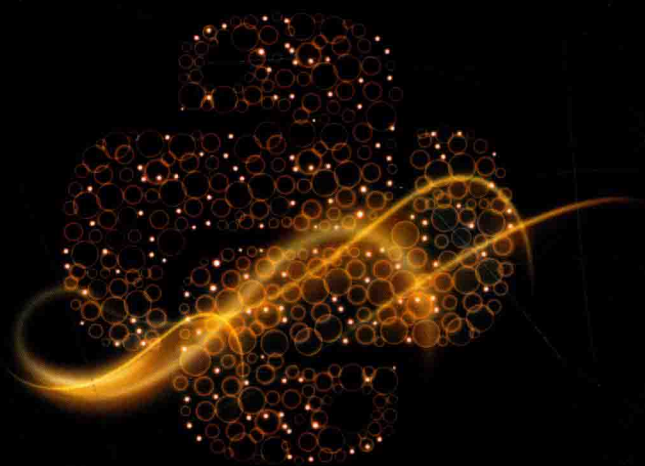
博文金融

金融科技
丛书

Python
机器学习与量化投资

何海群 著



Win Or Out

本书采用黑箱模式和案例教学模式，并结合大量经典案例，
介绍sklearn机器学习模块在金融领域的应用

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

Python
机器学习与量化投资

何海群 著

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

内 容 简 介

本书采用生动活泼的语言，从入门者的角度，讲解了 Python 语言和 sklearn 模块库内置的各种经典机器学习算法；介绍了股市外汇、比特币等实盘交易数据在金融量化方面的具体分析与应用，包括对未来股票价格的预测、大盘指数趋势分析等。简单风趣的实际案例让广大读者能够快速掌握机器学习在量化分析方面的编程，为进一步学习金融科技奠定扎实的基础。

图书在版编目（CIP）数据

Python 机器学习与量化投资 / 何海群著. —北京：电子工业出版社，2018.12
（金融科技丛书）
ISBN 978-7-121-35210-2

Ⅰ. ①P… Ⅱ. ①何… 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（2018）第 238872 号

策划编辑：黄爱萍
责任编辑：张彦红
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
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：18.75　字数：
版　　次：2018 年 12 月第 1 版
印　　次：2018 年 12 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

本书特色

本书全程采用黑箱模式和 MBA 案例模式，结合大量经典案例，介绍 sklearn 机器学习模块库和常用的机器学习算法，懂 Excel 就能看懂本书；逆向式课件模式，结合大量案例、图表，层层剖析；三位一体的课件模式：图书+开发平台+成套的教学案例，系统讲解、逐步深入。

本书是《零起点 Python 机器学习快速入门》的后续之作，为了节省篇幅，省略了 Python 基础教程，以及 sklearn 等机器学习方面的入门内容，没有经验的读者，建议先阅读《零起点 Python 机器学习快速入门》，再阅读本书，这样会收到事半功倍的效果。

本书简单实用，书中配备大量的图表说明，本书特点如下。

- IT 零起点：无须任何电脑编程基础，只要会打字、会使用 Excel，就能看懂本书，利用本书配套的 Python 软件包，轻松学会如何利用 Python 对股票数据进行专业分析和量化投资分析。
- 投资零起点：无须购买任何专业软件，本书配套的 zwPython 软件包，采用开源模式，提供 100%全功能、全免费的工业级数据分析平台。

- 配置零起点：所有软件、数据全部采用“开箱即用”模式，绿色版本，无须安装，解压缩后即可直接运行系统。
- 理财零起点：采用通俗易懂的语言，配合大量专业的图表和实盘操作案例，无须任何专业金融背景，轻松掌握各种量化投资策略。
- 数学零起点：全书没有任何复杂的数学公式，只有最基本的加、减、乘、除，轻轻松松就能看懂全书。

网络资源

本书的案例程序，已经做过优化处理，无须 GPU 显卡，全部支持单 CPU 平台，不过为避免版本冲突，请尽量使用 zwPython2017m6 版本运行本书的案例程序。

使用其他运行环境的读者，如 Linux、Mac 平台的用户，请尽量使用 Python 3 版本，自行安装其他所需的模块库，如 Numpy、Pandas、Tushare 等第三方模块库。

此外需要注意的是，大家运行本书案例得到的结果可能与书中略有差别；甚至多次运行同一案例，结果都有所差异。这属于正常情况，因为很多机器学习函数，内部使用了随机数作为种子数，用于系统变量初始化等操作，每次分析的起点或者中间参数会有所不同。

