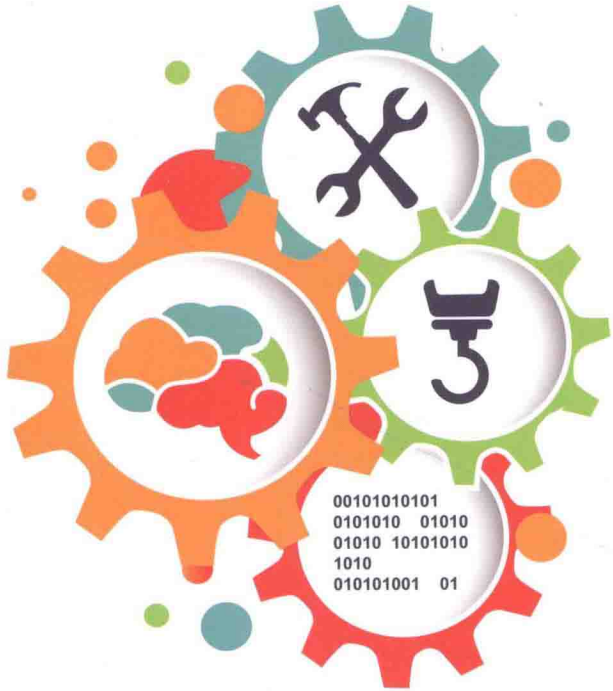
Broadview®
www.broadview.com.cn



机器学习之路

Caffe、Keras、scikit-learn实战

阿布 胥嘉幸 编著

绕过理论障碍，理解机器学习，
打通一条由浅入深的机器学习之路。
丰富的实战案例讲解，
介绍如何将机器学习技术运用到
股票量化交易、图片渲染等领域。

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

机器学习之路

Caffe、Keras、scikit-learn实战

阿布 胥嘉幸 编著

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

内 容 简 介

机器学习需要一条脱离过高理论门槛的入门之路。

本书《机器学习篇》从小红帽采蘑菇的故事开篇，介绍了基础的机器学习分类模型的训练（第 1 章）。如何评估、调试模型？如何合理地发掘事物的特征？如何利用几个模型共同发挥作用？后续章节一步一步讲述了如何优化模型，更好地完成分类预测任务（第 2 章），并且初步尝试将这些技术运用到金融股票交易中（第 3 章）。

自然界最好的非线性模型莫过于人类的大脑。《深度学习篇》从介绍并对比一些常见的深度学习框架开始（第 4 章），讲解了 DNN 模型的直观原理，尝试给出一些简单的生物学解释，完成简单的图片识别任务（第 5 章）。后续章节在此基础上，完成更为复杂的图片识别 CNN 模型（第 6 章）。接着，本书展示了使用 Caffe 完成一个完整的图片识别项目，从准备数据集，到完成识别任务（第 7 章）。后面简单描述了 RNN 模型（第 8 章），接着展示了一个将深度学习技术落地到图片处理领域的项目（第 9 章）。

本书适合能看懂 Python 代码，对机器学习感兴趣，期望入门的读者。

图书在版编目（CIP）数据

机器学习之路：Caffe、Keras、scikit-learn 实战 / 阿布，胥嘉幸编著. —北京：电子工业出版社，2017.8
ISBN 978-7-121-32160-3

Ⅰ. ①机… Ⅱ. ①阿… ②胥… Ⅲ. ①机器学习 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字(2017)第 165543 号

责任编辑：安　娜
印　　刷：三河市良远印务有限公司
装　　订：三河市良远印务有限公司
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：20.5　字数：405 千字
版　　次：2017 年 8 月第 1 版
印　　次：2017 年 8 月第 1 次印刷
定　　价：79.00 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

越来越多的人期待能挤进机器学习这一行业，这些人往往有一些编程和自学能力，但数学等基础理论能力不足。对于这些人群，从头开始学习概率统计等基础学科是痛苦的，如果直接上手使用机器学习工具往往又感到理解不足，缺少点什么。本书就是面向这一人群，避过数学推导等复杂的理论推衍，介绍模型背后的一些简单直观的理解，以及如何上手使用。本书希望能够得到这些人的喜爱。

