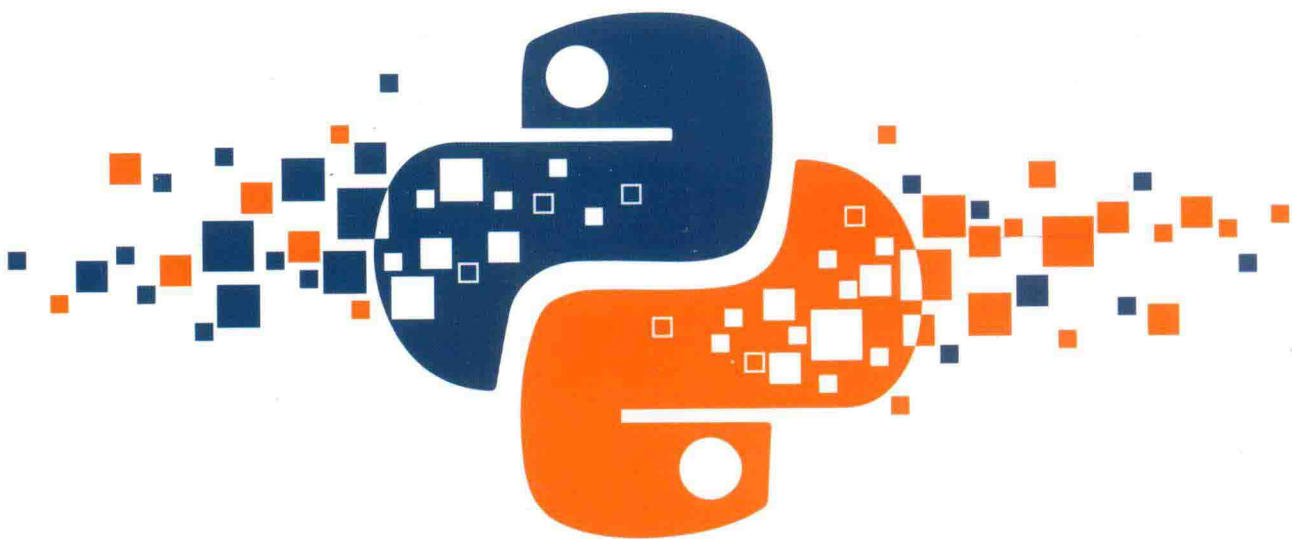


从零开始实践Python
深入浅出运用机器学习

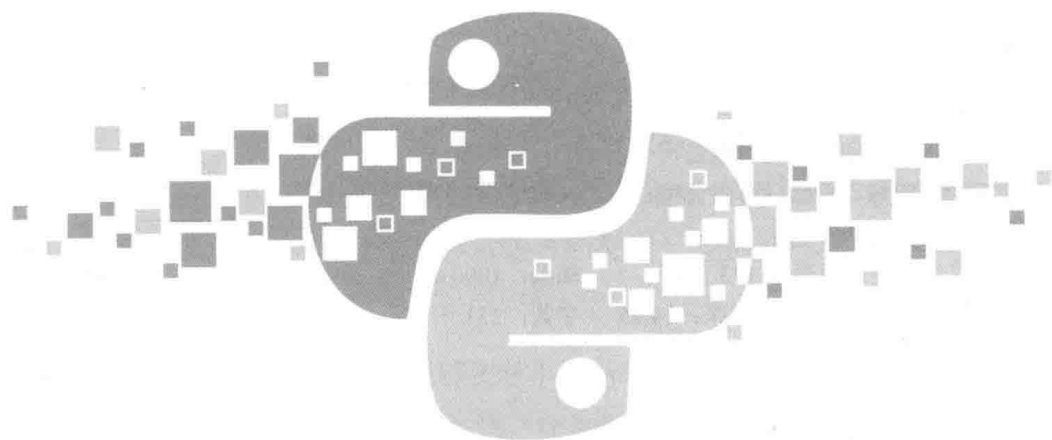


机器学习 Python实践

魏贞原 ■ 著

机器学习 Python实践

魏贞原■著



电子工业出版社
Publishing House of Electronics Industry
北京·BEIJING

内 容 简 介

本书系统地讲解了机器学习的基本知识，以及在实际项目中使用机器学习的基本步骤和方法；详细地介绍了在进行数据处理、分析时怎样选择合适的算法，以及建立模型并优化等方法，通过不同的例子展示了机器学习在具体项目中的应用和实践经验，是一本非常好的机器学习入门和实践的书籍。

