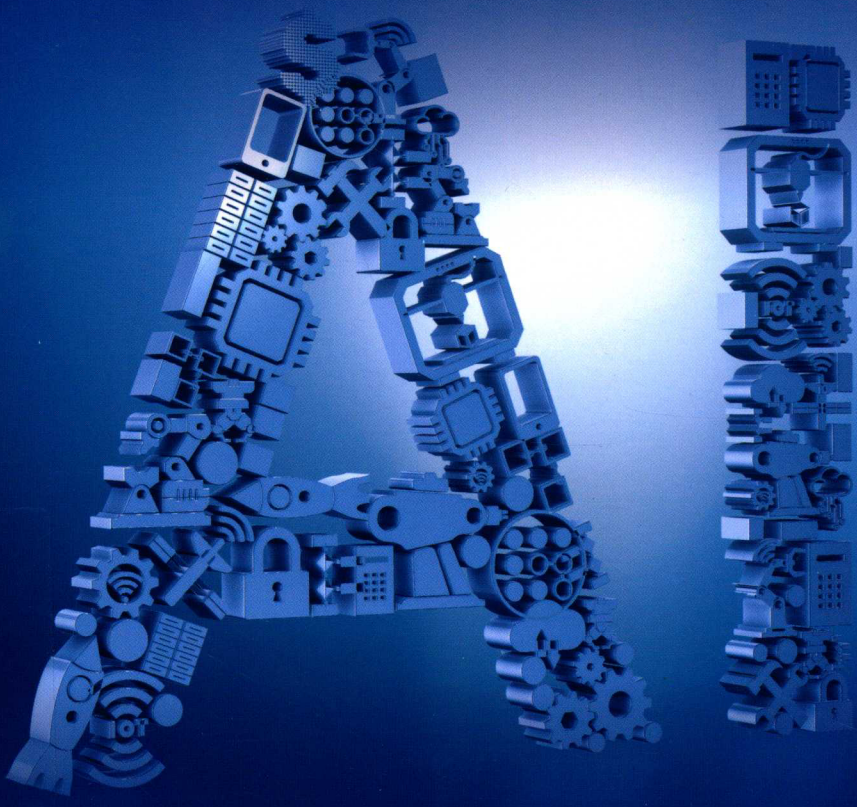
Broadview®
www.broadview.com.cn

有趣的引言故事：**激起兴趣**
清晰的思维导图：**明晰结构**
创意的自画图表：**更易理解**
详细的算法推导：**讲透原理**

快乐机器学习

[新加坡] 王圣元　著



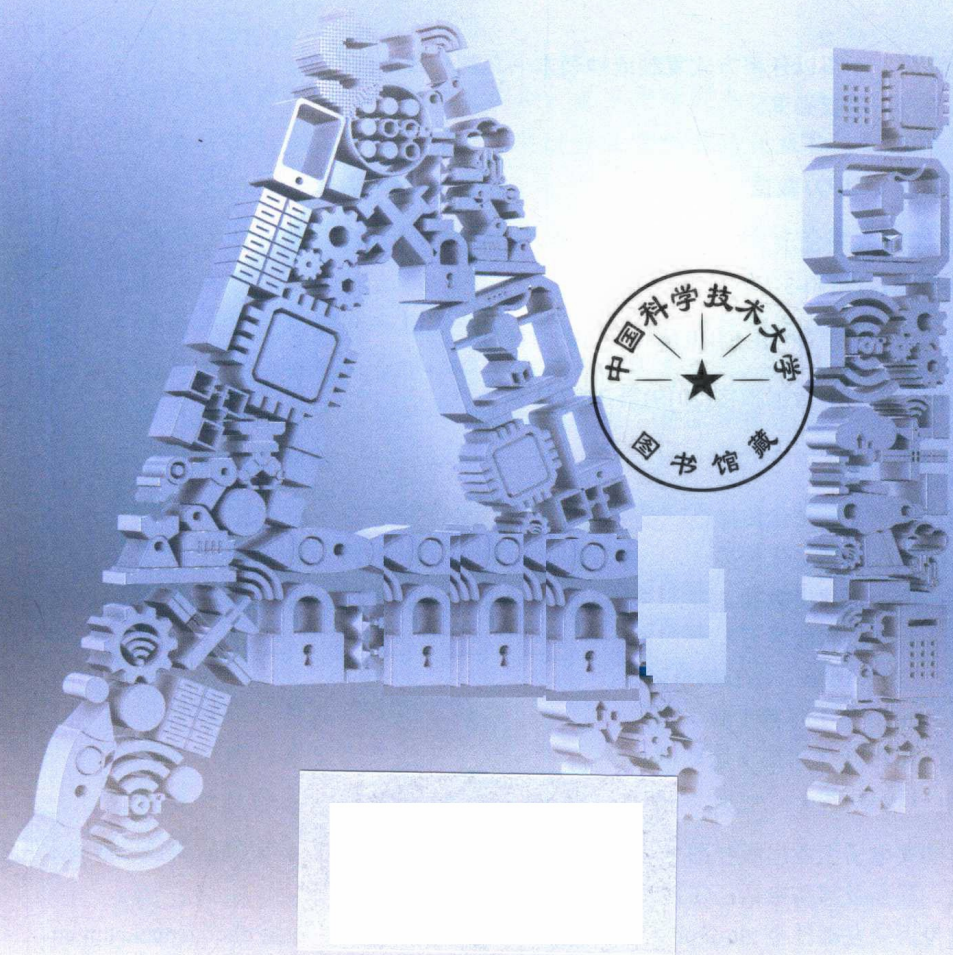
每个知识点都是理论和实践相结合，
既有严谨的算法推导，又有多样的代码展示，图文并茂

全彩印刷

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

快乐机器学习

[新加坡] 王圣元 著



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

内 容 简 介

学习并精通任何一门学科无外乎要经过四个步骤：它是什么？它可行吗？怎么学它？如何学好它？机器学习也不例外，本书就以这四个步骤来介绍机器学习。

本书第 1 章介绍“机器学习**是什么**”，即从定义开始，详细介绍机器学习涉及的知识、数据和性能度量。第 2 章介绍“机器学习**可行吗**”，即介绍机器具备学习样本以外的数据的能力。第 3 章介绍“机器学习**怎么学**”，即介绍机器如何选择出最优模型。作者在这 3 章的写作上花费的时间最多，光这 3 章的内容就绝对会让读者有所收获。第 4~14 章介绍“**如何学好**机器学习”，重点介绍机器学习的各类算法和调参技巧。第 15 章介绍机器学习中的一些非常实用的经验，包括学习策略、目标设定、误差分析和偏差与方差分析。

