HZ BOOKS
华章IT

智能系统与技术丛书

百度安全专家撰写，二十多位业界专家联袂推荐

Machine Learning for the Web Security

Web安全之机器学习入门

刘焱 编著

机械工业出版社
China Machine Press

智能系统与技术丛书

Machine Learning for the Web Security

Web安全之机器学习入门

刘焱 编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Web 安全之机器学习入门 / 刘焱编著 . —北京：机械工业出版社，2017.7
（智能系统与技术丛书）

ISBN 978-7-111-57642-6

I. W… II. 刘… III. 机器学习 IV. TP181

中国版本图书馆 CIP 数据核字（2017）第 179868 号

Web 安全之机器学习入门

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：吴 怡 责任校对：李秋荣
印 刷：北京诚信伟业印刷有限公司 版 次：2017 年 8 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：16.25
书 号：ISBN 978-7-111-57642-6 定 价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

内容简介

在现今的互联网公司中，产品线绵延复杂，安全防御体系无时无刻不在应对新的挑战。哪怕是拥有丰富工作经验的安全从业者，在面对层出不穷的攻击手段和海量的日志数据时也会望洋兴叹。机器学习是这些问题天然契合的解决方案，在数据量以指数级不断增长的未来，甚至有可能是唯一的出路。

本书首先介绍主流的机器学习工具，以及Python应用于机器学习的优势，并介绍Scikit-Learn环境搭建、TensorFlow环境搭建。接着介绍机器学习的基本概念和Web安全基础知识。然后深入讲解多个机器学习算法在Web安全领域的实际应用，如K近邻、决策树、朴素贝叶斯、逻辑回归、支持向量机、K-Means算法、FP-growth、隐式马尔可夫、有向图、神经网络等，最后介绍深度学习算法CNN、RNN。本书针对每一个算法都给出了具体案例，理论结合实际，深入浅出，讲解清晰，适合有信息安全基础知识的网络开发与运维技术人员参考。

作者简介

刘 焱 百度安全Web防护产品线负责人，负责百度安全的Web安全产品，包括防DDoS、Web应用防火墙、Web威胁感知、服务器安全以及安全数据分析等，具有近十年云安全及企业安全从业经历，全程参与了百度企业安全建设。研究兴趣包括机器学习、Web安全、僵尸网络、威胁情报等。他是FreeBuf专栏作家、i春秋知名讲师，多次在OWASP 、电子学会年会等发表演讲，参与编写了《大数据安全标准白皮书》。他还建立了微信公众号“兜哥带你学安全”，分享了大量信息安全技术知识。

封面设计 锡 彬

Praise 对本书的赞誉

此亦笃信之年，此亦大惑之年。此亦多丽之阳春，此亦绝念之穷冬。人或万事俱备，人或一事无成。我辈其青云直上，我辈其黄泉永坠。——《双城记》狄更斯著，魏易译

如今是一个人工智能兴起的年代，也是一个黑产猖獗的年代；是一个机器学习算法百花齐放的年代，也是一个隐私泄露、恶意代码传播、网络攻击肆虐的年代。AlphaGo 碾压柯洁之后，不少人担心 AI 会抢了人类的工作，然而信息安全领域专业人才严重匮乏，极其需要 AI 来补充专业缺口。

兜哥的这本书展示了丰富多彩的机器学习算法在错综复杂的 Web 安全中的应用，是一本非常及时的人工智能在信息安全领域的入门读物。正如书中所述，没有最好的算法，只有最合适的算法。虽然这几年深度学习呼声很高，但各种机器学习算法依然在形形色色的应用场景中有着各自独特的价值，熟悉并用好这些算法在安全领域的实战中会起到重要的作用。

