Packt

# 精通数据科学算法

## Data Science Algorithms in a Week

[英] 戴维·纳蒂加（David Natingga）著
封强 赵运枫 范东来 译

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

# 精通数据科学算法

**Data Science Algorithms in a Week**

[英] 戴维·纳蒂加（David Natingga）著
封强 赵运枫 范东来 译

人民邮电出版社
北京

**图书在版编目（CIP）数据**

精通数据科学算法 / (英) 戴维·纳蒂加
(David Natingga) 著 ; 封强, 赵运枫, 范东来译. --
北京 : 人民邮电出版社, 2019.5
ISBN 978-7-115-49816-8

Ⅰ. ①精… Ⅱ. ①戴… ②封… ③赵… ④范… Ⅲ.
①数据处理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第246378号

## 版权声明



◆ 著　[英] 戴维·纳蒂加（David Natingga）
译　封　强　赵运枫　范东来
责任编辑　武晓燕
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
雅迪云印（天津）科技有限公司印刷

◆ 开本：720×960　1/16
印张：11.25
字数：193 千字　2019 年 5 月第 1 版
印数：1－2 500 册　2019 年 5 月天津第 1 次印刷

著作权合同登记号　图字：01-2017-9024 号

定价：59.00 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**

# 内容提要

数据科学（Data Science）是从数据中提取知识的技术，是一门有关机器学习、统计学与数据挖掘的交叉学科。数据科学包含了多种领域的不同元素，包括信号处理、数学、概率模型技术和理论、计算机编程、统计学等。

本书讲解了7种重要的数据分析方法，它们分别是k最近邻算法、朴素贝叶斯算法、决策树、随机森林、k-means聚类、回归分析以及时间序列分析。全书共7章，每一章都以一个简单的例子开始，先讲解算法的基本概念与知识，然后通过对案例进行扩展以讲解一些特殊的分析算法。这种方式有益于读者深刻理解算法。

本书适合数据分析人员、机器学习领域的从业人员以及对算法感兴趣的读者阅读。

# 作者简介

Dávid Natingga于2014年毕业于伦敦帝国理工学院的计算与人工智能专业，并获工程硕士学位。2011年，他在印度班加罗尔的Infosys实验室工作，研究机器学习算法的优化。2012~2013年，他在美国帕罗奥图的Palantir技术公司从事大数据算法的开发工作。2014年，作为英国伦敦Pact Coffee公司的数据科学家，他设计了一种基于顾客口味偏好和咖啡结构的推荐算法。2017年，他在荷兰阿姆斯特丹的TomTom工作，处理导航平台的地图数据。

他是英国利兹大学计算理论专业的博士研究生，研究纯数学如何推进人工智能。2016年，他在日本高等科学技术学院当了8个月的访问学者。

# 致　　谢

我很感谢Packt出版社为我提供的这个机会，通过本书分享我在数据科学方面的知识和经验。我由衷地感谢我的妻子Rheslyn，她的耐心、爱与支持贯穿了本书的整个写作过程。

# 评阅者简介

Surendra Pepakayala是一位经验丰富的技术专家和企业家，在美国和印度有超过19年的工作经验。他在印度和美国的公司担任开发人员、架构师、软件工程经理和产品经理，在构建企业/Web软件产品方面拥有丰富的经验。他同时是一位在企业/Web应用程序开发、云计算、大数据、数据科学、深度学习和人工智能方面具有深厚兴趣和专业知识的技术人员/黑客。

他在美国企业工作了11年之后成为企业家，他成立了一个公司，为美国提供BI/DSS产品。随后他出售了该公司，开始从事云计算、大数据和数据科学咨询业务，帮助初创企业和IT组织简化其开发工作，缩短产品或解决方案的上市时间。此外，Surendra还通过自己丰富的IT经验将亏损的产品/项目变得盈利，他为此感到自豪。

他同时是eTeki（一个按需采访平台）的顾问，他在面试环节的贡献使eTeki在招聘和留住世界级IT专业人员方面处于领先地位。他对CGEIT、CRISC、MSP和TOGAF等各种IT认证草案的修改建议和相关问题进行了审查。他目前的工作重点是将深度学习应用于招聘流程的各个阶段，帮助人事部门（Human Resource，HR）找到最佳人才，并减少招聘过程中的摩擦。

# 前　言

数据科学是一门有关机器学习、统计学与数据挖掘的交叉学科，它的目标是通过算法和统计分析方法从现存数据中获取新知识。在本书中，你将会学习数据科学中7种重要的数据分析方法。每章将首先通过一个简单的例子解释某算法或分析某概念，然后用更多的例子与练习建立与拓展一些特殊的分析方法。

