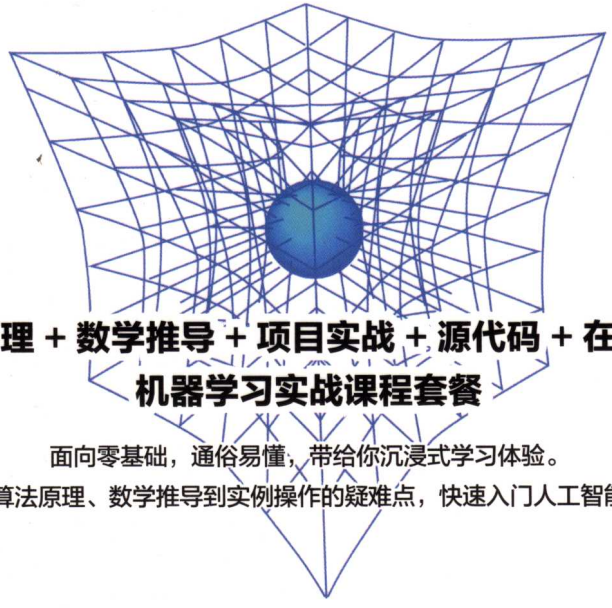
人气视频课程讲师

累计超过30万学员的选择

跟着迪哥学Python

数据分析与机器学习实战

唐宇迪／著



算法原理＋数学推导＋项目实战＋源代码＋在线服务

机器学习实战课程套餐

面向零基础，通俗易懂，带给你沉浸式学习体验。

打通从算法原理、数学推导到实例操作的疑难点，快速入门人工智能领域。

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

人气视频课程讲师

累计超过30万学员的选择

跟着迪哥学Python

数据分析与机器学习实战

唐宇迪／著

人民邮电出版社

北京

图书在版编目（CIP）数据

跟着迪哥学Python数据分析与机器学习实战 / 唐宇迪著. -- 北京 : 人民邮电出版社, 2019.9
ISBN 978-7-115-51244-4

Ⅰ. ①跟… Ⅱ. ①唐… 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(2019)第087638号

内 容 提 要

本书结合了机器学习、数据分析和 Python 语言，通过案例以通俗易懂的方式讲解了如何将算法应用到实际任务。

全书共 20 章，大致分为 4 个部分。第 1 部分介绍了 Python 必备的工具包，包括科学计算库 Numpy、数据分析库 Pandas、可视化库 Matplotlib；第 2 部分讲解了机器学习中的经典算法，例如回归算法、决策树、集成算法、支持向量机、聚类算法等；第 3 部分介绍了深度学习中的常用算法，包括神经网格、卷积神经网络、递归神经网络；第 4 部分是项目实战，基于真实数据集，将算法模型应用到实际业务中。

本书适合对人工智能、机器学习、数据分析等方向感兴趣的初学者和爱好者。

◆ 著　　唐宇迪
责任编辑　俞　彬
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
北京市艺辉印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：28.75
字数：765 千字　　2019 年 9 月第 1 版
印数：1 – 3 000 册　　2019 年 9 月北京第 1 次印刷

定价：89.00 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

前言
PREFACE

人工智能的飞速发展，带来了丰富的机遇与挑战。机器学习算法工程师、数据挖掘工程师、大数据工程师等岗位的薪资在 IT 行业也颇丰。面对高薪与前沿技术的诱惑，越来越多的大学毕业生准备投身其中，但苦于缺乏指导性教材进行系统学习，非科班出身的大学毕业生更是缺乏相关数学基础。

很多同学认识我是通过在线课程或线下培训，机器学习培训工作已经伴我走过了近 4 个年头。在这期间，开发的线上就业课程 40 余门，参与的学员累计超过 30 万人，顺利完成企业与高校讲师培训 30 余场，直播课程百余场。忙碌之余，最大的收获就是收到同学们晒出的各大企业的 offer 与认可。

