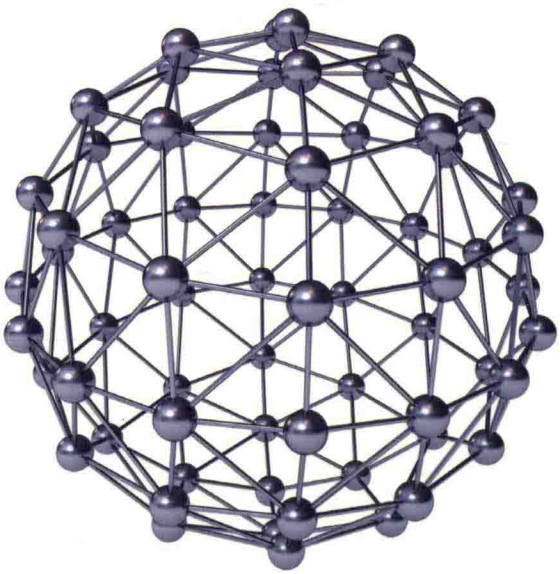
教育部高等学校计算机类专业教学指导委员会-华为ICT产学合作项目

数据科学与大数据技术专业系列规划教材

华为信息与网络技术学院指定教材

机器学习

赵卫东 董亮 ◉ 编著



系统、**完整**的数据科学与大数据技术**专业解决方案**

名校名师打造**大数据领域**精品力作

强调**基本概念**和机器学习**算法**

兼顾机器学习**经典**内容，突出深度学习**前沿**

中国工信出版集团

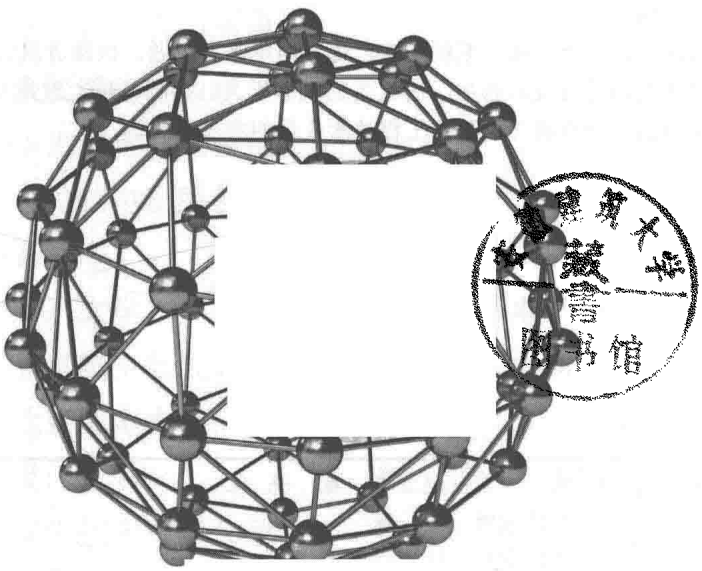
人民邮电出版社 POSTS & TELECOM PRESS

学校计算机类专业教学指导委员会-华为ICT产学合作项目
与大数据技术专业系列规划教材

华为信息与网络
技术学院指定教材

机器学习

赵卫东 董亮 ◉ 编著



人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

机器学习 / 赵卫东，董亮编著. -- 北京 : 人民邮电出版社，2018.8
数据科学与大数据技术专业系列规划教材
ISBN 978-7-115-48300-3

Ⅰ. ①机… Ⅱ. ①赵… ②董… Ⅲ. ①机器学习－教材 Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2018)第163066号

内 容 提 要

机器学习是人工智能的重要技术基础，涉及的内容十分广泛。本书涵盖了机器学习的基础知识，主要包括机器学习的概述、统计学基础、分类、聚类、神经网络、贝叶斯网络、支持向量机、进化计算、文本分析等经典的机器学习基础知识，还包括用于大数据机器学习的分布式机器学习算法、深度学习等高级内容。此外，本书还介绍了机器学习的热门应用领域推荐系统，并给出了华为机器学习平台上的实验。

