总　　序

2022 年 9 月，国务院学位委员会和教育部印发的《研究生教育学科专业目录（2022 年）》将原来的一级学科“图书情报与档案管理”更名为“信息资源管理”，使十余年前关于图书情报档案高等教育改革的集体共识在建制化学科认定上变成现实[①]。信息资源管理一级学科下设图书馆学、情报学、档案学、数据管理与数据科学、信息分析、数字人文、公共文化管理、出版管理、古籍保护与文献学、健康信息学、保密管理共 11 个二级学科，除图书情报档案学科群之外，还拓展到数据管理相关学科群和公共文化相关学科群。毫无疑问，这是中国图书情报档案学科及其教育在新时代进行的一次方向性变革和结构性调整，不仅意味着学科范畴、概念体系、理论形态、知识体系的重塑，也意味着职业群体、行业业态、实践模式的变革，以及在更广泛的能指领域与相关学科的交叉融合和核心能力竞争。

信息资源管理是一个国际化的概念，20 世纪 70 年代末兴起于美国，发轫于政府信息资源管理领域，成势于企业信息资源管理领域，欧洲除政府部门和企业界之外，图书情报界也是推动信息资源管理学科发展的主导者之一。信息资源管理在中国最先由图书情报界引入，历经三十余年的发展已经成为一个中国化的学科术语和概念，蕴含了图书情报档案特色，同时又超出了图书情报档案领域，并随着信息和数字信息技术发展不断拓展深化。采用“信息资源管理”作为一级学科名称体现了建设中国特色学科体系、学术体系和话语体系的重要举措[②]。

站在新的时间节点，信息资源管理学科建设需要学术共同体以更加开放的学术姿态在守正中强化创新，以创新性的知识生产和知识体系建构予以推进，其要者有以下三端。其一，追求信息资源管理学科内外部发展逻辑的统一，实现两者的同构共生。一般来说，从学科发展内在逻辑来看，学科是知识的客观发展和人类需求的主动建构相结合的产物，既有客观性，也有人为建构性；从学科发展的外在逻辑来看，学科发展中既有知识增长和深化的本体价值，也有推动社会发展的功用价值。信息资源管理一级学科的设立有可能在原有的图书情报与档案管理基础上围绕“四个面向”[③]发展出新的学科领域和新的学科方向，秉持以开放性为核心特征的新的知识生产模式，通过组织和参与者的多样性、

① 数字时代中国图书情报与档案学类教育发展方向及行动纲要[J]. 图书情报知识，2007，（1）：112-113.

② 马费成. 凝聚共识，推动信息资源管理一级学科建设[J]. 信息资源管理学报，2023，13（1）：4-8.

③ 四个面向是指面向世界科技前沿、面向经济主战场、面向国家重大需求、面向人民生命健康。

跨学科性、知识生产的应用情境性等，为信息资源管理学科内外部逻辑的协调统一带来新的可能性。其二，围绕信息、资源、管理三个核心概念范畴及其相互关联拓展深化学科内涵，发展信息资源管理理论、方法和应用。①信息概念范畴向信息链两端拓展，向前拓展到数据这一更加具有物理性质的层面，向后拓展到知识和智慧这一更加具有精神性质的层面，在“云大物移智”等新一代数字信息技术的加持下，这些领域的知识正在呈现显著的增长趋势。②资源概念范畴揭示了与侧重公益性质的图书情报与档案管理不同的学科认知和价值取向，需要建立数据信息的资源、资产、资本、生产要素等概念谱系，以及围绕所有权、使用权、收益权信息管理实践的重要问题解答。③管理概念范畴的集成化发展和向治理拓展，集成化发展包括对信息资源管理的管理对象、管理手段、管理流程、管理职能、管理模式等的集成；向治理拓展既包括信息资源管理本身的价值重塑和多元主体协同等治理方式的引入，也包括对循证治理、证据支持的决策等治理实践的关注。其三，在知识生产方法论上可借鉴拉卡托斯提出的“科学研究纲领”，在理论硬核、实践假设及方法与案例启示三个层面努力。理论硬核需要围绕信息、资源、管理三个核心概念范畴及其关联回答基础性、本质性的理论问题；实践假设形成辅助性假说组成的保护带，通过类型化的经验介入实现向理论硬核的转变；方法与案例启示总结中国实践场域的鲜活经验及其启发力，同时与世界范围的最佳实践形成有效对话和互鉴，促进中国信息资源管理自主知识体系和话语体系的建构。

从 2005 年 11 月起，根据高等学校信息资源管理类学科发展与专业教育改革的需要和图书市场的需求，为了建立结构合理、系统科学的学科体系和专业课程体系，创建符合信息资源管理的学科发展和教学改革要求的著作体系，进一步推动本学科领域的教学和科研工作的全面、健康和可持续发展，武汉大学二级教授和珞珈杰出学者邱均平老师率先发起、全程策划，联合兰州大学、中山大学、吉林大学、华中师范大学等 10 多所重点高校的专家、学者共同编著 17 本分册组成的“现代信息资源管理丛书”（以下简称“丛书”），由科学出版社正式出版。

