TURING 图灵程序设计丛书

MANNING

Machine Learning in Action

机器学习实战

[美] Peter Harrington 著
李锐 李鹏 曲亚东 王斌 译

使用Python阐述机器学习概念
介绍并实现机器学习的主流算法
面向日常任务的高效实战内容

人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

Machine Learning in Action

机器学习实战

[美] Peter Harrington 著
李锐 李鹏 曲亚东 王斌 译

人民邮电出版社
北京

图书在版编目（CIP）数据

机器学习实战 / (美) 哈林顿 (Harrington,P.) 著；
李锐等译. -- 北京 : 人民邮电出版社, 2013.6
（图灵程序设计丛书）
书名原文: Machine learning in action
ISBN 978-7-115-31795-7

Ⅰ. ①机… Ⅱ. ①哈… ②李… Ⅲ. ①机器学习－研究 Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2013)第095060号

内 容 提 要

机器学习是人工智能研究领域中的一个极其重要的方向。在现今大数据时代的背景下，捕获数据并从中萃取有价值的信息或模式，使得这一过去为分析师与数学家所专属的研究领域越来越为人们瞩目。

本书通过精心编排的实例，切入日常工作任务，摒弃学术化语言，利用高效可复用的 Python 代码阐释如何处理统计数据，进行数据分析及可视化。读者可从中学到一些核心的机器学习算法，并将其运用于某些策略性任务中，如分类、预测及推荐等。

本书适合机器学习相关研究人员及互联网从业人员学习参考。

◆ 著 [美] Peter Harrington
译 李 锐 李 鹏 曲亚东 王 斌
责任编辑 丁晓昀
执行编辑 李 鑫 龚 雪
责任印制 焦志炜

◆ 人民邮电出版社出版发行 北京市崇文区夕照寺街14号
邮编 100061 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
北京天宇星印刷厂印刷

◆ 开本：800×1000 1/16
印张：20.75
字数：490千字 2013年6月第1版
印数：1－4 000册 2013年6月北京第1次印刷
著作权合同登记号 图字：01-2012-4878号

定价：69.00元

读者服务热线：(010)51095186转604 印装质量热线：(010)67129223
反盗版热线：(010)67171154
广告经营许可证：京崇工商广字第 0021 号

译 者 序

这是我翻译的第三本书了，前两本分别是《信息检索导论》和《大数据：大规模互联网数据挖掘与分布式处理》。与图灵公司有了这两次合作后，我们一直保持着十分密切的联系。2012年11月，图灵的编辑和我说，这本书的原译者不能继续翻译了，问我能否续译后面的十二章。我翻阅了一下，觉得这本书不错，能帮助不少人，于是很快就接下了这个翻译任务，并在11月底启动了我的第三次图灵翻译之旅。

我翻译的这三本书分别涉及信息检索、数据挖掘和机器学习。虽然这几个领域各不相同，但是它们之间有着十分密切的关联。简单地说，机器学习算法在包含信息检索和数据挖掘在内的多个领域中都有着十分广泛的应用。现代互联网中的搜索引擎、社交网络、推荐引擎、计算广告、电子商务等应用中，都包含大量的机器学习算法。“机器学习”已经成为学术界和工业界炙手可热的术语。了解机器学习算法，是很多研究人员和互联网从业人员的基本要求。

翻译本书期间，业界和研究界也出现了大量热点名词，包括“大数据”（big data）、“深度学习”（deep learning）、“知识图谱”（knowledge graph）等，基于社交网络的研究和应用也层出不穷。可以说，机器学习与这些名词之间都具有十分密切的联系，了解机器学习对于把握业界和研究界的脉搏至关重要。

本书没有从理论角度来揭示机器学习算法背后的数学原理，而是通过“原理简述+问题实例+实际代码+运行效果”来介绍每一个算法。学习计算机的人都知道，计算机是一门实践学科，没有真正实现运行，很难真正理解算法的精髓。这本书的最大好处就是边学边用，非常适合于急需迈进机器学习领域的人员学习。实际上，即使对于那些对机器学习有所了解的人来说，通过代码实现也能进一步加深对机器学习算法的理解。

