TURING 图灵程序设计丛书

利用机器学习开发算法交易系统

[韩] 安明浩 著　王雪珂 译

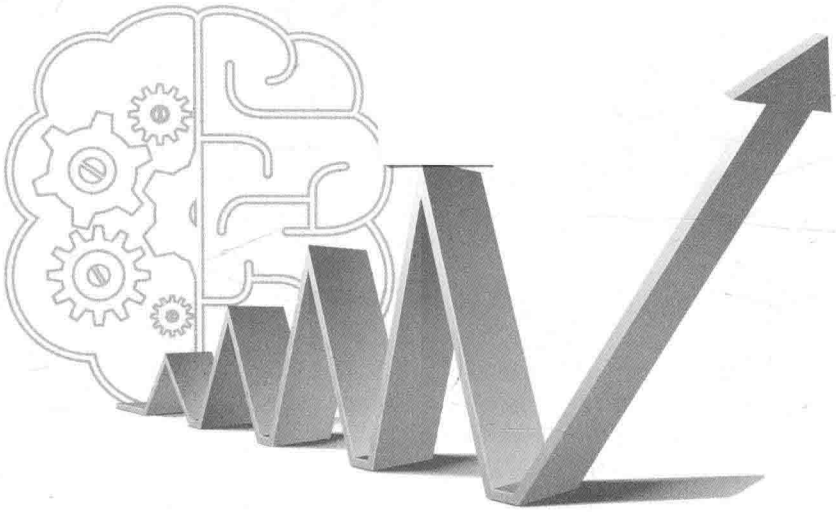
金融与机器学习的美好相遇，集中习得机器学习相关理论！
从数据爬取到系统构建实操，提供易于掌握的学习指南！

中国工信出版集团　人民邮电出版社 POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

利用机器学习开发算法交易系统

[韩] 安明浩 著　王雪珂 译



人民邮电出版社
北京

图书在版编目（CIP）数据

利用机器学习开发算法交易系统 /（韩）安明浩著；
王雪珂译. -- 北京：人民邮电出版社，2019.5
（图灵程序设计丛书）
ISBN 978-7-115-50404-3

Ⅰ. ①利… Ⅱ. ①安… ②王… Ⅲ. ①机器学习
Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2018)第287226号

内 容 提 要

本书介绍了适用机器学习的统计与概率方面的数学理论，以及其他相关领域知识，同时收录了实现代码。利用机器学习编写程序时，机器学习算法所占的比例并不大，重要的是理解数据并掌握特性。在此过程中，如果具备统计与概率相关的数学知识和机器学习应用领域的专业知识，则能大大节约时间并简化问题。经过这些过程，机器学习才能获得良好的应用效果。

本书适合机器学习入门者、具备编程能力的机器学习关注者、对股票交易原理感兴趣并乐于实践的读者。

◆ 著　　[韩] 安明浩
译　　王雪珂
责任编辑　陈　曦
责任印制　周昇亮
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
涿州市京南印刷厂印刷
◆ 开本：800×1000　1/16
印张：11.75
字数：168千字　　2019年5月第1版
印数：1-3 500册　　2019年5月河北第1次印刷
著作权合同登记号　图字：01-2018-0947号

定价：49.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字20170147号

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

前 言

范式的变化：从软件到数据

数据时代方兴未艾。人们不久前还在强调软件的重要性，“软件正在吞噬世间万物”这一多少有些挑衅性的话语也曾脍炙人口。媒体也紧跟潮流，报道了软件对我们的生活和产业造成的影响，介绍了新的软件技术，还刊登了对具有影响力的开发人员的采访。诚然，软件的重要性毋容置疑。现在如此，将来亦是如此。

我想说的是竞争力和价值。开源的普及可以称为软件发展史上的里程碑事件，因为开源是范式从软件转变为数据的核心原因。过去，软件本身就是一种竞争力。与现在不同，当时软件的数量不多，公开源代码的软件数量也很少。软件的功能和质量越好，竞争力越强。

现在，得益于开源，值得一用的软件数不胜数。从功能简单的开源软件到运行庞大IT服务时使用的复杂开源软件，覆盖多个领域。

同时，最近开源项目的水平不断提高，远超过去。比如谷歌、Facebook等跨国IT产业的龙头企业，都竞相将出于自身需要开发的、用于公司服务的高品质强性能软件进行了源代码公开。谷歌开源的TensorFlow深度学习库是进行图像识别等机器学习相关操作时使用的库，绝对不是功能弱、训练结果质量差的粗劣系统。

开源的普及使得软件无法仅凭功能和质量形成较高竞争力。想要实现机器学习，只要利用谷歌的TensorFlow即可使用全球最高水平的机器学习库。想要实时分析大数据，利用领英的Pinot，同样可以轻松开发高级实时大数据分析程序。由此，我们只需要搜索几次，就能获取优质软件。

