用 Python 分析数据、
预测结果的简单高效的方式

WILEY 异步图书 www.epubit.com.cn

Python 机器学习

预测分析核心算法

[美] Michael Bowles 著
沙赢 李鹏 译

MACHINE LEARNING IN PYTHON

ESSENTIAL TECHNIQUES
FOR PREDICTIVE ANALYSIS

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

WILEY

Python 机器学习

预测分析核心算法

[美] Michael Bowles 著
沙赢 李鹏 译

MACHINE LEARNING IN PYTHON

ESSENTIAL TECHNIQUES FOR PREDICTIVE ANALYSIS

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

Python机器学习 : 预测分析核心算法 / (美) 鲍尔斯 (Michael Bowles) 著 ; 沙嬴, 李鹏译. -- 北京 : 人民邮电出版社, 2017.1
ISBN 978-7-115-43373-2

Ⅰ. ①P… Ⅱ. ①鲍… ②沙… ③李… Ⅲ. ①软件工具－程序设计 Ⅳ. ①TP311.56

中国版本图书馆CIP数据核字(2016)第245035号

版权声明

◆ 著　[美] Michael Bowles
译　沙　嬴　李　鹏
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
北京艺辉印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：21
字数：445 千字　2017 年 1 月第 1 版
印数：1－3 000 册　2017 年 1 月北京第 1 次印刷

著作权合同登记号　图字：01-2015-4678 号

定价：69.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315

内容提要

在学习和研究机器学习的时候，面临令人眼花缭乱的算法，机器学习新手往往会不知所措。本书从算法和 Python 语言实现的角度，帮助读者认识机器学习。

本书专注于两类核心的“算法族”，即惩罚线性回归和集成方法，并通过代码实例来展示所讨论的算法的使用原则。全书共分为 7 章，详细讨论了预测模型的两类核心算法、预测模型的构建、惩罚线性回归和集成方法的具体应用和实现。

本书主要针对想提高机器学习技能的 Python 开发人员，帮助他们解决某一特定的项目或是提升相关的技能。

致我的孩子——Scott、Seth 和 Cayley，他们那如花的生命带给我世界上最美妙的时光。

致我亲密的朋友——David 和 Ron，他们给我带来了无私慷慨、坚定不移的友谊。

致我在加州山景城黑客道场的朋友和同事，他们给我带来了技术挑战以及机敏的应答。

致我的攀岩伙伴。正如一个伙伴凯瑟琳所说通过攀岩可以交上最好的朋友，因为“他们见证过因恐惧而僵硬的面庞，见证过相互鼓励搀扶，见证过攀登成功后的喜悦”。

作者简介

Michael Bowles 拥有机械工程学士和硕士学位、仪器仪表博士学位以及 MBA 学位。他的履历涉及学术界、工业界以及商业界。他目前在一家初创公司工作，其中机器学习技术至关重要。他是多个管理团队的成员、咨询师以及顾问。他也曾经在加州山景城的黑客道场、创业公司孵化器和办公场所教授机器学习课程。

他出生于俄克拉荷马州并在那里获得学士和硕士学位。在东南亚待了一段时间后，他前往剑桥攻读博士学位，毕业后任职于 MIT 的 Charles Stark Draper 实验室[①]。之后他离开波士顿前往南加州的休斯飞机公司开发通信卫星。在 UCLA 获得 MBA 学位后，他前往旧金山的湾区工作。作为创始人以及 CEO，他目前经营两家公司，这两家公司都已获风险投资。

他目前仍然积极参与技术以及创业相关的工作。近期项目包括使用机器学习技术进行自动交易，基于基因信息进行生物预测，使用自然语言处理技术进行网站优化，利用人口统计学及实验室数据预测医疗效果，在机器学习和大数据相关领域的公司里尽心尽责。可以通过 www.mbowles.com 联系到他。

[①] Charles Stark Draper 是一名美国科学家和工程师，被称为“惯性导航之父”。他是 MIT 仪器实验室的创始人，后来此实验室用其名来命名，此实验室设计了阿波罗登月计划中的阿波罗导航计算机。

技术编辑简介

Daniel Posner 拥有经济学学士以及硕士学位，在波士顿大学完成生物统计学博士学位。他曾为医药生物领域的公司以及帕罗奥图退伍军人事务部医院的研究人员提供统计学方面的咨询。

Daniel 和本书作者就书中相关主题有过广泛的合作。他们曾经一起撰写过开发 Web 级梯度提升算法的项目申请书。最近，他们合作利用随机森林和样条基扩展技术解决药物研制过程中的关键变量识别问题，其目标是提升预测效果以减少试验所需样本规模。

