大数据人才培养规划教材

以解决实际问题为**学习目标**
以实战案例贯穿为**学习手段**

Python
数据分析与应用

Python Data Analysis and Application

黄红梅 张良均 ◉ 主编
张凌 施兴 周东平 ◉ 副主编

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

大数据人才培养规划教材

Python
数据分析与应用

Python Programming

张

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

Python数据分析与应用 / 黄红梅，张良均主编. -- 北京 : 人民邮电出版社，2018.4
大数据人才培养规划教材
ISBN 978-7-115-37304-5

Ⅰ. ①P… Ⅱ. ①黄… ②张… Ⅲ. ①软件工具—程序设计 Ⅳ. ①TP311.561

中国版本图书馆CIP数据核字(2017)第311638号

内 容 提 要

本书以任务为导向，全面地介绍数据分析的流程和 Python 数据分析库的应用，详细讲解利用 Python 解决企业实际问题的方法。全书共 9 章，第 1 章介绍了数据分析的基本概念等相关知识；第 2～6 章介绍了 Python 数据分析的常用库及其应用，涵盖 NumPy 数值计算、Matplotlib 数据可视化、pandas 统计分析、使用 pandas 进行数据预处理、使用 scikit-learn 构建模型，较为全面地阐述了 Python 数据分析方法；第 7～9 章结合之前所学的数据分析技术，进行企业综合案例数据分析。除第 1 章外，本书各章都包含了实训与课后习题，通过练习和操作实践，帮助读者巩固所学的内容。

本书可以作为高校大数据技术类专业的教材，也可作为大数据技术爱好者的自学用书。

◆ 主　　编　黄红梅　张良均
副 主 编　张　凌　施　兴　周东平
责任编辑　左仲海
责任印制　马振武

◆ 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：787×1092　1/16
印张：19.25　　2018 年 4 月第 1 版
字数：440 千字　　2018 年 4 月北京第 1 次印刷

定价：49.80 元

读者服务热线：(010)81055256　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

大数据专业系列图书
编写委员会

序 PREFACE

随着大数据时代的到来，移动互联网络和智能手机迅速普及，多种形态的移动互联应用蓬勃发展，电子商务、云计算、互联网金融、物联网等不断渗透并重塑传统产业，大数据当之无愧地成为了新的产业革命核心。

未来 5～10 年，我国大数据产业将会是一个飞速发展时期，社会对大数据相关专业人才有着巨大的需求。目前，国内各大高校都在争相设立或准备设立大数据相关专业，以适应地方产业发展对战略性新兴产业的人才需求。

人才培养离不开教材，大数据专业是 2016 年才获批的新专业，目前还没有成套的系列教材，已有教材也存在企业案例缺失等亟须解决的问题。由广州泰迪智能科技有限公司和人民邮电出版社策划，校企联合编写的这套图书，尤如大旱中的甘露，可以有效解决高校大数据相关专业教材紧缺的困境。

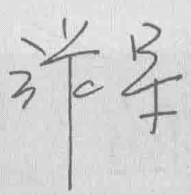
实践教学是在一定的理论指导下，通过引导学习者的实践活动，从而传承实践知识、形成技能、发展实践能力、提高综合素质的教学活动。目前，高校教学体系的设置有诸多限制因素，过多地偏向理论教学，课程设置与企业实际应用切合度不高，学生无法把理论转化为实践应用技能。课程内容设置方面看似繁多又各自为“政”，课程冗余、缺漏，体系不健全。本套图书的第一大特点就是注重学生的实践能力培养，根据高校实践教学中的痛点，首次提出“鱼骨教学法”的概念。以企业真实需求为导向，学生学习技能紧紧围绕企业实际应用需求，将学生需掌握的理论知识，通过企业案例的形式进行衔接，达到知行合一、以用促学的目的。