本书包含两部分：机器学习篇和深度学习篇。

机器学习篇（1～3 章）主要从零开始，介绍什么是数据特征，什么是机器学习模型，如何训练模型、调试模型，以及如何评估模型的成绩。通过一些简单的任务例子，讲解在使用模型时如何分析并处理任务数据的特征，如何组合多个模型共同完成任务，并在第 3 章初步尝试将机器学习技术运用到股票交易中，重复熟悉这些技术的同时，感受机器学习技术在落地到专业领域时常犯的错误。

深度学习篇（4～9 章）则主要介绍了一些很基础的深度学习模型，如 DNN、CNN 等，简单涵盖了一些 RNN 的概念描述。我们更关注模型的直观原理和背后的生物学设计理念，希望读者能够带着这些理解，直接上手应用深度学习框架。

说一点关于阅读本书的建议。本书在编写时不关注模型技术的数学推导及严谨表述，转而关注其背后的直观原理理解。建议读者以互动执行代码的方式学习，所有示例使用 IPython Notebook 编写。读者可在 Git 上找到对应章节的内容，一步一步运行书中讲解的知识点，直观感受每一步的执行效果。具体代码下载地址：https://github.com/bbfamily/abu。

本书适合有 Python 编程能力的读者。如果读者有简单的数学基础，了解概率、矩阵则更佳。使用过 Numpy、pandas 等数据处理工具的读者读起来也会更轻松，但这些都不是必需的。如果读者缺乏 Python 编程能力，或者希望进一步获得 Numpy、pandas 等工具

使用相关的知识，可以关注公众号：abu_quant，获得一些技术资料及文章。

感谢出版社提供机会让我们编写本书，感谢编辑不辞辛苦地和我沟通排版等细节问题。

本书的完成同样需要感谢我们的几位朋友：吴汶（老虎美股）、刘兆丹（百度金融），感谢你们在本书编写作过程中提供的有力支持。感谢本书的试读人员：蔡志威、李寅龙。

目录

第一篇　机器学习篇

第 1 章　初识机器学习 2

1.1　机器学习——赋予机器“学习”的灵魂 2

1.1.1　小红帽识别毒蘑菇 2

1.1.2　三种机器学习问题 6

1.1.3　常用符号 6

1.1.4　回顾 7

1.2　KNN——相似的邻居请投票 7

1.2.1　模型原理 7

1.2.2　鸢尾花卉数据集（IRIS） 9

1.2.3　训练模型 9

1.2.4　评估模型 12

1.2.5　关于 KNN 14

1.2.6　运用 KNN 模型 15

1.2.7　回顾 16

1.3　逻辑分类 I：线性分类模型 16

1.3.1　参数化的模型 16

1.3.2　逻辑分类：预测 18

1.3.3　逻辑分类：评估 22

1.3.4　逻辑分类：训练 23

1.3.5　回顾 24

1.4　逻辑分类 II：线性分类模型 24

1.4.1　寻找模型的权重 24

1.4.2 去均值和归一化 31
1.4.3 实现 33
1.4.4 回顾 34

第 2 章 机器学习进阶 35

2.1 特征工程 35
2.1.1 泰坦尼克号生存预测 35
2.1.2 两类特征 38
2.1.3 构造非线性特征 41
2.1.4 回顾 45
2.2 调试模型 46
2.2.1 模型调试的目标 46
2.2.2 调试模型 49
2.2.3 回顾 52
2.3 分类模型评估指标 53
2.3.1 混淆矩阵系指标 53
2.3.2 评估曲线 58
2.3.3 回顾 61
2.4 回归模型 61
2.4.1 回归与分类 61
2.4.2 线性回归 62
2.4.3 波士顿房价预测 66
2.4.4 泰坦尼克号生存预测：回归预测特征年龄 Age 69
2.4.5 线性模型与非线性模型 72
2.4.6 回顾 73
2.5 决策树模型 73
2.5.1 信息与编码 74
2.5.2 决策树 76
2.5.3 对比线性模型和决策树模型的表现 77
2.5.4 回顾 79
2.6 模型融合 80
2.6.1 融合成群体（Ensamble） 80
2.6.2 Bagging：随机森林（Random Forest） 82

