Broadview®
www.broadview.com.cn

机器学习
从入门到入职

用Sklearn与Keras搭建人工智能模型

张威 著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

机器学习
从入门到入职

用Sklearn与Keras搭建人工智能模型

张威 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

近年来机器学习是一个热门的技术方向，但机器学习本身并不是一门新兴学科，而是多门成熟学科（微积分、统计学与概率论、线性代数等）的集合。其知识体系结构庞大而复杂，为了使读者朋友能够把握机器学习的清晰的脉络，本书尽可能从整体上对机器学习的知识架构进行整理，并以 Sklearn 和 Keras 等机器学习框架对涉及的相关理论概念进行代码实现，使理论与实践相结合。

本书分为 4 个部分：第 1 章至第 3 章主要介绍机器学习的概念、开发环境的搭建及模型开发的基本流程等；第 4 章至第 7 章涵盖回归、分类、聚类、降维的实现原理，以及机器学习框架 Sklearn 的具体实现与应用；第 8 章至第 12 章主要阐述深度学习，如卷积神经网络、生成性对抗网络、循环神经网络的实现原理，以及深度学习框架 Keras 的具体实现与应用；第 13 章简单介绍机器学习岗位的入职技巧。

本书可作为机器学习入门者、对机器学习感兴趣的群体和相关岗位求职者的参考用书。

图书在版编目（CIP）数据

机器学习从入门到入职：用 sklearn 与 keras 搭建人工智能模型 / 张威著. —北京：电子工业出版社，2020.1

ISBN 978-7-121-38199-7

Ⅰ. ①机… Ⅱ. ①张… Ⅲ. ①机器学习②软件工具－程序设计 Ⅳ. ①TP181②TP311.561

中国版本图书馆 CIP 数据核字(2019)第 298114 号

责任编辑：张春雨
特约编辑：田学清
印　　刷：三河市良远印务有限公司
装　　订：三河市良远印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　　印张：29.5　　字数：714 千字
版　　次：2020 年 1 月第 1 版
印　　次：2020 年 1 月第 1 次印刷
印　　数：3000 册　　定价：99.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件到 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

前　言

为什么要写本书

人工智能是目前比较火热的概念，存在诸多观点。例如，有人认为人工智能系统的智能水平即将全面超越人类的水平，甚至在不久的将来，超级人工智能（“天网”“终结者”）将会出现，人类将会被消灭或者奴役；也有人认为人工智能只是一个炒作概念，因为人工智能的相关理论在20世纪六七十年代就引发过一波热潮，但是在之后的若干年中，由于互联网革命，人工智能的热度逐渐消退，而如今热潮再起，只不过是老调重弹。

上述观点都有失偏颇，人工智能在某些特定领域具有良好的表现：AlphaGo 在对阵柯洁时以巨大优势获胜；在特定医疗影像处理上，无论处理速度还是诊断的准确度，人工智能都完胜经验丰富的医师；人脸识别现在不仅可以对近期人脸进行判断，还可以根据以往面貌特征推断将来的面貌，并且在失踪人口的搜索上已经出现了成功案例。从上述例子来看，人工智能似乎在不久的将来会统治人类，其实不然，针对上述应用场景的人工智能通常被称为专用人工智能，是局限于某个特定任务（任务单一、需求明确、应用边界清晰、领域知识丰富、建模相对简单）进行的专门应用。总的来说，人工智能的应用现如今局限于专用性，远远没有达到诸如“终结者”这种具有通用性任务能力的程度。

人工智能总体发展水平仍处于起步阶段——通用人工智能领域的研究与应用仍然任重道远，对超级人工智能（“天网”“终结者”）的担忧大可不必，因为超级人工智能本质上是通用人工智能的高级形式，而通用人工智能尚处于起步阶段。例如，人脑是一个通用的智能系统，能够归纳与推理，具有诸多感官信息，可以处理视觉、听觉、判断、推理、学习、思考、规划、设计等各类问题。当前的人工智能系统在信息感知、机器学习等“浅层智能”方面进步显著，但是在概念抽象和推理决策等“深层智能”方面的能力还很薄弱。

人工智能的热潮是否是另外一个炒作概念？为什么 20 世纪六七十年代出现的学科到现在又被炒作了一番？其实任何一项技术都要经历以下几个阶段：一是理论的提出，理论的产生奠定整个学科的框架，由此衍生出 3 个学派，即符号主义、连接主义及行为主义；二是理论的实验室应用或者小规模应用，主要解决理论落地的问题，这些应用通常与大规模商业应用有一步之遥，之所以没有投入大规模商业应用主要是由于生态的限制；三是大规模的商业应用，该阶段已经具备一定的技术及商业生态，如互联网技术的诞生比计算机技术晚，但互联网革命爆发于 20 世纪 90

