大话Python机器学习

张居营◎编著

在Python实战中学习算法，让你爱上机器学习！

简单

注重原理的
理解与现实应用
方便读者快速入门

实用

包括20多个案例
5000+行代码示例
具有极大实操性

系统

基于十大经典算法
兼顾Python应用
与数据思维

有趣

10余个生活中
有趣的例子来诠释算法
极具趣味性

中国水利水电出版社
www.waterpub.com.cn

大话Python机器学习

张居营◎编著

中国水利水电出版社
www.waterpub.com.cn
·北京·

内 容 提 要

《大话 Python 机器学习》从机器学习的基础知识讲起，全面、系统地介绍了机器学习算法的主要脉络与框架，并在每个算法原理、应用等内容基础上，结合 Python 编程语言深入浅出地介绍了机器学习中的数据处理、特征选择、算法应用等技巧，是一本兼具专业性与入门性的 Python 机器学习书籍。

《大话 Python 机器学习》分为 13 章，主要内容有机器学习入门基础、应用 Python 实现机器学习前的准备、单变量线性回归算法、线性回归算法进阶、逻辑回归算法、贝叶斯分类算法、基于决策树的分类算法、K 近邻算法、支持向量机、人工神经网络、聚类算法、降维技术与关联规则挖掘，在具体介绍时侧重于机器学习原理、思想的理解，注重算法的应用，并辅助以相关的数据案例，方便读者快速入门。最后一章从一个关于房价预测的机器学习项目出发，系统展示了数据处理、特征提取、建模训练等机器学习完整流程，带领读者完成从零基础到入门数据科学家的飞跃。

《大话 Python 机器学习》条理清晰，内容深入浅出，以生活、工作中常见的例子来解释机器学习中的相关概念、算法原理和运算思维等，特别适合互联网创业者、数据挖掘相关人员、Python 程序员、人工智能从业者、数据分析师、计算机专业的学生学习，任何对机器学习、人工智能感兴趣的读者均可选择本书作为入门图书参考学习。

图书在版编目（CIP）数据

大话 Python 机器学习 / 张居营编著 . —北京：中国水利水电出版社，2019. 6
ISBN 978-7-5170-7434-2

Ⅰ. ①大… Ⅱ. ①张… Ⅲ. ①软件工具 - 程序设计②机器学习
Ⅳ. ①TP311. 561②TP181

中国版本图书馆 CIP 数据核字（2019）第 029352 号

书　　名	大话 Python 机器学习 DAHUA Python JIQI XUEXI
作　　者	张居营　编著
出版发行	中国水利水电出版社 （北京市海淀区玉渊潭南路 1 号 D 座　100038） 网址：www.waterpub.com.cn E-mail：zhiboshangshu@163.com 电话：（010）62572966-2205/2266/2201（营销中心）
经　　售	北京科水图书销售中心（零售） 电话：（010）88383994、63202643、68545874 全国各地新华书店和相关出版物销售网点
排　　版	北京智博尚书文化传媒有限公司
印　　刷	三河市龙大印装有限公司
规　　格	170mm × 230mm　16 开本　26. 5 印张　433 千字　1 插页
版　　次	2019 年 6 月第 1 版　2019 年 6 月第 1 次印刷
印　　数	0001—5000 册
定　　价	89. 80 元

前　言

Preface

有这样一个职场故事：小明和小刚同时应聘一家食品企业，在面试的最后阶段，老板给他们出了一道题：让他们分别到市场上去买一袋芒果来。不到一会儿，小明很快买了一袋芒果回来，当老板问他：市场上有几家芒果店，价位如何？小明却愣住了，回答不上来。又过了一会，小刚也带着一袋芒果回来了，但他完美地回答了老板的问题，同时提供了店家的联系方式。

看完这两个人的表现，相信大多数人都认为老板会选择小刚，因为他做事积极主动、注重市场信息的搜集。但是，老板最后选择了小明，因为小明买的芒果个个果汁饱满，味道甜嫩。虽然小刚买的芒果颜色、个头也不差且经过了多方比较，但是芒果的味道差了些，甚至有酸的、不熟的。

从这个案例可以看出，小刚虽然做事积极主动，考虑问题周全，但是在核心业务——挑选优质的芒果上，反而顾此失彼。而小明为什么能够在短时间内抓住核心问题，并有效解决呢？

小明说，最初我在吃芒果的时候，也倾向于认为个大、颜色金黄的芒果味道要甜一些，但是这个规律并不时时准确。尤其当我在外地买当地自产的芒果时，发现个小、浅黄色的芒果反而更甜。随着我吃的芒果越来越多，面对不同颜色、大小、形状、产地、成熟度等特征的芒果，我自己却很难摸清楚挑选美味芒果的规律了。这时候我采用了机器学习的算法，它是一种让机器来学习、挖掘数据内在规律并能够进行预测的方法。