作者写作本书的目的是深入浅出介绍机器学习，使看似复杂、晦涩的专业知识变得通俗易懂，让那些想入门的读者感觉门槛没有那么高，让有基础的读者感觉内容也很丰富。为了达到这两个目的，本书用有趣的引言故事来激起读者的阅读兴趣，用清晰的思维导图来明晰结构，用自画图表来增强美感，用公式推导来讲透原理，达到**趣、美、准、全**，让每位读者从本书中获益，快乐地学习机器学习。

本书非常适合机器学习初学者、高校相关专业学生及有一定数学和统计学基础的高中生学习。

图书在版编目（CIP）数据

快乐机器学习 /（新加坡）王圣元著. —北京：电子工业出版社，2020.1
ISBN 978-7-121-37590-3

Ⅰ. ①快… Ⅱ. ①王… Ⅲ. ①机器学习 Ⅳ.①TP181

中国版本图书馆 CIP 数据核字(2019)第 219771 号

责任编辑：王　静
印　　刷：天津千鹤文化传播有限公司
装　　订：天津千鹤文化传播有限公司
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：720×1000　1/16　印张：22.25　字数：445 千字
版　　次：2020 年 1 月第 1 版
印　　次：2020 年 1 月第 1 次印刷
定　　价：119.00 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：010-51260888-819，faq@phei.com.cn。

