HZ BOOKS 华章IT

[PACKT] PUBLISHING

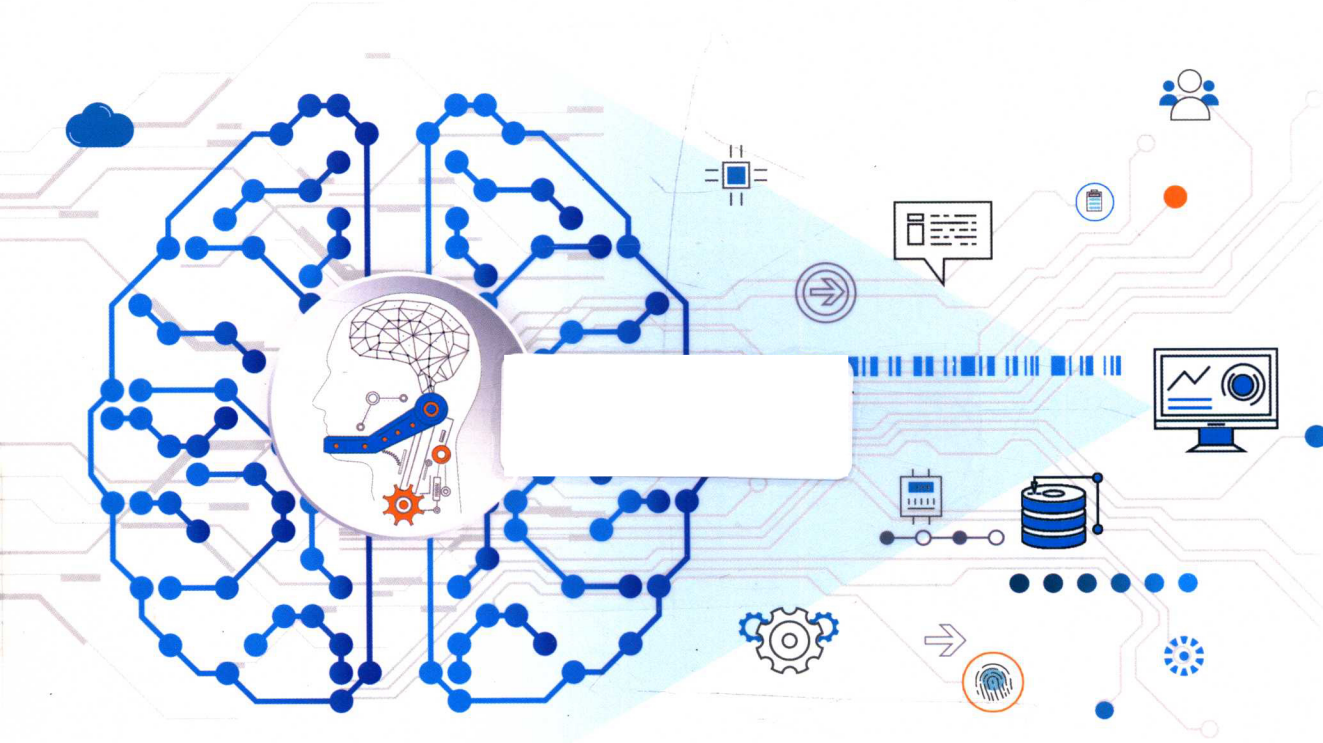
智能系统与技术丛书

Mastering Java Machine Learning

Java机器学习

[美] 乌黛·卡马特（Uday Kamath）
克里希纳·肖佩拉（Krishna Choppella） 著

陈瑶 陈峰 刘江一 等译



机械工业出版社
China Machine Press

智能系统与技术丛书

Mastering Java Machine Learning

Java机器学习

[美] 乌黛·卡马特（Uday Kamath）
克里希纳·肖佩拉（Krishna Choppella） 著
陈翔 陈峰 刘江一 等译

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Java 机器学习 /（美）乌黛·卡马特（Uday Kamath），（美）克里希纳·肖佩拉（Krishna Choppella）著；陈瑶等译 . —北京：机械工业出版社，2018.9
（智能系统与技术丛书）
书名原文：Mastering Java Machine Learning