本书的代码采用Python语言编写。Python代码简单优雅、易于上手，科学计算软件包众多，已经成为不少大学和研究机构进行计算机教学和科学计算的语言。相信Python编写的机器学习代码也能让读者尽快领略到这门学科的精妙之处。

由于个人精力有限，加上时间紧迫，和前两本书都是独立翻译有所不同，本书邀请了多名颇具实力的译者共同完成。全书共包括15章4个附录，曲亚东翻译第1～3章，李鹏博士翻译第4、10、11、12章及附录A、B，李锐博士翻译第5、8、9、15章及附录C、D，王斌翻译第6、7、13、14章及其他部分并审校全文。

感谢翻译过程中图灵公司谢工、傅志红、李鑫、郭志敏、刘紫凤等人给予的帮助，感谢所有译者的家人朋友一如既往的支持和鼓励，感谢所有帮助和指导过我们的人。

致约瑟夫与米洛。

由于译者水平有限，书中难免会有疏漏，还望读者不吝提出意见和建议。同前几本书一样，本书的勘误也会在网上及时公布，地址在：http://ir.ict.ac.cn/~wangbin/mli-book。读者可以通过邮件wbxjj2008@gmail.com或者新浪微博和我联系。

王斌

2013年1月15日凌晨于中关村

前　言

大学毕业后，我先后在加利福尼亚和中国大陆的Intel公司工作。最初，我打算工作两年之后回学校读研究生，但是幸福时光飞逝而过，转眼就过去了六年。那时，我意识到我必须回到校园。我不想上夜校或进行在线学习，我就想坐在大学校园里吸纳学校传授的所有知识。在大学里，最好的方面不是你研修的课程或从事的研究，而是一些外围活动：与人会面、参加研讨会、加入组织、旁听课程，以及学习未知的知识。

在2008年，我帮助筹备一个招聘会。我同一个大型金融机构的人交谈，他们希望我去应聘他们机构的一个对信用卡建模（判断某人是否会偿还贷款）的岗位。他们问我对随机分析了解多少，那时，我并不能确定“随机”一词的意思。他们提出的工作地点令我无法接受，所以我决定不再考虑了。但是，他们说的“随机”让我很感兴趣，于是我拿来课程目录，寻找含有“随机”字样的课程，我看到了“离散随机系统”。我没有注册就直接旁听了这门课，完成课后作业，参加考试，最终被授课教授发现。但是她很仁慈，让我继续学习，这让我非常感激。上这门课，是我第一次看到将概率应用到算法中。在这之前，我见过一些算法将平均值作为外部输入，但这次不同，方差和均值都是这些算法中的内部值。这门课主要讨论时间序列数据，其中每一段数据都是一个均匀间隔样本。我还找到了名称中包含“机器学习”的另一门课程。该课程中的数据并不假设满足时间的均匀间隔分布，它包含更多的算法，但严谨性有所降低。再后来我意识到，在经济系、电子工程系和计算机科学系的课程中都会讲授类似的算法。

2009年初，我顺利毕业，并在硅谷谋得了一份软件咨询的工作。接下来的两年，我先后在涉及不同技术的八家公司工作，发现了最终构成这本书主题的两种趋势：第一，为了开发出竞争力强的应用，不能仅仅连接数据源，而需要做更多事情；第二，用人单位希望员工既懂理论也能编程。

程序员的大部分工作可以类比于连接管道，所不同的是，程序员连接的是数据流，这也为人们带了巨大的财富。举一个例子，我们要开发一个在线出售商品的应用，其中主要部分是允许用户来发布商品并浏览其他人发布的商品。为此，我们需要建立一个Web表单，允许用户输入所售商品的信息，然后将该信息传到一个数据存储区。要让用户看到其他用户所售商品的信息，就要从数据存储区获取这些数据并适当地显示出来。我可以确信，人们会通过这种方式挣钱，但是如果让要应用更好，需要加入一些智能因素。这些智能因素包括自动删除不适当的发布信息、检测不正当交易、给出用户可能喜欢的商品以及预测网站的流量等。为了实现这些目标，我们需要应用机器学习方法。对于最终用户而言，他们并不了解幕后的“魔法”，他们关心的是应用能有效运行，这也是好产品的标志。