在培训工作中，同学们给我最多的反馈就是虽然能参考的资料有很多，但是都很难理解，尤其对于初学者而言，看各种公式就要晕掉了。这几年我也一直在思考如何讲解才能让大家更深刻、更轻松地理解机器学习中的每一个算法。

本书是我多年培训教学和学习心得的总结，最大的特色就是以接地气的方式向大家通俗地讲解算法原理与应用方法，让读者能够更轻松地去理解其中每一个复杂的算法。学习的目的肯定要在实际任务中发挥作用，我写作的初衷也是希望更多读者能将理论与实战方法应用到自己的业务中，所以本书整体风格是以实战为主，通过案例来解读如何将机器学习应用在实际的数据挖掘任务中。

本书面向的读者

本书主要面向对人工智能、机器学习、数据分析等方面有强烈兴趣的初学者和爱好者，通过本书的学习，读者能够掌握机器学习中经典算法原理推导、整体流程以及其中数学公式与各种参数的作用。案例全部采用当下流行的 Python 语言，从最基础的工具包开始讲起，让大家熟练使用 Python 及其数据科学工具包进行机器学习和数据挖掘领域的项目实战任务，并处理其中遇到的种种问题。

路线图

本书内容大体可以分为以下 4 个部分。

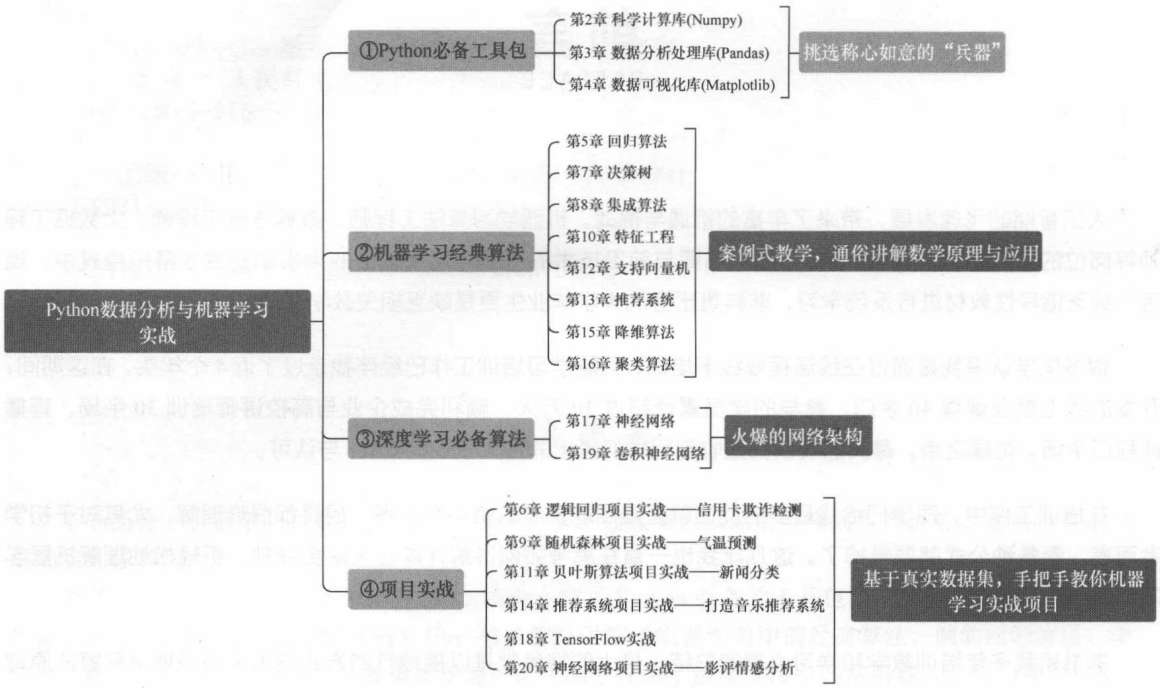


图 0-1 本书学习路线图

总结起来比较合适的学习路线如下。

第①步： Python 工具包的使用，先把称心如意的“兵器”准备好，它们是实战中的好帮手。

第②步： 理解机器学习算法，建模分析的核心就是其中的算法了，打牢基础才能走得更远。

第③步： 项目实战应用，将算法模型应用到实际业务中，通过实际任务来进行提升。

可能很多读者都觉得应当先把 Python 的基础打牢固再进行后续的学习，我觉得这样可能会花费较多时间，从而耽搁后续重点内容学习，建议读者对于编程语言通过实际案例边练边学，把重点放在机器学习原理与应用中。

