大数据人才培养规划教材

以解决实际问题为**学习目标**
以实战案例贯穿为**学习手段**

大数据数学基础

R语言描述

Mathematical Basis of Big Data (R)

程丹 张良均 ◉ 主编
叶提芳 柳扬 刘晓玲 ◉ 副主编

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

大数据人才培养规划教材

大数据数学基础

R语言描述

Mathematical Basis of Big Data (R)

程丹 张良均◉主编
叶提芳 柳扬 刘晓玲◉副主编

人民邮电出版社
北京

图书在版编目（CIP）数据

大数据数学基础 : R语言描述 / 程丹, 张良均主编
. -- 北京 : 人民邮电出版社, 2019.3
大数据人才培养规划教材
ISBN 978-7-115-49922-6

Ⅰ. ①大… Ⅱ. ①程… ②张… Ⅲ. ①计算机科学-数学－教材 Ⅳ. ①TP301.6

中国版本图书馆CIP数据核字(2018)第244914号

内 容 提 要

本书全面地讲解了在科学领域运用广泛的数据微积分、线性代数、统计学、数值计算、多元统计分析等数学基础知识。全书共6章：第1章介绍了大数据与数学、数学与R语言的关系；第2章介绍了微积分的基础知识，包括函数、极限、导数、微分、不定积分与定积分及其应用；第3章介绍了线性代数的基础知识，包括矩阵的运算、行列式、特征分解、奇异值分解；第4章介绍了统计学的基础知识，包括数据分布特征、概率论、随机变量的数字特征、参数估计、假设检验；第5章介绍了数值计算的基础知识，包括插值方法、函数逼近与拟合、非线性方程（组）求根；第6章介绍了常用的多元统计分析方法，包括回归分析、聚类分析、判别分析、主成分分析、因子分析和典型相关分析。本书中的几乎所有实例都结合R语言进行求解分析，所有章后都有课后习题，可以帮助读者巩固所学的内容。

本书可以作为高校大数据技术类专业的教材，也可作为大数据技术爱好者的自学用书。

◆ 主　　编　程　丹　张良均
副 主 编　叶提芳　柳　扬　刘晓玲
责任编辑　左仲海
责任印制　马振武

◆ 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
大厂聚鑫印刷有限责任公司印刷

◆ 开本：787×1092　1/16
印张：16.25　　2019年3月第1版
字数：373千字　　2019年3月河北第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

大数据专业系列图书
编写委员会

序 PREFACE

随着大数据时代的到来，移动互联网络和智能手机迅速普及，多种形态的移动互联应用蓬勃发展，电子商务、云计算、互联网金融、物联网等不断渗透并重塑传统产业，大数据当之无愧地成了新的产业革命核心。

未来 5～10 年，我国大数据产业将会是一个飞速发展时期，社会对大数据相关专业人才有着巨大的需求。目前，国内各大高校都在争相设立或准备设立大数据相关专业，以适应地方产业发展对战略性新兴产业的人才需求。

人才培养离不开教材，大数据专业是 2016 年才获批的新专业，目前还没有成套的系列教材，已有教材也存在企业案例缺失等亟须解决的问题。由广州泰迪智能科技有限公司和人民邮电出版社策划，校企联合编写的这套图书，犹如大旱中的甘露，可以有效解决高校大数据相关专业教材紧缺的困难。

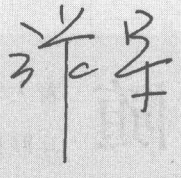
实践教学是在一定的理论指导下，通过引导学习者的实践活动，从而传承实践知识、形成技能、发展实践能力、提高综合素质的教学活动。目前，高校教学体系的设置有诸多限制因素，过多地偏向理论教学，课程设置与企业实际应用契合度不高，学生无法把理论转化为实践应用技能。课程内容设置方面看似繁多又各自为“政”，课程冗余、缺漏，体系不健全。本套图书的第一大特点就是注重学生的实践能力培养，根据高校实践教学中的痛点，首次提出“鱼骨教学法”的概念。以企业真实需求为导向，学生所学技能紧紧围绕企业实际应用需求，将学生需掌握的理论知识，通过企业案例的形式进行衔接，达到知行合一、以用促学的目的。