——Lenx，百度首席安全科学家，安全实验室负责人

存储和计算能力的爆发式增长，让我们获得了比以往更全面、实时地获取以及分析数据的潜在能力，但面对产生的海量信息如何快速准确地转化为业务需求则需要依赖一些非传统的手段。就安全领域来说，原先依赖于规则的问题解法过于受限于编写规则的安全专家自身知识领域的广度和深度，以及对于问题本质的理解能力。但我们都知道，安全漏洞层出不穷，攻击利用的方式多种多样，仅仅依赖于规则进行问题的发现在现阶段的威胁形势下慢慢变得捉襟见肘。面对威胁，企业安全人员需要打造这样一种能力，它能够让我们脱离单纯的点对点的竞争，case by case 的对抗，转而从更高的维度上来审视业务，发现潜在的异常事件。这些异常事件可能会作为安全人员深入调查的起点，让我们具备找到原有安全能力盲区以及发现新威胁的能力，使我们的技能水平以及对威胁的响应速度能持续提升。同时这种能力和防御体系结合，也有可能让我们达到在面对某些未知威胁时，以不变应万变，获得天然免疫的理想状态。兜哥的这本书或许是开启我们这种能力的一把钥匙。本书通过介绍通俗易懂的机

器学习原理，结合实际企业中的安全业务需求场景，让广大安全人员能够感受到这种“如日中天”的技术在传统安全领域内如何大放异彩。最后，May the force be with you。

——王宇，蚂蚁金服安全总监

百度是拥有海量互联网数据的几家公司之一，兜哥是百度前 IT 安全负责人，现 Web 安全产品负责人，研发的产品不仅应用于百度公司内部检测网络攻击，也应用在多个百度的商业安全产品中，服务于数万站长。兜哥的团队是国内最早一批将机器学习算法应用于网络安全场景的团队之一，本书聚集了兜哥及其团队多年的安全实践经验，覆盖了互联网公司可能会遇到的多个安全场景，比如用图算法检测 WebShell 等，非常好地解决了百度商业安全客户被入侵留后门的问题。兜哥将自己的技术选型、算法、代码倾囊相授，我相信本书的出版将会大大降低安全研发工程师转型安全数据分析专家的难度，值得推荐。

——黄正，百度安全实验室 X-Team 负责人，MSRC 2016 中国区第一

在大数据时代，犯罪分子作案的手段越来越高明，手动分析的成本越来越高，效率也越来越低；与此同时，人工智能技术越来越成熟，安全与人工智能技术相结合，才能适应新的环境，推荐安全从业者学习这本书。

——桑文峰，神策数据创始人 &CEO

网络世界的攻击与防护对抗发展到今天，各种技术已经日趋专业和精细，通过古老的 string-match 的防御方式越来越不能适应新的攻击环境，对于想尝试着把机器学习和安全相结合的从业人员来说，阅读本书是个很好的开端。

——赵林林，微步在线技术合伙人，前美团、高德安全负责人

就我有限的了解，在很长一段时间里，安全技术和机器学习技术都是分别演进的。安全问题几乎会伴随着任何新生事物而来，并与之同生长，这也使得安全研究人员往往会把注意力局限于关注事物的个体特征而忽略其群体特征；而有的时候，即使有意于群体特征的研究，也可能会囿于工具和方法以致难于寸进，这对安全问题的解决形成了事实上的约束。机器学习作为一种可以从另一个维度来解决问题的技术，则对此约束进行了相当程度的突破。用新工具去解决老问题，这要求对这两者都有比较深入的了解（例如 AlphaGo），基于了解而进行实践，基于实践而予以总结分享，这样的知识分享和传递，正是刘焱这本书的价值所在。

——张宇平，数盟 CTO

在安全分析中要想用好机器学习，需要精通算法、懂得攻防、理解数据，三方面的知识缺一不可。这样的人固然难找，兜哥却恰好是其中的一员。兜哥凭借在一线互联网企业多年的安全实战经验给读者奉上了这本满满都是干货的大作。书中没有烦琐的公式推演，一切用