版本冲突是开源项目常见的问题，为了解决这个问题，本书的源码是独立保存的。此外，我们还特意设计了 zwPython 教学版。

建议初学者先使用 zwPython 教学版，有关的课件程序，已经经过版本兼容测试，并且集成了 zwDat 金融数据集。

本书的读者 QQ 互动群：QQ 1 群的群号是 124134140；QQ 2 群的群号是 650924099；QQ 3 群的群号是 450853713。

资源下载地址：TopQuant 极宽量化网站“资源中心”。

请浏览以下网站，获取最新的网络资源地址：

- TopQuant.vip 极宽量化社区
- ziwang.com 字王网站

目前两个网站的指向都是一样的。

另外还可以在博文视点网站下载：http://www.broadview.com.cn。

目录设置

为运行本书课件程序，用户需要下载以下三个软件，并设置好目录：

- zwPython，必须放在根目录，是 Python 开发平台，为避免版本冲突，请尽量使用 zwPython2017m6 版本。
- kb_demo，本书 sklearn 机器学习配套课件源码。
- pg_demo，本书 Python 入门学习配套课件源码。

以上软件、程序最好保存在固态硬盘，这样速度会快很多；目录名称不要使用中文名称，压缩文件当中的中文名称只是为了便于用户下载。

zwPython 开发平台必须放在根目录，课件程序可以放在其他自定义目录，建议放在 zwPython 目录下面，作为二级目录。

致谢

特别感谢电子工业出版社的黄爱萍和陈林编辑在选题策划和稿件整理方面所做的大量工作。

同时，在本书创作过程中，极宽开源量化团队和培训班的全体成员提出很多宝贵的意见，并对部分课件程序做了中文注解。

特别是吴娜、余勤、邢梦来、孙励、王硕几位成员，为 TOP 极宽开源量化文库和开源软件编写文档，以及在团队成员管理方面做了大量工作，对他们的付出表示感谢。

何海群（字王）

TOP 极宽量化开源组发起人

2018 年 10 月 1 日

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/35210*

目　录

第 1 章

Python 与机器学习

目前神经网络、深度学习大热，谷歌、脸书、微软、IBM、亚马逊等企业巨头，纷纷投入巨资，各种深度学习的开发平台层出不穷：TensorFlow、PyTorch、CNTK、MXNet、Keras 等。

与此同时，Python 语言成为人工智能第一开发语言。

在传统的机器学习领域，或者说古典人工智能领域却波澜不惊，scikit-learn 始终居于王者般的统治地位。

图 1.1 是 scikit-learn 网站首页截图，在网站首页抬头右侧，写着：

Machine Learning in Python（Python 中的机器学习）

换句话说：

scikit-learn=Python 机器学习

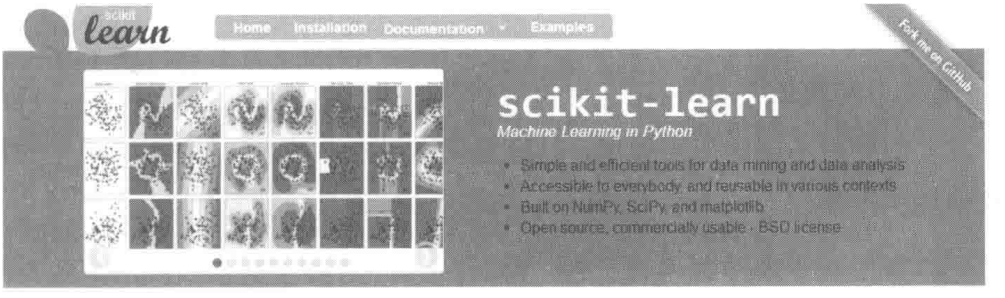


图 1.1　scikit-learn 网站首页截图

1.1 scikit-learn模块库

scikit-learn，简称 sklearn，是用 Python 开发的机器学习模块库，其中包含大量机器学习算法、数据集。

scikit-learn 模块库优点很多：简单易用、API 接口完善、案例文档丰富等。内置大量经过筛选的、高质量的机器学习模型；模块库覆盖了大多数机器学习任务；系统可扩展至较大的数据规模。

在 scikit-learn.org 网站首页抬头，scikit-learn 官方自我总结的优点如下。

- 简单高效的数据挖掘和数据分析工具。
- 可在各种环境中重复使用。
- 建立在 NumPy、SciPy 和 Matplotlib 上。
- 开放源码，可免费商业使用 BSD license。

scikit-learn 模块库是老牌的开源 Python 机器学习算法框架，源于 2007 年谷歌公司的 Google Summer of Code 项目，最早由数据科学家 David Cournapeau 发起，是 Python 语言中专门针对机器学习应用而发展起来的一款开源框架。

scikit-learn 模块库是一个简洁、高效的算法库，提供一系列的监督学习和无监督学习的算法，以用于数据挖掘和数据分析。scikit-learn 几乎覆盖了机器学习的所有主流算法，这为其在 Python 开源世界中奠定了江湖地位。

scikit-learn 的算法库是建立在 SciPy（Scientific Python）之上的，这也是其命名的由来。

SciPy 模块库是 Python 语言的基础科学计算工具包，基于 SciPy，目前开发者们针对不同的应用领域，已经发展出了众多的分支版本，SciPy 的扩展和模块在传统上被命名为 Scikits，即 SciPy 工具包的意思。