在这种环境下想要进一步提升竞争力，只能实现“智能化”。如果说现有软件提供的是功能，那么智能化软件则通过自动化消除繁琐的操作，通过判断传递新的价值。以本书对股票的讲解为例，如果现有的股票相关软件可以提供以功能为中心的价值（“随时便捷进行股票交易”），那么智能化软件则提供以利益为中心的价值（“现在买入 XX 股票可以获得 10% 的利润”）。

智能化需要数据，利用机器学习可以将数据转换为智能。具有代表性的机器学习技术之一 ——深度学习就是很好的例子。深度学习以 1943 年发布的神经网络技术为基础，2000 ~ 2005 年还因为性能低下和技术上的制约而备受冷遇，但是在 2006 ~ 2010 年华丽复活。虽然算法的发展在此过程中起到了重要的作用，但如果没有计算能力，尤其是没有数据的话，这是不可能实现的。

机器学习三大要素是算法、计算能力和数据。算法通过网络就能轻松实现，计算能力亦是如此，只要有预算，就能使用亚马逊等公有云轻松搞定。

而数据与前两种要素不同。对于以神经网络为首的机器学习算法，拥有的数据越多，性能越好，所以拥有越多优质数据就越有利。虽然网络使获取数据变得更简单，但是因为数据所有权的问题，通过网络能够获得的数据是有限的。大多数情况下，个人信息或信用卡使用明细等拥有重要价值的信息是不会公开的。

最终，智能化的质量取决于谁拥有更多值得使用的优质数据。向消费者呈现他们想要购买的商品的打折广告，比不停发送消费者不关心的产品广告更受青睐。今后的竞争只能以利益为中心展开，而不再以功能为中心，其中的核心必然是数据。

金融、IT、机器学习

金融是关于数字的产业，IT 技术在其中发挥了非常重要的作用。存取款、查询账户

余额、给他人转账时，都需要运用IT技术。并不是我说“要转账”，实际存在的钱就会被转交给收款人。在对方取钱之前，记录在IT系统某处数据库中的“我的存款”会减少转走的金额，而收款人数据库中记录的存款数则会增加相应金额。

如果没有IT技术，我们现在享受的多个金融服务要么会消耗更多时间，要么根本不可能实现。此前，IT系统在金融领域多用于记录、删除、修改交易相关的数据。但是最近，业界开始用更高标准要求IT系统。

可以简单地称其为“利用机器学习的智能化”。与数据库只能使用金融机构拥有的庞大数据不同，机器学习用于从数据中监测风险程度、预测未来的股价走势、寻找新的商业机会等，是完全不同的角度和技术。

机器学习需要许多数据和强大的运算能力作为支撑，金融界已经满足这些条件。金融数据是数字数据，比其他数据更准确，非常适合进行机器学习。例如，分析社交媒体时，需要分析人们撰写的文章，将其转换为恰当的数字，然后利用机器学习算法进行分析或预测，在此过程中，必然会发生一定程度的损失和偏差。况且，我们很难通过社交媒体分析明确了解文章内容是100%表达了作者的想法，还是因为其他原因而未能做到这一点。

世界级投资银行高盛集团很好地展现了这种潮流。Business Insider在2015年4月12日的报道中，以“Goldman Sachs is a tech company”为题，详细介绍了高盛集团的转型。

报道称，高盛集团在全球拥有33 000名员工，其中有9000名是工程师和程序员。也就是说，将近30%的员工从事与IT相关的工作。Facebook的IT相关职员为9199人，Twitter有3638名员工，领英有6897人。由此可以看出，高盛集团IT领域的员工数是多么庞大。

在金融界，具有代表性的机器学习应用就是算法交易。算法交易不依赖于人，而是

靠程序完成操作的。早在 2012 年，算法交易就占据了美国全部交易的 85%。毫不夸张地说，我们经常在股票市场提到的“外国人”其实不是“人”，而是机构或者对冲基金使用的算法交易程序。

随着 IT 在金融领域的广泛应用，当今在全球金融圣地华尔街上，除了 MBA、财务、经济专业等传统意义上的金融人士以外，也很容易找到看上去十分“突兀”的数学家、天文学家、物理学家和计算机科学家。使用机器学习等 IT 技术不仅能够增加金融机构的收益，还能找到新的商机，使效率最大化。这一点很早之前就已经得到证明，因此，金融与机器学习等 IT 技术的结合越来越深。

