Packt>

异步图书
www.epubit.com.cn

用Python探索Web机器学习系统开发　让系统做出更加聪明的预测

机器学习 Web 应用

Machine Learning for the Web

eBay公司EU Analytics部门负责人Davide Cervellin作序

[意] Andrea Isoni 著

杜春晓 译

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

机器学习 Web 应用

Machine Learning for the Web

[意] Andrea Isoni 著

杜春晓 译

人 民 邮 电 出 版 社

北 京

图书在版编目（CIP）数据

机器学习Web应用 /（意）爱索尼克（Andrea Isoni）著；杜春晓译. -- 北京 : 人民邮电出版社，2017.8
ISBN 978-7-115-45852-0

Ⅰ. ①机… Ⅱ. ①爱… ②杜… Ⅲ. ①机器学习
Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2017)第141915号

版权声明

◆ 著 [意] Andrea Isoni
译 杜春晓
责任编辑 陈冀康
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
北京鑫正大印刷有限公司印刷

◆ 开本：800×1000 1/16
印张：14.5
字数：280 千字 2017 年 8 月第 1 版
印数：1 – 2 400 册 2017 年 8 月北京第 1 次印刷

著作权合同登记号 图字：01-2016-8595 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

机器学习可用来处理由用户产生的、数量不断增长的 Web 数据。

本书讲解如何用 Python 语言、Django 框架开发一款 Web 商业应用，以及如何用一些现成的库和工具（sklearn、scipy、nltk 和 Django 等）处理和分析应用所生成或使用的数据。本书不仅涉及机器学习的核心概念，还介绍了如何将数据部署到用 Django 框架开发的 Web 应用，包括 Web、文档和服务器端数据的挖掘和推荐引擎的搭建方法。

本书适合有志于成为或刚刚成为数据科学家的读者学习，也适合对机器学习、Web 数据挖掘等技术实践感兴趣的读者参考阅读。

序

机器学习是什么？2016 年，无论参加大会、研讨会，还是接受采访，很多人都让我给机器学习下个定义。人们对机器学习是什么，抱有诸多疑问。理解这一新鲜事物可能为生活带来的潜在影响以及它日后对我们有何种意义之前，天性要求我们先给出其定义。

跟其他陡升为显学的学科类似，机器学习并不是新生事物。科学社区多年来一直致力于研制算法，实现重复性工作的自动化。参数固定的算法叫作静态算法，其输出是可预测的，输出只是输入变量的函数。还有一种情况，算法的参数是动态变化的，算法的输出是外部因素（最常见的是同一算法先前的输出）的函数，这种算法叫作动态算法，其输出不仅仅是输入变量的函数。动态算法是机器学习的支柱：从先前迭代生成的数据中，学习到一组规则，以改善之后的输出。

科学家、开发人员和工程师研究和使用模糊逻辑、神经网络和其他类型的机器学习技术已有多个年头，但直到今天，随着机器学习应用离开实验室，进入市场营销、销售和金融行业，这门学科才流行起来，基本上来讲，需要重复执行相同运算的活动都可以受益于机器学习。

机器学习的影响很容易理解，它将给我们的社会带来巨大冲击。关于下一个 5 到 10 年，机器学习将给我们带来什么，我能想到的最佳描述方式是：不妨回想工业革命时期发生了什么。蒸汽机发明之前，很多人从事高度重复性的体力工作。为了赚取少得不能再少的工资，他们往往要冒着生命危险或以牺牲健康为代价。工业革命出现后，社会得以发展，机器接管了生产过程的重要步骤，这带来了产量的增加，并且产出的可预测性更强和更稳定。与之相应的是，产品质量的提升和新工种的出现，操控机器这类新兴的工作取代了体力劳动。我们将造物的责任委托给由我们设计和发明的工具，这在人类历史上可是第一次。

机器学习将以相同的方式，改变执行数据运算的方式，减少人工干预的需要，将优化的工作交给机器和算法。数据处理人员将不再直接控制数据，而是通过控制算法间接控制数据。因此，运算的执行速度将会变得更快，更少的人将能控制规模更大的数据集，错误将会减少，从而结果的稳定性更高和可预测性更强。跟其他对我们生活产生重大影响的事物一样，爱慕和憎恶它的人都有。爱慕者称赞机器学习为他们生活带来便利；憎恶者批评，机器学习方法要有效，需要大量的迭代，因此需要大量数据。而通常来讲，我们“喂给”算法的数据可是我们的个人信息[①]。

