Web安全之机器学习入门

Web安全之机器学习入门_下载链接1_

著者:刘焱

出版者:

出版时间:2017-8

装帧:平装

isbn:9787111576426

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。

作者介绍:

目录: 对本书的赞誉
序一
序二
序三
前言
第1章 通向智能安全的旅程 1
1.1 人工智能、机器学习与深度学习 1
1.2 人工智能的发展 2
1.3 国内外网络安全形势 3
1.4 人工智能在安全领域的应用 5
1.5 算法和数据的辩证关系 9
1.6 本章小结 9
参考资源 10
第2章 打造机器学习工具箱 11
2.1 Python在机器学习领域的优势 11
2.1.1 NumPy 11
2.1.2 SciPy 15
2.1.3 NLTK 16
2.1.4 Scikit-Learn 17
2.2 TensorFlow简介与环境搭建 18
2.3 本章小结 19
参考资源 20
第3章 机器学习概述 21
3.1 机器学习基本概念 21

3.2 数据集 22
3.2.1 KDD 99数据 22
3.2.2 HTTP DATASET CSIC 2010 26
3.2.3 SEA数据集 26
3.2.4 ADFA-LD数据集 27
3.2.5 Alexa域名数据 29
3.2.6 Scikit-Learn数据集 29
3.2.7 MNIST数据集 30
3.2.8 Movie Review Data 31
3.2.9 SpamBase数据集 32
3.2.10 Enron数据集 33
3.3 特征提取 35
3.3.1 数字型特征提取 35
3.3.2 文本型特征提取 36
3.3.3 数据读取 37
3.4 效果验证 38
3.5 本章小结 40
参考资源 40
第4章 Web安全基础 41
4.1 XSS攻击概述 41
4.1.1 XSS的分类 43
4.1.2 XSS特殊攻击方式 48
4.1.3 XSS平台简介 50
4.1.4 近年典型XSS攻击事件分析 51
4.2 SQL注入概述 53
4.2.1 常见SQL注入攻击 54
4.2.2 常见SQL注入攻击载荷 55
4.2.3 SQL常见工具 56
4.2.4 近年典型SQL注入事件分析 60
4.3 WebShell概述 63
4.3.1 WebShell功能 64
4.3.2 常见WebShell 64
4.4 僵尸网络概述 67
4.4.1 僵尸网络的危害 68
4.4.2 近年典型僵尸网络攻击事件分析 69
4.5 本章小结 72
参考资源 72
第5章 K近邻算法 74
5.1 K近邻算法概述 74
5.2 示例：hello world！K近邻 75
5.3 示例：使用K近邻算法检测异常操作（一） 76
5.4 示例：使用K近邻算法检测异常操作（二） 80
5.5 示例：使用K近邻算法检测Rootkit 81
5.6 示例：使用K近邻算法检测WebShell 83
5.7 本章小结 85
参考资源 86
第6章 决策树与随机森林算法 87
6.1 决策树算法概述 87
6.2 示例：hello world！决策树 88
6.3 示例：使用决策树算法检测POP3暴力破解 89
6.4 示例：使用决策树算法检测FTP暴力破解 91
6.5 随机森林算法概述 93
6.6 示例：hello world！随机森林 93
6.7 示例：使用随机森林算法检测FTP暴力破解 95

6.8 本章小结 96
参考资源 96
第7章 朴素贝叶斯算法 97
7.1 朴素贝叶斯算法概述 97
7.2 示例：hello world！朴素贝叶斯 98
7.3 示例：检测异常操作 99
7.4 示例：检测WebShell（一） 100
7.5 示例：检测WebShell（二） 102
7.6 示例：检测DGA域名 103
7.7 示例：检测针对Apache的DDoS攻击 104
7.8 示例：识别验证码 107
7.9 本章小结 108
参考资源 108
第8章 逻辑回归算法 109
8.1 逻辑回归算法概述 109
8.2 示例：hello world！逻辑回归 110
8.3 示例：使用逻辑回归算法检测Java溢出攻击 111
8.4 示例：识别验证码 113
8.5 本章小结 114
参考资源 114
第9章 支持向量机算法 115
9.1 支持向量机算法概述 115
9.2 示例：hello world！支持向量机 118
9.3 示例：使用支持向量机算法识别XSS 120
9.4 示例：使用支持向量机算法区分僵尸网络DGA家族 124
9.4.1 数据搜集和数据清洗 124
9.4.2 特征化 125
9.4.3 模型验证 129
9.5 本章小结 130
参考资源 130
第10章 K-Means与DBSCAN算法 131
10.1 K-Means算法概述 131
10.2 示例：hello world！K-Means 132
10.3 示例：使用K-Means算法检测DGA域名 133
10.4 DBSCAN算法概述 135
10.5 示例：hello world！DBSCAN 135
10.6 本章小结 137
参考资源 137
第11章 Apriori与FP-growth算法 138
11.1 Apriori算法概述 138
11.2 示例：hello world！Apriori 140
11.3 示例：使用Apriori算法挖掘XSS相关参数 141
11.4 FP-growth算法概述 143
11.5 示例：hello world！FP-growth 144
11.6 示例：使用FP-growth算法挖掘疑似僵尸主机 145
11.7 本章小结 146
参考资源 146
第12章 隐式马尔可夫算法 147
12.1 隐式马尔可夫算法概述 147
12.2 hello world! 隐式马尔可夫 148
12.3 示例：使用隐式马尔可夫算法识别XSS攻击（一） 150
12.4 示例：使用隐式马尔可夫算法识别XSS攻击（二） 153
12.5 示例：使用隐式马尔可夫算法识别DGA域名 159
12.6 本章小结 162