和其他众多的开源项目一样，scikit-learn 目前主要由社区成员自发维护。可能是由于维护成本的限制，scikit-learn 相比其他项目要显得更为保守。

这种保守主要体现在两个方面。

- scikit-learn 从来不做除机器学习领域之外的其他扩展。
- scikit-learn 从来不采用未经广泛验证的算法。

1.1.1　scikit-learn 的缺点

虽然 scikit-learn 模块库功能强大，目前已经是机器学习最重要的模块库，但是，scikit-learn 也有缺点。

scikit-learn 模块库的缺点主要包括以下几个方面。

- 不支持深度学习和强化学习。
- 不支持 PyPy 加速，也不支持 GPU 加速。
- 不支持除 Python 之外的其他编程语言。

这些缺点主要是由历史原因造成的，scikit-learn 毕竟是 2007 年的作品，已经有超过十年的历史，其整体架构不适用于目前的 GPU 编程、神经网络和深度学习。不过这些缺点都不是大问题。

在深度学习、神经网络领域，目前有众多的优秀平台：TensorFlow、MXNet、CNTK、PyTorch 等，scikit-learn 模块库与这些平台配合，主要用于数据预处理和结果验证。

GPU 加速和 PyPy 优化属于 Python 底层优化，特别是 NumPy 基础科学计算库的优化，目前基于 GPU 的优化版本已经不断涌现，其中 MXNet、PyTorch 底层模块库，都是基于 NumPy 的 GPU 优化版本。

从工程角度而言，scikit-learn 的性能表现是非常不错的。

究其原因，一方面是因为其内部算法的实现十分高效，另一方面或许可以归功于 Cython 编译器：通过 Cython 在 scikit-learn 框架内部生成 C 语言代码的运行方式，scikit-learn 消除了大部分的性能瓶颈。

至于最后一个缺点，scikit-learn 模块库不支持除 Python 之外的其他编

程语言。这更加不是问题，目前 Python 已经是人工智能、机器学习领域的标准编程语言，强如 Facebook，坚持多年，最终还是把 Lua 语言开发的 Torch 项目，全部采用 Python 改写，并重新命名为 PyTorch 项目。

1.1.2 scikit-learn 算法模块

为了方便学习，scikit-learn 开发团队还提供了一个 scikit-learn 算法模块图，如图 1.2 所示。

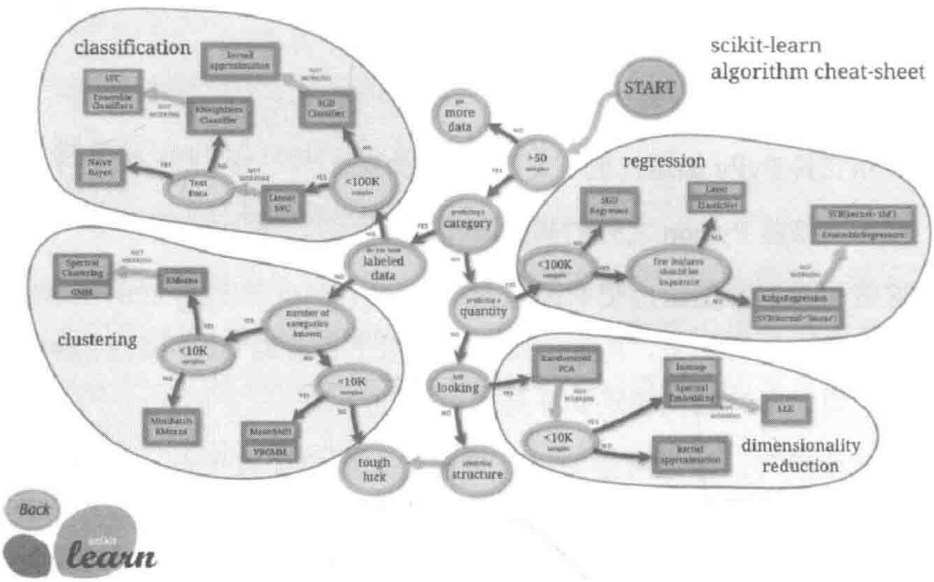


图 1.2 scikit-learn 算法模块图

图 1.3 是 scikit-learn 算法模块图的汉化版本。

scikit-learn 模块库实现了一整套用于数据降维、模型选择、特征提取和归一化的完整算法和模块，并且针对每个算法和模块，底层都进行了高度优化以提升速度，与此同时，模块库提供了丰富的参考案例和详细的说明文档。

这些案例程序，内容覆盖全面，讲解细致，并且很多案例都使用了真实的数据，绝大多数案例还配有 Matplotlib 绘制的数据图表。