新工科建设

数据科学与大数据技术系列

电子工业出版社版权所有
盗版必究

Python机器学习

数据建模与分析（第2版）

薛　薇◎编著

電子工業出版社

Publishing House of Electronics Industry

北京 · BEIJING

内 容 简 介

本书将引领读者进入 Python 机器学习领域。机器学习是一套先进、深刻且内容丰富的算法集合，已成为数据科学中数据建模与分析的重要方法。Python 是一款简明、高效且功能强大的开源工具，也是数据科学实践中最常用的计算机语言。学好机器学习的理论方法，掌握 Python 这个实用工具，是成长为数据科学人才所必需的。

本书采用理论与实践相结合的方式，理论上突出可读性并兼具知识深度和广度，实践上强调可操作性并兼具应用广泛性，对机器学习的原理部分进行了深入透彻的讲解，对机器学习的算法部分给出了 Python 代码，并且在各章中设置了 Python 编程示例。全彩呈现机器学习的数据建模可视化图例(80 多幅彩图)，扫描书中相应二维码即可查看。提供配套数据集、源代码、教学 PPT 等学习资源，登录华信教育资源网(www.hxedu.com.cn)即可免费下载。

本书可作为高等院校机器学习、数据分析等专业课程的教材，也可作为数据科学应用研究者及对 Python 机器学习感兴趣的数据建模与分析从业者的参考书。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目(CIP)数据

Python 机器学习：数据建模与分析 / 薛薇编著. —2 版. —北京：电子工业出版社，2023.7
ISBN 978-7-121-45935-1

Ⅰ. ①P… Ⅱ. ①薛… Ⅲ. ①软件工具－程序设计－高等学校－教材②机器学习－高等学校－教材
Ⅳ. ①TP311.561②TP181

中国国家版本馆 CIP 数据核字(2023)第 123916 号

责任编辑：秦淑灵
印　　刷：
装　　订：
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×1 092　1/16　　印张：24.5　　字数：647 千字
版　　次：2021 年 4 月第 1 版
　　2023 年 7 月第 2 版
印　　次：2023 年 7 月第 1 次印刷
定　　价：99.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：qinshl@phei.com.cn。

电子工业出版社版权所有
盗版必究

前　言

机器学习是数据科学中数据建模与分析的重要方法，既是当前大数据分析的基础和主流工具，也是迈向深度学习和通往人工智能的必经之路。Python 是数据科学实践中最常用的计算机语言，是当前最流行的机器学习实现工具，并且因在理论和应用中的不断发展完善而拥有长期竞争优势。因此，学好机器学习的理论方法，掌握 Python 这个实用工具，是成长为数据科学人才所必需的。

“实施科教兴国战略，强化现代化建设人才支撑”的论述中强调，教育、科技、人才是全面建设社会主义现代化国家的基础性、战略性支撑。必须坚持科技是第一生产力、人才是第一资源、创新是第一动力，深入实施科教兴国战略、人才强国战略、创新驱动发展战略，开辟发展新领域新赛道，不断塑造发展新动能新优势。近年来，人工智能及机器学习已经成为推动全球科技发展的强大引擎之一，因此相关人才培养和配套教材体系建设显得尤为重要。

作者将多年来在机器学习、数据挖掘、统计学、计算机语言和统计应用软件等课程中的教学经验与科研实践经验进行归纳总结，精心编写了这本实用的优质图书，希望将经验和心得分享给广大从事数据科学领域工作的同人，以及从事 Python 机器学习教与学的高校师生们。

本书的特点如下。

1．对原理部分进行清晰的讲解

机器学习是一门交叉性很强的学科，涉及统计学、数据科学、计算机科学等多个领域的知识。作者认为，读者要掌握每个模型或算法的精髓和实践，需要由浅入深地关注其直观含义、方法原理、公式推导、算法实现和适用场景等多个递进层面。本书正是基于这样的层面展开论述的。

2．对实践部分进行全面的实现

机器学习又是一门实操性很强的学科。作者认为，读者需要边学边做才能获得更加深刻的认知。正因如此，本书在各章中设置了 Python 编程示例。一方面，通过 Python 代码和可再现的各种图形，帮助读者理解抽象理论背后的直观含义和方法原理。另一方面，通过 Python 代码，帮助读者掌握和拓展机器学习的算法实现与应用实践。全书所有模型和算法都有相应的 Python 代码，每章结尾还给出了本章总结、本章相关函数列表及本章习题。

