深入浅出数据科学

Principles of Data Science

[美]斯楠·奥兹德米尔（Sinan Ozdemir） 著　张星辰 译

掌握必需的软件技能和数学知识，真正理解你的数据

Packt>

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

深入浅出数据科学

Principles of Data Science

[美]斯楠·奥兹德米尔（Sinan Ozdemir） 著　张星辰 译

人民邮电出版社
北京

图书在版编目（CIP）数据

深入浅出数据科学 / (美) 斯楠·奥兹德米尔 (Sinan Ozdemir) 著 ; 张星辰译. -- 北京 : 人民邮电出版社, 2018.10
ISBN 978-7-115-48126-9

Ⅰ. ①深… Ⅱ. ①斯… ②张… Ⅲ. ①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第056068号

版权声明

◆ 著 [美] 斯楠·奥兹德米尔（Sinan Ozdemir）
译 张星辰
责任编辑 王峰松
责任印制 焦志炜

◆ 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
北京市艺辉印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：20.75 彩插：2
字数：332 千字 2018 年 10 月第 1 版
印数：1－3 000 册 2018 年 10 月北京第 1 次印刷

著作权合同登记号 图字：01-2016-8085 号

定价：69.00 元

读者服务热线：(010) 81055410 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

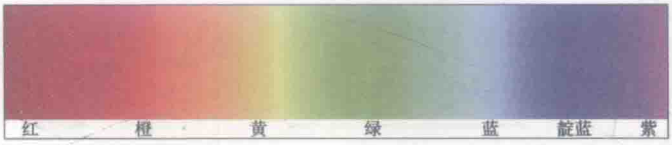

图 2.2　自然排序规则

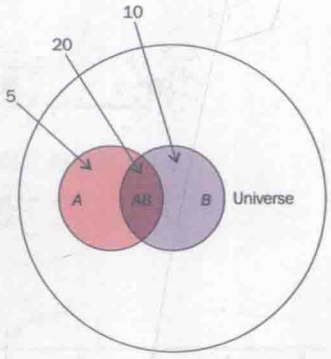

图 5.6　事时 *A* 和事件 *B* 的并集

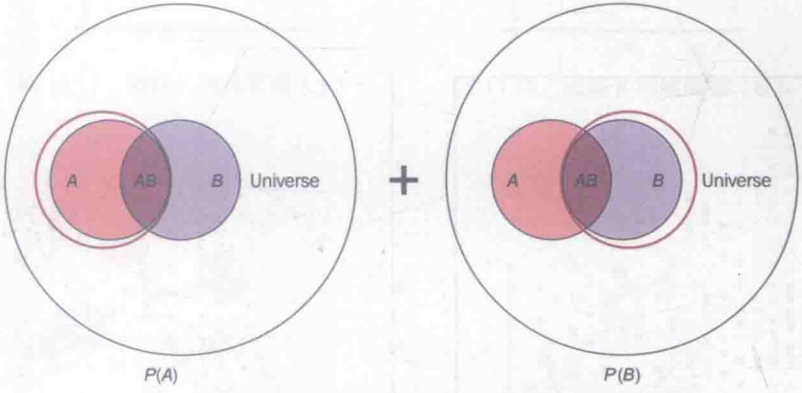

图 5.8　将事件 *A* 和事件 *B* 的圆形区域相加

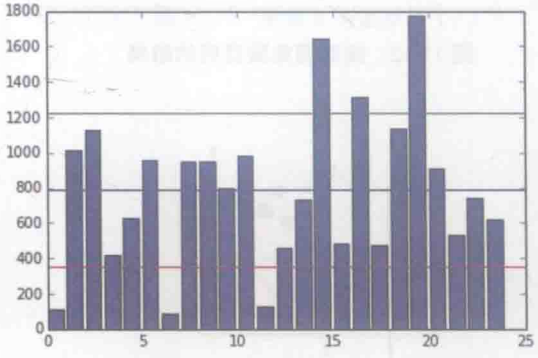

图 7.1　用条形图可视化用户及其好友数量

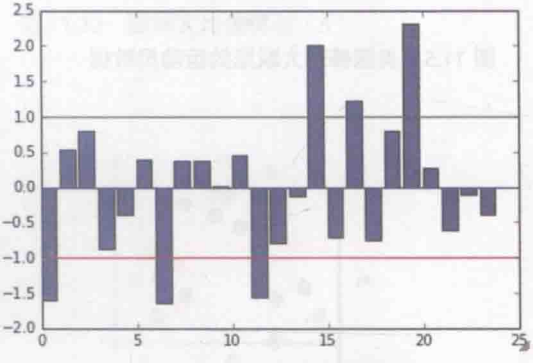

图 7.3　对图 7.2 加入 3 条辅助线

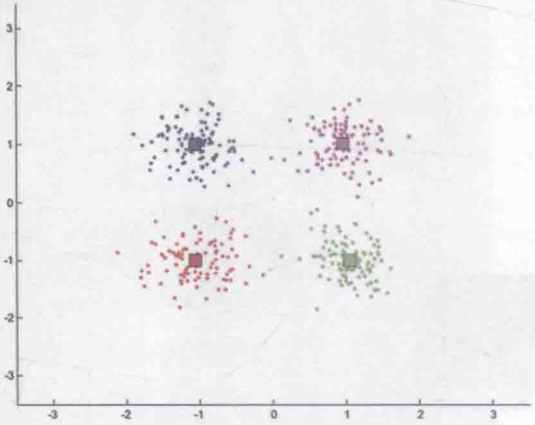

图 10.7　聚类模型的输出结果

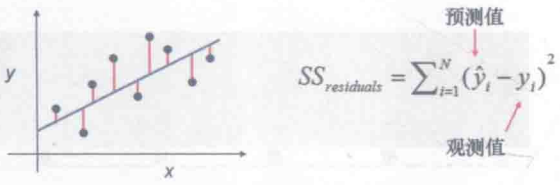

图 10.13　线性回归模型

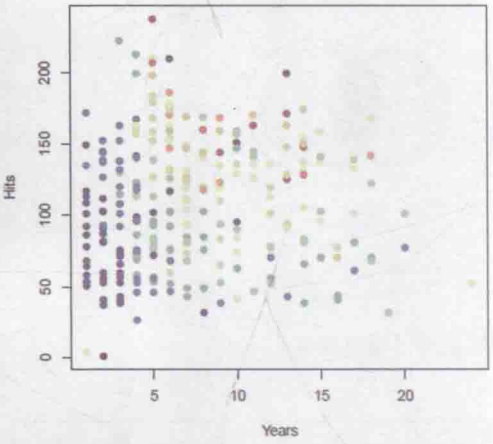

图 11.5　美国棒球大联盟的运动员数据

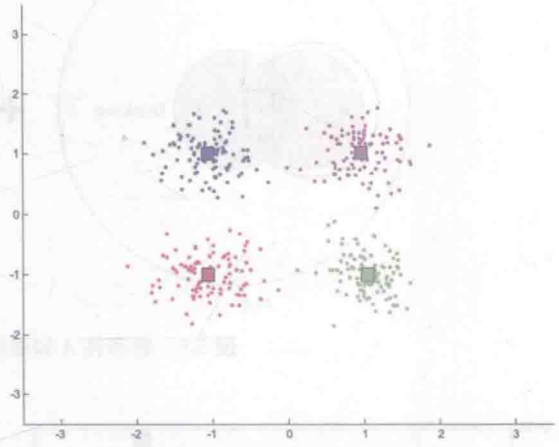

图 11.12　某数据集聚类后的结果

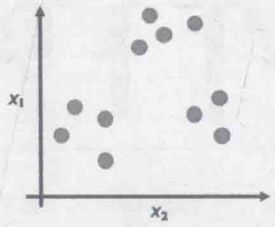

图 11.13　图解 K 均值聚类（1）

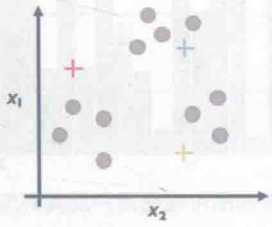

图 11.14　图解 K 均值聚类（2）

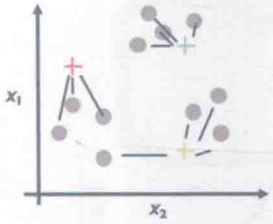

图 11.15　图解 K 均值聚类（3）

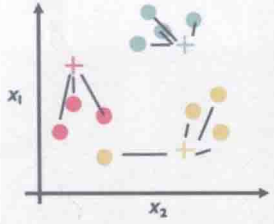

图 11.16　图解 K 均值聚类（4）

