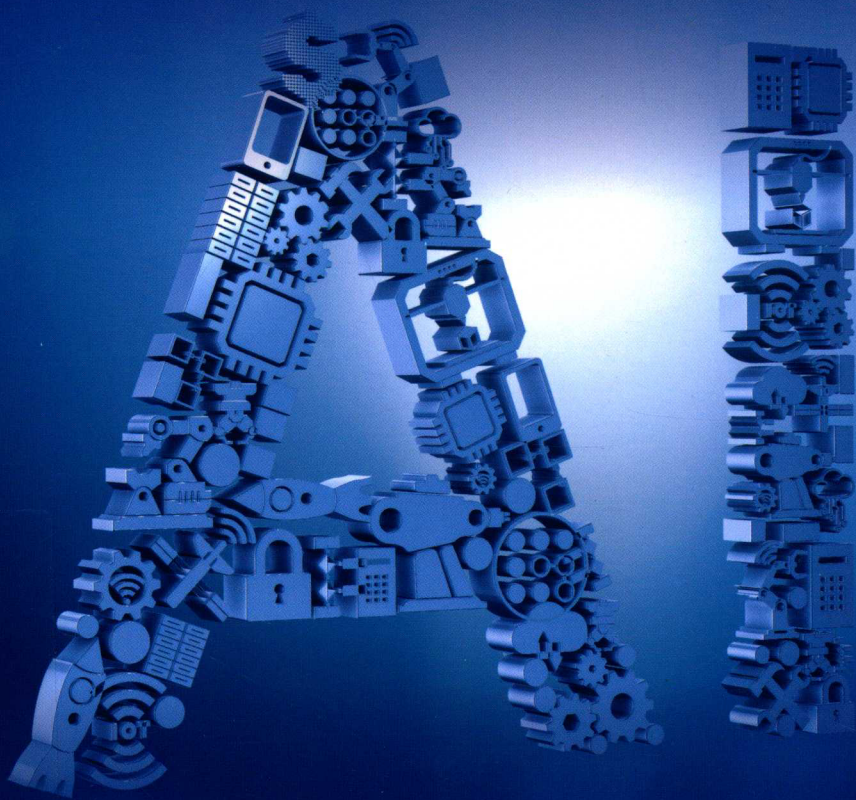


Broadview[®]
www.broadview.com.cn

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，又有多样性的代码展示，图文并茂

全彩
印刷



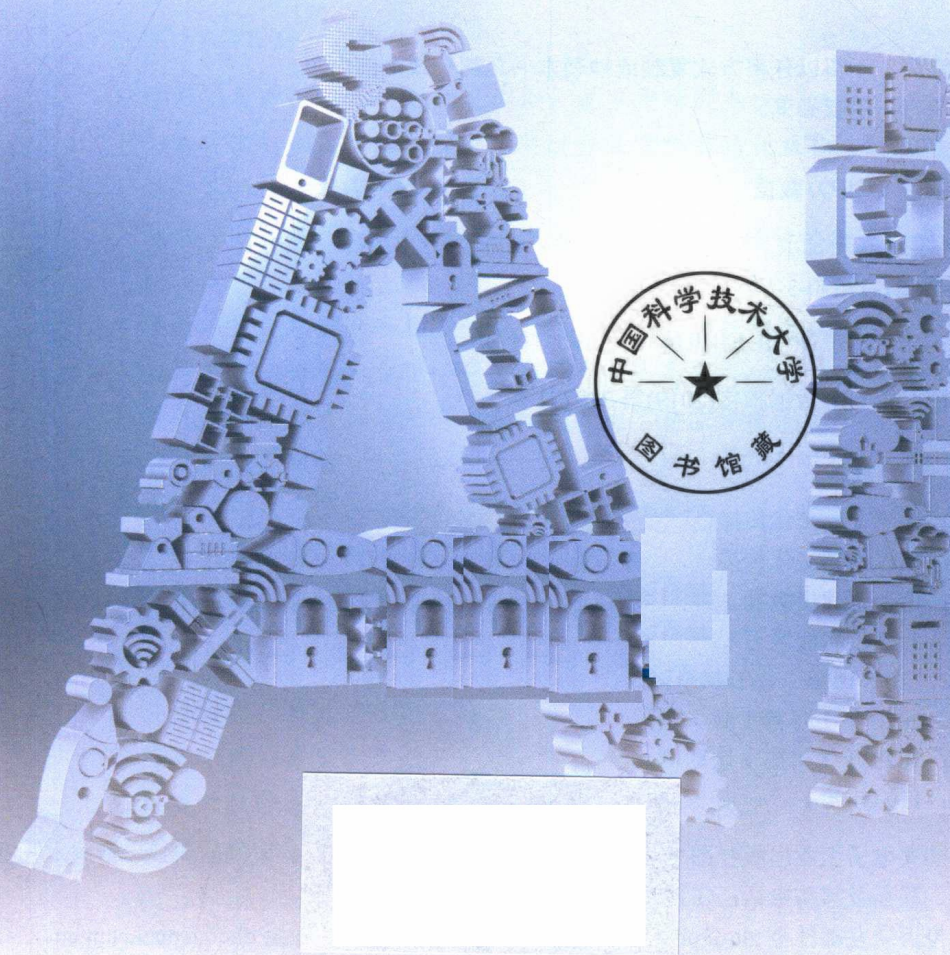
中国工信出版集团



电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>

快乐机器学习

[新加坡] 王圣元 著



电子工业出版社

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。

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

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 zltz@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看懂、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，这种以不清不白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。



本书阅读说明

书中符号说明

符 号	表达的意思
$\mathbf{x} = (x_1, x_2, \dots, x_n)$	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
(\mathbf{x}, y)	数据点
$D: \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(m)}, y^{(m)})\}$	数据集, m 表示数据的个数
$c: c(\mathbf{x}) = y$	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
$h: h(\mathbf{x}) = y$	假设函数 (假想函数), 被尝试的函数
$H = \{h_1, h_2, \dots, h_M\}$	假设函数集合, M 表示函数的个数
$g: g(\mathbf{x}) = y$	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

第 1 章 机器学习是什么——机器学习定义	1
引言	2
1.1 数据	5
1.1.1 结构型与非结构型数据	5
1.1.2 原始数据与加工	7
1.1.3 样本内数据与样本外数据	9
1.2 机器学习类别	9
1.2.1 有监督学习	10
1.2.2 无监督学习	10
1.2.3 半监督学习	11
1.2.4 增强学习	11
1.2.5 深度学习	11
1.2.6 迁移学习	12
1.3 性能度量	12
1.3.1 误差函数	13
1.3.2 回归度量	14
1.3.3 分类度量	15
1.4 总结	19
参考资料	20