一个机构会雇用一些理论家（思考者）以及一些做实际工作的人（执行者）。前者可能会将大部分时间花在学术工作上，他们的日常工作就是基于论文产生思路，然后通过高级工具或数学进行建模。后者则通过编写代码与真实世界交互，处理非理想世界中的瑕疵，比如崩溃的机器或者带噪声的数据。完全区分这两类人并不是个好想法，很多成功的机构都认识到这一点。（精益生产的一个原则就是，思考者应该自己动手去做实际工作。）当招聘经费有限时，谁更能得到工作，思考者还是执行者？很可能是执行者，但是现实中用人单位希望两种人都要。很多事情都需要做，但当应用需要更高要求的算法时，那么需要的人员就必须能够阅读论文，领会论文思路并通过代码实现，如此反复下去。

在这之前，我没有看到在机器学习算法方面缩小思考者和执行者之间差距的书籍。本书的目的就是填补这个空白，同时介绍机器学习算法的使用，使得读者能够构建更成功的应用。

关 于 本 书

本书讲述重要的机器学习算法，并介绍那些使用这些算法的应用和工具，以及如何在实际环境中使用它们。市面上已经出版了很多关于机器学习的书籍，大多数讨论的是其背后的数学理论，很少涉及如何使用编程语言实现机器学习算法。本书恰恰相反，更多地讨论如何编码实现机器学习算法，而尽量减少讨论数学理论。如何将数学矩阵描述的机器学习算法转化为可以实际工作的应用程序，是本书的主要目的。

读者对象

机器学习是什么？谁需要使用机器学习算法？简而言之，机器学习可以揭示数据背后的真实含义。这本书适合有数据需要处理的读者，也适合于想要获得并理解数据的读者。如果读者有一些编程概念（比如递归），并且了解一些数据结构（比如树结构），那么将有助于本书的阅读。尽管机器学习领域的专家不一定能从本书获益，但是如果读者具有线性代数和概率论的入门知识，那么也会利于本书的阅读。此外，本书使用Python语言进行编程，它过去也被称作“可执行的伪代码”。本书假定读者有一些基本的Python编程知识，不过不知道如何使用Python也没有关系，只要具备基本的编程思想，学习Python也不困难。

数据挖掘十大算法

数据以及基于数据做出决策是非常重要的，本书内容也是来源于数据——“数据挖掘十大算法”是IEEE数据挖掘国际会议（ICDM）上的一篇论文，2007年12月在*Journal of Knowledge and Information Systems*杂志上发表。依据知识发现和数据挖掘国际会议（KDD）获奖者的问卷调查结果，论文统计出排名前十的数据挖掘算法。本书的基本框架与论文中提到的算法基本一致。聪明的读者可能已经注意到，虽然论文只给出了十个重要的数据挖掘算法，但本书却有十五章。下面我会给出解释，这里我们先看看排名前十的数据挖掘算法。

论文选出的机器学习算法包括：C4.5决策树、K-均值（K-mean）、支持向量机（SVM）、Apriori、最大期望算法（EM）、PageRank算法、AdaBoost算法、*k*-近邻算法（kNN）、朴素贝叶斯算法（NB）和分类回归树（CART）算法。本书包含了其中的8个算法，没有包括最大期望算法和PageRank算法。本书没有包括PageRank算法，是因为搜索引擎巨头Google引入的PageRank算法已经在很多著作里得到了充分的论述，没有必要进一步累述；而最大期望算法没有纳入，是因为涉及太多的

