦学习的概念、由来、发展历史、架构思想、应用场景、优势、规范与标准、社区与生态等基础内容。

第二部分　原理（第 3～5 章）

详细讲解了联邦学习的工作原理、算法、加密机制、激励机制等核心技术。

第三部分　实战（第 6～7 章）

主要讲解了 PySyft、TFF、CrypTen 等主流联邦学习开源框架的部署实践，并给出了联邦学习在智慧金融、智慧医疗、智慧城市、物联网等领域的具体解决方案。

第四部分　拓展（第 8～9 章）

概述了联邦学习的形态、联邦学习的系统架构、当前面临的挑战等，并探讨了联邦学习的发展前景和趋势。

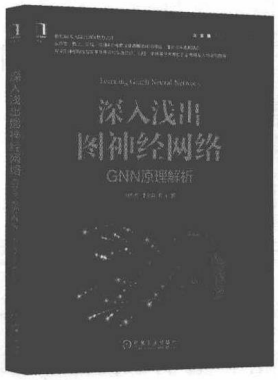
勘误与支持

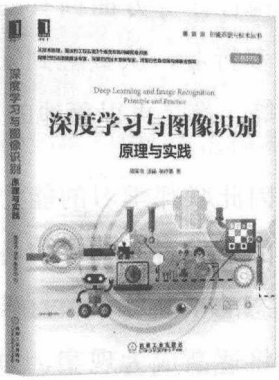
联邦学习的概念很新，更新很快，虽然我们已尽可能使本书内容准确、全面、紧跟技术前沿，但书中仍难免存在遗漏或不妥之处，恳请读者批评指正。如果你有关于本书的任何意见或建议，欢迎发送邮件到 yfc@hzbook.com，期待你的反馈。

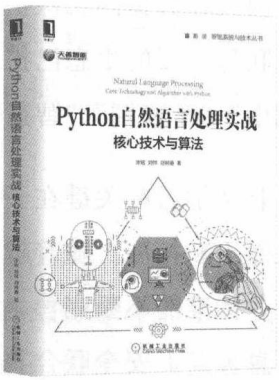
致谢

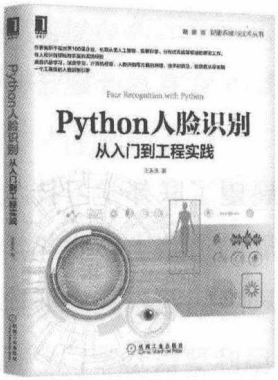
本书的写作占用了我们大量的业余时间，在此特别感谢家人、朋友的理解和支持。另外，在本书写作过程中，机械工业出版社华章公司的编辑们给予了精心指导和大力支持，没有他们细致的工作，本书无法如此顺利地出版，特此感谢。

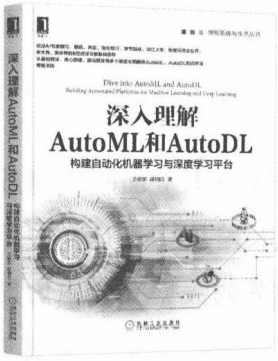
推荐阅读

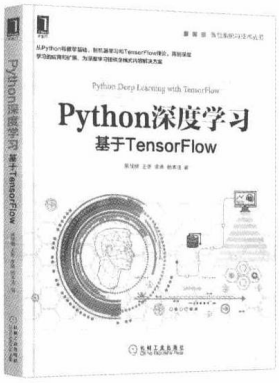


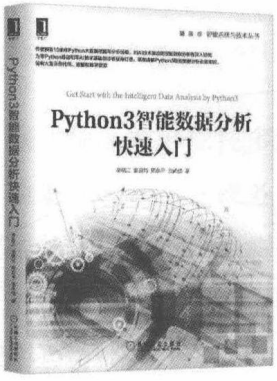


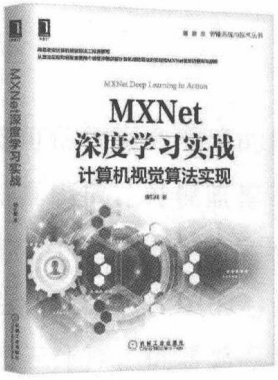


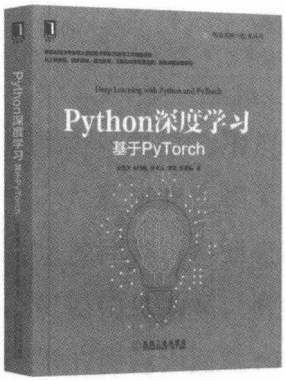












CONTENTS

目　录

第一部分

基　　础

CHAPTER 1

第 1 章

联邦学习的前世今生

联邦学习作为一种强调数据安全和隐私保护的分布式机器学习技术，在大数据与人工智能广泛发挥作用的背景下，受到具有数据监管和隐私保护需求行业的广泛关注。本章将主要介绍联邦学习的由来、发展历程及现状，并详细阐释联邦学习涉及的技术门类以及现有的生态与标准。

