大数据应用与技术丛书

数据科学实战入门
使用 Python 和 R

[法] 尚塔尔 • D • 拉罗斯(Chantal D. Larose)
丹尼尔 • T • 拉罗斯(Daniel T. Larose) 著
王海涛　宋丽华　邢长友 译

清华大学出版社
北　京

北京市版权局著作权合同登记号　图字：01-2019-4943

Chantal D. Larose, Daniel T. Larose

Data Science Using Python and R

EISBN：978-1-119-52681-0

图书在版编目(CIP)数据

数据科学实战入门：使用Python和R/(法)尚塔尔 • D • 拉罗斯(Chantal D. Larose)，(法)丹尼尔 • T • 拉罗斯(Daniel T. Larose) 著；王海涛，宋丽华，邢长友 译. —北京：清华大学出版社，2020.6

(大数据应用与技术丛书)

书名原文：Data Science Using Python and R

ISBN 978-7-302-55379-3

I.①数… II.①尚… ②丹… ③王… ④宋… ⑤邢… III. ①软件工具—程序设计 ①程序语言—程序设计 IV.①TP311.561 ②TP312

中国版本图书馆 CIP 数据核字(2020)第 070586 号

责任编辑：王　军
装帧设计：孔祥峰
责任校对：牛艳敏
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 装 者：小森印刷霸州有限公司
经　　销：全国新华书店
开　　本：170mm×240mm　　**印　　张：**15.75　　**字　　数：**340 千字
版　　次：2020 年 7 月第 1 版　　**印　　次：**2020 年 7 月第 1 次印刷
定　　价：69.80 元

产品编号：083301-01

译 者 序

进入 21 世纪以来，信息技术的发展突飞猛进，人类从信息时代步入数字时代，又马不停蹄地进入了数据时代。自 2008 年大数据被业界正式提出后，围绕大数据的科学研究和产业应用如火如荼，快速实现了从名词炒作到应用落地，数据采集、数据处理、数据建模、数据分析和数据可视化等大数据相关技术在越来越多的行业中得到了广泛研究和普遍应用。与此同时，我国政府高度重视大数据的理论研究和产业应用，并大力支持高校开设大数据科学与技术相关学科专业，以应对持续井喷的大数据人才需求。译者作为这一伟大时代的普通见证者和实践者，深深感到大数据技术将在未来的数据时代和人工智能时代发挥举足轻重的作用。

大数据不仅是一项技术，又是一门理论和实践性都很强的学科，更是一种创新性思维和理念。大数据行业不仅需要低端的数据标注师，也需要中端的数据工程师和数据分析师，还需要处于大数据人才结构最顶层的数据科学家。遗憾的是，数据科学家的培养绝非一蹴而就，需要经过大量的系统学习和专业培训。最近几年，大数据相关的书籍层出不穷、琳琅满目，有针对管理人员的大数据思维类图书，有面向高校学生的大数据技术原理类图书，还有面向企业技术人员的大数据实践实训类图书。诚然，这些图书中不乏概念清晰、思路新颖、内容全面的好书，但市场上真正能够很好地将理论、技术、工具和实际应用紧密融合的大数据图书少之又少。

本书的两名作者是大数据业界的知名专家学者，也是一对令人敬仰和羡慕的父女搭档，他们对大数据理论技术和行业应用有着深刻独到的理解。针对大数据业界人才培养的痛点，本书在讲透大数据科学基本原理的同时，非常重视面向实际问题的实战教学，希望借助当前世界上最流行、最好用的两大开源数据科学工具——Python 和 R 语言，来解决可能遇到的各种数据科学问题，这无疑有助于提高有志于大数据研究和应用的广大读者在这个前沿领域的专业技能。正如本书作者所言，通过本书，读者将亲身体验使用业界最先进的技术来逐步寻求针对实际业务问题的解决方案。换句话说，读者将通过数据科学的亲手实践来认识、学习和研究数据科学。另一方面，本书作者通过生活中的实际案例，将复杂枯燥的数据问题转化为有趣易懂的实践操作，对读者的专业背景要求较低，因而有着较广泛的受众群体。此外，本书作者精心组织内容，并提供了翔实的学习

