

计算机与智能科学丛书

机器学习和深度学习： 原理、算法、实战 (使用 Python 和 TensorFlow)

[印] 文卡塔·雷迪·科纳萨尼(Venkata Reddy Konasani)
沙伦德拉·卡德雷(Shailendra Kadre)

著

秦 婧 韩雨童

译

清华大学出版社

北 京

北京市版权局著作权合同登记号: 01-2023-0681

Venkata Reddy Konasani, Shailendra Kadre

Machine Learning and Deep Learning Using Python and TensorFlow

ISBN: 978-1-260-46229-6

Copyright © 2021 by McGraw-Hill Education.

All Rights reserved. No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including without limitation photocopying, recording, taping, or any database, information or retrieval system, without the prior written permission of the publisher.

This authorized Chinese translation edition is published by Tsinghua University Press Limited in arrangement with McGraw-Hill Education(Singapore)Pte.Ltd. This edition is authorized for sale in the People's Republic of China only, excluding Hong Kong, Macao SAR and Taiwan.

Translation Copyright © 2023 by McGraw-Hill Education(Singapore) Pte.Ltd and Tsinghua University Press Limited.

版权所有。未经出版人事先书面许可, 对本出版物的任何部分不得以任何方式或途径复制传播, 包括但不限于复印、录制、录音, 或通过任何数据库、信息或可检索的系统。

此中文简体翻译版本经授权仅限在中华人民共和国境内(不包括香港特别行政区、澳门特别行政区和台湾地区)销售。

翻译版权©2023 由麦格劳-希尔教育(新加坡)有限公司与清华大学出版社有限公司所有。

本书封面贴有 McGraw-Hill Education 公司防伪标签, 无标签者不得销售。

版权所有, 侵权必究。举报: 010-62782989, beiqinquan@tup.tsinghua.edu.cn

图书在版编目(CIP)数据

机器学习和深度学习: 原理、算法、实战: 使用Python 和 TensorFlow / (印) 文卡塔·雷迪·科纳萨尼, (印) 沙伦德拉·卡德雷著; 秦婧, 韩雨童译. —北京: 清华大学出版社, 2023.1
(计算机与智能科学丛书)

书名原文: Machine Learning and Deep Learning Using Python and TensorFlow

ISBN 978-7-302-62479-0

I. ①机… II. ①文… ②沙… ③秦… ④韩… III. ①机器学习 ②软件工具—程序设计IV. ①TP181
②TP311.561

中国国家版本馆 CIP 数据核字(2023)第 017024 号

责任编辑: 王 军

装帧设计: 孔祥峰

责任校对: 成凤进

责任印制: 沈 露

出版发行: 清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址: 北京清华大学学研大厦 A 座 邮 编: 100084

社 总 机: 010-83470000 邮 购: 010-62786544

投稿与读者服务: 010-62776969, c-service@tup.tsinghua.edu.cn

质 量 反 馈: 010-62772015, zhiliang@tup.tsinghua.edu.cn

印 装 者: 北京鑫海金澳胶印有限公司

经 销: 全国新华书店

开 本: 170mm×240mm 印 张: 34.75 字 数: 893 千字

版 次: 2023 年 3 月第 1 版 印 次: 2023 年 3 月第 1 次印刷

定 价: 128.00 元

产品编号: 093138-01

译者序

机器学习和深度学习已经成为从业人员在人工智能时代必备的技术，被广泛应用于图像识别、自然语言理解、推荐系统、语音识别等多个领域，并取得了丰硕的成果。目前，很多高校的人工智能、软件工程、计算机应用等专业均已开设了机器学习和深度学习的课程，此外，为便于让学生掌握一些大数据的分析和可视化技术，有一些非计算机专业也开设了与机器学习相关的课程。同时，企业中的一些软件开发人员也想学习相关行业中机器学习和深度学习技术应用的真实案例，而不用过多关注机器学习和深度学习中的数学知识。但是，在市面上能够同时满足上述需求的参考书并不多，而《机器学习和深度学习：原理、算法、实战(使用 Python 和 TensorFlow)》恰好是一本能满足高校学生及相关从业人员需求的机器学习和深度学习的参考书。

本书的作者是两位资深的行业专家并且拥有多年的企业培训经验，他们以案例式的编写方式由浅入深地讲解了机器学习和深度学习技术，让读者能够快速掌握机器学习和深度学习的原理及相关应用。在本书的翻译过程中，我们翻阅了大量资料，力求为读者献上一部贴近实际应用且通俗易懂的机器学习与深度学习方面的参考书。

最后，我要感谢给予本书指导的各位编辑。本书由秦婧和韩雨童翻译，虽然在翻译过程中进行了仔细的推敲和研究，但由于水平有限，翻译中难免有疏漏和争议之处，敬请广大读者提供宝贵意见和建议。

作者简介



Venkata Reddy Konasani 是一位数据科学领域的培训讲师。他拥有 10 年的数据分析和数据科学方面的经验，并且有 6 年培训讲师经验。他擅长构建信用风险模型，擅长营销分析及社交媒体分析。

目前，Venkata 为 Oracle、IBM、CTS 及美国银行等公司定期举办数据科学方面的培训。他组织了逾 100 次关于机器学习和深度学习主题的企业培训。

Venkata 是 Statifer.com(数据科学在线学习公司)的联合创始人。他毕业于孟买印度理工学院，获得应用统计和信息学硕士学位，目前居住在印度的维杰亚瓦达。



Shailendra Kadre 在印度的班加罗尔从事管理和信息技术顾问工作。他拥有在多个领域逾 20 年的工作经验，包括数字转型、机器学习、预售、信息技术项目和产品领域。他在这些领域出色地担任了许多关键的领导角色。他曾就职于多家全球性的一流公司，如惠普有限公司、塔塔咨询服务公司和马衡达信息技术有限公司等。Shailendra 是 *Going Corporate: A Geek's Guide* 一书的作者，也是 *Practical Business Analytics Using SAS: A Hands-on Guide* 一书的合著者。

他毕业于德里印度理工学院，获得设计工程专业硕士学位。他与妻子 Meenakshi、女儿 Neha 和儿子 Vivek 在班加罗尔生活。

致 谢

本书是团队成员始终如一、不辞辛苦工作的结晶，具有重要的意义。我们衷心感谢麦格劳-希尔教育有限公司的编辑 Lara Zoble，她一直鼓励我们做高质量的工作，并在规定的时间内完成本书。她适时的建议对这本书的形成起到了重要作用。我们也要感谢文字编辑和整个编辑团队。没有他们的鼓励和努力，就不可能完成本书。特别是文字编辑，在完成本书的收尾工作时，表现出了不厌其烦的工作态度并高质量地完成了本书的编辑。最后，我们衷心感谢麦格劳-希尔教育有限公司的全部生产、分销和营销团队，他们在本书的背后做出了很大贡献。

——Venkata Reddy Konasani
Shailendra Kadre

我感谢帮助我完成本书的所有人。数学系学生 Vrunda Paranjape，他耐心地测试了所有代码文件，并确保代码准确无误。我要感谢 Pradeep Venkataramu、Mohan Silaparasetty、Amiya Ranjan Bhowmick、Bhunnesh Kumar 和 Vijay Krishna(V2K)，他们鼓励我，使我成为更好的企业培训师。

我衷心地感谢我的朋友 Debendra 和 Sumit，以及我的培训项目中的很多学生，他们总是鼓励我将我的授课讲义改编成本书。本书的顺利完成离不开我的妻子、儿子、小女儿和其他家庭成员的牺牲和耐心，他们总是看到我工作到很晚。

我还要特别感谢 Patrick Winston(计算机科学家)、Andrew Ng(计算机科学家)和 Nando de Freitas(教授)，感谢他们给了我无尽的启发。

——Venkata

我的咨询工作始于我在 2011 年出版的 *Going Corporate: A Geek's Guide* 一书。这是一本面向极客的综合管理咨询图书。我的第二本书是 2015 年出版的 *Practical Business Analytics Using SAS: A Hands-on Guide*，该书是我与 Venkata 合著的。2021 年出版的 *Machine Learning and Deep Learning Using Python and TensorFlow* 是我在美国出版的第三本书。这是我和 Venkata 合著的第二本书。Venkata 多年从事机器学习领域的专业培训工作。他的培训经验对本书的内容至关重要。

对于本书，我要衷心地感谢我的家人和朋友 Srinivas D.、Milind Kolhatkar、Laxmi Sahu 和 Anup Parkhi，在他们不知疲倦的评审中提出了许多有益的建议，给予我很多帮助和鼓励。我的兄弟 Shailesh Kadre 和他的妻子 Neena Kadre 也在评审和建议方面提供了很多帮助，我衷心地感谢他们。与其他技术类图书一样，本书参考了该领域众多专家和杰出研究人员的工作。我必须感谢这些激励我的专家。最后但同样重要的是，我要感谢所有直接或间接帮助我完成这个项目的人。由于篇幅有限，这里不一一列出他们的名字。

