大数据/人工智能系列规划教材



| 轻松入门 | 专业实用 |
| 图文并茂 | 角度新颖 | 实战导向 |

图说图解机器学习

耿煜 李钦 杨耿 邱婉 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

需要全本请访问 www.ertongbook.com

大数据/人工智能系列规划教材

图说图解机器学习

耿煜　李钦　杨耿　邱婉　著

Publishing House of Electronics Industry
北京 · BEIJING

内 容 简 介

本书采用图形化的方法讲解人工智能和机器学习的知识与技术，并且借用图形化软件KNIME采用拖、拉、拽等“傻瓜式”的操作完成从简单到复杂的机器学习项目。全书分为三个部分，分别是人工智能技术入门、传统机器学习和深度学习。本书具有图说图解、自上而下、够用即止、实战掌握的特点，适合于作为人工智能入门者、人工智能技术应用者及高职高专院校理工科、本科院校非理工科专业学生的教材。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

图说图解机器学习 / 耿煜等著. — 北京 : 电子工业出版社, 2019.7
ISBN 978-7-121-36826-4
Ⅰ. ①图… Ⅱ. ①耿… Ⅲ. ①机器学习—图解 Ⅳ. ①TP181-64
中国版本图书馆CIP数据核字（2019）第112644号

责任编辑：贺志洪（hzh@phei.com.cn）
印　　刷：北京富诚彩色印刷有限公司
装　　订：北京富诚彩色印刷有限公司
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 × 1092　1/16　　印张：13.75　　字数：352 千字
版　　次：2019年7月第1版
印　　次：2019年7月第1次印刷
定　　价：56.00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至 dbqq@phei.com.cn。

本书咨询联系方式：（010）88254609 或 hzh@phei.com.cn 或 QQ6291419。

个人简介



耿煜

深圳信息职业技术学院软件学院移动互联应用技术专业教师，深圳市科创委入库专家，香港科技大学博士，从事半导体激光数据分析，大数据、人工智能与手机 App 应用开发。发表论文 20 余篇，其中以第一作者发表数篇 SCI 论文，发明专利 4 项，承担纵横向课题多项。



李钦

深圳信息职业技术学院专业教师，高级工程师。在图像处理与模式识别领域有十余年研究经验，主要研究方向为生物特征识别（包括掌纹识别、人脸识别、视网膜病变自动诊断等）。发表 SCI/EI 论文 30 余篇。撰写生物特征识别方向科研专著一部。主持广东省自然科学基金一项。获授权国家发明专利多项、美国发明专利一项。于 2009 年在美国举办的视网膜病变在线识别挑战赛上获得第二名，于 2011 年在日内瓦举办的日内瓦国际发明展上获得金奖，于 2016 年在日内瓦举办的日内瓦国际发明展上获得银奖。



杨耿

深圳信息职业技术学院专业教师，高级工程师，香港理工大学博士（人工智能、图像识别方向），2016 年、2017 年 ISO 国际标准智能交通分会第五工作组中国代表团专家，深圳市科技专家委员会科技项目评审专家，《交通运输研究》杂志审稿专家。主要研究方向为深度

学习、图像识别，发表论文10篇，获授权国家发明专利20项，主持多项横向课题。



邱婉

深圳市兆阳信息技术研究院理事长、董事，清华大学EMBA，浙江大学电子信息工程学士。曾任深圳市高新协会培训部部长（主要负责人工智能在智能制造、智能管理的应用以及数据挖掘），参与深圳市科创委国高评审相关工作。

需要全本请访问 www.ertongbook.com

前言

PREFACE

1. 创作经历

本书作者团队大部分成员就职于深圳信息职业技术学院（以下简称信息学院）。作者们初入职信息学院时，感觉凭借自己高学历教专科生简直是杀鸡用牛刀，但是往往几节课后就深感到拳头打在棉花上，甚至是如履薄冰。几乎任何一个公式都能让学生们面面相觑，几乎任何一个算法都能让学生们变成大眼瞪小眼。学生经常单刀直入地提问题，简单粗暴："老师，我学了这个能干什么"，而鲜有学生会问："这个问题怎么做""这个问题哪里出错了""我这样做行不行"。但是一旦"能干什么"的问题明确了，随之而来会源源不断地问"怎样做"。各种"惨痛"的经历告诉我们，不解决目标问题，大多数学生没有兴趣；不解决复杂度问题，大多数学生无法掌握。

