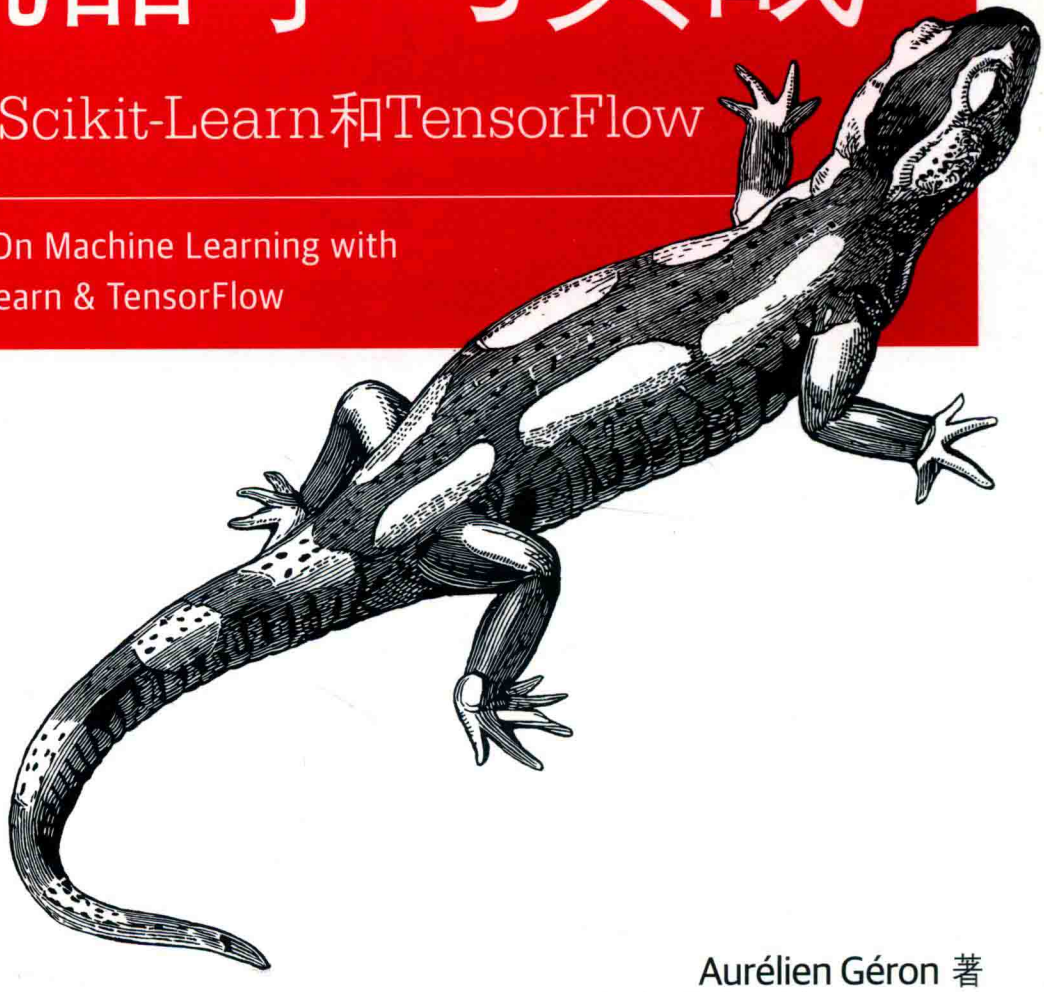
O'REILLY®

HZ BOOKS
华章IT

机器学习实战

基于Scikit-Learn和TensorFlow

Hands-On Machine Learning with
Scikit-Learn & TensorFlow



Aurélien Géron 著
王静源 贾玮 边蕤 邱俊涛 译

机械工业出版社
China Machine Press

机器学习实战：基于 Scikit-Learn 和 TensorFlow

Aurélien Géron 著

王静源 贾玮 边蕤 邱俊涛 译

Beijing · Boston · Farnham · Sebastopol · Tokyo O'REILLY®

dia, Inc. 授权机械工业出版社出版

机械工业出版社

图书在版编目（CIP）数据

机器学习实战：基于 Scikit-Learn 和 TensorFlow/（法）奥雷利安·杰龙著；王静源等译 . —北京：机械工业出版社，2018.7

(O'Reilly 精品图书系列)

书名原文：Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques for Building Intelligent Systems