不站在读者的角度编写的图书是不会受欢迎的。在设计和编写本书时，我们始终牢记专业人员

IV 机器学习和深度学习：原理、算法、实战(使用 Python 和 TensorFlow)

和全球学生群体的需求。我们让本书的内容保持实践性和以问题为导向。本书的全部内容几乎都是根据真实案例研究编写的。所有实用的数据分析概念的解释都与案例研究相结合。本书的语言通俗易懂，便于读者理解。我希望你会喜欢本书，本书是机器学习领域的一本入门书，涵盖了高级机器学习的主题，用 5 章专门讨论了神经网络和深度学习。

编写任何图书都会给作者的家庭带来一定的困难，编写本书也不例外。我能按时完成这本书，要归功于我的妻子 Meenakshi、女儿 Neha 和儿子 Vivek。在整个图书编写过程中，他们每个人都坚定地支持我。在编写本书的最后几个月，由于 COVID-19 导致的封闭状态，家中没有了家政服务。Meenakshi 不仅白天完成了日常工作，还出色地完成了所有的家务，照顾我们的孩子和年迈的父母。没有她的支持和鼓励，我是无法按时完成本书的。

祝你好运！

——Shailendra

前言

目前，市面上有关机器学习的图书主要包括两类，分别是面向学术研究的介绍机器学习理论方面的图书和代码手册类的图书。面向学术研究的机器学习理论的图书介绍了在机器学习算法中涉及的数学推导与公式，但对数据的实际应用涉及得很少。对于没有良好的统计或数学方面理论基础的读者来说，很难理解该类图书的内容。这些涉及机器学习原理的图书介绍了数据科学从业人员所面临的现实挑战，却极少谈到机器学习方面的实践。代码手册类图书主要包含代码和相关文档，缺少编码的原因和执行具体任务的逻辑方面的内容。机器学习的学术研究和它如何在工业界使用之间是有一段距离的。因此，我们需要一本书能以机器学习理论为基础并包含与其相关的在工业界的实践，而且在这些实际案例中有符合逻辑的讲解。本书的宗旨是弥补上述两类图书的空白(学术研究与工业界应用之间的空白)。

我们准备写一本让普通读者容易读懂的书。任何初学机器学习的读者都可以从本书开始学习。本书每章的内容分为三部分，第一部分通过类比、实例及可视化的方式来介绍该章涉及的内容，第二部分利用数学公式以一种学术风格来帮助加深理解，第三部分结合数据介绍真实的案例并通过编写代码来解决问题，从而更深入地理解相关概念。

本书以通俗易懂的方式编写，向普通读者解释机器学习与深度学习的概念。作为本书的作者，我们保证无论你是否具有学术和编码的背景，都能从第 1 章学习到最后一章。有时读者可能会觉得解释这些概念用了太多的示例。这是因为本书严格遵循了 Python 的规则，即：

- 明了胜于晦涩
- 简洁胜于复杂
- 复杂胜于凌乱
- 间隔胜于紧凑

本书的主要内容

- (1) Python 基础和统计
- (2) 基本的机器学习模型
- (3) 高级的机器学习模型
- (4) 必要的深度学习模型

在开始学习机器学习前，必须要学习 Python 基础知识和统计方法。本书涵盖了 Python 基础知识，让较少接触编程和统计学的读者与其他读者站在同一起跑线上。

本书讨论了基本的机器学习算法。基本的机器学习算法分为线性回归、逻辑回归、决策树及聚类分析算法。这些方法并不复杂，易于创建、解释及可视化。对于这些主题，本书将以理论和实践

相结合的方式给你带来轻松的阅读体验。

本书深入介绍了一些高级机器学习方法，如随机森林、Boosting(提升方法)及神经网络。你将学会使用这些模型，以及深入了解这些模型中必要的超参数。你还将学习如何在工业界的实际应用中创建和验证这些模型。

本书还将介绍深度学习的概念。与讲解机器学习时使用的框架不同，在介绍深度学习时使用的框架是 TensorFlow 和 Keras。通过在 TensorFlow 和 Keras 框架中对深度学习中概念的实践，你将学会使用 CNN、RNN 和 LSTM 等深度学习模型。

我们旨在编写一本工作手册，使你可以掌握利用 Python 语言实现机器学习和深度学习的技能。为了达到阅读本书的最佳效果，建议读者能在阅读本书时动手编写并运行相应的代码。

本书的主要特点

- 深入详尽地涵盖了机器学习和深度学习的概念。
- 作者是具备多年工作经验的该领域的工业界专家。
- 涵盖了原理、工业界的最佳实践，以及专业人员在综合应用这些模型时所遇到的问题。
- 书中实例来源于真实的工业界案例，如银行、保险、电子商务、医疗服务及自动驾驶。
- 即使是较少接触统计、数学和编程的读者，也很容易阅读本书。
- 利用可视化和类比的讲解方法，让复杂的概念变得简单易懂。
- 不需要参考或阅读其他资源即可学习本书的内容。
- 提供了数据集、代码及项目实例的下载链接。

完成本书学习后，你将具备的能力

- 使用 Python 进行数据处理。
- 使用 Python 中的统计方法和生成报告进行数据探索。
- 掌握线性回归和逻辑回归模型的创建和验证，并能使用这些模型进行预测。
- 掌握基于树的模型的创建和验证，如决策树和随机森林。
- 理解模型创建的实际应用，如特征工程和模型选择。
- 掌握高级机器学习算法的专业知识，如 Boosting 和人工神经网络方法。
- 使用 TensorFlow 和 Keras 工具编程。
- 在创建深度学习模型时处理超参数。
- 理解计算机视觉并使用 CNN 模型对数据分类。
- 序列模型的创建和验证，如 RNN 和 LSTM。

目标读者

- 任何想要学习机器学习和深度学习的人
- 数据科学的爱好者和从业人员
- 具有数学或统计学背景的本科生和研究生

- 希望进入数据科学领域的报表分析师
- 希望利用机器学习和深度学习方法创建预测模型的人
- 希望利用机器学习和深度学习方法进行数据可视化的专业人员
- 计算机视觉的爱好者
- 深度学习的爱好者
- 计算机科学工程专业的学生

预备知识

- 本书是机器学习的启蒙内容。没有严格的预备知识。
- 任何具备学士学位的读者都可以阅读本书。
- 具备高中数学知识基础即可阅读本书。
- 不需要具备高级的统计知识。
- 不需要具备高级的编程知识。

书中的实例

本书涵盖了大量的实例和案例研究。具有代表性的案例如下：

- 航空旅客的案例研究——预测一家航空公司的旅客数量
- 客户流失的案例研究——基于客户的使用情况预测一个电信公司的客户流失情况
- 客服中心客户调查案例研究——预测客户对客服中心的满意度
- 金县(King County)房屋价格预测案例——基于房屋特征预测金县的房屋价格
- 皮马印第安人糖尿病案例研究——基于诊断指标预测患者患糖尿病的概率
- 银行贷款案例研究——提供贷款前预测风险客户
- 零售业顾客细分的案例研究——基于零售业公司的顾客购买行为实现顾客细分
- 交通事故预测——基于传感器数据预测致命的交通事故
- 基于美国人口普查数据的收入预测——基于人口普查数据预测高收入人群
- 通过输入带数字的图片实现图片中的数字识别
- 通过输入目标图像实现目标检测
- 根据输入的单词序列预测下一个单词
- 机器翻译——将英语翻译成目标语言

软件和硬件的准备

软件下载链接

- Anaconda: 从 <https://www.anaconda.com/distribution/> 网站下载 Anaconda

软件版本

- Python 3.7 及以上版本
- TensorFlow 2.0 及以上版本

源码和数据集的下载

源码、数据集和参考文献请扫描封底二维码下载。

更新和勘误

GitHub 地址：https://github.com/venkatareddykonasani/ML_DL_py_TF_errata。

目 录

第 1 章 机器学习与深度学习概述	1	2.2.1 使用 Spyder IDE	20
1.1 人工智能与机器学习的历史	2	2.2.2 第一个代码示例	21
1.2 机器学习项目的基础	3	2.2.3 错误和错误信息	23
1.3 机器学习算法与传统计算机程序	4	2.2.4 命名约定	24
1.4 深度学习的工作原理	4	2.2.5 使用 print 输出消息	25
1.5 机器学习与深度学习的应用	5	2.3 Python 中的数据类型	25
1.5.1 日常生活中的应用	5	2.3.1 数值	25
1.5.2 机器学习在制造业中的应用	5	2.3.2 字符串	26
1.5.3 机器人技术的应用	6	2.3.3 列表	27
1.5.4 银行与金融领域的应用	6	2.3.4 字典	28
1.5.5 深度学习的应用	7	2.4 Python 中的包	30
1.6 本书的组织结构	7	2.4.1 NumPy	32
1.7 预备知识——数学基础	9	2.4.2 Pandas	33
1.8 术语	9	2.4.3 Matplotlib	34
1.9 机器学习——扩展视野	10	2.4.4 ScikitLearn	35
1.9.1 人工智能发展现状	10	2.4.5 nltk	35
1.9.2 人工智能未来的影响力	11	2.4.6 TensorFlow	35
1.9.3 与人工智能相关的伦理、社会和 法律问题	12	2.5 Python 中的条件语句和循环 语句	35
1.10 Python 及其作为机器学习语言的 潜力	13	2.6 数据处理与使用 Pandas 深入 分析数据	37
1.11 关于 TensorFlow	13	2.6.1 数据导入和数据集的基本信息	38
1.12 本章小结	13	2.6.2 子集和数据筛选器	41
第 2 章 Python 编程与统计学基础	15	2.6.3 Pandas 中实用的函数	45
2.1 认识 Python	15	2.7 基本的描述性统计	47
2.1.1 为什么使用 Python	16	2.7.1 平均值	48
2.1.2 Python 的版本	16	2.7.2 中位数	49
2.1.3 Python IDE	17	2.7.3 方差和标准差	50
2.1.4 安装 Python	18	2.8 数据探索	52
2.2 Python 编程入门	20	2.8.1 探索数值型连续变量	53
		2.8.2 探索离散变量和分类变量	61