大数据专业应该以大数据技术应用为核心，紧紧围绕大数据应用闭环的流程进行教学，才能够使学生从宏观上理解大数据技术在行业中的具体应用场景及应用方法。高校现有的大数据课程集中在如何进行数据处理、建模分析、调整参数，使得模型的结果更加准确。但是，完整的大数据应用却是一个容易被忽视的部分。本套图书的第二大特点就是围绕大数据应用的整个流程，从数据采集、数据迁移、数据存储、数据

分析与挖掘，最终到数据可视化，覆盖完整的大数据应用流程，涵盖企业大数据应用中的各个环节，符合企业大数据应用真实场景。

我很高兴看到这套书的出版，也希望这套书能给更多的高校师生带来教学上的便利，帮助读者尽快掌握本领，成为有用之才！

教育部长江学者特聘教授
国家杰出青年基金获得者
IEEE Fellow
华南理工大学计算机与工程学院院长



2017 年 12 月

前 言 FOREWORD

随着云时代的来临，数据分析技术将帮助企业用户在合理时间内获取、管理、处理以及整理海量数据，为企业经营决策提供积极的帮助。数据分析作为一门前沿技术，广泛应用于物联网、云计算、移动互联网等战略性新兴产业。虽然大数据目前在国内还处于初级阶段，但是其商业价值已经显现出来，特别是有实践经验的数据分析人才更是各企业争夺的热门。为了满足日益增长的数据分析人才需求，很多高校开始尝试开设不同程度的数据分析课程。“数据分析”作为大数据时代的核心技术，必将成为高校大数据相关专业的重要课程之一。

本书特色

本书以任务为导向，结合大量数据分析工程案例及教学经验，以 Python 数据分析常用技术和真实案例相结合的方式，深入浅出地介绍使用 Python 进行数据分析及应用的重要内容。除第 1 章外，本书各章都由任务描述、任务分析、任务实现、实训和课后练习等部分组成。设计思路以应用为导向，让读者明确如何利用所学知识来解决问题，通过实训和课后练习巩固所学知识，读者真正理解并能够应用所学知识。本书的内容由浅入深，第 1 章介绍数据分析的基本概念等相关知识，让读者在宏观上理解数据分析能够帮助我们解决什么问题。第 2～6 章结合具体的任务，介绍了 Python 数据分析常用的 NumPy、Matplotlib、pandas、scikit-learn 库的应用。第 7～9 章是前面几章内容基础上的综合应用，包括航空公司客户价值分析、财政收入预测分析、家用热水器用户行为分析与事件识别这 3 个综合案例，帮助读者搭建一条最佳的数据分析学习路线图。

本书适用对象

- 开设有数据分析课程的高校的教师和学生。

目前国内不少高校将数据分析引入教学中，在数学、计算机、自动化、电子信息、金融等专业开设了与数据分析技术相关的课程，但目前这一课程的教学仍然主要限于理论介绍。因为单纯的理论教学过于抽象，学生理解起来往往比较困难，教学效果也不甚理想。本书提供的基于实战案例和建模实践的教学模式，能够使师生充分发挥互动性和创造性，获得最佳的教学效果。

- 需求分析及系统设计人员。

这类人员可以在理解数据分析原理及建模过程的基础上，结合数据分析案例完成

精确营销、客户分群、交叉销售、流失分析、客户信用记分、欺诈发现、智能推荐等数据分析应用的需求分析和设计。

- 数据分析应用的开发人员。

这类人员可以在理解数据分析应用需求和设计方案的基础上，结合图书提供的基于第三方接口快速完成数据分析应用的编程实现。

- 进行数据分析应用研究的科研人员。

许多科研院所为了更好地对科研工作进行管理，纷纷开发了适应自身特点的科研业务管理系统，并在使用过程中积累了大量的科研信息数据。但是，这些科研业务管理系统一般没有对这些数据进行深入分析，对数据所隐藏的价值并没有充分分析利用。科研人员需要数据分析工具及有关方法来深挖科研信息的价值，从而提高科研水平。