3．本书适合作为机器学习或相关课程的教学及自学用书

本书将引领读者进入 Python 机器学习领域，理论上突出可读性并兼具知识深度和广度，实践上强调可操作性并兼具应用广泛性。本书采用一种有效而独特的方式讲解机器学习：首先，以数据建模与分析中的问题为导向，依知识点的难度由浅入深地讨论众多主流机器学习算法的核心原理；其次，通过 Python 编程和可视化图形，直观展示抽象理论背后的精髓和朴素道理；最后，通过应用实践示例强化算法的应用实践。

电子工业出版社版权所有 盗版必究

在章节安排上，本书共分为 13 章。在第 1 章以机器学习概述开篇和第 2 章对 Python 机器学习基础知识进行必要的提炼与总结后，第 3 章集中对数据预测与预测建模的各个方面进行了总览性论述，旨在帮助读者把握机器学习的整体知识框架。第 4～9 章按照由易到难的内在逻辑，顺序展开数据预测建模方法的讨论，涉及贝叶斯分类器、近邻分析、决策树、集成学习、人工神经网络、支持向量机等经典机器学习算法。第 10、11 章聚焦数据建模中不可或缺的重要环节，即特征工程，分别论述了特征选择和特征提取。第 12、13 章关注机器学习中的聚类算法。

在内容设计上，各章均由基本原理、Python 模拟和启示、Python 应用实践、本章总结、本章相关函数列表及本章习题等部分组成。基本原理部分，详细论述机器学习算法，旨在使读者能够知其然更知其所以然；Python 模拟和启示部分，通过数据模拟直观展示抽象理论背后的精髓和朴素道理，从而帮助读者进一步加深对理论精髓的理解；Python 应用实践部分，展现机器学习在环境污染、法律裁决、大众娱乐、医药健康、人工智能和商业分析等众多领域的应用价值，旨在提升读者的算法实践水平；本章总结、本章相关函数列表及本章习题部分，简要重述本章理论，归纳本章所涉及的 Python 函数，并通过习题强化知识要点和丰富 Python 应用实践内容。

本书第 1 版自问世以来受到了广大读者的欢迎和喜爱。在广泛听取高校老师和学生，以及其他读者的意见和建议的基础上，我们修订出版了第 2 版，修订说明如下。

1. 对第 1 版的章节顺序进行了较大调整

为更好地借助 Python 编程加深读者对机器学习理论的直观理解，我们将第 1 版集中编排的 Python 编程章节进行了拆解，分别将其调整到相应理论讲解的后面，以便于读者阅读和理解。同时，调整了部分章节理论讲解的逻辑顺序，使得理论论述更加有层次、更加清晰。

2. 每章分别设置了 Python 模拟和启示、Python 应用实践等部分

设置 Python 模拟和启示部分的目的是通过对较为理想的模拟数据的建模分析，更好地突显数据建模方法的理论精髓和特色，以便于读者理解。设置 Python 应用实践部分的目的是展示理论方法的实际应用价值，以便于读者举一反三地应用。为此，我们将第 1 版集中编排的 Python 代码分门别类地分章节重新编排。

3. 对关键 Python 代码进行了解释说明

增加对关键 Python 代码的解释说明，一方面便于读者了解 Python 函数中核心参数的含义，另一方面可帮助读者厘清编程思路，以便更好地了解相关理论的深层内涵。为此，我们以注释的形式对第 1 版中的 Python 代码增加了说明。同时为节约篇幅，第 2 版中略去了部分重复出现的代码，完整 Python 代码参见本书的配套电子资源，读者可登录华信教育资源网(www.hxedu.com.cn)免费下载。

4. 设置了选讲章节

我们希望以每周 3 或 4 课时共计约 17 周的课时数安排本书体量，但第 1 版的体量偏大。为此，我们在第 2 版中设置了选讲章节(目录中带星号的为选讲章节)，以便于任课教师依不同专业学生的先修课内容或知识点难度，有参考地删减课程内容。