2.8.3 探索其他类型的变量	63	4.4 实例研究：基于客服中心的 数据对客户分类(细分)	132
2.9 本章小结	65	4.4.1 研究目标和数据探索	132
2.10 本章习题	65	4.4.2 使用 Python 构建模型	135
第 3 章 回归与逻辑回归	67	4.4.3 绘制决策树	136
3.1 什么是回归	67	4.4.4 理解树的输出	137
3.2 构建回归模型	70	4.4.5 验证树的准确率	142
3.2.1 获取回归系数	70	4.5 过拟合的问题	143
3.2.2 回归线示例	74	4.5.1 巨大的决策树	143
3.3 R-squared	77	4.5.2 训练数据和测试数据	145
3.4 多元回归	80	4.5.3 过拟合	146
3.5 回归中的多重共线性	83	4.6 决策树的剪枝	149
3.5.1 什么是多重共线性	85	4.7 欠拟合的挑战	152
3.5.2 多重共线性的检测	87	4.8 使用二分搜索法设置剪枝参数	152
3.5.3 计算 VIF	88	4.9 其他剪枝参数	154
3.6 回归中各变量的影响力	93	4.9.1 叶子节点的最大数量	154
3.6.1 P-value	93	4.9.2 min_samples_in_leaf_node 参数	155
3.6.2 P-value 的理论	96	4.10 构建决策树模型的步骤	156
3.7 构建多元回归模型的步骤	97	4.11 本章小结	156
3.8 逻辑回归模型	97	4.12 本章习题	157
3.9 构建逻辑回归模型	102	第 5 章 模型选择和交叉验证	159
3.10 逻辑回归线的准确率	104	5.1 构建模型的步骤	159
3.11 多元逻辑回归线	107	5.2 模型验证指标：回归	161
3.12 逻辑回归中的多重共线性	111	5.3 实例：华盛顿金县的房屋销售	162
3.13 变量的个体影响力	113	5.3.1 研究目标和数据	162
3.14 构建逻辑回归模型的步骤	116	5.3.2 构建和验证模型	165
3.15 比较线性回归模型与逻辑 回归模型	116	5.4 模型验证指标：分类	167
3.16 本章小结	117	5.4.1 混淆矩阵和准确率	167
3.17 本章习题	118	5.4.2 不平衡分类的度量	169
第 4 章 决策树	119	5.5 权衡偏差与方差	189
4.1 什么是决策树	119	5.5.1 过拟合的问题：方差	189
4.2 划分准则：熵和信息增益	123	5.5.2 欠拟合问题：偏差	190
4.2.1 熵：不纯度的度量	124	5.5.3 权衡偏差与方差	190
4.2.2 信息增益	125	5.5.4 案例研究：皮马印第安人糖尿病 预测	192
4.2.3 基尼系数：与熵类似的方法	127	5.6 交叉验证	195
4.3 决策树算法	128	5.6.1 交叉验证：一个示例	195
		5.6.2 K-折交叉验证	197

5.6.3 训练集-验证集-留出交叉验证方法	199	7.4 案例研究：交通事故预测	265
5.7 特征工程的技巧和诀窍	201	7.4.1 研究背景和目标	266
5.7.1 什么是特征工程	202	7.4.2 数据探索	266
5.7.2 创建哑变量或独热编码	205	7.4.3 构建模型和验证模型	268
5.7.3 处理经度和纬度	214	7.5 Boosting 算法	271
5.7.4 处理日期类型的变量	219	7.6 AdaBoosting 算法	271
5.7.5 数据转换	225	7.7 Gradient Boosting 算法	273
5.8 处理类的不均衡	229	7.7.1 Gradient Boosting 算法	273
5.9 本章小结	234	7.7.2 Gradient Boosting 算法在决策树上的应用	273
5.10 本章习题	234	7.7.3 Boosting 算法中的超参数	274
第 6 章 聚类分析	235	7.7.4 Gradient Boosting 算法的图解	275
6.1 无监督学习	235	7.8 案例研究：基于人口普查数据的收入预测	278
6.1.1 聚类分析	236	7.8.1 研究背景和目标	279
6.1.2 案例研究：批发客户数据的客户细分	237	7.8.2 数据探索	280
6.2 距离的度量	241	7.8.3 数据清洗和特征工程	283
6.2.1 欧几里得距离	242	7.8.4 构建模型与验证模型	286
6.2.2 距离矩阵	243	7.9 本章小结	289
6.3 K-means 聚类算法	244	7.10 本章习题	289
6.3.1 聚类算法的步骤	245	第 8 章 神经网络	291
6.3.2 K-means 算法：图解	247	8.1 逻辑回归的网络图	291
6.3.3 K-means 聚类算法：输出	248	8.2 决策边界的概念	293
6.4 构建 K-means 聚类模型	248	8.3 多决策边界问题	299
6.5 聚类簇数量的选取	254	8.4 多决策边界问题的解决方案	302
6.6 本章小结	257	8.5 直观理解神经网络	309
6.7 本章习题	257	8.6 神经网络算法	312
第 7 章 随机森林和 Boosting	259	8.6.1 神经网络算法：非技术方式	313
7.1 集成模型	259	8.6.2 神经网络算法：数学公式	315
7.1.1 群体智慧	259	8.6.3 神经网络算法：示例	316
7.1.2 构建集成模型的方法	261	8.7 梯度下降的概念	319
7.2 Bagging	261	8.7.1 梯度下降在回归模型中的应用	320
7.2.1 自助采样法	262	8.7.2 学习率	321
7.2.2 Bagging 算法	262	8.7.3 梯度下降在回归模型中的编码	321
7.3 随机森林	263	8.7.4 一题多解	323
7.3.1 随机森林算法	264	8.8 案例研究：手写数字识别	324
7.3.2 随机森林中的超参数	265	8.8.1 研究背景和目标	324
		8.8.2 数据	326

8.8.3 构建模型	329	10.5.1 Sigmoid 函数	386
8.8.4 模型预测与验证模型	334	10.5.2 Tanh 函数	387
8.9 深度神经网络	336	10.5.3 ReLU 函数	390
8.10 本章小结	337	10.5.4 Softmax 函数	393
8.11 本章习题	337	10.5.5 激活函数的编码	393
第 9 章 TensorFlow 和 Keras	339	10.6 学习率	395
9.1 深度神经网络	339	10.6.1 学习率的应用示例	397
9.2 深度学习框架	340	10.6.2 设置学习率的代码	401
9.2.1 什么是 TensorFlow	341	10.6.3 动量	405
9.2.2 计算图	341	10.7 优化器	407
9.2.3 Python Notebook	343	10.7.1 SGD—Stochastic Gradient Descent	407
9.2.4 安装 TensorFlow	348	10.7.2 小批量梯度下降法(Mini-Batch Gradient Descent)	408
9.3 TensorFlow 中的关键术语	349	10.8 本章小结	410
9.4 使用 TensorFlow 构建模型	352	第 11 章 卷积神经网络	411
9.4.1 使用 TensorFlow 构建回归 模型	352	11.1 用于图像的人工神经网络	412
9.4.2 使用 TensorFlow 构建逻辑回归 模型	355	11.1.1 空间依赖性	412
9.5 Keras	357	11.1.2 ANNs 中自由参数的数量	414
9.5.1 什么是 Keras	357	11.2 卷积核	414
9.5.2 使用 Keras	358	11.2.1 卷积核的工作原理	415
9.5.3 在 Keras 中应用 MNIST 数据集的 示例	358	11.2.2 利用卷积核执行特征检测	417
9.6 本章小结	361	11.2.3 卷积核中的权重	422
第 10 章 深度学习中的超参数	363	11.3 卷积层	422
10.1 正则化	363	11.3.1 在 Keras 中使用卷积层	423
10.1.1 回归模型中的正则化	364	11.3.2 彩色图像的卷积核	425
10.1.2 L1 正则化与 L2 正则化	368	11.3.3 零填充	429
10.1.3 在神经网络中应用正则化	368	11.3.4 跨步卷积(stride)	435
10.1.4 L1 正则化和 L2 正则化的 编码	375	11.4 池化层	437
10.1.5 在 L1 正则化和 L2 正则化中 应用数据标准化	377	11.4.1 池化的工作原理	437
10.2 随机丢弃正则化	377	11.4.2 为什么进行池化	437
10.3 早停法	381	11.5 CNN 模型的架构	440
10.4 损失函数	384	11.5.1 CNN 模型中的权重	441
10.5 激活函数	386	11.5.2 CNN 模型的编码	443
		11.6 案例研究：从图像中识别 手势	447
		11.6.1 项目背景和研究目标	447
		11.6.2 数据	448

