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

R语言基础与数据科学应用

沈刚 ◉ 主编

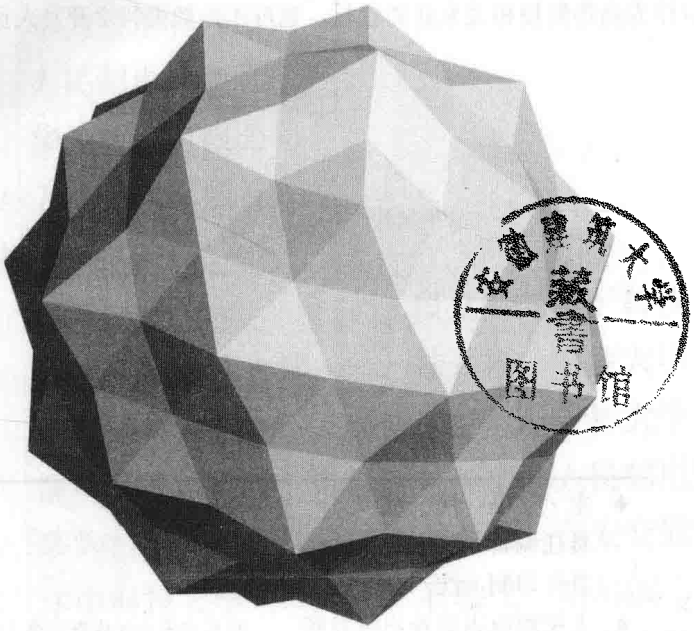
系统、完整的数据科学与大数据技术专业解决方案
名校名师打造大数据领域精品力作
提供数据挖掘与机器学习算法及R语言实现
掌握使用R语言进行大数据分析的基本过程

中国工信出版集团
人民邮电出版社 POSTS & TELECOM PRESS

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

R语言基础与数据科学应用

沈刚◉主编



人民邮电出版社
北京

图书在版编目（CIP）数据

R语言基础与数据科学应用 / 沈刚主编. -- 北京 : 人民邮电出版社, 2018.7
数据科学与大数据技术专业系列规划教材
ISBN 978-7-115-48302-7

Ⅰ. ①R… Ⅱ. ①沈… Ⅲ. ①程序语言－程序设计－教材 Ⅳ. ①TP312

中国版本图书馆CIP数据核字(2018)第091991号

内 容 提 要

本书是为初学者学习 R 语言基础及其在数据科学中的应用而编写的。全书内容包括三个部分，分别介绍了 R 语言的编程基础知识，数据处理、可视化和统计分析的实用技术，以及在机器学习、神经网络和深度学习中的具体应用。读者可以通过本书了解和体验 R 语言的风格特点和强大功能。本书中所有程序均在 R 3.4.3 环境下调试通过。

本书既可以作为高等院校相关专业的教材，也可作为数据科学开发人员的参考书。

◆ 主　　编　沈　刚
责任编辑　邹文波
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
北京圣夫亚美印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：19　　2018 年 7 月第 1 版
字数：461 千字　　2018 年 7 月北京第 1 次印刷

定价：49.80 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。

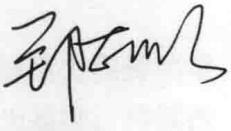
习近平总书记多次强调“人才是创新的根基”“创新驱动实质上是人才驱动”。绘制大数据发展的宏伟蓝图迫切需要创新人才培养体制机制的支撑。因此，需要把高端人才队伍建设作为大数据技术和产业发展的重中之重，需要进一步完善大数据教育体系，加强人才储备和梯队建设，将以大数据为代表的新兴产业发展对人才的创新性、实践性需求渗透融入人才培养各个环节，加快形成我国大数据人才高地。

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。

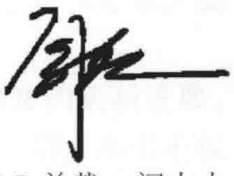
国家大数据战略和企业数字化转型成功的关键是培养多层次的大数据人才，然而，根据计世资讯的研究，2018 年中国大数据领域的人才缺口将超过 150 万人，人才短缺已成为制约产业发展的突出问题。

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

随着大数据技术的快速发展，数据科学已成为支撑其他科学研究领域，以及交通、旅游、金融和其他商业服务业中的大量热门应用的重要力量。目前的就业市场对具有数据科学专业背景和相关技能人员的需求一直居高不下。针对上述情况，截至 2018 年 4 月，我国已有 280 多所高等院校陆续开设了数据科学与大数据技术专业。

R 是在 GNU 协议框架下的一种自由、免费的开源软件，它既是一个功能强大的统计计算与可视化环境，也是一门解释型的、面向对象的程序设计语言。自 20 世纪 90 年代诞生以来，经过 20 多年的持续演化后，R 语言已经成为当前数据科学研究与应用中最受欢迎的程序设计语言之一。掌握 R 语言对成为一名成功的数据科学工作者无疑有着重要的意义。

