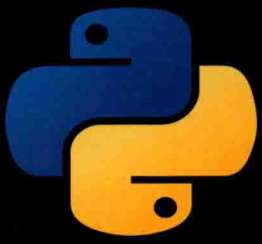
Top Quant
CHRD前海智库



零起点 Python机器学习 快速入门

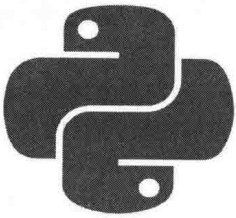
何海群 著

Win Or Home

读完本书内容和配套的教学代码
就能够轻松编写机器学习程序

无需任何编程、交易经验，也不需要具备超强的数据分析能力，
只要会使用Excel就可以轻松学会本书讲解的知识点。

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



零起点 Python机器学习 快速入门

何海群 著

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

内 容 简 介

本书采用独创的黑箱模式，MBA 案例教学机制，结合一线实战案例，介绍 Sklearn 人工智能模块库和常用的机器学习算法。书中配备大量图表说明，没有枯燥的数学公式，只要懂 Word、Excel，就能够轻松阅读全书，并学习使用书中的知识，分析大数据。本书具有以下特色：

- 独创的黑箱教学模式，全书无任何抽象理论和深奥的数学公式。
- 首次系统化融合 Sklearn 人工智能软件和 Pandas 数据分析软件，不用再直接使用复杂的 Numpy 数学矩阵模块。
- 系统化的 Sklearn 函数 API 接口中文文档，可作为案头工具书随时查阅。
- 基于 Sklearn+Pandas 架构，全程采用 MBA 案例模式，无需任何理论基础，懂 Excel 就可看懂。

图书在版编目（CIP）数据

零起点 Python 机器学习快速入门 / 何海群著. —北京：电子工业出版社，2017.5

ISBN 978-7-121-31141-3

Ⅰ. ①零… 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（2017）第 059262 号

责任编辑：黄爱萍
印　　刷：三河市兴达印务有限公司
装　　订：三河市兴达印务有限公司
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：720×1000　1/16　　印张：17.5　　字数：252 千字
版　　次：2017 年 5 月第 1 版
印　　次：2017 年 5 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 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

本书是一部意外之作，也是一部惊喜之作。

这是一本非常简单的 Python 机器学习入门教程，具有以下特色。

- 独创的黑箱教学模式，全书无任何抽象理论和深奥的数学公式。
- 首次系统化融合 Sklearn 人工智能软件和 Pandas 数据分析软件，无须使用复杂的 Numpy 数学矩阵模块。
- 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。
- 系统化的 Sklearn 函数 API 接口中文文档，可作为案头工具书随时查阅。
- 基于 Sklearn+Pandas 架构，全程采用 MBA 案例模式，无需任何理论基础，懂 Excel 就可看懂。

本书内容原本是《零起点 Python 足彩大数据与机器学习实盘分析》中的章节，在我们内部小范围使用时，深受学员喜爱，于是，将书中与机器学习相关的内容和案例单独抽取出来，加入部分 Python 入门内容，形成了《零起点 Python 机器学习快速入门》一书。

