中国高校
创意创新创业
教育系列丛书

范淼 李超 编著

Python
机器学习及实践

——从零开始通往Kaggle竞赛之路

清华大学出版社

中国高校创意创新创业教育系列丛书

Python 机器学习及实践

——从零开始通往Kaggle竞赛之路

范淼 李超 编著

清华大学出版社
北京

内 容 简 介

本书面向所有对机器学习与数据挖掘的实践及竞赛感兴趣的读者，从零开始，以 Python 编程语言为基础，在不涉及大量数学模型与复杂编程知识的前提下，逐步带领读者熟悉并且掌握当下最流行的机器学习、数据挖掘与自然语言处理工具，如 Scikit-learn、NLTK、Pandas、gensim、XGBoost、Google Tensorflow 等。

全书共分 4 章。第 1 章简介篇，介绍机器学习概念与 Python 编程知识；第 2 章基础篇，讲述如何使用 Scikit-learn 作为基础机器学习工具；第 3 章进阶篇，涉及怎样借助高级技术或者模型进一步提升既有机器学习系统的性能；第 4 章竞赛篇，以 Kaggle 平台为对象，帮助读者一步步使用本书介绍过的模型和技巧，完成三项具有代表性的竞赛任务。

图书在版编目(CIP)数据

Python 机器学习及实践——从零开始通往 Kaggle 竞赛之路/范淼，李超编著．--北京：清华大学出版社，2016
(中国高校创意创新创业教育系列丛书)
ISBN 978-7-302-44287-5

Ⅰ．①P…　Ⅱ．①范…　②李…　Ⅲ．①软件工具－程序设计　Ⅳ．①TP311.56

中国版本图书馆 CIP 数据核字(2016)第 164306 号

责任编辑：谢　琛
封面设计：常雪影
责任校对：李建庄
责任印制：王静怡

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 刷 者：清华大学印刷厂
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：210mm×235mm　　**印　张**：12.5　　**彩　插**：4　　**字　数**：274 千字
版　　次：2016 年 10 月第 1 版　　**印　次**：2016 年 10 月第1次印刷
印　　数：1～3000
定　　价：49.00 元

产品编号：069392-01

编委会名单

丛书总顾问

顾国彪　中国工程院院士，中国科学院电工研究所研究员、博士生导师

丛书顾问（按姓名拼音排序）

陈雪涛　麟玺创业投资管理有限公司总裁、北京创客空间科技有限公司副董事长
党德鹏　北京师范大学计算机系系主任，信息学院教授、博士生导师
黄　英　中关村科技园区海淀园管理委员会副主任、新闻发言人
李　超　清华大学信研院 Web 与软件研究中心副主任、副研究员
李　涛　Geek2Startup 联合创始人、曾任 CSDN 移动业务拓展总监
李卫平　知金教育咨询有限公司总裁、北京理工大学兼职教授、2016 中国互联网教育领军人物
刘峰峰　富士康工业工程学院华北分院院长、富士康廊坊厂区制造负责人
龙　林　亚都北京科技有限公司总裁
沈　拓　清华大学 x-lab 未来生活中心创始人、互联网＋研究院创始人
苏　茹　车库咖啡创始人、U＋联合创始人
陶　锋　清华大学 x-lab 互联网与信息技术创新中心执行主管兼培育顾问
滕桂法　河北省高等院校计算机教育研究会理事长
宋述强　《现代教育技术》杂志副主编、清华大学创客教育实验室 Co-director
宋跃武　创新知识体系创始人、海淀区创业园企业家训练营总裁实战导师
王　津　可穿戴计算产业联盟云计算大数据负责人、中民社会捐助发展中心副主任
王丽华　北京航空航天大学软件学院副院长、教授
王卫宁　中国人工智能学会秘书长
王　霞　清华大数据产业联合会秘书长
文　辉　清华阳光科技有限公司总裁
向　东　清华大学机械工程系副教授、博士生导师，中国绿色制造技术标准化技术委员会委员兼副秘书长，中国机械制造工艺协会常务理事

