HZ BOOKS
华章IT

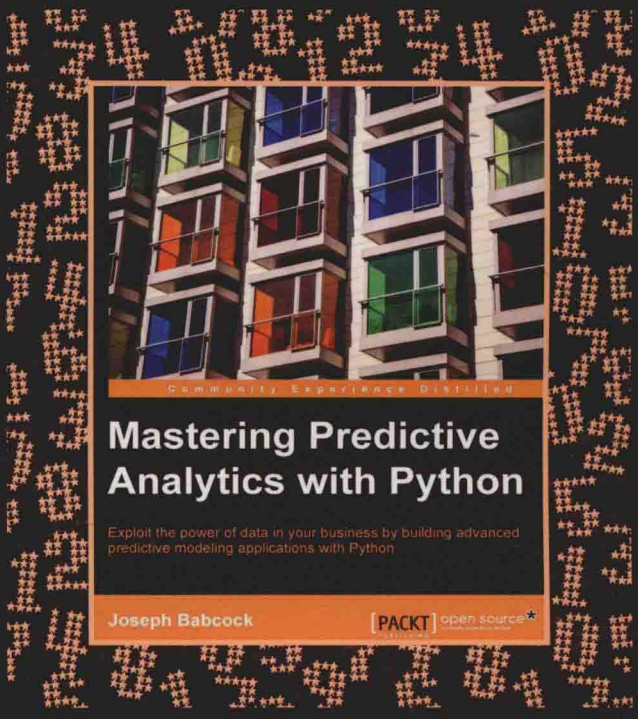
[PACKT]
PUBLISHING

数据科学与工程技术丛书

预测分析

Python语言实现

[美] 约瑟夫·巴布科克(Joseph Babcock) 著
余水清 译



MASTERING PREDICTIVE ANALYTICS WITH PYTHON

机械工业出版社
China Machine Press

数据科学与工程技术丛书

MASTERING PREDICTIVE ANALYTICS WITH PYTHON

预测分析

Python语言实现

[美] 约瑟夫·巴布科克(Joseph Babcock) 著
余水清 译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

预测分析：Python 语言实现 /（美）约瑟夫·巴布科克（Joseph Babcock）著；余水清译．
—北京：机械工业出版社，2017.6
（数据科学与工程技术丛书）
书名原文：Mastering Predictive Analytics with Python

ISBN 978-7-111-57389-0

I. 预… II. ① 约… ② 余… III. 软件工具—程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2017）第 165467 号

本书版权登记号：图字：01-2016-8644

本书着重介绍预测性分析技术，先概述了数据分析系统的基本架构和主要处理流程，然后从分类和无监督学习开始，逐一讲解每种机器学习算法的工作原理，并在每章的最后给出了详细的案例讨论。高质量的数据是能够进行正确分析的前提，为了便于后期分析模型的构建，本书还会介绍对于不同类型数据的清洗和过滤等内容。通过学习本书的内容，读者将了解将原始数据转化为重要结论的过程，并掌握快速将其中涉及的模型应用到自有数据中的方法。

预测分析：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：吴晋瑜　　责任校对：李秋荣
印　　刷：中国电影出版社印刷厂　　版　　次：2017 年 8 月第 1 版第 1 次印刷
开　　本：185mm × 260mm　1/16　　印　　张：13.25
书　　号：ISBN 978-7-111-57389-0　　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

本书不仅涵盖了分类、回归、聚类等众多算法以及诸如深度学习等前端技术，还讲解了这些方法的工作原理，以及如何在实践中应用它们。你将学到如何针对特定的问题选择正确的解决方法，以及如何开发具有吸引力的可视化界面，从而将预测性模型中所获得的洞见应用在实际工作中。

全书共9章，第1章介绍一个分析管道的核心组件以及它们之间的交互方式；第2章讨论着手搭建分析型应用所需完成的工作；第3章演示如何将一个数据集里的相似项定义成组；第4章探讨几种回归模型的拟合，包括将输入参数调整到正确数值范围并对类别特征做出正确说明；第5章阐述如何使用分类模型并介绍几种提升模型性能的策略；第6章讨论非结构化数据，涉及文本数据、图像数据及降维技术等；第7章介绍如何用深度神经网络来处理复杂数据；第8章讲述一个基本预测服务所包含的三个组件——客户端、服务器和Web应用；第9章介绍对预测模型的性能进行监控的若干策略。