Python 量化三部曲

Python 量化三部曲包括：

- 《零起点 Python 大数据与量化交易》（入门教材）
- 《零起点 Python 量化与机器学习实盘分析》（重点分析 Sklearn）

- 《零起点 Python 量化与 TensorFlow 深度学习实盘分析》（重点分析 TensorFlow）

此外，还有两部补充作品：

- 《零起点 Python 足彩大数据与机器学习实盘分析》
- 《零起点 Python 机器学习快速入门》

Python 学习路线

机器学习、人工智能和金融量化基本原理都是相通的，本质上都是数据分析。

本书虽然属于“零起点 Python”系列，但要更好地学习本书，掌握相关的配套程序，最好具备以下基础。

- Python 编程基础，不懂 Python 语言的读者，先花一周时间学习 Python 基本知识以及 Pandas（潘达思）数据分析软件基础操作。
- Top 极宽量化社区有“Python 量化与 zwQuant 学习路线图”，大家可以参考，网址是：http://topquant.vip/forum.php?mod=viewthread&tid=6。
- 先花一周时间学习 Python 基础，再阅读《zwPython 中文手册》，可以少走很多弯路。
- 学好 Python、Pandas 基础后，先将本书通读 1~2 遍。通读时，碰到问题没关系，记录一下跳过去，然后进行精读；正式学习每章的代码时，一定要将代码运行 1~2 遍，培养编程感觉。
- 根据代码学习画流程图，有了流程图就可以把握程序逻辑，重点是程序和策略的逻辑。
- 入门后，多看看配套的课件和程序源码，全套 zwQuant 量化开源程序都带有函数一级的中文注解。
- 在学习过程当中一定要多问，可以在论坛提问，这样大家都受益。

- 本书读者 QQ 群号：124134140。在群文件中有一个小软件，即 Python 流程图绘制软件 CODINGRAPH，使用该软件，拷贝代码就可以自动绘制流程图，对其他编程语言也适用。

网络资源

与本书有关的程序和数据下载，请浏览网站：http://TopQuant.vip，极宽量化社区中的【下载中心】栏目。

本书在 TopQuant.vip 极宽量化社区设有专栏，对本书、人工智能和机器学习有任何建议都可在社区相关栏目发布信息，我们会在第一时间进行反馈和答复。

“零起点 Python”系列丛书

本书继续保持了“零起点 Python”系列丛书的一贯风格——简单实用。书中配备了大量的图表说明，没有枯燥的数学公式，只要懂 Word、Excel，就能够轻松阅读全书。

- IT 零起点：无需任何电脑编程基础，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 对股票和足彩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、数据全部采用苹果“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：无需任何专业金融背景，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，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。

致谢

本书的出版要特别感谢电子工业出版社的黄爱萍编辑，感谢她在选题策划和稿件整理方面做出的大量工作。

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，也提出过很多宝贵的意见，并对部分内容程序做了中文注解。

特别是吴娜、余勤两位同学，为极宽开源量化文库和 zwQuant 开源量化软件编写文档，并在团队成员管理方面做了大量工作，为他们的付出表示感谢。

何海群（字王）

北京极宽科技有限公司 CTO

2017 年 2 月 25 日

目　录

第 1 章

从阿尔法狗开始说起

1.1 阿尔法狗的前世今生

百度百科的“阿尔法狗”词条是：

阿尔法狗（AlphaGo）是一款围棋人工智能程序，由谷歌（Google）旗下 DeepMind 公司的戴密斯·哈萨比斯、大卫·席尔瓦、黄士杰与他们的团队开发，其主要工作原理是“深度学习”。

2016 年 3 月，该程序与围棋世界冠军、职业九段选手李世石进行人机大战，并以 4:1 的总比分获胜。2016 年年末至 2017 年年初，该程序在中国棋类网站上以“大师”（Master）为注册账号与中国、日本、韩国数十位围棋高手进行快棋对决，连续 60 局无一败绩。不少职业围棋手认为，阿尔法围棋的棋力已经达到甚至超过围棋职业九段水平，在世界职业围棋排名中，其等级分曾经超过排名人类第一的棋手柯洁。

2017 年 1 月，谷歌 Deep Mind 公司 CEO 哈萨比斯在德国慕尼黑 DLD（数字、生活、设计）创新大会上宣布推出真正 2.0 版本的 AlphaGo。其特点是摈弃了人类棋谱，只靠深度学习的方式成长起来，挑战围棋的极限。

现代社会已经进入后互联网时代，信息资源随手可得，大数据产业风生水起，科技创新层出不穷，不过类似工业革命、卫星登月、原子弹爆炸、Internet 信息高速公路等级别的重大科技突破却一直没有出现，甚至有学者认为，目前的社会处于科技停滞阶段。