致谢

首先感谢 Wiley 出版社工作人员在本书创作期间提供的大量帮助。最早是组稿编辑 Robert Elliot 和我联系写作本书，他很容易相处。之后是 Jennifer Lynn 为本书的编辑。她对每个问题都积极响应，写作期间非常耐心地和我联系，保证我的写作计划如期完成。非常感谢你们的工作。

我也非常感谢如此敏锐、缜密的统计学家兼程序员 Daniel Posner 作为本书的技术编辑，从他那获得了巨大的安慰和帮助。感谢你的工作，也感谢你在机器学习、统计学以及算法上所做的有趣讨论。我还没见过其他谁的思维可以达到如此深入、如此迅捷。

前言

从数据中提取有助于决策的信息正在改变着现代商业的组织，同时也对软件开发人员产生了直接的影响。一方面是对新的软件开发技能的需求，市场分析师预计到 2018 年对具有高级统计和机器学习技术的人才需求缺口将达 140000 ～ 190000 人。这对具有上述技能的人员来说意味着丰厚的薪水和可供选择的多种有趣的项目。另一方面对开发人员的影响就是逐步出现了统计和机器学习相关的核心工具，这减轻了开发人员的负担。当他们尝试新的算法时，不需要重复发明“轮子”。在所有通用计算机语言中，Python 的开发人员已经站在了最前沿，他们已开发了当前最先进的机器学习工具包，但从拥有这些工具包到如何有效地使用它们仍然存在一定的距离。

开发人员可以通过在线课程、阅读质量上乘的书籍等方式来获得机器学习的相关知识。它们通常都对机器学习的算法、应用实例给出了精彩的阐述，但是因为当前算法如此之多，以至于很难在一门课程或一本书中覆盖全部算法的相关细节。

这给实践者带来了困难。当面临众多算法时，机器学习新手可能需要多次尝试才能做出决定，这往往需要开发人员来填补从问题提炼到最终问题解决之间的所有算法使用方便的细节。

本书尝试填补这一鸿沟，所采用的方法就是只集中于两类核心的“算法族”，这两类算法族已在广泛的应用领域中证明了其最佳的性能。此论断的证据如下：这两类算法在众多机器学习算法竞争者中已获得支配性地位，新开发的机器学习工具包都会率先支持此两类算法，以及研究工作给出的性能对比结论（见第 1 章）。重点关注这两类算法使我们可以更详细地介绍算法的使用原则，并通过一系列的示例细节来展示针对不同问题如何使用这些算法。

本书主要通过代码实例来展示所讨论的算法的使用原则。以我在加州山景城的黑客道场（Hacker Dojo）授课的经验来看，我发现开发人员更愿意通过直接看代码示例来了解算法原理，而不是通过数学公式推导。

本书使用 Python 语言，因为它能提供将功能和专业性良好结合的机器学习算法包。Python 是一种经常使用的计算机语言，以产生精炼、可读性代码而著称。这导致目前已有相当数量的业界旗舰公司采用 Python 语言进行原型系统开发和部署。Python 语言开发人员获得了广泛的支持，包括大量的同业开发人员组成的社区、各种开发工具、扩展库等等。Python 广泛应用于企业级应用和科学研究领域。它有相当数量的工具包支持计算密集型应用，如机器学习。它也收集了当前机器学习领域的代表性算法（这样就可以省去重复性劳动）。Python 相比专门的统计语言如 R 或 SAS（Statistical Analysis System）是一门更通用的语言，它的机器学习算法包吸收了当前一流的算法，并且在一直扩充。

本书的目标读者

本书主要面向想提高机器学习技能的 Python 开发人员，不管是针对某一特定的项目，还是只想提升相关技能。开发人员很可能在工作中遇到新问题需要使用机器学习的方法来解决。当今机器学习的应用领域如此之广，使其已成为简历中一项十分有用的技能。

本书为 Python 开发人员提供如下内容：

- 机器学习所解决的基本问题的描述；
- 当前几种最先进的算法；
- 这些算法的应用原则；
- 一个机器学习系统的选型、设计和评估的流程；
- 流程、算法的示例；
- 可进一步修改扩展的代码。

为了能够顺利地理解这本书，读者所需的背景知识包括：了解编程、能够读写代码。因为本书的代码示例、库、包都是 Python 语言的，所以本书主要适用于 Python 开发人员。本书通过运行算法的核心代码来展示算法的使用原则，然后使用含有此算法的工具包来展示如何应用此算法来解决问题。开发人员通过源代码可以获得对算法的直观感受，就像其他人通过数学公式的推导来掌握算法。一旦掌握了算法的核心原理，应用示例就直接使用 Python 工具包，这些工具包都包含了能够有效使用这些算法必需的辅助模块（如错误检测、输入输出处理、模型所需数据结构的处理、基于训练模型的预测方法的处理，等等）。