代码说话，特别适合了解算法原理、不知道如何在实际中应用的人学习。我个人读后深受启发，也推荐给身边每一个做安全数据分析的同行。

——周涛，启明星辰“鸿雁”计划首席研究员

在大数据时代潮流中，如何将大数据思想应用于网络安全技术是一个非常重要的主题。本书将 Web 安全与机器学习相结合，提出以数据驱动为基础，利用海量的数据资源分析 Web 恶意攻击，以通俗易懂的语言讲述了机器学习在 Web 安全领域多个方面的应用。对 Web 安全以及大数据安全感兴趣的人，这本书是一个很好的选择。

——李琦，清华大学副研究员

伴随着互联网的爆炸式发展，网络安全已上升到国家层面，按效果说话的安全能力建设得到高度重视。与此同时，安全团队却又不得不面对百花齐放的业务场景、大规模的数据中心，以及愈加剧烈、复杂和不确定性的网络攻击。如何在传统攻防对抗之外，寻找更有效、可落地的对抗方式，已成为各大企业安全团队思考的重点。所幸，近些年来，计算和存储资源已不是安全团队的瓶颈，安全团队自身在工程能力上也已非昔日吴下阿蒙。机器学习成为近些年来安全领域里第一批从学术走向工业的应用方向，并已有很多阶段性的实践成果。很欣喜地看到兜哥一直在推进机器学习系列的文章并编写了此书。此书重点讲解了常见机器学习算法在不同场景下的潜在应用和实践，非常适合初学者入门。希望此书能够启发更多的同行继续实践和深耕机器学习应用这个方向，并给安全行业带来更多的反馈和讨论。

——程岩，京东安全首席架构师

人工智能的技术发展正在不断加速，是时候探讨如何将机器学习应用于安全领域了。人工智能真的能在未来对抗网络攻击、自主保护我们的系统吗？这本书打开了一道门。这不仅是一部机器学习的科普书，一部机器学习在安全场景下的实战书，更是一部鼓励技术创新应用的行动指南。

——谢忱，FreeBuf 联合创始人，斗象科技 COO

安全正在发生变革，已经从专家模式逐渐演变成系统化、平台化，而随着机器学习和大数据技术的发展，未来安全将逐渐智能化。而这种变化已经得到了验证，在业务安全领域风控系统的基础数据中，如 IP 和用户画像、设备指纹生成和识别、规则的挖掘都使用大量的机器学习算法；在网络安全领域，如何入侵检测系统发现未知的攻击、如何解决无效的攻击行为也采用了大量的分类和关联规则算法。兜哥作为互联网企业的安全界资深研究者，一直对新技术的运用进行探索，这本书将为你打开安全智能化的大门。

——吴圣，58 同城高级架构师

机器学习一直是高大上的领域，作者结合自己的实际工作和研究，把机器学习在安全领域的应用讲得深入浅出、很接地气，稍具基础知识的读者就可动手体验应用机器学习的美妙感受。

——姚志武，借贷宝安全总监

纵观安全行业近十余年的攻击方式，从最早的单机小工具到如今分布式、大数据、自动化等攻击方式，防御的方式不得不随之不断升级，于是出现各种云安全产品，这些产品都能产生大量有价值的数据，但却少有产品能够真正利用这些数据实现联动防御，所以这些云都是单朵的小云。我们需要利用人工智能将这些数据进行联动，进行多维度、高精度的深入分析，还原攻击路径，才能真正实现态势感知，防御未知攻击。而人工智能的基础就是机器学习，让机器自适应、自分析、自决策，未来的安全防御必须具备这样的特性。本书采用实例的方式讲解机器学习在安全领域的应用，不仅能让读者了解到机器学习，还能让读者了解到攻击方式的检测手段，是一本难得的好书。

——尹毅，Sobug 技术合伙人，《代码审计：企业级 Web 代码安全架构》作者

在机器学习领域中，大多数的实用方向都表现在图像识别、广告推荐和个性画像等方面，但很少看见安全领域相关的机器学习方法和介绍，因为“安全”的概念是很模糊的，有的场景中，很少有合适的模型、适用的参数，甚至没有明确的算法。这本书介绍了基础的机器学习应用和方法，并结合部分特殊的场景对安全领域中较为常见和较为烦琐的分析提供了很好的例子和思维模型，不论是安全从业人员或者是机器学习领域的研究者，都值得阅读，可以从本书中获得很多好的启发和灵感。