不同于很多讲解机器学习的书籍，本书以实践为导向，使用 `scikit-learn` 作为编程框架，强调简单、快速地建立模型，解决实际项目问题。读者通过对本书的学习，可以迅速上手实践机器学习，并利用机器学习解决实际问题。本书非常适合于项目经理、有意从事机器学习开发的程序员，以及高校相关专业在读学生阅读。

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

机器学习：Python 实践 / 魏贞原著. —北京：电子工业出版社，2018.1

ISBN 978-7-121-33110-7

I. ①机… II. ①魏… III. ①机器学习②软件工具—程序设计 IV. ①TP311.561②TP181

中国版本图书馆 CIP 数据核字（2017）第 288527 号

策划编辑：石 倩

责任编辑：徐津平

特约编辑：赵树刚

印 刷：三河市双峰印刷装订有限公司

装 订：三河市双峰印刷装订有限公司

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：14.25 字数：251 千字

版 次：2018 年 1 月第 1 版

印 次：2018 年 1 月第 1 次印刷

定 价：59.00 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到 dbqq@phei.com.cn。

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



序言

人工智能作为一种新技术，它的发展变迁可以用著名的 S 曲线来表示。具有划时代意义的新技术（**Disruptive Innovation**），往往出现在社会发展的成熟期。这时国家的 GDP 增长开始减缓，生产成本居高不下，相比之下社会需求却没有显著增加。现在的人工智能技术发展正处于这样一个时期，可以说是应运而热。但我们更关心的是它是否到了腾飞的前期。我个人觉得时机已经到来，原因有三：数字化、物联网及人工智能（AI）技术。

第一，社会生活的高度数字化进程已经让人与人之间的联结几乎完全可以用数字化的方式来描述。这一切极大地归功于智能手机和社交媒体的深度普及。关于我们的信息都被记录在信息系统里，无论是以格式化的形式（ERP、银行系统、电商系统等），还是以非格式化的形式（社交媒体上的文字或语音的交流），这些都已成为机器可以分析解读的数据，而且还在不断积累中。

第二，物联网技术使信息系统能够实时不间断地从我们的日常活动中获取新的信息，并且可以通过双向通信机制给予实时反馈。比如智能手表和智能扫地机器人等。

第三，人工智能技术本身的发展已经到了足以支撑大规模商业化应用的阶段。无论是人脸识别还是医疗文献分析，人工智能已经作为工具出现在我们的身边。

很多人还在争论人工智能是否会成为人类的敌人，尤其是在 AI 技术发展到

泛人工智能（Artificial General Intelligence, AGI）甚至超级人工智能（Artificial Super Intelligence, ASI）的时候。这是否会成为现实或者什么时候会成为现实谁也说不清楚。但在人机同行的今天，如果还不赶快学习充实自己，明天我们肯定会遇到已经用 AI 版本升级了的人类对手。这种痛也许只有百多年前抵抗英法联军的八旗子弟才领悟过，但为时已晚。如何不掉队，赶上甚至超越时代发展潮流，学习新技术是唯一手段。

就像前几次工业革命一样，人工智能带来的技术革命也有三个关键成功因素（3M）：数据（原材料：Raw Material）、技术（机器：Machine）及商业模式（Business Model）。在现实世界中，数据的金矿已经大量积累，并等待我们去开采和精炼。技术，如 Python、卷积神经网络等都被成熟应用。商界精英更是想出了很多商业化的应用场景，在同声翻译、文本分析、医疗影像分析等领域展开了众多的投资和商业化应用。本书作为一本介绍 Python 的技术类专业书籍，立足机器学习中的监督式学习，围绕着课程、项目和方法深入浅出地介绍了如何使用 Python 来完成机器学习的相关工作。内容涵盖了人工智能的三个关键成功因素，以体系化和细致的讲解方便读者理解和学习有关机器学习的全过程。本书是一本实战性很强的参考书籍，帮助我们在人工智能时代迅速掌握新技术的精髓。