“丛书”的显著特点是：①定位高，创新性强；②范围广，综合性强；③水平高，学术性强；④发行量大，学术影响广泛。“丛书”大多数分册都多次重印，最多的先后印刷 8 次，被不少高校采用作为教材；同时，按照理论、方法、应用三结合的思路构建各著作的内容体系，体现内容上的前瞻性、科学性、系统性和实用性。特别值得一提的是，邱均平、沙勇忠编著出版的《信息资源管理学》分册是国内外正式出版的第一部带“学”字的“信息资源管理学”专著，开创了从专业领域向一级学科迈进的新阶段，具有学科奠基性作用和开创性的重要意义。

“丛书”的第二次整体改版是回应一级学科更名后知识发展与学科建设的需要，是在 2008 年至 2011 年出版“丛书”的基础上学界同仁所做的共同探索和努力。“丛书”的第二次整体改版的目的有四个：一是开展有组织的科学研究，力争产出一批具有奠基性、创新性和标志性的成果；二是争取构建一个比较全面、完整、系统的信息资源管理学科体系，为学科建设和发展奠定基础，充实内涵；三是锻炼和造就人才队伍，推出新秀，培养一大批信息资源管理专家、学者乃至“信息资源管理学家”，落实教育、科技、人才一体化发展战略，为把信息资源管理学科及相关行业做大做强提供人才保障；四是

分册以“著”或“编著”形式出版，可作教材使用，满足当前本学科教学和科研工作的需要，为提高人才培养质量提供支撑条件，为构建信息资源管理的自主知识体系做出应有的贡献。

“丛书”的第二次整体改版由邱均平、沙勇忠担任总主编，包括 15 本分册，其与邱均平、沙勇忠同时担任总主编的中国社会科学出版社出版的“数智时代信息资源管理丛书”的 12 本分册互补形成相对完整的信息资源管理著作体系。“丛书”的第二次整体改版后的 15 本分册的名称和作者分别是：《信息资源管理学（第二版）》（邱均平、沙勇忠、丁敬达、王琳）；《政府信息资源管理（第二版）》（王新才、谭必勇、吕元智等）；《数字资源建设与管理（第二版）》（毕达天、薛春香、毕强等）；《信息检索原理与技术（第二版）》（夏立新、叶光辉、翟姗姗）；《网络信息资源开发与利用（第二版）》（张洋、侯剑华、余厚强）；《信息分析（第三版）》（沙勇忠、牛春华）；《电子商务信息管理（第二版）》（王伟军、池毛毛、刘蕤等）；《竞争情报学（第二版）》（董克、赵蓉英）；《信息资源管理政策与法规（第二版）》（马海群、周丽霞、汪传雷）；《信息咨询与智能决策（第二版）》（文庭孝、张蕊、罗贤春）；《信息资源管理技术与信息系统》（刘永、窦永香、付永华）；《网络信息安全与保密管理》（吴国华、唐晓波、张蕊）；《国外信息资源管理》（舒非、余波）；《知识获取与用户研究》（王曰芬、邓胜利、岑咏华）；《数据科学与大数据分析》（朝乐门、步一、余云龙）。各分册除阐述各自领域相对成熟的知识积累和知识体系之外，还力图反映国内外的前沿理论和技术方法；既有编著者的独到见解和新的研究成果，又突出面向实践场域的应用，兼具专著与教材的双重风格。

“丛书”的出版得到了科学出版社的大力支持，同时还得到了各分册负责人、各位著者和参编院校的鼎力帮助。在编写过程中我们还参阅和引用了一些国内外文献，在此一并对相关作者表示衷心感谢!

信息资源管理学科发展归根到底有赖于学术共同体面向国家科技创新战略、文化强国战略、数字中国战略、国家安全战略等重大战略需求的知识生产和自主知识体系的构建，这既是历史性任务和时代性课题，也是推进中国式现代化过程中学术共同体的崇高使命。我们所做的工作是一种集智广思的探索和努力，不足之处在所难免，诚望同行专家及读者批评指正。

杭州电子科技大学资深教授、院长、博士生导师
邱均平
兰州大学副校长、二级教授、博士生导师
沙勇忠

2025 年 7 月 1 日

前　言

数据科学与大数据分析已成为现代信息资源管理领域中不可或缺的组成部分。在推动数字中国建设中，数据不仅是重要的生产要素，更是驱动创新与决策的重要力量。本书正是在这样的背景下应运而生，旨在为读者提供系统的理论知识和实用的分析方法。