——Kevin1986，搜狐资深安全研究员

不知道十余年前，在兜哥刚刚踏上安全这条“不归路”时，有没有想过如今的工作会面对多么错综复杂的环境，担负着数亿用户的信赖。在大型互联网公司中，产品线绵延复杂，每一个新产品的上线，每一次版本的迭代更新，都有不可预知的安全问题出现，安全防御体系也无时无刻不在应对新的挑战。哪怕是拥有丰富工作经验的安全从业者，在面对层出不穷的攻击手段和海量的日志数据时也会望洋兴叹。机器学习是这些问题天然契合的解决方案，在数据量以指数级不断增长的未来，甚至有可能是唯一的出路。机器学习如今可以说是如日中天的热词，但对于初学者来说可能并不是很容易就能掌握的技能。将学习到的内容应用在安全工作中更是难上加难。这或许也是机器学习经常出现在安全从业者视野中，却鲜有人愿意深入研究的原因之一吧。兜哥作为互联网安全领域内机器学习的先导践行者，可以说是从零开始，在不断尝试中摸索出了一条新的道路。毫无疑问，这是一个艰难而有价值的过程。

这本书作为走过这段历程之后的总结与分享，兜哥将多年的工作经验毫无保留地倾注于其中，以一个甲方安全人员的视角，将机器学习如何应用在 Web 安全工作的各个方面，用诙谐易懂的语言娓娓道来。在一气呵成地读罢兜哥亲手递过的样章之后，我只有一个想法：如今的 Web 安全领域，太需要这样一本佳作了。或许未来的某一天，机器学习或者说人工智能已经成为了保护互联网安全的中坚力量。回头一看，正是本书在路途的起点为我们指明了方向。

——幻泉，i 春秋教研中心总监

识别各类攻击一直是安全领域内难以解决的问题，由于语言的多样性，利用传统规则匹配来识别攻击已经过时，传统安全技术的发展也已经到了瓶颈，而本书提出机器学习结合 Web 安全的思路为安全技术发展指出了新的方向。相信读者阅读本书后能受益匪浅。

——西瓜，四叶草 CTO

安全监控的建立产生海量安全日志，人工查看审计日志已经无法解决企业实际安全需求。随着硬件成本降低，大数据技术成熟，机器学习在企业安全中的实践应运而生。本书详细介绍了如何通过机器学习分析海量安全日志，发现隐匿的攻击，本书是企业安全建设中不可多得的孙子兵法。

——廖威，易宝支付安全总监

早在 2009 在百度工作时，就因为工作交集认识了本书作者刘焱。期间经常讨论安全技术问题，为他的渊博知识与钻研精神所折服。近年来，Web 安全被越来越多的人所重视，攻防对抗上升到一个新的高度。各种新的攻击方法层出不穷，传统的检测与防御方式已不再适应，迫切需要更加智能的方法。随着机器学习的爆发式发展，两者的结合将是未来的趋势。在本书中，刘焱将枯燥复杂的算法、概念以简单易懂的图文结合方式呈现出来，并夹杂着他一贯的幽默风格，内容由浅入深、循序渐进。应用机器学习是未来的发展趋势，学习掌握它使创造出新一代的安全产品成为可能。希望大家喜欢这本书，并从中受益。

——刘袁君，医渡云安全总监

通过机器学习分析海量 Web 日志，进而发现业务异常和安全问题已经是安全监控平台的标配。然而，市场上信息安全和机器学习结合的工具书却很少，本书从基础知识和实际案例出发，逐步抽丝剥茧带你进入自动化安全的殿堂。书中的算法和思路是经过大规模部署和商业验证的，具备很强的可操作性。