大数据专业应该以大数据技术应用为核心，紧紧围绕大数据应用闭环的流程进行教学，才能够使学生从宏观上理解大数据技术在行业中的具体应用场景及应用方法。高校现有的大数据课程集中在如何进行数据处理、建模分析、参数调整，使得模型的结果更加准确。但是，完整的大数据应用却是一个容易被忽视的部分。本套图书的第二大特点就是围绕大数据应用的整个流程，从数据采集、数据迁移、数据存储、数据

分析与挖掘，最终到数据可视化，覆盖完整的大数据应用流程，涵盖企业大数据应用中的各个环节，符合企业大数据应用真实场景。

我很高兴看到这套书的出版，也希望这套书能给更多的高校师生带来教学上的便利，帮助读者尽快掌握本领，成为有用之才！

教育部长江学者特聘教授
国家杰出青年基金获得者
IEEE Fellow
华南理工大学计算机与工程学院院长



2017 年 12 月

前 言 FOREWORD

随着云时代的来临，大数据分析技术将帮助企业用户在合理的时间内获取、管理、处理及整理海量数据，为企业经营决策提供积极的帮助。大数据分析作为一门前沿技术，广泛应用于物联网、云计算、移动互联网等战略性新兴产业。在大数据的研究和应用中，数学是其坚实的理论基础，在数据处理、数据挖掘、评判分析等过程中，数学方法扮演着至关重要的角色。

本书致力于传播大数据分析技术的基础数学知识，以期通过理论结合实践的方式，帮助读者运用相关数学知识解决一些实际问题。

本书特色

本书将理论与实践相结合，通过实例与大量代码实现，深入浅出地介绍了大数据分析技术所需的数学基础，并引导读者利用所学知识解决问题。本书通过例题和课后习题帮助读者巩固所学知识，使读者真正理解并能够应用所学知识。本书内容由浅入深，第 1 章介绍大数据与数学之间的联系，让读者在宏观上了解学习大数据分析技术所需要的数学知识，以及 R 语言中常用于数学计算和统计计算的程序包；第 2～5 章全面地介绍了微积分、线性代数、统计学、数值计算在数据科学领域的简单应用；第 6 章结合前 5 章的知识，介绍了数据分析过程中常用的数学方法，并结合 R 语言对实例进行求解分析。

本书适用对象

- 开设有大数据分析课程的高校的教师和学生

目前，国内不少高校将大数据分析引入教学，在数学、计算机、自动化、电子信息、金融等专业开设与大数据分析技术相关的课程，但目前这一课程的前置基础课——数学的教学仍然主要限于理论介绍。因为单纯的理论教学过于抽象，学生理解起来往往比较困难，教学效果也不甚理想。本书提供的基于 R 语言的实践教学模式，能够使师生充分发挥互动性和创造性，实现最佳的教学效果。

- 大数据分析相关从业人员

这类人员可以通过本书理解大数据分析技术中常见算法背后的理论原理，并掌握相关实现方法等知识，从而能够对算法有一个全面且深入的了解。同时，也能够通过阅读本书对大数据分析方法等有所启发。

- 机器学习与数据挖掘从业人员

这类人员可以通过本书理解机器学习常用算法的基本实现方法，从而设计出更加高效、流畅的算法，以辅助生产，为决策提供依据。

- 数学爱好者

本书不仅介绍了微积分、线性代数、统计学、数值计算和多元统计分析的基础知识，还用 R 语言实现了绝大部分的理论与算法，可满足数学爱好者的求知欲望。

代码下载及问题反馈

为了帮助读者更好地使用《大数据数学基础（R 语言描述）》这本书，我们配套提供了相关计算过程的原始数据文件、R 语言程序代码，读者可以从泰迪云课堂（https://edu.tipdm.org/course/96）免费下载，也可登录人民邮电出版社教育社区（http://www.ryjiaoyu.com）下载。此外，为了帮助读者更好地学习，泰迪云课堂（https://edu.tipdm.org）还提供了配套的教学视频。

为满足教师授课需要，我们还提供了 PPT 课件，读者可以从泰迪云课堂（https://edu.tipdm.org/course/96）下载申请表，填写后发送至指定邮箱索取所需课件。对于其他图书资源，读者也可通过拨打热线电话（40068-40020）或扫描以下二维码关注微信公众号后咨询获取。

