

Python数据挖掘入门与实践

作者：[澳] Robert Layton

版权信息

书名：Python数据挖掘入门与实践

作者：[澳] Robert Layton

译者：杜春晓

ISBN：978-7-115-42710-6

本书由北京图灵文化发展有限公司发行数字版。版权所有，侵权必究。

您购买的图灵电子书仅供您个人使用，未经授权，不得以任何方式复制和传播本书内容。

我们愿意相信读者具有这样的良知和觉悟，与我们共同保护知识产权。

如果购买者有侵权行为，我们可能对该用户实施包括但不限于关闭该帐号等维权措施，并可能追究法律责任。

图灵社区会员 憨豆战神（vikecom@163.com）专享 尊重版权

[版权声明](#)

[译者序](#)

[前言](#)

[本书主要内容](#)

[本书的阅读前提](#)

[本书的目标读者](#)

[排版约定](#)

[读者反馈](#)

[客户支持](#)

[下载示例代码](#)

[下载配套PDF文件](#)

[勘误表](#)

[侵权](#)

[问题](#)

[第1章 开始数据挖掘之旅](#)

[1.1 数据挖掘简介](#)

[1.2 使用Python和IPython Notebook](#)

[1.2.1 安装Python](#)

[1.2.2 安装IPython](#)

[1.2.3 安装scikit-learn库](#)

[1.3 亲和性分析示例](#)

[1.3.1 什么是亲和性分析](#)

[1.3.2 商品推荐](#)

[1.3.3 在NumPy中加载数据集](#)

[1.3.4 实现简单的排序规则](#)

[1.3.5 排序找出最佳规则](#)

[1.4 分类问题的简单示例](#)

[1.5 什么是分类](#)

[1.5.1 准备数据集](#)

[1.5.2 实现OneR算法](#)

[1.5.3 测试算法](#)

[1.6 小结](#)

[第2章 用scikit-learn估计器分类](#)

[2.1 scikit-learn估计器](#)

[2.1.1 近邻算法](#)

[2.1.2 距离度量](#)

[2.1.3 加载数据集](#)

[2.1.4 努力实现流程标准化](#)

[2.1.5 运行算法](#)

[2.1.6 设置参数](#)

[2.2 流水线在预处理中的应用](#)

[2.2.1 预处理示例](#)

[2.2.2 标准预处理](#)

[2.2.3 组装起来](#)

[2.3 流水线](#)

[2.4 小结](#)

[第3章 用决策树预测获胜球队](#)

[3.1 加载数据集](#)

[3.1.1 采集数据](#)

[3.1.2 用pandas加载数据集](#)

[3.1.3 数据集清洗](#)

[3.1.4 提取新特征](#)

[3.2 决策树](#)

[3.2.1 决策树中的参数](#)

[3.2.2 使用决策树](#)

[3.3 NBA比赛结果预测](#)

[组装起来](#)

[3.4 随机森林](#)

[3.4.1 决策树的集成效果如何](#)

[3.4.2 随机森林算法的参数](#)

[3.4.3 使用随机森林算法](#)

[3.4.4 创建新特征](#)

[3.5 小结](#)

[第4章 用亲和性分析方法推荐电影](#)

[4.1 亲和性分析](#)

[4.1.1 亲和性分析算法](#)

[4.1.2 选择参数](#)

[4.2 电影推荐问题](#)

[4.2.1 获取数据集](#)

[4.2.2 用pandas加载数据](#)

[4.2.3 稀疏数据格式](#)

[4.3 Apriori算法的实现](#)

[4.3.1 Apriori算法](#)

[4.3.2 实现](#)

[4.4 抽取关联规则](#)

[评估](#)

[4.5 小结](#)

[第5章 用转换器抽取特征](#)

[5.1 特征抽取](#)

[5.1.1 在模型中表示事实](#)

[5.1.2 通用的特征创建模式](#)

[5.1.3 创建好的特征](#)

[5.2 特征选择](#)

[选择最佳特征](#)

[5.3 创建特征](#)

[主成分分析](#)

[5.4 创建自己的转换器](#)

[5.4.1 转换器API](#)

[5.4.2 实现细节](#)

[5.4.3 单元测试](#)

[5.4.4 组装起来](#)

[5.5 小结](#)

[第6章 使用朴素贝叶斯进行社交媒体挖掘](#)

[6.1 消歧](#)