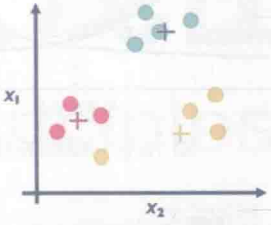

图 11.17　图解 K 均值聚类（5）

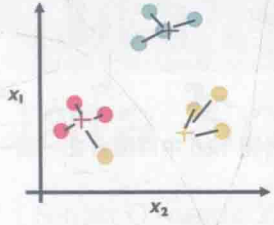

图 11.18　图解 K 均值聚类（6）

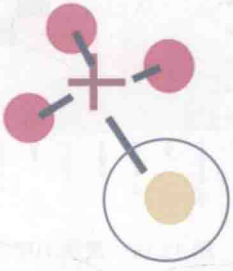

图 11.19　图解 K 均值聚类（7）

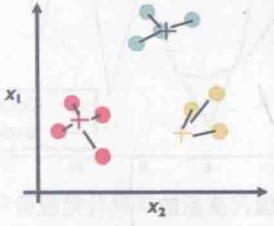

图 11.20　图解 K 均值聚类（8）

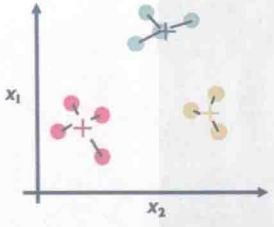

图 11.21　图解 K 均值聚类（9）

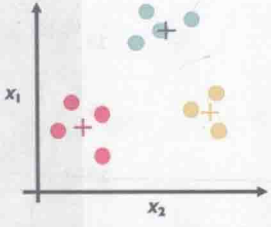

图 11.22　图解 K 均值聚类（10）

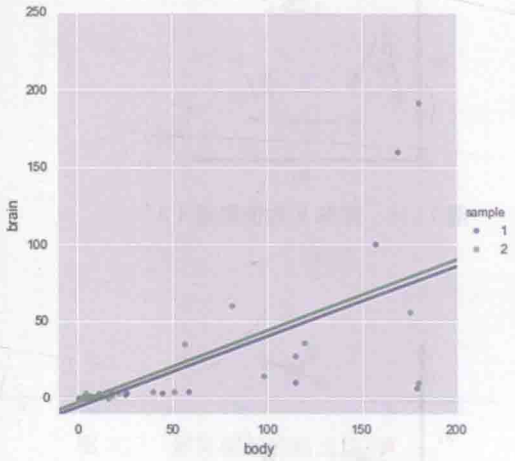

图 12.7　将图 12.6 中两图合在一起

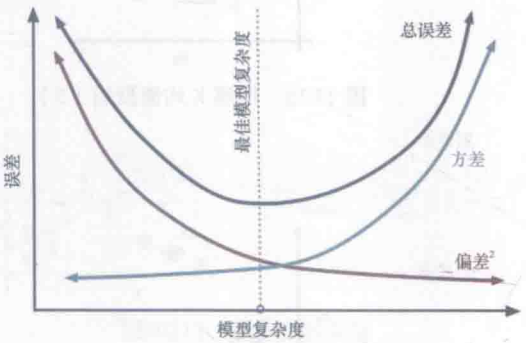

图 12.10　模型的总误差

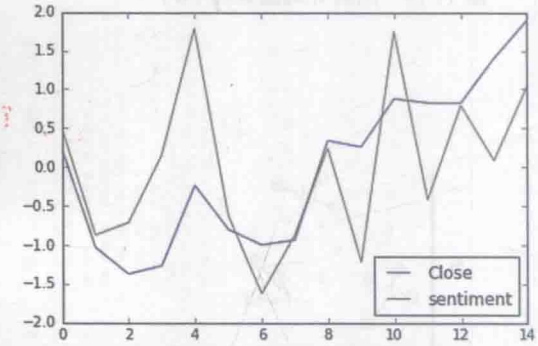

图 13.12　调整特征值尺度后的苹果公司股价和成交量

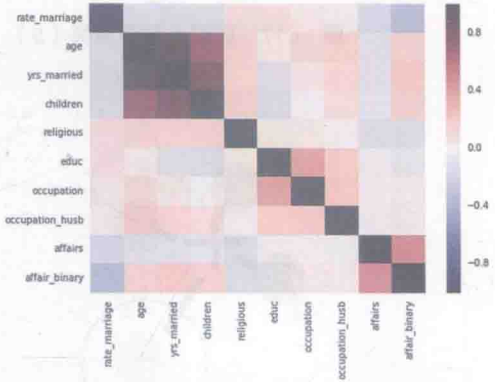

图 13.18　用热力图表示的相关性矩阵

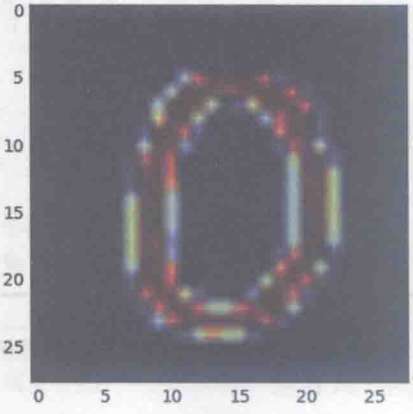

图 13.22　数据集中的图片

内容提要

数据科学家是目前最热门的职业之一。本书全面介绍了成为合格数据科学家所需的必备知识、技能和工作流程，是一本内容全面的实用性技术图书。