周德标

副合伙人

IBM 大中华区董事长执行助理

“周教授谈人工智能”微信公众号作者



前言

“这是最好的时代，也是最坏的时代”，这是英国文豪狄更斯的名著《双城记》开篇的第一句话，一百多年来不断被人引用。这里再次引用它来形容智能革命给我们带来的未来社会。从 2016 年 AlphaGo 在围棋比赛中战胜韩国选手李世石，到 2017 年 Master 战胜世界排名第一的围棋选手柯洁，人工智能再一次引起了世人的注意。在大数据出现之前，人工智能的概念虽然一直存在，但是计算机一直不擅长处理需要依赖人类的智慧解决的问题，现在换个思路就可以解决这些问题，其核心就是变智能问题为数据问题。由此，全世界开始了新一轮的技术革命——智能革命。

自从 1687 年艾萨克·牛顿发表了论文《自然定律》，对万有引力和三大运动定律进行了描述，人类社会进入了科学时代。在此之后，瓦特通过科学原理直接改进蒸汽机，开启了工业革命的篇章，由于机器的发明及运用成为这个时代的标志，因此历史学家称这个时代为“机器时代”。机器时代是利用机器代替人力，在原有的产业基础上加上蒸汽机形成新的产业，例如马车加上蒸汽机成为火车，改变了人的出行方式；帆船加上蒸汽机成为轮船，让货物的运输变得更加便捷。同时，原有的工匠被更加便宜的工人替代，社会的财富分配不均，社会进入动荡期，如英国大约花费了半个世纪的时间才完成了工业革命的变革。同样，第二次工业革命和信息革命，每一次变革都让财富更加集中，给社会带来动荡。第二次工业革命同样花费了半个世纪的时间，一代人才消除工业革命带来的影响，让大部分人受益。当前的智能革命也会带来财富的重新分配和社会的动荡，

当然目前的政府对这次革命的过程都有了足够的了解，能够把社会的动荡控制在最小范围，但是在变革中的人依然需要经受这次变革带来的动荡。

每一次变革都是一次思维方式的改进，工业革命是机器思维替代了农耕时代的思想；信息革命是香农博士（1916—2001 年）的信息论带来的思想方法替代机器思维，并成为社会主导思想；在这次智能革命中，以大数据为核心的思维方式将会主导这次变革。在历次的技术革命中，一个人、一家企业，甚至一个国家，可以选择的道路只有两条：要么加入变革的浪潮，成为前 2% 的弄潮儿；要么观望徘徊，被淘汰。

要成为 2% 的弄潮儿，需要积极拥抱这次智能变革，掌握在未来社会不会被淘汰的技能。在以大数据为基石的智能社会，利用机器学习算法对数据进行挖掘，是使机器更智能的关键，掌握数据挖掘是拥抱智能社会的举措之一。本书就将介绍如何利用机器学习算法来解决问题，对数据进行挖掘。

作者

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **下载资源**：本书如提供示例代码及资源文件，均可在 [下载资源](#) 处下载。
- **提交勘误**：您对书中内容的修改意见可在 [提交勘误](#) 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方 [读者评论](#) 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：<http://www.broadview.com.cn/33110>





目录

第一部分 初始

1	初识机器学习	2
1.1	学习机器学习的误区	2
1.2	什么是机器学习	3
1.3	Python 中的机器学习	3
1.4	学习机器学习的原则	5
1.5	学习机器学习的技巧	5
1.6	这本书不涵盖以下内容	6
1.7	代码说明	6
1.8	总结	6
2	Python 机器学习的生态圈	7
2.1	Python.....	7
2.2	SciPy.....	9
2.3	scikit-learn	9
2.4	环境安装	10
2.4.1	安装 Python.....	10