11.6.3 构建模型和验证模型	450	12.4 RNN 模型	491
11.7 规划理想的 CNN 模型	455	12.4.1 时间反向传播	492
11.7.1 卷积层和池化层的数量	455	12.4.2 计算参数的数量: 一个 示例	494
11.7.2 卷积层中卷积核的数量	456	12.4.3 构建 RNN 模型的示例代码	496
11.7.3 批量归一化(Batch Normalization)	464	12.4.4 使用 RNN 模型预测单词	498
11.7.4 选择优化器	471	12.5 使用 RNN 模型处理长序列	501
11.8 构建 CNN 模型的步骤	475	12.5.1 案例研究: 预测组成下一个 单词的字符	501
11.9 本章小结	475	12.5.2 梯度消失问题	511
11.10 本章习题	476	12.6 长短期记忆模型	512
第 12 章 RNN 与 LSTM	477	12.6.1 LSTM 的控制门	513
12.1 横截面数据和序列数据	477	12.6.2 直观理解 LSTM	518
12.1.1 横截面数据	477	12.6.3 应用 LSTM 的案例研究	519
12.1.2 序列数据	477	12.7 序列到序列模型	523
12.2 在序列数据上构建模型	478	12.8 案例研究: 语言翻译	529
12.2.1 在序列数据上使用 ANN 模型	478	12.8.1 研究目标和数据	529
12.2.2 在序列数据上使用 CNN 模型	479	12.8.2 数据预处理	530
12.2.3 处理序列数据的 ANN (Sequential ANN)	479	12.8.3 编码和解码	532
12.3 案例研究: 单词预测	481	12.8.4 构建模型	533
12.3.1 研究目标和数据	481	12.8.5 用模型做预测	534
12.3.2 数据预处理	484	12.9 本章小结	538
12.3.3 构建模型	487	12.10 本章习题	538
12.3.4 预测	490	第 13 章 参考文献(在线资源)	539

第1章

机器学习与深度学习概述

人工智能(Artificial Intelligence, AI)专家 Winston 在 1992 年就提出了人工智能的定义：人工智能是“使感知、推理、执行成为可能的计算研究”。这个定义至今仍然是成立的。然而，随着过去 25 年科技翻天覆地的变化，人工智能技术以前所未有的速度发展。早期的类人机器人(robot)就像我们一样直立行走，执行一些人类特有的日常任务。2015 年后，社交媒体上到处都是类人机器人表演的各种特技——跳舞、跳跃、倒立、翻筋斗。这些机器人表演的一些特技，即使对人类来说也是不容易的。近来，基于人工智能的人形机器人索菲亚被授予沙特阿拉伯公民身份，成为有史以来第一个拥有国籍的机器人。索菲亚以采访世界领导人和采访名人而闻名。索菲亚甚至在联合国发言，当然，这一切都是没有任何人为干预的。毫无疑问，这正是今天的人工智能发挥了作用。

然而，大多数用于工业和家庭的机器人并不是类人机器人。它们的外型取决于它们所执行的任务。不过，有一个事实是共同的——它们都内置了某种程度的智力。所有这些机器人和其他智能自动机背后的大脑就是我们所说的人工智能(本书中通篇与机器学习同义使用)，我们可以放心地称机器学习和深度学习算法是人工智能研究的基石。在后面的章节中，我们将详细讨论其中的诸多算法，但本章的内容仅限于作为对本书所讲内容的介绍。本章还会介绍一些在现阶段听起来可能有些陌生的术语，会对其中的部分术语加以解释。

人工智能是一个具有许多基础功能的多学科交叉研究领域。这里选取几个简单的场景进行讨论。哲学帮助人工智能科学家回答这样的问题：我们能用形式规则得出有效的结论吗？我们能利用今天现有的知识库让机器采取类似人类的行动吗？数学是人工智能的核心支柱，它帮助人工智能专业人员将规则形式化，从而得出有效的结论，我们如何才能用模糊信息进行推理呢？整个信息技术(IT)和计算领域(包括机器学习)严重依赖于现代数学的发展。人工智能在很大程度上借鉴(依赖)当今的神经科学，神经科学涉及神经系统，特别是大脑的研究。所有先进的人工智能系统的目标都是尽可能精确地模仿人类的思维。但在人工智能研究现状下，我们离这个目标还很遥远。想要达到这个目标可能还需要几十年的时间。

人工智能系统从心理学，特别是认知心理学中得到了很多启发，认知心理学把人的思维看作信息或数据的处理装置。动物和人类如何思考(和行动)的研究正在进行。所有这些知识都有助于人工智能系统的发展。控制理论与神经机械学领域也对机器学习和人工智能领域的相关研究作出了重要贡献。计算机科学与工程领域帮助构建高效的计算资源，来运行目前复杂且资源高度密集的人工智能算法(例如，具有数百万参数的深度学习算法)。

人工智能的前沿研究在当今的专家和学生中非常流行。全球许多领先的教育机构在其计算机科

学系提供人工智能研究生学位。麻省理工学院(MIT)是国际领先的大学之一，甚至成立了一个单独的学院来推动机器学习和人工智能领域的发展。

1.1 人工智能与机器学习的历史

“机器学习”这个术语是在 20 世纪 50 年代被创造出来的。直到 20 世纪 80 年代，人们一直在做各种尝试，让计算机程序能像简易版人脑一样工作。正如我们所知，在 20 世纪 90 年代，机器学习从知识驱动方法转向了数据驱动方法。1997 年是 IBM 的“深蓝”计算机在国际象棋上击败世界冠军的一年，而国际象棋是一项需要达到一定的人类智力才能进行的比赛。机器学习被公认是一个独立的领域，并在 20 世纪 90 年代开始蓬勃发展。它开始借鉴统计学和概率理论领域方法进行发展。机器学习的研究发展还受到数字化过程中数据可用性的不断增加以及互联网作为分发渠道的可用性推动。Geoffrey Hinton 在 2006 年创造了“深度学习”一词。深度学习的早期应用之一是，在图像和视频计算机被训练去观察和分辨物体。从那时起，人工智能的开发不断被领先的大学和像 IBM、谷歌和亚马逊这样的技术主力推向更高的水平。

在 21 世纪 10 年代的后半期，我们正在目睹从企业决策中使用的简单分析方法，转向在各种机器中交付更复杂的任务并使用更复杂的深度学习算法，其中包括更先进的类人机器人。人工智能逐渐被认为是未来几年人们最有希望涉足的领域之一。

下面给出了工业界流行的机器学习算法列表：

- 线性回归
- 逻辑回归
- 决策树
- 随机森林
- 梯度提升机(gradient boosted machines)
- 人工神经网络(Artificial Neural Networks, ANNs)
- 卷积神经网络(Convolution Neural Networks, CNNs)
- 循环神经网络(Recurrent Neural Networks, RNNs)
- 贝叶斯技术
- 支持向量机(Support Vector Machines, SVMs)
- 进化方法(Evolutionary approaches)
- 马尔可夫逻辑网络(Markov logic networks)
- 隐马尔可夫模型(hidden Markov model)
- 生成对抗网络(Generative Adversarial Networks, GANs)

下面是数据科学家广泛应用的编程语言列表：

- Python
- R
- SQL
- Java
- JavaScript

- MATLAB
- Scala
- Julia
- Go
- C/C++
- Ruby
- PHP
- SAS

根据项目需要，一名机器学习专家可能需要使用多种工具与编程语言。

1.2 机器学习项目的基础

机器学习，也被称为增强分析，被认为是人工智能的一个子集，与计算统计学密切相关。机器学习算法可以说是从某一组任务的经验中学习的计算机程序。这些任务执行的准确率会随着经验的积累而提高。为了达到目标准确率，机器学习算法需要用训练数据来进行训练。一旦经过训练，这些算法就会产生可以预测未来结果的能力。

机器学习在方法上与统计学的关联非常紧密，但它们的底层目标不同。描述性统计技术用于对种群(或一组数据)进行推断，而机器学习算法则寻找一般的预测模式。我们中的许多人已经了解一些关于机器学习的基本概念及简史。在各个行业中，机器学习可以用于管理和改进业务运营。然而，在使用模型做出一个预测之前，机器学习模型背后还有很多艰苦的工作。就像许多其他软件开发项目一样，为了构建一个有效的机器学习模型，数据科学团队需要通过对数据进行采集、处理及转换，来制定恰当的问题解决策略并创建原始模型。在将模型部署到实际生产环境之前，需要用大量的样本数据来训练该模型并验证结果。图 1.1 是模型构建过程的示意图，将在后面详细讨论。