[6.1.1 从社交网站下载数据](#)

[6.1.2 加载数据集并对其分类](#)

[6.1.3 Twitter数据集重建](#)

[6.2 文本转换器](#)

[6.2.1 词袋](#)

[6.2.2 N元语法](#)

[6.2.3 其他特征](#)

[6.3 朴素贝叶斯](#)

[6.3.1 贝叶斯定理](#)

[6.3.2 朴素贝叶斯算法](#)

[6.3.3 算法应用示例](#)

[6.4 应用](#)

[6.4.1 抽取特征](#)

[6.4.2 将字典转换为矩阵](#)

[6.4.3 训练朴素贝叶斯分类器](#)

[6.4.4 组装起来](#)

[6.4.5 用F1值评估](#)

[6.4.6 从模型中获取更多有用的特征](#)

[6.5 小结](#)

[第7章 用图挖掘找到感兴趣的人](#)

[7.1 加载数据集](#)

[7.1.1 用现有模型进行分类](#)

[7.1.2 获取Twitter好友信息](#)

[7.1.3 构建网络](#)

[7.1.4 创建图](#)

[7.1.5 创建用户相似度图](#)

[7.2 寻找子图](#)

[7.2.1 连通分支](#)

[7.2.2 优化参数选取准则](#)

[7.3 小结](#)

[第8章 用神经网络破解验证码](#)

[8.1 人工神经网络](#)

[神经网络简介](#)

[8.2 创建数据集](#)

[8.2.1 绘制验证码](#)

[8.2.2 将图像切分为单个的字母](#)

[8.2.3 创建训练集](#)

[8.2.4 根据抽取方法调整训练数据集](#)

[8.3 训练和分类](#)

[8.3.1 反向传播算法](#)

[8.3.2 预测单词](#)

[8.4 用词典提升正确率](#)

[8.4.1 寻找最相似的单词](#)

[8.4.2 组装起来](#)

[8.5 小结](#)

[第9章 作者归属问题](#)

[9.1 为作品找作者](#)

[9.1.1 相关应用和使用场景](#)

[9.1.2 作者归属](#)

[9.1.3 获取数据](#)

[9.2 功能词](#)

[9.2.1 统计功能词](#)

[9.2.2 用功能词进行分类](#)

[9.3 支持向量机](#)

[9.3.1 用SVM分类](#)

[9.3.2 内核](#)

[9.4 字符N元语法](#)

[抽取字符N元语法](#)

[9.5 使用安然公司数据集](#)

[9.5.1 获取安然数据集](#)

[9.5.2 创建数据集加载工具](#)

[9.5.3 组装起来](#)

[9.5.4 评估](#)

[9.6 小结](#)

[第10章 新闻语料分类](#)

[10.1 获取新闻文章](#)

[10.1.1 使用Web API获取数据](#)

[10.1.2 数据资源宝库reddit](#)

[10.1.3 获取数据](#)

[10.2 从任意网站抽取文本](#)

[10.2.1 寻找任意网站网页中的主要内容](#)

[10.2.2 组装起来](#)

[10.3 新闻语料聚类](#)

[10.3.1 k-means算法](#)

[10.3.2 评估结果](#)

[10.3.3 从簇中抽取主题信息](#)

[10.3.4 用聚类算法做转换器](#)

[10.4 聚类融合](#)

[10.4.1 证据累积](#)

[10.4.2 工作原理](#)

[10.4.3 实现](#)

[10.5 线上学习](#)

[10.5.1 线上学习简介](#)

[10.5.2 实现](#)

[10.6 小结](#)

[第11章 用深度学习方法为图像中的物体进行分类](#)

[11.1 物体分类](#)

[11.2 应用场景和目标](#)

[使用场景](#)

[11.3 深度神经网络](#)

[11.3.1 直观感受](#)

[11.3.2 实现](#)

[11.3.3 Theano简介](#)

[11.3.4 Lasagne简介](#)

[11.3.5 用nolearn实现神经网络](#)

[11.4 GPU优化](#)

[11.4.1 什么时候使用GPU进行计算](#)

[11.4.2 用GPU运行代码](#)

[11.5 环境搭建](#)

[11.6 应用](#)

[11.6.1 获取数据](#)

[11.6.2 创建神经网络](#)

[11.6.3 组装起来](#)

[11.7 小结](#)