ISBN 978-7-111-60919-3

I. J… II. ①乌… ②克… ③陈… III. JAVA 语言 – 程序设计 IV. TP312.8

中国版本图书馆 CIP 数据核字（2018）第 213305 号

本书版权登记号：图字 01-2017-7498

Java 机器学习

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：刘 锋
责任校对：李秋荣
印 刷：三河市宏图印务有限公司
版 次：2018 年 9 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16
印 张：21.25
书 号：ISBN 978-7-111-60919-3
定 价：89.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066
投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259
读者信箱：hzit@hzbook.com

FOREWORD

推 荐 序

Uday Kamath 博士是一个拥有奇思妙想的人。每一次他到我的办公室，我们都会展开很多有意义和富有成效的讨论。我已经在乔治梅森大学（GMU）作为计算机科学的副教授任教了15年，主要研究方向是机器学习和数据挖掘。认识 Uday 5年，第一次见面是在我的数据挖掘课上，他那时还是个学生，然后我们成了同事，以及大规模机器学习的项目和论文的共同作者。当 Uday 作为 BAE 系统应用智能公司（BAE Systems Applied Intelligence）[⊖]的首席数据科学家时，他获得了进化计算和机器学习方向的博士学位。对 Uday 来说，似乎拥有两个高要求的工作还不够，他的经历异常丰富，在 GMU 任职期间，他分别和计算机学院的四个人合作发表了多篇论文，这是不常见的。鉴于 Uday 的这种特质，不到四年他就博士毕业，对此我并不感到惊讶，现在我正为他的这本使用 Java 语言掌握高级机器学习技术的书写推荐序。Uday 对于新的富有刺激性的挑战的渴望再次出现，所以你才会看到手中这本超棒的书。

这本书是 Uday 浓厚的兴趣和全面的、夯实的理论知识的产物，同时也是他对书中所推荐的方法的实际可行性的敏锐领会。虽然已经有一些关于机器学习和数据分析的书，但 Uday 的书填补了理论和实际之间大量的空白。它提供了对于经典又高级的机器学习技术全面的、系统的分析，将重点放在技术的优点和局限性，以及技术的实际使用和实现上。对于从事数据科学和分析的人，以及热衷于想要掌握机器学习技术实用、有效实现的本科生和研究生来说，本书都是一份不可多得的好书、好资料。

这本书涵盖了机器学习中的经典技术，如分类、聚类、降维、异常检测、半监督学习和主动学习。同时介绍了新的高级主题，包括流数据学习、深度学习以及大数据学习的挑战。每一章指定一个主题，通过案例研究，介绍最前沿的基于 Java 的工具和软件，以及完整的知识发现周期——数据采集、实验设计、建模、结果及评估。每一章都是独立的，提供了很大的使用灵活性。附带的网站提供了源码和数据。对于学生和数据分析从业员来说，这确实很难得，大家可以直接用刚学到的方法进行实验，或者通过将这些方法应用到真实环境中加深对它们的理解。

⊖ BAE 系统应用智能公司是全球最大的军品公司之一，曾在 2011 年位居第一。——译者注

当阅读这本书的各个章节时，我想起了 Uday 对于学习和知识的热情。他在本书中描述的概念依旧清晰又饱含热情。我确信，作为读者，你也会感同身受。我一定会将这本书作为我所教课程的推荐资源，强烈推荐给学生。

Carlotta Domeniconi 博士

乔治梅森大学计算机科学副教授

THE TRANSLATOR'S WORDS

译 者 序

本书由陈瑶、陈峰、刘江一、刘旭斌、李勤5位译者共同翻译，其中3位已经在机器学习领域应用了本书所探讨的相关知识体系。英文版原书有500多页，其中介绍了大量的实例。每一章的结构基本相同，首先是基础理论介绍，而后是案例学习，所以对于初学者或者机器学习领域的同行，本书既可以作为查询机器学习现有理论的“字典”类手册，又可以作为想要通过阅读一些实例加深知识理解的工具书。如果想要了解某一个细分领域问题，有深入阅读需求的读者还可以查询在各类国际会议上发表的论文。在翻译过程中，我们对原书中的错误也进行了校阅。如果各位读者发现了中文版中的错误，欢迎大家积极指正。