除了编程背景，懂得相关数学、统计的知识将有助于掌握本书的内容。相关数学知识包括大学本科水平的微分学（知道如何求导，少量线性代数知识）、矩阵符号的意义、矩阵乘、求逆矩阵。这些知识主要是帮助理解一些算法中的求导部分，很多情况下就是一个简单函数的求导或基本的矩阵操作。能够理解概念层面上的数学计算将有助于对算法的理解。明白推导各步的由来有助于理解算法的强项和弱项，也帮助读者面对具体的问题时，决定哪个算法是最佳选择。

本书也用到了一些概率和统计知识。对这方面的要求包括熟悉大学本科水平的概率知识和概念，如实数序列的均值、方差和相关系数。当然即使这些知识对读者来说有些陌生，也不会影响读者对代码的理解。

本书涵盖了机器学习算法的两大类：惩罚线性回归（penalized linear regression），如岭回归算法（ridge）、Lasso 算法；集成方法（ensemble methods），如随机森林算法（random forests）、梯度提升算法（gradient boosting）。上述两大类算法有一些变体，都可以解决回归和分类的问题（在本书开始部分将会介绍分类和回归的区别）。

如果读者已熟悉机器学习并只对其中的一类算法感兴趣，可以直接跳到相关的二章。每类算法由两章组成，一章介绍基本原理，另外一章介绍针对不同类型问题的用法。惩罚线性回归由下列两章组成：第 4 章“惩罚线性回归模型”和第 5 章“利用惩罚线性回归方法来构建预测模型”；集成方法由下列两章组成：第 6 章“集成方法”和第 7 章“用 Python 构建集成模型”。快速浏览第 2 章“通过理解数据来了解问题”将有助于理解算法应用章节中的问题。刚刚进入机器学习领域准备从头到尾通读的读者可以把第 2 章留到阅读那些算法应用章节前。

本书包含的内容

如上所述，本书涵盖两大类算法，这些算法近期都获得了发展，并将仍然获得持续性研究，它们都起源于早期的技术，但已使这些早期技术黯然失色。

惩罚线性回归代表了对最小二乘法回归方法（least squares regression）的相对较新的改善和提高。惩罚线性回归具有的几个特征使其成为预测分析的首选。惩罚线性回归引入了一个可调参数，使最终的模型在过拟合与欠拟合之间达到了平衡。它还提供不同的输入特征对预测结果的相对贡献的信息。上述这些特征对于构建预测模型都是十分重要的。而且，对于某些问题惩罚线性回归可以产生最佳的预测性能，特别是对于欠定的问题以及具有很多输入参数的问题，如基因领域、文本挖掘等。进一步，坐标下降法（coordinate descent methods）等新方法可以使惩罚线性回归模型训练过程运行得更快。

为了帮助读者更好地理解惩罚线性回归，本书也概要介绍了线性回归及其扩展，如逐步回归（stepwise regression），主要是希望能够培养读者对算法的直观感受。

集成方法是目前最有力的预测分析工具之一。它可以对特别复杂的行为进行建模，特别是过定的问题，通常这些都是与互联网有关的预测问题（如返回搜索结果和预测点击率）。由于集成方法的性能，许多经验丰富的数据科学家在做第一次尝试时都使用该方法。集成方法使用相对简单，而且可以依据对预测的贡献程度对输入特征排序。

目前集成方法与惩罚线性回归齐头并进。然而惩罚线性回归是从克服一般回归方法的局限性进化而来的，集成方法是从克服二元决策树的局限性进化而来的。因此本书介绍集成方法时，也会涉及二元决策树的背景知识，因为集成方法继承了二元决策树的一些属性。了解这些将有助于培养对集成方法的直觉。

本书的组织

本书遵循了着手解决一个预测问题的基本流程。开始阶段包括对数据的理解、如何形式化表示问题，然后开始尝试使用算法解决问题，评估其性能。在这个过程中，本书将

概要描述每一步采用的方法及其原因。第 1 章给出本书涵盖的问题和所用方法的完整描述，本书使用来自 UC Irvine 数据仓库的数据集作为例子；第 2 章展示了一些数据分析的方法和工具，帮助读者对新数据集具有一定的洞察力。第 3 章“预测模型的构建：平衡性能、复杂性以及大数据”主要介绍由上述三者带给预测分析技术的困难以及所采用的技术，勾勒了问题复杂度、模型复杂度、数据规模和预测性能之间的关系，讨论了过拟合问题以及如何可靠地感知到过拟合，以及不同类型问题下的性能评价标准。第 4 章、第 5 章分别介绍惩罚线性回归的背景及其应用，即如何解决第 2 章所述的问题。第 6 章、第 7 章分别介绍集成方法的背景及其应用。

