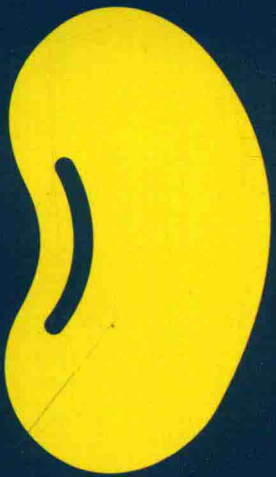
图灵程序设计丛书 TURING

白话机器学习算法



[新加坡] 黄莉婷 苏川集 / 著
武传海 / 译

斯坦福大学大数据基础课程教材
文科生也看得懂的算法及数据科学入门书
涵盖回归分析、神经网络、决策树、A/B测试等重要主题

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

TURING 图灵程序设计丛书

Numsense! Data Science for the Layman: No Math Added

白话机器学习算法

[新加坡] 黄莉婷　苏川集　著
武传海　译

人民邮电出版社
北　京

图书在版编目（CIP）数据

白话机器学习算法 /（新加坡）黄莉婷，（新加坡）苏川集著 ；武传海译. -- 北京 ：人民邮电出版社，2019.2
（图灵程序设计丛书）
ISBN 978-7-115-50664-1

Ⅰ. ①白… Ⅱ. ①黄… ②苏… ③武… Ⅲ. ①机器学习－算法 Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2019)第020041号

内 容 提 要

与使用数学语言或计算机编程语言讲解算法的书不同，本书另辟蹊径，用通俗易懂的人类语言以及大量有趣的示例和插图讲解 10 多种前沿的机器学习算法。内容涵盖 k 均值聚类、主成分分析、关联规则、社会网络分析等无监督学习算法，以及回归分析、k 最近邻、支持向量机、决策树、随机森林、神经网络等监督学习算法，并概述强化学习算法的思想。

任何对机器学习和数据科学怀有好奇心的人都可以通过本书构建知识体系。

◆ 著 [新加坡] 黄莉婷 苏川集
译 武传海
责任编辑 谢婷婷
责任印制 周昇亮

◆ 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
北京瑞禾彩色印刷有限公司印刷

◆ 开本：880×1230 1/32
印张：4
字数：139千字 2019年 2 月第 1 版
印数：1－3 500册 2019年 2 月北京第 1 次印刷
著作权合同登记号 图字：01-2018-3264号

定价：49.00元

读者服务热线：(010)51095186转600 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

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

版权声明

序

如今，大数据已经成为一大产业。随着数据逐渐主导我们的生活，“炼数成金”几乎成为每个机构都关注的焦点，各种模式识别和预测技术也成为提升业务能力的新手段。比如，商品推荐系统对消费者和商家都有好处，它会提醒消费者关注自己可能感兴趣的商品，同时也会帮助商家赚取更多的利润。

然而，大数据并非数据科学的全貌。数据科学是分析和利用数据的一门综合性学科，其范围涵盖机器学习、统计学和相关的数学分支。其中，机器学习占据首要位置，它是驱动模式识别和预测技术的主动力。机器学习算法是数据科学的力量之源，它和数据一起产生极其宝贵的知识，并且帮助我们以新的方式利用已有信息。

对于外行而言，要想理解数据科学如何推动当前的数据革命，就需要对这个领域有更好的认识。尽管现在对数据素养的需求很大，但是由于担心缺乏相关技能，一些人对数据科学领域敬而远之。

这正是莉婷和川集写作本书的缘由所在。我对两位作者的写作风格较为熟悉；在拜读本书之后，我发现这的确是专为外行写的数据科学书，两位作者特意省略了复杂的数学内容，从较高的层次讲解相关概念。但请不要误会，这并不意味着本书没有实质内容；相反，“干货”还不少，并且简洁精练。

你可能会问：本书采用的讲解方法有什么好处呢？实际上好处多多，并且对于外行来说，这种方法比普通的方法更可取。假设你对汽车的工作原理颇感兴趣，但是一窍不通，那么相比阅读深奥的燃烧学内容，你可能更容易接受对汽车零部件的概括性介绍。了解数据科学也是

如此：如果你对这个领域颇感兴趣，那么在深入研究数学公式之前，先从宽泛的概念入手比较容易。

第 1 章通过短小的篇幅讲了数据科学的一些基本概念，让每一位想入门数据科学的读者都拥有相同的知识基础；接着阐述算法选择等常被入门类读物所忽略的重要概念，以此促使读者进一步了解数据科学领域，并为读者提供一个完整的学习框架。

两位作者本来可以在书中讲解各种数据科学概念，而且讲解方法也有很多。但是，他们特意把讲解重点放在了对数据科学极其重要的机器学习算法上，并辅以相应的任务场景，这真是明智之举。k 均值聚类、决策树、最近邻等算法得到了应有的重视。此外，两位作者还对高级的分类和集成算法（比如支持向量机，它常常因为复杂的数学问题而令人生畏）以及随机森林做了讲解。当然，书中还讲了神经网络，它是当前的深度学习热潮背后的驱动力。

本书的另一个优点是，每个算法的讲解都配有直观的示例，比如通过预测犯罪行为介绍随机森林，以及在分析影迷性格特征时讲聚类。这些示例都是作者精心挑选的，有助于理解相关算法。与此同时，讲解并没有涉及高等数学知识，这样做有利于保持你对数据科学的兴趣和学习动力。

