O'REILLY®

Broadview®
www.broadview.com.cn

Python 机器学习手册

从数据预处理到深度学习

Machine Learning with Python Cookbook

[美] Chris Albon 著
韩慧昌 林然 徐江 译

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

O'REILLY®

Python 机器学习手册

从数据预处理到深度学习

Machine Learning with Python Cookbook

[美]Chris Albon 著
韩慧昌 林然 徐江 译

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

内容简介

本书采用基于任务的方式来介绍如何在机器学习中使用Python。书中有近200个独立的解决方案，针对的都是数据科学家或机器学习工程师在构建模型时可能遇到的常见任务，涵盖从简单的矩阵和向量运算到特征工程以及神经网络的构建。所有方案都提供了相关代码，读者可以复制并粘贴这些代码，用在自己的程序中。

本书不是机器学习的入门书，适合熟悉机器学习理论和概念的读者阅读。你可以将本书作为案头参考书，在机器学习的日常开发中遇到问题时，随时借鉴书中代码，快速解决问题。

©2018 by Chris Albon
Simplified Chinese Edition, jointly published by O'Reilly Media, Inc. and Publishing House of Electronics Industry, 2019. Authorized translation of the English edition, 2018 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.
All rights reserved including the rights of reproduction in whole or in part in any form.

本书简体中文版专有出版权由O'Reilly Media, Inc. 授予电子工业出版社。未经许可，不得以任何方式复制或抄袭本书的任何部分。专有出版权受法律保护。

版权贸易合同登记号　图字：01-2018-5011

图书在版编目（CIP）数据

Python机器学习手册：从数据预处理到深度学习 /（美）克里斯·阿尔本（Chris Albon）著；韩慧昌，林然，徐江译. —北京：电子工业出版社，2019.7
书名原文：Machine Learning with Python Cookbook
ISBN 978-7-121-36962-9

Ⅰ. ①P… Ⅱ. ①克… ②韩… ③林… ④徐… Ⅲ. ①软件工具－程序设计 Ⅳ. ①TP311.561

中国版本图书馆CIP数据核字（2019）第124840号

责任编辑：许　艳
封面设计：Karen Montgomery　张　健
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：22.75　字数：498千字
版　　次：2019年7月第1版
印　　次：2019年7月第1次印刷
定　　价：89.00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。

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

序

在过去的几年中，机器学习已经渗透到企业、非营利组织和政府的日常运作中。随着机器学习热度的增加，在对机器学习从业者的指导方面，涌现出一批高质量的文献。这些文献培养了整整一代的数据科学家和机器学习工程师。这些文献提供了学习资源，为人们讲解机器学习是什么及其工作原理。尽管这种方法富有成效，但却遗漏了一部分内容：机器学习日常开发中的细节。这就是笔者写本书的动机——本书不是写给学生读者学习机器学习理论的大部头，而是写给专业人士的“扳手型”工具书。我希望你把它放在书桌上，把你感兴趣的某些页折起来，在日常开发中需要解决实际问题时就拿过来翻一翻。

更具体地说，本书采用基于任务的方式来介绍机器学习，有近 200 个独立的解决方案（你可以复制并粘贴这些代码，它们都是可以正常运行的），针对的都是数据科学家或机器学习工程师在构建模型时可能遇到的常见任务。

本书的最终目标是成为人们在构建真实的机器学习系统时的参考书。例如，假设你有一个 JSON 文件，其中包含 1000 个数据分类特征和数值型特征，并且目标向量的分类不均衡，你想得到一个可解释的模型，那么使用本书提供的解决方案可以帮助你解决如下问题：

- 加载 JSON 文件（2.5 节）
- 对特征进行标准化（4.2 节）
- 对特征字典编码（5.3 节）
- 填充缺失的分类值（5.4 节）
- 使用主成分进行特征降维（9.1 节）
- 使用随机搜索选择最佳模型（12.2 节）
- 训练随机森林分类器（14.4 节）
- 选择随机森林中的重要特征（14.7 节）

试读结束：需要全本请在线购买： www.ertongbook.com

本书的目标是让你：

1．复制 / 粘贴代码，并确信它们能很好地运行在玩具数据集（toy dataset）[1] 上。

2．阅读每一节后面的“讨论”以增进对代码背后理论的理解，并了解哪些参数需要重点考虑。

3．对书中的代码进行各种组合与修改，以构建实际的应用。

哪些人适合阅读本书

本书不是机器学习的入门书。如果你对机器学习的基本概念还不太了解，或者从未花时间学习过机器学习，请不要购买本书。本书适合机器学习的实践者阅读，他们熟悉机器学习的理论和概念，可以参考书中的代码快速解决在日常开发中遇到的问题。