本书是“现代信息资源管理丛书”中的一部，涵盖了数据科学与大数据分析的核心内容。全书分为七章，由中国人民大学朝乐门（第 1～4 章）、北京大学步一（第 5～6 章）和杭州电子科技大学余云龙（第 7 章）撰写，其中：第 1 章绪论主要讨论了数据科学与大数据分析的基本概念，包括数据、模型、问题、数据科学及大数据分析，旨在为读者奠定坚实的基础。第 2 章数据科学理论探讨了数据科学的理论框架，涵盖研究目的、知识体系、主要原则、基本流程和人才培养，为读者提供了系统的理论指导。第 3 章数据科学方法分析了数据科学的方法论，包括统计分析方法、机器学习方法、数据可视化方法和大模型方法，帮助读者掌握各类数据分析技术。第 4 章数据科学技术涵盖了大数据管理技术、大数据计算技术、大数据存储技术和开源工具，系统讲解了数据处理的关键技术。第 5 章大数据分析理论主要讨论了大数据分析的理论基础，介绍了大数据分析的主要特点、基本流程、常用算法及可视化，帮助读者理解大数据分析的核心思想。第 6 章大数据分析工具及平台介绍了各种大数据分析工具及平台，包括数据采集工具、开源数据工具、数据可视化工具、情感分析工具、开源数据库及大数据分析平台，为读者提供了丰富的工具选择。第 7 章综合应用案例分析通过多个实际案例，展示了数据科学与大数据分析在学科发展态势挖掘、主题挖掘和引文分析等方面的应用，帮助读者将理论知识应用于实践，提升数据分析能力。

与国内外同类教材相比，本书具有以下显著特点。①目标读者定位：本书主要面向信息资源管理及相关学科领域的读者。②内容结构与章节覆盖：本书在国内率先将信息资源管理与数据科学及大数据分析结合，不仅包含数据科学，还增加了大数据分析的内容，具有更高的实用性和操作性，适合培养大数据应用型和复合型人才。③知识点选择：本书聚焦数据科学和大数据技术的应用，侧重为信息资源管理学科类专业讲解数据科学的核心理论、方法、技术与应用，提供方法、技术和工具层次的支持。④内容新颖性：本书集中体现了数据科学和大数据分析领域的最新理论成果和实践案例。

本书不仅适用于高等院校数据科学与大数据分析相关专业的师生，也适合从事数据分析、数据挖掘、数据管理等领域的专业人士阅读。希望通过本书的学习，读者能够掌握数据科学与大数据分析的核心知识与技能，并在实际工作中有效应用，推动相关领域的发展与进步。

在本书的撰写过程中，我们参阅了大量中外文献资料，并在很大程度上借鉴了这些文献。在此，我们向本书参考文献的作者表示衷心的感谢。然而，由于篇幅所限，未能完整列出所有参考文献，我们向未能一一列出的参考文献作者表示诚挚的歉意。

我们深知自己的水平有限，加之撰写时间相对紧迫，书中难免存在一些疏漏。因此，我们诚恳地请教专家和广大读者，期待您的宝贵意见和批评指正，以便我们在未来的工作中不断完善和提高。

最后，特别感谢本系列丛书的主编邱均平教授和沙勇忠教授的指导与支持，感谢科学出版社对本书的鼎力支持。中国人民大学硕士研究生李美静同学参与了本书的排版和校对工作，在此表示感谢。同时，衷心感谢中国人民大学、北京大学和杭州电子科技大学的支持，也感谢各位读者对本书的支持与关注。我们相信，在数据驱动的新时代，本书将成为读者学习和应用数据科学与大数据分析的重要参考。

目　录

第1章 绪 论

数据科学以数据、模型和问题为三大基本要素。首先，数据构成了数据科学的研究对象，包括数字、文本、图像及声音等多种形式，其目的在于从这些数据中抽取有价值的信息与知识。其次，模型是数据科学中分析、解释和利用数据的主要工具。通过构建模型，数据科学家能够进行数据分析、预测和分类，进而揭示数据的内在结构和规律，实现数据价值化的目的。最后，问题是数据科学的出发点，解决问题是数据科学的主要目的。深入理解问题及其背景和需求，是有效实施数据科学项目的关键。数据、模型和问题三者相互联系、相互依赖，共同构建了数据科学的核心框架，使数据科学能够综合应用统计学、计算机科学与特定领域的知识，有效地解决问题。

1.1 数 据

数据科学反映了人们对数据的认识的三种变化：从“数”到“据”的转变、从“数据”到“大数据”的转变以及从“结构化数据”到“非结构化数据”的转变。

1.1.1 从“数”到“据”的转变

从“数”到“据”的转变是数据科学中对数据认识的一个突破。数据科学中的“数据”不仅仅是“数”，而是更广泛地指记录、证据和凭证等“据”。“数值”仅仅是“数据”的一种存在形式。除了“数值”，数据科学中所说的“数据”还包括文字、图形、图像、动画、文本、语音、视频、多媒体和富媒体等多种类型，如图 1-1 所示。

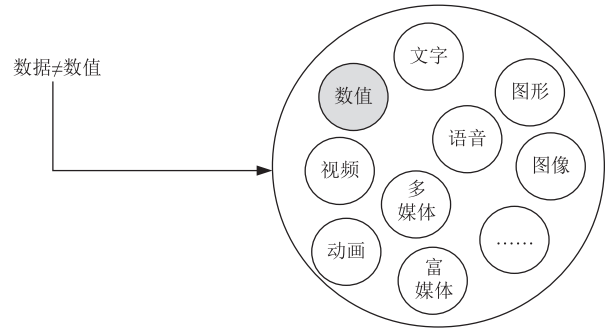

图 1-1 从“数”到“据”的转变

（1）数据科学中的“计算”不仅仅是加减乘除等基本的数学计算，还包括数据的查询、挖掘、发现、分析和可视化等多种处理方式。

（2）数据科学关注的并非单一学科的问题，而是涉及多个学科领域（如统计学、计算机科学等），强调跨学科的研究视角。

1.1.2 从“数据”到“大数据”转变

从“数据”到“大数据”的转变是数据科学中对数据认识的另一个突破。大数据是指对现有理论、方法、技术和工具构成挑战的数据。人们通常使用以字母 V 开头的特征（如 volume、variety、velocity 等）来解释大数据带来的不同挑战。大数据的三种典型定义方法均体现了这些特征。