谢　琛　清华大学出版社资深策划人
谢将相　易创互联创始人、易创学院创始人
邢春晓　清华大学信研院副院长、教授、博士生导师
熊　斌　中国科学院电工研究所蒸发冷却技术研究发展中心副研究员
胥克谦　中国教育技术协会教育游戏专业委员会常务理事、皮影客创始人
杨士强　清华大学计算机系教授、博士生导师
张有明　顶你学堂创始人、中国高科股份有限公司教育事业部副总经理
赵　龙　创业沙拉联合创始人
赵　鑫　精一天使公社合伙人
郑　莉　清华大学计算机系教授、教育部教育信息化技术标准委员会专家兼秘书长
祝智庭　华东师范大学终身教授、教育技术学博士生导师

出版说明

在产业升级急需、区域发展呼吁、国家政策引导、社会舆论支持、成功者亲身鼓励等多方的推动下，“大众创新、万众创业”已成为一股年轻人普遍关注和参与的热潮。“大众创业、万众创新”作为创新驱动发展战略的最主要实施方案，有效改善了我国就业困难、行业升级效率低下等局面。

变化、颠覆、创新早已成为我们这个时代的主旋律，这是我们酝酿这套丛书的背景。

创新的关键在于人才培养。2015 年 11 月底，教育部下发了《教育部关于做好 2016 届全国普通高等学校毕业生就业创业工作的通知》，通知规定，从 2016 年起所有高校都要设置创意创新创业教育课程，对全体学生开发开设创意创新创业教育（以下简称三创教育）必修课和选修课，并纳入学分管理。对有创业意愿的学生，开设创业指导及实训类课程，对已经开展创业实践的学生，开展企业经营管理类培训。各地各高校要配齐配强创意创新创业教育专职教师，建立以课堂教学为主渠道，讲座、论坛、培训为补充的多形式就业指导课程体系。

强调突破和实践的三创教育正式成为国家创新驱动战略及人才培养计划的重要组成部分，这是我们策划这套丛书的契机。

中关村汇聚了中国最顶级高校，无论是知识、技术、人才的优质及密集程度，还是知识性企业、企业专利、创业项目及创业服务机构的数量及质量，均处于全国领先的地位；可以说，中关村早已积累了深厚的创新文化和实践经验。而北京市海淀区的各个高校，也近水楼台，在师资和课程、学业评价、校企合作等多个维度对创意创新创业教育已经展开了有益的探索和尝试。这是这套丛书诞生和成长的土壤。

于是，我们立足于中关村的知识资源和创业实践，整合多方资源，广邀优质活跃的创业导师、三创组织、孵化器等，和教育出版机构，成立了“中关村融智三创丛书工作室”，希望通过非盈利机构的形式，助力高校需求驱动型人才培养。我们计划创造性地撰写一套通识性的三创丛书，以跨学科的主题学习和任务驱动形式，跨专业地服务于三创起航教育，并将同步策划基于互联网和社交媒体的一体化新型教学模式及资源服务，我们诚挚希望创意创新创业教育能在中国生长出内生力量，真正形成气候，实现繁荣，实现创新驱动。

丛书遵循教学与创业的规律进展，从有意启发学生创意及创新思维开始，引导高校学生思考为什么要创意创新创业、“我”适合什么样的行业及发展道路，直至给出具体的创业

方法论及实战指导。因此，丛书将首先策划和编写涉及三创方面的用于实现创意、支撑创新的实用前沿技术类和技能类书籍，为三创教育夯实“硬实力”。在由浅入深组织丛书撰写的同时，考虑到各个领域专业化门槛较高，市场需求、政策导向、学校师资力度不一，我们同时还兼顾了不同行业的特点，力争覆盖新兴行业及国家重点发展的领域，如互联网、新能源、汽车等。

此外，丛书还将策划另外两大类书籍：一类将涵盖创业启程相关的商业模式设计、产品营销、团队创建与维持、投融资、产品运营管理、法规遵从、知识产权策略、沟通与表达等众多技术和技能之外的方面，即服务于三创的“软实力”；另一类还将从国际国内一流高校、创客空间、孵化器及知名产业园区的创意、创新、创业及投资经历中筛选出众多生动鲜活的经典案例，提供综合全面的“好案例”。

