异步图书
www.epubit.com.cn

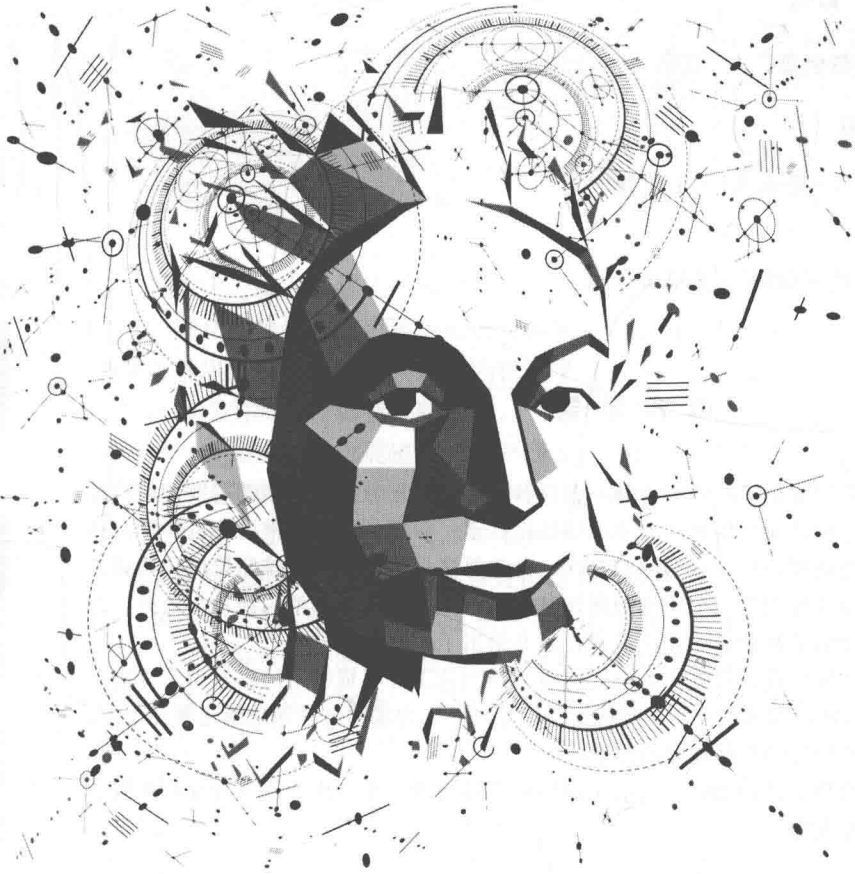
“十三五”
国家重点图书出版规划项目

Practical Machine Learning
实用机器学习

孙亮 黄倩 / 著

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

“十三五”
国家重点图书出版规划项目



Practical Machine Learning
实用机器学习

孙亮 黄倩 / 著

人民邮电出版社
北 京

图书在版编目（CIP）数据

实用机器学习 / 孙亮，黄倩著. -- 北京 : 人民邮电出版社，2017.5
ISBN 978-7-115-44646-6

Ⅰ. ①实… Ⅱ. ①孙… ②黄… Ⅲ. ①机器学习
Ⅳ. ①TP181

中国版本图书馆CIP数据核字(2017)第027041号

内 容 提 要

大数据时代为机器学习的应用提供了广阔的空间，各行各业涉及数据分析的工作都需要使用机器学习算法。本书围绕实际数据分析的流程展开，着重介绍数据探索、数据预处理和常用的机器学习算法模型。本书从解决实际问题的角度出发，介绍回归算法、分类算法、推荐算法、排序算法和集成学习算法。在介绍每种机器学习算法模型时，书中不但阐述基本原理，而且讨论模型的评价与选择。为方便读者学习各种算法，本书介绍了R语言中相应的软件包并给出了示例程序。

本书的最大特色就是贴近工程实践。首先，本书仅侧重介绍当前工业界最常用的机器学习算法，而不追求知识内容的覆盖面；其次，本书在介绍每类机器学习算法时，力求通俗易懂地阐述算法思想，而不追求理论的深度，让读者借助代码获得直观的体验。