代码下载及问题反馈

为了帮助读者更好地使用《Python 数据分析与应用》这本书，我们配套提供了原始数据文件和程序代码，读者可以从“泰迪杯”数据挖掘挑战赛网站（http://www.tipdm.org/tj/1266.jhtml）免费下载，也可登录人民邮电出版社教育社区下载（http://www.ryjiaoyu.com）。另外，为方便教师授课，我们还提供了 PPT 课件，读者可以从“泰迪杯”数据挖掘挑战赛网站（http://www.tipdm.org/tj/840.jhtml）下载申请表，填写后发送至指定邮箱；其他图书资源，读者可通过热线电话（40068-40020）或以下微信公众号咨询获取。

我们已经尽最大努力避免在文本和代码中出现错误，但是由于水平有限，编写时间仓促，书中难免出现一些疏漏和不足的地方。如果您有更多的宝贵意见，欢迎发送邮件至邮箱 13560356095@qq.com，期待能够得到您真挚的反馈。同时，本书内容更新将及时在“泰迪杯”全国数据挖掘挑战赛网站上发布，读者可以登录网站或关注泰迪大数据挖掘微信公众号（TipDataMining）查阅相关信息。

编　者

2017 年 10 月

目录 CONTENTS

第 1 章 Python 数据分析概述

当今社会，网络和信息技术开始渗透进人类日常生活的方方面面，产生的数据量也呈现指数型增长的态势。现有数据的量级已经远远超越了目前人力所能处理的范畴。如何管理和使用这些数据，逐渐成为数据科学领域中一个全新的研究课题。Python 语言在最近十年发展迅猛，大量的数据科学领域的从业者使用 Python 完成数据科学相关的工作，其中最突出的就是数据分析师。

学习目标

（1）掌握数据分析的概念与流程。
（2）了解数据分析的应用场景。
（3）了解 Python 在数据分析领域的优势。
（4）了解 Python 数据分析常用类库。
（5）掌握 Windows/Linux 系统中 Anaconda 的安装。
（6）掌握 Jupyter Notebook 的常用功能。

任务 1.1 认识数据分析

任务描述

数据分析作为大数据技术的重要组成部分，近年来随着大数据技术逐渐发展和成熟。数据分析技能，被认为是数据科学领域中数据从业人员需要具备的技能之一。与此同时，数据分析师也成了时下最热门的职业之一。数据分析技能的掌握是一个循序渐进的过程。明确数据分析概念、分析流程和分析方法等相关知识是迈出数据分析的第一步。

任务分析

（1）掌握广义的数据分析和狭义的数据分析的概念。
（2）掌握典型的数据分析流程。
（3）了解七大类常见的数据分析应用场景。

1.1.1 掌握数据分析的概念

数据分析是指用适当的分析方法对收集来的大量数据进行分析，提取有用信息和形成结论，对数据加以详细研究和概括总结的过程。随着计算机技术的全面发展，企业生产、收集、存储和处理数据的能力大大提高，数据量与日俱增。而在现实生活中，需要把这些繁多、复杂的数据通过统计分析进行提炼，以此研究出数据的发展规律，进而帮助企业管理层做出决策。

广义的数据分析包括狭义数据分析和数据挖掘。狭义的数据分析是指根据分析目的，采用对比分析、分组分析、交叉分析和回归分析等分析方法，对收集的数据进行处理与分析，提取有价值的信息，发挥数据的作用，得到一个特征统计量结果的过程。数据挖掘则是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，通过应用聚类模型、分类模型、回归和关联规则等技术，挖掘潜在价值的过程。

