HZ BOOKS 华章IT
[PACKT] PUBLISHING

智能系统与技术丛书

Designing Machine Learning Systems with Python

机器学习系统设计

Python语言实现

[美] 戴维·朱利安（David Julian）著
李洋 译

机械工业出版社
China Machine Press

智能系统与技术丛书

Designing Machine Learning Systems with Python

机器学习系统设计

Python语言实现

[美] 戴维·朱利安（David Julian）著

李洋 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

机器学习系统设计：Python 语言实现 /（美）戴维·朱利安（David Julian）著；李洋译．
—北京：机械工业出版社，2017.5
（智能系统与技术丛书）
书名原文：Designing Machine Learning Systems with Python

ISBN 978-7-111-56945-9

I. 机… II. ①戴… ②李… III. 机器学习－系统设计 IV. TP181

中国版本图书馆 CIP 数据核字（2017）第 096698 号

本书版权登记号：图字：01-2017-0486

机器学习系统设计：Python 语言实现

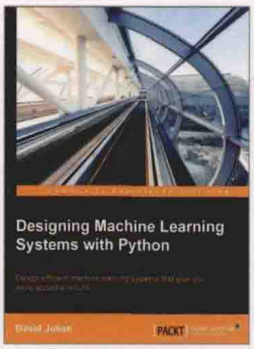
出版发行：机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码：100037）
责任编辑：陈佳媛　　责任校对：殷　虹
印　　刷：三河市宏图印务有限公司　　版　　次：2017 年 6 月第 1 版第 1 次印刷
开　　本：186mm×240mm　1/16　　印　　张：12.5
书　　号：ISBN 978-7-111-56945-9　　定　　价：59.00 元

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

本书介绍了机器学习系统设计的整个过程，以及相关的Python库，并在各个知识环节中都给出了Python示例，为设计高效机器学习系统提供详实指南。

本书共9章，第1章介绍机器学习的设计原理和相关模型；第2章讲解Python中众多针对机器学习任务的程序包；第3章涵盖大数据、数据属性、数据源、数据处理和分析等主题，介绍基本的数据类型、结构和属性；第4章探索最常见的机器学习模型，即逻辑模型、树状模型和规则模型；第5章研究机器学习最常用的技术，创建线性回归和Logistic回归的假设语句；第6章介绍人工神经网络算法；第7章讨论特征的不同类型，即定量特征、有序特征和分类特征，以及如何结构化和变换特征；第8章介绍主要的集成方法及其在Scikit-learn中的实现；第9章介绍模型选择和参数调优技术，并将这些技术应用于一些案例研究之中。



原书封面

作者简介

戴维·朱利安（David Julian）数据分析员、信息系统咨询顾问和培训讲师。他目前正致力于Urban Ecological Systems和Blue Smart Farms（http://www.bluesmartforms.com.au）的机器学习项目，该项目旨在发现和预测温室作物虫害。

封面设计 锡 彬

华章IT
HZBOOKS | Information Technology

THE TRANSLATOR'S WORDS

译 者 序

2016 年，对于计算机相关从业者（和职业围棋手）而言，毋庸置疑，最具冲击力的大事件就是 AlphaGo 的成功了。对此，即便是如我本人这样最迟钝的计算机工程师，也终于不能无动于衷，感觉是时候跳出 if-else 的懒惰，捡起尘封多年乃至遗忘的线性规划和微积分等知识，投身于人工智能的汪洋了。历经 60 载的孕育，人工智能的时代终于到来了。

回想起本世纪初，我曾参与了电信公司的一个营销项目，这个项目的目标是建立一系列客户指标，以反映客户的价值和分类，使营销人员能够进行精准营销和客户关怀。对于这个项目，当时的术语是，数据仓库和集市，旋转、切片、透视等统计分析，分类和聚类等数据挖掘，等等。当工作作风一向是直接有效（简单粗暴）的市场营销专家，了解到数据仓库和统计工具软硬件的昂贵、数据挖掘工作的繁杂之后，他们提出直接拿一套指标变量和决策阈值，然后用 if-else 来决定对付客户的营销手段。好吧，指标变量还好，但是优化的决策边界怎么拿？最终，一份虚构臆想的报告出炉了，对此，我至今仍怀有深深的罪恶感。

如今，市场营销的专家作风依旧吧？但是，即便是初出茅庐（大有可为）的软件工程师，也完全能够用触手可得的开源工具和计算环境，建立起一个机器学习系统，获得一些令人信服的决策边界优化解，让那些令人哭笑不得的推销短信变得更少，让短信垃圾成为雪中送炭，想要获取信息的人们无须再从一些衣冠楚楚、侃侃而谈的顾问手里购买一纸空洞的报告了。这就是人工智能的时代，在自动驾驶成为投资大鳄眼中的香饽饽时，人工智能已经无所不在了。本书也是如此，对于计算机科学专业的小伙伴们来说，书中的内容都不陌生，但当这些都成为随手可得、随时要用的东西时，就证明了我们已经身