2.6.3 Boosting：GBDT 83
2.6.4 Stacking 86
2.6.5 泰坦尼克号生存预测：小结 93
2.6.6 回顾 94

第 3 章 实战：股票量化 95

3.1 第一步：构造童话世界 95
3.1.1 股票是什么 95
3.1.2 当机器学习与量化交易走在一起 96
3.1.3 构造一个童话世界 96
3.1.4 回顾 100
3.2 第二步：应用机器学习 100
3.2.1 构建特征数据 100
3.2.2 回归预测股票价格 103
3.2.3 分类预测股票涨跌 108
3.2.4 通过决策树分类，绘制决策图 112
3.2.5 回顾 114
3.3 第三步：在真实世界应用机器学习 114
3.3.1 回测 115
3.3.2 基于特征的交易预测 119
3.3.3 破灭的童话——真实世界的机器学习 122

第二篇 深度学习篇

第 4 章 深度学习：背景和工具 126

4.1 背景 126
4.1.1 人工智能——为机器赋予人的智能 126
4.1.2 图灵测试 126
4.1.3 强人工智能 vs 弱人工智能 127
4.1.4 机器学习和深度学习 128
4.1.5 过度的幻想 128
4.1.6 回顾 129

4.2 深度学习框架简介 129
4.2.1 评测方式 130
4.2.2 评测对象 131
4.2.3 深度学习框架评测 131
4.2.4 小结 135
4.3 深度学习框架快速上手 135
4.3.1 符号主义 135
4.3.2 MNIST 136
4.3.3 Keras 完成逻辑分类 138
4.3.4 回顾 141
4.4 Caffe 实现逻辑分类模型 141
4.4.1 Caffe 训练 MNIST 概览 142
4.4.2 Caffe 简介 144
4.4.3 准备数据集 145
4.4.4 准备模型 146
4.4.5 模型训练流程 149
4.4.6 使用模型 149
4.4.7 Caffe 的 Python 接口 150
4.4.8 回顾 151

第 5 章 深层学习模型 152

5.1 解密生物智能 154
5.1.1 实验一：大脑的材料 154
5.1.2 实验二：探索脑皮层的功能区域 156
5.1.3 实验三：不同的皮层组织——区别在于函数算法 158
5.1.4 实验四：可替换的皮层模块——神经元组成的学习模型 161
5.1.5 模拟神经元 162
5.1.6 生物结构带来的启发 163
5.1.7 回顾 164
5.2 DNN 神经网络模型 164
5.2.1 线性内核和非线性激活 164
5.2.2 DNN、CNN、RNN 165
5.2.3 逻辑分类：一层神经网络 166

5.2.4 更多的神经元 167
5.2.5 增加 Hidden Layer（隐层） 168
5.2.6 ReLu 激活函数 170
5.2.7 理解隐层 171
5.2.8 回顾 172
5.3 神经元的深层网络结构 172
5.3.1 问题：更宽 or 更深 172
5.3.2 链式法则：深层模型训练更快 173
5.3.3 生物：深层模型匹配生物的层级识别模式 175
5.3.4 深层网络结构 177
5.3.5 回顾 178
5.4 典型的 DNN 深层网络模型：MLP 178
5.4.1 优化梯度下降 179
5.4.2 处理过拟合：Dropout 181
5.4.3 MLP 模型 182
5.4.4 回顾 185
5.5 Caffe 实现 MLP 185
5.5.1 搭建 MLP 185
5.5.2 训练模型 189
5.5.3 回顾 190

第 6 章 学习空间特征 191

6.1 预处理空间数据 192
6.1.1 像素排列展开的特征向量带来的问题 192
6.1.2 过滤冗余 194
6.1.3 生成数据 195
6.1.4 回顾 198
6.2 描述图片的空间特征：特征图 199
6.2.1 图片的卷积运算 199
6.2.2 卷积指令和特征图 201
6.2.3 回顾 206
6.3 CNN 模型 I：卷积神经网络原理 206
6.3.1 卷积神经元 207