前　言

作者写作本书的目的就是用通俗的文字来讲解机器学习，最好通俗得如作者在女儿生日时给她写的信：

亲爱的欣玥：

从2020年开始，愿你：

- **学习**不要死记硬背，避免过拟合；也不要蜻蜓点水，避免欠拟合。
- **心态**像随机梯度下降一样，不要过分注重眼前的利益和一时的得失，进而看不清大局而被假象蒙骗。
- **抉择**像随机森林一样，各取所长，集思广益，这样你才能做出最正确的决定。
- **操行**像自适应提升一样，知错能改，这样你才能越来越优秀。
- **说话**像奥卡姆剃刀原理一样，牢记“少就是多”，当一个好的聆听者。
- **脾气**不要像梯度爆炸一样越来越大，也不要像梯度消失一样没有，要稳定地敢爱敢恨。
- **容忍**像支持向量机一样，最大化你的容错间隔。有一些错误是在所难免的，要学会将硬间隔变成软间隔。
- **生活**像偏差和方差达到最优点一样，不偏不倚，不骄不躁。

从2020年开始，爸爸会

- 最初辅导你有监督学习。
- 然后锻炼你半监督学习。
- 接着放任你无监督学习。
- 不断评估你要增强学习。

当学习到了某个临界点时，不管外界资源多么丰富，你的表现一定会趋于稳定，这时必须靠深度学习才能最大程度地突破自我，最终获得迁移学习的能力。

学习并精通一门学科无外乎要经过四个步骤：它是什么？它可行吗？怎么学它？如何学好它？学习机器学习也不例外，本书就以这四个步骤来解读机器学习。

- 第 1 章介绍“机器学习**是什么**”，即从定义开始，详细介绍机器学习涉及的知识、数据和性能度量。
- 第 2 章介绍“机器学习**可行吗**”，即机器具备学习样本以外的数据的能力。本章从概率的角度证明样本内误差和样本外误差的关系。
- 第 3 章介绍“机器学习**怎么学**”，即机器如何选出最优模型。本章介绍机器学习版本的样本内误差（训练误差）和样本外误差（测试误差），再通过验证误差来选择模型。

前 3 章属于机器学习的概述：第 1 章介绍机器学习的概念，为了让读者打好基础；第 2 章为证明机器学习是可行的，让读者做到心中有数；第 3 章运用机器学习性能指标而构建框架，看懂它们不需要精通任何机器学习的算法。作者在这 3 章的写作上花费的时间最多，光这 3 章的内容就绝对让读者有所收获。

第 4~14 章介绍“**如何学好**机器学习”，重点介绍机器学习的各类算法和调参技巧。在本书中，机器学习模型分为线性模型、非线性模型和集成模型。

- 第 4~8 章介绍线性模型，包括线性回归模型、对率回归模型、正则化回归模型、支持向量机模型。
- 第 9~11 章介绍非线性模型，包括朴素贝叶斯模型、决策树模型、人工神经网络模型、正向/反向传播模型。
- 第 12~14 章介绍集成模型，包括随机森林模型、提升树模型、极度梯度提升模型。

第 15 章介绍机器学习中一些非常实用的经验，包括学习策略、目标设定、误差分析、偏差和方差分析。

为了帮助读者阅读，下面的流程图展示了整本书的大框架。

本书的每一章都以通俗的引言开始，吸引读者；以精美的思维导图过渡，让讲解思路更清晰；以简要的总结结束，让读者巩固所学的知识。此外，每个知识点都是理论和实践相结合，既有严谨的数学推导，又有多样（Python 和 MATLAB）的代码展示，图文并茂，最好地服务各类读者。