5．优化了部分 Python 代码

第 2 版对 Python 代码进行了重新梳理，从全书角度统一了 Python 代码的书写风格，优化了部分 Python 代码的写法，并对第 1 版中的 Python 代码进行了适当删减和压缩。

总之，本书无论从内容设计上还是从体量安排上，都更贴近数据科学与大数据技术的专业课程设置，也可满足人工智能、统计学及计算机应用等相关专业课程的要求。此外，本书也可作为 Python 机器学习研究应用人员的参考用书。

在本书编写过程中，陈欢歌老师参与了部分章节的编写和文献资料与数据的整理，电子工业出版社高等教育分社的秦淑灵老师从选题策划到章节安排都提出了宝贵建议，在此一并表示感谢。

在以大数据与人工智能技术为代表的新一轮科技浪潮的推动下，Python 编程与机器学习也迅猛发展并快速迭代，形成了方法丰富、分支多样、应用广泛的整体态势。因此，想要全面而深入地掌握其全貌，需要不断学习与完善、不断跟进与提高。希望各位读者不吝赐教，对本书中的疏漏之处提出宝贵意见。

作者
中国人民大学应用统计研究中心
中国人民大学统计学院

电子工业出版社版权所有
盗版必究

电子工业出版社版权所有
盗版必究

目　　录

电子工业出版社版权所有 盗版必究

电子工业出版社版权所有
盗版必究

电子工业出版社版权所有
盗版必究

电子工业出版社版权所有 盗版必究

电子工业出版社版权所有
盗版必究

电子工业出版社版权所有 盗版必究

电子工业出版社版权所有
盗版必究

电子工业出版社版权所有 盗版必究

第 1 章 机器学习概述

移动互联技术、物联网技术和云计算技术的蓬勃发展，不但将人类社会与物理世界有效连接起来，还创造性地建立了一个数字化的网络体系。运行其上的搜索引擎、大型电子商务平台、互联网金融平台、社交网络平台和各类应用程序(App)等不断改变着社会生产方式、企业管理服务方式及人类生活方式，伴随着巨大比特流的随时随地的海量释放，一个数据收集、存储、处理能力空前的大数据时代已经到来。

大数据分析是围绕具有典型 5V(Volume，海量数据规模；Velocity，快速流转且动态激增的数据体系；Variety，多样异构的数据类型；Value，潜力大但密度低的数据价值；Veracity，有噪声影响的数据质量)特征的大数据集展开的。大数据广义上包括大数据理论、大数据技术、大数据应用和大数据生态等方面的组合架构。其中，大数据理论从计算机科学、统计学、数学及实践等方面汲取营养，旨在探索独立且关联于自然世界和人类社会的新的数据空间，构筑数据科学的理论基础和认知体系，具有鲜明的跨学科色彩；大数据技术是推动大数据发展最活跃的因素，包括大数据采集和传输、大数据集成和存储、云计算与大数据分析、大数据平台构建和大数据隐私与安全等众多技术方面；多领域大数据应用场景的有效开发成为带动大数据发展的重要引擎，涉及个人、企业与行业、政府及时空综合应用等若干方面；大数据生态通常是指大数据与其相关环境所形成的相互作用、相互影响的系统，如大数据市场需求、政策法规、人才培养、产业配套与行业协调、区域协同与国际合作等共生系统。

本书将聚焦大数据分析中的经典方法和主流实现技术：机器学习的基本原理，以及基于 Python 编程和机器学习的数据建模与分析。

1.1 机器学习的发展：人工智能中的机器学习

大数据深层次量化分析的实际需求、大数据存储力和计算力的空前卓越，使得作为人工智能重要组成部分和人工智能研究发展重要阶段的机器学习的理论及应用在当今社会得到了前所未有的发展并大放异彩。

诞生于 20 世纪 50 年代的人工智能(Artificial Intelligence，AI)，因旨在实现人脑部分思维的计算机模拟，以及人类智力任务的自动化完成，所以从研究伊始就具有浓厚的神秘色彩。人工智能研究经历了从符号主义人工智能(Symbolic AI)，到机器学习(Machine Learning)，再到深度学习(Deep Learning)的发展阶段。

1.1.1 符号主义人工智能

20 世纪 50 年代到 80 年代末，人工智能的主流实现范式是符号主义人工智能，即基于“一切都可规则化编码”的基本信念，通过让计算机执行事先编写好的程序，也称硬编码，依指定规则自动完成相应的处理任务，实现与人类水平相当的人工智能。