本书适合需要应用机器学习算法解决实际问题的工程技术人员阅读，也可作为相关专业高年级本科生或研究生的入门教材或课外读物。

◆ 著　　孙 亮 黄 倩
责任编辑　杨海玲
责任印制　焦志炜

◆ 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
北京市艺辉印刷有限公司印刷

◆ 开本：800×1000　1/16
印张：22
字数：490 千字　　2017 年 5 月第 1 版
印数：1 – 4 000 册　　2017 年 5 月北京第 1 次印刷

定价：79.00 元

读者服务热线：(010)81055410　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广字第 8052 号

序一

机器学习是人工智能领域最成功、发展最迅速的分支之一。然而多年来，除了人机对弈之外，成功的机器学习应用对普通百姓来说似乎始终显得遥不可及。大数据时代的来临对机器学习提出了新的挑战，同时也为机器学习打开了一扇通向实用化舞台的大门。近年来，源于实际需求的海量数据集不断出现，一方面为技术交流和算法验证提供了有力的基础数据支撑，另一方面也有助于促进相关科学研究更贴近我们的日常生活。例如，在滴滴，我们正致力于结合机器学习技术和大数据技术来预测各地用户需求，并进行实时的运力调度和订单分配，不断为广大司机和用户提供更满意的服务。可以说，机器学习技术和大数据技术的结合正在显著地拉近科研人员和普通百姓的距离，正在显著地改变我们的生产方式和生活方式。

2016 年 8 月，我在中国人工智能大会上担任“机器学习的明天”专题论坛联席主席时曾指出，人类很多难题的解决都离不开大数据和人工智能的结合。今天，我欣喜地看到，本书两位作者的合作恰恰体现了实用机器学习的这一重要发展趋势。两位作者长期在学术界和工业界工作，其中孙亮博士是我的学生，他从 2006 年开始就一直从事降维算法、稀疏学习、数值最优化算法等方向的机器学习研究，其快速降维算法研究曾获得机器学习领域顶级会议 SIGKDD 的最佳论文提名奖；黄倩博士是高文院士的学生，他从 2004 年开始就一直从事视频大数据的压缩和增强显示研究，其研究成果曾在视频处理领域旗舰期刊 IEEE T-CSVT 发表并在工业界获得实际应用。两位作者博士毕业后均有在万人以上规模企业的相关研发经历，接触过各类实际问题和数据，对机器学习如何在实际中应用有着深刻的理解和成功的经验。

这本书不厚，但却覆盖了用机器学习技术解决实际问题的主要步骤和常用算法。两位作者从实际应用出发，介绍了数据探索、数据预处理、算法应用、性能评价等具体内容，并深入浅出地介绍了模型复杂度、损失函数等机器学习领域的基本概念。由于集成学习在实际中应用较为广泛，因此本书专列一章加以讨论。考虑到实践中大家更关注的是如何选择和使用算法，两位作者还使用 R 语言软件包来引导读者实际操作。与市面上对机器学习作一般性介绍的书籍相比，本书介绍的算法稍稍复杂一些，但也更加实用，书中讨论的内容正是实际应用机器学习解决问题时所需要掌握的内容。对于广大业界爱好者和相关专业研究生来说，这是一本理想的入门读物和参考书，因此我非常乐意向大家推荐本书。

叶杰平
滴滴研究院副院长，密歇根大学终身教授
2016 年 12 月

序二

2016 年 10 月 3 日，我从宿州匆匆赶回上海，试图从连日飨宴中恢复精力，等待 5 日凌晨谷歌公司的 Pixel 发布会。业界预测这将是一个划时代的发布会，标志着“Mobile First”向“AI First”的转型。这一场发布会，让我难掩激动地在朋友圈中留下了洋洋洒洒但未必成熟的千字技术点评。

我为什么会如此期待这场盛会，要从十多年前说起。那时我还是一个学生，学着和 IT 八竿子打不着的天文学。在本科结束后，响应高中读《时间简史》时内心深处的召唤，我继续学了天文学。经历短暂的欢愉之后，我终于意识到这条路的艰辛，我是不可能做这行了。我开始思考，有没有一条路能让我既接近最前沿的理论，又有足够好的实践环境呢？我在看了《AI》（《人工智能》）这部电影之后被震撼得天旋地转——人工智能可以创造出如此的美好，人工智能可以引发无穷的深思。