本书深入浅出、内容全面、案例丰富，每章结尾都有习题，供读者巩固所学知识。

本书适合作为高等院校本科生、研究生的机器学习、数据分析、数据挖掘等课程的教材，也可作为对机器学习感兴趣的研究人员和工程技术人员的参考资料。

◆ 编　　著　赵卫东　董　亮
责任编辑　张　斌
责任印制　沈　蓉　彭志环

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
三河市潮河印业有限公司印刷

◆ 开本：787×1092　1/16
印张：23.25　　2018 年 8 月第 1 版
字数：611 千字　　2018 年 8 月河北第 1 次印刷

定价：59.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

教育部高等学校计算机类专业教学指导委员会-华为 ICT 产学合作项目
数据科学与大数据技术专业系列规划教材

编　委　会

丛书序一 PREFACE

毫无疑问，我们正处在一个新时代。新一轮科技革命和产业变革正在加速推进，技术创新日益成为重塑经济发展模式和促进经济增长的重要驱动力量，而“大数据”无疑是第一核心推动力。

当前，发展大数据已经成为国家战略，大数据在引领经济社会发展中的新引擎作用更加突显。大数据重塑了传统产业的结构和形态，催生了众多的新产业、新业态、新模式，推动了共享经济的蓬勃发展，也给我们的衣食住行带来根本改变。同时，大数据是带动国家竞争力整体跃升和跨越式发展的巨大推动力，已成为全球科技和产业竞争的重要制高点。可以大胆预测，未来，大数据将会进一步激起全球科技和产业发展浪潮，进一步渗透到我们国计民生的各个领域，其发展扩张势不可挡。可以说，我们处在一个“大数据”时代。

大数据不仅仅是单一的技术发展领域和战略新兴产业，它还涉及科技、社会、伦理等诸多方面。发展大数据是一个复杂的系统工程，需要科技界、教育界和产业界等社会各界的广泛参与和通力合作，需要我们以更加开放的心态，以进步发展的理念，积极主动适应大数据时代所带来的深刻变革。总体而言，从全面协调可持续健康发展的角度，推动大数据发展需要注重以下五个方面的辩证统一和统筹兼顾。

一是要注重“长与短结合”。所谓“长”就是要目标长远，要注重制定大数据发展的顶层设计和中长期发展规划，明确发展方向和总体目标；所谓“短”就是要着眼当前，注重短期收益，从实处着手，快速起效，并形成效益反哺的良性循环。

二是要注重“快与慢结合”。所谓“快”就是要注重发挥新一代信息技术产业爆炸性增长的特点，发展大数据要时不我待，以实际应用需求为牵引加快推进，力争快速占领大数据技术和产业制高点；所谓“慢”就是防止急功近利，欲速而不达，要注重夯实大数据发展的基础，着重积累发展大数据基础理论与核心共性关键技术，培养行业领域发展中的大数据思维，潜心培育大数据专业人才。

三是要注重“高与低结合”。所谓“高”就是要打造大数据创新发展高地，要结合国家重大战略需求和国民经济主战场核心需求，部署高端大数据公共服务平台，组织开展国家级大数据重大示范工程，提升国民经济重点领域和标志性行业的大数据技术水平和应用能力；所谓“低”就是要坚持“润物细无声”，推进大数据在各行各业和民生领域的广泛应用，推进大数据发展的广度和深度。

四是要注重“内与外结合”。所谓“内”就是要向内深度挖掘和深入研究大数据作为一门学科领域的深刻技术内涵，构建和完善大数据发展的完整理论体系和技术支撑体系；所谓“外”就是要加强开放创新，由于大数据涉及众多学科领域和产业行业门类，也涉及国家、社会、个人等诸多问题，因此，需要推动国际国内科技界、产业界的深入合作和各级政府广泛参与，共同研究制定标准规范，推动大数据与人工智能、云计算、物联网、网络安全等信息技术领域的协同发展，促进数据科学与计算机科学、基础科学和各种应用科学的深度融合。

五是要注重“开与闭结合”。所谓“开”就是要坚持开放共享，要鼓励打破现有体制机制障碍，推动政府建立完善开放共享的大数据平台，加强科研机构、企业间技术交流和合作，推动大数据资源高效利用，打破数据壁垒，普惠数据服务，缩小数据鸿沟，破除数据孤岛；所谓“闭”就是要形成价值链生态闭环，充分发挥大数据发展中技术驱动与需求牵引的双引擎作用，积极运用市场机制，形成技术创新链、产业发展链和资金服务链协同发展的态势，构建大数据产业良性发展的闭环生态圈。