我们已经尽最大努力避免在文本和代码中出现错误，但是由于水平有限，而且编写时间仓促，书中难免出现一些疏漏和不足的地方。如果您有更多的宝贵意见，欢迎发送邮件至邮箱 13560356095@qq.com 给予反馈。同时，本书内容更新将及时在泰迪云课堂（https://edu.tipdm.org/course/96）上发布，读者可以登录网站或关注泰迪大数据挖掘微信公众号（TipDataMining）查阅相关信息。更多本系列图书的信息可以在“泰迪杯”数据挖掘挑战赛网站（http://www.tipdm.org/tj/index.jhtml）查阅。

编　者

2018 年 10 月

目录 CONTENTS

第 1 章　绪论 …… 1

1.1　大数据与数学 …… 1

1.1.1　大数据的定义 …… 1

1.1.2　数学在大数据领域的作用 …… 2

1.2　数学与 R 语言 …… 4

1.2.1　base …… 5

1.2.2　stats …… 5

小结 …… 6

课后习题 …… 6

第 2 章　微积分基础 …… 8

2.1　函数与极限 …… 8

2.1.1　映射与函数 …… 9

2.1.2　数列与函数的极限 …… 14

2.1.3　极限运算法则与存在法则 …… 17

2.1.4　连续函数的运算与初等函数的连续性 …… 18

2.2　导数与微分 …… 19

2.2.1　导数的概念 …… 19

2.2.2　函数的求导法则 …… 24

2.2.3　微分的概念 …… 26

2.3　微分中值定理与导数的应用 …… 30

2.3.1　微分中值定理 …… 30

2.3.2　函数的单调性与曲线的凹凸性 …… 31

2.3.3　函数的极值与最值 …… 34

2.4　不定积分与定积分 …… 39

2.4.1　不定积分的概念与性质 …… 40

2.4.2　换元积分法与分部积分法 …… 44

2.4.3　定积分的概念与性质 …… 46

2.4.4　定积分的换元法与分部积分法 …… 50

2.4.5　不定积分与定积分的实际应用 …… 51

小结 …… 53

课后习题 …… 54

第 3 章　线性代数基础 …… 56

3.1　矩阵及其运算 …… 56

3.1.1　矩阵的定义 …… 56

3.1.2　特殊矩阵 …… 57

3.1.3　矩阵的运算 …… 61

3.1.4　矩阵行列式 …… 65

3.1.5　矩阵的逆 …… 78

3.1.6　矩阵的秩 …… 80

3.2　矩阵的特征分解与奇异值分解 …… 84

3.2.1　特征分解 …… 84

3.2.2　奇异值分解 …… 96

小结 …… 100

课后习题 …… 101

第 4 章　概率论与数理统计基础 …… 103

4.1　数据分布特征的统计描述 …… 103

4.1.1　集中趋势度量 …… 103

4.1.2　离散趋势度量 …… 110

4.1.3　偏度与峰度的度量 …… 115

4.2　随机事件及其概率 …… 117

4.2.1　随机事件的定义 …… 117

4.2.2　随机事件的概率 …… 119

4.3　随机变量与概率分布 …… 122

4.3.1　随机变量的定义 …… 122

4.3.2　随机变量的分布函数 …… 122

4.4　随机变量的数字特征 …… 127

4.4.1　随机变量的数学期望 …… 127

4.4.2　随机变量的方差 …… 130

4.4.3 协方差与相关系数 …… 132
4.4.4 协方差矩阵与相关矩阵 …… 134
4.5 参数估计与假设检验 …… 137
4.5.1 参数估计 …… 137
4.5.2 假设检验 …… 139
小结 …… 142
课后习题 …… 142
第 5 章 数值计算基础 …… 144
5.1 数值计算的基本概念 …… 144
5.1.1 误差的来源 …… 144
5.1.2 误差分类 …… 146
5.1.3 数值计算的衡量标准 …… 147
5.2 插值法 …… 147
5.2.1 Lagrange 插值 …… 147
5.2.2 线性插值 …… 150
5.2.3 样条插值 …… 152
5.3 函数逼近与拟合 …… 153
5.3.1 数据的最小二乘线性拟合 …… 153
5.3.2 函数的最佳平方逼近 …… 155