参考资源 162
第13章 图算法与知识图谱 163
13.1 图算法概述 163
13.2 示例：hello world！有向图 164
13.3 示例：使用有向图识别WebShell 169
13.4 示例：使用有向图识别僵尸网络 173
13.5 知识图谱概述 176
13.6 示例：知识图谱在风控领域的应用 177
13.6.1 检测疑似账号被盗 178
13.6.2 检测疑似撞库攻击 179
13.6.3 检测疑似刷单 181
13.7 示例：知识图谱在威胁情报领域的应用 183
13.7.1 挖掘后门文件潜在联系 184
13.7.2 挖掘域名潜在联系 185
13.8 本章小结 187
参考资源 187
第14章 神经网络算法 188
14.1 神经网络算法概述 188
14.2 示例：hello world！神经网络 190
14.3 示例：使用神经网络算法识别验证码 190
14.4 示例：使用神经网络算法检测Java溢出攻击 191
14.5 本章小结 193
参考资源 194
第15章 多层感知机与DNN算法 195
15.1 神经网络与深度学习 195
15.2 TensorFlow编程模型 196
15.2.1 操作 197
15.2.2 张量 197
15.2.3 变量 198
15.2.4 会话 198
15.3 TensorFlow的运行模式 198
15.4 示例：在TensorFlow下识别验证码（一） 199
15.5 示例：在TensorFlow下识别验证码（二） 202
15.6 示例：在TensorFlow下识别验证码（三） 205
15.7 示例：在TensorFlow下识别垃圾邮件（一） 207
15.8 示例：在TensorFlow下识别垃圾邮件（二） 209
15.9 本章小结 210
参考资源 210
第16章 循环神经网络算法 212
16.1 循环神经网络算法概述 212
16.2 示例：识别验证码 213
16.3 示例：识别恶意评论 216
16.4 示例：生成城市名称 220
16.5 示例：识别WebShell 222
16.6 示例：生成常用密码 225
16.7 示例：识别异常操作 227
16.8 本章小结 230
参考资源 230
第17章 卷积神经网络算法 231
17.1 卷积神经网络算法概述 231
17.2 示例：hello world！卷积神经网络 234
17.3 示例：识别恶意评论 235
17.4 示例：识别垃圾邮件 237
17.5 本章小结 240

参考资源 242
・・・・・・・(收起)

Web安全之机器学习入门_下载链接1_

标签

机器学习

安全

人工智能

计算科学

计算机科学

web

HTTP

HAOK

评论

师父领进门修行在个人，兜哥已经尽力了，要把这些东西消化吸收，应用在自己学习工作中，要下一番苦工，这本书就像书名额，入门，要深入，还是需要看其他的书和开源的机器学习框架。

书上代码好多是错的，作者对一些包都调不对，有点像是圈钱的

侵权洗稿请先道歉

的确只是入门，介绍了几个安全方面的数据集，以及如何特征化处理这些数据集，简略介绍了各个机器算法，简单的Python包调用这些算法在数据集上的使用。而实际工作中如何筛选出黑白样本？安全类的黑白样本数据量差距之大，黑样本的变异进化，如何处理都没有介绍，只是适合入门，符合书名。

内容很简略，有基础后翻翻就好

凑合，虽然是兜哥的书，真的是凑合，以后看书买书就得避开所有『入门』、『速成』、『简史』的字样

瞅了一眼目录

很多东西说的太简单了，需要自己去查更多的知识点。书里的代码并没有过多的解释这样写的原因

很简略 ..

入门谈不上，只能对基本算法的简单了解，只谈表相，不聊原理，例子全是罗列的代码。。。。

写的太过简单。

数据集居然介绍了10页...套路是
概念概述(摘录其他)/安全数据集/算法实验/注释,200多页真没那么多干货而且79的定价,

对比周老师的88的机器学习,真的没什么诚意

幸好没买…

Web安全之机器学习入门_下载链接1_

书评

兜哥在安全圈大名鼎鼎，早有耳闻，看到这本书，赞誉部分，互联网小半个圈子的安全负责人不吝誉美之词，虽然不知道是否真正读过。不过，作为可以上手练习的实操人工智能机器学习算法的入门指南，本书还是值得推荐。
基于特征和签名的传统扫描和识别算法，对未知威胁的无能为力，...

Web安全之机器学习入门_下载链接1_