我开始去选修《机器人》的课程，开始去听张学工老师的《模式识别》课，开始把自己的知识往这个方向靠。所幸因为玻尔兹曼，这两个专业还是有些联系的。在一次选假期小班课程的时候，我选了一个 SVM 的讲座，后来当孙亮和我说他也听过这堂课的时候，我才意识到原来高人就在身边，只是我一直不曾留意。我曾看到过孙亮床位下一箱一箱的计算机书，大多与机器学习有关，只是人的神经系统是很奇妙的，当你不关注的时候，这些事情会自动在你的世界被屏蔽；当你关注的时候，这个概念会在极短的时间以各种方式在你面前展现。不久后我认识了同样以机器学习见长的黄倩，那时在周末大家都会找时间一起聚聚，当然话题大多是关于计算机的发展方向和算法的，然后我就作为旁听者被他们带进了门。

孙亮硕士毕业后去美国继续读机器学习的博士，最后到微软公司从事他擅长的机器学习和数据科学的研究；黄倩在高文院士指导下硕博毕业，最终回到高校任教，带领学子探索最前沿的算法。这些年他们一直躬耕一线，从未中断，我也在辗转中断断续续地做着数据相关的工作，最终转到 Hadoop/Spark/TensorFlow 开源平台。这几年大数据如火如荼，机器学习热浪汹涌，AlphaGo 的成功进一步激励了业界，TensorFlow 的开源、“AI First”的概念终于在坎坎坷坷 10 年之后开始盛行，高校的学子和一线的工程师开始被这个全新的世界吸引。同事和实习生不断地要我给他们推荐一些大数据和机器学习方面的书，我也曾给同事买过一些我看过的书。遗憾的是，这些书大多要么是纯理论的“屠龙术”，离实际应用还有一段距离，要么就是针对算法模块搞些例子，使学习者只知其然不知其所以然，还有的就是与目前业界普遍应用的算法不吻合。

对于一线的 IT 从业者和想要实践算法的学生来说，一本理论与实践相结合、能展现目前业界通用算法使用技巧的书，应该是大家最喜闻乐见的。这要求作者要有极深厚的理论功底，

能够把算法娓娓道来，还需要有足够的实践经验，娴熟业界各领域现在通用的算法。《实用机器学习》正是这样的一部佳作。本书详细讨论了机器学习中回归、分类、推荐、排序 4 类经典问题，详述了每一类问题中常用算法的理论来源，以及在 R 环境中如何去使用、评测和可视化展现。作为一线的实践者，书中对数据预处理也做了独立讲解，就理论过渡、全面性、实用性、易读性来说，本书都做了充分的考量。R 语言作为一个机器学习的平台工具，具有使用简单、分析方便、可用库完备以及可视化容易等特点，为进行问题分析和寻找原因提供了足够的便捷性。即使部分算法没有分布式的解决方案，本书讲授的实用机器学习算法，也极易在 Mahout 或 Spark MLLib 上平移对应的接口。在算法和实践平台上，本书倡导的方法和环境，几乎都可以和工业界做到无缝对接。

人工智能的世界来了，浩浩汤汤，开启了 IT 业者的另外一个世界。或许，我们只是需要一个开始，而《实用机器学习》来得适逢其时，你的一小步将来难说不是人类机器学习世界的一大步！

Fantasy（裴少芳）
威比网络科技（上海）有限公司大数据部总监
2016 年 12 月于上海

前　言

本书侧重于数据分析和机器学习的实践，涉及从原始数据搜集到建立模型解决问题再到算法性能评估的全过程。书中主要介绍实践中最常用的 4 类算法，包括回归算法、分类算法、推荐算法和排序算法。此外，书中还会介绍集成学习。集成学习是一类通过综合多个模型取长补短以取得更好效果的方法，对于回归、分类、推荐和排序问题都适用。在实践中，充分掌握这 4 类算法和集成学习即可解决相当多的实际问题。由于篇幅所限，聚类分析、关联规则等其他相关内容书中并没有一一介绍。