5.3.3 数据的多变量拟合 …… 158
5.3.4 数据的非线性曲线拟合 …… 160
5.4 非线性方程（组）求根 …… 162
5.4.1 二分法求解非线性方程 …… 163
5.4.2 Newton 法求解非线性方程 …… 165
5.4.3 Newton 法求解非线性方程组 …… 166
小结 …… 169
课后习题 …… 170
第 6 章 多元统计分析 …… 172
6.1 回归分析 …… 172
6.1.1 一元线性回归 …… 172
6.1.2 多元线性回归 …… 178
6.1.3 Logistic 回归 …… 184
6.2 聚类分析 …… 189
6.2.1 距离和相似系数 …… 189
6.2.2 系统聚类法 …… 193
6.2.3 动态聚类法 …… 198
6.3 判别分析 …… 200
6.3.1 距离判别 …… 200
6.3.2 贝叶斯判别 …… 204
6.3.3 费希尔判别 …… 205
6.4 主成分分析 …… 206
6.4.1 总体主成分 …… 207
6.4.2 样本主成分 …… 209
6.5 因子分析 …… 211
6.5.1 正交因子模型 …… 212
6.5.2 参数估计 …… 214
6.5.3 因子旋转 …… 218
6.5.4 因子得分 …… 220
6.6 典型相关分析 …… 222
6.6.1 总体典型相关 …… 222
6.6.2 样本典型相关 …… 223
6.6.3 典型相关系数的显著性检验 …… 228
小结 …… 229
课后习题 …… 230
附录 I *t* 分布表 …… 236
附录 II *F* 分布表 …… 238
参考文献 …… 250

第1章 绪论

当今社会，几乎所有的人类活动都会产生数据，例如，各类具备全球定位系统（Global Positioning System，GPS）功能的交通工具会定时产生位置数据；家用智能热水器能够记录用户每日用水的各项数据；手机中的各类 App 能够收集用户不同领域的偏好数据等。管理和使用这些数据，促进了一个全新的领域——数据科学领域的发展，而数据科学领域的基石就是数学。

本章将通过介绍大数据的概念，进一步说明微积分、线性代数、统计学、数值计算及多元统计分析在数据科学领域的重要作用。

1.1 大数据与数学

最早提出大数据概念的是全球知名咨询公司麦肯锡。该公司称："数据已经渗透到当今每一个行业和业务职能领域，成了重要的生产因素。"人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。实则，大数据在物理学、生物学、环境生态学等学科领域，以及军事、金融、通信等行业已有些时日，只是由于近年来互联网和信息行业采用了大数据技术，使得这一名词的曝光度提高，进而变得火热起来。本节通过介绍大数据的定义与数学各分支在大数据中的作用，阐述大数据与数学的关系。

1.1.1 大数据的定义

多数人认为"大数据"是一个新兴词汇，实则不然，早在 1980 年，著名的未来学家阿尔文·托夫勒便在《第三次浪潮》一书中将大数据赞颂为"第三次浪潮的华彩乐章"。大数据一词大约是从 2009 年开始被引入公众视线的。

1. 大数据的特征

虽然"大数据"这一个词汇已经诞生了近 40 年，但是目前为止并没有一个明确的定义。维克托·迈尔·舍恩伯格在《大数据时代》一书中提到了大数据应该具备以下 3 种特征。

（1）不是随机样本，而是全体数据。过去，因为记录、存储和分析数据的工具不够好，为了让分析变得简单，人们只能收集或者抽取尽量少的数据进行分析。如今，技术条件已经有了非常大的提高，虽然人类可以处理的数据依然是有限的，也永远是有限的，但是处理的数据量已经大大增加，而且未来会越来越多。在条件允许的情况下，使用全体数据往往能够得到一个更加准确、更接近真实的结果。

（2）不是精确性，而是混杂性。执迷于精确性是信息缺乏时代和模拟时代的产物。大约只有 5%的数据是结构化且能适用于传统数据库的，如果不接受混乱，剩下 95%的非结

构化数据就无法被利用。所以只有接受不精确性，才能从数据中获取更大的价值。需要特别注意的是，不精确性并非大数据固有的，它只是用来测量、记录和交流数据的一个缺陷。因为拥有更大的数据量所能带来的商业利益远远超过增加一点的精确性，所以通常不会通过大量增加成本提升数据的精确性。