图 1-1 所示为广义数据分析的概念。广义数据分析是指依据一定的目标，通过统计分析、聚类、分类等方法发现大量数据中的目标隐含信息的过程。

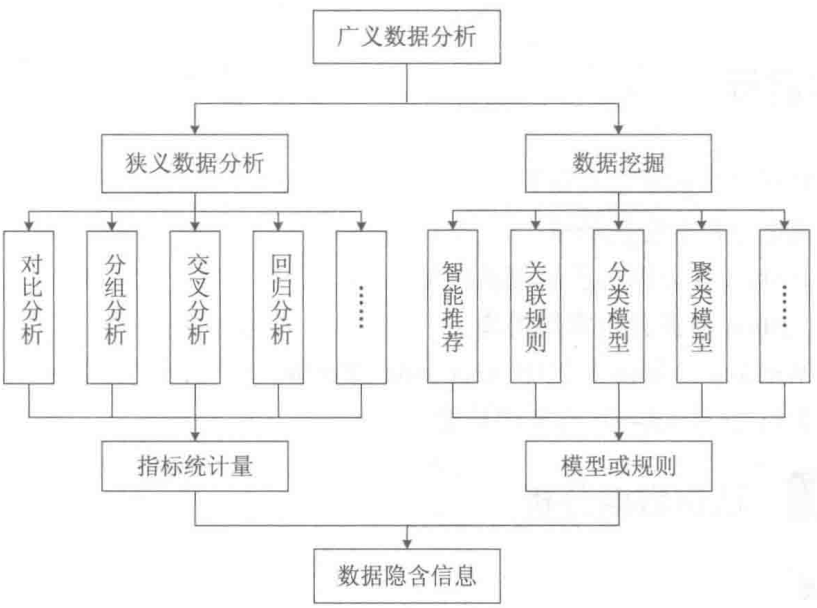


图 1-1 广义数据分析的概念

1.1.2 掌握数据分析的流程

数据分析已经逐渐演化为一种解决问题的过程，甚至是一种方法论。虽然每个公司都会根据自身需求和目标创建最适合的数据分析流程，但数据分析的核心步骤是一致的。图 1-2 所示是一个典型的数据分析的流程。

1. 需求分析

需求分析一词来源于产品设计，主要是指从用户提出的需求出发，挖掘用户内心的真实意图，并转化为产品需求的过程。产品设计的第一步就是需求分析，也是最关键的一步，因为需求分析决定了产品方向。错误的需求分析可能导致在产品实现过程中走入错误方向，甚至对企业造成损失。

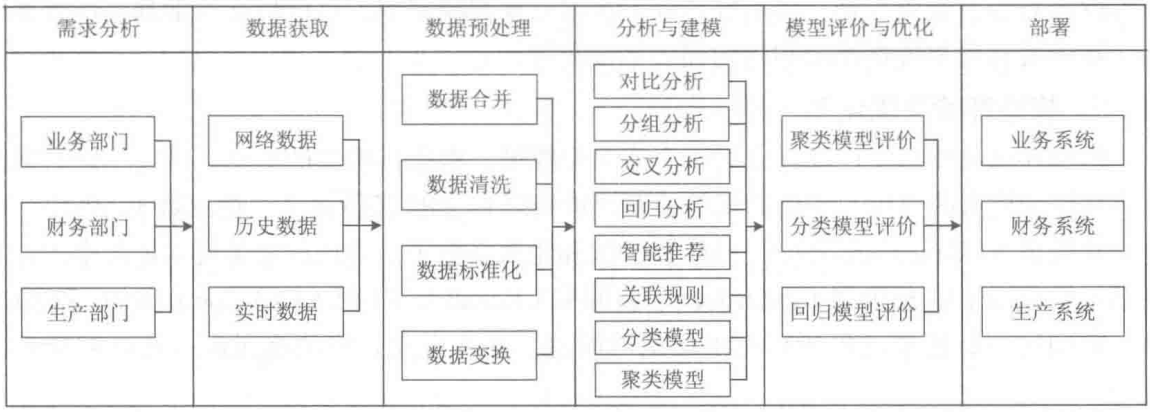


图 1-2 数据分析流程

数据分析中的需求分析是数据分析环节的第一步，也是非常重要的一步，决定了后续的分析方向和方法。数据分析中的需求分析的主要内容是，根据业务、生产和财务等部门的需要，结合现有的数据情况，提出数据分析需求的整体分析方向、分析内容，最终和需求方达成一致意见。