作者简介

约瑟夫·巴布科克（Joseph Babcock）

现为AQR Capital Management机器学习研究员，之前曾是Netflix高级数据科学家。他有近10年的复杂数据集研究经验，解决了来自医疗健康和娱乐行业的众多大数据挑战。他毕业于美国约翰·霍普金斯大学医学院，获得了该校所罗门·斯奈德神经系统学科的博士学位，在该校就读期间，他运用机器学习预测了毒品对心脏方面的副作用。

华章IT
HZBOOKS | Information Technology

译 者 序

2016 年初发生的一件事，让人们对于人工智能的迅速发展刮目相看，即由谷歌开发的人工智能机器人程序 AlphaGo 战胜了人类围棋世界冠军职业九段棋手李世石，最终比分为 4∶1。在此之后，AlphaGo 一路过关斩将，与世界上数十位围棋高手对战，竟然无一失败，可谓是战绩辉煌。

毫无疑问，在围棋领域，AlphaGo 已经明显超越了一般人类的智能水平。它的设计和开发涉及了很多的核心技术，深度学习就是其中之一。所谓深度学习，就是指多层的人工神经网络以及对神经网络进行训练和优化的方法。前面一层神经网络把大量数据矩阵化之后作为输入，通过非线性激活方法计算权重，再产生另一个数据集合，以此作为输出，同时也作为后面一层神经网络的输入。这个过程就像生物神经大脑的工作机理一样，通过合适的矩阵数量，多层组织链接在一起，形成神经网络“大脑”，对数据进行精准复杂的处理，既使得处理速度得以提高，又使得处理的准确度和精确度也达到了一个非常高的水平。

深度学习和人工神经网络是典型的无监督学习算法之一，其他典型的无监督学习算法还包括分组和聚类，在这些算法当中，由于不需要事先有典型的输入和输出作为训练样本，因此具有极大的灵活性。此外还有分类和回归等有监督学习算法和半监督学习方法。这些机器学习算法在大数据处理方面有着各自的应用场景，如果应用得当，将会使大数据处理如虎添翼。

AlphaGo 只是机器学习技术与大数据相结合的典型案例之一。移动互联网和物联网（IOT）的飞速发展，使人类社会累积数据的速度达到前所未有的程度，这里所谓的大数据包括几乎一切形式的结构化、半结构化以及非结构化的数据，例如网络日志、音频、视频、图片、地理位置信息等。因此，如何对这些数据进行高效的采集、存储、处理并从中发掘到有价值的信息，就是大数据分析处理需要解决的问题。具体来说就是时下比较热门的几个技术热点：物联网和工业物联网（IIoT），主要解决的是数据的采集问题；云计算，主要解决高效存储和计算问题；数据分析技术，主要解决如何有效对数据进行挖掘，并从中发现应用价值的问题。本书主要关注数据分析技术之一，即预测性分析技术，以及如何将分析结果以可视化的方式展现给利益相关者。当然，数据能够进行正确分析的前提是有高质量的数据，因此，本书也会适当地提及对于不同类型的数据如何进行清洗和过滤，从而为后期分析模型的构建打好基础。

说到大数据分析，不得不提及当前对大数据进行处理分析的编程语言，首屈一指的当属 R 了。作为 MATLAB 和 SAS 等昂贵而复杂的统计软件的免费替代品，简单易用

的 R 迅速风靡全球，在金融街的表现尤为突出。但是 R 的优点也许也是其缺点，例如在建模技术上并不是很完善，处理海量数据时显得有些笨重，等等。而 Python 结合了 R 语言的快速性、处理复杂数据的能力以及更务实的语言特质，迅速地成为主流，尤其是近几年成长得很快。由于直观、易于学习，以及生态系统近年来急剧增长，Python 在统计分析领域迅速占有一席之地。IPython notebook 和 NumPy 可以用作快速进行数据分析和处理的一种轻便工具，而 Python 可以作为中等规模数据处理的强大工具。丰富的数据社区也是 Python 的优势，因为可供用户随时获取到大量的工具包和功能。另外，Python 也可以与多种关系型数据库（例如 MySQL、PostGreSQL 等）和非关系型数据库（例如 MongoDB、Hadoop 等）进行无缝集成，再加上对于分布式计算框架（例如 Spark）的支持，让基于 Python 构建的数据分析和预测系统可以很容易地扩展到大规模数据集上。因此，基于 Python 构建大数据的分析预测系统无疑是比较好的解决方案之一。