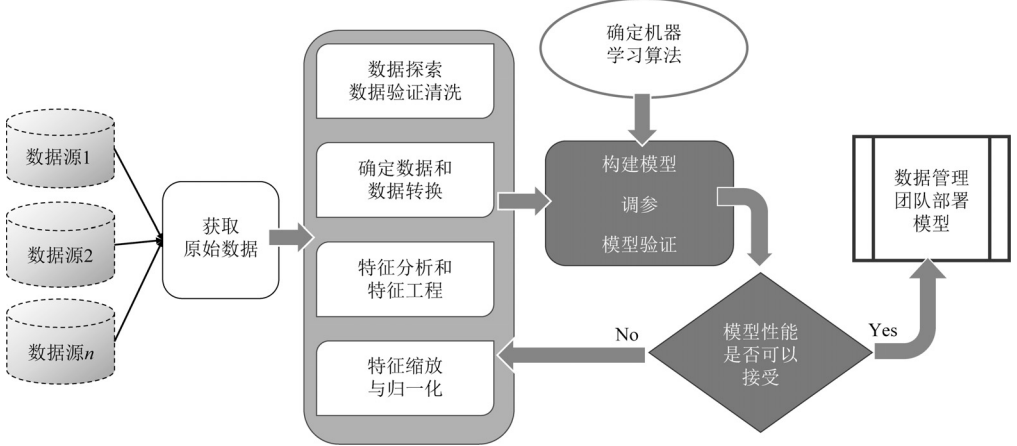


图 1.1 典型的机器学习示例

数据科学团队通常与 IT 部门一起工作。软件开发项目和数据科学项目之间有着根本的区别。软件开发项目围绕着代码的设计、开发和测试，而数据科学项目则是针对大量的数据，涉及的编码

工作较少。在所有机器学习项目中，都有 70% 的时间和资源会分配给数据采集、清洗，然后将数据放到可用的数据结构中(例如以表格的形式存储)。一旦数据和模型准备就绪，最重要的工作就是训练模型，训练过程通常需要大量的训练数据。一般银行和像亚马逊这样的零售网站都有大量可靠的用户交易数据，开发和部署机器学习技术是其日常业务运作的重要部分之一。

像很多其他项目一样，每个机器学习项目都需要解决一个业务问题。自然，第一步是阐述业务问题并概述战略业务目标；然后是确定任务范围。一旦任务完成并进行了适当的审查，项目所有人(通常是从业务方面)就会确定真正的利益相关者，并最终从高级管理层获得资金和所需资源的支持。

一个机器学习项目需要在项目组中有许多专家，其中一些人是数据专家，职责是以可用的格式给出干净的可用数据。业务分析师代表业务方帮助确定需求，而解决方案架构师则负责新解决方案的设计。项目团队中的数据科学家最初可能会研究多个机器学习模型，并对它们进行训练，评估哪个模型给出的结果最准确。模型评估和测试阶段的目标是得到一个最简单的模型，在能达到目标准确率的同时快速完成任务。

一旦数据科学家最终确定了最可靠的模型并量化了其性能需求，就可以将解决方案部署到生产中，这意味着它可以应用到实际业务中。数据库管理员通常负责将最终的解决方案部署到生产中。本节讨论的所有团队角色取决于项目规模和团队结构，可能因组织的不同而有所差异。

1.3 机器学习算法与传统计算机程序

众所周知，传统编程已经经过了一段时间的发展。我们都很熟悉计算机程序，大概每个人在高中或本科期间都做过一些编程(计算机程序)。正如我们所知，传统的编程通常是手工过程。一个人(程序员)编写计算机代码不过是使用许多现有编程语言中的规则。计算机能理解这个代码；然后，一个人将输入数据提供给计算机程序，根据代码规则得到所需要的输出。

另一方面，机器学习算法基于样本数据(称为训练数据)建立数学模型。模型一旦训练好，该模型就可以用来做某类预测或决策；执行任务则不需要太多显式的编程。随着输入的训练数据越来越多，预测或决策的准确性会越来越高。历史数据的输入和输出都是已知的，预测的准确性可以通过历史数据来衡量。一旦得到了期望的准确率水平，训练好的模型就可以用来对新的输入数据进行预测。这就像人类通过经验积累变得更好一样。机器学习模型几乎在每个行业都有应用，银行和其他金融机构用已知的历史数据训练模型，然后使用这些模型来预测新客户拖欠贷款的概率，这些模型中输入的数据是客户的统计数据，如收入、信用卡数量、资产和一些默认参数。

1.4 深度学习的工作原理

虽然有些机器学习模型只是简单的统计练习，但深度学习可能会引起很多人的兴趣，因为它是一个相对较新的现象，“深度学习”这个术语是在 2006 年提出的。深度学习，又称分层学习，是机器学习算法的一个子集，给定原始输入数据，深度学习逐步从多个层次中提取更高层次的表征。深度学习技术的一个常用领域是图像处理。例如，在医学图像处理应用中，给定原始像素矩阵的图像，第一个深度学习层可以仅识别图像边缘，第二、第三和第四层可以加密和合成眼睛、鼻子和面部。

深度学习过程可以自行决定哪一层提取边缘、人脸(faces)、眼睛、鼻子、面部(face); 换句话说, 它足够智能, 可以自行决定表征提取的不同层次。

一般来说, 深度学习过程可以学会解决任何需要人类思考才能解决的问题。在这项技术中, 人工神经网络技术和算法从大量数据中学习时会模仿人类大脑。我们将在本书后续章节中介绍神经网络和不同类型的学习方法。深度学习需要大量的数据来训练模型, 然后才能将其投入生产中使用。随着数字时代飞速发展, 大量的数据及其处理技术不断发展, 这是深度学习在最近几年才得到进一步发展的原因。

1.5 机器学习与深度学习的应用

机器学习在各个行业中都有广泛的应用, 无论是金融业和银行、制造业, 还是一般自动化或机器人技术等, 不一而足。到 2021 年, 至少 1/5 的领先制造商将依靠嵌入式智能技术, 并且使用人工智能和其他相关技术来实现流程自动化。

1.5.1 日常生活中的应用

自动驾驶汽车如今大受欢迎。它们装有摄像机、传感器、激光器和测距仪来感知环境。自动驾驶汽车基于人工智能软件控制着转向、制动和加速, 它检测和区分道路上的物体, 并像人类驾驶员一样调整自己。预计这些自动驾驶汽车很可能会大大降低由人为错误引起的事故发生率。语音识别是另一个基于机器学习且具有很潜力的领域, 它可以影响我们与机器, 甚至是与使用我们不熟悉的语言的人的交互方式。来自美国的旅行者可以在度假网站上利用自动语音识别技术和对话管理系统引导的整个对话来预订机票。现在这是一个很现实的应用。语音识别在许多传统意义上以人为主导的领域得到了广泛的应用, 如大公司的客户服务中心。

1997 年, IBM 著名的“深蓝”计算机成为第一台在表演赛中击败国际象棋世界冠军加里·卡斯帕罗夫的人造机器。人工智能技术在游戏中发挥了很大的作用, 并且让比赛变得更加精彩。你的电子邮件消息系统正用人工智能过滤垃圾邮件。通过学习算法, 每天都有数十亿条信息被归类为垃圾邮件。机器翻译是计算机自动将一种语言翻译成另一种语言(翻译质量可以接受的情况)的另一个人工智能领域。翻译软件从本地语言到目标语言的翻译样本构建了一个复杂的统计模型, 例如从意大利语到英语。这些示例样本可能会处理数万亿字的数据。

1.5.2 机器学习在制造业中的应用

利用机器学习算法, 工厂正在减少计划外停机时间, 降低维护成本, 提高生产产量, 并实现高达 30% 的产品质量提升。机器学习正被用于简化生产的每个阶段——优先领域是供应链优化、提高原材料的产量、质量控制、工作进度, 以及从众多组合中选择最优的产品配置。利用机器学习算法, 制造商可以发现一个产品及其包装中的异常。它能显著提高产品质量, 防止有缺陷的货物离开制造设施。在产品质量控制中, 使用机器学习的缺陷检测效率几乎可以比许多行业仍然盛行的人工检测技术的效率提高一倍。机器学习和人工智能技术具有巨大的潜力, 可以提高生产吞吐量, 最大限度地提高每小时的收入潜力, 并通过识别和预防浪费将能源利用率提高到最佳水平。在制造业, 这些

技术的早期采用者看到，由于效率的提高，他们的现金流甚至翻了一番。几乎在其业务周期的每一个阶段都可以看到显著的改进，包括提高产品质量，最大限度地利用工厂和人力资源，减少几乎所有地方的浪费。