事实上，机器学习作为一种工具得以迅速发展，其主要应用在于提升市场营销和顾客支持的效率。为顾客提供个性化服务，促使他们购买而不只是浏览，或让他们高兴而不是失望，需要对顾客有着深入的理解。

例如，就市场营销而言，如今市场营销人员开始考虑位置、设备、购买历史、访问过的网站、天气状况等信息（仅举几个例子）来决定公司是否向一组特定顾客展示广告。

通过电视或报纸这样无法追踪的媒体传播营销信息的日子已然成为遥远的过去。如今，市场营销人员希望知道谁点击和购买了他们商品等一切信息，他们好优化创意和投入，合理分配预算，以充分利用他们手中的资源。这就要求提供前所未有的高度个性化服务，若使用合理，可以让顾客感到他们是受尊重的个体而不只是某一社会人口学分组的一部分。

机器学习既吸引人又充满挑战，但无疑下一个十年的赢家，将会是那些能够理解非结构化数据，并且能够基于这些数据以可扩展的方式做出决策的公司或个人：除了机器学习，我还没有看到哪种方式能实现这样的伟业。

Andrea Isoni 的这本书朝这个世界迈出了一步；读它就好像是向下窥视兔子的洞穴，你能从中看到用机器学习技术实现的几个应用，作者将机器学习技术整合到 Web 应用中。访问用机器学习技术创建的个性化服务网站，顾客能从中体验到为他们个人提供的优化过的服务。

如果你想为日后的职业生涯提前做好准备，该书是你必须要读的；下一个十年跟数据打交道的任何人若想成功的话，都需要熟练掌握这些技术。

Davide Cervellin，@ingdave

eBay 公司 EU Analytics 部门负责人

[①] 言外之意，隐私受到威胁。——译者注

译者序

20 年前，IBM 研制的深蓝计算机勉强战胜俄罗斯棋王卡斯帕罗夫，它在体力上的优势似乎比智力方面更明显。但刚刚过去的这一年，谷歌的 AlphaGo 计算机程序打败了围棋高手李世石，它的升级版 Master 威力更是了得，横扫中日韩高手，它擅长走快棋，招法狠毒，令人类高手胆颤。由此可见，近年来，人工智能技术随着硬件、大数据、机器学习技术的发展，取得了长足的进步。

机器学习技术作为人工智能的一个子领域，研究和应用热潮不减，研讨会、学习班和创业项目层出不穷；国内学者入选 AAAI Fellow；该领域的书籍一印再印；人脸识别、自动驾驶、机器翻译、智能客服、物流无人机和家居、医疗、教育机器人等各种应用不断推向市场。从以上种种表现来看，我们处在人工智能时代的风口和前夕。作为该领域的从业者，我们不能满足于看热闹，应努力掌握背后的核心技术——机器学习，力求弄懂该技术，并努力探索其他可能的实现人工智能的方法，把人类智慧的边界向前推进一步。更令人鼓舞的是，大数据产业发展已上升到国家战略层面，我国要实现从数据大国向数据强国的转变，需要一批掌握了数据挖掘、机器学习等相关技术的人才。

本书讲解的是商业网站数据分析和挖掘所用到的机器学习理论和技术。作者先介绍了机器学习的基本概念、Python 机器学习工具栈（NumPy、pandas 和 matplotlib 等），接着分别讲解了无监督和有监督机器学习理论，每种方法都给出形式化描述，其间用到了大量概率统计、线性代数等数学知识，比如最小二乘、相关性、贝叶斯概率和奇异值分解等。作者的统计学背景在这一点上得到了很好的体现。这部分数学知识能够较好地满足有志于深入学习的读者的需要，水平高的读者可以从中感受机器学习模型的数学魅力。介绍完这两大类机器学习理论，作者又从 Web 结构和内容两个方面讲解了 Web 挖掘技术；介绍了信息