在机器学习领域，新的理论和研究成果层出不穷，然而这些研究和应用领域之间存在着共识的差距，每一个在机器学习类书籍上贡献出宝贵时间的译者，都有一个共同的愿望，即希望通过知识的扩大化传播，能让更多的人加入这个还需要前仆后继的行业，构建行业间的桥梁，帮助人们生活得越来越好，提高研究成果的可应用率，最后真正实现产业化。

PREFACE

前 言

关于机器学习有许多著名的书籍，有的是从数据中学习的理论教材；有的是某个特定领域的标准参考书，例如聚类、离群值检测或概率图建模；有的是操作指导，提供使用某种编程语言及相应的工具和库函数进行实践的示例。其中那些对理论涵盖面比较广的书，对细节的阐述有所欠缺，而那些专注于某个课题或工具的书也有短板，例如，可能解释了很多在流环境和在批处理环境中的方法有什么不同之处。另外，对于一些有机器学习方面的经验，而且喜欢用 Java 工具的读者，若他们希望通过某本书来拓展他们的知识，从本质上获得提升，那么合适的书籍尤其凤毛麟角。

在一本书中，他们希望找到：

- 各种不同技术之间的差别，对于不同场景中的数据——有标签数据和无标签数据、数据流或批处理、本地数据或者分布式存储的数据、结构化的或非结构化的数据，每种技术有哪些长处和短处。
- 应用某种机器学习理论的成熟的方法示例，使用最合适的技术，包括该理论最重要的数学表达式，以及这些技术如何能够最大限度地发挥该理论的优势。
- 对成熟的基于 Java 的框架、库、可视化工具的描述性的介绍，以及如何把这些技术应用到实践中。

据我们所知，目前为止，这样的书一本也没有。

鉴于以上情况，本书的核心思想就是要填补这个空白，力图在理论和实践中取得平衡：一方面使用概率论、统计学、基础线性代数、初等微积分等解释机器学习的理论；另一方面强调方法论、实例研究、工具和代码示例，作为实践的支撑。

根据 KDnuggets 2016 年的软件调查报告显示，在机器学习使用的编程语言中，有 16.8% 的人投了 Java 一票，它是第二受欢迎的语言，仅次于 Python。更重要的是，比起 2015 年，Java 的受欢迎程度提高了 19%！显然，在建立和部署与机器学习相关的系统方面，Java 仍然是一种重要且高效的工具，偶尔的支持率下降也不影响大局。在本书中，我们的目标是让有一定 Java 编程经验和机器学习基础知识的读者，成为该领域中既专业又热情的爱好者。本书的目的就是铺一条阳光大道，以便读者向成为资深的数据科学工作者这个方向迈进。为使读者

的进阶之路更加顺利，本书囊括了一个名副其实的机器学习技术弹药仓库，包括数据分析方法、学习算法、模型性能评估以及更多的监督学习和无监督学习、聚类和异常检测、半监督学习和主动学习等相关的内容。读者可能对其中一些已十分熟悉，对另一些没那么熟悉，而只是粗略地了解。本书还讲述了一些特别的话题，例如概率图模型、文本挖掘和深度学习。鉴于如今企业级别的系统越来越受重视，本书也涵盖了这方面的独特挑战，包括从数据流中学习、可应用于实时系统的工具和技术，以及大数据世界的必要架构：

- 机器学习如何在大规模分布式环境下工作？
- 在上述条件下有哪些必要的权衡？
- 算法需要做哪些必要的调整？
- 上述这类系统如何与强大的 Hadoop 生态系统的其他技术交互操作？