2.4.2	安装 SciPy	10
2.4.3	安装 scikit-learn.....	11
2.4.4	更加便捷的安装方式.....	11
2.5	总结.....	12
3	第一个机器学习项目	13
3.1	机器学习中的 Hello World 项目	13
3.2	导入数据.....	14
3.2.1	导入类库.....	14
3.2.2	导入数据集.....	15
3.3	概述数据.....	15
3.3.1	数据维度.....	16
3.3.2	查看数据自身.....	16
3.3.3	统计描述数据.....	17
3.3.4	数据分类分布.....	17
3.4	数据可视化.....	18
3.4.1	单变量图表.....	18
3.4.2	多变量图表.....	20
3.5	评估算法.....	20
3.5.1	分离出评估数据集.....	21
3.5.2	评估模式.....	21
3.5.3	创建模型.....	21
3.5.4	选择最优模型.....	22
3.6	实施预测.....	23
3.7	总结.....	24
4	Python 和 SciPy 速成	25
4.1	Python 速成	25
4.1.1	基本数据类型和赋值运算.....	26

4.1.2	控制语句.....	28
4.1.3	复杂数据类型.....	29
4.1.4	函数.....	32
4.1.5	with 语句.....	33
4.2	NumPy 速成.....	34
4.2.1	创建数组.....	34
4.2.2	访问数据.....	35
4.2.3	算数运算.....	35
4.3	Matplotlib 速成.....	36
4.3.1	绘制线条图.....	36
4.3.2	散点图.....	37
4.4	Pandas 速成.....	39
4.4.1	Series.....	39
4.4.2	DataFrame.....	40
4.5	总结.....	41

第二部分 数据理解

5	数据导入.....	44
5.1	CSV 文件.....	44
5.1.1	文件头.....	45
5.1.2	文件中的注释.....	45
5.1.3	分隔符.....	45
5.1.4	引号.....	45
5.2	Pima Indians 数据集.....	45
5.3	采用标准 Python 类库导入数据.....	46
5.4	采用 NumPy 导入数据.....	46
5.5	采用 Pandas 导入数据.....	47
5.6	总结.....	47

6 数据理解 48

6.1 简单地查看数据.....	48
6.2 数据的维度.....	49
6.3 数据属性和类型.....	50
6.4 描述性统计.....	50
6.5 数据分组分布（适用于分类算法）.....	51
6.6 数据属性的相关性.....	52
6.7 数据的分布分析.....	53
6.8 总结.....	54

7 数据可视化 55

7.1 单一图表.....	55
7.1.1 直方图.....	55
7.1.2 密度图.....	56
7.1.3 箱线图.....	57
7.2 多重图表.....	58
7.2.1 相关矩阵图.....	58
7.2.2 散点矩阵图.....	60
7.3 总结.....	61

第三部分 数据准备

8 数据预处理 64

8.1 为什么需要数据预处理.....	64
8.2 格式化数据.....	65
8.3 调整数据尺度.....	65
8.4 正态化数据.....	67
8.5 标准化数据.....	68

8.6	二值数据	69
8.7	总结	70

9 数据特征选定

9.1	特征选定	72
9.2	单变量特征选定	72
9.3	递归特征消除	73
9.4	主要成分分析	75
9.5	特征重要性	76
9.6	总结	76

第四部分 选择模型

10 评估算法

10.1	评估算法的方法	78
10.2	分离训练数据集和评估数据集	79
10.3	K 折交叉验证分离	80
10.4	弃一交叉验证分离	81
10.5	重复随机分离评估数据集与训练数据集	82
10.6	总结	83

11 算法评估矩阵

11.1	算法评估矩阵	85
11.2	分类算法矩阵	86
11.2.1	分类准确度	86
11.2.2	对数损失函数	87
11.2.3	AUC 图	88
11.2.4	混淆矩阵	90