本书大多数作者都有较高的学历和较多的研究经历，深知学习的艰难和痛苦，更对"知识就是力量"有深切的体会。在作者们的学习阶段，往往发现找到一本合适的入门书是那么得难，开始就啃业界大牛的著作往往是一个从入门到放弃的过程，或者因为很多书公式过多而影响了核心理念的掌握，导致入门如登天。结合在信息学院的授课经验，我们总结出"图说图解、自上而下、够用即止、实战掌握"的教学方法，希望带给大家一本不一样的入门书。

借用凯撒的一句名言，希望大家"我来，我见，我征服"。

2. 创作背景

我们如何才能迈向新时代呢？答案就是拥抱新动能，而人工智能是新动能中最有代表性的一个。虽然人工智能和机器学习作为热门词汇早已进入了人们的视野，但是它们究竟是什么，可能社会上绝大多数人还是不知道，更别说知道它们能干什么了。但是在国家顶层，早

就预见到了其力量，“十九大”报告指出：“加快发展先进制造业，推动互联网、大数据、人工智能和实体经济深度融合”。在这个大时代，为了让更多的人搭上新时代的电梯，我们精心设计所有内容，确保能学、会用、可进阶。

3. 知识体系

全书分为三个部分，分别是人工智能技术入门、传统机器学习和深度学习。

人工智能技术入门：这部分包括第 1 章及第 2 章，第 1 章主要从历史发展角度讲述人工智能，第 2 章在技术层面上为今后的学习打下“够用”的数学基础和 KNIME 操作基础。

传统机器学习：这部分是本书的主要部分，包括第 3 章到第 9 章，分别介绍线性回归、模型优化、逻辑回归、支持向量机、决策树、深入理解决策树和贝叶斯模型。在这部分中，我们会逐步掌握 KNIME 的使用，更重要的是逐步掌握机器学习的流程、数据的处理、模型的使用等技术。其中第 5 章模型优化和第 8 章深入理解决策树不是入门必备知识，可以作为选修内容。

深度学习：这部分内容只包括第 10 章，仅仅简单介绍深度学习的入门知识和应用技术。

4. 特色

本书的特色可以概括为“图说图解，自上而下，够用即止，实战掌握”。

“图说图解”是本书最显而易见的特色。本书将所有深奥难懂的机器学习原理图形化地展现及讲解，让读者能够从直觉上理解而不是从概念或者公式上理解。配合图形化的机器学习工具 KNIME，使读者能够更方便快速地入门机器学习，免去了编程这个令初学者望而却步的前提条件。为了保证这一点，我们制作了大量的原创图片用于讲解，精挑细选了开源、免费、影响力大并且功能无限制的 KNIME 作为工具。

“自上而下”是本书的核心特色。本书从内容上以机器学习模型为明线，在这条明线外还有两条按照“自上而下”的教学理念设计的暗线。第一条暗线就是使用 KNIME 工具，从大致流程的掌握到细节的掌握，保证读者能够从大局上知道自己在干什么，进而知道细节上应该怎样设置。第二条暗线就是机器学习知识和技术由宏观理解到具体技能掌握，保证读者能够从宏观上理解一个机器学习项目的流程，接着再去理解具体技术的细节。为保证这一点，本书从简单模型入手逐步深入到复杂模型，从使用“干净”数据逐步扩展到使用“脏”数据。

“够用即止”是本书降低读者入门门槛和学习负担的保障。本书省略了大量的理论推导和公式计算，仅通过图形化方法从概念上让读者理解算法的原理，具体操作时直接使用 KNIME 工具即可完成。而且对于机器学习应用者来说，大多数人也没有必要理解背后的理论及公式，即使之后不用 KNIME 而使用 Python，也是编写几句代码就可以解决的，没有必要去深究理论。为了保证这一点，我们删除了大量的公式，增加了大量的图解。

“实战掌握”是验证本书是否成功的关键。纸上谈兵毕竟不是真本事，本书采用 Kaggle 机器学习平台的真实竞赛为实战项目，从入门项目泰坦尼克号到复杂项目银行客户分类，在实战中使读者逐步掌握机器学习的流程、模型的设置、数据清洗、非平衡数据的处理等问题。