ISBN 978-7-111-60302-3

I. 机… II. ①奥… ②王… III. 机器学习 IV. TP181

中国版本图书馆 CIP 数据核字（2018）第 174102 号

北京市版权局著作权合同登记

图字：01-2017-3412 号

书　　名 / 机器学习实战：基于 Scikit-Learn 和 TensorFlow

书　　号 / ISBN 978-7-111-60302-3

责任编辑 / 陈佳媛

封面设计 / Randy Comer，张健

出版发行 / 机械工业出版社

地　　址 / 北京市西城区百万庄大街 22 号（邮政编码 100037）

印　　刷 / 北京市兆成印刷有限责任公司

开　　本 / 178 毫米 ×233 毫米　16 开本　29.75 印张

版　　次 / 2018 年 9 月第 1 版　2018 年 9 月第 1 次印刷

定　　价 / 119.00 元（册）

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：(010)88379426；88361066

购书热线：(010)68326294；88379649；68995259

投稿热线：(010)88379604

读者信箱：hzit@hzbook.com

O'Reilly Media, Inc. 介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 Make 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过书籍出版，在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar 博客有口皆碑。”

——Wired

“O'Reilly 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——Business 2.0

“O'Reilly Conference 是聚集关键思想领袖的绝对典范。”

——CRN

“一本 O'Reilly 的书就代表一个有用、有前途、需要学习的主题。”

——Irish Times

“Tim 是位特立独行的商人，他不光放眼于最长远、最广阔的视野并且切实地按照 Yogi Berra 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去 Tim 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——Linux Journal

译者序

随着 AlphaGo 在人机大战中一举成名，关于机器学习的研究开始广受关注，数据科学家也一跃成为“21 世纪最性感的职业”。关于机器学习和神经网络的广泛应用虽然兴起不久，但是对这两个密切关联的领域的研究其实已经持续了好几十年，早已形成了系统化的知识体系。对于想要踏入机器学习领域的初学者而言，理论知识的获取并非难事。

本书作者 Aurélien Géron 曾经是谷歌工程师，在 2013 年至 2016 年，主导了 YouTube 的视频分类工程，拥有丰富的机器学习项目经验。作者的写作初衷是希望从实践出发，手把手地帮助开发者从零开始搭建起一个神经网络。这也正构成了本书区别于其他机器学习教程的最重要的特质——不再偏向于原理研究的角度，而是从开发者的实践角度出发，在动手写代码的过程中，循序渐进地了解机器学习的理论知识和工具的实践技巧。对于想要快速上手机器学习的开发者来说，本书不啻为一个非常值得尝试的起点项目。

本书主要分为两个部分。第一部分为第 1～8 章，涵盖机器学习的基础理论知识和基本算法——从线性回归到随机森林等，帮助读者掌握 Scikit-Learn 的常用方法；第二部分为第 9～16 章，探讨深度学习和常用框架 TensorFlow，一步一个脚印地带领读者使用 TensorFlow 搭建和训练深度神经网络，以及卷积神经网络。

书中涉及不少数学公式，作者对抽象的公式背后的含义也都一一做出了阐释，因此即便是对数学不敏感的初学者，也同样能够理解机器学习任务的实质。

书中所涉及的专业术语与概念较多，部分概念及术语尚无公认的中文译法，因此我们较多地参考了网络上和研究论文中常用的译法，若有不适合读者朋友的术语，可根据原词确认其原始词意。在翻译过程中虽然力求准确地反映原著内容，但由于译者水平有限，如有错漏之处，恳请读者批评指正。

本书译者均来自于 ThoughtWorks 咨询公司，王静源翻译了第 1～8 章，贾玮翻译了第

13～16章，边蕤翻译了第11章和第12章，邱俊涛翻译了第9章和第10章，并对全书译文进行校对和最后定稿。

最后要感谢华章公司的编辑，他们为保证本书的质量做出了大量的编辑和校正工作，在此深表谢意。

译者

2018年3月

目录

前言 1

第一部分 机器学习基础

第 1 章 机器学习概览 11

什么是机器学习 12

为什么要使用机器学习 12

机器学习系统的种类 15

监督式 / 无监督式学习 16

批量学习和在线学习 21

基于实例与基于模型的学习 24

机器学习的主要挑战 29

训练数据的数量不足 29

训练数据不具代表性 30

质量差的数据 32

无关特征 32

训练数据过度拟合 33

训练数据拟合不足 34

退后一步 35

测试与验证 35

练习 37

第 2 章 端到端的机器学习项目 39

使用真实数据 39

..... 40

框架问题 41
选择性能指标 42
检查假设 45
获取数据 45
创建工作区 45
下载数据 48
快速查看数据结构 49
创建测试集 52
从数据探索和可视化中获得洞见 56
将地理数据可视化 57
寻找相关性 59
试验不同属性的组合 61
机器学习算法的数据准备 62
数据清理 63
处理文本和分类属性 65
自定义转换器 67
特征缩放 68
转换流水线 68
选择和训练模型 70
培训和评估训练集 70
使用交叉验证来更好地进行评估 72
微调模型 74
网格搜索 74
随机搜索 76
集成方法 76
分析最佳模型及其错误 76
通过测试集评估系统 77
启动、监控和维护系统 78
试试看 79
练习 79