第 2 章 机器学习可行吗——计算学习理论	22
引言	23
2.1 基础知识	25

2.1.1	二分类	25
2.1.2	对分	26
2.1.3	增长函数	29
2.1.4	突破点	30
2.2	核心推导	31
2.2.1	机器学习可行条件	31
2.2.2	从已知推未知	33
2.2.3	从民意调查到机器学习	35
2.2.4	从单一到有限	36
2.2.5	从有限到无限	37
2.2.6	从无限到有限	38
2.3	结论应用	39
2.3.1	VC 不等式	39
2.3.2	VC 维度	40
2.3.3	模型复杂度	40
2.3.4	样本复杂度	41
2.4	总结	42
	参考资料	43
	技术附录	43
第3章	机器学习怎么学——模型评估选择	47
	引言	48
3.1	模型评估	52
3.2	训练误差和测试误差	52
3.2.1	训练误差	52
3.2.2	真实误差	54
3.2.3	测试误差	57
3.2.4	学习理论	57
3.3	验证误差和交叉验证误差	60
3.3.1	验证误差	60
3.3.2	交叉验证误差	61

3.3.3 学习理论.....	62
3.4 误差剖析.....	64
3.4.1 误差来源.....	64
3.4.2 偏差—方差权衡.....	66
3.5 模型选择.....	67
3.6 总结.....	70
参考资料.....	71
技术附录.....	71
第4章 线性回归.....	73
引言.....	74
4.1 基础知识.....	75
4.1.1 标量微积分.....	75
4.1.2 向量微积分.....	76
4.2 模型介绍.....	77
4.2.1 核心问题.....	77
4.2.2 通用线性回归模型.....	83
4.2.3 特征缩放.....	84
4.2.4 学习率设定.....	86
4.2.5 数值算法比较.....	87
4.2.6 代码实现.....	89
4.3 总结.....	90
参考资料.....	90
第5章 对率回归.....	92
引言.....	93
5.1 基础内容.....	94
5.1.1 联系函数.....	94
5.1.2 函数绘图.....	95
5.2 模型介绍.....	96
5.2.1 核心问题.....	96

5.2.2 查准和查全	102
5.2.3 类别不平衡	104
5.2.4 线性不可分	105
5.2.5 多分类问题	106
5.2.6 代码实现	109
5.3 总结	110
参考资料	111
第 6 章 正则化回归	112
引言	113
6.1 基础知识	114
6.1.1 等值线图	114
6.1.2 坐标下降	116
6.2 模型介绍	116
6.2.1 核心问题	116
6.2.2 模型对比	122
6.2.3 最佳模型	125
6.2.4 代码实现	126
6.3 总结	126
参考资料	127
第 7 章 支持向量机	128
引言	129
7.1 基础知识	133
7.1.1 向量初体验	133
7.1.2 拉格朗日量	136
7.1.3 原始和对偶	137
7.2 模型介绍	138
7.2.1 硬间隔 SVM 原始问题	138
7.2.2 硬间隔 SVM 对偶问题	144
7.2.3 软间隔 SVM 原始问题	148