检索模型、主题抽取模型 LDA。讲解完机器学习理论和技术之后，作者引入了为 Web 开发完美主义者准备的 Django 框架，让昔日在幕后默默奉献的数据分析高手有机会走到台前，用自己研制的算法驱动一款 Web 产品。作者带领我们利用前面讲解的算法和挖掘技术，用 Django 框架搭建推荐系统和影评分析系统。学到这里，你会不由地感叹 Python 真是全栈工程师的好朋友。数据分析师用 Python 就能从头到尾打造一款智能 Web 产品，可见 Python 的应用范围之广。年初，Facebook 更是开源了 PyTorch 深度学习框架，进一步巩固了 Python 在机器学习领域的地位。

当然，我们最终开发出的产品还比较初级，离最终面向用户的产品在用户体验上还有较大差距，但稍加打磨至少可作为一个最小可行性产品（MVP）先行投入市场，收集用户反馈，日后再图大的改进。此外，限于篇幅，作者也没有讲怎么将系统部署到生产服务器。感兴趣的读者可以试试 Heroku、SAE 等云应用平台，也可以尝试用 Apache、mod_wsgi 在自己的计算机上搭服务器。你可能还需要申请一个域名。这样，你就可以向朋友推荐自己开发的产品了，你具备了向全球用户提供智能 Web 产品的能力！嘿，伙计，快来看，这是我刚刚上线的 Web 推荐系统！用机器学习算法驱动的哦！

感谢人民邮电出版社的陈冀康编辑等为本书编校、出版辛勤付出的各位朋友。读者罗导运行了第 1 章的代码，并指出了原书及译者注中的几处问题。师妹瞿乔阅读了第 2 章译文，她本人也是一本 Python 图书的译者。泰安读者陈新光阅读了第 6 章译文，他正努力学习数据科学知识，祝他学有所成。翻译过程中，我向北京大学冷含莹、东京大学范超、上海健康学院姜萌等朋友请教过问题；我旁听了北大的统计学基础、随机过程等课程，了解了很多统计学概念，参考了市面上现有的多本著作，其中包括大名鼎鼎的西瓜书，查询了 CSDN 等网站的文章，在此一并表示衷心的感谢。感谢西安工业大学的李刚老师、重庆大学杨刘洋同学等读者对翻译工作的支持。最后，感谢我的家人，我翻译图书的时间是用他们的辛勤劳动换来的，因此也更加宝贵。

本人学识有限，且时间仓促，书中翻译错误、不当和疏漏之处在所难免，敬请读者批评指正。

杜春晓

2017 年 2 月

作者简介

Andrea Isoni 博士是一名数据科学家、物理学家，他在软件开发领域有着丰富的经验，在机器学习算法和技术方面，拥有广博的知识。此外，他还有多种语言的使用经验，如 Python、C/C++、Java、JavaScript、C#、SQL、HTML。他还用过 Hadoop 框架。

译者简介

杜春晓，英语语言文学学士，软件工程硕士。其他译著有《Python 数据挖掘入门与实践》《Python 数据分析实战》和《电子达人——我的第一本 Raspberry Pi 入门手册》等。新浪微博：@宜_生。

技术审稿人简介

Chetan Khatri 是一名数据科学研究员，他共有 4 年半的研究和开发经验。他在 Nazara Technologies Pvt. Ltd 公司担任数据和机器学习方面的首席工程师，主导在游戏和电信订阅业务从事数据科学实践。他曾在一家顶尖的数据公司和印度四大公司其中一家工作，管理数据科学实践平台和后者的资源团队。在这之前，他曾供职于 R & D Lab 和 Eccella Corporation。他拥有印度喀奇大学（KSKV Kachchh University）的计算机科学硕士学位，辅修数据科学，是该学校的金牌得主。

他积极以多种方式为社会做贡献，其中包括为大二学生做讲座，在学术以及其他各种会议上介绍数据科学相关知识，还援助社区一个数据平台。他在学术研究和行业最佳实践两方面均有着相关的专业知识。他喜欢参加数据科学马拉松比赛。他参与发起了 Python 社区——PyKutch。他目前正在探究深层神经网络和增强学习，学习使用并行和分布式计算管理数据。

感谢喀奇大学计算机科学系主任 Devji Chhanga 教授，感谢他引领我走上数据科学研究的正确道路，并给予宝贵的指导意见。同样把感谢送给我亲爱的家人。

Pavan Kumar Kolluru 是一名交叉学科工程师，他是大数据、数字图像和处理、遥感（高光谱数据和图像）方面的专家，精通 Python、R 和 MATLAB 编程。他的研究重点在于如何用机器学习技术、编制算法处理大数据。