11.2.5	分类报告.....	91
11.3	回归算法矩阵.....	93
11.3.1	平均绝对误差.....	93
11.3.2	均方误差.....	94
11.3.3	决定系数 (R^2)	95
11.4	总结.....	96
12	审查分类算法.....	97
12.1	算法审查.....	97
12.2	算法概述.....	98
12.3	线性算法.....	98
12.3.1	逻辑回归.....	99
12.3.2	线性判别分析.....	100
12.4	非线性算法.....	101
12.4.1	K 近邻算法.....	101
12.4.2	贝叶斯分类器.....	102
12.4.3	分类与回归树.....	103
12.4.4	支持向量机.....	104
12.5	总结.....	105
13	审查回归算法.....	106
13.1	算法概述.....	106
13.2	线性算法.....	107
13.2.1	线性回归算法.....	107
13.2.2	岭回归算法.....	108
13.2.3	套索回归算法.....	109
13.2.4	弹性网络回归算法.....	110
13.3	非线性算法.....	111
13.3.1	K 近邻算法.....	111

13.3.2	分类与回归树	112
13.3.3	支持向量机	112
13.4	总结	113
14	算法比较	115
14.1	选择最佳的机器学习算法	115
14.2	机器学习算法的比较	116
14.3	总结	118
15	自动流程	119
15.1	机器学习的自动流程	119
15.2	数据准备和生成模型的 Pipeline	120
15.3	特征选择和生成模型的 Pipeline	121
15.4	总结	122
第五部分 优化模型		
16	集成算法	124
16.1	集成的方法	124
16.2	装袋算法	125
16.2.1	装袋决策树	125
16.2.2	随机森林	126
16.2.3	极端随机树	127
16.3	提升算法	129
16.3.1	AdaBoost	129
16.3.2	随机梯度提升	130
16.4	投票算法	131
16.5	总结	132

17	算法调参	133
17.1	机器学习算法调参	133
17.2	网格搜索优化参数	134
17.3	随机搜索优化参数	135
17.4	总结	136

第六部分 结果部署

18	持久化加载模型	138
18.1	通过 pickle 序列化和反序列化机器学习的模型	138
18.2	通过 joblib 序列化和反序列化机器学习的模型	140
18.3	生成模型的技巧	141
18.4	总结	141

第七部分 项目实践

19	预测模型项目模板	144
19.1	在项目中实践机器学习	145
19.2	机器学习项目的 Python 模板	145
19.3	各步骤的详细说明	146
	步骤 1: 定义问题	147
	步骤 2: 理解数据	147
	步骤 3: 数据准备	147
	步骤 4: 评估算法	147
	步骤 5: 优化模型	148
	步骤 6: 结果部署	148

19.4	使用模板的小技巧	148
19.5	总结	149
20	回归项目实例	150
20.1	定义问题	150
20.2	导入数据	151
20.3	理解数据	152
20.4	数据可视化	155
20.4.1	单一特征图表	155
20.4.2	多重数据图表	157
20.4.3	思路总结	159
20.5	分离评估数据集	159
20.6	评估算法	160
20.6.1	评估算法——原始数据	160
20.6.2	评估算法——正态化数据	162
20.7	调参改善算法	164
20.8	集成算法	165
20.9	集成算法调参	167
20.10	确定最终模型	168
20.11	总结	169
21	二分类实例	170
21.1	问题定义	170
21.2	导入数据	171
21.3	分析数据	172
21.3.1	描述性统计	172
21.3.2	数据可视化	177
21.4	分离评估数据集	180
21.5	评估算法	180

21.6	算法调参.....	184
21.6.1	K 近邻算法调参.....	184
21.6.2	支持向量机调参.....	185
21.7	集成算法.....	187
21.8	确定最终模型.....	190
21.9	总结.....	190
22	文本分类实例.....	192
22.1	问题定义.....	192
22.2	导入数据.....	193
22.3	文本特征提取.....	195
22.4	评估算法.....	196
22.5	算法调参.....	198
22.5.1	逻辑回归调参.....	199
22.5.2	朴素贝叶斯分类器调参.....	199
22.6	集成算法.....	200
22.7	集成算法调参.....	201
22.8	确定最终模型.....	202
22.9	总结.....	203
附录 A	205
A.1	IDE PyCharm 介绍.....	205
A.2	Python 文档.....	206
A.3	SciPy、NumPy、Matplotlib 和 Pandas 文档.....	206
A.4	树模型可视化.....	206
A.5	scikit-learn 的算法选择路径.....	209
A.6	聚类分析.....	209