直到 2016 年 3 月，谷歌公司的 AlphaGo 横空出世，与围棋世界冠军、职业九段选手李世石进行人机大战，并以 4:1 的总比分获胜。

这个结果震惊了整个社会，特别是学术界人工智能领域的从业人员。过去人们一直认为：围棋是人类智慧的最后堡垒，19×19 的围棋棋盘矩阵可以衍生出天文数字的组合变化，至少在 50 年内，由于计算机技术的限制，人工智能无法达到人类职业选手的标准。没想到，AlphaGo 轻松战胜了人类的围棋冠军。

AlphaGo 程序虽然神秘，但其核心算法却很简单，源自古老的 Monte Carlo 蒙特卡洛算法。

2006 年，欧洲数学家 Rémi Coulomb、Kocsis 和 Szepervari 等学者，在研究围棋程序时，结合蒙特卡洛算法与对手树搜索算法，设计出一种全新的算法：MCTS 蒙特卡罗树搜索算法。

MCTS 全称是 Monte Carlo Tree Search，即蒙特卡罗树搜索算法，是一种在人工智能问题中做出最优决策的方法，它结合了随机模拟的一般性和树搜索的准确性。

而古老的蒙特卡洛算法，正是源自世界上最古老的国际大赌场蒙特卡洛。事实上，现代统计学、博弈学甚至金融领域的量化投资，都不同程度地源自赌场和蒙特卡洛算法。

1.2 机器学习是什么

百度百科的“机器学习”词条是：

机器学习（Machine Learning，ML）是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构，使之不断改善自身的性能。它是人工智能的核心，也是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域，其主要使用归纳、综合而不是演绎。

Machine Learning（中文版是《计算机科学丛书：机器学习》）是机器学习的经典图书，被卡内基梅隆等许多大学作为机器学习课程的教材，书中主要涵盖了目前机器学习中各种最实用的理论和算法，包括概念学习、决策树、神经网络、贝叶斯学习、基于实例的学习、遗传算法、规则学习、基于解释的学习和增强学习等。

该书作者 Tom Mitchell 在他这本书的序言开场白中给出了一个机器学习的定义。

机器学习这门学科所关注的问题是：计算机程序如何随着经验积累自动提高性能。

Marshland 也在 *Machine Learning: An Algorithmic Perspective* 一书的序言中，从算法角度这样描述机器学习。

机器学习最有趣的特征之一就是它介于几个不同理论学科之间，主要是计算机科学、统计学、数学和工程学。机器学习经常被作为人工智能的一部分来进行研究，这把它牢牢地置于计算机科学中。若要理解为什么这些算法能够有效工作，则需要一定的统计学和数学头脑，这往往是计算机科学专业的本科生所缺少的能力。

1.3　机器学习大史记

机器学习是人工智能研究较为年轻的分支，它的发展过程大致经历了

三次高潮：

- 20 世纪 40 年代到 60 年代的萌芽期。
- 20 世纪 60 年代末期到 80 年代的摸索期。
- 20 世纪 90 年代到目前的崛起期。

1. 萌芽期

20 世纪 40 年代到 60 年代属于机器学习的萌芽期，以下是期间的一些标志性事件。

（1）最早的人工神经网络原型

1943 年，神经科学家和控制论专家 Warren McCulloch 与逻辑学家 Walter Pitts 基于数理逻辑算法创造了一种神经网络计算模型，这是最早的人工神经网络原型。

（2）Hebb 赫布学习规则

1949 年，心理学家赫布（Donald Hebb）基于神经心理学的学习机制，提出了一种学习假说，即 Hebb 赫布学习规则，开启了机器学习的第一步，

Hebb 赫布学习规则，是一种典型的无监督学习规则，与“条件反射”机理一致，并且已经得到了神经细胞学说的证实。

1948 年，研究人员将这种计算模型的思想应用到了 B 型图灵机上。

1954 年，Farley 和物理学家 Wesley A. Clark 在 MIT 首次用计算机模拟了一个赫布网络。

（3）图灵测试

图灵测试是人工智能哲学方面第一个严肃的提案，图灵测试一词源自阿兰·图灵 1950 年的论文《计算机器与智能》。

阿兰·图灵（Alan Turing）是英国数学家、逻辑学家，被视为计算机科学之父。1931 年，阿兰·图灵进入剑桥大学国王学院，毕业后到美国普

林斯顿大学攻读博士学位。

1950 年，图灵发表了一篇划时代的论文《计算机器与智能》，文中提出了著名的图灵测试：如果一台机器能够与人类展开对话（通过电传设备）而不能被辨别出其机器身份，那么称这台机器具有智能。