本套丛书具有开放性强、通识全面、面向实践、行业案例新鲜丰富的特点。正如它所服务的主题，这套丛书本身就是一个三创教育的探索。翻开顾问名单会发现，顾问中既有德高望重的院士，也有富有产业园区建设经验的官员，既有创业成功的企业家，还有不断探索三创教育的学者——我们希望这套丛书成为教育家、研究者、企业家、投资方、创业者等多方合作的载体，营造一个充满活力、良性互动、可持续发展的教育生态系统，全方位地为高校教师、学生、创客空间、创新创业团队提供权威性、高品质的三创教育服务；我们也希望这套丛书的出版，不仅能够填补全国2800余所高等院校所面临的急迫巨大的教材缺口，更能为高校创新创业教育体系的建立和完善、创业实践指导、产学研转化等略尽绵薄之力。

最后，感谢海淀园管委会和清华大学出版社的领导们，他们在本套丛书的策划、撰写、编辑出版的过程中提供了大量帮助。由于创意创新创业主题宏大，瞬息万变，本套丛书难免存在疏漏不足，有待今后进一步补充和完善，恳请读者批评指正。如有读者愿意分享更精彩的理念或案例，也欢迎联系。

中关村融智三创丛书工作室

2016年9月

推荐语[①]

过去近二十年，计算机科学的发展是被大量的数据推动的。海量数据提供了认识世界的新视角，同时也带来了分析和理解数据的巨大挑战。如何从数据中获得知识，并利用这些知识帮助设计和创造更满足用户需求的产品，希望将来自新的人工智能算法。大数据的核心思想体现在整个工业流程中从决策到执行数据的重要性，其重要性的发挥依赖于现代计算方法——机器学习。机器学习可以利用数据做很多决策，这些在统计意义上都是好的决策，比如要不要把这首歌推荐给那个用户。更惊奇的是当数据足够大，计算能力足够强，机器也可以学得比人更好。清华大学范淼和李超的新著《Python 机器学习及实践》很契合实际，从零开始介绍简单的 Python 语法以及如何用 Python 语言来构建机器学习的模型。每一个章节环环相扣，配合代码样例，非常适合希望了解机器学习领域的初学者，甚至没有编程基础的学生。大数据要求机器学习应该更普及，而普及的途径则是降低相关工具的使用难度。希望看到这本新书能推动机器学习的普及。

——今日头条实验室科学家，前百度美国深度学习实验室少帅科学家　李磊

这是一本面向机器学习实践并且具有很强实用性的好书。每个章节，在简要介绍一种机器学习模型的基础上，结合具体的例子，给出了详细的 Python 程序的编程方法，有利于读者对机器学习方法细节的掌握。跟随本书，读者将一步步跨入机器学习的殿堂，掌握用机器学习方法求解实际问题的技能。本书适合于想使用机器学习方法求解实际问题的博士生、硕士生、高年级本科生，以及在企业工作的工程技术人员阅读，是一本快速掌握机器学习方法求解实际问题的入门读物，相信读者将从本书中获益匪浅。

——清华大学计算机系教授 马少平

① 按照推荐人的姓名拼音排序。

机器学习是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性能的一门学科，也是当前科研机构及企业开展应用研究的热点之一。随着“互联网+”概念在中国的提出，科研及工程技术人员迫切需要将机器学习技术与互联网技术结合起来，把互联网与机器学习技术应用到人类生活中。但机器学习作为一门技术，具有一定的门槛，如何提供一本通俗易懂、快速入门的技术书籍，让在职科技人员及在校学生能够尽快熟悉机器学习的内容，理解机器学习的含义及本质，是需要尽快解决的问题。

本书前两部分采用通俗的语言，借助于现实生活的例子及开源库包，介绍了机器学习的基本概念及开源库包的安装、使用和编程调用方法，通过实例展现了使用经典算法模型的分析过程及思考问题的方法。第三及第四部分介绍了在解决实际问题时如何通过抽取或者筛选数据特征、优化模型配置，进一步提升经典模型的性能表现，从而达到能够将机器学习的经典算法应用到解决现实问题的目的。