化工和制药等基于流程的制造商通过提高生产效率和产品质量，利用数据和人工智能技术(如机器学习)，使利润大大增加。即使在办公室工作中，机器学习和其他人工智能技术也可以减少重复性和常规性任务，如搜索发票号码、纸张处理和估计订单价值。在另一个不同的例子中，欧洲的一家制造厂对其所有重型机械(如用于钣金生产的气动压力机)实施了预测性维护解决方案。在此之前，工程师浪费了很多时间处理故障，而不是进行计划的维护活动。新的维护解决方案使他们能够以 90% 以上的准确率预测设备故障，并实施更好的维护规划。这不仅提高了设备的可靠性和正常运行时间，而且提高了产品的整体质量。

在制造业应用中，因为有监督的机器学习算法随后产生预定义的目标，实践中经常使用这类算法。分类和回归类型的机器学习算法在预测性设备维护中得到了广泛的应用。

1.5.3 机器人技术的应用

机器人技术也在某种程度上受到机器学习技术的影响和指导。它在自动驾驶汽车、无人驾驶飞机和其他视觉敏感任务中得到了应用，比如，检测和评估硅片识别和物体分类中的故障。为了保持生产率，机器人使用机器学习技术进行动态交互和障碍规避。如我们所知，医疗领域的机器人广泛使用机器学习技术。如今，医疗机器人能够完成手术和缝合等任务，胜过目前最佳的人类外科医生。这些工作一般是由训练有素的医生在医疗监督下进行的。亚马逊等在线零售商非常有效地在巨型仓库中使用基于机器学习的机器人来挑选和放置数十万形状大小各异的货物。周围有很多这样的机器人，它们使用机器学习，通过视觉、抓握、运动控制、识别物理和逻辑数据模式，采取主动行为来完成不同的任务。就机器人技术而言，机器学习技术仍处于起步阶段，但已经产生了重大影响。

我们将在后面的章节中讨论基于卷积神经网络(CNNs)检测和测量硅片中故障的系统。另外，一些机器人与控制任务使用自主学习模型，其中涉及深度学习和无监督的方法。

1.5.4 银行与金融领域的应用

银行和金融界可能是机器学习模型最重要的用户，是最早采用这些技术的领域之一。世界各地的银行和金融公司通常拥有大量准确的客户交易历史记录与人口统计信息。这些领域以定量计算为主，因此它们可以有效地使用机器学习算法进行各种金融建模。随着机器学习工具可访问性和准确性的提高、精确数据的可用性不断提升以及计算系统的改进，机器学习应用领域越来越多。机器学习在金融风险建模、客户信用评分、贷款审批等方面都有应用。一些机器学习算法也用来标定客户的投资组合，使其符合设定的财务目标和风险偏好。例如，一位职业妇女设定了一个在 50 岁时储存 40 万美元退休金的目标。根据她目前的收入、资产和风险偏好，算法专家可以非常有效地建议她如何将投资分散到不同的资产类别，以实现这一既定目标。在全球范围内，许多高级财务顾问通常使用这种基于机器学习的软件，整个过程在经验丰富的财务顾问的协助下可以更高效地运作。由于这一过程的大部分是自动化的，当客户的业务目标、风险偏好和任意控制参数随时间变化时，顾问可以轻松地建议许多投资组合，这是一个很自然的过程。机器学习模型在银行业的另一个主要应

用是信用卡或贷款还款中的欺诈检测。基于输入数据(如客户人口统计信息和客户的历史交易记录),这些模型以 90%以上的准确率提醒安全团队客户可能的违约或欺诈行为。为了达到这一准确率,大型金融机构甚至使用数以百万计的客户数据样本来训练它们的欺诈检测模型,这些数据包括客户的收入、信用卡数量、贷款数量、工作、年龄、婚姻状况、保险详情等。现在,为了实现利益最大化,几乎所有全球领先的银行都具备人工智能功能,这些银行正在积极部署基于机器学习的模型和其他基于人工智能的技术。

简单线性回归在预测(或预报)和财务分析中得到了广泛的应用。逻辑回归是一种广泛使用的统计技术,用于计算意向分数。这类算法中所有选定的协变量同时包含在一个逻辑回归模型中,用来预测任务状态。意向分数是每个单元的合成预测概率。在金融、投资或商业领域中,经常利用决策树形成一个或多个行为过程。深度学习在金融领域中用于制定策略,如高频交易(HFT),其同样可以应用于银行欺诈检测。后续章节会详细讨论这些机器学习算法。

1.5.5 深度学习的应用

深度学习是机器学习的一个子集。上面讨论的大多数应用可能已经涉及深度学习算法。机器人和其他人工智能机器将深度学习技术应用于手写识别、物体检测和分类、图像分割及图像重建等视觉任务。新闻机构利用深度学习算法来识别并过滤掉负面新闻。利用数十万个基于深度学习的模型(相互协调),自动驾驶汽车可以在繁忙的道路上行驶,同时有效地避免交通事故的发生。例如,一个模型可能擅长识别行人,而同一辆自动驾驶汽车中的其他模型能精确地识别道路上的交通信号灯及快速移动的物体。所有这些工作都需要相互协调来避免意外发生;而在非自动驾驶车辆中,这些工作通常由人脑在眼睛和其他感知器官的帮助下完成。

你可能已经意识到了另一个深度学习应用案例,自动化的机器人记者已经在为各个新闻机构撰写故事和报道。这种现象已经存在了两三年。自动将一种语言翻译为另一种语言的技术已经存在了很多年,例如将德语翻译成英语。而将深度学习算法应用到这项任务中可以获得极高的准确度。现在可以将给定的单词、短语或句子更准确地翻译成另一种语言,同时确保翻译文字的经典性。用户越多地使用该程序进行翻译,翻译的准确性就会越高。深度学习在游戏行业也有着广泛的应用,在游戏领域中模型学习了一个非常复杂的任务,即只根据游戏屏幕上显示的像素来玩一个计算机游戏。一些性能最好的深度学习系统是在过去十年才出现的,例如苹果手机与谷歌最新的自动翻译器等智能手机上的语音识别器。

1.6 本书的组织结构

在这本书中,我们尝试涵盖了几乎所有重要的机器学习和深度学习算法。围绕各种机器学习和深度学习算法的讨论分为 12 章。每一章都是一个完整的单元,涵盖了必要的理论和数学知识,以及案例研究和在实际数据上执行的算法的代码。

第 1 章:机器学习与深度学习概述。本章介绍了机器学习和开发机器学习模型的各个阶段。举例讨论了机器学习模型如何应用于工业领域。本章列出了所有流行的机器学习和深度学习算法,以及用于开发这些算法的工具。

第2章：Python 编程与统计学基础。本章介绍 Python 编程，包括用于开发的 Python 中的基本命令。讨论了 Python 中必要的软件包，如数据处理技术(子集、筛选和创建新变量)。阐述了统计学的一些基本概念，如均值、中位数和百分位数。本章还对 Python 中的数据填充做了一些介绍。

第3章：回归与逻辑回归。本章解释了什么是回归线，以及在项目中如何创建一个回归模型。阐述了 R-squared 作为模型性能度量的概念。然后讨论如何建立多元回归模型，包括多重共线性等概念。本章还涉及如何衡量变量的个别影响等问题，介绍了逻辑回归直线和逻辑回归模型的发展，讨论了逻辑回归中的多重共线性，以及如何衡量单个变量对整体模型的影响。

第4章：决策树。本章介绍了一些分割(拆分)技术、熵的概念和信息增益。详细讨论了决策树算法，并基于 Python 对该算法建立了决策树模型。验证决策树、决策树模型的剪枝和剪枝参数等概念也得到了足够详细的讨论。

第5章：模型选择与交叉验证。本章解释了模型验证技术、训练误差和测试误差。介绍了受试者工作特征(ROC)曲线的产生和曲线下面积(AUC)的测量方法；以及模型过拟合和欠拟合的概念。本章还引入了偏差和方差之间的权衡，以及模型交叉验证技术，如 K-折交叉验证。本章阐述的工程技巧在实际项目中非常有用。

第6章：聚类分析。本章介绍了无监督学习。本章结合距离测度和距离矩阵的概念，深入讨论了 K-means 聚类算法。

第7章：随机森林和 Boosting。本章首先讨论了群体和集成模型的智慧之处。本章深入讨论了随机森林算法，它是 Bagging 算法的一种特殊形式。本章还详细讨论了随机森林中的超参数以及如何对其进行微调。在此基础上，讨论了自适应提升算法(AdaBoosting)与梯度提升算法(Gradient Boosting)，以及提升算法中的关键超参数。

第8章：人工神经网络。本章从对决策边界的解释开始，延伸到隐藏层与反向传播算法的问题。本章讨论了建立神经网络模型的所有步骤。本章对梯度下降以及它如何解决神经网络优化问题也有深入的说明。

第9章：TensorFlow 和 Keras。本章讨论了几种深度学习软件包，并进行了比较。我们沿着 TensorFlow 编程范例的构建模块学习了 TensorFlow 框架。接着学习了 Keras 及其特性。读完这一章，将可以熟悉 TensorFlow 和 Keras 中一些非常有用的命令。