R 语言具有简洁优雅、功能扩展性好、代码可读性强等特点，很适合作为零基础的初学者学习编程的入门语言。同时，R 可以在几乎所有的主流操作系统平台上运行。在 R 社区里，为数众多的开发者不断为 R 免费提供各种先进的开发包，正是他们的努力与贡献使得 R 日益完善。

本书是为初学者学习 R 语言基础以及在数据科学中的应用而编写的。全书内容分为如下三部分。

第 1 部分介绍 R 的基础知识，由第 1 章、第 2 章、第 3 章、第 4 章组成，内容包括引言、R 的数据与运算、R 程序设计基础、R 语言面向对象编程等知识。

第 2 部分介绍 R 编程的一些实用技术，由第 5 章、第 6 章、第 7 章组成，内容包括 R 支持的数据结构与数据处理、绘图与数据可视化编程、统计与回归分析等知识。

第 3 部分是关于 R 在数据科学的具体应用，由第 8 章、第 9 章组成，内容包括统计机器学习、神经网络与深度学习。第 9 章还特别介绍了与深度学习有关的几个 R 语言包。

本书配有大量的程序示例和使用 R 绘制的图形，所有代码均在 R 3.4.3 环境下调试通过。本书还提供教学 PPT 课件和源程序文件等，若读者需要，可以登录人民邮电出版社教育社区（http://www.ryjiaoyu.com.cn）免费下载。

本书在内容的选择及深度的把握上同时考虑了初学者的需要和 R 的最新进展，在内容安排上力求循序渐进，通过可以重现的示例引导初学者动手实践。本书不仅适合于教学，也可供相关专业领域各类人员自学使用。

最后，特别感谢杨潇、杨明祺和郭义同学在本书编写过程中给予的支持和协助，特别是在审阅初稿时提出了宝贵的意见和建议。由于编者的知识和水平有限，书中难免存在不足之处，敬请广大读者批评指正。

编　者

2018 年 4 月

目 录 CONTENTS

1 第1章 引言

Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.

——Mike Loukides

R 是一个广泛适用于统计计算、数据分析和其他科学研究的软件环境，也是一种支持复杂的数据处理、数据可视化及机器学习的编程语言。换句话说，R 是数据科学家的得力助手。为了避免对软件环境和编程语言的混淆，本书会把前者称为 R 系统，而把后者称作 R 语言以示区分。

本章是关于 R 的一些入门知识。首先，我们简单回顾 R 的起源与发展，然后介绍如何下载、安装、运行 R 系统。当读者在自己的计算机上建立好 R 环境后，就可以开始学习使用 R 语言来实现简单的编程。在本章的最后，还会讨论使用 R 语言开发程序时可以用来提高工作效率的一些方法，例如，用集成开发环境（Intergrated Development Environment，IDE）编写并调试程序，以及如何获取必要的帮助信息等。

本章所介绍的内容是关于 R 的基础知识。已经具备其他编程语言开发经验或者其他数据分析处理软件知识的读者，可以有选择地阅读本章的相关部分。

本章的主要内容包括：

（1）R 环境的安装与启动；

（2）包的使用与管理；

（3）工作空间。

1.1 R 的起源与发展

近几年来，无论是在哪一种公认的最受欢迎编程语言的排行榜上都不难找到 R 的名字。和其他流行的编程语言相比，不少初学者可能会觉得 R 比较陌生，把它当作一门略显小众的编程语言。有相关统计分析与数据挖掘经验的人都知道，R 除了拥有其他编程语言不具备的强大的统计与绘图功能之外，还能提供友好的用户交互方式和良好的语法表达能力，已经被大量的统计学家、数据分析师、市场营销人员和科研工作者用于数据的检索、清洗、分析、可视化和呈现。这里对 R 的产生和特点做一些初步的介绍。

1.1.1 R 的产生与演化

R 语言是一种开源的脚本语言，一直以来在数据分析与预测，以及数据可视化等方面享有良好的声誉。早在 1993 年，R 的最初版本就发布给统计学家以及其他掌握了深入编程能力的研究人员使用，去解决他们面对的复杂数据统计分析任务，并用多样化的图形来展示结果。据说，R 的名字来源于它当时的两名开发者，新西兰奥克兰大学的 Ross Ihaka 和 Robert Gentleman，两人名字的首字母都是 R。

传统上，在统计分析与计算领域存在着三大主流软件：SAS、SPSS 和 S。SAS（Statistical Analysis System，统计分析系统）是最早由北卡罗来纳州立大学开发，现在由 SAS 研究所维护与销售的一种统计分析软件；SPSS（Statistical Product and Service Solutions，统计产品与服务解决方案软件）起初是由斯坦福大学的几名学生开发的，现在由 SPSS 公司经营；S 语言是 John Chambers 和他的同事们于 1976 年在 AT&T 贝尔实验室开发的一种专用于统计分析的解释型语言。