尽管目前市场上关于机器学习的书籍很多，但很少具有能够将开发语言及机器学习理论紧密结合，利用开源技术，采用类似“实训”方式进行实践教学的书籍。而本书的作者根据自己的学习经历及学习过程的体会，把自己的学习经验充分融入书本之中，采用由浅入深的方法，结合机器学习的内容，把算法学习的每一步都给读者以详细展现，减少了学生的学习难度，加快了学生学习的进度，是一本适合在校学生及工程技术人员在机器学习方面快速入门的指导书。

——北京邮电大学软件学院教授，教研中心主任　吴国仕

人工智能的发展日新月异，机器学习的应用如火如荼。在这个变革的时代，大众特别需要一本既能帮助读者理解机器学习理论，又能让人快速上手实践的入门级图书。这是一本侧重于 Python 机器学习具体实践与实战的入门级好书。不同于多数专业性的书籍，该书拥有更低的阅读门槛。即便不是计算机科学技术专业出身的读者，也可以跟随本书借助基本的 Python 编程，快速上手最新并且最有效的机器学习模型。作为在一线从事机器学习理论与技术的研发人员，该书的作者整合了当下数据科学所使用的最为流行的资源，如 Scikit-learn、Pandas、XGBoost 和 Tensorflow 等，一步步带领读者从零基础快速成长为一位能够独立分析数据并且参与机器学习竞赛的兴趣爱好者。同时，这本书的作者记录下大量在机器学习实践过程中的心得体会。全书深入浅出，让人在实践中获得知识。

——香港科技大学计算机与工程系讲座教授，系主任，IEEE,AAAI Fellow，
国际人工智能协会（IJCAI，AAAI）常务理事，中国人工智能协会副理事，
ACM KDD China ACM 数据挖掘委员会 中国分会主席 杨强

机器学习的每一次进步带动了很多学科的大力发展。这是一本由在读博士生撰写，侧重于 Python 机器学习具体实践和实战的入门级教科书。不同于多数专业书籍，该书的作者从初学者的视角，一步步带领读者从零基础快速成长为一位能够独立进行数据分析并且参与机器学习竞赛的兴趣爱好者。全书深入浅出，特别是有意了解机器学习，又不想被复杂的数学理论困扰的读者，可会从此书中获益。

——苏州大学计算机科学与技术学院副院长，人类语言技术研究所所长，特聘教授，国家杰出青年科学基金获得者　张民

不同于多数专业性的书籍，该书拥有更低的阅读门槛。即便不是计算机科学技术专业出身的读者，也可以跟随本书借助 Python 编程快速上手最新并且最有效的机器学习模型。如果说机器学习会主导信息产业的下一波浪潮，那么在这波浪潮来临之前，我们是否有必要对其一窥究竟。我很高兴看到有这样一本零基础实战的好书服务广大读者，为普及这一潮流尽绵薄之力。就像过去几十年间我们不懈普及计算机与互联网一样，人工智能，特别是机器学习的核心思想也应该走出象牙塔，拥抱普罗大众，尽可能让更多的兴趣爱好者参与到实践当中。

——清华大学语音和语言技术中心主任，教授　郑方

这是一本讲解利用 Python 进行机器学习实战的入门级好书。该书带领刚入门的读者，从零开始，一步步学习数据分析并掌握机器学习竞赛技能。如果你想学习机器学习方法又不想被复杂的数学理论所困扰，相信你会从本书中获益。该书适合于从事机器学习研究和应用的在校生和科研工作者。

——微软研究院首席研究员，自然语言处理资深专家　周明

前　言

致广大读者：

欢迎各位购买和阅读《Python 机器学习及实践》！

本书的编写旨在帮助大量对机器学习和数据挖掘应用感兴趣的读者朋友，整合并实践时下最流行的基于 Python 语言的程序库，如 Scikit-learn、Pandas、NLTK、gensim、XGBoost、Tensorflow 等；针对现实中的科研问题，甚至是 Kaggle 竞赛（当前世界最流行的机器学习竞赛平台）中的分析任务，快速搭建有效的机器学习系统。

读者在阅读了几个章节之后，就会发现这本书的特别之处。作者力求减少读者对编程技能和数学知识的过分依赖，进而降低理解本书与实践机器学习模型的门槛；并试图让更多的兴趣爱好者体会到使用经典模型，乃至更加高效的方法解决实际问题的乐趣。同时，作者对书中每一处的关键术语都提供了标准的英文表述，也方便读者快速查阅和理解相关的英文文献。