——宋文宽，联想集团信息安全高级经理

序 — Preface

兜哥是网络安全行业的老兵，早在成为自媒体人之前，他所带领的团队在 Web 入侵检测、WebShell 识别等技术上就是百度安全防御的重要组成部分。他是一位十分难得的拥有敏感产品神经的技术人，在百度这些年，不仅将许多新产品、新技术引入百度，丰富百度防御能力，更通过自己的努力将百度在威胁检测等方面的经验不断传播出去。他通过自己的智能安全三部曲将他在人工智能方向的探索向业界做了系统性分享，在安全技术亟待突破的今天，有着很深的借鉴意义。本书是他的第一部著作，重点介绍如何在安全场景下进行机器学习。

回顾网络安全行业这十年来的发展，从防火墙、下一代防火墙、入侵检测到威胁情报，安全厂商一次次将新的概念引入，将新的技术包装，但安全威胁的现状却一天天恶化着，当我们看到越来越多的安全入侵事件发生，其波及范围也已经不仅仅是互联网业务，更有大量的国民基础设施深陷泥潭，我们不禁要问，是黑客越来越强大，还是我们的技术不够先进？我们被眼花缭乱的技术所困，却忽略了安全的本质——对抗。今天炙手可热的人工智能是否也会是一枚“银弹”？

诚然，人工智能在自然语言处理、图像识别、棋类对抗领域的成绩有目共睹，而安全能否成为下一个人工智能的突破口？现在看来，一切正方兴未艾，在有监督学习方向，能否大幅简化安全工程师的工作量，让准实时对抗成为可能；在无监督学习方向，能否突破安全对抗的猫鼠游戏，让安全由被动变为主动。等待我们的将是一场令人激动的技术探索。

冯景辉，安全宝联合创始人，百度商业安全总监

马杰，安全宝创始人，百度安全总经理

Preface 序　二

兜哥在通过数据分析进行安全检测的技术方面已经积累了很长时间，从最初我们合作建立国内最大 TB 级别日志分析系统开始，在这几年中，他一直在不断尝试使用更合适的技术来解决问题，这次欣喜地看到了他又有新的突破。

拿到样章当看到“通向智能安全的旅程”这一章时，着实被深深地吸引住了，在新技术中尝试使用机器学习的能力，借助 AI，能让系统变得更加聪明更好用，从而更好地解决问题。Gartner 在 2015 年就提出过“自适应安全架构来应对高级定向攻击”的概念，其中实现这套架构很重要的一个阶段就是让系统具备对攻击的预测能力，“预测”是一种更接近人的思考方法，通过机器学习及人工智能的技术迭代，这将有可能实现。

安全数据分析已经从搭建大数据分析系统过渡到使用机器学习的过程中了，通过机器学习算法对安全事件的分析在一段时间内也许并不能突出优势，就像我们面对一个天才少年一样，因为阅历原因暂时他不能超越你，但他一定会用非常短的时间就能追上并且更好地帮助你。

阅读过程中常常感叹于兜哥的细心和他对此系列书的撰写决心，兜哥是一位不折不扣的技术实践者，全书使用了超过 15 种机器学习的算法，收集整理了大量或知名、或在真实环境下出现过的案例，并一一详细给出了使用机器学习算法进行分析的方法。书中还包含了丰富的数据集以及大量的实例，能帮助入门的同学降低学习成本，快速进入技术实践中。阅读的过程中，我常常在假想，如果回到几年前看到这本书，现在我们搭建的分析系统又将是另一番景象。

本书的写作风格是实战型的，围绕常见的安全问题，通过代码导读的方式，把每个问题与机器学习算法相关联，循序渐进，揭开了机器学习的神秘面纱。对于立志从事信息安全技术的同学来说，这种实战型的案例更直观，更能激发学习兴趣，推动机器学习在安全分析上的应用。