5. 本书是什么

本书是一本机器学习的入门书。

本书是一本关于机器学习应用的书。

本书的目标是使一个高中水平的读者通过本书能够入门机器学习，并掌握足够的进一步提升的能力。

本书也可以看成是机器学习图解的 KNIME 软件教程。

6. 本书不是什么

本书不研究任何机器学习公式、理论。

本书不覆盖任何机器学习模型。

本书字不多。

7. 如何使用本书

对于具有理工科背景的同学来说，建议从头至尾学习每章内容以了解每个模型的原理及其应用。对于非理工科背景或者仅仅关心模型应用的同学，可以直接阅读模型使用部分，而将模型原理部分当作手册参考即可。

每章的最后都有课后练习部分，请大家仔细思考。所有答案、模型源文件和数据都可以扫描下面的二维码索取。



关注公众号查看本书所有答案、模型源文件和数据

更多反馈可以加作者微信进入图说图解机器学习交流群交流。



作者微信

8. 编写分工

耿煜：主笔，主要负责全书的组织设计、案例分析和整体结构。
李钦：案例搜集整理与筛选。
杨耿：深度学习案例与应用。
邱婉：图解设计及绘图。

9. 致谢

感谢深圳信息职业技术学院各位老师和同学的帮助，感谢深圳兆阳信息技术研究院各位工程师的协助，感谢我们的家人、朋友。没有你们的帮助就没有这本书的问世。

目录

CONTENT

▶ 第 1 章　人工智能及机器学习概述 1

1.1　人工智能概述 1
1.1.1　人工智能简史 2
1.1.2　人工智能是什么 4
1.1.3　人工智能的能力 5
1.2　机器学习概述 5
1.2.1　机器学习是什么 5
1.2.2　以监督学习为例 6
1.2.3　学习任务 7
1.2.4　机器学习要解决的基本问题 7
1.2.5　机器学习如何优化模型 7
1.2.6　机器学习工作流程 7
1.2.7　机器学习的各大流派 8
1.2.8　机器学习算法选择 8
1.2.9　需要的知识 9
1.3　深度学习概述 9
1.4　机器学习与统计学 9
1.5　课后练习 10

▶ 第 2 章　机器学习基础知识 11

2.1　数学基础 11
2.1.1　数据的分类 12
2.1.2　基本统计学术语 12
2.1.3　回归 14
2.1.4　最小二乘法 14
2.1.5　判断拟合好坏 15
2.1.6　小结 17

2.2 读图.....17
2.2.1 数值数据的分布.....17
2.2.2 分类数据的分布.....18
2.3 KNIME.....21
2.3.1 KNIME 简介.....21
2.3.2 下载和安装.....21
2.3.3 KNIME 基本使用.....21
2.3.4 小结.....28
2.4 课后练习.....28

▶ 第 3 章 线性回归.....29

3.1 简单线性回归.....30
3.1.1 场景说明.....30
3.1.2 KNIME 建立工作流.....30
3.1.3 数据获取.....30
3.1.4 观察数据.....31
3.1.5 数据划分.....33
3.1.6 模型训练.....34
3.1.7 模型测试.....37
3.1.8 损失函数.....37
3.2 多元线性回归初步.....38
3.2.1 任务及数据说明.....38
3.2.2 建立基本的工作流.....38
3.2.3 读取并观察数据.....39
3.2.4 整合界面.....49
3.3 多元线性回归进阶.....51
3.3.1 优化模型.....51
3.3.2 正向选择节点.....55
3.3.3 反向消除.....58
3.3.4 模型解释.....58
3.3.5 特征归一化.....59
3.3.6 使用 KNIME 具体实现归一化.....59
3.3.7 相关系数.....60
3.4 课后练习.....61

▶ 第 4 章 逻辑回归.....63

4.1 逻辑回归基本概念.....63
4.1.1 分类问题.....63
4.1.2 从线性回归到逻辑回归.....65

需要全本请访问 www.ertongbook.com

4.1.3 判定边界 66
4.1.4 KNIME 工作流 66
4.1.5 读取数据 67
4.1.6 数据处理 67
4.1.7 模型训练及测试 68
4.1.8 模型评价 69
4.2 逻辑回归实战 71
4.2.1 泰坦尼克号生存问题背景介绍 71
4.2.2 读取数据 72
4.2.3 数据处理 73
4.2.4 数据可视化及删除无关列 75
4.2.5 模型训练和测试 82
4.2.6 模型评价 83
4.2.7 提交结果 85
4.2.8 模型解释 89
4.3 课后练习 90