6.3.2 卷积层 208
6.3.3 多层卷积 211
6.3.4 回顾 216
6.4 CNN 模型 II：图片识别 216
6.4.1 连接分类模型 216
6.4.2 猫狗分类 217
6.4.3 反思 CNN 与 DNN 的结合：融合训练 221
6.4.4 深度学习与生物视觉 222
6.4.5 回顾 224
6.5 CNN 的实现模型 224
6.5.1 ImageNet 简介 224
6.5.2 Googlenet 模型和 Inception 结构 226
6.5.3 VGG 模型 228
6.5.4 其他模型 231
6.5.5 回顾 232
6.6 微训练模型（fine-tuning） 232
6.6.1 二次训练一个成熟的模型 232
6.6.2 微训练在 ImageNet 训练好的模型 233
6.6.3 回顾 239

第 7 章 Caffe 实例：狗狗品种辨别 240

7.1 准备图片数据 240
7.1.1 搜集狗狗图片 240
7.1.2 清洗数据 241
7.1.3 标准化数据 242
7.1.4 回顾 243
7.2 训练模型 243
7.2.1 生成样本集 244
7.2.2 生成训练、测试数据集 245
7.2.3 生成 lmdb 246
7.2.4 生成去均值文件 247
7.2.5 更改 prototxt 文件 247
7.2.6 训练模型 249

7.2.7 回顾 249
7.3 使用生成的模型进行分类 249
7.3.1 更改 deploy.prototxt 249
7.3.2 加载模型 250
7.3.3 回顾 257

第 8 章 漫谈时间序列模型 258

8.1 Embedding 259
8.1.1 简单的文本识别 260
8.1.2 深度学习从读懂词义开始 261
8.1.3 游戏：词义运算 264
8.1.4 回顾 264
8.2 输出序列的模型 265
8.2.1 RNN 265
8.2.2 LSTM 266
8.2.3 并用人工特征和深度学习特征——一个 NLP 模型的优化历程 268
8.2.4 反思：让模型拥有不同的能力 270
8.2.5 回顾 273
8.3 深度学习：原理篇总结 273
8.3.1 原理小结 273
8.3.2 使用建议 275

第 9 章 用深度学习做个艺术画家——模仿实现 PRISMA 277

9.1 机器学习初探艺术作画 278
9.1.1 艺术作画概念基础 278
9.1.2 直观感受一下机器艺术家 279
9.1.3 一个有意思的实验 280
9.1.4 机器艺术作画的愿景 281
9.1.5 回顾 282
9.2 实现秒级艺术作画 282
9.2.1 主要实现思路分解讲解 283
9.2.2 使用统计参数期望与标准差寻找 mask 290

9.2.3 工程代码封装结构及使用示例 299
9.2.4 回顾和后记 302

附录 A 机器学习环境部署 303

附录 B 深度学习环境部署 307

附录 C 随书代码运行环境部署 312

第一篇

机器学习篇

机器学习篇主要面向机器学习零基础的读者，已有相关知识的读者可以直接跳过这一篇。

第 1 章

初识机器学习

本章将介绍几个浅层的学习模型，并尝试解释这些模型背后的“直观”原理。通过对本章的学习，相信你将有能力将这些模型运用到自己的工程中。

1.1 机器学习——赋予机器“学习”的灵魂

当人类用感情和希望去创造一样东西，那一样东西就会被赋予灵魂。

——宫崎骏《猫的报恩》

本节将对比机器学习和人类学习的过程。

400 多年前，我们发明了望远镜，拓展了“视觉”的能力；320 年前，我们发明了代步工具——自行车，提升了“步行”的能力；100 多年前，从热气球到莱特兄弟的飞机，我们具备了新的能力——“飞行”。在信息技术日益成熟的今天，机器学习将带我们步入更加神奇的世界——扩展“学习”的能力。