2. 数据获取

数据获取是数据分析工作的基础，是指根据需求分析的结果提取、收集数据。数据获取主要有两种方式：网络数据与本地数据。网络数据是指存储在互联网中的各类视频、图片、语音和文字等信息；本地数据则是指存储在本地数据库中的生产、营销和财务等系统的数据。本地数据按照数据时间又可以划分为两部分：历史数据与实时数据。历史数据是指系统在运行过程中遗存下来的数据，其数据量随系统运行时间的增加而增长；实时数据是指最近一个单位时间周期（月、周、日、小时等）内产生的数据。

在数据分析过程中，具体使用哪种数据获取方式，依据需求分析的结果而定。

3. 数据预处理

数据预处理是指对数据进行数据合并、数据清洗、数据标准化和数据变换，并直接用于分析建模的这一过程的总称。其中，数据合并可以将多张互相关联的表格合并为一张；数据清洗可以去掉重复、缺失、异常、不一致的数据；数据标准化可以去除特征间的量纲差异；数据变换则可以通过离散化、哑变量处理等技术满足后期分析与建模的数据要求。在数据分析的过程中，数据预处理的各个过程互相交叉，并没有明确的先后顺序。

4. 分析与建模

分析与建模是指通过对比分析、分组分析、交叉分析、回归分析等分析方法，以及聚类模型、分类模型、关联规则、智能推荐等模型与算法，发现数据中的有价值信息，并得出结论的过程。

分析与建模的方法按照目标不同可以分为几大类。如果分析目标是描述客户行为模式的，可采用描述型数据分析方法，同时还可以考虑关联规则、序列规则和聚类模型等。如果分析目标是量化未来一段时间内某个事件发生概率的，则可以使用两大预测分析模型，即分类预测模型和回归预测模型。在常见的分类预测模型中，目标特征通常都是二元数据，

例如欺诈与否、流失与否、信用好坏等。在回归预测模型中，目标特征通常都是连续型数据，常见的有股票价格预测和违约损失率预测等。

5. 模型评价与优化

模型评价是指对于已经建立的一个或多个模型，根据其模型的类别，使用不同的指标评价其性能优劣的过程。常用的聚类模型评价指标有 ARI 评价法（兰德系数）、AMI 评价法（互信息）、V-measure 评分、FMI 评价法和轮廓系数等。常用的分类模型评价指标有准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 值（F1 Value）、ROC 和 AUC 等。常用的回归模型评价指标有平均绝对误差、均方误差、中值绝对误差和可解释方差值等。

模型优化则是指模型性能在经过模型评价后已经达到了要求，但在实际生产环境应用过程中，发现模型的性能并不理想，继而对模型进行重构与优化的过程。在多数情况下，模型优化和分析与建模的过程基本一致。

6. 部署

部署是指将数据分析结果与结论应用至实际生产系统的过程。根据需求的不同，部署阶段可以是一份包含了现状具体整改措施的数据分析报告，也可以是将模型部署在整个生产系统的解决方案。在多数项目中，数据分析师提供的是一份数据分析报告或者一套解决方案，实际执行与部署的是需求方。

1.1.3 了解数据分析应用场景

企业使用数据分析解决不同的问题，实际应用的数据分析场景主要分为以下 7 类。

1. 客户分析（Customer Analytics）

客户分析主要是根据客户的基本数据信息进行商业行为分析，首先界定目标客户，根据客户的需求、目标客户的性质、所处行业的特征以及客户的经济状况等基本信息，使用统计分析方法和预测验证法分析目标客户，提高销售效率。其次了解客户的采购过程，根据客户采购类型、采购性质进行分类分析，制定不同的营销策略。最后还可以根据已有的客户特征进行客户特征分析、客户忠诚度分析、客户注意力分析、客户营销分析和客户收益分析。通过有效的客户分析能够掌握客户的具体行为特征，将客户细分，使得运营策略达到最优，提升企业整体效益等。