处其时。

本书涵盖了建立机器学习系统的方方面面，相对比较基础，其中最有价值的是，书中介绍了机器学习系统设计的整个过程，以及相关的 Python 库，并在各个知识环节中都给出了 Python 示例。无论对于机器学习系统的新兵还是老手，本书都有一定的参考价值。对于机器学习系统的初学者而言，本书较为系统地介绍了相关知识，同时也在一开始就给出了语言和环境，能够让大家甩开膀子，撸起袖子，伸手开干；而对于机器学习系统的老手而言，其更多的参考价值在于如何使用 Python 来实现那些概念。

但需要注意的是，本书绝不是机器学习的学科教材，也不是 Python 库的用户手册，更不是实际项目的设计文档。因此，本书并没有对各种模型提供完整的解释和严格的推导，也没有对 Python 库的各种对象和函数提供完整详尽的说明，更不会对实际问题给出详细的解决方案和实现。但本书确实是一个简明的指引，并富有逻辑，让我们能够按图索骥，由此及彼，较为系统地了解 Python 机器学习系统设计的方方面面，并以此为线索，展开更多的阅读和深入的学习。同时，书中的诸多示例也能在一定程度上为我们解决类似问题提供思路。

在人工智能的时代，翻译一本机器学习的书籍，对译者而言也是幸甚至哉，借此与各路志士同仁共勉。

李洋

2017 年 2 月

PREFACE

前　言

机器学习是计算世界所见的最大趋势之一。机器学习系统具有意义深远且令人兴奋的能力，能够在各种应用领域为人们提供重要的洞察力，从具有开创性的挽救生命的医学研究到宇宙基础物理方面的发现，从为我们提供更健康、更清洁的食物到互联网分析和建立经济模型，等等。事实上，就某种意义而言，这项技术在我们的生活中已经无所不在。要想进入机器学习的领域，并且对其具有充分的认知，就必须能够理解和设计服务于某一项目需要的机器学习系统。

本书的主要内容

第 1 章从机器学习的基础知识开始，帮助你用机器学习的范式进行思考。你将学到机器学习的设计原理和相关模型。

第 2 章讲解了 Python 中众多针对机器学习任务的程序包。本章会让你初步了解一些大型库，包括 NumPy、SciPy、Matplotlib 和 Scilit-learn 等。

第 3 章讲解了原始数据可能有多种不同格式，其数量和质量也可能各不相同。有时，我们会被数据淹没；而有时，我们希望从数据中榨取最后一滴信息。数据要成为信息，需要有意义的结构。本章我们介绍了一些宽泛的主题，如大数据、数据属性、数据源、数据处理和分析等。

第 4 章在逻辑模型中探索了逻辑语言，并创建了假设空间映射；在树状模型中，我们发现其具有广泛作用域并易于描述和理解；在规则模型中，我们讨论了基于有序规则列表和无序规则集的模型。

第 5 章介绍了线性模型，它是使用最广泛的模型之一。线性模型是众多高级非线性技术的基础，例如，支持向量机（SVM）和神经网络。本章还研究了机器学习最常用的技术，创建线性回归和 logistic 回归的假设语句。

第 6 章介绍了机器学习最强大的人工神经网络算法。我们将看到这些网络如何成为大脑神经元的简化模型。

第 7 章讨论了特征的不同类型，即定量特征、有序特征和分类特征。我们还将详细学习如何结构化和变换特征。

第 8 章解释了集成机器学习背后的动机和成因，其来源于清晰的直觉并具有丰富的理论历史基础。集成机器学习的类型在于模型本身，以及围绕着三个主要问题（如何划分数据、如何选择模型、如何组合其结果）的考量。

第 9 章着眼于一些设计策略，以确保你的机器学习系统最优。我们将学习模型选择和参数调优技术，并将所学知识应用于一些案例研究之中。

阅读前的准备工作

你需要有学习机器学习的意愿，并需要下载安装 Python 3。Python 3 的下载地址是：https://www.python.org/downloads/ 。