指导和大量配套习题，很适合作为高职高专或本科高校教材使用，教师可以针对不同层次和不同专业的学生合理选取教学内容。

本书译者王海涛、宋丽华和邢长友均来自南京高校，长期从事信息技术领域的教学和科研工作，在计算机网络、大数据和系统分析等领域有较深入的研究，并翻译出版过多本专业技术类图书，保证了本部译著的质量。在翻译图书的过程中，南京审计大学金审学院和陆军工程大学的各级领导为译者提供了许多帮助。此外，清华大学出版社的王军编辑也为本书的翻译出版付出了大量心血，在此一并表示感谢！

由于译者水平有限，难免存在翻译上的纰漏和理解上的偏差，希望广大读者批评指正。最后，真诚希望本书的出版能为我国的大数据人才培养和大数据行业发展提供一点帮助，为我国的科技进步尽绵薄之力。

译者　王海涛

作 者 简 介

本书的两位作者 Chantal D. Larose 博士和 Daniel T. Larose 博士是一对罕见的父女数据科学家。本书是他们合著的第三本书，在此之前他们还共同编写了如下两本图书：

- *Data Mining and Predictive Analytics*, Second Edition, Wiley, 2015

这是一本非常适合作为本书参考书的著作，通过本书可以深入研究数据挖掘和预测分析。

- *Discovering Knowledge in Data: An Introduction to Data Mining*, Second Edition, Wiley, 2014

Chantal D. Larose 于 2015 年在康涅狄格大学取得统计学博士学位，博士论文是 *Model-Based Clustering of Incomplete Data*。作为纽约州立大学新帕尔兹分校决策科学系的助理教授，她曾帮助学校创立了商务分析学理学学士学位。现在，作为东康涅狄格州立大学统计学与数据科学系的助理教授，她正在参与开发数据科学系的数据科学课程。

Daniel T. Larose 于 1996 年在康涅狄格大学取得统计学博士学位，其博士论文题目是 *Bayesian Approaches to Meta-Analysis*。他是中央康涅狄格州立大学统计学和数据科学系的教授。2001 年，他设立了世界上第一个数据挖掘方向的线上理学硕士学位。本书是他编写或合著的第 12 本书。他经营一家名为 DataMiningConsultant.com 的小型咨询公司。他还负责指导 CCSU 大学的线上数据科学硕士学位项目。

致　　谢

Chantal 的致谢

最诚挚的感谢献给我的父亲 Daniel，感谢他校对书稿时孜孜不倦和连珠妙语。他对本书编写的指导和热情感染我并提高了图书编写质量，我和他一起工作是一种享受。还要多多感谢我的小妹妹 Ravel，她对我充满无限的爱，并且她具有不可思议的音乐和科学天赋。她是我的同路人，并给予我写作灵感。感谢我的弟弟 Tristan，他在学校里勤奋刻苦。感谢我的母亲 Debra，感谢她对我无微不至的关爱。当然，还要感谢伴我写作的良友——咖啡。

Chantal D. Larose 博士
统计学与数据科学系的助理教授
东康涅狄格州立大学

Daniel 的致谢

我的所有感谢都是针对本人家庭的。我首先要感谢我的女儿 Chantal，感谢她极具洞察力的头脑、温柔的外表以及她每天带给我的快乐。感谢我的女儿 Ravel，感谢她的独特品质，她有勇气追随她的梦想并成为一名化学家。感谢我的儿子 Tristan，他具有数学和计算机方面的天赋，并且他在后院帮我清理石头。最后，感谢我亲爱的妻子 Debra，感谢这些年来她对我们一家人的深情关爱。我非常爱所有家人。

Daniel T. Larose 博士
统计学和数据科学系的教授
中央康涅狄格州立大学

序　言

为什么要阅读本书

原因之一：数据科学非常热门。数据科学现在确实炙手可热。彭博社称数据科学家为“美国最热门的职业”；商业内参称其为“目前美国最棒的工作”；Glassdoor.com 连续三年将其评为 2018 年全球最佳职位；哈佛商业评论称数据科学家为“21 世纪最光彩夺目的职业”。

原因之二：两大最流行的开源工具。Python 和 R 语言是世界上最流行的两个开源数据科学工具。世界各地的分析师和程序员协作开发了大量数据分析软件包，提供给 Python 和 R 用户免费使用。