他目前正在探索如何找到不同学科之间的联系，以降低数据处理过程在计算和自动化方面的难度。

作为一名数据（图像和信号）处理方面的专业人士和老师，他一直在处理多 / 高光谱数据，该项工作使得他在数据处理、信息抽取和分割方面积累了很多专业知识。他用到的

高级处理技术有 OOA、随机集和马尔可夫随机场。

作为一名程序员和教师，他专注于 Python 和 R 语言，他执教于企业和教育行业的兄弟会。他培训过多批学员，教他们使用 Python 和各种包（信号、图像和数据分析等）。

作为一名机器学习研究员 / 教练，他是分类（有监督和无监督）、建模和数据理解、回归以及数据降维方面的专家。他曾开发出一套大数据（图像或信号）方面的新型机器学习算法，作为他理科硕士阶段的研究成果，该算法将数据降维和分类纳入同一框架，这为他赢得了很高的分数。他培训过多家大型公司的员工，教他们用 Hadoop 和 MapReduce 分析大数据。他的大数据分析专业知识包括 HDFS、Pig、Hive 和 Spark。

Dipanjan Sarkar 是 Intel 公司的一名数据科学家。Intel 是世界上最大的半导体公司，它的使命是让世界更加连通和更具效率。他主要从事分析、商业智能、应用开发和构建大规模的智能系统方面的工作。他从班加罗尔的印度信息技术学院（IIIT）获得信息技术硕士学位。他的专业领域包括软件工程、数据科学、机器学习和文本分析。

Dipanjan 的兴趣包括学习新技术、数据科学和最近的深度学习以及了解具有颠覆性的初创企业动态。业余时间，他喜欢阅读、写作、玩游戏和看情景喜剧。他写过一本关于机器学习的书 *R Machine Learning by Example*，该书由 Packt Publishing 出版。他还为 Packt Publishing 出版的几本机器学习和数据科学图书做过技术评审。

前言

数据科学，尤其是机器学习，成为当下科技商业领域人们热议的议题。这类技术可用来处理用户产生的、数量在不断增长的数据。本书将讲解如何用 Python 语言、Django 框架开发一款 Web 商业应用，还将讲解如何用一些现成的库（sklearn、scipy、NLTK 和 Django 等）处理和分析（通过机器学习技术）应用生成或使用的数据。

本书主要内容

第 1 章，Python 机器学习实践入门，讨论机器学习的主要概念以及数据科学专业人士用 Python 处理数据所使用的几个库。

第 2 章，无监督机器学习，讲解为数据集分簇和从数据中抽取主要特征所用到的算法。

第 3 章，有监督机器学习，讲解预测数据集标签最常用的有监督机器学习算法。

第 4 章，Web 挖掘技术，讨论 Web 数据的组织、分析和从中提取信息的主要技术。

第 5 章，推荐系统，详细介绍当今商业领域所使用的几种最流行的推荐系统。

第 6 章，开始 Django 之旅，介绍开发 Web 应用所用到的 Django 的主要功能和特点。

第 7 章，电影推荐系统 Web 应用，将介绍的机器学习概念付诸实践，动手实现为 Web 用户推荐电影的应用。

第 8 章，影评情感分析应用，再次通过一个实例，使用讲述的知识，分析在线影评的情感倾向和相关性。

本书的阅读前提

读者应该准备一台计算机，装好 Python 2.7，能够运行（和修改）书中各章讲解的代码。

本书的目标读者

任何有一定编程经验（Python）和统计学背景，对机器学习感兴趣和/或希望从事数据科学职业的读者均可从本书受益。

排版约定

本书使用不同的文本样式来区分不同类别的内容。以下是常用样式及其用途说明。

正文中的代码、数据库表名、文件夹名、文件名、文件扩展名、路径名、URL 地址、用户输入的内容和 Twitter 用户名显示方式如下：

“在终端输入以下命令，安装 Django 这个库：sudo pip install django。”

代码块样式如下：

```
INSTALLED_APPS = (
...
'rest_framework',
'rest_framework_swagger',
'nameapp',
)
```

