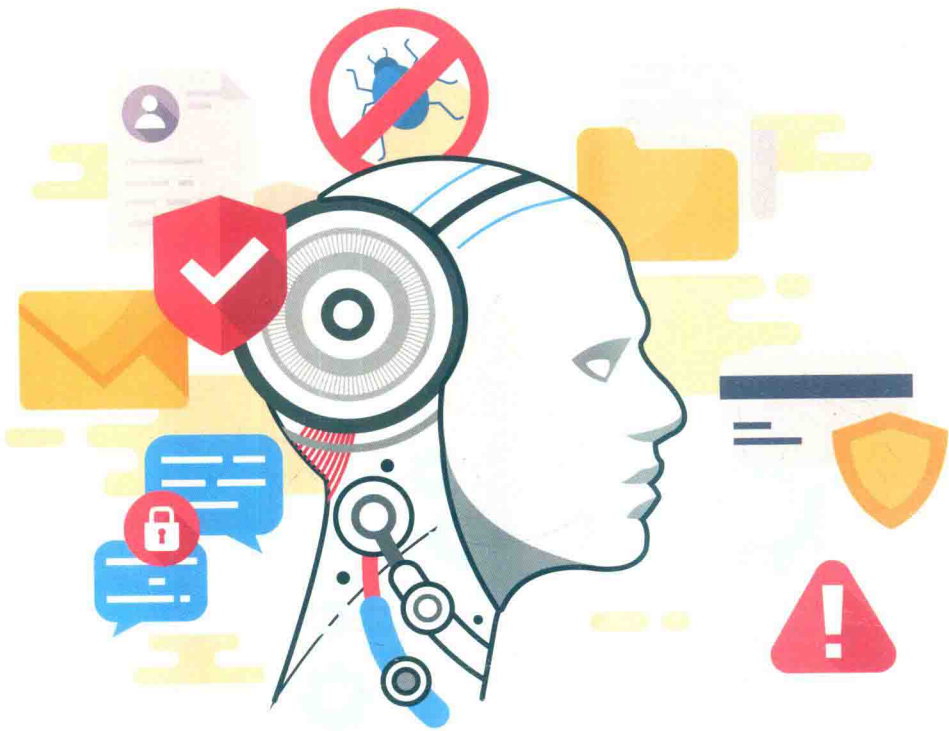
Broadview®
www.broadview.com.cn

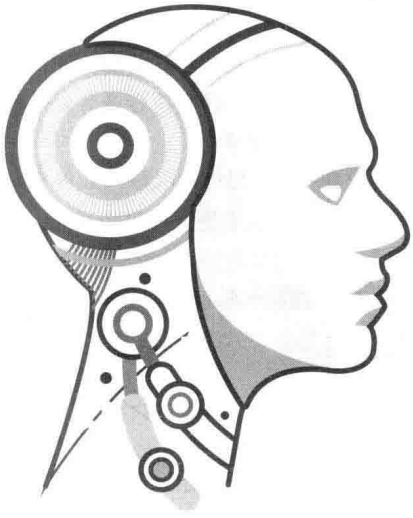


机器学习

互联网业务安全实践

王 帅 吴哲夫 著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn



机器学习

互联网业务安全实践

王 帅 吴哲夫 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

互联网产业正在从 IT 时代迈入 DT 时代（数据时代），同时互联网产业的繁荣也催生了黑灰产这样的群体。那么，在数据时代应该如何应对互联网业务安全威胁？机器学习技术在互联网业务安全领域的应用正是答案。

本书首先从机器学习技术的原理入手，自成体系地介绍了机器学习的基础知识，从数学的角度揭示了算法模型背后的基本原理；然后介绍了互联网业务安全所涉及的重要业务场景，以及机器学习技术在这些场景中的应用实践；最后介绍了如何应用互联网技术栈来建设业务安全技术架构。作者根据多年的一线互联网公司从业经验给出了很多独到的见解，供读者参考。

本书既适合机器学习从业者作为入门参考书，也适合互联网业务安全从业者学习黑灰产对抗手段，帮助他们做到知己知彼，了解如何应用机器学习技术来提高与黑灰产对抗的能力。

图书在版编目（CIP）数据

机器学习互联网业务安全实践 / 王帅，吴哲夫著. —北京：电子工业出版社，2019.9
ISBN 978-7-121-35568-4

Ⅰ. ①机… Ⅱ. ①王… ②吴… Ⅲ. ①机器学习－应用－互联网络－网络安全－研究 Ⅳ. ①TP393.08

中国版本图书馆 CIP 数据核字（2018）第 260304 号

责任编辑：张春雨
文字编辑：许　艳
印　　刷：三河市良远印务有限公司
装　　订：三河市良远印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编：100036
开　　本：787×980　1/16　印张：32　字数：553 千字
版　　次：2019 年 9 月第 1 版
印　　次：2019 年 9 月第 1 次印刷
定　　价：128.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