《数据科学实战入门　使用 Python 和 R》一书将使用世界上应用最广泛的开源分析工具来培养你在这个前沿领域的专业技能。在本书中，你将亲身体验使用业界最先进的技术来逐步寻求针对实际业务问题的解决方案。简而言之，你将通过数据科学实践来学习数据科学。

本书适用于初学者和非初学者

《数据科学实战入门　使用 Python 和 R》一书是为普通读者编写的，读者不必具备数据分析和编程经验。我们知道，信息时代的经济正在驱使许多英语专业和历史专业的学生重新调整知识结构，以把握数据科学家巨大的市场需求的机会[1]。正是基于这种考虑，我们在本书提供了如下素材以帮助那些刚接触数据科学的读者能够快速入门。

- 专门为初学者编写了一章来学习 Python 和 R 的基础知识。初学者可以了解使用什么样的平台，下载哪些软件包以及了解入门阶段需要掌握的各种内容。
- 提供了附录 A“数据汇总与可视化”，专门用于填补读者在数据分析入门知识学习中存在的各种漏洞。附录 B“参考文献”扫描封底二维码获取。

1 例如，2017 年 5 月 IBM 公司预测，到 2020 年底对“数据科学家、数据开发人员和数据工程师”的年需求岗位将达到近 70 万个。

- 遍及全书的分步指导，每个步骤都提供了详细说明。
- 每章都有针对性的练习，可以通过练习检查自己的理解程度和学习进展情况。

对于那些有数据分析或编程经验的读者而言，他们将享受一站式学习如何使用 Python 和 R 进行数据科学实践的机会。无论是企业经理、信息总监(CIO)，还是首席执行官(CEO)和财务总监(CFO)，他们都乐意与数据分析师和数据库分析人员进行更好的沟通。本书重点强调要准确核算数据模型成本，这将有助于每个读者从庞杂的数据中发掘最有价值的知识，同时避免可能使贵公司蒙受数百万美元损失的潜在陷阱。

此外，《数据科学实战入门　使用 Python 和 R》一书涵盖了如下一些令人兴奋的新主题：

- 随机森林
- 广义线性模型
- 用于提高利润的数据驱动的错误成本

本书中使用的所有数据集都可扫描封底二维码获取。

《数据科学实战入门　使用 Python 和 R》一书可作为教材使用

《数据科学实战入门　使用 Python 和 R》一书很适合作为教材使用，既可以用于一学期的入门级课程教材，也可以用于两学期的入门和提高级系列课程教材。指导教师将会受益于每章末尾提供的练习，本书中总共有 500 多道习题。有三类习题，从对基本知识理解的测试到对新的具有挑战性的应用问题的更实际的分析。

- 概念辨析题。这些练习用于测试学生对书中基本知识的理解，以确保学生已掌握了所学习的内容。
- 数据处理题。这些应用类练习要求学生按照各章中给出的分步说明，使用 Python 和 R 处理数据。
- 实践分析题。这是学生在学习过程中真正需要学到的内容，学生将会使用新近掌握的知识和技能来发现新数据集中的潜在模式和趋势。学生将在接近实际的环境中培养他们的专业技能。本书中一半以上的练习都是实践分析类习题。