▶ 第 5 章 模型优化 91

5.1 梯度下降 91
5.1.1 损失函数 92
5.1.2 使用 KNIME 优化模型 96
5.2 正则化 98
5.2.1 准确性和健壮性 98
5.2.2 复杂的模型 98
5.2.3 欠拟合和过拟合 98
5.2.4 正则化防止过拟合 100
5.2.5 使用 KNIME 设置正则化 100
5.3 模型评价 101
5.3.1 混淆矩阵 101
5.3.2 F1 103
5.3.3 ROC 曲线和 AUC 104
5.4 课后练习 106

▶ 第 6 章 支持向量机 107

6.1 支持向量机基本概念 107
6.1.1 支持向量机是什么 107
6.1.2 支持向量是什么 108
6.1.3 逻辑回归与支持向量机的比较 108
6.1.4 核 110

6.1.5 线性核模型调参 111
6.1.6 非线性核模型调参 113
6.1.7 C 与 γ 114
6.2 SVM 初战 114
6.2.1 问题说明 114
6.2.2 建立工作流 114
6.2.3 数据观察 115
6.2.4 模型训练与测试 117
6.2.5 观察结果 118
6.3 支持向量机解决泰坦尼克号问题 119
6.3.1 归一化 119
6.3.2 核函数 120
6.3.3 新建工作流 120
6.3.4 C 参数 123
6.4 一个重要的问题 124
6.5 课后练习 124

▶ 第 7 章 决策树 125

7.1 决策树简介 125
7.1.1 决策树的优点 125
7.1.2 决策树的缺点 126
7.1.3 防止过拟合 126
7.1.4 问题解析 126
7.1.5 奥卡姆剃刀 128
7.1.6 提前结束 128
7.1.7 剪枝 130
7.1.8 组合算法 131
7.1.9 Adaboosting 133
7.2 使用决策树解决泰坦尼克号生存问题 135
7.3 决策树高级应用实战——特征工程 137
7.3.1 数据探寻 137
7.3.2 特征工程 143
7.3.3 异常数据处理 146
7.4 决策树高级应用实战——模型建立与比较 149
7.4.1 决策树 149
7.4.2 袋装 153
7.4.3 随机森林 157
7.4.4 提升 159
7.5 课后练习 160

▶ 第 8 章 深入理解决策树 161

8.1 决策树进阶 161
8.1.1 如何构建决策树 161
8.1.2 ID3 算法决定什么是最好的 162
8.1.3 CART 算法决定什么是最好的 164
8.1.4 KNIME 设置 165
8.2 数据不平衡问题优化 165
8.2.1 多数数据降采样 166
8.2.2 少数数据过采样 168
8.2.3 SMOTE 算法 170
8.3 课后练习 172

▶ 第 9 章 贝叶斯分析 173

9.1 贝叶斯定理 173
9.1.1 基本术语 173
9.1.2 条件概率 174
9.1.3 全概率和贝叶斯 176
9.1.4 贝叶斯定理 176
9.1.5 贝叶斯定理在机器学习中的应用 177
9.2 贝叶斯算法解决银行客户分类问题 178
9.2.1 工作流 178
9.2.2 贝叶斯算法的学习器节点 178
9.3 情感分析案例 179
9.3.1 安装插件 179
9.3.2 建立工作流 180
9.4 课后练习 183

▶ 第 10 章 深度学习 185

10.1 深度学习简介 185
10.1.1 深度学习的关键 186
10.1.2 我们的目标 186
10.1.3 深度学习图像识别原理概述 187
10.1.4 图像识别探析 187
10.2 卷积神经网络（CNN） 189
10.2.1 CNN 基本原理 189
10.2.2 常用 CNN 模型 193
10.3 KNIME 实现卷积神经网络 195
10.3.1 环境构建 195
10.3.2 安装所需的工具 195

10.3.3 步骤分析 198
10.4 深度学习开源应用举例 199
10.5 深度学习工商业应用举例 201
10.6 课后练习 205

▶ 参考文献 206

第1章 人工智能及机器学习概述



它（人工智能）将成为终极搜索引擎，可以理解网络上的一切信息。它会准确地理解你想要什么，给你你需要的东西。我们现在还远远没有做到这一点。然而，我们能够逐渐接近，我们目前正在为此努力。