本书分为 13 章，其中第 1～3 章介绍数据科学；第 4～8 章介绍数学知识，包括统计学和概率论；第 9 章介绍数据可视化；第 10～12 章介绍机器学习；第 13 章介绍案例。各个章节内容均由浅入深，同时通过案例和 Python 代码，使读者掌握实战技能。

本书适合有志于成为数据科学家的师生或业界新手，同时也适合经验丰富的职场老手参考。

关于作者

Sinan Ozdemir 是一名数据科学家、创业者和教育工作者。Sinan 的学术生涯在约翰·霍普金斯大学（The Johns Hopkins University）渡过，主修数学专业。随后他从事教育事业，曾经在约翰·霍普金斯大学和 General Assembly 公司举办多次数据科学讲座。在此之后，他创立了旨在通过人工智能技术和数据科学力量帮助企业销售团队的创业公司 Legion Analytics。

在完成 Y Combinator 创业加速器之后，Sinan 的大部分时间都在经营这家快速成长的公司，同时也会制作一些数据科学教育资料。

我要感谢父母和姐姐对我的帮助，同时还要感谢我的导师们，包括约翰·霍普金斯大学的 Pam Sheff 博士和 Sigma Chi 兄弟会的 Nathan Neal。

感谢 Packt 出版社给予我和大家分享数据科学原理的机会，我非常激动这个领域将在未来数年改变我们所有人的生活。

译者简介

张星辰，北京荣之联科技股份有限公司 BI 技术顾问，毕业于重庆邮电大学，具有 5 年数据相关工作经验，熟悉商业智能和数据可视化，通过了微软数据科学专业认证。

中文版审校人

鲜思东，重庆邮电大学教授，硕士生导师，复杂系统智能分析与决策重庆市高校重点实验室副主任，中国商业统计学会理事。现任国际期刊《Advancements in Case Studies》编辑，担任《Knowledge-Based Systems》和《IEEE Transactions on Systems、Man and Cybernetics: Systems》等多个国际期刊的审稿人。