下面的一些补充材料也可以免费提供给本书的指导教师。

- 完整的解决方案手册。该手册不仅提供习题答案，还给出了详细的题解说明。
- 书中每章的幻灯课件。这些课件不仅便于学生阅读，还有助于学生理解书中的内容。

若想获取这些资料，可以联系当地 Wiley 出版社的销售代理，请求他们邮寄资料，前提是你必须选用本书作为教材。

《数据科学实战入门 使用 Python 和 R》一书适合本科高年级学生或研究生，学生不需要掌握统计学、计算机编程或数据库专业知识，唯一的要求是学生的上进心。

本书内容组织

《数据科学实战入门　使用Python和R》一书基于数据科学方法论进行内容的组织。数据科学方法是一种在科学框架体系内进行数据分析的阶段性、自适应和迭代式方法。

1. 问题理解阶段。首先，需要清晰地阐明项目目标；然后将这些目标转化为一种可以用数据科学解决的问题。

2. 数据准备阶段。数据清洗/准备阶段很可能是整个数据科学处理过程中最费力气的阶段。

- 相关内容参见第3章：“数据准备”。

3. 探索性数据分析阶段。在此阶段通过图形化探索方法获得对数据的初步认识。

- 相关内容参见第4章：“探索性数据分析”。

4. 设置阶段。建立数据模型的性能基准，如果需要，可以对数据进行分割和平衡处理。

- 相关内容详见第5章：“为建模数据做准备”。

5. 建模阶段。建模阶段是数据科学研究过程的核心，在此阶段应用各种先进的算法来发现隐藏在数据中的一些确实具有价值的关系。

- 相关内容参见第6章以及第8~14章。

6. 评估阶段。确定设计的模型是否有价值，在此阶段需要从一系列可选的模型中选择性能最佳的模型。

- 相关内容参见第7章：“模型评估”。

7. 部署应用阶段。在此阶段需要与管理层协作来调整模型以适应实际部署。

目　录

第 1 章
数据科学导引

1.1 为何学习数据科学

数据科学(data science)是当今全球发展最快的研究领域之一，该领域在 2017 年提供的就业机会已是 2012 年的 6.5 倍。预计未来对数据科学家的需求将持续井喷。举例来说，2017 年 5 月 IBM 公司预测，到 2020 年底对“数据科学家、数据开发人员和数据工程师”的年需求岗位将达到近 70 万个。根据 http://Infoworld.com 报告，“数据科学家在美国依然是最高端职业”的一个重要原因是“顶尖人才的短缺”。这正是我们撰写本书的动机——帮助培养合格的数据科学家。

1.2 何为数据科学

简而言之，数据科学就是在科学框架下对数据进行系统的分析。也就是说，数据科学的主要工作包括：

- 数据分析的自适应、迭代和分阶段方法；
- 在系统框架内对数据进行分析；
- 发现最优模型；
- 评估并核算预测误差的实际成本。

此外，数据科学结合了：

- 数据驱动的数据统计分析方法；
- 计算机科学的计算能力和编程活力；
- 领域相关的商务智能。

目的是从庞大的数据库中发掘具有实际操作意义和市场价值的有用信息。

换句话说，数据科学可以帮助人们从现有未充分利用的数据库中提取可操作的知识。因此，现在可以充分利用沉寂已久的数据仓库来发现数据中隐藏的价值并提高人们对数据的认知。通过数据科学，人们能够利用大量数据和强大的计算能力解决复杂的问题，或只有凭借数据的分析才能找到既定模式。这些发现可以带来令人激动的结果，例如对疾病患者进行更有效的治疗或为一个企业创造更多的利润。

1.3　数据科学方法论

遵循数据科学方法论(Data Science Methodology, DSM)[1]，有助于数据分析师了解自身正在执行数据分析的哪个阶段。图 1.1 通过如下几个阶段说明了 DSM 的自适应和迭代特性。

1. 问题理解阶段。开发团队是否经常发现他们之前竭尽全力解决的某个问题并非预期的问题呢？此外，营销团队和分析团队的工作目标是否常常并未达成一致呢？这一阶段我们试图避免这些易犯的错误。

a. 首先，必须清晰阐明项目的目标；

b. 然后，将这些目标转化为一种可以用数据科学加以解决的问题。

2. 数据准备阶段。各种数据来源的原始数据很少能直接用于数据分析算法。相反，原始数据需要被清洗以便执行后续数据分析。当数据分析师首次检查数据时，他们就会发现难以避免的数据质量问题，并且这些问题似乎总会发生。在数据准备阶段，我们需要解决上述问题。数据清洗/准备可能是整个数据科学处理过程中最困难的阶段。下面给出数据准备阶段需要完成的一个非完备的任务清单。