推荐语

对于安全行业来说，业务安全正随着业务形态的复杂化而变得越来越具有挑战性，本书从概念到实例都有比较详细的讲解，能够帮助阅读者更好地思考和学习，提供业务安全需求相关的更多技术选择。

——张作裕

阿里巴巴钉钉 CRO

平台型互联网公司都面临着垃圾注册、刷单、“薅羊毛”、信息泄露等业务安全方面的威胁，与黑灰产的对抗需要构建一套有效的业务安全模型体系，而对这个垂直领域，业内的关注度较低。本书作者结合自己多年的实践经验，从技术角度讲解了构建这套模型体系所涉及的常用算法和工具，适合从事业务安全算法领域的初学者学习，也适合中高阶的从业者参考。

——陈朝钢

资深风控架构师

机器学习是多学科交叉的领域，有极广泛的应用。作为一名互联网行业的从业者，很高兴从本书中看到了知识与正义的共鸣：从安全的视角探索机器学习的应用，以高端的技术构筑互联网业务的防护盾。本书作者对业务安全的理解深刻，从数学基础、模型算法、系统应用方面对机器学习知识进行了梳理，值得点赞和学习。

——陈景东

蚂蚁金服高级技术专家

机器学习、安全都是目前互联网领域的热门研究方向，两者的交叉更是最近的研究热点之一。这本书深入浅出，从机器学习的基础理论、模型出发，一步步揭示如何将机器学习应用在业务安全中，书中既有理论的讲解也有经验的总结，非常值得机器学习和业务安全的开发人员学习和借鉴。

——邓钦华

网易云音乐算法智能部负责人

在业务安全形势日益严峻的今天，如何利用机器学习扩大风险的识别范围，提高风险识别的准确度，提升业务安全的自动化水平，是业务安全从业人员高度关注的问题。本书作者对业务安全中常用的机器学习算法和模型进行了深入的讲解，并通过反欺诈、反爬虫、账户安全、内容安全、信贷安全等实际案例指引读者在业务安全工作中选择和应用合适的算法和模型。作者将他们的丰富实战经验完整教授给读者，不仅授人以鱼，更授人以渔。

阅读本书，做业务安全工作的同学可以快速为自己的业务选择合适的机器学习方案；设计算法的同学可以了解机器学习在业务安全应用中的独特之处，激发灵感，对机器学习在业务安全中的应用进行更深入的研究。

市面上鲜有图书既深入讲解机器学习算法和模型，又毫无保留地分享在业务安全实践中应用机器学习的经验，本书尤其值得业务安全的初学者深入研读。

——许瑞

唯品会业务安全负责人

通常我认为有两类（机器学习）算法书比较优秀：一类是书中的知识体系是自洽的，读者不需要同时查阅其他资料就可以学习；另一类是将知识和实践有机结合，不会让读者有学习屠龙术之感，能很快上手实践。非常难得的是，这两个特征都体现在本书中，因而此书是进入智能网络安全领域的一本非常棒的入门书籍。知易行难，让我们像作者一样在机器学习这条路上漫漫求索吧。

——张金

阿里巴巴搜索事业部

高级算法专家

推荐序一

2016 年 3 月，AlphaGo 战胜李世石，人工智能一下子又成为被广泛讨论的热门话题。这两年人工智能发展得非常快，深度学习为语音、文本和图像处理带来了很多突破。机器学习在各种业务场景中有很重要的应用价值。市面上介绍机器学习的书有不少，介绍互联网业务安全的书也有一些，但是介绍机器学习在互联网业务安全领域实践的书，并不多。

本书开篇概述了互联网业务安全的内涵，接着就进入正题介绍机器学习的内容，从机器学习的入门知识到模型再到具体工程的实施，让缺少相关经验的读者能够比较容易地顺着这个思路了解对应内容。后面的章节更多的是围绕具体业务安全工作而展开的，内容十分丰富。更重要的是，这些都是两位作者在实际业务场景中的实战经验的总结。从我个人的角度看，这些在业务场景中积累的经验更加宝贵，就好比是真的上了战场而且打了胜仗的高手所分享的经验，弥足珍贵。