数学知识，如果它像其他算法那样简化成一章，就无法讲述清楚算法的核心，有兴趣的读者可以参阅相关材料。

本书结构

本书由四大部分15章和4个附录组成。

第一部分 分类

本书并没有按照“数据挖掘十大算法”的次序来介绍机器学习算法。第一部分首先介绍了机器学习的基础知识，然后讨论如何使用机器学习算法进行分类。第2章介绍了基本的机器学习算法：*k*-近邻算法；第3章是本书第一次讲述决策树；第4章讨论如何使用概率分布算法进行分类以及朴素贝叶斯算法；第5章介绍的Logistic回归算法虽然不在排名前十的列表中，但是引入了算法优化的主题，也是非常重要的，这一章最后还讨论了如何处理数据集合中的缺失值；第6章讨论了强大而流行的支持向量机；第7章讨论AdaBoost集成方法，它也是本书讨论分类机器学习算法的最后一章，这一章还讨论了训练样本非均匀分布时所引发的非均衡分类问题。

第二部分 利用回归预测数值型数据

第二部分包含两章，讨论连续型数值的回归预测问题。第8章主要讨论了回归、去噪和局部加权线性回归，此外还讨论了机器学习算法必须考虑的偏差方差折中问题。第9章讨论了基于树的回归算法和分类回归树（CART）算法。

第三部分 无监督学习

前两部分讨论的监督学习需要用户知道目标值，简单地说就是知道在数据中寻找什么。而第三部分开始讨论的无监督学习则无需用户知道搜寻的目标，只需要从算法程序中得到这些数据的共同特征。第10章讨论的无监督学习算法是K-均值聚类算法；第11章研究用于关联分析的Apriori算法；第12章讨论如何使用FP–Growth算法改进关联分析。

第四部分 其他工具

本书的第四部分介绍机器学习算法使用到的附属工具。第13章和第14章引入的两个数学运算工具用于消除数据噪声，分别是主成分分析和奇异值分解。一旦机器学习算法处理的数据集扩张到无法在一台计算机上完全处理时，就必须引入分布式计算的概念，本书最后一章将介绍MapReduce架构。

示例

本书的许多示例演示了如何在现实世界中使用机器学习算法，通常我们按照下面的步骤保证算法应用的正确性：

(1) 确保算法应用可以正确处理简单的数据；

(2) 将现实世界中得到的数据格式化为算法可以处理的格式；

(3) 将步骤2得到的数据输入到步骤1的算法中，检验算法的运行结果。

千万不要忽略前两个步骤而直接跳到步骤3来检验算法处理真实数据的效果。任何复杂系统都是由基础工程构成的，尤其是算法出现问题时，增量地搭建系统可以确保我们及时找到问题出现的位置和原因。如果刚开始就把这些堆砌在一起，我们就很难发现到底是不准确的算法实现引发的问题还是数据格式的问题。此外，本书在实现算法的过程中，记录了很多注意事项，将有助于读者深入了解机器学习算法。

代码约定和下载

本书正文和程序清单中的源代码都使用等宽字体。一些程序清单中包含了代码注解，以突出其中蕴含的重要概念。在某些场合，带编号的程序注释会在程序清单之后进一步解释说明。

本书所有源代码均可在英文版出版商的网站上下载：www.manning.com/MachineLearninginAction。[①]

作者在线

本书的读者还可以访问出版商Manning的网络论坛。在论坛上读者可以评论本书的内容，讨论技术问题，得到作者或其他用户的帮助。为了使用和订阅论坛，请访问http:/www.manning.com/MachineLearninginAction，该网页包含如何注册论坛、如何获取帮助以及论坛的行为规则。

出版商Manning对读者承诺，为读者和作者提供讨论的空间。作者自愿参与作者在线论坛，我们也不承诺作者参与论坛讨论的次数。建议读者尽量问作者具有挑战性的问题，以免浪费作者的宝贵时间。