第 3 章 分类 80

MNIST 80
训练一个二元分类器 82
性能考核 83

使用交叉验证测量精度 83
混淆矩阵 84
精度和召回率 86
精度 / 召回率权衡 87
ROC 曲线 90
多类别分类器 93
错误分析 95
多标签分类 98
多输出分类 99
练习 100

第 4 章 训练模型 102

线性回归 103
标准方程 104
计算复杂度 106
梯度下降 107
批量梯度下降 110
随机梯度下降 112
小批量梯度下降 114
多项式回归 115
学习曲线 117
正则线性模型 121
岭回归 121
套索回归 123
弹性网络 125
早期停止法 126
逻辑回归 127
概率估算 127
训练和成本函数 128
决策边界 129
Softmax 回归 131
练习 134

第 5 章 支持向量机 136

线性 SVM 分类 136

软间隔分类 137
非线性 SVM 分类 139
多项式核 140
添加相似特征 141
高斯 RBF 核函数 142
计算复杂度 143
SVM 回归 144
工作原理 145
决策函数和预测 146
训练目标 146
二次规划 148
对偶问题 149
核化 SVM 149
在线 SVM 151
练习 152

第 6 章 决策树 154

决策树训练和可视化 154
做出预测 155
估算类别概率 157
CART 训练算法 158
计算复杂度 158
基尼不纯度还是信息熵 159
正则化超参数 159
回归 161
不稳定性 162
练习 163

第 7 章 集成学习和随机森林 165

投票分类器 165
bagging 和 pasting 168
Scikit-Learn 的 bagging 和 pasting 169
包外评估 170
Random Patches 和随机子空间 171

随机森林 172
极端随机树 173
特征重要性 173
提升法 174
AdaBoost 175
梯度提升 177
堆叠法 181
练习 184

第 8 章 降维 185
维度的诅咒 186
数据降维的主要方法 187
投影 187
流形学习 189
PCA 190
保留差异性 190
主成分 191
低维度投影 192
使用 Scikit-Learn 192
方差解释率 193
选择正确数量的维度 193
PCA 压缩 194
增量 PCA 195
随机 PCA 195
核主成分分析 196
选择核函数和调整超参数 197
局部线性嵌入 199
其他降维技巧 200
练习 201

第二部分 神经网络和深度学习

第 9 章 运行 TensorFlow 205
安装 207
创建一个计算图并在会话中执行 208

管理图 209
节点值的生命周期 210
TensorFlow 中的线性回归 211
实现梯度下降 211
手工计算梯度 212
使用自动微分 212
使用优化器 214
给训练算法提供数据 214
保存和恢复模型 215
用 TensorBoard 来可视化图和训练曲线 216
命名作用域 219
模块化 220
共享变量 222
练习 225

第 10 章 人工神经网络简介 227

从生物神经元到人工神经元 227
生物神经元 228
具有神经元的逻辑计算 229
感知器 230
多层感知器和反向传播 233
用 TensorFlow 的高级 API 来训练 MLP 236
使用纯 TensorFlow 训练 DNN 237
构建阶段 237
执行阶段 240
使用神经网络 241
微调神经网络的超参数 242
隐藏层的个数 242
每个隐藏层中的神经元数 243
激活函数 243
练习 244

第 11 章 训练深度神经网络 245

梯度消失 / 爆炸问题 245

Xavier 初始化和 He 初始化 246
非饱和激活函数 248
批量归一化 250
梯度剪裁 254
重用预训练图层 255
重用 TensorFlow 模型 255
重用其他框架的模型 256
冻结低层 257
缓存冻结层 257
调整、丢弃或替换高层 258
模型动物园 258
无监督的预训练 259
辅助任务中的预训练 260
快速优化器 261
Momentum 优化 261
Nesterov 梯度加速 262
AdaGrad 263
RMSProp 265
Adam 优化 265
学习速率调度 267
通过正则化避免过度拟合 269
提前停止 269
ℓ_1和ℓ_2正则化 269
dropout 270
最大范数正则化 273
数据扩充 274
实用指南 275
练习 276