2014 年 6 月 8 日，一台计算机（计算机尤金·古斯特曼是一个聊天机器人，一个电脑程序）成功地让人类相信它是一个 13 岁的男孩，成为有史以来首台通过图灵测试的计算机。这被认为是人工智能发展的一个里程碑事件。

（4）第一台神经网络机

20 世纪 50 年代初期，科学家 Walter Pitts 和 Warren McCulloch 是最早的“神经网络”研究学者，他们分析了理想化的人工神经元网络，并且指出了它们进行简单逻辑运算的机制。当时 24 岁的研究生马文·闵斯基是他们的学生，1951 年他与 Dean Edmonds 一起建造了第一台神经网络机，称为 SNARC。在接下来的 50 年中，闵斯基是 AI 领域最重要的领导者和创新者之一。

（5）第一批机器人棋手

游戏 AI 一直被认为是评价 AI 进展的一种标准。

1951 年，Christopher Strachey 使用曼彻斯特大学的 Ferranti Mark 1 机器写出了一个西洋跳棋（Checkers）程序；Dietrich Prinz 则写出了一个国际象棋程序。1952 年，IBM 科学家亚瑟·塞缪尔开发了一个跳棋程序。该程序能够通过观察当前位置，并学习一个隐含的模型，从而为后续动作提供更好的指导。塞缪尔发现，伴随着该游戏程序运行时间的增加，其可以实现越来越好的后续指导，最终，该程序的棋力甚至可以挑战专业棋手。

通过这个程序，塞缪尔驳倒了普罗维登斯提出的“机器无法超越人类，且无法像人类一样写代码和学习的模式”。他创造了“机器学习”一词，并将它定义为“可以提供计算机能力而无须显式编程的研究领域”。

奇努克（Chinook）是由美国艾尔伯特大学计算机科学家 Jonathan Schaeffer 及其同事开发的一款国际跳棋人工智能程序。1994 年这个程序击败了人类国际跳棋世界冠军马里恩·廷斯利（Marion Tinsley），这是机器程序第一次在竞技游戏中获得官方世界冠军。2007 年这个程序完成了国际跳棋每一步最佳解决方案的信息数据库，成为国际跳棋领域不可能被击败的存在。

（6）逻辑理论家

1955 年，Newell 和 Simon（后来荣获诺贝尔奖）在 J. C. Shaw 的协助下开发了“逻辑理论家（Logic Theorist）”。这个程序能够证明《数学原理》中前 52 个定理中的 38 个，其中某些证明比原著更加新颖和精巧。Simon 认为他们已经“解决了神秘的心/身问题，解释了物质构成的系统如何获得心灵的性质。”（这一断言的哲学立场后来被 John Searle 称为“强人工智能”，即机器可以像人一样具有思想）。

（7）达特矛斯会议

1956 年的达特矛斯会议，是人工智能历史上一个里程碑，这次会议确定了 AI 人工智能的名称和目标，被业界认为是 AI 人工智能诞生的标志。

与会者包括 Ray Solomonoff、Oliver Selfridge、Trenchard More、Arthur Samuel、Newell 和 Simon，他们中的每一位都是未来 AI 领域的殿堂级专家。

会上纽厄尔和西蒙讨论了“逻辑理论家”，而 McCarthy 则说服与会者接受“人工智能”一词作为本领域的名称。

（8）感知机与最小二乘法

1957 年，心理学家罗森·布拉特基于神经感知科学背景提出了第二模型，非常的类似于今天的机器学习模型，它比 Hebb 赫布学习规则更加实用。

基于该模型罗森·布拉特设计出了第一个计算机神经网络——感知机（The Perceptron），它模拟了人脑的运作方式。