另外，本书假定读者熟悉 Python 编程语言和包管理。

哪些人不适合阅读本书

如前所述，本书不是机器学习的入门书，因此它不应该是你的第一本机器学习书。如果你还不熟悉交叉验证、随机森林和梯度下降等概念，那么建议你先阅读一本入门类机器学习图书，再阅读本书来学习机器学习的实用解决方案。

书中用到的术语

机器学习会用到很多领域的技术，包括计算机科学、统计学和数学，因此在关于机器学习的讨论中会使用各种各样的术语：

观察值[2]（*observation*）

我们观察到的单个单位——例如，一个人、一次销售或一条记录。

学习算法（*learning algorithm*）

用来学习模型的最佳参数的算法——例如，线性回归、朴素贝叶斯或决策树。

模型（*model*）

学习算法的输出。学习算法训练出的模型可以用来做预测。

1　译者注：可以理解为小型数据集，比真实世界的大数据集小得多、干净得多。
2　译者注：也可理解为样本，本书中某些地方将 observation 译作“样本”。

参数（*parameter*）

一个模型在训练过程中学习到的权重或系数。

超参数（*hyperparameter*）

一个学习算法在训练前需要设置的一组参数。

性能（*performance*）

用来评估模型的指标。

损失（*loss*）

一个需要在训练中最小化或最大化的指标。

训练（*train*）

使用类似梯度下降之类的数学方法将一个学习算法应用到数据上。

拟合（*fit*）

使用分析方法将一个机器学习算法应用到数据上。

数据（*data*）

一组观察值。

致谢

没有一些朋友和陌生人的帮助，这本书是不可能完成的。很难列出提供过帮助的所有人的名字，但我想至少提一下这些人：Angela Bassa、Teresa Borcuch、Justin Bozonier、Andre deBruin、Numa Dhamani、Dan Friedman、Joel Grus、Sarah Guido、Bill Kambouroglou、Mat Kelcey、Lizzie Kumar、Hilary Parker、Niti Paudyal、Sebastian Raschka 和 Shreya Shankar。

我欠他们所有人一瓶啤酒，或者五瓶。

关于作者

Chris Albon 是一位有十年经验的数据科学家和政治学家，他将统计学习、人工智能和软件工程应用到政治和社会活动以及人道主义活动中，譬如监查选举情况、灾难救助等。目前，Chris 是肯尼亚创业公司 BRCK 的首席数据科学家。这家公司致力于为前沿市场的互联网用户构建一个稳健的网络。

封面说明

本书封面上的动物是绿颊非洲咬鹃（Narina trogon 或 Apaloderma narina），Narina trogon 是以法国鸟类学家 Fran Çois Levaillant 情妇的名字命名的，她的名字取自科伊科伊（Khoikhoi）语中的一个词，意为“花朵”，不过这个词读起来很费劲。绿颊非洲咬鹃在非洲地区很常见，无论是低地还是高原，无论是热带还是温带气候，它们都能适应。这些鸟儿通常把巢穴筑在树洞里。它们分布广泛，因此也很少被认为需要保护。

绿颊非洲咬鹃主要以昆虫、小型无脊椎动物、小型啮齿动物和小型爬行动物为食。雄鸟羽毛鲜亮，会不断发出刺耳但低声的鸣叫以守卫领地和吸引雌鸟。无论是雌鸟还是雄鸟，它们的上半身都为绿色，尾部的羽毛为带金属光泽的蓝绿色。雌鸟的面颊和胸部羽毛为棕色，而雄鸟的下半身羽毛为鲜红色。雏鸟羽毛的颜色与雌鸟类似，不过翅膀尖部为独特的白色。

O’Reilly 图书封面上的很多动物都是濒危动物，它们对于这个世界很重要。如果想了解如何提供帮助，请访问网站 animals.oreilly.com。

本书封面图片来自 *Wood’s Animate Creation*。

目录

第 1 章　向量、矩阵和数组 1

1.0　简介 1
1.1　创建一个向量 1
1.2　创建一个矩阵 2
1.3　创建一个稀疏矩阵 3
1.4　选择元素 5
1.5　展示一个矩阵的属性 6
1.6　对多个元素同时应用某个操作 7
1.7　找到最大值和最小值 8
1.8　计算平均值、方差和标准差 9
1.9　矩阵变形 10
1.10　转置向量或矩阵 11
1.11　展开一个矩阵 12
1.12　计算矩阵的秩 13
1.13　计算行列式 14
1.14　获取矩阵的对角线元素 14
1.15　计算矩阵的迹 15
1.16　计算特征值和特征向量 16
1.17　计算点积 17
1.18　矩阵的相加或相减 18
1.19　矩阵的乘法 19