第 12 章 跨设备和服务器的分布式 TensorFlow 279

一台机器上的多个运算资源 280
安装 280
管理 GPU RAM 282
在设备上操作 284
并行执行 287

控制依赖 288
多设备跨多服务器 288
开启一个会话 290
master 和 worker 服务 290
分配跨任务操作 291
跨多参数服务器分片变量 291
用资源容器跨会话共享状态 292
使用 TensorFlow 队列进行异步通信 294
直接从图中加载数据 299
在 TensorFlow 集群上并行化神经网络 305
一台设备一个神经网络 305
图内与图间复制 306
模型并行化 308
数据并行化 309
练习 314

第 13 章 卷积神经网络 315

视觉皮层的组织结构 315
卷积层 317
过滤器 318
多个特征图的叠加 319
TensorFlow 实现 321
内存需求 323
池化层 323
CNN 架构 325
LeNet-5 326
AlexNet 327
GoogLeNet 328
ResNet 331
练习 334

第 14 章 循环神经网络 337

循环神经元 337
记忆单元 339
输入和输出序列 340

TensorFlow 中的基本 RNN 341
通过时间静态展开 342
通过时间动态展开 344
处理长度可变输入序列 344
处理长度可变输出序列 345
训练 RNN 346
训练序列分类器 346
训练预测时间序列 348
创造性的 RNN 352
深层 RNN 353
在多个 GPU 中分配一个深层 RNN 354
应用丢弃机制 355
多个时间迭代训练的难点 356
LSTM 单元 357
窥视孔连接 359
GRU 单元 359
自然语言处理 361
单词嵌入 361
用于机器翻译的编码器 – 解码器网络 362
练习 364

第 15 章 自动编码器 366

高效的数据表示 366
使用不完整的线性自动编码器实现 PCA 368
栈式自动编码器 369
TensorFlow 实现 370
权重绑定 371
一次训练一个自动编码器 372
重建可视化 374
特征可视化 375
使用堆叠的自动编码器进行无监控的预训练 376
去噪自动编码器 377
TensorFlow 实现 378
稀疏自动编码器 379
TensorFlow 实现 380

变分自动编码器 381
生成数字 384
其他自动编码器 385
练习 386

第 16 章 强化学习 388

学习奖励最优化 389
策略搜索 390
OpenAI gym 介绍 391
神经网络策略 394
评估行为：信用分配问题 396
策略梯度 397
马尔可夫决策过程 401
时间差分学习与 Q 学习 405
探索策略 406
逼近 Q 学习 407
使用深度 Q 学习玩吃豆人游戏 407
练习 414
致谢 415

附录 A 练习答案 416

附录 B 机器学习项目清单 438

附录 C SVM 对偶问题 444

附录 D 自动微分 447

附录 E 其他流行的 ANN 架构 453

前言

机器学习浪潮

2006 年，Geoffrey Hinton 等人发表了一篇论文[注1]，展示了如何训练能够高精度（>98%）识别手写数字的深度神经网络。他们将这种技术称为“深度学习”。在当时，深度神经网络的训练被普遍认为是不可能的。这篇论文[注2]重新激起了科学界的兴趣，不久之后，许多新的论文展示了深度学习不仅是可行的，而且（在超级计算能力和大数据的帮助下）能够取得令人瞩目的成就，是其他机器学习技术所难以企及的。这种热情很快扩展到机器学习相关的许多其他领域。

历经十年的快速发展，机器学习已经征服了整个行业。它已经成为众多高科技产品中的黑科技核心：对你的网络搜索结果进行排名，实现智能手机的语音识别，为你推荐视频，打败世界围棋冠军。在你意识到之前，它甚至会驾驶你的汽车。

你的项目中的机器学习

你已对机器学习充满了兴趣，并迫不及待想要投入其中了吧？

可能你想让自己制作的机器人有思想？让它可以识别人脸？或者学会走路？

或者你们公司有大量的数据（用户日志、财务数据、产品数据、机器传感器数据、热线数据、HR 报告等），如果知道去哪儿找，你会挖掘出一些隐藏的宝石，比如：

- 细分客户群，并为每个群体设置最佳的市场策略

注 1：详见 Hinton 的主页 *http://www.cs.toronto.edu/~hinton/*。

注 2：虽然 Yann Lecun 的深卷积神经网络自 20 世纪 90 年代以来在图像识别方面运作良好，但这并非一般用途。