洪贤斌，西交利物浦大学、英国利物浦大学机器学习方向博士生，苏州谷歌开发者社区组织者。

英文版审稿人

Samir Madhavan 有 6 年的数学科学实战经验，著有《Mastering Python for Data Science》一书。Samir 曾是 Mindtree 公司反欺诈算法团队 Aadhar 的一员，参与了 UID（Unique Identification）项目。此后，他作为首批员工加入 Flutura Decision Sciences and Analytics 公司，作为核心成员帮助团队壮大至数百人。在此期间，他利用大数据和机器学习技术帮助 Flutura 公司进行商业拓展。目前，他是总部设在波士顿的高科技医药公司 Zapprx 的数据分析团队负责人，帮助 Zapprx 打造数据驱动的产品，以便更好地服务客户。

Oleg Okun 是机器学习专家，曾以作者和编辑身份参与 4 本图书的出版，以及多篇文章、会议论文的发表。Oleg 的职业生涯超过 25 年，他活跃在白俄罗斯的学术界和工业界，也曾在芬兰、瑞士和德国工作过。Oleg 的工作经验包括图片分析、指纹生物识别、生物信息、在线/离线营销分析和信用评分分析。

Oleg 对机器学习和物联网（IoT）充满了热情，目前在德国汉堡工作。

我要向父母表达最深切的感谢，感谢他们为我做的一切。

前言

本书的主题是数据科学。在过去几十年，这个领域的研究和应用都获得了飞速发展。作为一个快速发展的领域，数据科学正在吸引媒体和就业市场的关注。2015 年，美国政府任命 DJ Patil 为史上第一任首席数据科学家。坦白讲，这一举动是对科技公司最近大举招募数据团队行为的效仿。数据科学技能受到广泛欢迎，其人才市场需求必将远远超过今天的就业市场。

这本书将力求弥合数学、编程和专业领域之间的差距。很多人是某一个（或者两个）领域的专家，但合理地使用数据科学需要同时精通以上 3 个领域。我们将深入讨论这 3 个领域并解决复杂的问题。我们将清洗、探索和分析数据，得出科学、准确的结论。我们还将利用机器学习和深度学习技术解决更加复杂的数据问题。

本书涵盖的内容

第 1 章：如何听起来像数据科学家。本章将介绍数据科学家常用的专业术语和解决的问题类型。

第 2 章：数据的类型。本章将介绍不同类型和尺度的数据，以及如何处理这些数据。从本章起，我们将介绍数据科学必备的数学知识。

第 3 章：数据科学的 5 个步骤。本章将介绍实施数据科学的 5 个基本步骤，包括数据探索和数据获取、数据建模、可视化分享等。本章会通过案例对每个步骤进行详细介绍。

第 4 章：基本的数学知识。本章将介绍微积分、线性代数等数学知识，并用案例介

绍它们如何帮助数据科学家做出判断。

第 5 章：概率论入门：不可能，还是不太可能。本章将以初学者的视角介绍概率论的基本理论，以及如何利用概率论从随机世界中获取知识。

第 6 章：高等概率论。本章将尝试利用上一章介绍的概率论知识，比如贝叶斯推理，探索现实世界中隐藏的意义。

第 7 章：统计学入门。本章将使用统计推断法解决问题，包括基本的统计试验、正态分布和随机抽样等方法。

第 8 章：高等统计学。本章将使用假设检验、置信区间等方法对试验结果进行评价，学会正确理解 p 值和其他指标的含义。这些技能至关重要。

第 9 章：交流数据。本章将介绍相关性和因果关系如何影响我们对数据的理解，学会通过数据可视化和他人交流数据，分享数据科学分析结果。

第 10 章：机器学习精要：你的烤箱在学习吗。本章将介绍机器学习的定义，并通过真实案例介绍机器学习可以在何时、以何种方式被使用。本章还将介绍如何评价模型。

第 11 章：树上无预言，真的吗。本章将介绍更复杂的机器学习模型，比如决策树模型和贝叶斯预测模型，以解决更复杂的数据问题。

第 12 章：超越精要。本章将介绍数据科学中的一些神秘力量，包括偏差和方差。神经网络将作为一种流行的深度学习技术进行介绍。

第 13 章：案例。本章将通过一系列案例强化你对数据科学的理解。我们将通过预测股价、笔迹检测等案例，反复演练数据科学工作流的全过程。

你需要做的准备工作

本书使用 Python 代码演示所有案例。你需要一台安装了 Python 2.7 且有 UNIX 风格终端窗口的计算机（Linux/Mac/Windows）。我也推荐使用 Anaconda，它包含了本书案例中使用的大多数 Python 包。