回到本书的主要内容上，本书先对数据分析系统的基本架构和主要处理流程进行了扼要介绍，然后从分类和无监督学习开始，逐一讲解每种机器学习算法的主要工作原理，并且在每章的最后一节给出详细的案例讨论，从而将理论很好地落实到实现中。通过对每一种具体学习算法理论部分的介绍，我们可以窥见本书作者 Joseph Babcock 坚实的理论功底；而落实到具体的 Python 实现上，我们又可以体会到 Python 的强大和简洁。因此，我觉得对数据分析、数据建模或者机器学习感兴趣的读者都可以读一读本书。但是译者个人觉得，读者最好具备一定的统计学知识和基本的 Python 编程经验，不然会稍显吃力。

值得一提的是，机器学习和人工智能已经写进了 2017 年的政府工作报告，因此从某种程度上说，已经上升到了国家战略层面。我们有理由相信，在可以预见的未来，基于机器学习和人工智能的数据处理和分析预测技术必将为人们的生活和工作带来极大的改变。因此，本书可以为那些对机器学习算法和预测分析技术感兴趣的人指出大致的方向。而对于那些早已投身其中的先行者来说，本书也不失为一个比较好的参考。

对我本人而言，本书的翻译过程也是一次对于机器学习知识的重新学习和对于预测分析技术的全面梳理。感谢华章公司和静编辑的信任和支持，同时也感谢家人的理解和支持。由于译者水平有限，如有不当之处，尚请广大读者批评指正。

余水清

2017 年 3 月

关于审稿人

迪普尼詹·德比（Dipanjan Deb）是一位有着16年丰富经验的数据分析专家，他在机器/统计学习、数据挖掘以及预测分析方面均有深厚的经验，所涉及的领域横跨医疗健康、航海、汽车自动化、能源、消费品（CPG），以及人力资源等。他擅长使用开源和商业版软件设计开发最前沿的分析解决方案，以实现大规模的并行和系统优化。

Dipanjan擅长将数据科学家们组建成分析团队，并交付高质量的解决方案。他与行业专家、技术专家以及数据科学家通力合作，制订策略构建分析型解决方案，以缩短从概念验证（POC）到商业发布的转换过程。

他擅长实现基于R、Python、Vowpal Wabbit、Julia以及SAS的全监督、半监督或无监督学习算法，精通自行管理和云环境下的Hadoop和Spark分布式框架。他是一个业余Kaggler开发者，同时也是一位IoT/IIoT（物联网/工业物联网）爱好者（树莓派以及Arduino原型设计）。

前 言

通过学习本书，你将逐步掌握将原始数据转化为重要结论的过程。本书所涉及的大量案例学习和代码样例，均使用现下流行的开源 Python 库，阐述了分析应用完整的开发过程。详细的案例讲述了常见应用场合下健壮、可扩展的应用。你将学会如何快速将这些模型应用到自己的数据中去。

本书内容

第 1 章讲述了如何描述一个分析管道中的核心组件以及组件间的交互方式，也探讨了批处理和流处理之间的区别，以及每种应用最适用的一些情况，还讲解了基于两种范式的基础应用样例以及每一步所需的设计决策。

第 2 章讨论了着手搭建分析型应用所需完成的诸多工作。运用 IPython notebook，我们讨论了如何使用 pandas 将文件中的数据上传到数据帧中、重命名数据集中的列名、过滤掉不想要的行、转换类型以及创建新的列。另外，我们将整合不同来源的数据，并使用聚合和旋转进行一些基本的统计分析。

第 3 章将演示如何将一个数据集里的相似项定义成组。这种探索性分析是我们在理解新数据集过程中经常第一个使用的。我们探索计算数据点值间相似性的不同方法，并描述这些度量可能最适合于哪些数据。我们既探讨分裂聚类算法（将数据分解成一组一组更小的部分），也探讨凝聚聚类算法（每个数据点都是一个聚类的开始）。通过一系列数据集，我们将展示每种算法在哪些情景下性能更好或者更差，以及如何优化它们。我们也看到了首个（比较小的）数据管道——PySpark 中基于流数据的聚类应用。