（3）不是因果关系，而是相关关系。因果关系强调原因和结果必须同时具有必然的联系，即二者的关系属于引起和被引起的关系。而相关关系的核心是量化两个数据值之间的数理关系，相关关系强是指当一个数据值增加时，另一个数据值很有可能也会随之增加。

2. 大数据的定义

现阶段，大数据领域比较通用的大数据定义基于图 1-1 所示的 5V，其中每个 V 的具体定义如下。

（1）Volume：数据量大，即采集、存储和计算的数据量都非常大。真正大数据的起始计量单位往往是 TB（1 024GB）、PB（1 024TB）。

（2）Velocity：数据增长速度快，处理速度也快，时效性要求高。比如，搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法尽可能要求实时完成推荐。这是大数据区别于传统数据挖掘的显著特征。

（3）Variety：种类和来源多样化。种类上包括结构化、半结构化和非结构化数据，具体表现为网络日志、音频、视频、图片、地理位置信息等，数据的多类型对数据处理能力提出了更高的要求。数据可以由传感器等自动收集，也可以由人类手工记录。

（4）Value：数据价值密度相对较低。随着互联网及物联网的广泛应用，信息感知无处不在，信息量大，但价值密度较低。如何结合业务逻辑并通过强大的机器算法来挖掘数据的价值，是大数据时代最需要解决的问题。

（5）Veracity：数据的准确性和可信赖度高，即数据的质量高。数据本身如果是虚假的，那么它就失去了存在的意义，因为任何通过虚假数据得出的结论都可能是错误的，甚至是相反的。

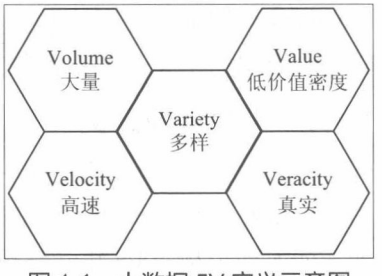


图 1-1　大数据 5V 定义示意图

1.1.2　数学在大数据领域的作用

信息化时代，大数据在各领域发挥着越来越重要的作用。人们使用大数据技术从海量数据中挖掘信息，发现规律，探索潜在价值。在大数据的研究和应用中，数学是坚实的理论基础。在数据预处理、分析与建模、模型评价与优化等过程中，数学方法扮演着至关重要的角色。

1. 微积分

从 17 世纪开始，随着社会的进步和生产力的发展，以及航海、天文、矿山建设等许多课题要解决，数学也开始研究变化的量，进入了“变量数学”时代，微积分也由此诞生。通过微积分可以描述运动的事物，描述一种变化的过程。由于微积分是研究变化规律的方法，所以只要是与变化、运动有关的研究，都或多或少地与微积分存在联系，都需要运用微积分的基本思想和方法。可以说，微积分的创立极大地推动了生活的进步。

微积分是整个近代数学的基础，有了微积分，才有了真正意义上的近代数学。统计学

中的概率论部分就是建立在微积分的基础之上的。比如，在函数关系的对应下，随机事件先是被简化为集合，继之被简化为实数，随着样本空间被简化为数集，概率相应地由奇函数约化为实函数。因此，微积分中有关函数的种种思想方法都可以畅通无阻地进入概率论领域。随机变量的数字特征、概率密度与分布函数的关系、连续型随机变量的计算等都是微积分现有成果的直接应用。

微积分的基础是极限论，在概率论中运用广泛，如分布函数的性质、大数定律、中心极限定理等。同时，在机器学习中，非常重要的各类最优化算法本质上就是在一定约束条件下求一个函数的最值，而这一概念和微积分基础中的极限论息息相关。

2. 线性代数

线性代数与大数据技术开发的关系很密切，线性代数领域的矩阵、秩、向量、正交矩阵、特征值与特征向量等概念在大数据分析、建模中发挥着巨大的作用。

在大数据中，许多应用场景的分析对象都可以抽象表示为矩阵。比如，大量 Web 页面及其关系、微博用户及其关系、文本数据中的文本与词汇的关系等都可以用矩阵表示。Web 页面及其关系用矩阵表示时，矩阵元素代表了页面 a 与页面 b 的关系。这种关系可以是指向关系，比如，1 表示 a 和 b 之间有超链接，0 表示 a 和 b 之间没有超链接。著名的 PageRank 算法就是基于这种矩阵进行页面重要性的量化，并证明其收敛性的。