a. 识别异常数据并决定如何处理它们；

b. 对数据进行转换和标准化；

c. 对类别变量重新分类；

d. 对数值变量进行分箱处理；

e. 添加索引字段。

1 改编自数据挖掘的跨行业标准实践(Cross-Industry Standard Practice for Data Mining, CRISP-DM)。

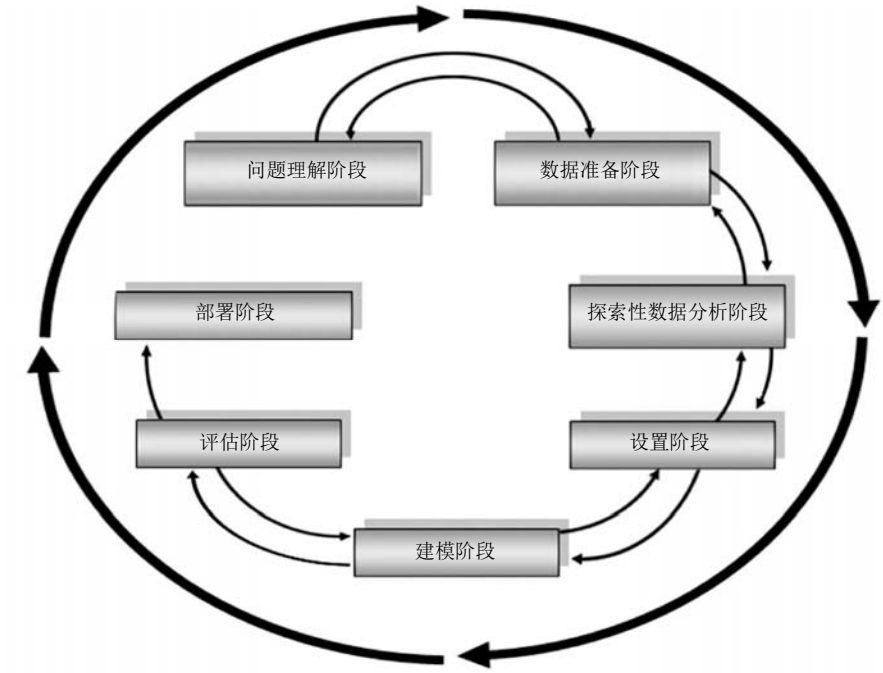

图 1.1　数据科学方法论：7 个阶段

数据准备阶段的详细内容参见第 3 章。

3. 探索性数据分析阶段。到这一阶段，待处理数据已变得干净且整齐，现在可以开始探索数据并试图获取一些基本信息。在此关注图形化数据探索。现在还不是应用复杂算法的时候，相反我们希望使用简单的探索方法帮助我们获得一些对数据的初步认识。在这一阶段，你很可能会发现只需要使用这些简单的方法，就能获悉很多信息。下面列出一些可供采用的方法。

a. 探索自变量与目标变量之间的一元关系；

b. 探究变量之间的多元关系；

c. 基于预测值的分箱以增强数据模型；

d. 根据现有变量的组合导出新变量。

我们将在第 4 章中阐述探索性数据分析阶段。

4. 设置阶段。此时，我们已基本为开始数据建模做好了准备。在这一阶段，我们还需要先处理少量重要且烦琐的事务，例如：

a. 交叉验证，可以是 2 折或 n 折，这对于避免数据疏浚是必需的。此外，还需要对数据的划分进行评估，以确保它们确实是随机的。

b. 平衡数据。这有助于提高某些算法揭示数据中所蕴含关系的能力。

c. 建立性能基准。假设曾告知你我们有一个模型能够以 99%的概率正确预测某一信用卡交易是否存在欺诈，你是否会感到吃惊呢？你应该不会，由于实际上非欺诈性的交易概率为 99.932%。因此，我们的模型可以简单地预测每一笔交易都是非欺诈性的，并且该模型的正确率可达 99.932%。这一事例说明了为数据模型建立适当性能基准的重要性，以便可以校准模型并确定它们是否有用。

第 5 章将对设置阶段加以介绍。

5. 建模阶段。在建模阶段将有机会应用各种先进的算法发现隐藏在数据中的一些确实具有价值的关系。建模阶段是对数据进行科学研究的核心，包括以下内容：

a. 选择和实施适当的建模算法。技术应用不当将导致不准确的分析结果，这可能会使你的公司损失大笔资金。

b. 确保我们采用的模型优于基准模型。

c. 对模型算法进行微调以优化结果。例如，是否应该加宽或加深我们的决策树？我们的神经网络应该含有一个还是两个隐藏层？最大化我们收益的临界点应该是什么？分析师往往需要花费一些时间对他们的模型进行微调，以便得到最佳的解决方案。