我把芒果的颜色、大小、形状、产地、成熟度、甜度、多汁程度等数据放进一个机器学习算法中，让它对芒果的物理特征与甜度等品质之间的关系进行学习，最后得到一个相关性模型。今天我就应用这个模型，找到了市场上的其中一家店，根据所卖芒果的特征与信息，马上就知道了哪家店芒果很甜的结论，然后我就很快买回来了。

这个故事告诉我们：一，在职场中，做事积极有责任心很重要，但是核心竞争力更重要；二，当我们掌握了机器学习的相关知识并能有所应用时，它将帮助我们挖掘出更多信息，并做出相对精确的决策，尤其身处当

今的大数据时代，信息量与数据量均以几何级别的水平不断增长，它更需要我们对信息量进行提炼，精准获取需要的信息。

机器学习应运而生

在现有技术的限制下，人脑处理信息的能力毕竟有限，机器学习却能突破这一限制，以它自己的学习方式，从繁杂的数据中发现规律，进行预测。这也是机器学习越来越重要的一个原因。一方面，在我们生活的很多方面，机器学习正在帮助我们解决问题，比如过滤垃圾邮件、预防疾病等；另一方面，以机器学习为基础的深度学习逐渐形成了模拟人类思维的人工智能，例如 AlphaGo 在围棋上的成功、汽车自动驾驶等。

以上这些情况导致了机器学习的火爆，企业对机器学习的人才需求旺盛，年薪动辄几十万元，甚至上百万元。著名研究机构 SlashData 曾对 2 万名开发者进行调查，结果显示，这些开发者未来一年最希望掌握的技能就是机器学习与数据科学。然而，想进入机器学习领域具有一定的门槛，需要掌握一定的数学知识与编程技巧等，这也是令很多非专业人员望而却步的主要原因。

高门槛并不代表你不需要关注它、了解它，尤其是在未来“人工智能时代”，缺乏一定的算法思维，可能就会在核心业务上缺少关键助手，落后于他人。本书是一本兼具专业性与入门性的书籍，通过结合 Python——当前最流行、最受欢迎的机器学习编程语言，同时辅以大量有趣的生活和工作中的例子，既能向非专业人员讲解机器学习算法的基本原理与应用，又能帮助专业学习者深入掌握相关算法、Python 编程，甚至是数据科学思维等。

笔者作为一名多年从事数据分析的专业人士，在数据分析、挖掘、机器学习算法使用上有自己的见解与经验，也曾做过有关数据分析与机器学习的大众化培训，反响不错；另外，笔者也在一些科技自媒体平台上分享过有关机器学习的相关文章，传递机器学习应用知识。这些都促使笔者对机器学习及 Python 使用的相关感悟进行总结，最终形成本书。因笔者水平和成书时间所限，书中难免存有疏漏和不当之处，敬请指正。

本书特色

1. 结构编排注重算法间的内在逻辑，为读者提供较好的阅读体验

本书从初学者的视角出发，在注重机器学习的主要原理与数学基础之上，以平实通俗的语言，带领读者了解机器学习的理论基础及 Python 使用

技巧。在介绍机器学习算法时条理清晰，按照从回归问题到分类问题、从监督学习到无监督学习的顺序展开，内容编排上注重算法间的内在逻辑，给读者提供较好的阅读体验。

2. 内容深入浅出，以实例引导，方便读者快速入门

本书注重对机器学习原理和思想的理解，注重算法的应用，每一部分均辅以相关数据案例，方便读者快速入门；内容深入浅出，以生活与工作中常见的例子来解释机器学习中的相关概念、算法原理、运算思维等，基本做到了对每个关键知识点的案例解释。

3. 知识涵盖范围广，强调项目实战中的数据科学思维

本书介绍机器学习但内容又不限于机器学习，注重 Python 编程应用但又不限于此，具体表现在对每个机器学习的实战项目上，不仅论述了算法解决问题的过程，还注重算法训练之前的数据处理与数据清洗、算法训练之后的评价与效果比较等。特别是本书最后一章，从一个机器学习项目出发，系统地展示了数据处理、特征选择、算法应用等完整流程，带领读者完成从零基础到入门数据科学家的飞跃。

本书内容及体系结构

第 1 章　机器学习入门基础

本章以日常生活中的案例为基础，通俗地讲解了机器学习的内涵与思维，并介绍了机器学习项目的实施流程与应用等，相信读者学完第 1 章就能够对机器学习有一个整体的了解。

第 2 章　应用 Python 实现机器学习前的准备

本章对应用 Python 实现机器学习之前需要做什么准备进行了详细介绍，包括为什么使用 Python、Python 机器学习的一些常用库、Python 集成工具 Anaconda 的安装与使用、应用平台 Jupyter Notebook 模式的介绍等，既为后边章节的学习打下基础，也让读者初步掌握 Python 使用的一些技巧。