第 4 章探讨了几种回归模型拟合模型，包括将输入参数调整到正确数值范围并对类别特征做出正确说明。我们对线性回归进行拟合、评估，也包括正则化回归模型。我们还研究树回归模型的用处，以及如何优化参数选项来拟合模型。最后，我们会讨论一个基于 PySpark 的简单随机森林模型，该模型也可以用于更大的数据集。

第 5 章阐述了如何使用分类模型并介绍几种提升模型性能的策略。除了转换类别特征之外，我们讨论了如何利用 ROC 曲线对逻辑回归准确性进行解释。为了尝试提升模型的性能，我们讲解了 SVM 的用处。最后，我们将使用梯度提升决策树算法，以期在测试数据集上可以取得较好的性能。

第 6 章讨论复杂的、非结构化的数据。其中还涉及了降维技术（例如 HashingVectorizer）、矩阵分解（例如 PCA、CUR 和 NMR）以及概率模型（例如 LDA），讨论了图

像数据，包括标准化操作和阈值转换操作，并介绍如何使用降维技术找出图像之间的共同模式。

第 7 章介绍了将深度神经网络作为一种生成模型的方法，来处理那些工程师难以处理其特征的复杂数据。我们将研究如何使用反向传播训练神经网络，并探究附加层难以达到最优的原因。

第 8 章描述了一个基本预测服务的三个组件，并探讨这种设计如何使我们与其他用户或者软件系统分享预测模型的结果。

第 9 章介绍几个监控初步设计后预测模型性能的策略。我们也会讨论一些模型的性能或组件会随时间变化的场景。

阅读准备

你需要安装好最新版的 Python、PySpark 以及 Jupyter notebook。

读者人群

本书主要针对业务分析员、BI 分析员、数据科学家，或是一些已经掌握高级分析员理论知识的初级数据分析员。通过阅读本书，上述读者将可以运用 Python 设计并构建高级分析解决方案。读者必须具备基础 Python 开发经验。

本书约定

在本书中，你会发现很多用以区别不同信息的文本样式。以下是一些文本样式的例子，以及每种样式所代表含义的解释。

正文中的代码、数据库表名、文件夹名称、文件名、文件扩展名、路径名、虚拟 URL、用户输入，以及 Twitter 用户名等均以以下模式展现：“使用 `head()` 和 `tail()` 来查看以下数据的开头和结尾。”

任何命令行的输入或者输出都会采用以下形式：

```
rdd_data.coalesce(2).getNumPartitions()
```

新名词和**重要文字**会以加粗格式给出。屏幕上的文字（例如菜单或者对话框）在文中以如下形式展现：“回到**文件**标签栏，你会注意到在右上角有两个选项。”

表示警示或重要提醒。

表示提示和技巧。

下载样例代码

你可以用自己的账户登录 http://www.packtpub.com 下载本书上的样例代码文

件。如果你是通过其他途径购买本书，可以访问 http://www.packtpub.com/support，注册账户申请这些文件。

你也可以访问华章官网 http://www.hzbook.com，通过注册并登录个人账号，下载本书的代码。

下载本书彩图

我们也提供本书的 PDF 文件，支持彩色版的截屏/图表。这些彩色图片会帮助你更好地理解输出的变化。文件下载地址：https://www.packtpub.com/sites/default/files/downloads/MasteringPredictiveAnalyticswithPython_ColorImages.pdf。

目　录

第1章
数据转换成决策——从分析应用着手

从季度财务预测到客户调查，数据分析帮助企业做出明智决策并制订未来工作计划。电子数据表中饼图和趋势图之类的数据可视化工具已经使用了数十年之久，而近几年，商业分析中可用的数据源的体量和多样性都大幅提升，与此同时，用来诠释数据中的信息的工具也日益成熟。

随着网络的迅猛发展，电子商务和社交媒体平台产生了丰富的数据，分析应用可以比以往更快速地获得这些数据。图片、搜索词条和网络论坛帖子都是典型的非结构化数据，简单使用传统的电子数据表程序无法处理这些数据。如果使用正确的工具，无论是否结合传统数据，这类数据都将为企业提供新的洞见。

通常，历史客户记录这样的结构化表单数据存储在电子数据仓库中，而且很容易导入电子表单程序中。即便是这样的表单数据，很多行业内数据的数据量和更新速度都在增长。即使分析员曾经通过交互式操作转换过原始数据，强大而可靠稳健的分析方法迫切需要自动化的处理，才能适应企业接收到的数据量和速度。