建模阶段是数据科学工作的核心，将在第 6 章和第 8~14 章进行详细介绍。

6. 评估阶段。你的同事可能觉得他对超级碗比赛的预测很有把握，但是他的预测究竟有用吗？这确实是一个问题。任何人都能做出预测，但是预测相对于实际数据的表现确实是真正的测试。在评估阶段，我们评价我们的模型的运行情况，模型是否有价值，或者我们是否需要返工并设法改善我们的预测模型。

a. 需要根据源自设置阶段的性能基准度量对你的模型进行评估。我们是否优于猴子投掷飞镖模型呢？如果没有，最好再尝试改进一下模型。

b. 需要确定你的模型是否真正解决了手头的问题。你的模型实际上是否达到了之前在问题理解阶段为其设定的目标？是否没有充分考虑待解决问题的某些重要方面呢？

c. 考虑数据固有的错误代价，因为数据驱动的成本评估是模拟实际成本的最佳方法。例如，在市场营销活动中，假阳性的代价不如假阴性的代价高。然而，对于抵押贷款机构来说，假阳性将付出高昂的代价。

d. 你应该定制一系列模型，并确定表现最好的模型。选择单个最佳模型或少量较优模型，然后进入部署阶段。

第 7 章将介绍评估阶段。

7. 部署阶段。至此，你的模型终于为部署应用的黄金时段做好了准备！向管理层上报你的最佳模型，并与管理层协作来调整模型以适应实际部署。

a. 编写一份结果报告可视为一个最简单的部署使用示例。在你的报告中，重点描述管理层感兴趣的结果，要向管理层表明你解决了问题，并且尽可能说明预估的收益。

b. 你应继续参与之前的项目！参与模型部署使用中涉及的各种会议和流程，以便使模型始终致力于解决手头的问题。

应该强调的是，DSM 是迭代和自适应的。所谓自适应，我们意指为了执行后续工作，有时根据当前阶段获得的一些知识，我们认为有必要返回之前的某个阶段。这也正是在图 1.1 中为什么大多数阶段之间都存在双向箭头的原因。例如，在评估阶段，我们可能会发现我们创建的模型实际上并没有解决最初提出的问题，那么就需要返回到建模阶段开发一个能胜任的模型。

此外，DSM 是迭代式的，因为有时可以利用我们在类似问题上的经验来构建一种有效的模型。也就是说，我们创建的模型可以用于调查相关问题的起点。这也正解释了图 1.1 中的外层一圈箭头展现了通过已有模型的持续循环，用于考察针对新问题的新解决方案。

1.4　数据科学任务

下面列出了一些最常见的数据科学任务：

- 描述
- 估计
- 分类
- 聚类
- 预测
- 关联

接下来，将说明每个任务的具体内容以及在哪些章节介绍这些任务。

1.4.1　描述

数据科学家最常见的一项任务就是描述隐含在数据中的模式和趋势。举例来说，数据科学家会将最可能放弃我们公司服务的客户群体描述为拨打客户服务电话次数较多且占线时间较长的那组客户。在描述了这类客户群体之后，数据科学家会解释说拨打客户服务电话次数较多意味着客户不满意。因此，通过与营销团队合作，数据分析师可以建议应该采取的干预措施以设法挽留此类客户。

数据描述任务在世界各地被专家和非专业人士广泛使用。例如，当体育播音员评论一名棒球运动员职业生涯中的平均击球率(击中数/击打数)为 0.350 时，他描述的是该运动员的职业生涯的击球表现。这是描述性统计的一个例子[1]，在“附录 A：数据汇总与可

1 参阅 *Discovering Statistics*, Daniel T. Larose, W.H. Freeman，2016。

视化”中可以找到更多的示例。此外，本书中几乎每一章都包含描述任务的例子，包括第 4 章中的图形化 EDA 方法、第 10 章中的数据聚集描述以及第 11 章中的二元关系。