对于每种算法，本书首先介绍算法的原理。在理解算法原理和算法优缺点的基础上，读者在实践中就可以根据数据的特点和问题的具体需求选用合适的算法。为了突出算法的实践性，本书使用 R 语言中的软件包来介绍机器学习算法，特别是介绍了如何使用各种算法。R 语言是一种开源和免费的解释型语言，其最大的优点是提供了各种软件包，实现了各种不同的算法。机器学习中很多强大的算法在 R 中都有相应的程序包。我们在讲解各种机器学习算法时，都介绍了 R 中相应的软件包，并提供了相应的 R 程序来帮助读者学习这些软件包的使用。这样读者就可以通过 R 来直接使用相应的算法，获得数据分析的第一手建模经验。

除了介绍这 4 类机器学习算法之外，本书涵盖了使用机器学习解决实际问题的整个流程，包括数据探索、数据预处理、使用机器学习算法所构建的模型的评价和选择等。在实际使用机器学习处理数据的过程中，数据的探索和预处理是非常重要的步骤，在很多场合甚至比建立模型本身更加重要，从原始数据中提取出一个好的特征在很多时候能够显著地提高模型的性能。得到构建的模型后，我们还需要评价和选择模型。本书还会介绍不同类型算法对应的评价标准以及如何进行模型选择，并介绍 R 中的相关工具（如 caret 包），以帮助读者直接上手。

我们尽量使用简单通俗的语言来介绍机器学习中的基本概念和各种常用算法，并通过介绍 R 中对应的软件包来帮助读者迅速了解和掌握各种算法的使用。为了准确地介绍各类算法，不可避免地要用到一些数学知识，本书在第 3 章特别介绍了一些相关的数学知识。

本书的所有 R 代码（包括生成书中图的大部分 R 代码）都可以从人民邮电出版社异步社区（www.epubit.com.cn）网站上获得。

本书的出版得到了国家自然科学基金（61300122、61502145）的支持，得到了人民邮电出版社编辑杨海玲女士的支持和帮助，在此表示诚挚的谢意。成稿的关键时期适逢我们各自的女儿降生，在此衷心感谢双方家人的理解与支持。因水平和时间所限，书中难免有错误或不当之处，恳请广大读者不吝指正。读者若有任何问题或建议，可发送电子邮件至 sun.liang@outlook.com 或 huangqian@gmail.com。

孙　亮　黄　倩

2016 年 12 月分别于华盛顿雷德蒙和南京

目　录

第1章 引论

随着计算机和互联网越来越深入到生活中的方方面面，人们搜集到的数据也呈指数级的增长。在这种情况下，*大数据*（big data）应运而生。大数据通常体量特别大，而且数据比较复杂，使得无法直接使用传统的数据库工具对其进行存储和管理。大数据带来了很多挑战，如数据的搜集、整理、存储、共享、分析和可视化等。广义的大数据处理涵盖了上述所有领域；狭义的大数据更多是指如何使用*机器学习*来分析大数据，从海量的数据中分析出有用的信息。

大数据分析的核心是机器学习算法。很多时候，我们有足够的数据，但是对如何利用这些数据缺乏理解。同时，实际问题往往比较复杂，并不能直接套用机器学习算法，我们需要对实际问题进行一些转化，使得机器学习算法可以应用。虽然实际问题表现形式各异，但是在将它们转化为机器学习能够处理的问题时，一般转化为如下4类问题：（1）回归问题；（2）分类问题；（3）推荐问题；（4）排序问题。这4类问题是实际应用中最主要的类型，覆盖了大部分实际问题。在1.3节，我们将详细介绍每类问题的具体例子。

1.1 什么是机器学习

机器学习（machine learning）是计算机科学的一个分支，也可以认为是*模式识别*（pattern recognition）、*人工智能*（artificial intelligence）、*统计学*（statistics）、*数据挖掘*（data mining）等多个学科的交叉学科。机器学习与*数值优化*（numerical optimization）也有很高的重合度。

机器学习研究如何从数据中学习出有效的模型，进而能对未来作出预测。例如，如果商店能够预测某一件商品在未来一段时间的销售量，就可以提前预订相应数量的商品，这样既可以避免缺货，又可以避免进太多货而造成积压。与传统的决策算法不同的是，机器学习算法依赖于数据。在前面的例子中，我们要从历史数据中学习出相应的模型以对未来进行预测。这样做有两个好处：第一，由于算法依赖于数据，可以使用新的数据来不停地更新模型，使得模型能够自适应地处理新的数据；第二，对人的介入要求少。在使用机器学习的过程中，虽然也会尽量利用人的经验，但更多地强调如何利用人的经验知识从数据中训练得到更好的模型。