云朋，百度无人车首席安全架构师

序　三 *Preface*

跟兜哥相识迄今一年有余，当时我还在一个跨境电商公司当码农头子，互金、电商也都还是资本圈炙手可热的概念，我们这个小而美的电商公司不能免俗，三天一小促，五天一大促地在玩着冲刺 GMV（日总交易金额）的游戏。玩命狂奔业绩的同时，我早早地就跟当时还身为独立安全公司的“安全宝”交了抗 DDoS 费，保证每次业务起起落落的时候，不会受到某一小撮别有用心的敌对势力的干扰。“安全宝”的服务接入不到半年就爆出新闻，百度全资收购了“安全宝”，推出了面向企业的百度安全服务体系。一直跟我对接的“安全宝”的销售朋友摇身一变，成为三巨头之一的金领员工。朋友高升遇喜自然要多多分享，于是某日就电话约了“百度资深安全工程师和销售团队”来我们这里做一个交流。产品介绍、业界八卦聊了半个多小时以后，一直安坐对方一角，眼睛闪着灵光的胖子始终没说话。我接受不了屋里仅有两个胖子，一个是我一直在聊；另一个胖子却如此沉默。于是我就开口问：“你们客户端的那个核心 xx 功能，就是这个角落里不说话的大黑客写的吧？”

“没有，没有，我们的 xx 功能不是那样的。”这哥们终于开口说了第一句话。

“不可能啊，因为 xx、xx、xx。”我又吧啦吧啦说了一通。

“嘿嘿，其实是 xx、xx。”他抬抬头，翻起眼睛看着我，一种内行跟内行言简意赅过招的感觉跃然而出。

“额，来，先留个微信吧。”我站起来，把手机递了过去。10 秒钟后，“中国相声界的小学生通过扫一扫添加你为好友”的消息弹了出来。“你太逗了。”我忍不住看着对方评论了一句，心想：这么有趣的码农朋友交定了。散会后，几个人站在办公室楼下，相声界的小学生朋友特别真诚地感谢了一下我提供如此好的机会，让他们有机会从中国互联网的物理核心交换地区后厂村来到事业线、大白腿比例明显高一个数量级的 CBD 地区。我则不失时机地指点了对方一下，应该步行走一段什么样的路线去地铁站，能更顺利地回到核心交换地区。这就是我跟兜哥的第一次见面。

接下来的一段时间，相声界的小学生朋友变成了我微信朋友圈中昵称更换频率最高的人，在目睹了“青青河边草”“小小铜豌豆”等花式变更之后，我知道蹭小学生朋友一顿大餐的机会成熟了，于是很愉快地约了一顿丰盛的晚饭。一向不胜酒力又闷骚的码农们碰到三观相近的同类，总是特别容易敞开心扉，觥筹交错间，关于奋斗、关于公司、关于互联网，当然，还有关于男男女女，让一次普通而平淡的饭局变得特别有记忆特质。尽管我的记忆力很难达到生活自理的标准，不过时至今日，还是经常想起与相声界的小学生朋友把酒言欢的许多细节，觉得有趣而温暖。

后来我们目睹了百度公司毅然启航进入人工智能的时代。其实对于搜索巨人百度公司，人工智能领域内常见的如最大熵、隐马尔科夫、卷积神经网络等数学模型，从第一天起就如血液一般，渗透进入分词、排序、分类、聚类等搜索业务的大部分领域，经过了十来年的高歌猛进，这些晦涩难懂的数学公式日益扩大了其应用范围，在安全领域也得到了越发深入的应用。

聪明、努力、专注是兜哥写作一本书的智力储备和保证。这个被摩尔定律不停推动、变革的时代，一本技术书籍本身的价值和生命周期总是有限的。然而，随着年纪渐长，我们越来越体会到，自己的时间消费中最有价值的部分，永远是与有趣的灵魂和思想共处的片段。人类天性讨厌无趣，毕生的使命都是在寻找与有趣共振的机会。一本精心写作的书籍中，饱含了作者倾注的时光和智慧，这些无形的精神宝藏是让我们手不释卷的核心吸引力。品一杯茶，我们的欢喜来自于能品到茶叶所经历的春夏秋冬和风霜雨雪；读一本书，我们的满足来自于通过书本连接到有趣的灵魂。有趣的人总会相遇，希望在读完本书后，你也能感受到书中纷繁枯燥的数学逻辑背后与你共振的有趣灵魂。