年代中后期，这是因为最初计算机占地面积较大，且算力较低，而价格又使大部分家庭难以负担，随着计算机的普及，以及计算机系统使用的简易化等诸多合力，互联网革命开始兴起。同样的发展逻辑也出现在移动互联网上，并且人工智能的发展也适用于该逻辑。

人工智能在 20 世纪的发展比较缓慢主要是因为相关的生态并没有很完善地建立起来，而现在已经截然不同。

成为人工智能的参与者是一个不错的想法。对大多数人来说，人工智能的印象源自科幻电影。像先知或者智囊团一样的计算机，或者像忍者一样灵活运动的机器人，充满了神秘；同样，掌握人工智能也并不是一件容易的事情，首先它是多门学科的综合，既各有体系，又相互交融，而相关的理论也日新月异，初学者在这多如牛毛的资料中容易迷失，最终只能“从入门到放弃”。

本书主要从以下几个方面使读者向“从入门到入职”的目标逐步靠近。

- 从知识体系及算法原理上尽可能全面而翔实地进行介绍，使读者能够形成完整的知识架构，同时对相关算法的底层逻辑有清晰的了解。
- 对基于成熟的机器学习框架（Sklearn、Keras）的使用方法进行讲解，避免读者在学习过程中重复“造轮子”，使其在阅读完本书之后可以拥有基本的模型开发能力。
- 从职业的角度，首先对工作中实际模型开发流程的 4 个步骤进行详细介绍，并搭配相应的代码讲解，而不是局限于模型的实现；然后对目前人工智能岗位做出了详细的分析，使读者在求职过程中能够精确地匹配自己的能力和有意愿担任的职位，并且给出了具有参考意义的进阶之路。

如何阅读本书

本书包括 13 章，可分为 4 个部分：第 1 章至第 3 章主要介绍机器学习的概念、开发环境的搭建及模型开发的基本流程等；第 4 章至第 7 章涵盖回归、分类、聚类、降维的实现原理，以及机器学习框架 Sklearn 的具体实现与应用；第 8 章至第 12 章主要阐述深度学习，如卷积神经网络、生成性对抗网络、循环神经网络的实现原理，以及深度学习框架 Keras 的具体实现与应用；第 13 章简单介绍机器学习岗位的入职技巧。

第一部分

第 1 章：机器学习概述，主要介绍机器学习的概念、发展历程及趋势。

第 2 章：机器学习的准备工作，主要介绍机器学习的知识准备和环境准备步骤，以及机器学习模型开发的工作流程。

第 3 章：Sklearn 概述，主要介绍 Sklearn 的环境搭建与安装，以及 Sklearn 常用类及其结构。

第二部分

第 4 章：Sklearn 之数据预处理，主要介绍常用的数据预处理方法，如缺失值处理、数据的规

范化、非线性变换、自定义预处理及非结构性数据预处理等。

第 5 章：Sklearn 之建立模型（上），主要介绍监督学习的概念及相关的监督学习模型，如线性回归、广义线性模型、稳健回归、支持向量机、高斯过程、梯度下降、决策树及分类等算法原理。

第 6 章：Sklearn 之建立模型（下），主要介绍无监督学习的概念及相关的无监督学习模型，主要包括聚类和降维。聚类方法包括 K-mean 聚类、均值偏移聚类、DBSCAN 聚类等；降维方法包括主成因分析、隐含狄利克雷分布、流形学习等。

第 7 章：Sklearn 之模型优化，模型优化有很多种方法，有针对数据本身的优化，如一些采样方法；有针对模型本身的优化，如超参数调整；还有一些集成学习方法，用于提高模型的健壮性及表现能力。

第三部分

第 8 章：Keras 主要 API 及架构介绍，主要阐述 Keras 的环境搭建与安装，以及 Keras 的架构、API 与模型形式。

第 9 章：一个神经网络的迭代优化，主要介绍神经网络的组成、结构、学习机制，以及通用的调优方法。

第 10 章：卷积神经网络，主要介绍卷积神经网络的重要组件及其实现原理，并列举常用的卷积神经网络模型，如 LeNet、AlexNet、VGGNet、残差网络等。

第 11 章：生成性对抗网络，主要介绍生成性对抗网络的原理、常见的生成性对抗网络、自动编码器模型及代码实现。

第 12 章：循环神经网络，主要介绍循环神经网络的常见模型层及代码实现。

第四部分

第 13 章：机器学习的入职准备，主要介绍机器学习岗位及求职者的分布、机器学习岗位的面试技巧，以及机器学习相关岗位的技能侧重点。

勘误和支持

鉴于笔者对机器学习部分理论的理解与认知存在局限性，并且由于机器学习的理论体系发展迅速，故无法做到完备且详尽地将相关理论收录到本书中。针对特定方向（如人工视觉、推荐系统、语义识别等）的相关知识将在后续系列丛书中详细阐述。