## 本书涵盖的内容

第1章，用k最近邻算法解决分类问题，基于 $k$ 个最相似的项对数据项分类。

第2章，朴素贝叶斯，学习用贝叶斯定理来计算某个数据项属于某一个特定类的概率。

第3章，决策树，将决策准则整理、归纳成树的分支，并用一个决策树将数据项分类到叶节点所在的类中。

第4章，随机森林，用决策树集成的方式来划分数据项，通过减少偏差的负面影响来提高算法的准确率。

第5章，k-means聚类，将数据划分成 $k$ 个簇来寻找模式和数据项之间的相似度，并应用这些模式划分新的数据。

第6章，回归分析，通过一个方程对数据进行建模，并以这种简单的方式对

未知数据进行预测。

第7章，时间序列分析，通过揭示依赖时间的数据的发展趋势和重复模式来预测未来的股票市场、比特币价格和其他的时间事件。

附录A，统计，提供一个对数据科学家实用的统计方法和分析工具的概要。

附录B，R参考，涉及基本的R语言结构。

附录C，Python 参考，涉及基本的Python语言结构、整本书所用到的命令和函数。

附录D，数据科学中的算法和方法术语，提供数据科学与机器学习领域中一些非常重要并且实用的算法和方法术语。

## 阅读本书所需要的开发工具

最重要的是，保持一个积极的态度去思考问题——许多新的知识隐藏在练习中。同时，你也需要在自己选择的系统中运行Python和R程序。本书的作者是在Linux操作系统中使用命令行来运行编程语言的。

## 本书适合的读者

本书是为熟悉 Python 和 R 语言并且有统计背景、期望成为一名数据科学专业人士的读者准备的。那些目前正在开发一两种数据科学算法，并且现在想学习更多的知识以扩展他们技能的开发人员将会发现这本书是非常有用的。

## 体例约定

本书应用了不同的文本样式以便区别不同种类的信息。这里列举部分的示例并对其含义做出解释。文本中的代码、数据库表名、文件夹名称、文件的扩展名、路径名、虚拟 URL、用户输入和 Twitter 句柄如下所示：“对于这章前面的可视化描述部分，将会用到 matplotlib 库。”

代码块如下所示：

```
import sys
sys.path.append('..')
sys.path.append('../../common')
import knn # noqa
import common # noqa
```

任意的命令行输入或者输出如下所示：

```
$ python knn_to_data.py mary_and_temperature_preferences.data
mary_and_temperature_preferences_completed.data 1 5 30 0 10
```

警告信息或者重要注释的标志。

温馨提示和小技巧的标志。

# 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：

- 配套代码。

要获得以上配套资源，请在异步社区本书页面中单击 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，按书名搜索，进入本书页面，单击“提交勘误”，输入勘误信息，单击“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、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；有意出版图书的作者也可以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益的保护，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。异步社区创办于2015年8月，提供大量精品IT技术图书和电子书，以及高品质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。

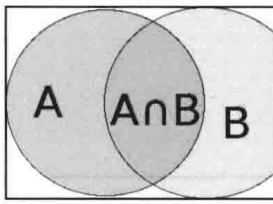
异步社区

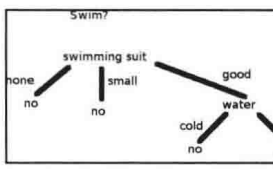
微信服务号

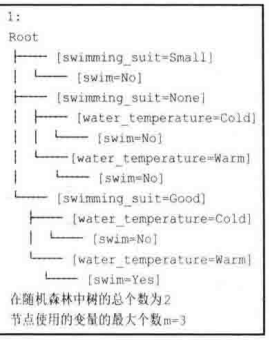
# 目 录

CONTENTS

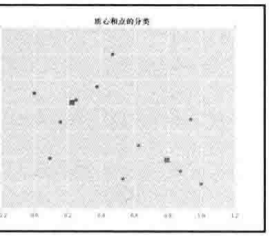


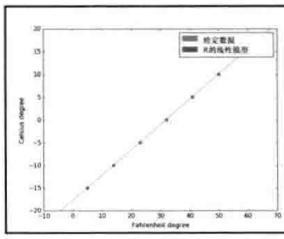


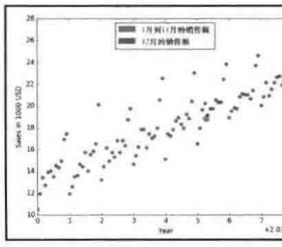










# 第1章
# 用k最近邻算法解决分类问题

最近邻算法可以基于某数据实例的邻居来判定该实例的类型。*k* 最近邻算法从距离该实例最近的 *k* 个邻居中找出最具代表性的类型，并将其赋给该数据实例。

本章将介绍 k-NN 算法的基础知识，并通过一个简单的例子——Mary 对温度的偏好来理解和实现 k-NN 算法。在意大利的示例地图上，您将学习如何选择正确的 *k* 值，以使算法正确执行并达到最高的准确率。您将从房屋偏好的例子中学习如何重新调整 k-NN 算法的数值参数。在文本分类的例子中，您将学习如何选择一个好的标准来衡量数据点之间的距离，以及如何消除高维空间中不相关的维度以保证算法的正确执行。

## 1.1 Mary 对温度的感觉

举个例子，如果 Mary 在 10℃的时候感觉冷，但在 25℃的时候感觉热，那么在 22℃的房间里，最近邻算法猜测她会感到温暖，因为 22℃比 10℃更接近 25℃。

前面的例子可以知道 Mary 什么时候感觉到热或冷，但当 Mary 被问及是否感到热或冷时，风速也是一个影响因素，如表 1-1 所示。