如何使用本书

为了运行书中的代码示例，需要有 Python2.x、SciPy、NumPy、Pandas 和 scikit-learn。由于交叉依赖和版本的问题，这些软件的安装可以会有些困难。为了简化上述软件安装过程，可以使用来自 Continuum Analytics（http://continuum.io/）的这些包的免费分发版。Continuum Analytics 提供的 Anaconda 软件包可自由下载并且包含 Python2.x 在内的运行本书代码所需的全部软件包。我在 Ubuntu 14.04 Linux 发行版上测试了本书的代码，但是没有在其他的操作系统上测试过。

约定

为了便于对本书的理解，本书通篇采用如下的约定。

警告 这些方框表示是与其周围文本直接相关的不能忘记的重要信息。

注释 表示与当前讨论相关的注释、说明、提示和技巧。

源代码

研究本书示例代码时，可以选择手工敲入这些代码，也可以直接使用随书带的源代码文件。本书用到的所有源代码都可以从 http://www.wiley.com/go/pythonmachinelearning 中下载获得。在本书代码片段旁会有一个下载的小图标，并注明文件名，这样就可以知道此文件在下载的源代码中，并且可以很轻松地从下载源代码中找到此文件。读者可访问上述网站，定位到本书书名（可以使用搜索框或书名列表），在本书详细介绍页面点击“Download Code”链接就可以获得本书的全部源代码。

注释 因为很多书的书名都非常相似，最简单的方法就是通过 ISBN 来查找，本书的 ISBN 是 978-1-118-96174-2。

下载源代码后，只需用你惯用的解压缩工具解压缩即可。

勘误表

我们已尽最大可能避免在文本和代码中出现错误。但是没有人是完美的，同样错误也难免会发生。如果读者在书中发现了错误，诸如拼写错误、代码错误等等，能及时反馈，我们将非常感谢。提交勘误表，能减少其他读者的困惑，同时也帮助我们提供更高质量的内容。

获得本书勘误表的方法：访问 http://www.wiley.com，通过搜索框或书名列表定位到本书；然后进入本书的详细介绍页面，点击“Book Errata”链接，可以看到关于本书所有已提交的并由 Wiley 出版社编辑上传的勘误表。

目录

第 1 章
关于预测的两类核心算法

本书集中于机器学习领域，只关注那些最有效和获得广泛使用的算法。不会提供关于机器学习技术领域的全面综述。这种全面性的综述往往会提供太多的算法，但是这些算法并没有在从业者中获得积极的应用。

本书涉及的机器学习问题通常是指“函数逼近（function approximation）”问题。函数逼近问题是有监督学习（supervised learning）问题的一个子集。线性回归和逻辑回归是解决此类函数逼近问题最常见的算法。函数逼近问题包含了各种领域中的分类问题和回归问题，如文本分类、搜索响应、广告放置、垃圾邮件过滤、用户行为预测、诊断等。这个列表几乎可以一直列下去。

从广义上说，本书涵盖了解决函数逼近问题的两类算法：惩罚线性回归和集成方法。本章将介绍这些算法，概述它们的特性，回顾算法性能对比研究的结果，以证明这些算法始终如一的高性能。

然后本章讨论了构建预测模型的过程，描述了这里介绍的算法可以解决的问题类型，以及如何灵活地构建问题模型，选择用于做预测的特征。本章也描述了应用算法的具体步骤，包括预测模型的构建、面向部署的性能评估等。

1.1 为什么这两类算法如此有用

有几个因素造就了惩罚线性回归和集成方法成为有用的算法集。简单地说，面对实践中遇到的绝大多数预测分析（函数逼近）问题，这两类算法都具有最优或接近最优的性能。这些问题包含：大数据集、小数据集、宽数据集（wide data sets）[①]、高瘦数据集（tall skinny data sets）[②]、复杂问题、简单问题，等等。Rich Caruana 及其同事的两篇论文为上述论断提供了证据。

1．“An Empirical Comparison of Supervised Learning Algorithms,” Rich Caruana, Alexandru Niculescu-Mizi。

① 宽数据集（wide data set）指每次观测时有大量的测量项，但是观测次数有限的数据。若把数据看成表格形式，则此类数据集列数很多，而行数有限。典型的此类数据集包括神经影像、基因组以及其他生物医学方面的。——译者注。

② 高瘦数据集（tall skinny data set）指每次观测时测量项有限，但是进行了大量的观测。若把数据看成表格的形式，则此类数据集列数有限，行数很多。典型的此类数据集包括临床试验数据、社交网络数据等。——译者注。