本书面向的读者

本书面向希望学习数据科学，并在各自领域中使用它的人。

读者需要有基本的数学基础（代数或概率论），能够阅读 R/Python 脚本以及伪码。读者不一定要有数据领域的经验，但必须对学习和使用本书所讲的技能具有热情，无论是对自己的数据集还是其他数据集。

约定

本书运用了多种不同的文本格式，以便对相关信息进行区分。以下是部分文本格式和对它们的解释。

代码块的格式如下：

```
tweet = "RT @j_o_n_dnger: $TWTR now top holding for Andor, unseating $AAPL"
words_in_tweet = first_tweet.split(' ')     # list of words in tweet
```

当我想提醒你注意某段代码时，相关代码行和信息将被加粗：

```
for word in words_in_tweet:              # for each word in list
  if "$" in word:                        # if word has a "cashtag"
  print "THIS TWEET IS ABOUT", word      # alert the user
```

警告或重要备注出现在这种类型的提示框。

提示和技巧出现在这种类型的提示框。

资源与支持

本书由异步社区出品，社区（https://www.epubit.com/）为您提供相关资源和后续服务。

配套资源

本书提供如下资源：

- 本书源代码；
- 书中彩图文件。

要获得以上配套资源，请在异步社区本书页面中点击 配套资源，跳转到下载界面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

提交勘误

作者和编辑尽最大努力来确保书中内容的准确性，但难免还会存在疏漏。欢迎您将发现的问题反馈给我们，帮助我们提升图书的质量。

当您发现错误时，请登录异步社区，搜索到本书页面，点击“提交勘误”，输入相关信息，点击“提交”按钮即可。本书的作者和编辑会对您提交的勘误进行审核，确认并接受后，您将获赠异步社区的 100 积分。积分可用于在异步社区兑换优惠券，或者用于兑换样书或奖品。

扫码关注本书

扫描下方二维码，您将会在异步社区微信服务号中看到本书信息及相关的服务提示。

与我们联系

我们的联系邮箱是 contact@epubit.com.cn。

如果您对本书有任何疑问或建议，请您发邮件给我们，并请在邮件标题中注明本书书名，以便我们更高效地做出反馈。

如果您有兴趣出版图书、录制教学视频，或者参与图书翻译、技术审校等工作，可以发邮件给我们，或者到异步社区在线提交投稿（直接访问 www.epubit.com/selfpublish/submission 即可）。

如果您是学校、培训机构或企业，想批量购买本书或异步社区出版的其他图书，也可以发邮件给我们。

如果您在网上发现有针对异步社区出品图书的各种形式的盗版行为，包括对图书全部或部分内容的非授权传播，请您将怀疑有侵权行为的链接发邮件给我们。您的这一举动是对作者权利的保护，也是我们持续为您提供有价值的内容的动力之源。

关于异步社区和异步图书

“异步社区”是人民邮电出版社旗下IT专业图书社区，致力于出版精品IT技术图书和相关学习产品，为作译者提供优质出版服务。社区创办于2015年8月，提供超过1000种图书、近千种电子书，以及众多技术文章和视频课程。更多详情请访问异步社区官网https://www.epubit.com。

“异步图书”是由异步社区编辑团队策划出版的精品IT专业图书的品牌，依托于人民邮电出版社近30年的计算机图书出版积累和专业编辑团队，相关图书在封面上印有异步图书的LOGO。异步图书的出版领域包括软件开发、大数据、AI、测试、前端、网络技术等。

社区二维码	服务号二维码