该实现范式的顶峰成果是 20 世纪 80 年代盛行的专家系统(Expert System)及不断涌现的各类计算机博弈系统。事实上，符号主义人工智能适合解决规则能够明确定义的逻辑问题，但尚有许多无法解决的研究难题。

例如，专家系统在某种意义上能够代替专家给病人看病，帮助人们甄别矿藏，但系统建立过程中的知识获取和知识表示问题一直没能得到很好的解决。知识获取的难点在于，如何全面系统地获取专家的领域知识，如何有效克服知识传递过程中的思维跳跃性和随意性。知识表示问题则更为复杂。传统“如果……则……”的简单因果式的计算机知识表示方式，显然无法表达形式多样的领域知识。更糟糕的是，专家系统几乎不存储常识性知识。爱德华·阿尔伯特·费根鲍姆(Edward Albert Feigenbaum)[①]曾估计，一般人拥有的常识存入计算机大约有 100 万条事实和抽象经验。将数量如此庞大的事实和抽象经验整理、表示并存储在计算机中，难度是极大的，而没有常识的专家系统的智能水平是令人担忧的。

又如，作为符号主义人工智能另一重大应用研究成果的计算机博弈系统，自 20 世纪 70 年代开始，主要体现在国际象棋、中国象棋、五子棋、围棋等棋类应用上。其顶峰成果是 IBM 研制的“深蓝”(Deep Blue)超级计算机。1997 年 5 月，“深蓝”与国际象棋大师加里·卡斯帕罗夫(Garry Kasparov)进行了 6 局制比赛，结果“深蓝”以两胜三平一负的成绩获胜。“深蓝”出神入化的棋艺依赖于能快速评估每种可能走法的利弊的评估系统。该系统背后除了有高性能计算机硬件系统的支撑，还有基于数千种经典对局和残局数据库的一般规则，以及国际象棋大师乔约尔·本杰明等人组成的人类棋手参谋团队针对卡斯帕罗夫的套路而专门设置的应对策略。计算机博弈系统的最大“死穴”是人类“不按套路出牌”。在这一点上人类的智能水平是计算机远不能及的。

由于符号主义人工智能很难解决语言翻译、语音识别、图像分类等更加复杂和模糊的、没有明确规则定义的逻辑问题，因此需要一种更迭符号主义人工智能的新策略，这就是机器学习。

事实上，机器学习对计算机博弈系统同样有着革命性的卓越贡献[②]。因使用特定算法和编程方法实现人工智能，基于机器学习的计算机博弈系统不仅能够极大地压缩原先数百万行的代码(包括所有的棋盘边缘情况，对手棋子的所有可能的移动等)，而且能够从以前的游戏中学习策略并提高其未来的性能。

1.1.2 基于机器学习的人工智能

如何将大型数据库、机器学习算法和分布式计算整合在一个体系架构下，创造比人类

① 爱德华·阿尔伯特·费根鲍姆：计算机人工智能领域的科学家，被誉为“专家系统之父”，1994 年获得计算机科学领域最负盛名的奖项——图灵奖。

② 1952 年，亚瑟·塞缪尔(Arthur Samuel)创建了第一个真正的基于机器学习的棋盘游戏；1963 年，唐纳·米基(Donald Michie)提出了基于强化学习的井字游戏(Tic-Tac-Toe)。

更好的、能够完成判断与认知推理等更为复杂和模糊的任务(如自然语言理解、图像识别和分类等)的机器，成为人工智能探索的新热点。其中的机器学习是关键。

机器学习概念的提出源于“人工智能之父”阿兰·图灵(Alan Turing)于 1950 年进行的图灵测试[①]。该测试令人信服地表明“会思考的机器”是可能存在的，计算机能够具有学习与创新能力。与符号主义人工智能策略截然不同的是，机器学习的出发点是，与其明确地编写程序让计算机按规则完成智能任务，不如教计算机借助某些算法完成任务。