成书原因

我初次接触机器学习的时候，它奇妙的概念和令人惊讶的结果让我感到十分兴奋。我曾经在一个项目中开发过识别物体的程序，整个过程充满艰难困苦。用人眼能够轻易区分物体，而想要在计算机中用程序识别物体，不仅需要丰富的数学知识，还要想办法通过代码实现。

通过计算机识别物体时，需要把识别物体所需的全部内容用代码编入程序，还必须包含可能误识别的相关物体和图像特点。若要识别不同形态的物体，则需要将以上的辛苦过程重复一遍。如果再辛苦一次能够出现好的结果自然很幸福，但现实往往很残酷。虽然在特定条件下识别率很高，但只要有一点点偏离轨道，识别率就会严重下滑。

但是，十分常见的机器学习示例 MNIST Digit Recognition 却能够轻松识别数字。它的示例代码不长，准确度达到 95% ~ 98%，这在过去是完全无法想象的。因此，我抱着对机器学习的期待，开始了学习之旅。每一段旅程都是相似的，想要遇到旅行指南中出现的梦幻般的美景并非易事，没过多久我就明白了这一点。我也再次意识到，

示例只是示例。但是，对机器学习抱有浓厚兴趣的我并不想就此放弃，所以开始了正式的学习。

以我个人的标准看，机器学习图书可以分为两大类：第一类对机器学习进行介绍，第二类对机器学习的数学背景进行说明。前者着重介绍机器学习算法和使用方法，虽然有助于学习机器学习的使用方法，但是在理解和使用机器学习的概念时，内容略显不足；后者则着重介绍机器学习算法中使用的数学概念和各种定理，过于专业。如果想开发新的机器学习算法，或者改善现有的机器学习算法，那么可以通过这些图书获得帮助。但是，如果重点在于使用机器学习算法，那么这些书则太过笼统。

我以口渴之人挖井的心情默默地钻研了机器学习之后发现，如果想要灵活运用机器学习，必须理解核心的数学概念，掌握想要适用领域的专业知识。

其实，使用机器学习编写程序时，机器学习算法所占的比例并不大，重要的是对数据的理解和对特征的把握。这个过程称为探索性数据分析（EDA，exploratory data analysis），要想切实执行该过程，需要具备统计和概率相关的数学知识。在有效进行EDA时，如果拥有专业知识，可以大幅缩减时间并简化问题。只有经历了这些过程，应用机器学习时才能呈现出好的结果。

我希望大家能够通过本书集中学习机器学习相关的数学理论、专业知识，以及实现机器学习的代码。如前所述，数据是灵活使用机器学习时更重要的要素，如果没有确定适用领域，那么只算完成了一半。因此，我选择的应用对象是股票，因为能够轻易获得数据，而且数据自身可信度高、有难度。

股市是典型的难预测领域，拥有学习机器学习需要的所有因素。在股市可以体验利用数学理论构建预测模型、处理模型所需的数据、预测股价、分析并改善训练结果等机器学习整体流程。

书中探讨的主题——统计、时间序列和算法交易的内容十分丰富，每一个主题

都很难用一本书完成说明。因此，本书将以和算法交易有直接联系的必知事项为中心进行讲解。

至于本书未涉及的内容，需要各位通过其他图书进行学习。本书假设读者均已具备编程能力，故不再单独讲解编程相关知识。

本书结构

本书大致分为三部分。

第一部分　机器学习概要。介绍机器学习的概念、功能、种类。

第 1 章　机器学习

这一章比较重要的内容有对机器学习概念的把握、机器学习能够解决的问题、机器学习的过程，以及“没有免费的午餐”定理（NFL，no free lunch theorem）。

第二部分　通过介绍与算法交易有关的数学背景知识探讨统计和时间序列。要想进行算法交易，需要构建决定股票买入和卖出的“模型”。该部分讲解构建模型所需的最基本的统计概念和时间序列概念。

第 2 章　统计

这一章包括机器学习的全部基础知识，阐述正确使用机器学习时必须清楚的概念——标准差、直方图、正态分布等。

第 3 章　时间序列数据

这一章介绍算法交易的理论基础——数学概念，最后添加实现算法时所需的必知内容。

第三部分　利用 Python 实现简单的算法交易。尝试实现以机器学习为基础的模型和以时间序列为基础的模型，对实现结果进行分析并讨论改善方法。

第 4 章　算法交易

这一章对算法交易的概念进行介绍，说明算法交易中使用的两种模型。

第 5 章　实现算法交易系统

这一章利用 Python 和库，实现第 4 章中的两种模型。

第 6 章　性能评价与优化

这一章评价第 5 章中实现的模型的预测性能，以及为了实现更高的预测性应该如何进行优化。

实操所需软 / 硬件

机器学习需要同时具备软件和硬件。对于所需的硬件，应当尽可能选择速度快的 CPU。因为机器学习算法进行的计算比其他软件多，执行时间少则数秒，多则数月，所以 CPU 越快越能尽早取得结果。