由于本书不涉及对大量数学模型和复杂编程知识的讲解，因此受众非常广泛。这其中就包括：在互联网、IT 相关领域从事机器学习和数据挖掘相关任务的研发人员；于高校就读的博士、硕士研究生，甚至是对计算机编程有初步了解的本科生；以及对机器学习与数据挖掘竞赛感兴趣的计算机业余爱好者等。

感激父母长久以来对我的关爱。也非常感谢我在清华大学和纽约大学的导师们：郑方、周强以及 Ralph Grishman 教授，对于我利用业余时间编写本书的理解和支持。特别致谢纽约大学的 Emma Zhu 同学，在我写书期间所给予计算设备的帮助。最后，感谢中国国家留学基金委为本人在美国留学期间所提供的生活资助。

最后，衷心地希望各位读者朋友能够从本书获益，同时这也是对我最大的鼓励和支持。全书代码下载地址为：http://pan.baidu.com/s/1bGp15G。对于书中的错误，欢迎大家批评指正，并发送至电邮：fanmiao.cslt.thu@gmail.com。我们会在本书的勘误网站

https://coding.net/u/fanmiao_thu/p/Python_ML_and_Kaggle/topic 上记录下您的重要贡献。

写于美国纽约中央公园
2015年12月25日

目　录

第 1 章

简 介 篇

本章介绍机器学习的基本理论和必要的编程准备。首先，借由美国卡内基梅隆大学(Carnegie Mellon University)著名教授 Tom Mitchell 对机器学习(Machine Learning)的经典定义，在“1.1 机器学习综述”节中进行阐述，并力求通俗易懂。然后以“良/恶性乳腺癌肿瘤预测”问题为实例，向读者朋友更加细致地剖析机器学习理论中的关键概念。而后，在“1.2 Python 编程库”节中解释之所以选择 Python 搭建机器学习平台的原因和优势，同时为读者朋友推介一系列用于快速搭建机器学习系统的 Python 编程库，并且这些编程库都会在本书的后续章节中详加讨论。“1.3 Python 环境配置”节将一步步教会大家如何在最常见的两大 PC 操作系统平台(Windows 和 Mac OS)上配置所需的编程环境，包括如何架设 Python 2.x 解释器环境和所需的编程库等。最后，利用配置好的编程环境，在“1.4 Python 编程基础” 节中，我们要向读者朋友提供这门当下最流行的计算机编程语言的编程规范和基本要素讲解，目的在于方便各位理解和进一步实践本书后续的代码。

1.1 机器学习综述

机器学习是一门既“古老”又“新兴”的计算机科学技术，隶属于人工智能(Artificial Intelligence)研究与应用的一个分支。

早在计算机发明之初，一些科学家就开始构想拥有一台可以具备人类智慧的机器。这其中就包括计算机结构理论的先驱、人工智能之父艾伦·麦席森·图灵(Alan Mathison Turing)。图灵在 1950 年发表的论文《计算机器与智能》(*Computing Machinery and Intelligence*)[1] 中提出了具有开创意义的“图灵测试”(Turing Test)，用来判断一台计算机是否达到具备人工智能的标准。我们将有关描述“图灵测试”的原文节选如下：

The new form of the problem can be described in terms of a game which we call the "imitation game". It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart front the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman.

We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, "Can machines think?"

这段英文原文的意思概括来讲就是："如果通过问答这种方式，我们已经无法区分对话那端到底是机器还是人类，那么就可以说这样的机器已经具备人工智能。"，如图 1-1 所示。尽管仍然有一些科学家并不完全赞同这种测试标准；但是不得不承认，在计算机刚刚发明不到 10 年的时间里，图灵能够具有这种前瞻性的构想，甚至为我们提供了用来测试人工智能的蓝图，是极为难能可贵的。