总之，推动大数据的创新发展，已经成为了新时代的新诉求。刚刚闭幕的党的十九大更是明确提出要推动大数据、人工智能等信息技术产业与实体经济深度融合，培育新增长点，为建设网络强国、数字中国、智慧社会形成新动能。这一指导思想为我们未来发展大数据技术和产业指明了前进方向，提供了根本遵循。

习近平总书记多次强调“人才是创新的根基”“创新驱动实质上是人才驱动”。绘制大数据发展的宏伟蓝图迫切需要创新人才培养体制机制的支撑。因此，需要把高端人才队伍建设作为大数据技术和产业发展的重中之重，需要进一步完善大数据教育体系，加强人才储备和梯队建设，将以大数据为代表的新兴产业发展对人才的创新性、实践性需求渗透融入人才培养各个环节，加快形成我国大数据人才高地。

国家有关部门“与时俱进，因时施策”。近期，国务院办公厅正式印发《关于深化产教融合的若干意见》，推进人才和人力资源供给侧结构性改革，以适应创新驱动发展战略的新形势、新任务、新要求。教育部高等学校计算机类专业教学指导委员会、华为公司和人民邮电出版社组织编写的《教育部高等学校计算机类专业教学指导委员会-华为 ICT 产学合作项目——数据科学与大数据技术专业系列规划教材》的出版发行，就是落实国务院文件精神，深化教育供给

侧结构性改革的积极探索和实践。它是国内第一套成专业课程体系规划的数据科学与大数据技术专业系列教材，作者均来自国内一流高校，且具有丰富的大数据教学、科研、实践经验。它的出版发行，对完善大数据人才培养体系，加强人才储备和梯队建设，推进贯通大数据理论、方法、技术、产品与应用等的复合型人才培养，完善大数据领域学科布局，推动大数据领域学科建设具有重要意义。同时，本次产教融合的成功经验，对其他学科领域的人才培养也具有重要的参考价值。

我们有理由相信，在国家战略指引下，在社会各界的广泛参与和推动下，我国的大数据技术和产业发展一定会有光明的未来。

是为序。

中国科学院院士　郑志明

2018 年 4 月 16 日

丛书序二 PREFACE

在500年前的大航海时代，哥伦布发现了新大陆，麦哲伦实现了环球航行，全球各大洲从此连接了起来，人类文明的进程得以推进。今天，在云计算、大数据、物联网、人工智能等新技术推动下，人类开启了智能时代。

面对这个以“万物感知、万物互联、万物智能”为特征的智能时代，“数字化转型”已是企业寻求突破和创新的必由之路，数字化带来的海量数据成为企业乃至整个社会最重要的核心资产。大数据已上升为国家战略，成为推动经济社会发展的新引擎，如何获取、存储、分析、应用这些大数据将是这个时代最热门的话题。

国家大数据战略和企业数字化转型成功的关键是培养多层次的大数据人才，然而，根据计世资讯的研究，2018年中国大数据领域的人才缺口将超过150万人，人才短缺已成为制约产业发展的突出问题。

2018年初，华为公司提出新的愿景与使命，即“把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界”，它承载了华为公司的历史使命和社会责任。华为企业BG将长期坚持“平台+生态”战略，协同生态伙伴，共同为行业客户打造云计算、大数据、物联网和传统ICT技术高度融合的数字化转型平台。

人才生态建设是支撑“平台+生态”战略的核心基石，是保持产业链活力和持续增长的根本，华为以ICT产业长期积累的技术、知识、经验和成功实践为基础，持续投入，构建ICT人才生态良性发展的使能平台，打造全球有影响力的ICT人才认证标准。面对未来人才的挑战，华为坚持与全球广大院校、伙伴加强合作，打造引领未来的ICT人才生态，助力行业数字化转型。

一套好的教材是人才培养的基础，也是教学质量的重要保障。本套教材的出版，是华为在大数据人才培养领域的重要举措，是华为集合产业与教育界的高端智力，全力奉献的结晶和成果。在此，让我对本套教材的各位作者表示由衷的感谢！此外，我们还要特别感谢教育部高等学校计算机类专业教学指导委员会副主任、北京大学陈钟教授以及秘书长、北京航空航天大学马殿富教授，没有你们的努力和推动，本套教材无法成型！