目前，机器学习已成为研究和应用的热点之一。一些能够使用机器学习解决的实际问题包括：

- 根据信用卡交易的历史数据，判定哪些交易是欺诈交易；
- 从字母、数字或者汉字图像中有效地识别出相应的字符；

- 根据用户以往的购物历史来给用户推荐新的商品；
- 根据用户当前的查询和以往的消费历史向其推荐适合的网页、商品等；
- 根据汽车的发动机排量、年份、类型、重量等信息估计汽车的耗油量。

虽然这些问题的具体形式不同，但是均可转化成机器学习可以解答的问题形式。

从概念上讲，在机器学习中，我们的目标是从给定的数据集中学习出一个模型 f，使得它能够有效地从输入数据中预测我们感兴趣的量。根据问题的不同，我们感兴趣的量（或者叫目标值）可以有不同的形式。例如，在分类问题中，目标值就是若干类别之一；在排序问题中，目标值就是关于文档的一个序列。

在机器学习中，通常我们解决问题的流程如下：

（1）搜集足够多的数据；

（2）通过分析问题本身或者分析数据，我们认为模型 f 是可以从数据中学习出来的；

（3）选择合适的模型和算法，从数据中学习出模型 f；

（4）评价模型 f，并将其利用在实际中处理新的数据。

在实际中，还需要根据应用的实际情况及时更新模型 f。例如，若数据发生了显著变化，则需要更新模型 f。因此，在实际部署机器学习模型时，上面的第 3 步和第 4 步是一个循环反复的过程。

一个经常与机器学习同时提起的相关领域是*数据挖掘*（data mining）。数据挖掘和机器学习在很多时候都被（不严格地）混用，因为这两者有很多重叠的地方。传统意义上，机器学习更加注重于算法和理论方面，而数据挖掘更加注重实践方面。数据挖掘中的很多算法都来自于机器学习或者相关领域，少数来自于数据挖掘领域，如*关联规则*（association rule）。

另一个与机器学习关联很深的领域是统计学。在统计学中，我们学习了很多传统的处理数据的方法，包括数据统计量的计算、模型的参数估计、假设检验等。但在实际问题中，很多情况下我们并不能直接使用统计学中的方法来解决问题。一方面，随着数据规模的扩大，统计学中很多传统的数据分析方法需要通过大量的计算才能得到结果，时效性不高；另一方面，传统的统计学方法更多地考虑了算法在数学上的性质，而忽略了如何在实际中更好地应用这些算法。

1.2　机器学习算法的分类

在机器学习中，常用的算法可以分为监督型学习（supervised learning）和非监督型学习（unsupervised learning）[①]。

- 在监督型学习中，除了输入数据 x 外，我们还知道对应的输出 y。我们的目标是构建一个函数 $f(x)$，使得 $f(x)$ 能够预测输出 y。

① 在很多资料中还有第三类称为强化学习（reinforcement learning），近年来还有半监督型学习（semi-supervised learning）提出。本书主要涉及监督型学习和非监督型学习，不讨论强化学习和半监督型学习。

- 在非监督型学习中，我们只有输入数据 x，没有对应的输出 y。我们的目标是从数据中学习数据本身存在的模式（pattern）。例如，聚类分析（cluster analysis）就是一个非监督型学习的典型例子，它通过分析样本之间的相似度来将样本划分为几个不同的聚类。

在监督型学习中，输出 y 一般称为目标变量（target variable）或者因变量（dependent variable），而输入 x 称为解释变量（explanatory variable）或者自变量（independent variable）。

在实际中，在条件允许的情况下，我们偏好监督型学习。因为我们知道相应的目标变量的值，所以能够更加准确地构建模型，取得更好的效果。对于非监督型学习，在实际中，我们可以直接将其结果作为输出，但更多地是将其结果作为新的特征，再应用到监督型学习的算法中。例如，对于一组数据，可以先使用 k 均值算法对数据进行聚类分析，然后将聚类分析的结果作为新的特征。本书将主要讨论监督型学习。