1.20 计算矩阵的逆 20
1.21 生成随机数 21

第 2 章 加载数据 23

2.0 简介 23
2.1 加载样本数据集 23
2.2 创建仿真数据集 25
2.3 加载 CSV 文件 28
2.4 加载 Excel 文件 29
2.5 加载 JSON 文件 29
2.6 查询 SQL 数据库 31

第 3 章 数据整理 33

3.0 简介 33
3.1 创建一个数据帧 34
3.2 描述数据 35
3.3 浏览数据帧 37
3.4 根据条件语句来选择行 39
3.5 替换值 40
3.6 重命名列 41
3.7 计算最小值、最大值、总和、平均值与计数值 43
3.8 查找唯一值 44
3.9 处理缺失值 45
3.10 删除一列 47
3.11 删除一行 48
3.12 删除重复行 49
3.13 根据值对行分组 51
3.14 按时间段对行分组 52
3.15 遍历一个列的数据 54
3.16 对一列的所有元素应用某个函数 55
3.17 对所有分组应用一个函数 56
3.18 连接多个数据帧 57
3.19 合并两个数据帧 59

第 4 章 处理数值型数据 63
4.0 简介 63
4.1 特征的缩放 63
4.2 特征的标准化 65
4.3 归一化观察值 66
4.4 生成多项式和交互特征 69
4.5 转换特征 70
4.6 识别异常值 71
4.7 处理异常值 73
4.8 将特征离散化 75
4.9 使用聚类的方式将观察值分组 77
4.10 删除带有缺失值的观察值 79
4.11 填充缺失值 81
第 5 章 处理分类数据 83
5.0 简介 83
5.1 对 nominal 型分类特征编码 84
5.2 对 ordinal 分类特征编码 86
5.3 对特征字典编码 88
5.4 填充缺失的分类值 91
5.5 处理不均衡分类 93
第 6 章 处理文本 97
6.0 简介 97
6.1 清洗文本 97
6.2 解析并清洗 HTML 99
6.3 移除标点 100
6.4 文本分词 101
6.5 删除停止词（stop word） 102
6.6 提取词干 103
6.7 标注词性 104
6.8 将文本编码成词袋（Bag of Words） 107
6.9 按单词的重要性加权 109

试读结束：需要全本请在线购买： www.ertongbook.com

第 7 章　处理日期和时间 113

7.0　简介 113
7.1　把字符串转换成日期 113
7.2　处理时区 115
7.3　选择日期和时间 116
7.4　将日期数据切分成多个特征 117
7.5　计算两个日期之间的时间差 118
7.6　对一周内的各天进行编码 119
7.7　创建一个滞后的特征 120
7.8　使用滚动时间窗口 121
7.9　处理时间序列中的缺失值 123

第 8 章　图像处理 127

8.0　简介 127
8.1　加载图像 128
8.2　保存图像 130
8.3　调整图像大小 131
8.4　裁剪图像 132
8.5　平滑处理图像 133
8.6　图像锐化 136
8.7　提升对比度 138
8.8　颜色分离 140
8.9　图像二值化 142
8.10　移除背景 144
8.11　边缘检测 148
8.12　角点检测 150
8.13　为机器学习创建特征 153
8.14　将颜色平均值编码成特征 156
8.15　将色彩直方图编码成特征 157

第 9 章　利用特征提取进行特征降维 161

9.0　简介 161
9.1　使用主成分进行特征降维 161

9.2 对线性不可分数据进行特征降维 164
9.3 通过最大化类间可分性进行特征降维 166
9.4 使用矩阵分解法进行特征降维 169
9.5 对稀疏数据进行特征降维 170

第 10 章 使用特征选择进行降维 173

10.0 简介 173
10.1 数值型特征方差的阈值化 173
10.2 二值特征的方差阈值化 175
10.3 处理高度相关性的特征 176
10.4 删除与分类任务不相关的特征 178
10.5 递归式特征消除 180

第 11 章 模型评估 183

11.0 简介 183
11.1 交叉验证模型 183
11.2 创建一个基准回归模型 187
11.3 创建一个基准分类模型 188
11.4 评估二元分类器 190
11.5 评估二元分类器的阈值 193
11.6 评估多元分类器 197
11.7 分类器性能的可视化 198
11.8 评估回归模型 201
11.9 评估聚类模型 203
11.10 创建自定义评估指标 204
11.11 可视化训练集规模的影响 206
11.12 生成对评估指标的报告 208
11.13 可视化超参数值的效果 209