1.1 联邦学习的由来

人工智能自 1956 年在达特茅斯会议上被正式提出以来，经历了三轮发展浪潮。第三轮浪潮起源于深度学习技术，并实现了飞跃。人工智能技术不断发展，在不同前沿领域体现出强大活力。然而，现阶段人工智能技术的发展受到数据的限制。不同的机构、组织、企业拥有不同量级和异构的数据，这些数据难以整合，形成了一座座数据孤岛。当前以深度学习为核心的人工智能技术，囿于数据缺乏，无法在智慧零售、智慧金融、智慧医疗、智慧城市、智慧工业等更多生产生活领域大展拳脚。

大数据时代，公众对于数据隐私更为敏感。为了加强数据监管和隐私保护，确保个人数据作为新型资产类别的法律效力，欧盟于 2018 年推行《通用数据保护条例》(GDPR)。中国也在不断完善相关法律法规以规范数据的使用，例如，2017 年实施《中华人民共和国网络安全法》和《中华人民共和国民法总则》，2019 年推出《互联网个人信息安

全保护指南》，2020年推出《中共中央国务院关于构建更加完善的要素市场化配置体制机制的意见》《中华人民共和国个人信息保护法(草案)》等。这些法律条目都表明，数据拥有者需要接受监管，具有保护数据的义务，不得泄露数据。

目前，一方面，数据孤岛和隐私问题的出现，使传统人工智能技术发展受限，大数据处理方法遭遇瓶颈；而另一方面，各机构、企业、组织所拥有的海量数据又有极大的潜在应用价值。于是，如何在满足数据隐私、安全和监管要求的前提下，利用多方异构数据进一步学习以推动人工智能的发展与落地，成为亟待解决的问题。保护隐私和数据安全的联邦学习技术应运而生。

1.2 联邦学习的发展历程

人工智能自被正式提出以来，经历了60多年的演进过程，现已成为一门应用广泛的前沿交叉学科。机器学习作为人工智能最重要的分支之一，应用场景丰富，落地应用众多。

随着大数据时代的到来，各行各业对数据分析的需求剧增，大数据、大模型、高计算复杂度的算法对机器的性能提出了更高的要求。在这样的背景下，单机可能无法很好地完成数据庞大、计算复杂度高的大模型训练，于是分布式机器学习技术应运而生。分布式机器学习使用大规模的异构计算设备(如GPU)和多机多卡集群进行训练，目标是协调和利用各分布式单机完成模型的快速迭代训练。

但是，之前传统的分布式机器学习技术需要先将集中管理的数据采取数据分块并行或者模型分块并行的方式进行学习，同样面临着数据管理方数据泄露的风险，这在一定程度上制约了分布式机器学习技术的实际应用和推广。

如何结合数据隐私保护与分布式机器学习，在保证数据安全的前提下合法合规地开展模型训练工作，是目前人工智能领域的研究热点问题之一。联邦学习技术在数据不出本地的前提下对多方模型进行联合训练，既保证了数据安全和隐私，又实现了分布式训练，是解决人工智能发展困境的可行途径。

本节将主要介绍联邦学习的发展历程。首先，由于联邦学习本质上属于一种分布式机器学习技术/框架的延伸，因此本节将简要介绍机器学习与分布式机器学习的概念和重要的发展节点；其次，由于联邦学习使用了很多数据隐私保护领域的研究成果，因此本节会介绍隐私保护相关研究的历程；最后，本节将概述正处于成长阶段的联邦学习发展过程。

1. 机器学习

机器学习的提出与发展可以追溯到20世纪40年代。早在1943年，Warren McCulloch和Walter Pitts就在其论文“A logical calculus of the ideas immanent in nervous activity”㊀中描述了神经网络的计算模型。该模型借鉴了生物细胞的工作原理，试图对大脑思维过程加以仿真，引起了许多学者对神经网络的研究兴趣。1956年达特茅斯会议正式提出人工智能概念。短短3年后，Arthur Samuel就给出了机器学习的概念。所谓机器学习，就是研究和构建一种特殊算法(而非某一个特定的算法)，能够让计算机自己在数据中学习从而进行预测。

然而，由于当时的神经网络设计不当、要求进行数量庞大的计算，再加上硬件计算能力的限制，神经网络被认为是不可能实现的，机器学习的研究长期陷入停滞。直到20世纪90年代，随着云计算、异构计算等高新技术的发展，许多传统的机器学习算法被提出，并取得了良好的效果。1990年，Robert Schapire发表论文“The strength of weak learnability”㊁，文中提出弱学习集可以生成强学习，推动了机器学习领域使用Boosting算法；1995年，Corinna Cortes和Vapnik发表论文“Support-vector networks”㊂，提出支持向量机的模型；2001年，Breinman发表论文“Random forests”㊃，提出随机森林算法。随着深层网络模型和反向传播算法的提出，神经网络也重回研究视野，进入繁荣发展阶段。

㊀ McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. The bulletin of mathematical biophysics, 1943, 5(4): 115-133.

㊁ Schapire R E. The strength of weak learnability[J]. Machine learning, 1990, 5(2): 197-227.

㊂ Cortes C, Vapnik V. Support-vector networks[J]. Machine learning, 1995, 20(3): 273-297.

㊃ Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.