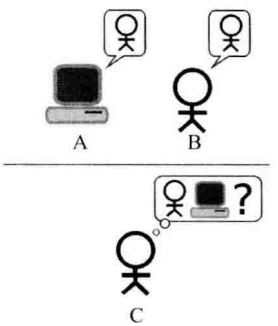

图 1-1 图灵测试

而机器学习，作为人工智能的分支，从 20 世纪 50 年代开始，也历经了几次具有标志性的事件，这其中包括：1959 年，美国的前 IBM 员工塞缪尔（Arthur Samuel）开发了一个西洋棋程序。这个程序可以在与人类棋手对弈的过程中，不断改善自己的棋艺。在 4 年之后，这个程序战胜了设计者本人；并且又过了 3 年，战胜了美国一位保持 8 年常胜不败的专业棋手。1997 年，IBM 公司的深蓝（Deep Blue）超级计算机在国际象棋比赛中力克俄罗斯（前苏联）专业大师卡斯帕罗夫（Garry Kimovich Kasparov），自此引起了全世界从业者的瞩目。同样是 IBM 公司，于 2011 年，她的沃森深度问答系统（Waston DeepQA）在美国知名的百科知识问答电视节目（Jeopardy）中一举击败多位优秀的人类选手成功夺冠，又使得我们朝着达成"图灵测试"更近了一步。最近的一轮浪潮来自于深度学习（Deep Learning）的兴起，也就是在笔者正在写这本书期间，谷歌公司 DeepMind 研究团队正式宣布[10]其创造和撰写的机器学习程序 AlphaGo① 以 4∶1 的总比分击败了世界顶级围棋选手李世石，见证了人工智能的极大进步。

按照机器学习理论先驱、塞缪尔先生的说法，他并没有编写具体的程序告诉西洋棋程

① https://en.wikipedia.org/wiki/AlphaGo

序如何行棋。事实上，这也是不可能的。因为下棋策略千变万化，我们无法通过编写完备的，哪怕是固定的执行规程来对战人类棋手。从塞缪尔的西洋棋程序，到谷歌的AlphaGo，我们可以总结出机器学习系统具备如下特点：

- 许多机器学习系统所解决的都是无法直接使用固定规则或者流程代码完成的问题，通常这类问题对人类而言却很简单。比如，计算机和手机中的计算器程序就不属于具备智能的系统，因为里面的计算方法都有清楚而且固定的规程；但是，如果要求一台机器去辨别一张相片中都有哪些人或者物体，这对我们人类来讲非常容易，然而机器却非常难做到。
- 所谓具备“学习”能力的程序都是指它能够不断地从经历和数据中吸取经验教训，从而应对未来的预测任务。我们习惯地把这种对未知的预测能力叫做泛化力(Generalization)。
- 机器学习系统更加诱人的地方在于，它具备不断改善自身应对具体任务的能力。我们习惯称这种完成任务的能力为性能(Performance)。塞缪尔的西洋棋程序和谷歌的 AlphaGo 都是典型的借助过去对弈的经验或者棋谱，不断提高自身性能的机器学习系统。

尽管我们通过西洋棋程序的例子总结了一些机器学习系统所具备的特性，但是作者仍然喜欢引述美国卡内基梅隆大学(Carnegie Mellon University)机器学习研究领域的著名教授 Tom Mitchell 的经典定义[2]来作为阐述机器学习理论的开篇：

> *A program can be said to learn from experience* ***E*** *with respect to some class of tasks* ***T*** *and performance measure* ***P****, if its performance at tasks in* ***T****, as measured by* ***P****, improves with experience* ***E****.*

真的是令人称道的表述，而且带有英文独特的韵脚和节律。我们尝试翻译一下：如果一个程序在使用既有的经验(E)执行某类任务(T)的过程中被认定为是“具备学习能力的”，那么它一定需要展现出：利用现有的经验(E)，不断改善其完成既定任务(T)的性能(P)的特质。

下面，我们会对其中的三个关键术语：任务(Task)、经验(Experience)、性能(Performance)逐一进行剖析，并将一个“良/恶性乳腺癌肿瘤预测”的经典机器学习问题引作开篇实例。

1.1.1　任务

机器学习的任务种类有很多，本书侧重于对两类经典的任务进行讲解与实践：监督学习(Supervised Learning)和无监督学习(Unsupervised Learning)。其中，监督学习关注对事物未知表现的预测，一般包括分类问题(Classification)和回归问题(Regression)；无监督学习则倾向于对事物本身特性的分析，常用的技术包括数据降维(Dimensionality