在监督型学习中，一般将整个数据集分为训练集（training set）和测试集（test set）。利用训练集中的数据，可以构建相应的模型（model）或者学习器（learner）。利用测试集，可以估计所构建模型的性能高低。在数据集中，我们使用样本（sample）、数据点（data point）或实例（instance）来称呼其中的每个点。监督型学习可以进一步分为回归问题、分类问题等。我们将在 1.3 节利用具体的例子来介绍监督型学习。

1.3 实际应用

在本节中，我们将会介绍一些可用机器学习解决的实际问题，包括病人住院时间预测、信用分数估计、Netflix 上的影片推荐和酒店推荐。每个例子都对应一类不同的机器学习问题。通过这些不同类型的机器学习问题，读者对机器学习可以有更多直观的感受。

1.3.1 病人住院时间预测

机器学习在医疗行业有着广泛的应用。我们以 Heritage Health Prize[①]竞赛作为例子以说明如何使用机器学习来预测病人未来的住院时间。

在美国每年都有超过 7000 万人次住院。根据相关统计，2006 年在护理病人住院上所花的无关费用就已经超过了 300 亿美元。如果我们能够根据病人的病历提前预测病人将来的住院时间，那么就可以根据病人的具体情况提前做好相关准备从而减少那些无谓的开销。同时，医院可以提前向病人发出预警，这样就能在降低医疗成本的同时提高服务质量。在从 2011 年开始的 Heritage Health Prize 竞赛（HHP）中，竞争者成功地使用机器学习的方法，由病人的历史记录预测了病人在未来一年的住院时间。图 1-1 显示了竞赛中使用的病历数据的一部分样本。

① http://www.heritagehealthprize.com/c/hhp

	MemberID	ProviderID	Vendor	PCP	Year	Specialty	PlaceSvc	PayDelay	DSFS	PCG	CharlsonIndex	ProcedureGroup	SupLOS
1	42286978	8013252	172193	3779	Y1	Surgery	Office	26	5- 9 months	NEURMENT	0	MED	0
2	97983248	3810650	726296	9300	Y3	Internal	Office	50	5- 9 months	NEURMENT	0	MED	1
3	2759427	2997752	140343	91952	Y3	Internal	Office	14	0- 1 month	METABS	0	EM	1
4	73970939	7953394	240043	70110	Y1	Laboratory	Independent Lab	24	0- 1 month	METABS	0	SURG	0
5	11837054	7557061	490247	68062	Y2	Surgery	Outpatient Hospital	27	4- 5 months	FXNTSLC	1-2	EM	1
6	45844561	1963486	4042	59031	Y3	Pediatrics	Office	25	3- 4 months	NEURMENT	0	EM	0

图 1-1　病历数据示例

1.3.2　信用分数估计

在现实生活中，向银行申请贷款是比较常见的，如房屋贷款、汽车贷款等。银行在办理个人贷款业务时，会根据申请人的经济情况来估计申请人的还款能力，并根据不同还款能力确定安全的借款金额和相应的条款（如不同的利率）。在美国，每个成年人都有相应的信用分数（credit score），用来衡量和评估借款者的还款能力和风险。

在估计申请者的还款能力时，需要搜集用户的多个方面的信息，包括：

- 收入情况；
- 年龄、性别；
- 职业；
- 家庭情况，如子女数量等；
- 还款历史，包括未按时还款的记录、还款金额等；
- 现有的各种贷款和欠款情况等。

如何将这些因素综合考虑从而决定借贷者的信用分数呢？直观地讲，可以使用一些简单的规则来确定信用分数。例如，某申请者的当前借款金额很高但收入一定，则进一步借款的风险很高，信用分数将会较低；又如，某申请者的某张信用卡在过去经常没有按时还款，则其信用分数也会较低。虽然使用简单的规则能够大致解决信用分数估计的问题，但是这个办法最大的问题是不能自适应地处理大量数据。随着时间的变化，申请者不还款的风险模型可能会发生变化，因此，相应的规则也需要修改。