对人工智能感兴趣的读者朋友，欢迎以电子邮件（bllizard@yeah.net）的方式与笔者取得联系，期待能够得到广大读者真挚的反馈，在技术的道路上互勉共进。

特别致谢

笔者花费大量时间总结了机器学习的相关理论框架及代码实现，在本书的编写过程中，首先

要感谢我的父母，感谢你们在我写书的过程中的支持和照顾。此外，还要感谢曾经的同事李毅、王志远、谢瑞杰、朱琦焱、李云海、温海、丁锦城、杨永邦、张择仪等人的支持和理解。在此一起表示衷心的感谢。

【读者服务】

- 获取博文视点学院 20 元付费内容抵扣券
- 获取免费增值资源
- 加入读者交流群，与更多读者互动
- 获取精选书单推荐

微信扫码回复：（**38199**）

目 录

第1章　机器学习概述

本章主要对机器学习的基本内容进行简单介绍。

什么是机器学习（Machine Learning，ML）？当大家学习人工智能（Artificial Intelligence，AI）技术时，会发现存在各种各样的“学习”，如深度学习、强化学习、机器学习、监督学习、集成学习等。这些“学习”摆在人们面前的时候，不由得让人产生一种迷茫感，这些学习有一些相互重合的地方，有一些又是单独的学科。有志从事人工智能行业或者对人工智能感兴趣的朋友，可以从厘清这些“学习”的概念开始。

在了解机器学习的定义之后，可以进一步了解机器学习的发展历程，从而对这门学科的发展演进脉络有一个大致的了解，这对把握人工智能的发展趋势具有一定的指导作用。

最后，读者还需要了解机器学习存在哪些商业应用，以及机器学习可能的发展趋势。由此，读者可以自行判断诸如以下问题。

- 人工智能是否仅仅是一个风口？
- 还有哪些应用场景没有被挖掘出来？
- 什么样的工作将来会被人工智能取代？

1.1 什么是机器学习

首先需要了解什么是人工智能。人工智能是计算机科学的一个分支，它企图了解智能的实质，并生产一种新的能以人类智能相似的方式做出反应的智能机器，该领域的研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等。

人工智能所包含的类目很多，其中包括但不限于诸如计划调度、专家系统、推荐系统等学科，其中机器学习就是人工智能中比较重要的学科之一（见图 1-1）。

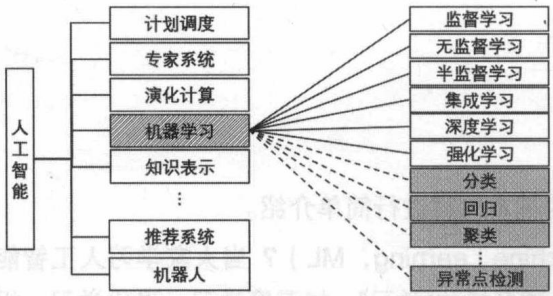

图 1-1　人工智能与机器学习的关系

机器学习，使用算法解析数据并从中学习，然后对真实世界中的事件做出决策和预测。与传统的用于解决特定任务、硬编码的软件程序不同，机器学习用大量的数据“训练”，并通过各种算法从数据中学习如何完成任务。如图 1-1 所示，之前所提到的监督学习、深度学习、强化学习等内容，实际上都属于机器学习的一个子类目，它们之间的算法并不是相互独立、相互排斥的，而是多门学科的交汇，都是基于特定的场景、任务及实现途径划分的。

机器学习来源于早期的人工智能领域，传统的算法包括决策树（Decision Tree）、聚类、分类、支持向量机（Support Vector Machine，SVM）、集成学习（Ensemble Learning）等。关于这些算法的详细内容会在第 5 章至第 7 章详细介绍。

根据学习数据源性质，机器学习算法可以分为监督学习（如分类问题）、无监督学习（如聚类问题）、半监督学习，它们之间的区别在于样本是否被标记（存在标签）。如果学习样本存在明确的类目或者对应的数值，则该学习为监督学习；如果样本不存在标签，则为无监督学习；半监督学习则介于两者之间，学习样本的标签不完整。

深度学习（Deep Learning）专门用于指定深度神经网络，是利用深度神经网络解决特征表达的一种学习过程。深度神经网络本身并不是一个全新的概念，可大致理解为包含多个隐含层的神