（1）Gartner（高德纳咨询公司）的定义方法：大数据是指无法使用传统流程或工具处理或分析的信息，是需要新处理模式才能具有更强的决策力、洞察力和流程优化能力的海量、高增长率和多样化的信息资产。

（2）GB/T 35295—2017 的定义方法：大数据是指具有体量巨大、来源多样、生成极快且多变等特征并且难以用传统数据体系结构有效处理的包含大量数据集的数据。

（3）IBM 的定义方法：大数据是拥有以下四个特点（又称为“4V”）中任意一个的数据源：极大的数据量级（volume）；以极快的速度（velocity）移动数据；极广泛的数据源类型（variety）；极高的准确性（veracity），确保数据源的真实性。

人们对大数据的 V 特征的认识经历了不断拓展和演化的过程。在 2001 年，Gartner 首次提出了大数据的 3V 特征——volume（数据体量）、velocity（数据速度）和 variety（数据多样性）。随后，2012 年，IBM 提出了大数据的第四个 V——veracity（数据真实性）。同年，Gartner 又提出了大数据的第五个 V——value（数据价值）。随着时间的推移，数据科学的 V 特征不断增多，反映了人们对大数据现象的认识不断深入和拓展的过程。

（1）数据体量大：相对于现有的计算和存储能力而言，当数据量达到 PB 级以上时，一般称之为“大数据”。然而，我们应该注意到，大数据的时间分布往往不均匀，近几年生成数据的占比最高。

（2）数据类型多：大数据涉及多种数据类型，包括结构化数据、非结构化数据和半结构化数据。有统计显示，在未来，非结构化数据的占比将达到 90%以上。非结构化数据包括的数据类型很多，如网络日志、音频、视频、图片、地理位置信息等。数据类型的多样性往往导致数据的异构性，进而增加数据处理的复杂性，对数据处理能力提出了更高要求。

（3）数据速度快：大数据中的“速度”通常指的是数据的产生速度和处理速度。一方面，大数据的产生速度快速增加。根据统计数据，2009 年至 2020 年数字宇宙的年均增长率达到 41%。另一方面，我们对大数据处理的速度（计算速度）也提出了更高的要求，特别是对于实时数据分析的需求越来越迫切，这成为一个热门话题。

（4）数据价值密度低：在大数据中，价值密度的高低与数据总量的大小之间并不存在线性关系。有价值的数据往往被淹没在海量无用数据之中，也就是人们常说的“我们

淹没在数据的海洋，却又在忍受着知识的饥渴”（We are drowning in a sea of data and thirsting for knowledge）。例如，一部长达 120 分钟的连续不间断的监控视频中，有用数据可能仅占几秒。因此，“如何从海量数据中洞见出有价值的数据”是数据科学的重要课题之一。

（5）数据真实性：大数据中的“真实性”指数据的准确性和可信度。随着数据量的增加和数据来源的多样化，确保数据的真实性变得越来越重要。数据可能受到错误、噪声、欺骗或不完整性的影响，这会影响到数据的价值和可信度。因此，处理大数据时需要关注数据的真实性，采取适当的数据质量控制措施，确保数据分析和决策的准确性和可靠性。

1.1.3 从“结构化数据”到“非结构化数据”的转变

从结构化程度看，通常将数据分为：结构化数据、非结构化数据和半结构化数据，如表 1-1 所示。在数据科学中，数据的结构化程度对于数据处理方法的选择具有重要影响。例如，结构化数据的管理可以采用传统关系数据库技术，而非结构化数据的管理往往采用 NoSQL、NewSQL 或关系云技术。

表 1-1 结构化数据、非结构化数据和半结构化数据的对比

类型	含义	特征	举例
结构化数据	直接可以用传统关系数据库存储和管理的数据	先有结构，后有数据	关系型数据库中的数据
非结构化数据	无法用关系数据库存储和管理的数据	没有（或难以发现）统一结构的数据	语音、图像文件等
半结构化数据	经过一定转换处理后可以用传统关系数据库存储和管理的数据	先有数据，后有结构（或较容易发现其结构）	HTML、XML 文件等

注：HTML，hyper text markup language，超文本标记语言；XML，extensible markup language，可扩展标记语言

（1）结构化数据：以“先有结构，后有数据”的方式生成的数据。通常，人们所说的“结构化数据”主要指的是在传统关系数据库中捕获、存储、计算和管理的数据。在关系数据库中，需要先定义数据结构（如表结构、字段的定义、完整性约束条件等），然后严格按照预定义的结构进行捕获、存储、计算和管理数据。当数据与数据结构不一致时，需要按照数据结构对数据进行转换处理。