本书将会解释如何把机器学习应用到真实世界的数据和相关领域中，并提供了正确的方法论、流程、应用软件以及分析。每一章都包含了案例研究，介绍如何使用最合适的开源 Java 工具来应用本章所学的技术。本书介绍了超过 15 种开源 Java 工具，广泛支持各种技术，既有代码示例，也有使用实践。所有的代码、数据和配置，读者都可以下载并进行实验。我们还展示了超过 10 个真实世界的机器学习案例，演示了数据科学家的工作流程。每个案例都有以下实验步骤的细节：数据提取、数据分析、数据清理、特征降维/选择、映射到机器学习、模型训练、模型选择、模型进化以及结果分析。读者可以将此作为实践指导，学习如何将各章介绍的工具和方法论用于解决手头的业务问题。

主要内容

第 1 章介绍了机器学习的基本概念和技术。读者在 Packt 的其他类似书籍中也可以看到这些内容，例如《Learning Machine Learning in Java》等。本章涉及的概念有：数据、数据转换、采样和偏移、特征及其重要性、监督学习、无监督学习、大数据学习、数据流和实时学习、概率图模型，以及半监督学习。

第 2 章单刀直入地展示了监督学习的广泛场景及其相关技术的全景，还涵盖了特征选择和降维、线性建模、逻辑模型、非线性模型、SVM 和核函数、集成学习技术（例如装袋算法和提升算法）、验证技术和评价指标，还有模型选择。本章的案例研究使用了 Weka 和 RapidMiner，包括从数据分析到模型性能分析的所有步骤。和其他各章一样，案例研究是作为示例来帮助读者理解本章介绍的技术是如何应用到真实生活中的。这个案例研究所使用的数据集来自 UCI Horse Colic。

第 3 章展示了多种先进的聚类和离群值技术及其应用。本章涵盖的主题包括无监督数据的特征选择和降维、聚类算法、聚类的模型评估，以及使用统计学方法、距离和分布式技术做异常检测。在本章末尾，我们展示了一个案例研究，使用一组真实世界的图像数据集 MNIST 进行聚类和离群值检测。另外，使用 Smile API 完成特征降维，使用 ELKI 进行学习。

第 4 章讲述了当只有少量的标签数据可以使用时，学习的算法和技术的细节。本章涵盖

的主题包括自训练、生成模型、转导 SVM 算法、协同训练、主动学习和多视角学习。案例研究使用了两种学习系统，基于 UCI 威斯康星乳腺癌数据集来展开。本章介绍的工具有 JKernelMachines、KEEL 和 JCLAL。

第 5 章涵盖了对实时呈现的独特环境下的数据流进行数据学习的问题。本章涉及的内容有：流机器学习和应用、监督的流学习、无监督聚类流学习、无监督离群值学习、流学习的评估技术以及评估使用的指标。本章末尾的详细案例研究说明了如何使用 MOA 框架。使用的数据集是 Electricity（ELEC）。

第 6 章展示了对多维空间中的复合关联概率分布进行编码，可以有效地表示许多现实问题。概率图模型提供了一个框架来表示、绘制推断，并在这种情况下有效地学习。本章大体上涵盖概率概念、PGM、贝叶斯网络、马尔可夫网络、图结构学习、隐马尔可夫模型和推断。本章末尾会使用真实的数据集进行详细的案例研究。案例研究中使用的工具有 OpenMarkov 和 Weka 的贝叶斯网络。数据集是 UCI Adult（Census Income）。

第 7 章介绍深度学习。如果今天在大家的想象中有一个机器学习的超级明星，那一定是深度学习，它已经在解决最复杂的 AI 问题的技术中占据了主导地位。本章的主题广泛地涵盖了神经网络、神经网络中的问题、深度信念网络、受限玻耳兹曼机、卷积网络、长短期记忆单元、降噪自动编码器、循环网络等。我们提供了一个详细的案例研究来展示如何实现深度学习网络、调整参数和执行学习。本章使用了 DeepLearning4J 和 MNIST 图像数据集。

第 8 章详细地介绍了在文本挖掘领域执行各种分析的技术、算法和工具。广泛地涵盖了文本挖掘、文本挖掘所需的组件、文本数据的表示、降维技术、主题建模、文本聚类、命名实体识别和深度学习等领域的主题。案例研究使用真实的非结构化文本数据（Reuters-21578 数据集）突出主题建模和文本分类，使用的工具是 Mallet 和 KNIME。