阅读本书需要准备什么 / 如何使用本书

对于初学者来说，可能在学习路线以及职业规划上有些迷茫，这里结合我对机器学习与数据科学领域的理解来进行阐述分析。首先无论从事人工智能中哪个方向，肯定要从工程师做起，那手里一定得有一个称心如意的“兵器”，本书选择的是 Python 语言，基于 3.x 版本进行实战演示。读者如果具备大学数学基础，学习起来会相对更容易一些，在学习过程中，难免遇到各种难以理解的算法问题，建议大家先对其整体流程进

行通俗理解，再结合实际案例进行思考，很多时候数学上的描述十分复杂，而代码中的解释却浅显易懂。项目实战的目的一方面是从应用的角度阐述如何进行实际任务建模与分析，另一方面也是一个积累的过程。人工智能行业发展迅速，不要停下学习的脚步，每天都要学习新的知识来充实自己。

配套资源

本书由异步社区（https://www.epubit.com/）为您提供相关资源。

本书提供配套的源代码和数据源文件。要想获得配套资源，请登陆异步社区，按书名搜索，进入本书页面，点击配套资源，跳转到下载页面，按提示进行操作即可。注意：为保证购书读者的权益，该操作会给出相关提示，要求输入提取码进行验证。

建议与反馈

由于作者水平有限，书中难免有错误和不当之处，欢迎读者指正。如果读者遇到问题需要帮助，也欢迎交流（微信号：digexiaozhushou），我期望与你共同成长。

目录

CONTENTS

第 1 章
人工智能入门指南

当今时代，人工智能迅速发展，高薪的诱惑、前沿的技术挑战使得越来越多的小伙伴想要学习人工智能，那么更大的问题也就随之产生了——如何学习人工智能呢？正所谓“万事开头难”，如何走好第一步十分关键。学习人工智能的成本还是蛮高的，一般来说，付出了大量的时间和精力，一定要有满意的收获才可以。作为 Python 开篇之讲，本章首先介绍机器学习处理问题的方法与流程，以及实战必备武器——Python 基础教程及其环境配置。

1.1 AI 时代首选 Python

人工智能就是用编程实现各种算法和数据建模。提起编程，以前大家可能更注重 C 语言和 Java 语言，但是现在，Python 在数据科学领域运用广泛，相信大家早已在各大媒体和圈子中看到 Python 与日俱增的发展前景，可以说，Python 已经成为当下最火的编程语言之一了（见图 1-1）。

图 1-1 AI 时代首选 Python

1.1.1 Python 的特点

Python 被当作“核心武器”肯定是有原因的，进入 AI 行业，大家最初给自己的定位基本都是工程师，办事效率肯定是越高越好，这跟 Python 的出发点也是一致的，试问：能用 1 行代码解决的问题，何必用 10 行呢？

如果大家学过 C 语言，肯定会觉得它用起来还是比较麻烦的，限制非常多。但是用 Python 写起程序来可以更随性一些，没有那么多的语法束缚，用起来容易，学起来也很简单。

当要实际完成一项编程任务时，肯定需要借助各种工具，Python 提供了非常丰富的工具包来解决各种数据处理、分析、建模等问题。我们只要调用工具包，就可以轻轻松松地完成任务，相当于前人已经种好了树，我们去乘凉就好了。

那么，Python 在其他领域应用得怎么样呢？大家可能听过“Python 全栈开发”这个概念，所以 Python 相当于“万金油”，只要把它学好了，应用还是十分广泛的。