银行通常可以得到海量的申请者数据和对应的历史数据。利用机器学习的方法，我们希望可以从这些申请者过去的还款记录中自适应地学习出相应的模型，从而能够“智能”地计算申请者的信用分数以了解贷款的风险。具体地讲，在机器学习模型中，将申请者的信息作为输入，我们可以计算申请者在未来能够按时还款的概率。作为一个典型的例子，FICO 分数[①]就是美国 FICO 公司利用机器学习模型开发出来的一个信用分数模型。

1.3.3　Netflix 上的影片推荐

Netflix 是美国的一家网络视频点播公司，成立于 1997 年，到 2015 年该公司已经有了近 7000 万的订阅者，并且在世界上超过 40 个国家或地区提供服务。Netflix 上的一项很重要的功能是根据用户的历史观看信息和喜好推荐相应的影片，如图 1-2 所示。2006 年 10 月至 2009

① http://www.myfico.com/CreditEducation/WhatsInYourScore.aspx

年 9 月，Netflix 公司举办了 Netflix Prize[①]比赛，要求参赛者根据用户对于一些电影的评价（1 星～5 星），推测用户对另外一些没有看过电影的评价。如果能够准确地预测用户对于那些没有看过的电影的评价，就可以相应地向这些用户推荐他们感兴趣的电影，从而显著提高推荐系统的性能和 Netflix 公司的盈利水平。

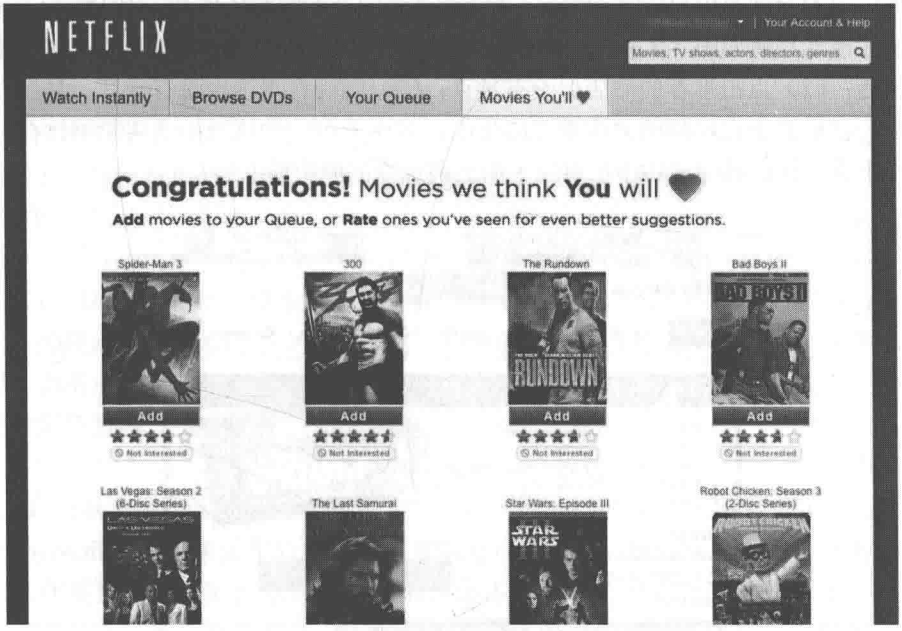


图 1-2 Netflix 上的电影推荐

在 Netflix Prize 比赛中，获胜的标准是将 Netflix 现有推荐系统的性能提高 10%。在 2009 年，BellKor's Pragmatic Chaos 队赢得了比赛。其主要方法是基于矩阵分解的推荐算法，并使用集成学习的方法综合了多种模型。Netflix Prize 比赛显著地推动了推荐算法的研究，特别是基于矩阵分解的推荐算法的研究。在本书中，我们也将详细介绍这些推荐算法。

1.3.4 酒店推荐

Expedia 是目前世界上最大的在线旅行代理（online travel agency，OTA）之一。它的一项很重要的业务是向用户提供酒店预订，作为用户和大量酒店之间的桥梁。对于用户的每个查询，Expedia 需要根据用户的喜好，提供最优的排序结果，这样用户能够方便地从中选出最合适的酒店。

Expedia 于 2013 年年底与国际数据挖掘大会（International Conference on Data Mining，ICDM）联合举办了酒店推荐比赛。在该项比赛中，Expedia 提供了实际数据，包括用户的查询以及其对所推荐结果点击或者购买的记录。在进行酒店推荐时，Expedia 考虑了如下因素：