本书的读者对象

本书的读者包括数据学家、科学家，或任何好奇的人。你需要具备一些线性代数和 Python 编程的基础，对机器学习的概念有基本了解。

CONTENTS

目　录

CHAPTER 1

第 1 章

机器学习的思维

机器学习系统具有意义深远且令人兴奋的能力，能够在各种应用领域为人们提供重要的洞察力；从具有开创性的挽救生命的医学研究到宇宙基础物理方面的发现，从为我们提供更健康、更清洁的食物到互联网分析和建立经济模型，等等。事实上，就某种意义而言，这项技术在我们的生活中已经无所不在。物联网的蔓延正产生着惊人的数据量，很显然，智能系统正以相当剧烈的方式改变着社会。Python 及其库等开源工具，以及以 Web 为代表的越来越多的开源知识库，使学习和应用这门技术有了新的和令人兴奋的途径，也使学习过程更为容易和廉价。本章将涵盖如下主题：

- 人机界面
- 设计原理
- 模型
- 统一建模语言

1.1 人机界面

如果你有幸用过微软 Office 套件的早期版本，你大概还能记得 Mr Clippy 办公助手。这一功能出现在 Office 97 中，每当你在文档开头输入“亲爱的”，它就会不请自来，从电脑屏幕的右下角蹦出来，询问“你好像在写信，需要帮助吗？”

在 Office 的早期版本中，Mr Clippy 是默认开启的，几乎被所有软件用户嘲笑过，这可以作为机器学习的第一次大败笔而载入史册。

那么，为什么这个欢乐的 Mr Clippy 会如此遭人痛恨呢？在日常办公任务中使用自动化助手不一定是个坏主意。实际上，自动化助手的后期版本，至少是最好的那几个，可以在后台无缝运行，并能明显提高工作效率。文本预测有很多例子，有些很搞笑，大错特错，但大多数并没有失败，它们悄无声息，已经成为我们正常工作流的一部分。

在这一点上，我们需要区分错误和失败的不同。Mr Clippy 的失败是因为它的突兀和差劲的设计，而它的预测并不一定是错误的；也就是说，它可能给出了正确的建议，但那时你已经知道你正在写一封信件。文本预测的错误率很高，经常会得出错误的预测，但这并没有失败，主要是因为它的失败方式被设计为悄无声息的。

设计任何与人机界面紧耦合（系统工程的说法）的系统都很困难。与一般的自然界事物一样，我们并非总能预测人类行为。表情识别系统、自然语言处理和手势识别技术等，开启了人机交互的新途径，对机器学习专家而言，所有这些都具有重要的应用。

每当设计需要人机输入的系统时，我们应当预见所有可能的人机交互方式，而不仅仅是我们所期望的那些方式。在本质上，我们对这些系统试图要做的是，培养它们对人类经验全景的一些理解。

在 Web 的早期，搜索引擎使用的是一种简单的系统，以文章中出现搜索条件的次数为基础。很快，Web 开发者就通过增加关键词与搜索引擎展开了博弈。显然，这将导致一场围绕关键词的竞赛，Web 将变得极为烦人。随后，为了提供更为准确的搜索结果，人们又设计了度量优质引用链接的页面排名系统。而今，现代搜索引擎都使用了更为复杂和秘密的算法。

对机器学习设计师同样重要的是，人机交互中所产生的数据量一直在增长。这会带

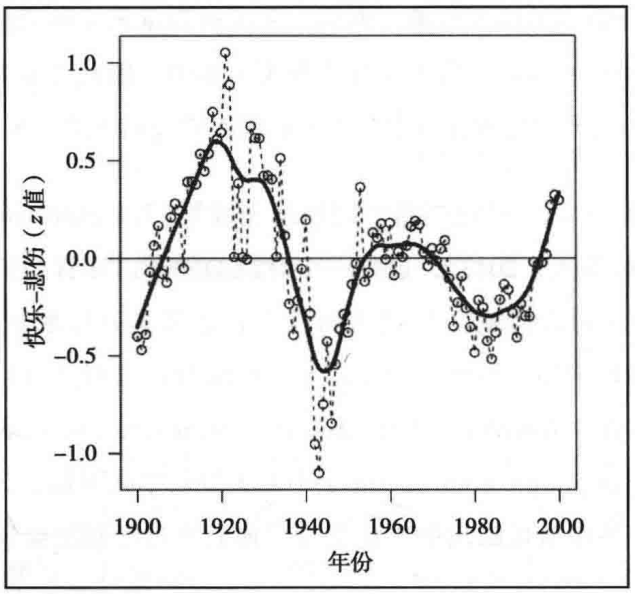
来诸多挑战，尤其是数据的庞大浩瀚。然而，算法的力量正是在于从海量数据中提取知识和洞察力，这对于较小规模的数据集几乎是不可能的。因此，如今大量的人机交互被数字化，而我们才刚刚开始理解和探索其中的数据能够被利用的众多途径。