总结起来就是一句话：简洁、高效，用起来舒服！对于初学者来说，Python 是很友好的，可以说它是最简单易学的编程语言。

1.1.2 Python 该怎么学

很多零基础的读者的第一想法可能就是先去买一本非常厚的 Python 教材，然后慢慢地从入门到精通……其实我认为语言只是用来帮助解决问题的工具，不建议去找一本特别厚的书，来个半年学习计划，用最短的

时间学习最基础的、暂时够用的知识就可以了，越高级的语法用到的概率越小，先入手用起来，然后边做案例边学习才是高效的学习方法。

推荐大家先熟悉Python的基础部分，到图书馆随便找本这方面的书，或者看看Python的在线课程都可以，有其他语言基础的同学学习2 ~ 3天就能用起来，第一次接触编程语言的人花一周的时间也会学得差不多了。

在后续的章节中本书还会涉及Python工具包的使用，其实这些工具的使用方法在其官方文档中都写得清清楚楚，并不需要全部背下来，只需要熟练操作即可，真正用到它的时候，还是要看看文档中每一个参数的具体含义。

1.2 人工智能的核心——机器学习

到底该如何学习人工智能呢？可以说，人工智能这个圈子太大了，各行各业都有涉及，可选择的方向也五花八门、各不相同，包括数据挖掘、计算机视觉、自然语言处理等各大领域。那么，是不是每个方向要学习的内容差别很大呢？不是的。其实最核心的就是机器学习，你要做的一切都离不开它，所以无论选择哪个领域，一定要把基础打牢。因此，第一个目标就是搞定机器学习的各大算法，并掌握其应用实践方法。

1.2.1 什么是机器学习

可能有些读者对机器学习还不是很熟悉，只不过因为最近这个词比较火才准备投身这个领域中。举一个小例子，我以前特别喜欢玩一款叫作《梦幻西游》的游戏。弃坑之后，游戏方的客服经理总给我打电话，说“迪哥能不能回来接着玩耍（充值）呀，帮派的小伙伴都十分想念你……”这时候我就想：他们为什么会给我打电话呢？这款游戏每天都有用户流失，不可能给每个用户都打电话吧，那么肯定是挑重点用户来沟通了。其后台肯定有玩家的各种数据，例如游戏时长、充值金额、战斗力等，通过这些数据就可以建立一个模型，用来预测哪些用户最有可能返回来接着玩啦！

机器学习要做的就是在数据中学习有价值的信息，例如先给计算机一堆数据，告诉它这些玩家都是重点客户，让计算机去学习一下这些重点客户的特点，以便之后在海量数据中能快速将它们识别出来。

机器学习能做的远不止这些，数据分析、图像识别、数据挖掘、自然语言处理、语音识别等都是以其为基础的，也可以说人工智能的各种应用都需要机器学习来支撑（见图1-2）。现在各大公司越来越注重数据的价值，人工成本也是越来越高，所以机器学习也就变得不可或缺了。

再给大家简单介绍一下学会机器学习之后可能从事的岗位，最常见的就是数据挖掘岗，即通过建立机器学习模型来解决实际业务问题，就业前景还是非常不错的，基本所有和数据打交道的公司都需要这个岗位。

图 1-2　机器学习的应用领域

接下来就是当下与人工智能结合最紧密的计算机视觉、自然语言处理和语音识别了。说白了就是要让计算机能看到、听到、读懂人类的数据。相对来说，我觉得计算机视觉领域发展会更快一些，因为随着深度学习技术的崛起，越来越多的研究人员加入这个行列，落地的项目更是与日俱增。自然语言处理和语音识别也是非常不错的方向，至于之后的路怎么走还是看大家的喜好吧，前提都是一样的——先把机器学习搞定！

1.2.2　机器学习的流程

上一小节简单介绍了机器学习的基本概念，那么机器学习是如何做事情的呢？下面通过一个简单例子来了解一下机器学习的流程（见图 1-3）。假设我们从网络上收集了很多新闻，有的是体育类新闻，有的是非体育类新闻，现在需要让机器准确地识别出新闻的类型。