第 3 章　从简单案例入手：单变量线性回归

本章从机器学习最基础的算法——单变量线性回归入手，在介绍其基本原理、求解过程等专业知识的基础上，展示了利用机器学习算法解决一个实际案例的基本流程，引领读者建立对机器学习的初步印象。

第 4 章　线性回归算法进阶

本章在第 3 章的基础上，进一步讲解了机器学习中的主要线性回归算法，包括多变量线性回归、岭回归、Lasso 回归，其中还对一些典型问题进行了拓

展，比如梯度下降法的原理与求解方法、正则化问题等。通过对本章的学习，读者可以对机器学习中的回归问题有个系统性的了解与知识构建。

第 5 章　逻辑回归算法

本章主要介绍机器学习中的逻辑回归算法。作为分类问题的基础算法，逻辑回归却又具有一定的回归特性，读者通过从线性回归过渡到逻辑回归的学习，能够理解回归、分类问题处理的共性与不同之处。

第 6 章　贝叶斯分类算法

本章进入对贝叶斯分类算法的学习，该算法强调了由事物发生的条件概率而构建的一种判定方法。本章通过对贝叶斯定理的介绍，对朴素贝叶斯分类算法的原理、参数估计、Python 实现，以及贝叶斯网络算法的基本原理与特点等的介绍，向读者全面展示了贝叶斯算法的相关内容。

第 7 章　基于决策树的分类算法

本章进入以信息论为基础的决策树算法的学习，通过对熵与信息熵等概念的介绍，系统展现了决策树算法中的 ID3、C4. 5、CART 等算法，同时详细介绍了剪枝方法、集成学习算法、随机森林算法等内容。

第 8 章　K 近邻算法

本章介绍了 K 近邻算法的原理与特点、算法学习要解决的问题等内容，并结合两个具体案例——文化公司推广活动的效果预估和解决交通拥堵问题，通俗易懂地讲解了 K 近邻算法的应用以及 Python 实现。K 近邻算法借鉴了空间映射的原理。通过对本章的学习，读者可以了解机器学习中应用空间维度解决问题的思维方法。

第 9 章　支持向量机

本章主要介绍机器学习中支持向量机的算法，该算法不同于 K 近邻算法，它是通过超平面、间隔等空间思维特征来实现的一种算法。内容包括在线性可分下、线性不可分下、非线性、多类分类等不同情形下的支持向量机算法以及支持向量回归机，同时介绍了各种情形算法的 Python 实现技巧。

第 10 章　人工神经网络

本章开始进入对人工神经网络算法的学习，它是基于生物学而采用的复杂的并行计算分析技术，其最大特点是能够拟合极其复杂的非线性函数。当前比较热门的人工智能基础——深度学习就是以此为基础的。本章不仅介绍了人工神经网络算法的原理与应用，还介绍了深度学习的相关内容，引导读者对深度学习有初步认识。

第 11 章　聚类算法

本章是由机器学习中的监督学习向无监督学习算法延伸的一章，聚类

算法就是典型的、应用最广泛的无监督算法。本章在对监督学习与无监督学习的原理和区别进行详细介绍的基础上，系统地讲解了聚类算法的原理、应用以及主要的聚类算法的实现。通过对本章的学习，读者能够快速理解并应用聚类算法。

第 12 章　降维技术与关联规则挖掘

本章介绍了无监督学习算法的另外几种处理方法——降维技术、关联规则挖掘等的基本原理与具体应用，并结合最基础的二维样本案例对各类算法进行讲解，有助于读者直观理解并掌握降维技术与关联规则挖掘的相关内容与算法实现。

第 13 章　机器学习项目实战全流程入门

本章介绍了从机器学习算法学习到项目实战的提升过程，同时以一个简单的项目实战全流程详细展示了机器学习项目解决方案的基本过程。通过本章的学习，读者不仅能够初步建立机器学习和数据科学思维，而且能够全面掌握 Python 的使用。

本书读者对象

- 互联网创业者
- 数据挖掘相关人员
- Python 程序员
- 人工智能从业者
- 数据分析师
- 计算机专业的学生

本书资源获取及联系方式

（1）扫描下面的微信公众号，关注后输入“74342”并发送到公众号后台，获取本书资源下载链接。然后将该链接粘贴到计算机浏览器地址栏中，按 Enter 键后即可进入资源下载页面，根据提示下载即可。

（2）推荐加入 QQ 群：981389956（若此群已满，请根据提示加入相应的群），可在线交流学习。

最后，祝您学习路上一帆风顺！

目　录

Contents