1.4.2 估计

估计就是指使用一组自变量粗略估算数值目标变量的值。估计模型是使用目标值已知的记录建立的，因此该模型不仅能够获悉哪些目标值与自变量的值相关联，而且该估计模型可以估计未知的新数据的目标值。例如，数据分析师可以根据一组个人和人群的统计数据，估算可以为某个潜在客户提供的抵押贷款金额。这种估计的模型是基于调研之前的为客户提供贷款数额的模型构建的，这种估计要求目标变量是数值型的。估计方法的具体内容参见第 9 章、第 11 章和第 13 章。

1.4.3 分类

分类与估计有些类似，区别在于其目标变量是离散的而不是连续的。分类很可能是数据科学中最常见的任务，也是最容易盈利的任务。例如，抵押贷款机构希望了解哪些客户有可能会拖欠抵押贷款，这种情况也同样适用于信用卡公司。分类模型可以显示包含既有客户实际违约状态的大量完整记录。因此，模型可以学习到哪些属性与违约的客户相关联。最后，可以将这些经过训练的模型应用到新的数据中，即申请贷款或信用卡的客户，期望这些模型有助于甄别哪些客户最可能拖欠贷款。分类方法详见第 6、第 8、第 9 和第 13 章。

1.4.4 聚类

聚类任务旨在识别相似的记录组。例如，在一组信用卡申请人的数据中，一个聚类(或一组数据)可能代表较年轻、受教育程度较高的客户，而另一个聚类可能代表较年长、受教育程度较低的客户。聚类的思想是，同一个聚类中的各个记录彼此相似，但不同聚类中的各个记录相差较大。寻找适当的聚类至少在两个方面是有用的：(1)你的客户可能对聚类说明感兴趣，即每组客户特征的详细描述；(2)聚类本身可以用作后续分类或估计模型的输入。第 10 章将介绍聚类方法。

1.4.5 预测

预测任务也与估计或分类相似，只是预测与未来有关。例如，一位金融分析师可能很有兴趣预测未来三个月苹果公司股票的价格。这种预测即代表估计，因为股票价格是一种数值变量，也是一种预测，因为它与未来有关。再举一个例子，药品研制化学家可

能会对某一特定成分能否有助于为制药公司研制出畅销的新药品感兴趣。这个例子中既有预测也有分类，因为目标变量是一种“是/否”变量，即表示药物是否能盈利。

1.4.6　关联

关联任务旨在确定哪些属性相互关联，即哪些属性“关系紧密”。数据科学家使用关联方法，试图揭示量化两个或多个属性之间关系的潜在规则。这些关联规则通常采取“先有前提后有结果”的形式，并且包含支持度和信任度测量。举例来说，试图避免客户流失的营销人员可能会发现如下关联规则：“如果顾客拨打客服电话超过三次，那么该顾客将流失”。支持度是指规则适用的记录比例，而信任度是指规则执行正确的比例。我们将在第 14 章中讨论关联任务。

1.5　习题

概念辨析

1. 简要说明数据科学的概念。
2. 数据科学涉及哪些研究领域？
3. 数据科学的目标有哪些？
4. 阐述 DSM 的 7 个阶段。
5. 含有一个问题理解阶段可以带来什么好处？
6. 为什么需要数据准备阶段？请说明本阶段需要处理哪三个问题。
7. 在哪个阶段数据分析师开始探索数据来了解一些简单的信息？
8. 用自己的话阐明为何需要为我们的模型确立一个性能基准。这一工作出现在哪个阶段？
9. 数据科学研究的核心是哪个阶段？解决一个特定问题为何往往需要采用多种算法？
10. 如何确定我们的预测是否有用？这一决定出现在哪个阶段？
11. 判断对错并解释原因：数据科学家的工作到评估阶段就结束了。
12. 解释 DSM 为何是自适应的。
13. 描述 DSM 的迭代特性。
14. 列举最常见的数据科学任务。
15. 上述数据科学任务中有哪些是许多非专业人士一直都在从事的任务？

16. 什么是数据估计？对于估计而言，目标变量必须满足什么条件？

17. 数据科学最常见的任务是哪一项任务？对于该任务，目标变量需要满足什么条件？

18. 什么是聚类说明？

19. 判断对错并解释原因：预测只能用于离散的目标变量。

20. 对于关联规则而言，支持度代表什么？