第10章：深度学习中的超参数。本章讨论了在处理深度学习算法时必须知道的一些非常重要的超参数。在本章中，我们讨论了一些基本的深度学习概念，如正则化、学习速率、动量、激活函数及各种优化函数。

第11章：卷积神经网络。本章讨论了卷积核、卷积层和池化层等概念，并且讨论了卷积神经网络的详细工作。你将了解 CNN 的各种参数以及如何对其进行微调。学习完本章之后，你将能够构建一个带有实时数据的最优卷积神经网络。

第12章：RNN 与 LSTM。本章首先讨论时序模型，并介绍循环神经网络模型。接着讨论一些基本的概念，如通过时间的反向传播，以及与 RNN 相关的许多其他概念。你将了解梯消失的问题，并学习长短期记忆(LSTM)的概念。你还将学习其他基本概念，如控制门以及 LSTM 如何使用这些控制门。

1.7 预备知识——数学基础

机器学习是关于数据和算法的。机器学习研究需要一定的数学基础。要理解这本书，只需要高中水平的数学能力。正如将在后续的章节中所看到的，最晦涩的数学分析步骤由软件来处理，所以你不必为错综复杂的数学推理或数学理论的细节而担忧。如果你的基本知识扎实，就一定能在更好地理解整个过程的同时轻松解析结果。为了在机器学习领域达到精通的水平，这里给出了一个基本数学主题的列表。想要对这些主题有一个全面的认识，需要参考标准的数学教科书。在更广泛的范围内，需要以下方面的一些基础技能。

- 函数、方程和图形
- 基础微积分
- 优化理论
- 向量和矩阵
- 基本统计学与概率论

为了快速上手，建议按具体主题攻克它。

- 线性代数：标量、向量及其运算的定义；特征值和特征向量；常用的矩阵运算
- 初等微积分：函数、导数、偏导数、微分、梯度、基本积分
- 基本统计学和概率论：总体、样本、变量及其分类、均值、中位数、众数、参数、统计量、分布测度、定量和定性分析实例、概率论的基本理论和相关概念

除了所有这些至关重要的数学基础，如果你还能学习微分方程、向量、矩阵微积分和梯度算法，将大有裨益。

1.8 术语

算法：它们可能是人工智能领域的第二个重要组成部分，第一个是可靠的数据。在人工智能语境中，算法是赋予机器学习和深度学习模型，或者机器中的神经网络的一组规则或命令，以便它能够进行自我训练，其中回归、分类和聚类是最常用的算法类型。

认知计算：任何一种能在数据挖掘、模式识别和自然语言处理等技术的帮助下像人类大脑一样进行自我训练的基于计算机的模型或算法。

自然语言处理(NLP)：自然语言处理技术通常使用人工智能来识别以任何语言(如英语或德语)进行的人类演说或对话。NLP 算法的语言解释应该与人类理解语言的方式非常相似。

监督学习：监督学习的工作方式与教师在课堂上训练学生非常相似，其中分类和回归算法就属于这一类。与另一类机器学习(无监督学习)算法相比，监督学习算法更常用。例如，我们训练一个机器学习模型来认知苹果的颜色和形状；下次遇到像这样的真实水果时，机器会将其与训练数据进行对比，并且能正确识别苹果。

无监督学习：与监督学习不同的是，无监督学习中不会向模型提供预先的训练。机器需要自己在给定的数据集中找到隐藏的结构。在无监督学习中，不提供分类或标注的数据给算法进行训练。算法必须在没有任何指导的情况下进行学习。最常用的无监督学习类型是聚类分析。

1.9 机器学习——扩展视野

就像其他知识分支一样，对人工智能的前景有更广阔的视角可以帮助你成为一名聪明的技术人员。对于机器学习来说尤其如此，因为它永远具有颠覆性，影响着我们的日常生活，这与其他技术都不同。本节将讨论机器学习(和人工智能)的当前状态和未来路线图，以及机器学习中一些触及日常生活的其他问题。

1.9.1 人工智能发展现状

几年前，一些公司匆忙地进行数据分析——准确地说是大数据分析。现在，为了追赶人工智能和机器学习技术的发展，类似的企业在爆炸式增长。机器学习仍然被认为是一项研发工作。直到近期，仍然不容易探寻到产品开发所需要的基础设施平台。

机器学习现在无疑被认为数字化转型的关键驱动因素。目前廉价的并行计算、大规模数据集和更优算法的广泛普及进一步推动了机器学习的发展。像 Netflix、Airbnb、特斯拉、Uber、阿里巴巴、腾讯、小米及许多其他公司，都已经用内部平台和模型构建了可控的机器学习功能平台。然而，大多数公司并不太容易获得这些内部尖端的机器学习能力。最近，新一代可用的机器学习平台很可能使机器学习更容易被更广泛的企业使用——不管是小型企业还是大型企业。因此，机器学习现在变得更加容易获得。现有可用的开源软件允许开发先进的自学系统。TensorFlow 就是谷歌开发的一个非常受欢迎的机器学习工具。我们将在本书的后半部分广泛使用它处理深度学习问题。

深度学习尤其受到大家的关注。谷歌、脸书、亚马逊、百度等科技巨头以及全球其他许多有前途的初创企业，通过在视觉领域和基于语言的事务中使用基于深度学习的技术取得了一些重要的成果。在我们看来，在短期内直到未来三到五年，深度学习的研究将会持续热下去。有趣的是，在另一个框架中，强化学习(RL)的发展迅猛。在过去的十年中，强化学习受到了研究者的广泛关注。谷歌和脸书等公司已经发布了多种强化学习框架。强化学习框架在训练人工智能模型模拟人类参与游戏方面取得了重大进展，这些游戏的表现与人类相当甚至优于人类。虚拟代理、NLP、图像识别和语音识别是基于人工智能的数据驱动型公司聚焦的重点领域。

kdnuggets 网站上《未来的机器学习趋势》一文中讲述了最近的研究热点。它发现了当今企业中有 3 种机器学习倾向占主导地位：基于云端的分析、算法经济和数据飞轮。数据飞轮涉及大量数据，其构建更好的基于算法的业务模型，且达到更好的业务效率和用户体验，最终获取更多的数据。这是一个数据-算法-数据的循环。许多专家断言，这些重要的发展趋势助力许多公司转变为今天所看到的数据公司。随着算法经济逐渐取代传统的业务流程管理模式，每个企业都将被迫转变为基于数据的企业。在这些所谓的数据公司中，像一些领先的大学一样，有一个可以进行内部和外部直接交流的研究社区，在这社区中可以与商业领导者共同寻找商业解决方案。

根据 McAfee Labs 的 2018 年危机报告，未来机器学习将用于检测网络入侵、在线欺诈和垃圾邮件。对于无服务器环境中的网络安全专家来说，它在检测恶意软件方面非常有用。随着网络攻击数量的不断增长，机器学习已经在协助企业改进安全方法。美国一家领先的办公打印机制造商已经在其企业级打印机和多功能设备上利用海量数据结合机器学习技术进行在线入侵(恶意软件)防御。在可预见的未来，开发人员或许可以开发基于区块链技术的工具，作为对抗网络入侵和确保数据安全

全的可行方式。

1.9.2 人工智能未来的影响力

企业数字化转型是当今世界的一大流行语。许多商业哲学和技术驱动着数字化转型。在所有这些推动变革的技术中，人工智能或许被认为是最具颠覆性的。说到人工智能对企业造成的影响，有必要指出的是，只有约 5% 的职业可以通过现有可行的基于人工智能或机器学习的技术实现完全自动化。

许多组织的 IT 运维已经在很大程度上实现了自动化。从 CEO 的角度来看，IT 运维只是一个成本中心，如果人工智能技术能够高效且经济地使其自动化，那将会很受欢迎。从最高管理层的角度来看，财务、市场营销或任何其他对组织不具有战略影响的职能也是如此。在这种情况下，人仍然会发挥重要作用，但事情永远不会像过去一样了。在快速变化的环境中，专业人员必须在人工智能(包括物理机器人)的协助下工作。自动驾驶汽车和卡车已经触手可及。我们可能很快就能实现卡车运输公司或基于应用程序的出租车公司(如 Uber)的所有商业功能完全自动化。到目前为止，这些功能的实现还是一个悬而未决的问题。

Mike Thomas 的文章“人工智能的未来”很好地描述了人工智能如何影响各行各业。

交通：完善自动驾驶汽车可能还需要十年左右的时间。到那个时候，自动驾驶技术很可能会完全取代人类司机。研究人员和商业机构会有很大的动力在涉及自动驾驶出租车和运输卡车等车辆的公司中实现 100% 的商业功能自动化。

制造业：人类工人与人工智能驱动的机器人一起执行一定范围的重复性任务，如组装和堆叠工作。机器学习和数据预测辅助的系统维护可以使设备保持平稳运行。为了降低错误率，提高可预测性，并在人类辅助下提高整个生产流程的效率，人工智能甚至在供应链中承担了许多重复性和劳动力集中的工作。