那么如何让机器像智能生物一样，获得学习的能力？早些时候的科学家一直试图让机械拥有真正意义上的智能，进而产生智能行为——学习，但最终这个方向并没有走通。今天的科学家走出了思维的桎梏，放弃纯粹的生物模仿，而是利用科学理论完成仿生——用数学模拟智能生物学习的过程。莱特兄弟发明的飞机是空气动力学的产物，而今天的机器学习是以数据为中心，通过训练模型发现数据中的某些内在模式，并将其运用到新数据中的技术。简单来说，机器学习就是研究数据，模拟生物学习的能力。

让我们先看一个简单的故事，直观对比一下机器学习和人类学习过程中的差异。

1.1.1 小红帽识别毒蘑菇

小红帽去森林采蘑菇，她希望自己能够了解哪些蘑菇是有毒的。第一天，采集了 5 朵蘑菇，小红帽观察蘑菇外表，发现其中 2 朵外表鲜艳的是毒蘑菇，3 朵外表朴素的是正

常蘑菇；于是小红帽学到一个新知识："外表鲜艳的蘑菇是有毒的！"如图 1-1 所示。

图 1-1　毒蘑菇

小红帽通过眼睛观察蘑菇，机器则通过输入的数值识别事物。让机器模拟小红帽的这段学习经验，从用数据描述样本的信息开始：小红帽采集的 5 朵蘑菇，按特征（feature）提取数值，可以描述为

$$X:\begin{matrix}\text{外表是否鲜艳}\\ \begin{bmatrix}1\\0\\1\\0\\0\end{bmatrix}\end{matrix}$$

注意，这里的数据行是不同的样本，数据列是"外表是否鲜艳"这一特征。其中编码信息：1-鲜艳，0-不鲜艳。得到的是否毒蘑菇的结果为

$$y:\begin{bmatrix}1\\0\\1\\0\\0\end{bmatrix}$$

其中，编码蘑菇的类别：1-毒蘑菇，0-正常蘑菇。y 就叫作 X 的标签（label）向量，而 5 朵蘑菇样本数据 X 叫作训练集（training set）。机器从训练集学习到识别类别的知识，这一过程叫作模型（model）的训练，即：

机器学到的知识 == 训练好的模型

第二天，小红帽为了检验她的新知识，又采集了 5 朵新蘑菇做实验，结果发现误判了 2 朵蘑菇类别，于是小红帽得出结论：之前总结的知识识别蘑菇的准确率只有 60%。

对于学习到的模型，很自然地我们希望考察它在未知样本数据上面的应用能力，所

以接下来拿一部分和训练集不同的新样本，让模型预测。新的检验知识的蘑菇集合叫作测试集。

$$\text{外表是否鲜艳}$$

$$\textbf{(test-set)}\ X'\text{: } \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \text{对应的测试结果 } y'\text{: } \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}。$$

小红帽通过“准确率”衡量模型的表现：

$$\text{准确率}=\frac{\text{正确分类的数量}}{\text{总数}}$$

小红帽反思了自己，仅仅通过“外表是否鲜艳”这一特征来辨别毒蘑菇是不可靠的。于是，小红帽学着从更多角度判读蘑菇是否有毒。例如，蘑菇的外表、生长地、尺寸。有的特征对辨识毒蘑菇很有帮助，有的则用处不大。慢慢地，小红帽学会了同时权衡这些特征，辨识一朵蘑菇是否有毒。毒蘑菇的生长环境如图 1-2 所示。

图 1-2　毒蘑菇的生长环境

从机器学习的角度来看：单个特征的模型太过简单，于是，提取了更多的蘑菇特征。

$$X = \begin{bmatrix} \text{外表是否黏滑} & \text{生长地是否潮湿} & \text{大小尺寸} & \cdots \\ 0 & 1 & 3.3 & \cdots \\ 1 & 1 & 7.3 & \cdots \\ 0 & 0 & 1.3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}。$$

接着，我们需要某种方式将这些特征组合起来，让它们一起发挥作用，帮助小红帽识别毒蘑菇。机器学习中的模型组合特征的方式有两类：

- 非参数化（如 1.2 节将要登场的 KNN）
- 参数化