希望本书能够给读者带来更多的帮助。

曾宪杰

蘑菇街副总裁

推荐序二

当我突然收到王帅同学的《机器学习互联网业务安全实践》初稿时，既感到惊讶也感到佩服。惊讶的是，在我印象中还是个毛头小伙子的他，已经能拿着自己的著作出现在我面前。佩服的是，写书毕竟是一件非常繁杂辛苦的事情，王帅同学虽然研究生毕业才 5 年，但却有勇气也乐意在繁忙的工作之余，花功夫将理论知识和自己的实践经验总结成书，造福读者。

回想起来，王帅读研期间一直都是一位发展全面、表现优秀的学生。学习成绩好，自不必说，他是 2010 年从哈尔滨工业大学保送到华中科技大学图像所（2013 年已与自控系合并为自动化学院）读研的，读研期间课程成绩名列前茅。最突出的是他的科研动手能力，那段时间我们刚好承担了一项国家工程的关键技术攻关任务，由于问题的特殊性，几乎没有可参考借鉴的资料，而且时间紧、任务重，王帅作为主力承担了其中的两项研究工作，均圆满完成任务，得到单位的好评。另一方面，王帅还是当时图像所研究生会的主席，积极为同学服务，把所里的学生工作开展得有声有色，除了组织日常的学术交流活动，文体活动也举办得丰富多彩，拿了学校不少的奖，很有影响力。

这几年，人工智能在媒体的高度关注下热度爆棚，技术发展极为迅速，新思想、新方法、新算法层出不穷，应用领域也在不断扩大。如果仅靠在学校学习的知识，显然是不能适应这个领域日新月异的发展的，每一个技术人员在工作中都必须有很强的自学能力，不断提高自身素质，才能跟上技术发展的步伐。显然，王帅做到了不断学习、不断进步。他能写这本书就是最好的证明。

这本书的意义不仅仅在于王帅同学对自己的前期工作做了很好的总结，更重要的是，业务安全是一个充满激烈对抗的领域，如何应对黑灰产对互联网平台的攻击是每一位相关技术从业者都需要思考的问题，本书对于那些刚入职场、刚进入业务

安全领域的新人来说，具有很强的指导意义，能让他们很快将书本知识和实际应用相连接，尽快达到工作要求。当然，这本书将机器学习理论与业务安全相结合，也能让这个领域的从业者受到启发，具有“抛砖引玉”的作用。

最后，希望王帅同学戒骄戒躁，继续努力，为机器学习在业务安全领域中的应用做出更多贡献。

曹治国

教授，华中科技大学自动化学院院长

序言一

写下本文的此刻，我正坐在从杭州前往北京的 G40 次列车上，准备参加第二天在北京理工大学举办的 MLA 2017 会议。北京是我开始参加工作的地方，也是我第一次实习的地方，对于北京，我是很有感情的。而对于杭州，则怀着难以名状的情愫，从古至今，无数文人墨客在此留下印记，其中李叔同先生的“未能抛得杭州去，一半勾留是此湖”给我的印象最为深刻。所以 2015 年春节后，我毅然从百度离开加入蘑菇街（现在的美丽联合集团，简称美联），在反作弊团队工作。工作的方向也从搜索算法策略转到了业务安全算法策略。我们的团队从最初仅有反作弊相关算法策略，到现在机器学习算法能够服务于主要的业务安全场景，算法技术的迭代与优化历经了近 3 年的时间。虽然与 BATJ 等巨头相比，我们的体量还有较大的差距，但是“麻雀虽小，五脏俱全”，当前我们的业务安全算法策略体系基本涵盖了统计机器学习方法、深度学习方法和复杂网络的相关算法。

在 2018 年 51CTO 组织的 WOT 峰会[1]和唯品会组织的城市沙龙上海站[2]中，我们的团队都分享了在美联业务安全场景中使用机器学习方法的一些心得体会和实践经验，收到了较好的反响。在会议期间，我们和同行们针对当前所面临的问题做了深入的交流。而我个人也在 CSDN 的博客上发表文章，剖析和分享生产环境中涉及的一些算法原理知识。正是因为这些文章，电子工业出版社的张春雨先生辗转找到我，希望我能写一本关于如何在业务安全中应用机器学习的书籍。说实话，一开始我是非常“紧张”的，一是考虑到业务安全的范围实在太大，自己平时接触的工作还是有一定的局限的；二是机器学习这个领域内的经典图书很多，李航博士的《统计学习方法》和周志华老师的《机器学习》（俗称“西瓜书”）都广受好评，我来写书岂