R 可以看作是对 S 语言的继承与发展。尽管 S 和 R 有一些显著的区别，但用 S 语言编写的大部分代码在 R 上依然可以运行。简单地说，R 是一个有着强大统计分析功能及绘图功能的软件环境，也是由 S 语言发展出来的编程语言。因此，R 既是一套软件系统，也包括了一种程序设计语言。现在，S 语言的商业版就是由 TIBCO 软件公司运营的 S-PLUS 软件。而 R 系统则是开源、免费的。R 在 GNU（General Public Licence）协议下开源并免费发行，R 语言社区中的大量开发者不断为其发展做出自己的贡献。目前 R 的开发及维护由 R 开发核心小组（R Development Core Team）具体负责。

从数据分析软件的角度来看，数据科学家、统计学家、分析师、金融工程师使用 R 作为一种数据处理的工具，对采集到的数据进行统计分析，完成可视化、建模和预测等任务。从程序设计语言的角度来看，人们可以使用 R 语言编写函数和脚本，来完成所需的数据分析工作。R 提供了完整的交互式的面向对象开发方法。在早期，R 是一种统计学家为统计学家专门设计的程序语言，现在的 R 语言支持对象、运算符和函数，已将数据的探索、建模与可视化等工作有机地融为一体。

正是由于 R 本身兼顾了软件环境与程序设计语言的特性，R 的使用者往往只要输入几

行代码，就可以实现复杂的数据分析目标。举例来看，在数据分析中经常会遇到的线性回归、非线性回归、聚类、分类等工作，以及画出相应结果的图形，都可以用几个 R 语言自带的函数或是 R 语言包中的函数以十分简洁的方式实现。通常，这些工作只需要简单的数据预处理和参数设置，就可以直接调用 R 函数。

由于其自身的吸引力，R 语言逐渐超越了学术界的小圈子，进入大众的视野，开始被一些企业选用来完成各自的商业目标。随着越来越多的数据分析师开始接触 R 并且向同行推荐 R，用户群的扩大增加了 R 的影响力，从而引发了更多研究者对 R 的兴趣。因此，除了一些标准的统计分析功能之外，很多数据科学领域最新的成果也被率先转化为 R 语言中的工具。伴随着学术界与相关产业对数据科学重视程度的日益增加，R 语言正在不断拓展自己的边界。例如，深度学习近来取得了很多受人瞩目的成果，原来在 Python 中率先实现的技术也就迅速地被纳入到 R 中来。

正是因为 R 开源、免费（SPSS 和 SAS 都是商业运行的软件），支持所有的主流操作系统平台，拥有活跃而数量庞大的用户社区（他们贡献的包是 R 语言非常重要的组成部分），才能使其成为集统计、数据分析、可视化和机器学习等功能于一身的一种适用于数据科学领域的强有力的工具。

1.1.2 R 的特点

R 语言起源于基于函数式编程范式而设计的 S 语言，采用的是面向数学函数的抽象，因此，用户在 R 控制台中的交互过程实质上类似于使用计算器，是一个用户提交需要计算的函数，由 R 环境完成对函数赋值的循环。同时，R 语言也在越来越多地支持面向对象的设计思想，无论是变量，还是函数，在 R 中都被视为对象。在这里，我们不去研究 R 语言本身的特点，仅从用户的角度来谈谈 R 对数据科学的研究与应用起到支撑作用的几个特色。

1. 适用于统计计算和机器学习

前面提到过，R 是用于统计计算和图形显示的开源软件环境和编程语言。与其他由计算机专家或工程师创造的主流编程语言不同，R 最初是由少数统计学家为其他统计学家开发的。因此，R 能有效地处理和存储数据，并且为统计分析提供了大量的功能。R 语言涉及统计分析和数据挖掘的众多应用领域，包括新闻传媒、市场分析和科学研究等。一些重要的科技公司已经在用户行为分析、新产品设计的决策支持、事件影响预判等工作中大量使用 R，另一些企业则把是否掌握 R 语言作为招聘数据科学家和量化分析师的重要依据。由于数据科学与大数据技术等领域的快速发展，近年来，R 的知名度得到了很大的提升。由于 R 具有非常活跃而且规模庞大的用户社区，很多开发者把最新的数据科学研究成果转化成 R 语言的包来扩展 R 的功能，供其他用户使用，这就使得 R 能不断适应用户的需要。

2. 简单易学，具有高度的灵活性

一些非专业人员刚开始接触 R 时可能会觉得 R 内容太多，包罗万象，似乎不易掌握。