① http://www.netflixprize.com/

- 用户的位置和酒店的位置；
- 酒店的特征，如酒店的价格、星级、位置吸引程度等；
- 用户过去预订酒店的历史，包括价格、酒店类型、酒店星级；
- 其他竞争对手的信息。

根据用户的查询及用户的背景信息，Expedia 返回推荐的酒店序列。在 Expedia.com 上，典型的酒店搜索界面如图 1-3 所示。根据返回的推荐结果，用户有 3 种选择：（1）付款预定推荐的酒店；（2）点击推荐的酒店但没有预订；（3）既没有点击也没有预订。显然，根据用户的反应，我们希望在理想的酒店推荐结果中，对应于第一种选择的酒店能够排在最前面，并且对应于第二种选择的酒店排在对应于第三种选择的酒店前面。

图 1-3　在 Expedia.com 上搜索酒店

1.3.5　讨论

上文中的 4 个例子分别对应于机器学习中的 4 类典型问题：

- 回归（regression）；
- 分类（classification）；
- 推荐（recommendation）；
- 排序（ranking）。

在第一类问题中，首先需要为每个病人构建一个特征向量 $\boldsymbol{x}$，然后构建一个函数 f，使得可以用 $f(\boldsymbol{x})$来预测病人的住院时间 y。注意，这里要预测的量（病人的住院时间 y）的范围是 0～

365（或者 366），我们可以将其转化为回归问题。在回归问题中，目标变量是一个连续值。

在第二类问题中，需要为每个申请者构建一个特征向量 $\boldsymbol{x}$，而输出 y 是 0 或者 1，代表批准贷款或者不批准贷款。事实上，输出 y 也可以是批准的概率。这是机器学习中典型的分类问题。在分类问题中，目标变量 y 是一个离散变量。与回归问题类似，我们的目标是构建一个函数 f，使得 $f(\boldsymbol{x})$可以预测真实的 y。在典型的两类分类（binary classification）问题中，目标变量的取值为 0 或者 1（有时是−1 或者 1）。在多类分类（multi-class classification）问题中，我们有多个类，而目标变量的取值是其中之一。

在第三类问题中，需要根据用户过去的历史为每个用户推荐相应的商品，这是一个典型的推荐问题。与回归和分类问题相比，我们需要为每个用户返回一个感兴趣的商品序列。

在第四类问题中，需要根据用户的输入（在上文的例子中是用户对于酒店的查询），从一系列对象（在这个例子中是酒店）中根据用户的需要返回一个对象的序列，使得该序列最前面的对象是用户最想要的。这类问题称为排序（ranking）问题。同前面的回归问题和分类问题相比，排序问题需要考虑整个返回序列。与前面的影片推荐例子相比，在排序问题中我们需要明确的用户输入，而在影片推荐中我们只是根据用户过去的历史信息来进行推荐，用户没有进行明确的输入。

在实际应用中，机器学习的应用远远超出上面的几个例子。例如，近期非常热门的 AlphaGo，谷歌公司在其中使用了深度学习（deep learning）来学习围棋对弈；德国的蒂森克虏伯（ThyssenKrupp）集团作为电梯的主要制造商之一，应用机器学习来预测电梯发生故障的时间从而提前维修，降低电梯的综合运营成本；美国的很多大型零售商在开设新店时，都要搜集各个地区的各种信息和历史销售数据，通过建立机器学习模型的形式选择最优的店址。

1.4 本书概述

本书主要从解决实际问题的角度来介绍常用的机器学习算法。在 1.3 节中我们讨论了机器学习中常见的 4 类典型问题，基本上覆盖了目前实际中可以使用机器学习算法来解决的主要问题类型。在本书中，我们将主要讨论对应的 4 类算法，包括：

- 回归算法；
- 分类算法；
- 推荐算法；
- 排序算法。

其中回归算法和分类算法是两类最常用的算法，也是其他很多算法的基础，因此我们首先予以介绍。推荐系统在目前有了越来越多的应用，而排序算法在搜索引擎等领域也获得了广泛的应用，因此我们也会对常用的推荐算法和排序算法进行介绍。

在上面的 4 个例子中，我们可以构建多个不同的模型，希望它们之间能够取长补短，使