以矩阵为基础的各种运算，如矩阵分解，是分析对象、特征提取的途径，因为矩阵代表了某种变换或映射，所以分解后得到的矩阵就代表了分析对象在新空间中的一些新特征。其中，特征分解（Eigen Decomposition）和奇异值分解（Singular Value Decomposition）等在大数据分析中应用十分广泛。

3. 统计学

统计学是一门基于数据的科学，是一种研究数据搜集、整理、分析与应用的方式和方法。数据是严谨的、枯燥的、冷冰冰的，同时，正确的数据又是丰富的、客观的、忠实的、从不会欺骗人的。

在当今的信息时代，数据是信息的载体，是统计学分析的对象。统计工作本身就是对数据进行搜集、整理、分析、解释这样一个系统的过程。离开了数据，统计学就失去了研究的意义和价值。同理，离开了统计学，数据就只是单纯的数据而已，几乎没有价值。通过统计的方法和原理整理及分析出来的数据，在精确度和适用度方面才会有较高的提升，才会实现数据的真正价值。

大数据的分析与挖掘等工作，从数据预处理开始，至建模得出结论，无不存在着统计学的身影。比如，统计分析所提供的诸如方差分析、假设检验、相关性分析等方法，都有助于数据分析前期的数据探索、数据预处理、特征工程等操作；朴素贝叶斯、Apriori 关联规则等算法本身的理论基础就来源于统计学。拥有扎实的统计基础，能够更加深入地理解算法，并解释结果。此外，在得出分析结果以后，研究者还需要通过统计分析来描述结果，以方便其他人理解。

4. 数值计算

数值计算是求解工程实际问题的重要方法之一，且随着工程问题规模的不断增大，相

比于理论研究和实验研究，其实用价值更大。在大数据时代的背景下，数据分析、数据挖掘、机器学习等算法中常见的插值、数值逼近、非线性方程求解等都属于数值计算的范畴。

从更高的层面看，数值计算指有效使用数字计算机求数学问题近似解的方法与过程，几乎涵盖了所有涉及复杂数学运算的计算机程序。数值计算主要研究如何利用计算机更好地解决各种数学问题，包括连续系统离散化和离散型方程的求解，并考虑误差、收敛性和稳定性等问题。

5. 多元统计分析

多元统计分析简称多元分析，是从经典统计学中发展起来的一个分支，是数理统计学中的一个重要的分支学科，是一种综合分析方法。20 世纪 30 年代，R.A.费希尔、H.霍特林、许宝碌及 S.N.罗伊等人做了一系列奠基性的工作，使多元分析在理论上得到迅速发展。20 世纪 50 年代中期，随着电子计算机的发展和普及，多元分析在地质、气象、生物、医学、图像处理及经济分析等领域得到了广泛的应用，同时也促进了理论的发展。

多元分析在大数据分析中有非常广泛的应用，能够在多个对象和多个指标互相关联的情况下分析出它们的统计规律。多元分析的主要方法包括回归分析、判别分析、聚类分析、主成分分析（Principal Component Analysis，PCA）、因子分析及典型相关分析等。这些分析方法在大数据领域都有着非常广泛的应用，其中，回归分析中的一元或多元线性回归可用于预测连续型数据，如股票价格预测和违约损失率预测等；判别分析与回归分析中的逻辑回归可用于预测类别型数据，这些数据通常都是二元数据，如欺诈与否、流失与否、信用好坏等；聚类分析是在不知道类标签的情况下将数据划分成有意义或有用的类，如客户细分等；主成分分析与因子分析都是用少数的几个变量（因子）来综合反映原始变量（因子）的主要信息，在大数据分析中常被用于对数据进行降维；利用典型相关分析方法可以快捷、高效地发现事物间的内在联系，如某种传染病与自然环境或社会环境的相关性等。

1.2 数学与 R 语言

R 语言是由新西兰奥克兰大学的 Ross Ihaka 与 Robert Gentleman 一起开发的一种面向对象的编程语言，是免费开源、能够有效用于统计计算和绘图的语言和环境。它是一套完整的数据处理、计算和制图软件系统，是一套开源的数据分析解决方案，由一个庞大且活跃的全球性研究型社区维护。其具有以下几点优势。