如果你正打算学习数据科学或相关算法，并且正在寻求一个切入点，那么我强烈建议你阅读本书。在我看来，本书是无与伦比的数据科学入门读物。有了它，数学不再是数据科学之路上的拦路虎。

Matthew Mayo

数据科学家、KDnuggets 编辑

前　言

本书由分别毕业于英国剑桥大学和美国斯坦福大学的数据科学爱好者黄莉婷和苏川集为你呈现。

我们发现，虽然数据科学被越来越多地用来改善决策，但是很多人对它知之甚少。鉴于此，我们把一些教程汇编成书，以便更多人学习。不管你是心怀抱负的学生，还是商业精英或其他什么人，只要你对数据科学充满好奇，都可以通过本书学习。

每篇教程介绍一种数据科学技术，并讲解其重要功能和基本思想，但内容不会涉及数学。此外，我们还将结合现实世界中的数据和实例对这些技术做具体阐释。

本书得到了不少朋友的帮助，没有他们，本书就无法面世。

首先，我们要感谢 Sonya Chan，她是本书英文版的文字编辑，也是我们的好朋友。她巧妙地把我们两人的写作风格融合在一起，确保将我们各自讲解的内容衔接得天衣无缝。

其次，感谢 Dora Tan，她是一位才华横溢的平面设计师，本书英文版的排版设计和封面设计都出自她之手。

感谢我们的朋友 Michelle Poh、Dennis Chew 和 Mark Ho，他们提出了许多宝贵的建议，使本书读起来更容易理解。

还要感谢密歇根大学安娜堡分校的 Long Nguyen 教授，以及斯坦福大学的 Percy Liang 教授和 Michal Kosinski 博士。他们耐心地培养我们，并且无私地分享自己的专业建议。

最后，我们还要感谢彼此。尽管有时会争吵，但我们仍然是好朋友。我们一起并肩作战，直至实现最初目标。

电子书

扫描如下二维码，即可购买本书电子版。

为何需要数据科学

假设你是年轻的医生。有位患者来到你的诊所，跟你抱怨说自己呼吸困难、胸部疼痛，并偶尔伴有胃灼热。于是，你给他检查血压和心率，发现一切正常，并且他没有其他病史。

然后，你发现他偏胖。由于他说的症状在体重超标的人群中普遍存在，因此你安慰他说，“不用担心，没什么大问题”，并且建议他抽空多锻炼身体。

上述诊断常常是误诊。心脏病患者与肥胖症患者表现出的症状相似，医生经常忽视这一点，而没有为患者做进一步检查。如果进一步检查，就可能查出更严重的疾病。

人类的判断力有一定的局限性，有限、主观的经验和不完备的知识都会影响它。这会破坏决策过程，那些缺乏经验的医生很可能就此放弃对患者做进一步检查，从而无法得到更准确的诊断结论。

在这种情况下，数据科学就能派上大用场。

数据科学技术不依赖于个人的判断力，它使得我们可以利用来自多个数据源的信息做出更好的决策。例如，可以查看记录着类似症状的病历，从中发现先前那些被忽视的诊断结果。

借助现代计算机和高级算法，我们能够做到以下几点。

- 从大型数据集中发现隐藏的趋势。
- 充分利用发现的趋势做预测。
- 计算每种结果出现的概率。
- 快速获取准确结果。

本书是数据科学及其算法的入门书，在讲解时采用了通俗易懂的语言。（不谈数学！）为了帮助你理解主要概念，本书采用了直观的解释方式，并且配有大量的插图。

每种算法各自成章，并且配有应用实例来解释其原理。书中用到的数据都可以从互联网上获得[①]。

每一章的最后都有小结，便于你复习这一章学过的内容。本书最后附有各种算法优缺点的比较，以及常用术语表，供你参考学习。

我们希望本书能够让你真正了解数据科学，并且帮助你正确地运用数据科学做出更好的决策。

让我们一道踏上数据科学之旅吧！

① 关于如何获得数据集，请访问图灵社区并点击页面右侧的“随书下载”：http://www.ituring.com.cn/book/2618。——编者注

目　录

第1章
基础知识

要想完全搞明白数据科学算法，必须先从基础知识学起。本章主要介绍数据科学的基础知识，它是本书最长的一章，篇幅大概是后续各章（讲解各种具体算法）的两倍。通过学习本章，你将对绝大多数数据科学研究涉及的基本步骤有大致的了解。这些基本步骤会帮助你评估上下文以及约束条件，并选出适合在研究中使用的算法。

数据科学研究有 4 个主要步骤。首先，必须处理和准备待分析的数据。其次，根据研究需求挑选合适的算法。再次，对算法的参数进行调优，以便优化结果。最后，创建模型，并比较各个模型，从中选出最好的一个。

1.1 准备数据

数据科学就是关于数据的科学。如果数据的质量差，那么分析得再精确也只能得到平淡无奇的结果。本节将介绍数据分析中常用的数据格式，还会涉及一些用来改进结果的数据处理方法。

1.1.1 数据格式

在数据分析中，表格是最常用的数据表示形式，如表 1-1 所示。表格中的每一行就是一个数据点，代表一个观测结果；每一列是一个变量，用来描述数据点。变量也叫属性、特征或维度。