（2）非结构化数据：不存在或难以发现统一结构的数据，即在未定义结构的情况下或并不按照预定义的结构要求捕获、存储、计算和管理的数据。通常，主要指无法在传统关系数据库中直接存储、管理和处理的数据，包括所有格式的办公文档、文本、图片、图像、音频或视频信息。

（3）半结构化数据：介于完全结构化数据（如关系型数据库、面向对象数据库中的数据）和完全无结构的数据（如语音、图像文件等）之间的数据。例如，HTML、XML 文件等，其数据的结构与内容耦合度高，进行转换处理后可发现其结构。

目前，非结构化数据占比最大，绝大部分数据或数据中的绝大部分属于非结构化数

据。因此，非结构化数据是数据科学中的重要研究对象之一，也是数据科学与传统数据管理的主要区别之一。

1.2 模　　型

在数据科学中，统计学和机器学习的模型扮演着至关重要的角色，它们用于分析和解释数据，以及进行预测和提供决策支持。模型可以帮助数据科学家从历史数据中抽取知识、识别模式和评估现象之间的关系，并用这些信息来预测未来事件或决策的后果。

1.2.1 数据科学中的模型、算法、参数与超参数

在数据科学中，算法（algorithm）是一组系统的步骤，用于训练模型（model）。模型通常由一组参数（parameter）定义，这些参数是通过算法训练得到的。训练过程有时涉及选择最优的超参数（hyperparameter），这些超参数会指导学习算法确定最适合的参数集合，以便准确地将输入特征（input feature）映射到输出标签或目标（output label or target）上。这一过程旨在调整模型的预测性能，具体示例如表 1-2 所示。

表 1-2　算法、模型、参数和超参数的区别与联系（以简单线性回归为例）

概念	功能	举例
算法	用于训练模型的方法	简单线性回归算法、KNN 算法、K-means 算法、SVM、逻辑回归算法
模型	算法所训练出的结果	$y = \beta_1 x + \beta_0$
参数	用于描述模型参数，其取值可以由训练集训练得出	β_1 和 β_0
超参数	控制机器学习过程并确定学习算法最终学习的模型参数值的参数	训练集和测试集的划分比例

注：KNN，K-nearest neighbor，K 最近邻；K-means，K 均值；SVM，support vector machine，支持向量机

（1）算法：用于训练模型的方法，可分为监督学习、无监督学习和半监督学习等类型，例如本书后续章节介绍的简单线性回归算法、KNN 算法、K-means 算法、SVM、逻辑回归算法等。

（2）模型：机器学习算法采用已知数据集训练出的结果，即模型是算法的输出，由训练数据和算法共同决定一个模型。由于同一个算法在不同训练集上训练出的模型可以不同，算法和模型之间并非为一一对应关系。例如，简单线性回归算法在不同训练集（如父子身高数据集、某企业广告投入与销售额数据集等）上训练出多个不同的模型，这些模型的区别在于参数取值不同。例如，基于父子身高数据集训练出的模型为 y=–0.04x+77.69，基于某企业广告投入与销售额数据集训练出的模型为 y=195.34x+23.5。

（3）参数：参数可以分为算法参数和模型参数。其中，算法参数又称为“超参数”，而模型参数简称为“参数”。算法参数和模型参数的区别在于前者由数据分析师手动指定（调参），而后者可以通过机器学习自动训练，如图 1-2 所示。因此，模型参数用于描述

一个具体的模型。通常，同一个算法所训练出的模型的参数个数和类型是一致的，区别在于参数取值。例如，简单回归分析中的斜率和截距项。

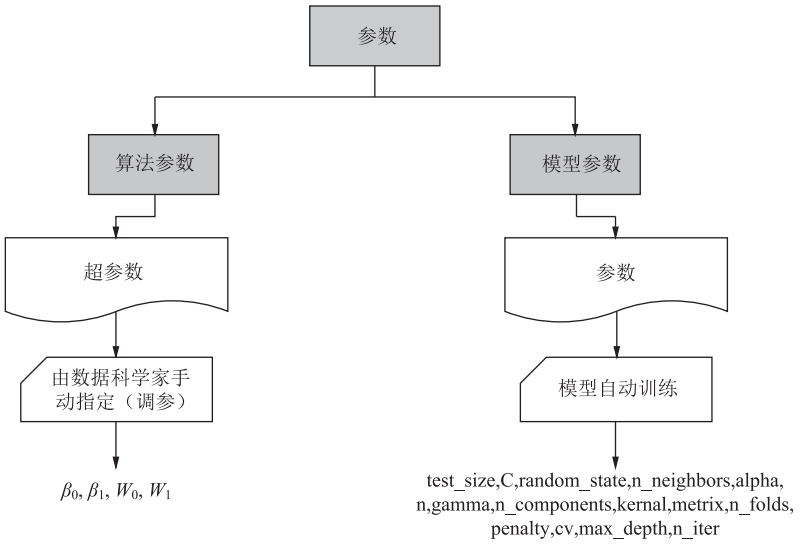

图 1-2 Scikit-learn 中常见的参数及超参数

（4）超参数：控制机器学习过程并确定学习算法最终学习的模型参数值的参数，如训练集和测试集的分割比例、优化算法中的学习率、聚类算法中的聚类数、多数算法中的损失函数的选择、神经网络学习中的激活函数的选择、神经网络中的隐含层数及迭代次数等。图 1-2 给出了 Python 机器学习第三方包 Scikit-learn 中常用的超参数和参数。