第 9 章讨论了当今最重要的挑战。当数据很大或者以非常高的速率增加时，可以使用哪些学习方案？如何处理可扩展性？主题涵盖了大数据集群部署框架、大数据存储选项、批数据处理、批数据机器学习、实时机器学习框架和实时流学习。在批量和实时大数据的案例研究中，我们选择了 UCI Covertype 数据集和机器学习库 H2O、Spark MLlib 和 SAMOA。

附录 A 涵盖了线性代数的概念，作为一个简单的复习。它的覆盖范围一定不是完整的，但是它粗略地包含了一些与本书所述的机器学习技术相关的重要概念。包括向量、矩阵、基本矩阵运算和属性、线性变换、矩阵逆、特征分解、正定矩阵和奇异值分解。

附录 B 提供了一个概率论的简要介绍。包括概率公理、贝叶斯定理、概率密度估计、平均值、方差、标准差、高斯标准差、协方差、相关系数、二项分布、泊松分布、高斯分布、中心极限定理和误差传播。

必备知识

本书假设你有一些 Java 编程经验，并对机器学习概念有基本的了解。如果你既没有经验也不太了解机器学习，但是你很好奇并且是一个自我激励的人，那么不要担心，继续阅读吧！

对于那些有一定相关背景的人来说，意味着熟悉简单的数据统计分析以及监督和无监督学习所涉及的概念。那些可能没有所需的必要数学技能或者必须唤醒他们遥远的记忆来重新记起那些奇怪的公式或有趣的符号的人，请不要沮丧。如果你是一个喜欢挑战的人，附录中的入门知识可能就是你启动引擎所需要的一切。一点点忍耐就能让你坚持下去！对于那些从未接触过机器学习的人来说，第1章就是为你和需要复习的人写的。这就是你的初学者工具包，先跳进去，然后找出它的全部内容。你可以尽可能地使用在线资源来扩充你的基础知识。最后，对于那些对Java没有感觉的人，有一个秘密是：本书中描述的许多工具都有强大的GUI（图形用户界面）。有一些包括类似向导的界面，使得它们可以非常易于使用，并且不需要任何Java知识。所以如果你是Java新手，只需跳过需要编码的例子，学习使用基于GUI的工具！

读者对象

本书的主要读者对象是负责处理数据的专业人士，其职责可能包括数据分析、数据可视化或转换、机器学习模型的训练、验证、测试和评估。大体是使用Java或基于Java的工具执行预测、描述或规范分析。Java的选择可能意味着个人偏好，也可能意味着以前有Java编程经验。另一方面，工作环境或公司政策也许限制了第三方工具的使用，所以只能使用Java和其他几种语言编写的工具。在第二种情况下，预期的读者可能没有Java编程经验。本书对待这类读者就像对待他们的同事——Java专家（最先提出策略的人）一样公平。

第二类读者可以通过具有两个属性的一类形象来定义：对机器学习具有求知欲的读者和对概念、实践技术和工具综合有期望的读者。这类读者可以选择略过数学和工具介绍，专注于学习最常见的监督和无监督机器学习算法。另一个可行建议是略读第1章、第2章、第3章和第7章，跳过所有其他的部分，然后直接阅读工具部分。如果你想快速分析客户所说的随时会出现的数据集，并给客户一个令人满意的分析，这是一种非常有建设性的办法。重要的是，通过重现本书中的实验所得到的一些实践经验，会让你提出只有大师才会问的正确问题！或者，你也许希望使用本书作为参考，以快速查找有关AP聚类算法（仿射传播）的详细信息（第3章），或者通过简要回顾原理图来回忆LSTM架构（第7章），或者要标记在基于流学习的异常值检测中基于距离的聚类方法的优缺点列表（第5章）。本书适用于所有读者，并且每个人都会发现很多可供学习的内容。

用户支持

本书的示例源码可以从http://www.packtpub.com通过个人账号下载，也可以访问华章图书官网http://www.hzbook.com，通过注册并登录个人账号下载。