有项研究的题目为《20 世纪书籍中的情绪表现》（*The expression of emotion in 20th century books*, Acerbi 等人，2013），这是一个有趣的例子。尽管严格地说，该研究属于数据分析而非机器学习，但就一些理由而言，它还是具有说明性的。该研究的目的是，从 20 世纪的书籍中抽取情绪内容文本，以情绪分值的形式进行图表化。通过访问 Gutenberg 数字图书馆、WordNet（http://wordnet.princeton.edu/wordnet/）和 Google 的 Ngram 数据库（books.google.com/ngrams）中的大量数字化书籍，该研究的作者能够绘制出 20 世纪文学作品中所反映出的文化变迁。他们通过绘制情绪词语使用的趋势来实现其研究目的。

在该研究中，作者对每个词语进行标记（1-gram 分词算法），并与情绪分值和出版年份进行关联。诸如快乐、悲伤、恐惧等情绪词语，可以依据其表达的正面或负面情绪进行评分。情绪分值可以从 WordNet（wordnet.princeton.edu）获得。WordNet 给每个情绪词语都赋予了情绪反应分值。最后，作者对每一情绪词语的出现次数进行了计数：

$$M = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{C_{the}}M_z = \frac{M-\mu_M}{\sigma_M}$$

在此式中，c_i 表示特定情绪词语的计数，n 表示情绪词语的总数（不是所有词语，仅包括具有情绪分值的词语），C_{the} 表示文本中 *the* 的计数。在归一化总和时，考虑到一些年份出版或数字化的书籍数量更多，同时晚期的书籍趋向于包含更多的技术语言，因此使用了词语 *the* 而不是所有词语的计数。对于在相当长的一段时期内的散文文本中的情绪，这种表示更为精确。最后，通过正态分布对分值进行归一化，即 M_z，减去均值后除以标准差。



上图摘自《20世纪书籍中的情绪表现》（*The expression of emotion in 20th century books, Alberto Acerbi, Vasileios Lampos, Phillip Garnett, R. Alexander Bentley*）美国科学公共图书馆。

这里，我们可以看到该项研究所生成的一张图表。该图显示了这一时期所著书籍的快乐－悲伤分值，从中可以明显看出二战时期的负面倾向。

这项研究之所以有趣，有如下一些原因。首先，它是一项数据驱动的科学研究，而在过去，类似的研究内容被认为是诸如社会学和人类学的软科学，但在该研究中，给出了坚实的实验基础。此外，尽管其研究结论令人印象深刻，但其实现过程相对容易。这主要得益于 WordNet 和 Google 已经完成的那些卓越努力。其亮点在于，如何使用互联网上免费的数据资源和软件工具，例如 Python 的数据和机器学习包等，任何具备数据技能和动机的人都能够从事这方面的研究。

1.2 设计原理

我们经常拿系统设计和其他事物的设计进行类比，例如建筑设计。在一定程度上，

这种类比是正确的，它们都是依据规格说明，在结构体中放置设计好的组件。但当我们考虑到它们各自的运行环境时，这种类比就会瓦解。在建筑设计上，通常会假设，当景观正确形成后就不会再改变。

软件环境则有些不同。系统是交互和动态的。我们设计的任何系统，诸如电子、物理，或人类，都会嵌入在其他系统中。同样，在计算机网络中有不同的层（应用层、传输层、物理层，等等），具有不同的含义和功能集，所以在项目中，需要在不同的层完成所需执行的活动。

作为这些系统的设计师，我们必须对其背景（即我们所工作的领域）具有强烈意识。领域知识能够赋予我们工作的背景，为我们在数据中发现模式提供线索。

机器学习项目可以分解为如下 6 项不同的活动：

- ❑ 定义目标和规格说明
- ❑ 准备和探索数据
- ❑ 建立模型
- ❑ 实现
- ❑ 测试
- ❑ 部署

设计师主要关注前三项活动。但是，他们通常需要在其他活动中扮演主要角色，并且在许多项目中必须如此。同时，这些活动在项目的时间表中不一定是线性序列。但重点是，这些都是明确不同的活动。这些活动可以并行进行，或者彼此相互作用，但通常会涉及不同类型的任务，在人力和其他资源、项目阶段和外在性上相互分离。而且，我们需要考虑到不同的活动所涉及的操作模式也彼此不同。想想看，我们在勾勒想法时、进行特定分析任务时，以及编写一段代码时，大脑工作方式的差异。

通常，最困难的是从何下手。我们可以先专研某一问题的不同要素，构思其中的特征集，或者考虑用什么模型。这样就能得出目标和规格说明的定义。或者我们可能不得