第 12 章 模型选择 213

12.0 简介 213
12.1 使用穷举搜索选择最佳模型 213
12.2 使用随机搜索选择最佳模型 216
12.3 从多种学习算法中选择最佳模型 218

12.4 将数据预处理加入模型选择过程 220
12.5 用并行化加速模型选择 221
12.6 使用针对特定算法的方法加速模型选择 223
12.7 模型选择后的性能评估 224

第 13 章 线性回归 227

13.0 简介 227
13.1 拟合一条直线 227
13.2 处理特征之间的影响 229
13.3 拟合非线性关系 231
13.4 通过正则化减少方差 233
13.5 使用套索回归减少特征 235

第 14 章 树和森林 237

14.0 简介 237
14.1 训练决策树分类器 237
14.2 训练决策树回归模型 239
14.3 可视化决策树模型 240
14.4 训练随机森林分类器 243
14.5 训练随机森林回归模型 244
14.6 识别随机森林中的重要特征 245
14.7 选择随机森林中的重要特征 248
14.8 处理不均衡的分类 249
14.9 控制决策树的规模 250
14.10 通过 boosting 提高性能 252
14.11 使用袋外误差（Out-of-Bag Error）评估随机森林模型 253

第 15 章 KNN 255

15.0 简介 255
15.1 找到一个观察值的最近邻 255
15.2 创建一个 KNN 分类器 258
15.3 确定最佳的邻域点集的大小 260
15.4 创建一个基于半径的最近邻分类器 261

第 16 章　逻辑回归 263
16.0　简介 263
16.1　训练二元分类器 263
16.2　训练多元分类器 265
16.3　通过正则化来减小方差 266
16.4　在超大数据集上训练分类器 267
16.5　处理不均衡的分类 269
第 17 章　支持向量机 271
17.0　简介 271
17.1　训练一个线性分类器 271
17.2　使用核函数处理线性不可分的数据 274
17.3　计算预测分类的概率 278
17.4　识别支持向量 279
17.5　处理不均衡的分类 281
第 18 章　朴素贝叶斯 283
18.0　简介 283
18.1　为连续的数据训练分类器 284
18.2　为离散数据和计数数据训练分类器 286
18.3　为具有二元特征的数据训练朴素贝叶斯分类器 287
18.4　校准预测概率 288
第 19 章　聚类 291
19.0　简介 291
19.1　使用 K-Means 聚类算法 291
19.2　加速 K-Means 聚类 294
19.3　使用 Meanshift 聚类算法 295
19.4　使用 DBSCAN 聚类算法 296
19.5　使用层次合并聚类算法 298
第 20 章　神经网络 301
20.0　简介 301
20.1　为神经网络预处理数据 302

20.2 设计一个神经网络 304
20.3 训练一个二元分类器 307
20.4 训练一个多元分类器 309
20.5 训练一个回归模型 311
20.6 做预测 313
20.7 可视化训练历史 315
20.8 通过权重调节减少过拟合 318
20.9 通过提前结束减少过拟合 320
20.10 通过 Dropout 减少过拟合 322
20.11 保存模型训练过程 324
20.12 使用 k 折交叉验证评估神经网络 326
20.13 调校神经网络 328
20.14 可视化神经网络 331
20.15 图像分类 333
20.16 通过图像增强来改善卷积神经网络的性能 337
20.17 文本分类 339

第 21 章 保存和加载训练后的模型 343

21.0 简介 343
21.1 保存和加载 scikit-learn 模型 343
21.2 保存和加载 Keras 模型 345

第 1 章

向量、矩阵和数组

1.0 简介

NumPy 是 Python 机器学习技术栈的基础。NumPy 能对机器学习中常用的数据结构——向量（vector）、矩阵（matrice）、张量（tensor）——进行高效的操作。NumPy 并不是本书的重点，不过它在后面的章节中也会频繁出现。本章将介绍在进行机器学习的过程中可能经常遇到的 NumPy 操作。

1.1 创建一个向量

问题描述

创建一个向量。

解决方案

使用 NumPy 创建一个一维数组：

```
# 加载库
import numPy as np

# 创建一个行向量
vector_row = np.array([1, 2, 3])

# 创建一个列向量
vector_column = np.array([[1],
                          [2],
                          [3]])
```


试读结束：需要全本请在线购买： www.ertongbook.com