2. 营销分析（Sales and Marketing Analytics）

营销分析囊括了产品分析、价格分析、渠道分析、广告与促销分析这 4 类分析。产品分析主要是竞争产品分析，通过对竞争产品的分析制定自身产品策略。价格分析又可以分为成本分析和售价分析。成本分析的目的是降低不必要的成本；售价分析的目的是制定符合市场的价格。渠道分析是指对产品的销售渠道进行分析，确定最优的渠道配比。广告与促销分析则能够结合客户分析，实现销量的提升、利润的增加。

3. 社交媒体分析（Social Media Analytics）

社交媒体分析是以不同的社交媒体渠道生成的内容为基础，实现不同社交媒体的用户分析、访问分析和互动分析等。用户分析主要根据用户注册信息、登录平台的时间点和平

时发表的内容等用户数据，分析用户个人画像和行为特征；访问分析则是通过用户平时访问的内容分析用户的兴趣爱好，进而分析潜在的商业价值；互动分析根据互相关注对象的行为预测该对象未来的某些行为特征。同时，社交媒体分析还能为情感和舆情监督提供丰富的资料。

4. 网络安全（Cyber Security）

大规模网络安全事件的发生，例如 2017 年 5 月席卷全球的 WannaCry 病毒，让企业意识到网络攻击发生时预先快速识别的重要性。传统的网络安全主要依靠静态防御，处理病毒的主要流程是发现威胁、分析威胁和处理威胁。这种情况下，往往在威胁发生以后才能做出反应。新型的病毒防御系统可使用数据分析技术，建立潜在攻击识别分析模型，监测大量网络活动数据和相应的访问行为，识别可能进行入侵的可疑模式，做到未雨绸缪。

5. 设备管理（Plant and Facility Management）

设备管理同样是企业关注的重点。设备维修一般采用标准修理法、定期修理法和检查后修理法等方法。其中，标准修理法可能会造成设备过剩修理，修理费用高；检查后修理法解决了修理费用成本问题，但是修理前的准备工作繁多，设备的停歇时间过长。目前企业能够通过物联网技术收集和分析设备上的数据流，包括连续用电、零部件温度、环境湿度和污染物颗粒等多种潜在特征，建立设备管理模型，从而预测设备故障，合理安排预防性的维护，以确保设备正常作业，降低因设备故障带来的安全风险。

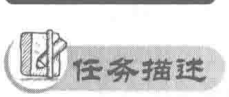
6. 交通物流分析（Transport and Logistics Analytics）

物流是物品从供应地向接收地的实体流动，是将运输、储存、装卸搬运、包装、流通加工、配送和信息处理等功能有机结合起来而实现用户要求的过程。用户可以通过业务系统和 GPS 定位系统获得数据，使用数据构建交通状况预测分析模型，有效预测实时路况、物流状况、车流量、客流量和货物吞吐量，进而提前补货，制定库存管理策略。

7. 欺诈行为检测（Fraud Detection）

身份信息泄露及盗用事件逐年增长，随之而来的是欺诈行为和交易的增多。公安机关、各大金融机构、电信部门可利用用户基本信息、用户交易信息和用户通话短信信息等数据，识别可能发生的潜在欺诈交易，做到提前预防、未雨绸缪。以大型金融机构为例，通过分类模型分析方法对非法集资和洗钱的逻辑路径进行分析，找到其行为特征。聚类模型分析方法可以分析相似价格的运动模式。例如对股票进行聚类，可能发现关联交易及内幕交易的可疑信息。关联规则分析方法可以监控多个用户的关联交易行为，为发现跨账号协同的金融诈骗行为提供依据。

任务 1.2 熟悉 Python 数据分析的工具



Python 已经有将近 30 年的历史。在过去的将近 30 年中，Python 在运维工程师群体中受到广泛欢迎，然而却极少有企业将 Python 作为生产环境的首选语言。在最近几年，这一