同学们、朋友们，翻过这篇序言，开启学习旅程，祝愿在大数据的海洋里，尽情展示你们的才华，实现你们的梦想！

华为公司董事、企业BG总裁　阎力大

2018年5月

前言 FOREWORD

机器学习是人工智能的技术基础，伴随着人工智能几十年的发展，期间有过几次大起大落。作为机器学习的高级阶段，最近几年，深度学习算法在自然语言处理、语音识别、图像处理等领域的突破，使得机器学习成为计算机学科非常热门的一个方向。这也标志着机器学习已经彻底迈出实验室大门，走向实践，推动着人工智能向更高阶段发展。

与机器学习十分密切的概念有数据挖掘、大数据分析等，这些数据分析技术使用了一些机器学习的方法和算法，解决了企业应用的一些问题，辅助业务人员和管理人员做出更好的决策。几种技术相辅相成，共同促进了数据分析技术和人工智能的进步。

从早期的统计学习，发展到联结主义的神经网络，直到深度神经网络的过程中，机器学习的基础一直是足量的高质量数据。“互联网+”的热潮推动了大数据的产生以及处理大数据的软硬件技术的迅猛发展，为机器学习提供了更好的数据和分析的技术基础，在一些应用领域，机器已经达到甚至超过人类的智能水平，从而引发机器学习在金融、智能制造、零售、电子商务、电信等众多行业的广泛应用。

大数据应用和人工智能发展已经引起了全球关注，企业对机器学习的人才需求增大，与之密切相关的数据科学家、数据挖掘工程师、大数据分析师、机器学习分析师等数据分析类人才成为21世纪最有吸引力的人才。根据相关企业估计，上述人才的需求增长将持续6～8年。为此，教育部批准了一批高校成立数据科学与大数据技术、大数据应用、智能科学与技术等相关专业，为业界培养相应的专业人才。

机器学习是一门理论和实践并重的课程，内容比较多，很多算法也有一定的难度。此外，机器学习的应用需要一定的经验和技巧。编写一本兼顾机器学习理论的系统性，又能体现机器学习的应用，使之适合新工科教学的机器学习教材，就成为本书编写的目标。

目前，市场上已有多种机器学习的专业图书和教材，大多数图书有两种倾向，要么理论内容繁杂，对初学者和任课教师来说挑战较大，短短的几十学时难以消化；要么偏重应用和实践，理论的系统性不够，理论与应用方面的内容不够平衡。针对上述问题，作者参阅了大量文献资料，结合过去多年数据分析的研究和实践，重新梳理了机器学习的整个课程体系，使内容基本覆盖机器学习的基础内容，深入浅出，读者在此基础上可以钻研机器学习的高级算法。

本书具有以下特点。

（1）大多数章节都有典型的 Python 算法和案例，深入浅出地解释理论，方便学习理解。本书最后附有主要参考文献，方便读者加深对教材内容的思考。此外，每章还配有思考题，以此检验读者对基本知识的理解和应用能力。

（2）在介绍传统的机器学习理论的基础上，突出了机器学习目前主流的一些内容，包括深度学习的典型算法与应用、知识图谱、机器学习在电子推荐技术的应用等。

（3）本书第 14 章的华为 FusionInsight 平台提供 3 个月的免费使用期，请读者根据实际情况开通使用。

本书在写作过程中，得到了教育部高等学校计算机类专业教学指导委员会-华为 ICT 产学合作项目组的大力支持，另外，研究生于召鑫、蒲实、朱荣斌、耿甲、袁雪如、陈子轩等在资料收集过程中做了一些工作，在此一并表示感谢。

赵卫东

2018 年 2 月

复旦大学

目 录 CONTENTS

01 第1章 机器学习概述

随着大数据的发展和计算机运算能力的不断提升，人工智能在最近几年取得了令人瞩目的成就。目前在很多行业中，都有企业开始应用机器学习技术，从而获取更深刻的洞察，为企业经营或日常生活提供帮助，提升产品服务水平。机器学习已经广泛应用于数据挖掘、搜索引擎、电子商务、自动驾驶、图像识别、量化投资、自然语言处理、计算机视觉、医学诊断、信用卡欺诈检测、证券金融市场分析、游戏和机器人等领域，机器学习相关技术的进步促进了人工智能在各个领域的发展。