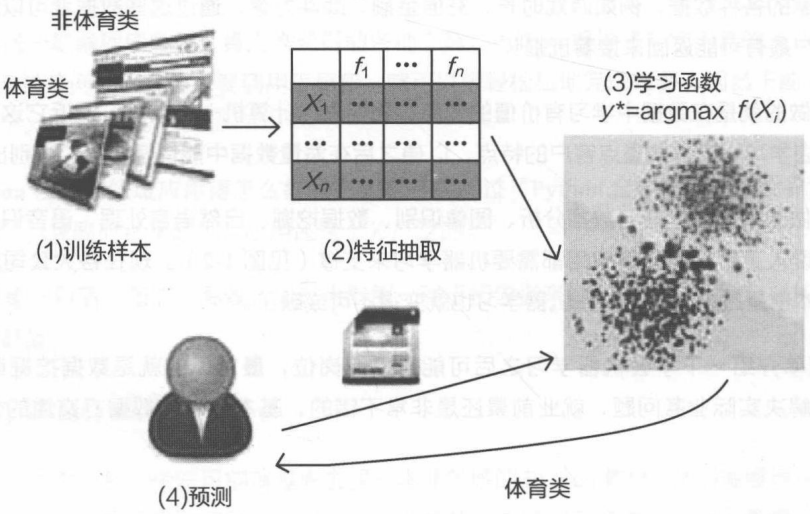


图 1-3　机器学习流程

一般来说，机器学习流程大致分为以下几步。

第①步：数据收集与预处理。例如，新闻中会掺杂很多特殊字符和广告等无关因素，要先把这些剔除掉。除此之外，可能还会用到对文章进行分词、提取关键词等操作，这些在后续案例中会进行详细分析。

第②步：特征工程，也叫作特征抽取。例如，有一段新闻，描述“科比职业生涯画上圆满句号，今天正式退役了”。显然这是一篇与体育相关的新闻，但是计算机可不认识科比，所以还需要将人能读懂的字符转换成计算机能识别的数值。这一步看起来容易，做起来就非常难了，如何构造合适的输入特征也是机器学习中非常重要的一部分。

第③步：模型构建。这一步只要训练一个分类器即可，当然，建模过程中还会涉及很多调参工作，随便建立一个差不多的模型很容易，但是想要将模型做得完美还需要大量的实验。

第④步：评估与预测。最后，模型构建完成就可以进行判断预测，一篇文章经过预处理再被传入模型中，机器就会告诉我们按照它所学数据得出的是什么结果。

1.2.3 机器学习该怎么学

很多读者可能都会有这种想法：工具包已经非常成熟了，是不是会调用工具包就可以了呢？笔者认为掌握算法原理与实际应用都是很重要的，很多人容易忽略算法的推导，这对之后的学习和应用肯定是不利的，因为做一件事情不能盲目去做，需要知道为什么要这么做！工具包也一样，不仅要学会使用它，更要知道其中每一个参数的作用，以及每一步操作在算法中都是什么含义。

这就需要熟悉每一个算法是怎么来的，每一步数学公式的目的是什么，数据是怎么一步步变成最后的决策结果的，每一步的参数又会对最终的结果产生什么样的影响。这几点都是非常重要的，所以在学习过程中需要深入其中每一步细节。

学习过程肯定有些枯燥，最好先从整体上理解其工作原理，然后再深入到每一处细节。这其中会涉及很多数学知识，对于初学者来说最头疼的就是这些公式和符号了，让大家从头到尾先学一遍数学可能有点不现实，所以遇到问题或者不理解的地方还需要大家勤动手，边学边查，也就是“哪里不会点哪里”。本书中所有知识点也都是按照笔者的理解跟大家分享的，所以不要惧怕数学，也不要过于钻牛角尖，理解即可。

1.3 环境配置

现在跟大家说一说本书所需的环境配置，也就是后续案例怎么玩起来，这个很重要，能给大家节省很多时间。我们要安装 Python 所需环境，不推荐去 Python 官网下载一个安装包，否则之后的配置和要安装的东西就太多了。