1.2.2 数据科学中的模型构建

数据科学中的模型构建至少遵循三个基本原则：奥卡姆剃刀（Occam’s razor）原则、没有免费的午餐定理（no free lunch theorem，NFL 定理）和偏差-方差权衡（bias-variance trade-off）。

1. 奥卡姆剃刀原则

奥卡姆剃刀原则的含义为“如无必要，勿增实体”（Entities should not be multiplied without necessity）。在数据科学的模型构建中，奥卡姆剃刀原则强调应当尽量避免不必要的复杂性。这一原则在数据科学模型构建中的重要性体现在以下几个方面。

（1）强调模型简洁有效性：建模时，若有多个模型可以解释同样的数据，应优先选择较简单的模型。简单模型通常更易于理解和维护，且降低训练数据的过拟合风险。

（2）强调模型的泛化能力：简单模型往往具有更好的泛化能力，意味着它们在未见过的数据上表现更佳。这是因为简单模型较少依赖训练数据的特定偶然特性，更能捕捉数据的基本规律。

（3）重视可解释性和可信赖性：简单模型通常更易于解释。在商业和科学研究中，

模型的可解释性极为重要，有助于理解模型的决策过程，并确保其决策可接受，进而建立对模型的信任。

（4）重视计算效率：较简单的模型通常需要较少的计算资源来训练和运行，这在资源有限或需要快速响应的场景中尤其重要。

在数据科学实践中，奥卡姆剃刀原则提醒研究人员和实践者寻找解释数据最简捷有效的方法，鼓励去除不必要的复杂性，使模型更加健壮和高效。

2. NFL 定理

在数据科学中的模型构建过程中，NFL 定理指出没有一种单一的机器学习算法能够在所有问题上都比其他算法更出色。因此，通常需要尝试多种模型，并根据具体问题找到最适合的模型。图 1-3 对比了通用模型和专用模型在不同问题类型上的性能表现。

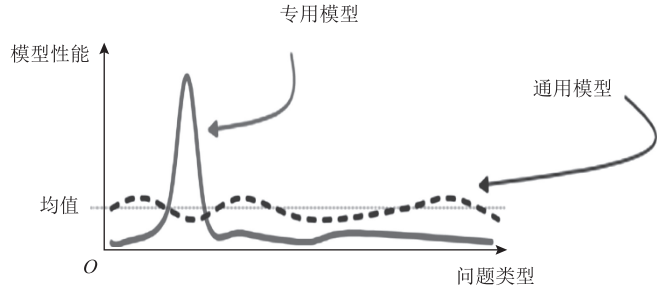

图 1-3　通用模型与专用模型的差异

（1）通用模型是指适用于多种问题类型的模型，如线性回归、决策树等。这些模型通常具有较强的泛化能力，能够在不同类型的数据和问题上表现良好。根据 NFL 定理，通用模型并没有超越其他专用模型的优势。NFL 定理指出，在所有问题上，没有一个算法能够比其他算法更好，这包括通用模型在内。

（2）专用模型是针对特定问题设计和优化的模型，通常使用领域专家的知识和先验信息。这些模型可能在特定问题上表现非常优秀，但在其他问题上可能并不适用或表现较差。根据 NFL 定理，没有一种模型能够在所有问题上都表现良好，专用模型在特定问题类型上可能有较好的性能，但不具备通用性。

3. 偏差-方差权衡

在数据科学的模型构建中，偏差-方差权衡是指在选择和调整模型时需要平衡偏差和方差之间的关系。

（1）偏差：模型对于真实关系的错误拟合程度。具有高偏差的模型倾向于对数据进行过度简化，可能无法捕捉数据中的复杂关系，导致欠拟合问题。

（2）方差：模型在不同数据集上预测结果的不稳定性。具有高方差的模型对训练数据的小变化敏感，可能导致过拟合问题，即模型在训练数据上表现良好，但在新数据上表现较差。

偏差-方差权衡意味着调整模型的复杂度以平衡偏差和方差之间的关系，从而获得更好的泛化能力。通常情况下，增加模型的复杂度可以降低偏差但增加方差，而降低模型的复杂度则可能降低方差但增加偏差，如图 1-4 所示。因此，需要根据具体问题和数据集的特点来选择适当的模型复杂度，以达到最佳的预测性能。