只要本书英文版在销售，读者都可以访问英文版出版商的作者在线论坛，阅读以前的讨论文档。

① 读者也可以访问图灵社区本书页面提交勘误或下载源代码，网址是ituring.com.cn/book/1021。

致　谢

这是目前为止本书最容易写的部分……

首先，我要感谢Manning出版社的工作人员，尤其是本书的编辑Troy Mott，如果没有他的支持和热情帮助，本书不会出版。我还要感谢Maureen Spencer，她对最终稿进行了润色，和她在一起工作相当愉快。

其次，我要感谢Arizona州立大学的Jennie Si老师，她允许我在未注册的情况下听她的“离散随机系统”课。还要感谢MIT的Cynthia Rudin，他给我推荐了论文“Top 10 Algorithms in Data Mining”[①]（数据挖掘十大算法），促成了本书的写作思路。Mark Bauer、Jerry Barkely、Jose Zero、Doug Chang、Wayne Carter以及Tyler Neylon对本书亦有贡献，在此一并感谢。

特别要感谢在成书过程当中提供珍贵反馈意见的评阅人，他们是：Keith Kim、Franco Lombardo、Patrick Toohey、Josef Lauri、Ryan Riley、Peter Venable、Patrick Goetz、Jeroen Benckhuijsen、Ian McAllister、Orhan Alkan、Joseph Ottinger、Fred Law、Karsten Strøbæk、Brian Lau、Stephen McKamey、Michael Brennan、Kevin Jackson、John Griffin、Sumit Pal、Alex Alves、Justin Tyler Wiley和John Stevenson。

技术校对人员Tricia Hoffman和Alex Ott在本书出版之前对技术内容进行了快速审阅，对于他们的意见和反馈我表示感谢。当阅读书中的代码时，Alex表现得像一个冷血杀手！谢谢他对本书的贡献。

我还要感谢那些通过MEAP购买和阅读早期版本的读者，以及对作者在线论坛做出贡献的人们（甚至是发“钓鱼贴”的用户）。如果没有这些人的帮助，这本书就不是现在这个样子。

我还要感谢我的家庭在写书期间给予的支持。感谢我爱人的鼓励以及在写书期间对我的非规律生活的宽容。

最后，我要感谢硅谷这个伟大的地方，我和我爱人在这里工作、交流思想和情感。

---

① Xindong Wu等，“Top 10 Algorithms in Data Mining”，*Journal of Knowledge and Information Systems* 14, no. 1 (December 2007)。

关于封面

本书封面插画的标题为“伊斯特里亚人”（“Man from Istria”，伊斯特里亚是克罗地亚面向亚得里亚海的一个很大半岛）。该插画来自克罗地亚斯普利特民族博物馆2008年出版的Balthasar Hacquet的《图说西南及东汪达尔人、伊利里亚人和斯拉夫人》（*Images and Descriptions of Southwestern and Eastern Wenda, Illyrians, and Slavs*）的最新重印版本。Hacquet（1739—1815）是一名奥地利内科医生及科学家，他花费数年时间去研究各地的植物、地质和人种，这些地方包括奥匈帝国的多个地区，以及伊利里亚部落过去居住的（罗马帝国的）威尼托地区、尤里安阿尔卑斯山脉及西巴尔干等地区。Hacquet发表的很多论文和书籍中都有手绘插图。

Hacquet出版物中丰富多样的插图生动地描绘了200年前西阿尔卑斯和巴尔干西北地区的独特性和个体性。那时候相距几英里的两个村庄村民的衣着都迥然不同，当有社交活动或交易时，不同地区的人们很容易通过着装来辨别。从那之后着装的要求发生了改变，不同地区的多样性也逐渐消亡。现在很难说出不同大陆的居民有多大区别，比如，现在很难区分斯洛文尼亚的阿尔卑斯山地区或巴尔干沿海那些美丽小镇或村庄里的居民和欧洲其他地区或美国的居民。

Manning出版社利用两个世纪之前的服装来设计书籍封面，以此来赞颂计算机产业所具有的创造性、主动性和趣味性。正如本书封面的图片一样，这些图片也把我们带回到过去的生活中去。

目　录

第一部分　分类