[第 12 章 大数据处理](#)

[12.1 大数据](#)

[12.2 大数据应用场景和目标](#)

[12.3 MapReduce](#)

[12.3.1 直观理解](#)

[12.3.2 单词统计示例](#)

[12.3.3 Hadoop MapReduce](#)

[12.4 应用](#)

[12.4.1 获取数据](#)

[12.4.2 朴素贝叶斯预测](#)

[12.5 小结](#)

[附录 接下来的方向](#)

[第 1 章——开始数据挖掘之旅](#)

[Scikit-learn教程](#)

[扩展IPython Notebook](#)

[第 2 章——用scikit-learn估计器分类](#)

[k近邻算法的扩展](#)

[更多复杂的流水线](#)

[比较分类器](#)

[第 3 章——用决策树预测获胜球队](#)

[pandas的更多内容](#)

[更多复杂特征](#)

[第 4 章——用亲和性分析方法推荐电影](#)

[新数据集](#)

[Eclat算法](#)

[第 5 章——用转换器抽取特征](#)

[增加噪音](#)

[Vowpal Wabbit](#)

[第 6 章——使用朴素贝叶斯进行社交媒体挖掘3](#)

[垃圾信息监测](#)

[自然语言处理和词性标注](#)

[第 7 章——用图挖掘找到感兴趣的人](#)

[更复杂的算法](#)

[NetworkX](#)

[第 8 章——用神经网络破解验证码](#)

[好（坏？）验证码](#)

[深度网络](#)

[增强学习](#)

[第 9 章——作者归属问题](#)

[增加数据量](#)

[博客语料](#)

[局部N元语法](#)

[第 10 章——新闻语料分类](#)

[算法评价](#)

[近期趋势分析](#)

[实时聚类](#)

[第 11 章——用深度学习方法为图像中的物体进行分类](#)

[Keras和Pyleam2](#)

[Mahotas](#)

[第 12 章——大数据处理](#)

[Hadoop课程](#)

[Pydoop](#)

[推荐引擎](#)

[更多资源](#)

版权声明

Copyright © 2015 Packt Publishing. First published in the English language under the title *Learning Data Mining with Python*.

Simplified Chinese-language edition copyright © 2016 by Posts & Telecom Press. All rights reserved.

本书中文简体字版由Packt Publishing授权人民邮电出版社独家出版。未经出版者书面许可，不得以任何方式复制或抄袭本书内容。

版权所有，侵权必究。

译者序

云计算、大数据、物联网，这几年很火。到现在为止，人们对云计算的激情已经回落到比较理智的水平，各种云基础设施已投入使用，支撑起关系国计民生的信息化应用。物联网还在建设中，家电智能化、个人健康信息数字化、交通智能化等趋势在我们身边悄然推进。开放互联的概念不再囿于传统的互联网思维，我们生活所触及的一切正在被编织到一张包罗万象的大网之中。它将会对社会产生何种影响，我们拭目以待。虽然大数据现在很火，各种大数据研究中心相继建立，但这只是刚刚开始。随着更多的人和设备接入互联网，随着人们对世界认识的加深和新工具的研发，数据规模将加速膨胀，超乎想象。大数据的春天才刚刚到来。数据采集能力上去之后，势必要求数据挖掘能力跟得上。正如作者在第12章中讲到的，大数据带来的一个挑战就是，重要信息可能被垃圾信息湮没。由此我们不难推断出数据挖掘技术在发现、突显和传承人类文明方面将起到不可替代的作用。本书讲解的正是大数据时代的核心技能——数据挖掘技术，可以预见该项技术将发挥出越来越重要的作用。