如果想正式进行机器学习，我强烈推荐各位使用 GPU，而不是 CPU。GPU 是数学和科学计算中的专业设备，运算速度相当快。CPU 能够串联处理，而 GPU 能够并行处理，所以在相同时间内，GPU 的处理速度更快。最新的 CPU 有 4 核和 8 核，而 GPU 拥有数千个核。并行处理时，GPU 能够在同一时间内用数千个核同时处理数千个计算。

软件的选择同样重要。使用硬件的过程中，如果需要更好的性能，可以通过升级轻松实现。但是，软件中使用的库和开发的机器学习代码隶属于语言和库，需要慎重选择。

人们过去认为，选择软件时要追求速度。因为机器学习需要进行非常多的计算，运行时间很长，所以能够尽快看到代码结果比长时间的等待有意义，因此经常使用 C、C++ 等语言。

但是，目前计算能力已经比以前提高了很多，如果短时间内需要强大的计算能力，可以使用云。最重要的是，GPU 的登场使速度的魅力逐渐消减（图像识别等领域的运行时间依旧很长，所以同样偏爱 C++）。因此，库正在成为软件选择的重要因素。

选择机器学习库时，应该考虑它是否支持所需的机器学习算法、是否拥有处理数据时所需的功能，以及能否持续升级等。

想要亲自实现机器学习算法，不仅需要数学知识，还要通过诸多测试和优化等进行检验。因此，如果不是为了创建新的机器学习算法或者改善现有的机器学习算法，那就没有必要亲自开发。

想要使用机器学习，只需理解机器学习算法的相关概念及其使用方法。从编写机器学习程序的整体过程可知，机器学习算法自身所占的时间比例并不多，只有 10% ~ 20%。将数据用于机器学习的预处理、分析并改善机器训练结果的后期处理等过程更加重要，而且需要更多时间，所以要清楚库中是否包含这些功能，因为预处理和后期处理决定了机器训练结果的质量。

选择机器学习库时，必须确认其能否支持 GPU。CPU 和 GPU 之间运算速度的差异小则数倍，大则数百倍、数千倍，因此，如果机器学习中使用的数据较大，或者要使用深度学习等计算量较多的算法时，GPU 必不可少。

本书推荐的算法交易硬件环境和软件整理如下。

- **硬件**：Intel i5 以上，内存 4 GB，HDD 256 GB 以上。
- **操作系统**：支持 Python（Windows、OS X、Linux）。
- **数据库**：MySQL（用于保存股价相关数据）。
- **编程语言**：Python，广泛用于 Web 开发、云、金融。简练易学，拥有多种库。
- **库**
  - NumPy：提供高维数学功能的开源 Python 库。NumPy 的核心功能是 ndarray，

它是 *n* 维数组数据类型，能够快速灵活地使用多维数组。NumPy 可以用作各种数学和科学运算中常用的向量和标量，同时能够与数据库联动使用。

- SCIPY：提供科学计算所需功能的库，提供优化、线性代数、积分、FFT 等功能。
- Pandas：处理金融数据的库，使用 Dataframe 类，拥有处理时间序列金融数据所需的各种功能。
- Matplotlib：拥有绘制图表或数据可视化所需的多种功能，同时提供保存或缩放图表所需的简单 UI。
- scikit-learn（sklearn）：Python 机器学习库，能够实现除深度学习以外的几乎所有机器学习算法，同时包含数据处理和分析训练结果的功能。使用方法不受算法影响，是能在短暂的学习时间内直观使用的具有代表性的 Python 机器学习库。
- Statsmodels：Python 统计库，支持数据挖掘、统计模型推测、统计测试等与统计相关的各种功能。

使用 Anaconda 安装库

Anaconda 程序能够一次性安装书中实操所需的包和相关程序，支持 Windows、OS X、Linux。如果不熟悉 Python，那么可能很难通过 Python 安装程序 pip 等逐个安装所需的库，此时可以使用 Anaconda 构建所需的实操环境。

在 Anaconda 主页可查看详细说明并下载。Anaconda 的安装十分简单，运行下载的文件即可完成。Anaconda 提供 Python 2.7 和 Python 3.5 两种版本，建议各位安装 Python 2.7 版本。本书中使用的库和代码均在 Python 2.7 版本中完成测试。

操作系统不同，Anaconda 安装的库也有所不同。Linux 能够安装 Anaconda 中支持的所有库，OS X 支持大部分，Windows 最少。

示例代码下载

参见“图灵社区”本书主页（http://www.ituring.com.cn/book/1929）“随书下载”。

目 录

第一部分

第二部分

第三部分