7.2.4	软间隔 SVM 对偶问题.....	150
7.2.5	空间转换.....	151
7.2.6	核技巧.....	155
7.2.7	核 SVM.....	158
7.2.8	SMO 算法.....	159
7.2.9	模型选择.....	161
7.3	总结.....	162
	参考资料.....	164
	技术附录.....	164
第 8 章	朴素贝叶斯.....	170
	引言.....	171
8.1	基础知识.....	174
8.1.1	两种概率学派.....	174
8.1.2	两种独立类别.....	174
8.1.3	两种学习算法.....	175
8.1.4	两种估计方法.....	176
8.1.5	两类概率分布.....	177
8.2	模型介绍.....	179
8.2.1	问题剖析.....	179
8.2.2	朴素贝叶斯算法.....	182
8.2.3	多元伯努利模型.....	183
8.2.4	多项事件模型.....	184
8.2.5	高斯判别分析模型.....	184
8.2.6	多分类问题.....	186
8.2.7	拉普拉斯校正.....	187
8.2.8	最大似然估计和最大后验估计.....	188
8.3	总结.....	190
	参考资料.....	191
	技术附录.....	191

第 9 章 决策树	195
引言	196
9.1 基础知识	198
9.1.1 多数规则	198
9.1.2 熵和条件熵	198
9.1.3 信息增益和信息增益比	200
9.1.4 基尼指数	201
9.2 模型介绍	201
9.2.1 二分类决策树	201
9.2.2 多分类决策树	209
9.2.3 连续值分裂	210
9.2.4 欠拟合和过拟合	211
9.2.5 预修剪和后修剪	212
9.2.6 数据缺失	215
9.2.7 代码实现	218
9.3 总结	219
参考资料	219
第 10 章 人工神经网络	220
引言	221
10.1 基本知识	223
10.1.1 转换函数	223
10.1.2 单输入单层单输出神经网络	224
10.1.3 多输入单层单输出神经网络	224
10.1.4 多输入单层多输出神经网络	225
10.1.5 多输入多层多输出神经网络	225
10.2 模型应用	227
10.2.1 创建神经网络模型	227
10.2.2 回归应用	230
10.2.3 分类应用	238

第 11 章 正向/反向传播	246
引言.....	247
11.1 基础知识.....	250
11.1.1 神经网络元素.....	250
11.1.2 链式法则.....	254
11.2 算法介绍.....	254
11.2.1 正向传播.....	254
11.2.2 梯度下降.....	257
11.2.3 反向传播.....	258
11.2.4 代码实现.....	262
11.3 总结.....	268
参考资料.....	268
技术附录.....	269
第 12 章 集成学习	272
引言.....	273
12.1 结合假设.....	277
12.1.1 语文和数学.....	277
12.1.2 准确和多样.....	278
12.1.3 独裁和民主.....	279
12.1.4 学习并结合.....	279
12.2 装袋法.....	280
12.2.1 基本概念.....	280
12.2.2 自助采样.....	280
12.2.3 结合假设.....	281
12.3 提升法.....	282
12.3.1 基本概念.....	282
12.3.2 最优加权.....	283
12.3.3 结合假设.....	285
12.4 集成方式.....	286

12.4.1 同质学习器	286
12.4.2 异质学习器	286
12.5 总结	288
参考资料	288
第 13 章 随机森林和提升树	289
引言	290
13.1 基础知识	293
13.1.1 分类回归树	293
13.1.2 前向分布算法	294
13.1.3 置换检验	295
13.2 模型介绍	296
13.2.1 随机森林	296
13.2.2 提升树	302
13.2.3 代码实现	306
13.3 总结	307
参考资料	307
第 14 章 极度梯度提升	309
引言	310
14.1 基础知识	311
14.1.1 树的重定义	311
14.1.2 树的复杂度	313
14.2 模型介绍	313
14.2.1 XGB 简介	313
14.2.2 XGB 的泛化度	314
14.2.3 XGB 的精确度	315
14.2.4 XGB 的速度	318
14.2.5 代码实现	324
14.3 总结	325
参考资料	326

第 15 章 本书总结.....	327
15.1 正交策略.....	328
15.2 单值评估指标.....	330
15.3 偏差和方差.....	332
15.3.1 理论定义.....	332
15.3.2 实用定义.....	334
15.3.3 最优误差.....	335
15.3.4 两者权衡.....	336
15.3.5 学习曲线.....	336
结语.....	339