——拉里·佩奇

知识点

- 人工智能的历史
- 人工智能的定义
- 人工智能的能力
- 机器学习的定义
- 机器学习与统计学

在这一章中，我们一起回顾一下人工智能的历史，理解什么是人工智能。然后初步了解机器学习和深度学习，并大致知道机器学习和统计学之间的关系。

1.1 人工智能概述

人工智能是近几年极其热门的一门技术，人们对人工智能的态度也各有不同。我们将会从其历史出发，一步一步地探寻人工智能到底是什么。

1.1.1 人工智能简史

我们先从人工智能的历史谈起，来看看历史上人工智能是如何一步一步地发展起来的，进而从感性上了解什么是人工智能。

1. 初创期：1940's～1950's

最初的人工智能研究是20世纪30年代末到50年代初的一系列科学进展交汇的产物。神经学研究发现大脑是由神经元组成的电子网络，其激励电平只存在“有”和“无”两种状态，不存在中间状态。维纳的控制论描述了电子网络的控制和稳定性。克劳德·香农提出的信息论则描述了数字信号（即高低电平代表的二进制信号）。图灵的计算理论证明数字信号足以描述任何形式的计算。这些密切相关的想法暗示了构建电子大脑的可能性。

这一阶段的工作包括一些机器人的研发，例如威廉·格雷·沃尔特的“乌龟”（Turtles），还有“约翰霍普金斯兽”（Johns Hopkins Beast）。这些机器并未使用计算机、数字电路和符号推理，控制它们的是纯粹的模拟电路。沃尔特·皮茨和沃伦·麦卡洛克分析了理想化的人工神经元网络，并且指出了它们进行简单逻辑运算的机制。他们是最早描述所谓“神经网络”的学者。马文·闵斯基是他们的学生，当时还是一名24岁的研究生。1951年他与其他研究者一道建造了第一台神经网络机，称为SNARC。在接下来的50年中，闵斯基是人工智能（AI）领域最重要的领导者和创新者之一。

1951年，克里斯托弗·斯特雷奇使用曼彻斯特大学的Ferranti Mark 1机器写出了一个西洋跳棋（checkers）程序；迪特里希·普林茨则写出了一个国际象棋程序。在20世纪50年代中期和60年代初亚瑟·塞缪尔开发的西洋棋程序已经可以挑战具有相当水平的业余爱好者。游戏人工智能一直被认为是评价人工智能进展的一种标准。

人工智能诞生的标志是1956年召开的达特茅斯会议，在这次会议上人工智能的名称和任务得以确定。该会议提出的断言之一是“学习或者智能的任何其他特性的每一个方面都应能被精确地加以描述，使得机器可以对其进行模拟。”与会者包括大量在AI领域重要的科学家，他们中的每一位都将在人工智能研究的第一个十年中做出重要贡献。在此次会议上纽厄尔和西蒙讨论了“逻辑理论家”，而麦卡锡则说服与会者接受“人工智能”一词作为本领域的名称。1956年达特茅斯会议上人工智能的名称和任务得以确定，同时出现了最初的成就和最早的一批研究者，因此这一事件被广泛地认为人工智能诞生的标志。

2. 黄金年代：1956～1974年

达特茅斯会议推动了全球第一次人工智能浪潮的出现。当时乐观的气氛弥漫着整个学界，在算法方面出现了很多世界级的发明，其中包括增强学习的雏形——贝尔曼公式，而增强学习是谷歌AlphaGo算法的核心思想内容。现在常听到的深度学习模型，其雏形叫做感知器，也是在那几年间发明的。除了算法和方法论有了新的进展，在第一次浪潮中，科学家们还造出了聪明的机器。其中，有一台叫做STUDENT（1964）的机器能证明应用题，还有一台叫做ELIZA（1966）的机器可以实现简单的人机对话。于是，人工智能界认为按照这样的发展速度，人工智能真的可以代替人类。

这个时候，计算机可以解决代数应用题、证明几何定理、学习和使用英语，研究者们在私下的交流和公开发表的论文中表达出相当乐观的情绪，认为具有完全智能的机器将在20年内出现。

需要全本请访问 www.ertongbook.com