1 http://wot.51cto.com/act/2017/innovation/page/agent

2 https://mp.weixin.qq.com/s/7t5zMuAscs_I8f1poMrJVA

不是班门弄斧？而与张春雨先生深入沟通后，我逐渐打消了顾虑，也明确了本书的定位。

幸运的是，我们团队内新加入的盖世（花名）同学对于此事非常感兴趣，再加上其个人在机器学习领域也积累了不少经验，所以我们一拍即合，欣然接受了张春雨先生的邀请，决定为互联网业务安全中的机器学习技术做一点小小的贡献。

本书旨在为工程技术人员提供一份在业务安全中实践机器学习技术的入门指南，内容包括业务的背景、机器学习算法的原理、算法的实现与优化，以及在生产环境中算法的上线与迭代方法。如果我们踩过的“坑”和积淀的经验能够为相关从业者带来一些启发，我们就心满意足了。

此时列车刚开过济南西站，窗外已经是茫茫黑夜，正如黑灰产和“羊毛党”们所处的隐蔽之处。与这些不法分子对抗是业务安全从业者的职责，而机器学习技术也许就是划破这黑暗的一束光，为我们赢得胜利带来可能。希望此书可以让这束光愈加明亮。

王帅

序言二

作为一个科班出身的计算机从业人员，深知在机器学习领域摸爬滚打的不易。在山东大学学习期间，我学的是软件工程，对编程有浓厚的兴趣，陈竹敏老师认可我的才能，并让我参加与美国得克萨斯州大学的合作项目，还推荐我继续读研究生。在读研的两年期间，北大的杨雅辉老师对我的学习给予了极多的指导。后来，我又跟随微软亚洲研究院的袁进辉老师学习，收获良多，从一个动手能力极弱的“小白”成长为能熟练编写代码的机器学习工程师。现在从业三年，也指导了许多学弟、学妹进入职场，希望自己也能像我的老师们一样无私地传授知识。

回想自己学习机器学习的经历，感慨良多。本科毕业时，尽管已经学习了《微积分》《线性代数》《离散数学》《数理统计》《计算机组成原理》《编译原理》《操作系统》《算法导论》《运筹学》等教材，我却并没有见到这些本应有极高价值的书本知识在实际工作中发挥多大的作用，因此十分迷茫。当时陈竹敏老师推荐我继续深造，从此折节读书，半载后来到梦寐以求的学府——北京大学。感谢我的室友，他们的专业（自然语言处理和机器学习）对我产生了极大的影响，也终于看到了自己投入时间学习的课程知识能够发挥的价值。为了不至于落后周围人太多，我深居简出，自学了《数值分析》《测度论》《代数》《统计学》《贝叶斯统计》《图论》《矩阵论》《凸优化》等教材，并且了解与学习了衍生的应用学科知识，研读《机器学习》《密码学》《应用回归分析》《组合数学》等书籍。工作以后，虽然有很多想要深入学习的细分理论知识，买了《实变函数与泛函分析》《博弈论》《拓扑学》等图书，但是一直苦于没有足够的时间，这些书籍已经在书架上落灰了。

上面罗列了一些教材，其实是想给在校的学生朋友学习机器学习提供一个书单。当然，纸上学来终觉浅，绝知此事要躬行，任何理论知识只有在实际场景中应用或实验，才能加深理解。

作为一个机器学习领域的新人，我也在不断认真学习机器学习的理论，希望能够在工作中充分应用所学知识。我曾在传统行业工作，后来进入大数据领域，在电商行业摸爬滚打。我觉得人应该脚踏实地，无论身处何种行业，都应该在一个专业领域深入地学习。现在，我已经是一个父亲，肩上的责任越来越重，但是我十分感恩。感谢家人，让我学会了真诚待人，享受生活中的一切美好。

吴哲夫

前言

机器学习学科的发展大体经历了规则学习、统计学习、深度学习这三大阶段。从最早的结构化的人机赛棋，到广泛领域的知识问答，再到当下红极一时的自动驾驶等工业领域，机器学习已经被成功应用到模式识别、数据挖掘、自然语言处理、人工智能、语音识别、图像识别等各个领域，并且被综合应用到信息检索、生物信息技术、自动驾驶、无人机、AR/VR、医疗、教育等各个行业。