1.3.1 Anaconda 大礼包

环境配置时只需下载 Anaconda 即可，它相当于一个“全家桶”，里面不仅有 Python 所需环境，而且还把后续要用到的工具包和编程环境全部搞定了。首先登录 Anaconda 官网（https://www.anaconda.com/download/），下载对应软件，如图 1-4 所示。

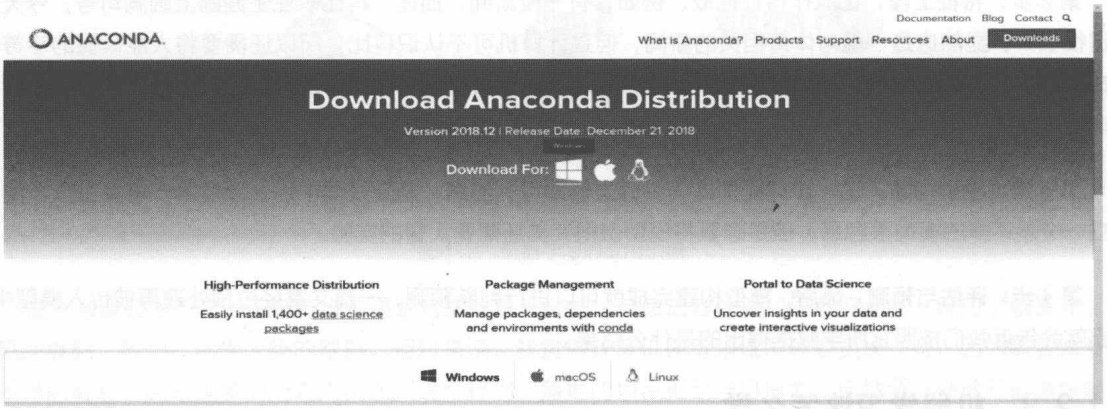


图 1-4 Anaconda 下载

然后根据自己的电脑选择不同的操作系统，并选择是 64 位的还是 32 位的。如果电脑是 32 位的，可以考虑换一换，因为很多工具包都不支持。

一定要选择 Python 3 版本（见图 1-5），几年前我在讲课和工作的时候用的是 Python 2.7 版本，当时，2.7 版本用的人比较多，而且相对稳定。但是从现在的角度出发，很多工具包都不支持 2.7 版本了，所以直接下载 3 版本即可。如果下载速度比较慢，读者可以登录镜像网址 https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive，下载对应版本软件。

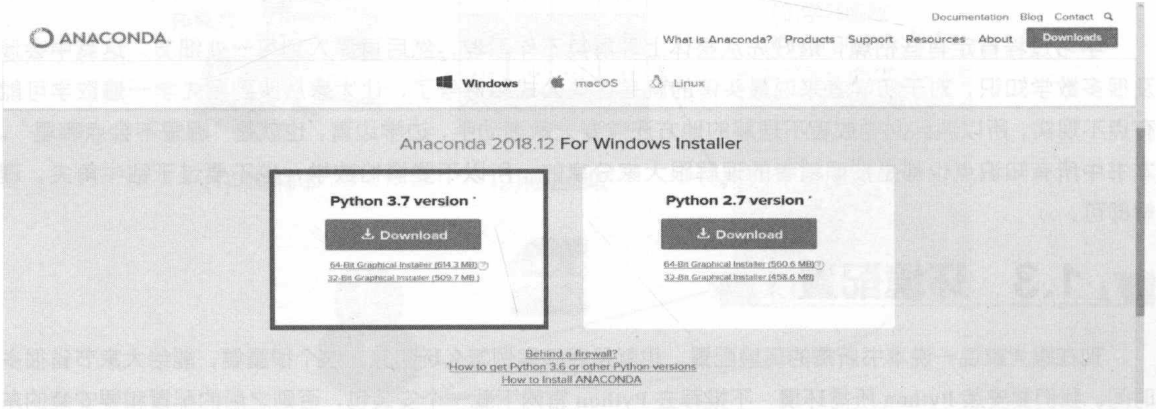


图 1-5 Python 版本选择