研究数据的方法正变得越来越强大和复杂。根据输入一些可变参数得出的趋势表安排未来工作或者总结历史模式，高级分析学强调运用复杂的预测模型（参见本章的“预测分析的目标”）理解现状，并预测近期和未来的结果。

用以实现上述预测效果的各种方法通常需要以下几个基本元素：

- 试图预测的结果或目标，例如购买交易或搜索结果中的**点击率**（CTR）。
- 一组包含了**特征值**的数据列——也被称为**预测因子**（例如，客户的人口统计信息、交易账户的历史事务，或者广告的点击方式），用以描述数据集中每条记录的单个属性（例如，账户或广告）。
- 能发现单个或整套模型的程序，并把这些特征恰到好处地映射到给定样本数据所关注的结果上。
- 一种在新数据上评估模型性能的方法。

将预测建模技术运用到强大的分析应用中，以发掘看似毫无关系的输入数据之间的复杂关系，这也给商业分析师们带来了一组新的挑战：

- 对于一个特定问题，最佳处理方法是什么？
- 如何在历史数据和新数据上正确地评估这些技术的性能？
- 调整某个特定方法的性能的首选策略是什么？
- 如何稳定地扩展这些技术，以同时涵盖一次性分析和持续性分析需求？

本书将会展示如何通过设计分析型解决方案，将企业的业务数据转化为强大的洞见，进而有效地应对上述挑战。构建这些分析应用需要完成以下几个主要任务：

- 将原始数据转换成模型需要的净化格式。这包括清洗异常数据以及将非结构化数据转换成结构化数据。
- 特征工程，将这些净化后的输入数据转换成设计预测模型所需的格式。
- 通过数据子集校准预测模型并评估其性能。
- 评估模型的持续性能时对新数据评分。
- 对于定期更新，自动实现转换和建模工作。
- 将模型的输出信息展现给其他系统和用户——通常通过 Web 应用实现。
- 为分析员和企业用户生成报告，从数据和模型中提取有规律的、强大的洞见。

贯穿本章始终，我们会使用基于 Python 编程语言的开源工具来搭建这些种类的应用。为什么是 Python 呢？因为 Python 语言很好地掌握了健壮的编译语言诸如 Java、C ++ 和 Scala 以及单纯的统计程序包（例如 R、SAS 或是 MATLAB）之间的优美平衡。交互式使用 Python 的方式有几种：使用命令行（抑或后面章节用到的基于浏览器的 notebook 环境）、绘制图形以及原型设计指令。Python 还提供扩展库，支持我们将探索性研究转换成 Web 应用（例如 Flask、CherryPy 和 Celery——在第 8 章会讨论到），或者扩展运用到大数据集上（使用 PySpark——在后面章节里研究）。因此，使用同一种语言，我们既可以分析数据，也可以开发软件应用。

在深入学习这些工具的技术细节之前，我们先在更高的层面上看看这些应用背后的概念和技术架构。

本章将介绍以下内容：

- 定义一个分析管道的基本元素，包括数据转换、完整性检查、预处理、模型开发、评估、自动化、部署和报告。
- 解释面向批处理和流处理的区别及其在开发管道过程中每一步的含义。
- 检验如何实现批处理和流处理在 Lambda 体系架构下数据处理的完美结合。
- 研究运用流处理管道执行社交媒体数据的情感分析案例。
- 研究运用批处理管道产生有针对性的电子邮件营销活动样例。

预测分析的目标

预测分析这一术语以及诸如**数据挖掘**和**机器学习**等术语，在本书中经常被用来描述搭建分析方案中的相关技术。然而，有一点需要谨记的是，这些方法可以用于实现两个完全不一样的目标。推理包括搭建模型以评估不同参数对不同结果的意义，强调结果的意义和透明度而不是预测性能。举个例子，回归模型的系数（第 4 章）被用来评估对特定模型输入（例如，客户年龄或收入）输出变量（例如，销售量）对应变化的影响。推理类模型生成的预测结果的准确性可能比其他技术差一些，但其提供的有价值的概念性洞见也许可以指导企业决策。相反，预测强调估测结果的准确性，即便模型本身是一个黑匣子，输入和结果输出之间的关联经常不明显。例如，深度学习（第 7 章）能产生最新模型，并能非常精确地给出复杂输入集的预测结果，但是输入参数之间的关联性以及预测结果很难解释清楚。