——罗翼，中国互联网资深码农，曾任去哪网高级总监，某著名跨境电商 CTO

前　言 *Preface*

近几年，人工智能无疑成为人们口中的热点话题，先是谷歌的 AlphaGo，后有百度的度秘、无人车，微软必应搜索推出的小冰。这一系列人工智能产品的推陈出新，令人眼花缭乱，一时间给人的感觉是人工智能遍地开花。无论人们接受还是不接受，人工智能都在迅速渗透各行各业。网络安全相比之下是一个传统行业，基于规则以及黑白名单的检测技术已经发展到了一定的瓶颈，而利益驱动的黑产团伙，其技术的发展已经远远超乎我们的想象。如何借助人工智能的力量，提升安全行业的整体检测与防护能力，成为各大安全厂商研究的课题。在国内安全行业，BAT 以及大量新兴的创业公司先后进入企业安全领域，他们凭借着自身数据搜集、处理、积累以及人工智能方面的优势，正在逐渐改变着整个安全行业。安全产品的形态也从硬件盒子逐步走向混合模式以及云端 SaaS 服务，安全技术从重防御逐步走向数据分析以及智能驱动。传统安全厂商也凭借其强大的安全人才储备，迅速推进人工智能在安全产品的落地。

我在网络安全这个行业搬了好几年砖，前五年做大型互联网公司的企业安全建设，从准入系统到 WAF、SIEM、IPS 等，基本都开发或者使用过，最近三年一直负责云安全产品，从抗 D、WAF 产品到、SIEM、入侵检测等，使用的技术从规则、黑白名单、模型、沙箱再到机器学习，从单机的 OSSIM 到 Hadoop、Storm、Spark、ELK，也算目睹了安全技术或者更准确地说是数据分析处理技术的迅猛发展。我深深感到，使用人工智能技术改变这个行业不是我们的选择，而是必经之路。我在真正意义上接触机器学习是 2014 年年底，当时带领了一个很小的团队尝试使用机器学习算法解决安全问题，磕磕绊绊一直走到现在，变成几十人的一个产品团队。

本书是我机器学习三部曲的第一部，主要以机器学习常见算法为主线，以生活中的例子和具体安全场景介绍机器学习常见算法，定位为机器学习入门书籍，便于大家可以快速上手。全部代码都能在普通 PC 上运行。第二部将重点介绍深度学习，并以具体的十个案例介绍机器

学习的应用，主要面向具有一定机器学习基础或致力于使用机器学习解决工作中问题的读者，全书的重点集中在问题的解决而不是算法的介绍。由于深度学习通常计算量已经超过了 PC 的能力，部分代码需要在服务器甚至 GPU 上运行，不过这不影响大家的阅读与学习。第三部将重点介绍强化学习和对抗网络，并以若干虚构安全产品或者项目介绍如何让机器真正具备 AlphaGo 级别的智能。

本书的第 1 章概括介绍了机器学习的发展以及互联网目前的安全形势。第 2 章介绍了如何打造自己的机器学习工具箱。第 3 章概括介绍机器学习的基本概念。第 4 章介绍 Web 安全的基础知识。第 5 章到第 13 章介绍浅层机器学习算法，包括常见的 K 近邻、决策树、朴素贝叶斯、逻辑回归、支持向量机、K-Means、FP-growth、Apriori、隐式马尔可夫、有向图。第 14 章到第 17 章介绍神经网络以及深度学习中常用的递归神经网络和卷积神经网络。每章都会以生活中的例子开头，让读者有一个感性的认识，然后简短介绍基础知识，最后以安全领域的 2~3 个例子讲解如何使用该算法解决问题。全书定位是能让更多的安全爱好者以及信息安全从业者了解机器学习，动手使用简单的机器学习算法解决实际问题。在写作中尽量避免生硬的说教，能用文字描述的尽量不用冷冰冰的公式，能用图和代码说明的尽量不用多余的文字。正如霍金所言“多写 1 个公式，少一半读者”，希望反之亦然。