本书讲解了如何用Python语言进行数据挖掘。Python是一种通用型编程语言，它简单易学，上手快，有着丰富的第三方库，社区氛围友好。从数据采集、分析一直到应用开发层面，Python都有成熟的库。使用Python语言进行开发，无需过多关注语言细节，开发者可以将主要精力放到业务本身。书中使用IPython Notebook作为开发环境，它将代码执行、富文本、公式编辑、绘图、多媒体等功能集合在一起，是科学计算和数据分析的好工具。书中所涉及的数据挖掘对象很丰富，有Iris鸢尾花卉数据集、Ionosphere电离层数据集、NBA比赛结果、MovieLens电影评分数据集、古登堡计划所收集的图书、安然公司邮件数据集、博客语料、CIFAR-10图像数据集等。从这些分属于不同行业的数据集，也能一窥数据挖掘应用之广。此外，作者还介绍了从Twitter、Reddit网站采集数据的方法。在算法方面，除了常见的决策树、朴素贝叶斯、支持向量机等，作者还介绍了最近几年非常热的深度学习。大数据、深度学习对计算能力要求很高，作者介绍了如何在亚马逊云主机上运行MapReduce任务。这本书由浅入深，以真实数据为研究对象，逐渐增大数据集规模，真刀实枪地向读者介绍了Python数据挖掘是怎么回事，并给读者进一步学习指出了多种可能的方向。工程实践之余，作者还不忘介绍数据挖掘常用思路，毫不保留地把自己积攒的宝贵经验传授给读者。这一点我在阅读过程中，深有体会。正如作者自己在前言里所写的，书中不会涉及大量公式推导，所有的算法都是以很直观的形式向读者介绍，所以即使你缺乏一定的数学基础，只要肯用功，也不用担心自己读不懂。

回到七八年前，当我还是一名英语专业学生的时候，我压根不会想到有一天会学编程，会去翻译这样一本书。后来有幸读了计算机辅助翻译这样一个专业，才开始接触到计算机知识，但是当父亲跟我提起数据建模时，我还是一脸茫然。研究生几年，系里为我们这些非计算机背景的学生开设了Python编程课。从那时起我就有事没事学点Python，一开始是照着*Natural Language Processing with Python*的示例敲，自那时起五年之后我竟想起给NLTK提交几处微小的改动。大约是为了激励我这个后生继续为他们服务，“居心叵测”的Steven Bird竟把我加入到贡献者名单里。去中关村图书大厦的时候，我常常喜欢浏览一下语言与程序设计书架上有没有关于Python的新书，碰到喜欢的就翻翻看，这几年眼看着Python书多了起来，很是欣慰。此外，我去北大、北外旁听过计算语言学、概率统计等课程，去北航旁听过计算机系统基础，看过Udacity的统计学入门和吴恩达老师的机器学习课程视频，兴致来了也曾捧着Rosen的《离散数学及其应用》读上几页。工作中，经常帮同事写个简单的Python程序处理数据，最近还帮他们爬取了一个网站。PyCon北京，我连着去了三四届了，每次都有或多或少的收获，2015年我见过一位大神行云流水般演示用pandas处理数据，很受震撼。以上就是我与Python、数据挖掘的交集。我想说的是，不要再用上学时读的那个专业的思维局限自己的发展，学科的界限在模糊，融合的趋势在增强，数学的重要性在提升。提到数学，今天还看了一个TED演讲视频，说的是借助计算机改变传统数学教育方法。这种理念什么时候能应用到一线教学，非常值得期待。生在这个充满变革的时代，倍感幸运。

以我有限的水平去翻译这样一本书，心里不免焦虑。遇到问题，我四处寻找能人相助。感谢作者Robert Layton，我每次发邮件向他求教或确认问题，他总能很快地回复我，有时第二天回复，还会说抱歉回复晚了。感谢我的同学黄子轩、孙伟、周星，他们在我学习计算机科学的路上给予了很多指导和帮助。翻译本书时，我还向子轩求证作者在第6章给出的示例是否有误。感谢上海大学研究生钱亦欣同学，他帮我审读了第3章，并给出若干很专业的修改意见，上述第6章那处问题，我也曾向他请教过，最终证明是原书弄错了。感谢李少华，他帮我弄明白了Python在Windows系统中的环境变量设置方法。感谢陈健锁，我曾就数据库相关术语向他请教。感谢图灵的朱巍编辑，是她促成了我最终去翻译这样一本趣味盎然的书！感谢图灵公司，我是你们忠实的读者！最后感谢我的妻子，她承担了照顾女儿的重任。女儿的出生，让我惊叹于生命的奇妙！感谢岳父岳母一家人帮忙照看孩子，我才有时间去做翻译。感谢我的父亲、姐姐，他们以我翻译本书为骄傲。

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

杜春晓

2016年1月3日

欢迎访问：电子书学习和下载网站 (<https://www.shgis.cn>)

文档名称：Python数据挖掘入门与实践 - [澳] Robert Layton.epub

请登录 <https://shgis.cn/post/1829.html> 下载完整文档。

手机端请扫码查看：