（1）可运行于多种平台之上，包括 Windows、UNIX、Mac OS X 和 Linux 等。

（2）在保证语法简单的同时，兼顾了程序设计语言的逻辑与自然的语言风格。

（3）拥有数目众多的程序包，能够轻松满足数据分析、数据挖掘、机器学习等领域的需求。

（4）可以通过程序包调用如 Python、Java、C、C++等语言，同时还提供了 Google、Twitter、微博等的 API 接口。

R 语言提供了各种数学计算、统计计算的函数，能够灵活地进行数据分析。常用于数学计算和统计计算的程序包（packages）有 base、stats，它们可以完成大部分数学计算工作。

此外，还可以使用 rootSolve、Ryacas、Deriv、prettyR、EnvStats、class、klaR、MASS 等程序包辅助完成数学计算工作。

1.2.1 base

base 程序包是 R 语言的基础包，其包含了 R 语言的基本功能，如算术、输入/输出、基本编程支持等。base 程序包中常用于数学计算的函数及说明如表 1-1 所示。

表 1-1 base 程序包中常用于数学计算的函数及说明

函 数 名	说 明
intersect	用于计算集合的并
union	用于计算集合的交
setdiff	用于计算集合的差
expression	用于表示函数的表达式
derive3	用于高阶求导
polyroot	用于求解实数多项式方程或复数多项式方程
matrix	用于创建矩阵
diag	用于创建单位矩阵，或提取矩阵的主对角线元素
lower.tri	用于提取矩阵的上三角矩阵
upper.tri	用于提取矩阵的下三角矩阵
t	用于矩阵的转置运算
det	用于求解行列式
solve	用于求矩阵的逆
eigen	用于求矩阵的特征值和特征向量
svd	用于对矩阵进行奇异值分解
max	用于求数据的最大值
min	用于求数据的最小值

1.2.2 stats

stats 程序包是 R 语言的统计包，具有统计计算和生成随机数的功能。stats 程序包中常用于统计计算的函数及说明如表 1-2 所示。

表 1-2 stats 程序包中常用于统计计算的函数及说明

函 数 名	说 明
D	用于求函数的导数或微分
integrate	用于求定积分
median	用于求中位数

续表

函 数 名	说 明
Quantile	用于求四分位数
mean	用于求均值
IQR	用于求四分位数的间距
var	用于求方差
sd	用于求标准差
dbinom	用于求二项分布的概率
dpois	用于求泊松分布的概率
dunif	用于求均匀分布的概率
dexp	用于求指数分布的概率
dnorm	用于求正态分布的概率
var	用于求协方差
cor	用于求相关系数

小结

本章作为全书的绪论部分，详细阐述了大数据的 3 个特性与 5V 理论，阐述了微积分、线性代数、统计学、数值计算与大数据之间的联系。同时，读者也需要注意，大数据相关的数学知识涵盖范围非常广，本书只是对其中最基础的部分做了介绍。

课后习题

选择题

（1）下列关于大数据特征的说法正确的是（　　）。

A．不是样本，而是全体数据

B．不是因果关系，而是相关关系

C．不是精确性，而是混杂性

D．不是符号计算，而是数值计算

（2）下列关于大数据 5V 概念的说法错误的是（　　）。

A．Velocity：数据增长速度快，处理速度也快，时效性要求高

B．Volume：数据量大

C．Value：数据价值密度相对较高

D．Variety：种类和来源多样化

（3）下列不属于多元统计分析的主要方法的是（　　）。

A．判别分析

B．因子分析

C．时间序列分析

D．典型相关分析

（4）下列关于微积分的说法正确的是（　　）。

A．统计学与微积分互相孤立，毫无关系

B．涉及运动事物的数学计算几乎都会涉及微积分

C．微积分可以解决所有数学问题

D．微积分是20世纪的伟大发明

（5）下列关于与数学相关的R语言程序包的说法错误的是（　　）。

A．能灵活地进行数据分析

B．base程序包可用于求二项分布、泊松分布等分布的概率

C．base程序包可用于计算集合的并、交、差

D．stats程序包用于求方差和相关系数