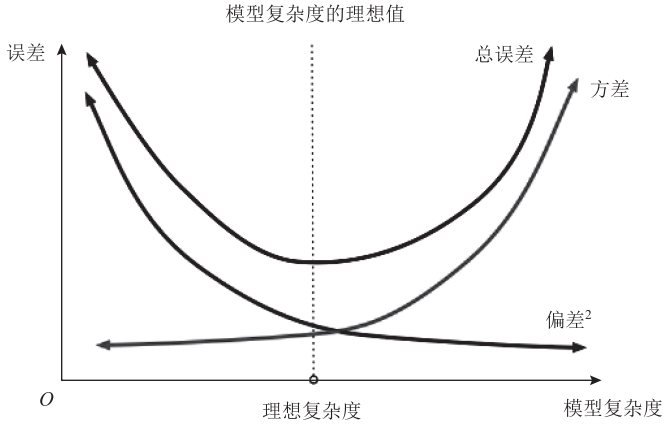

图 1-4 数据科学模型中的偏差-方差权衡

1.2.3 数据科学中的模型评估

机器学习在数据科学中的应用主要体现在通过调用机器学习算法训练出具体“模型”。在实际使用模型之前，需要对其进行评估。在模型评估中常用的方法有以下几种。

1. 学习曲线

学习曲线（learning curve）是用于可视化模型学习过程中性能变化的图表。它的横坐标代表训练样本量，纵坐标代表性能度量，如准确率或误差等。在学习曲线的示意图（图 1-5）中，可以看到随着训练样本量的增加，模型在训练集和交叉验证集上的得分逐

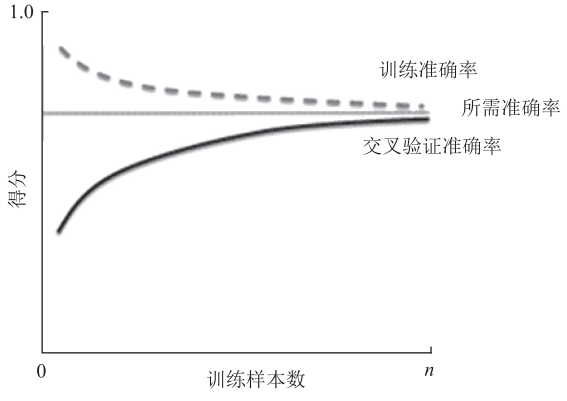

图 1-5 学习曲线的示意图

渐接近某个稳定值，这反映了模型的收敛行为和泛化能力。如果训练得分和交叉验证得分趋于一致，且都达到较高水平，表明模型具有良好的学习效果和泛化能力。通过学习曲线，我们能够判断模型是否适应了训练数据，并识别过拟合或欠拟合的问题。

（1）高偏差：在学习曲线中，当训练样本数量增加时，训练准确率和交叉验证准确率逐渐趋于收敛，但它们与所需准确率之间存在较大的偏差，如图 1-6 所示。高偏差通常表明模型在训练集和验证集上的表现都不佳，这种情况通常与“欠拟合”现象有关。欠拟合的主要原因有两个：首先，模型可能过于简单，无法捕捉数据的复杂性和底层模式；其次，所选特征可能不足以提供足够的信息，或与模型预测的目标关系不大。这些因素限制了模型学习数据的能力，导致预测性能不佳。

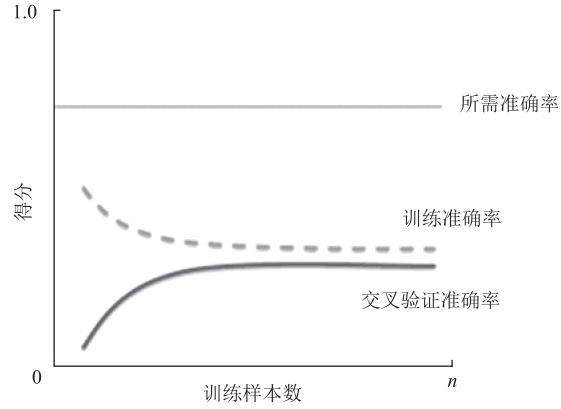

图 1-6　高偏差的学习曲线

（2）高方差：在学习曲线中，当训练样本数量增加时，训练得分可能接近理想值，但交叉验证得分却显著低于理想水平，如图 1-7 所示。高方差通常表明模型可能存在“过拟合”现象，即模型在训练集上表现优异，但在未见过的数据上表现不佳。过拟合的主要原因通常包括两个方面：首先，模型可能过于复杂，捕捉到训练数据中的噪声和异常值，而不仅仅是底层数据模式；其次，相对于可用的训练样本量来说，可能存在过多的特征属性，导致模型过度适应训练数据而无法泛化到新数据。

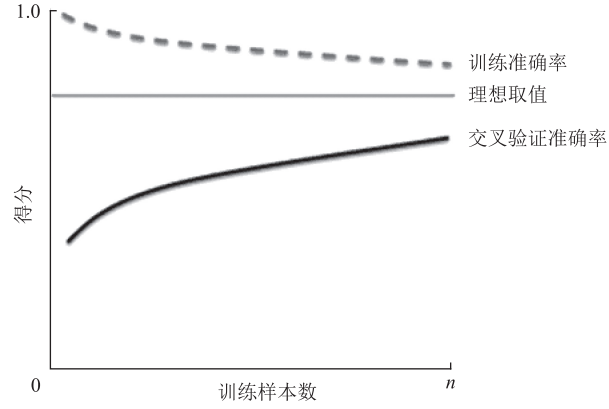

图 1-7　高方差的学习曲线

2. 混淆矩阵

混淆矩阵是常用于评估有监督学习算法性能的一种工具，可以快速计算精度和召回率等指标，是制作 ROC 曲线的基础，如图 1-8 所示。其中，用“正例”（positive）和“负例”（negative）来表示样本的“类别”；用“真”（true）和“假”（false）来表示“模型预测是否正确”。