作者非常欣赏谷歌大脑研究员 Chris Olah 的观点“I want to understand things clearly, and explain them well”，即力争把每个知识点弄懂、弄透，然后以通俗易懂的方式让其他人学会、学透。作者愿意做“把困难的东西研究透而简单展示给大众”的人（Research Distiller），因为学术界中的论文虽然“高大上”，但是很多会让读者读完还是一头雾水。用 Chris Olah 的话来讲，这种以不清不楚的方式来解释高难课题的做法欠下了太多研究债务（Research Debt）。

这本书能够完成，受到很多机器学习优质课程的启发，比如斯坦福大学 Andrew Ng 教授的 CS229 课程、加州理工学院 Yaser S. Abu-Mostafa 教授的 Learning from Data 课程、台湾大学林轩田教授的机器学习基石和技法、华盛顿大学 Emily Fox 和 Carlos Guestrin 教授的 Machine Learning Specialization。他们的课程都是理论结合实际，通俗而不失严谨，学习这些课程可以让我解决工作中的很多需求，可见这些课程的含金量之高，在这里我想对他们表达最真挚的感谢（即便他们也不认识我 ☺）!

此外，感谢父母无条件地支持我写书，感谢爷爷、大伯和姐夫经常阅读我的公众号文章，经常鼓励我，感谢夫人在我写书时帮着带娃，感谢女儿给我的无穷动力：想象着以后她拿着我写的书可以自豪地跟别的小朋友说“这是我爸爸写的书”。最后感谢所有“王的机器”公众号的读者，你们的支持和反馈一直激励着我不断进步，这本书是特别为你们而写的。

由于作者水平有限，书中难免会有错漏之处，欢迎诸位专家和广大读者斧正。

作　者

本书阅读说明

书中符号说明

符　　号	表达的意思
$\boldsymbol{x}=(x_1,x_2,\ldots,x_n)$	特征值 (输入,自变量)，通常用 n 维向量表示
y	标签 (输出,变量)，通常用标量表示
$(\boldsymbol{x},y)$	数据点
$D:\{(\boldsymbol{x}^{(1)},y^{(1)}),\ldots,(\boldsymbol{x}^{(m)},y^{(m)})\}$	数据集，m 表示数据的个数
$c:c(\boldsymbol{x})=y$	目标函数 (真实函数)，未知的函数，也是机器学习的原因和动力
$h:h(\boldsymbol{x})=y$	假设函数 (假想函数)，被尝试的函数
$H=\{h_1,h_2,\ldots,h_M\}$	假设函数集合，M 表示函数的个数
$g:g(\boldsymbol{x})=y$	最优假设函数，从 H 中选出最优 g (即 g 靠近 c)，机器学习的结果
A	机器学习算法
$M=\{A,H\}$	机器学习模型，模型由假设函数集合和算法组成
$e_{\text{out}}(h),e_{\text{in}}(h)$	假设函数 h 对应的样本外误差、样本内误差
$e_{\text{train}}(h),e_{\text{val}}(h),e_{\text{test}}(h)$	假设函数 h 对应的训练误差、验证误差、测试误差

使用代码示例

本书有一些简单的 Python, Matlab 和 Keras 的代码示例，各自对应的图标见下表以方便读者区分。

	Python 3.6		Matlab R2018a

本书中的一些补充代码，比如

- 线性回归玩转金郡房价预测（Linear Regression - King County Housing Price Prediction）
- 对率回归玩转亚马逊情感分析（Logistic Regression: Amazon Sentiment Analysis）

- 正则化回归玩转金郡房价预测（Regularized Regression - King County Housing Price Prediction）
- 决策树玩转借贷俱乐部（Decision Tree - Lending Club）
- 集成树玩转借贷俱乐部（Ensemble Tree - Lending Club）
- 极度梯度提升玩转借贷俱乐部（XGBoost - Lending Club）

均可从作者公众号“王的机器”中下载。

图标约定


表示对话


表示技巧


表示信息


表示警告

目　录