医疗：人工智能技术在医疗领域的应用相对来说处于发展初期。人工智能技术有助于实现快速精准的疾病诊断。制药公司正在加快和简化原本需要时间和资源密集型的药物研发过程。机器外科医生可以采用更好、更有效的方式协助医生进行复杂的手术。虚拟护理助手可以实时监测患者状态。借助大数据分析，会给患者带来更多个性化的治疗体验。

同样，所有其他业务，如教育、媒体、客户服务等，都受到了人工智能技术的影响。而且，我们可以确信地说，所有这些都只是冰山一角。人工智能的实际影响比人们想象的要广泛深远。

2018 年 8 月举办的三星 AI 论坛，展示了一些关于人工智能未来路线图的有趣理论。LeCun(人工智能科学家)认为，无监督学习(或自监督学习)掌握着人工智能的未来。LeCun 进一步解释了这是如何做到的。与主要依赖反复试验的 RL 模型相比，无监督学习模型可以熟练地表示心智能力，如我们所说的常识。Breazeal 的演讲聚焦于社交机器人。他解释说，在未来 10 到 20 年，我们可能会拥有具有社交能力和情感能力的智能机器人(所谓的关系型人工智能技术)，它们呈现出一系列令人兴奋的优势。这些机器人将成为我们日常生活中不可或缺的一部分，从而显著提高我们的生活质量。

关于人工智能带来的突破及其未来路线图是一个激动人心和无休止的讨论，大量文献从各个角度详细地讨论了人工智能技术。

1.9.3 与人工智能相关的伦理、社会和法律问题

在前面，我们已经看到了一些人工智能的颠覆性，它同时深刻地影响着个人和整个社会。任何不考虑由此产生的伦理和社会问题的人工智能和机器学习的讨论都是不全面的。

在一篇题为“人工智能和机器人的 10 个伦理问题”的文章中，Miguel González-Fierro 讨论了与人工智能和机器人相关的 10 个重要伦理问题：错误信息和假新闻，工作岗位转移，军用机器人，算法偏见，监管，超级智能，隐私，网络安全，人工智能的错误，机器人权利。其中大部分是不言自明的。为了有一个初步的了解，我们仍将进行简短的讨论。

人们对于现代社交媒体如何被严重地滥用来传播歪曲和虚假新闻进行了全方位的探讨。计算机视觉领域的最新技术可以完全伪造视频。事实上，你几乎可以人为地创建任何视频，证明这些视频的真实性对普通人来说非常困难。而且，普通公众几乎没有任何办法查明真相。对许多年轻人来说，新闻、事实和信息的唯一来源是社交媒体。前面已经讨论了使用人工智能替代人类工作的问题。在一封公开信中，25 000 名科学家和人工智能专业人员呼吁禁止不需要人类监督的自治武器(包括军用机器人)，以避免基于人工智能的武器的全球军备竞赛。在人工智能算法的开发过程中，坚持必要性、避免偏见和歧视是一件显而易见的事情。曾报道过这样一个案例：在算法有倾向性的情况下，特定肤色的人与其他人相比，检出率更低。不过，这里我们不讨论该算法的细节。

因为立法者在制定法律时从未考虑过这种人工智能爆炸，目前基于人工智能的产品和服务大多不受监管。需要制定新的法规，使人工智能系统可以对非理性做法造成的伤害承担更多责任。许多互联网公司正在收集比它们业务所需要的更多的用户信息。公司允许收集多少关于其用户的信息？又有多少是道德的？我们认为社会及立法者还没有给予人工智能系统足够的关注。另一个重要问题是隐私，剑桥分析丑闻的广泛传播，8700 万脸书的个人资料被泄露，并用于影响美国总统选举和英国退出欧盟竞选。在这个互联网民主政体下，对隐私问题的法规是否足够明确？

虽然人工智能可以在很大程度上帮助我们防范网络安全攻击，但与此同时，黑客也可以利用人工智能找到新的复杂的攻击方法。人工智能的另一个伦理问题是，机器人会犯错，但很难确定其责任。这引发了更多相关的法律问题。当机器人在情感上智能化时，它们应该有权利吗？它们将如何与人类联系起来？谁来定义机器人可以具备智能水平？我们需要对人造机器进行这样的限制吗？这些问题包括更多的公开问题，需要进行公开辩论，从而形成全球统一明确的政策指导方针。所有这些与人工智能相关的伦理、社会和法律问题并不是针对任何一个国家的，对整个地球来说都是典型的问题。一个国家开发有害的人工智能系统将影响所有人。这是一个严重的全球关切的问题，类似于今天核武器的迅速扩散，温室气体和污染造成不可逆转的环境变化。

目前，许多国家正在积极地为人工智能的技术发展建立法律环境，例如韩国从 2008 年开始实施《智能机器人开发和传播促进法》。这项法律旨在提高生活质量，促进国民经济的进步。它起草了必要的政策，以创造和促进智能机器人行业的可持续发展。另一个例子是，欧盟通过了一项关于机器人法规的里程碑式决议，这被广泛认为是监管人工智能技术的第一步。2015 年，欧盟还成立了与机器人技术和人工智能增长相关的法律咨询工作组。

1.10 Python 及其作为机器学习语言的潜力

Guido van Rossum 在 20 世纪 90 年代开发了 Python 编程语言。今天, Python 是最广泛应用的通用编程语言之一。Python 是一种现代脚本语言。用 Python 编写的程序直接输入解释器, Python 解释器不需要任何编译直接运行程序。Python 代码可以简单快速地获取反馈(如查找错误)。

Python 是一种跨平台编程语言。可以在 Mac、Windows、Linux 系统上运行 Python 程序。无论是在个人计算机上还是在大型服务器上, Python 程序均可以运行。即使在 iOS 和 Android 系统的平板电脑上也可以使用 Python。Python 平台上有很多第三方库, 它们进一步增加了该语言的通用性和实用性, 因为用于各种编程任务的大量代码已经是库形式编写的现成代码。

值得注意的是, Python 是免费软件。下载和使用 Python 不需要支付任何费用。使用 Python 编写的任何源代码都是你自己的, 你可以随意共享。Python 是一种非常通用的语言。使用 Python 可以加快软件开发速度。你可以在任何感兴趣的地方使用 Python。Python 使编程变得非常有趣。

甚至国际空间站的 Robonaut 2 机器人也使用 Python 开发其中央指挥系统。2020 年欧洲火星任务还计划使用与收集土壤样本任务相关的 Python 程序。数据科学家喜欢 Python 的内置库, 这些库使模型构建项目变得更加容易。Python 的易用性使数据科学软件项目的开发速度非常快。

1.11 关于 TensorFlow

TensorFlow 库是由 Google Brain Team 开发的。它是 Google Brain 的第二代系统。TensorFlow 的构建规模可以达到任何程度, 并且可以在多个 CPU 或 GPU 上运行。TensorFlow 甚至可以在移动操作系统上运行。TensorFlow 支持多种语言包, 如 Python、C++ 或 Java。TensorFlow 库集成了不同的应用程序编程接口(API), 支持像 CNN 或 RNN 这样的深度学习架构。

TensorFlow 是一个主要用于机器学习的开源平台。它是由工具、库和社区资源组成的灵活生态系统。它为数据科学家使用机器学习的最新技术提供支持。开发人员利用 TensorFlow 可以快速构建和部署机器学习的系统。

TensorFlow 是近年来盛行的深度学习软件包。它可以用来构建任何深度学习架构, 从简单的人工神经网络到 CNN 或 RNN。TensorFlow 被谷歌用于其旗下的几乎所有产品, 包括谷歌邮件、照片及谷歌搜索引擎。

1.12 本章小结

近年来, 搜索最多的岗位前三名分别是机器学习工程师、深度学习工程师、高级数据科学家。几年前, 数据科学家是最受欢迎的。据我们的观察, 2010 年前后, 基本分析开始向机器学习和深度学习发生明显转变。技术领域和职称每年都在发生更大的变化。然而, 现在与大约 10 年前的情况显然有所不同。

14 机器学习和深度学习：原理、算法、实战(使用 Python 和 TensorFlow)

值得注意的是，在过去的几年里，一种技术并没有被另一种技术所取代，但同一个领域正在不断发展。2015 年，我们为初学者撰写了一本分析书籍；2020 年，作为作者，随着技术的发展，我们用这本关于机器学习的进阶书籍来回馈读者。我们可以把 2010 年到 2019 年称为人工智能的 10 年，这项技术将在未来几十年持续存在，会随着每一天、每一周和每一年的推移而发展。作为机器学习专家，将永远面临与它并驾齐驱的挑战。

引用谷歌首席执行官 Sundar Pichai 的话，“人工智能是人类正在努力的最重要的事情之一。它比我不知道电或火更深刻。”作为作者，我们没有更好的词汇来强调人工智能技术对人类的重要性，再次重申，它将永远存在，不断发展。