所有的命令行输入或输出使用下面这种样式：

```
python manage.py migrate
```

新的术语和**重要的词语**使用黑体。出现在屏幕上的词语，例如菜单或对话框里，样式如下“如你所见，页面上有两个输入框，输入姓名和邮箱后，单击‘**添加**’，将其添加到数据库”。

此图标表示警告或重要信息。

此图标表示提示或技巧。

读者反馈

我们热忱地欢迎读者朋友给予我们反馈，告诉我们你对于这本书的所思所想——你喜欢或是不喜欢哪些内容。大家的反馈对我们来说至关重要，将帮助我们确定到底哪些内容是读者真正需要的。

如果你有一般性建议的话，请发邮件至 feedback@packtpub.com，请在邮件主题中写清书的名称。

如果你是某一方面的专家，对某个主题特别感兴趣，有意向自己或是与别人合作写一本书，请到 www.packtpub.com/authors 查阅我们为作者准备的帮助文档。

客户支持

为自己拥有一本 Packt 出版的书而自豪吧！为了让你的书物有所值，我们还为你准备了以下内容。

下载示例代码

如果你是从 www.packtpub.com 网站购买的图书，用自己的账号登录后，可以下载所有已购图书的示例代码。如果你是从其他地方购买的，请访问 http://www.packtpub.com/support 网站并注册，我们会用邮件把代码文件直接发给你。也可以访问 www.epubit.com.cn 来下载示例代码。

代码文件下载步骤如下。

1. 用邮箱和密码登录或注册我们的网站。
2. 鼠标移动到页面顶部的 **SUPPORT** 选项卡下。
3. 单击 **Code Downloads & Errata**。
4. 在搜索框 **Search** 中输入书名。
5. 选择你要下载代码文件的图书。

6. 从下拉菜单中选择你从何处购买该书。

7. 单击 **Code Download** 下载代码文件。

你还可以在 Packt Publishing 网站图书详情页，单击 **Code Files** 按钮下载代码文件。在 **Search** 搜索框中输入书名进行搜索可找到该书的图书详情页。请注意你需要登录网站。

代码下载下来之后，请确保用以下解压工具的最新版本进行解压或抽取文件：

- Windows 用户：WinRAR / 7-Zip；
- Mac 用户：Zipeg / iZip / UnRarX；
- Linux 用户：7-Zip / PeaZip。

本书的代码包在 GitHub 上也存储了一份：https://github.com/PacktPublishing/Machine-Learning-for-the-Web。我们很多其他图书和视频的代码包也存储到了 GitHub 上：https://github.com/PacktPublishing/。将它们检出到本地。

下载本书配套 PDF 文件

我们还为你准备了一个 PDF 文件，该文件包含书中的所有屏幕截图 / 图表。这些彩图能弥补书中黑白图像的不足，有助于你理解本书内容。该文件的下载地址为 http://www.packtpub.com/sites/default/files/downloads/MachineLearningfortheWeb_ColorImages.pdf。

勘误表

即使我们竭尽所能来保证图书内容的正确性，错误也在所难免。如果你在我们出版的任何一本书中发现错误——可能是在文本或代码中——倘若你能告诉我们，我们将会非常感激。你的善举足以减少其他读者在阅读出错位置时的纠结和不快，帮助我们在后续版本中更正错误。如果你发现任何错误，请访问 http://www.packtpub.com/submit-errata，选择相应书籍，单击“Errata Submission Form”链接，输入错误之处的具体信息。你提交的错误得到验证后，我们就会接受你的建议，该处错误信息将会上传到我们网站或是添加到已有勘误表的相应位置。

访问 https://www.packtpub.com/books/content/support，在搜索框中输入书名，可查看该书已有的勘误信息。这部分信息会在 Errata 部分显示。

版权保护

所有媒体在互联网上都面临的一个问题就是侵权。对 Packt 来说，我们严格保护我们

的版权和许可。如果你在网上发现针对我们出版物的任何形式的盗版产品，请立即告知我们地址或网站名称，以便我们进行补救。

请将盗版书籍的网址发送到 copyright@packtpub.com。

如果你能这么做，就是在保护我们的作者，保护我们，只有这样，我们才能继续以优质内容回馈像你这样热心的读者。

问题

你对本书有任何方面的问题，都可以通过 questions@packtpub.com 邮箱联系我们，我们也将尽最大努力来帮你答疑解惑。

目录