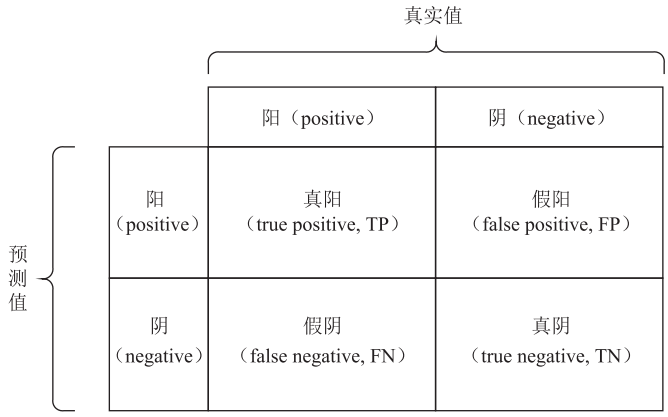

图 1-8 混淆矩阵的示意图

（1）TP（true positive）：模型“正确地（真/true）”预测了样本的类别，样本的预测类别为“正例”。

（2）FN（false negative）：模型“错误地（假/false）”预测了样本的类别，样本的预测类别为“负例”，即模型犯了类似于统计学上的第一类错误（type Ⅰ error）。

（3）FP（false positive）：模型“错误地（假/false）”预测了样本的类别，样本的预测类别为“正例”，即模型犯了类似于统计学上的第二类错误（type Ⅱ error）。

（4）TN（true negative）：模型“正确地（真/true）”预测了样本的类别，样本的预测类别为“负例”。

（5）模型的精度（precision）：在所有判别为正例的结果中，模型正确预测的样例所占的比例，即

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

（6）模型的召回率（recall）：在所有正例中，模型正确预测的样本所占的比例，即

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

除了模型的精度和召回率，基于混淆矩阵可以定义的模型评估指标有很多，包括正确率（accuracy）、错误率（misclassification/error rate）、特异性（specificity）、流行程度（prevalence）等，由于篇幅所限，在此不再逐一详解。

3. ROC 曲线与 AUC 面积

接受者操作特征（receiver operating characteristic，ROC）曲线是以“假正率（FP_rate）”和“真正率（TP_rate）”分别作为横坐标和纵坐标的曲线。通常，人们将 ROC 曲线与“假正率（FP_rate）”轴围成的面积称为“曲线之下的区域（area under curve，AUC）面积”。AUC 面积越大，说明模型的性能越好，如图 1-9 所示。在图 1-9 中，L2 曲线对应的性能优于曲线 L1 对应的性能，即曲线越靠近 A 点（左上方）性能越好，曲线越靠近 B 点（右下方）曲线性能越差。

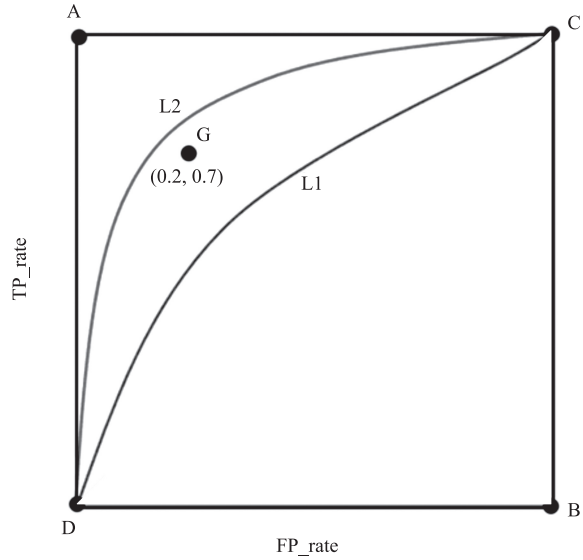

图 1-9　ROC 曲线与 AUC 面积

1.3　问　　题

解决问题是数据科学的主要目的之一。数据科学通过使用统计学、机器学习、数据分析和数据可视化等技术，将大量的数据转化为可行的洞察力和知识，以解决实际问题、支持决策过程、优化业务流程、提高效率和创造新的价值。无论是在商业、科研、政府政策、医疗保健还是其他领域，数据科学的核心目标都是利用数据来回答关键问题、解决实际问题，并帮助组织和个人做出基于数据的敏捷决策。数据科学中的问题可以从以下两个方面进行分类。

1.3.1　系统Ⅰ类问题和系统Ⅱ类问题

从问题的分析过程和解决方式看，数据科学中的问题可以分为系统Ⅰ类问题和系统Ⅱ类问题，如表 1-3 所示。