机器学习应用于安全领域遇到的最大问题就是缺乏大量的黑样本，即所谓的攻击样本，尤其相对于大量的正常业务访问，攻击行为尤其是成功的攻击行为是非常少的，这就给机器学习带来了很大挑战。本书很少对不同算法进行横向比较，也是因为确实在不同场景下不同算法表现差别很大，很难说深度学习就一定比朴素贝叶斯好，也很难说支持向量机就比不过卷积神经网络，拿某个具体场景进行横向比较意义不大，毕竟选择算法不像购买 SUV，可以拿几十个参数评头论足，最后还是需要大家结合实际问题去选择。

这里我要感谢我的家人对我的支持，本来工作就很忙，没有太多时间处理家务，写书以后更是花费了我大量的休息时间，我的妻子无条件承担起了全部家务，尤其是照料孩子等繁杂事务。我很感谢我的女儿，写书这段时间几乎没有时间陪她玩，她也很懂事地自己玩，我想用这本书作为她的生日礼物送给她。我还要感谢吴怡编辑对我的支持和鼓励，让我可以坚持把这本书写完。最后还要感谢各位业内好友尤其是我 boss 对我的支持，排名不分先后：马杰 @ 百度安全、冯景辉 @ 百度安全、林晓东 @ 百度基础架构、黄颖 @ 百度 IT、李振宇 @ 百度 AI、Lenx@ 百度安全、黄正 @ 百度安全、程岩 @ 百度云、郝轶 @ 百度云、云鹏 @ 百度无人车、赵林林 @ 微步在线、张宇平 @ 数盟、谢忱 @Freebuf、李新 @Freebuf、李琦 @ 清华、徐恪 @ 清华、王宇 @ 蚂蚁金服、王珉然 @ 蚂蚁金服、王龙 @ 蚂蚁金服、周涛 @ 启明星辰、姚志武 @ 借贷宝、刘静 @ 安天、刘袁君 @ 医渡云、廖威 @ 易宝支付、尹毅 @

sobug、宋文宽@联想、团长@宜人贷、齐鲁@搜狐安全、吴圣@58安全、康宇@新浪安全、幻泉@i春秋、雅驰@i春秋、王庆双@i春秋、张亚同@i春秋、王禾@微软、李臻@paloalto、西瓜@四叶草、郑伟@四叶草、朱利军@四叶草、土夫子@XSRC、英雄马@乐视云、sbilly@360、侯曼@360、高磊@滴滴、高磊@爱加密、高渐离@华为、刘洪善@华为云、宋柏林@一亩田、张昊@一亩田、张开@安恒、李硕@智联、阿杜@优信拍、李斌@房多多、李程@搜狗、Tony@京东安全、简单@京东安全、姚聪@face+、李鸣雷@金山云，最后我还要感谢我的亲密战友陈燕、康亮亮、蔡奇、哲超、新宇、子奇、月升、王磊、碳基体、刘璇、钱华钧、刘超、王胄、吴梅、冯侦探、冯永校。

本书面向信息安全从业人员、高等院校计算机相关专业学生以及信息安全爱好者，机器学习爱好者，对于想了解人工智能的CTO、运维总监、架构师同样也是一本不错的科普书籍。当读者在工作学习中遇到问题时可以想起本书中提到的一两种算法，那么我觉得就达到效果了，如果可以让读者像使用printf一样使用SVM、朴素贝叶斯等算法，那么这本书就相当成功了。

我平时在FreeBuf专栏以及i春秋分享企业安全建设以及人工智能相关经验与最新话题，同时也运营我的微信公众号“兜哥带你学安全”，欢迎大家关注并在线交流。

本书使用的代码和数据均在GitHub上发布，地址为：https://github.com/duoergun0729/1book，代码层面任何疑问可以在GitHub上直接反馈。