GitHub上也提供了本书的代码，网址是https://github.com/PacktPublishing/Practical-Predictive-Analytics。

ABOUT THE AUTHORS

作者简介

Uday Kamath 博士是 BAE 系统应用智能公司的首席数据科学家，专门研究可扩展机器学习，并在反洗钱 AML、金融犯罪欺诈检验、网络空间安全和生物信息学领域拥有 20 年的研究经验。Kamath 博士负责 BAE 系统应用智能公司 AI 部门核心产品的研究分析，这些产品涉及的领域有行为科学、社交网络和大数据机器学习方面。在 Kenneth De Jong 博士的指导下，他获得了乔治梅森大学的博士学位，他的论文研究聚焦于大数据和自动化序列挖掘的机器学习领域。

我要感谢我的朋友 Krishna Choppella 接受合著本书的邀请，在这漫长而令人满意的写作之旅中，他真是一个能干的搭档。

衷心感谢本书的审校者，尤其是 Samir Sahli 博士，感谢他提供的宝贵意见、建议并对各章节进行了深入的审校。感谢 Carlotta Domeniconi 教授，她的建议和意见帮助我们润色了本书的各个章节。还要感谢所有 Packt 的工作人员，尤其是 Divya Poojari、Mayur Pawanikar 和 Vivek Arora，他们帮助我们按时完成了撰写任务。这本书需要在很多方面做出个人牺牲，我要感谢我的妻子 Pratibha、我们的保姆 Evelyn 对我无条件的支持。最后，感谢我所有敬爱的老师和教授，他们不仅授业，还给我灌输了学习的乐趣。

KrishnaChoppella 在 BAE 系统应用智能公司的角色是作为解决方案架构师，构建工具和客户解决方案。他有 20 年的 Java 编程经验，主要兴趣是数据科学、函数编程和分布式计算。

ABOUT THE REVIEWERS

审校者简介

Samir Sahli 于2004 年获得法国 Nice Sophia-Antipolis 大学应用数学和信息科学的学士学位，还分别于2008 年和2013 年获得加拿大拉瓦尔大学和魁北克大学物理学（光学/光子学/图像科学）的硕士学位和博士学位。在其研究生学习期间，他曾与加拿大国防研究和发展部门（Defence Research and Development Canada，DRDC）合作研究了航拍图像的自动检测和目标识别，特别是在不可控环境和非最佳图像捕获条件情况下。自 2009 年以来，他一直在欧洲和北美洲的几家公司中担任顾问，专门从事智能、监视和侦察（Intelligence，Surveillance，and Reconnaissance，ISR）以及遥感领域的工作。

Sahli 博士于 2013 年作为博士后研究员加入 McMaster 光子学研究院（McMaster Biophotonics）。他的研究领域是光学、图像处理和机器学习。期间参与了几个项目，如新一代胃肠道成像装置的研发，用于个体化放疗的皮肤红斑高光谱成像研发，使用荧光寿命成像显微技术与多光子显微技术，自动检测癌症前期 Barrett 食道细胞的研发。

Sahli 博士于 2015 年加入 BAE 系统应用智能公司。他此后作为一名数据科学家，利用机器学习、统计和社会网络分析工具开发分析模型，以检测针对保险、银行和政府客户的复杂的欺诈模式和洗钱计划。

Prashant Verma 于 2011 年在电信领域的爱立信（Ericsson）公司，作为 Java 开发者开始了他的 IT 职业生涯。在拥有数年 Java EE 经验后，他转向了大数据领域，使用过几乎所有的大数据流行技术，如 Hadoop、Spark、Flume、Mongo、Cassandra 等，也使用过 Scala。他在 QA Infotech 公司作为首席数据工程师，负责通过分析和机器学习解决电子学习问题。

Prashant 先后在多家公司任职，如 Ericsson 和 QA Infotech，拥有电信和电子学习的领域知识。他在业余时间也从事过自由职业顾问。

我想感谢 Packt 出版社给了我审校本书的机会，同时感谢我的老板和家人，当我忙于此书时，他们付出了足够的耐心。

CONTENTS

目　录