机器学习的很多方法在原理上是相通的，只不过适用的领域不同。机器学习的能力比较强大，应用范围广泛，要解决的问题多且繁杂，因此并不存在一个适用于所有问题的结构化方法。这就要求机器学习工程师具备较高的素质，除了掌握计算机科学基础的三个方面的知识（系统、软件、理论），还要对机器学习算法有深入的了解，只有这样才可以搭建出一个适用于工业界应用的好框架。

基础决定深度。一般来讲，机器学习由**模型**（建模）、**策略**（学习方法）、**算法**（实现）三部分组成，叫作机器学习三要素。这三部分层层递进，推理的难度逐渐增加，对人的要求也不一样：在建模过程中需要有理解能力，在设计学习方法时需要有数学推理能力，最终将学习方法实现为算法时需要有转化能力。当然，一些资深的程序员或者 ACM（Association for Computing Machinery）竞赛的参赛者，本身有非常强的代码理解能力，这些能力能帮助他们理解算法，并进一步理解机器学习的过程。

要想成为一名优秀的机器学习工程师，必须有良好的数学基础。在本科阶段学习的数值分析、线性代数、概率与统计、离散数学等课程知识，对于理解大多数模型来说已经足够了。概率与统计及离散数学是理解模型的基础，线性代数决定了你实现算法的能力，数值分析决定了推衍过程。然而，如果希望更深入地理解模型算法的实现原理、掌握和学习更多的模型，还需要学习矩阵论、优化论、泛函分析、

贝叶斯统计、模糊数学等方面的知识。

在工业界，很多时候大家只是使用模型，最低的要求就是理解模型的输入、中间过程和输出。要快速掌握并使用模型，关键在于理解模型的适用条件，这样才能构造出符合模型要求的特征。客观地讲，并没有不好的模型，只有没有构造好的特征。这也牵扯到模型的适用性问题，有些任务可能非常难以转化成模型最适用的问题，因为有时候如果强制使用某种模型，可能需要对于任务与特征本身有深刻的理解，以及长时间的浸淫。

很多时候，我们并不一定要选择最合适的模型，究其原因：一是我们所选择的模型可能并不需要特别复杂的转化就可以用于此种问题；二是机器学习工程师不一定有时间对某个行业进行深入的分析和研究，提取出适用于模型的各种特征；三是强制使用某种模型可能会导致转化问题本身就是一个复杂的问题，需要对结构进行大量修改以及在工程上提供支持。由于业界竞争激烈，有时候我们需要的是快速迭代，因此这时更关键的是选择一个基本适用的模型，先验证得到问题的可解性和baseline，然后再不断优化。

本书适合那些从其他编程领域转入机器学习领域的工程师阅读，帮助他们快速掌握模型及其应用。本书假设读者已掌握微积分、概率论、线性代数以及离散数学的基础知识。书中简单介绍了机器学习的基本概念及其背后的数学原理，以机器学习在业务安全领域的应用为线，详细讲解如何将机器学习应用到业务安全工作中，对一些模型的策略和算法进行了深入介绍。

本书第 2 章、第 3 章中的一些数学基础知识（定义、原理等），引用自国内外名校采用的本科与研究生教材，笔者按照自己学习机器学习的路线对这些知识进行了编排，并统一了数学符号，方便大家快速了解或查询。

限于篇幅，本书只列出必需的数学基础知识，仅对某些定理给出了证明，并加入笔者的解释，帮助大家理解。机器学习是一门与数学联系十分紧密的学科，因此笔者更愿意用符号、公式和算法语言来介绍相应的内容，希望大家能通过定义理解函数，通过算法语言理解算法本身，培养看公式比看文字更高效的能力。希望大家能够理解算法的原理，了解如何恰当地将机器学习应用到实际场景中，既抛出问题，

又给出笔者积累的解决问题的思路。最后还要强调，数学是基础，数学概念字字珠玑，请大家认真理解，在此基础上你甚至能创造属于自己的算法。

鉴于写作时间仓促以及篇幅有限，书中有些地方的讲解可能比较晦涩或者不够全面，尽管笔者竭尽所能，有些疏漏也在所难免，希望大家能够在发现问题后第一时间联系笔者，笔者会在再版时更正，在此先表示感谢。下图所示为机器学习算法工程师需要具备的技能树。

目录