从计算机程序设计角度看，符号主义人工智能体现的是给计算机输入“规则”和“数据”，计算机处理数据，并依据以程序形式明确表达的“规则”自动输出“答案”；机器学习体现的是给计算机输入“数据”和从数据中预期得到的“答案”，计算机找到并输出“规则”，并依据“规则”给出对应新数据的“答案”，从而完成各种智能任务。相对于经典的编程范式，人们将机器学习视为一种新的编程范式。图 1.1 展示了谷歌(Google)人工智能研究员、深度学习工具 Keras 之父弗朗索瓦·肖莱(Francois Chollet)对两种编程范式的基本描述。

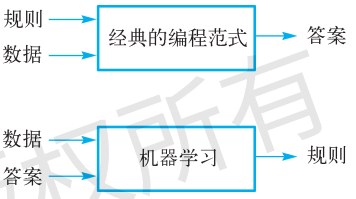

图 1.1　两种编程范式

在图 1.1 中，机器学习中的“规则”，是计算机基于大数据集，借助算法解析“数据”和“答案”关联性的结果，通常不会甚至根本无法通过人工程序事先明确编写出来。从某种意义上看，机器学习能够比人类做得更好。

基于速度更快的硬件与更大数据集的训练，机器学习在 20 世纪 90 年代开始蓬勃发展，并迅速成为人工智能最受欢迎且应用最成功的分支领域。其在自然语言理解领域的最高成就之一是 IBM 制造的、以公司创始人托马斯·约翰·沃森(Thomas J. Watson)的名字命名的智能计算机：沃森(Watson)。2011 年 4 月 1 日，美国著名的问答节目《危险边缘》拉开了沃森与人类的情人节人机大战的序幕。《危险边缘》是一个综合性智力竞猜电视节目，题目涵盖时事、历史、艺术、流行文化、哲学、体育、科学、生活常识等方面几乎所有人类已知的知识。与沃森同场竞技肯·詹宁斯(Ken Jennings)和布拉德·鲁特(Brad Rutter)是该节目有史以来成绩最好的两位人类参赛者。但是，基于数百万份图书、新闻、电影剧本、辞海、文选资料，借助深度快速问答(DeepQA)技术中的 100 多套算法，以及 3s 内的问题解析和候选答案搜索能力，沃森最终以近 8 万的得分，将两位得分均在 2 万左右的人类选手远远甩在了后面，成为《危险边缘》节目的新王者。

尽管如此，智能计算机面临的自然语言理解挑战仍是严峻的。事实上，与其他计算机一样，沃森完成的是文字符号的处理，无法真正理解其含义。例如，问答题“这个被信赖的朋友是一种非奶制的奶末”，标准答案应是咖啡伴侣。因为咖啡伴侣多是植物制的奶精而非奶制品，并且人类做这道题时会很快想到“朋友”对应“伴侣”。但计算机却只能在数据库里寻找“朋友”“非奶制”“奶末”这些词的关联词，结果关联最多的是牛奶。此外，如

① 图灵测试：在不接触对方的情况下，一个人通过某种特殊方式和计算机进行一系列问答，如果在相当长时间内，人无法根据这些问答判断对方是人还是计算机，则可认为这个计算机具有同人相当的智力，是具有智能和思维能力的。图灵测试用于判断机器是否具有智能。

何领悟双关、反讽之类的语言修辞，以及分析比语言理解本身更复杂的情感问题等，都是智能计算机面临的巨大挑战。

机器学习的最大突破是 2006 年提出的深度学习。深度学习是机器学习的重要分支领域，旨在从数据中学习数据表示的新方法。它强调基于训练数据，通过众多连续的神经网络层(Layer)，过滤和提取数据中服务于预测的重要特征。相对于拥有众多层的深度学习，某些经典的机器学习算法有时也被称为浅层学习(Shallow Learning)算法。

目前，包括强化学习(Reinforcement Learning，RL)策略在内的深度学习，已被广泛应用于自然语言理解和语音解析，不仅能够应对自然语言理解的挑战，还能够解决以图像识别和分类等为核心任务的众多感知问题。例如，在计算机博弈系统上的成功案例是谷歌旗下的 DeepMind 公司开发的人工智能围棋软件阿尔法围棋(AlphaGo)。2015 年 10 月，阿尔法围棋以 5∶0 完胜欧洲围棋冠军、职业二段选手樊麾；2016 年 3 月，阿尔法围棋对战世界围棋冠军、职业九段选手李世石，并以 4∶1 的总比分获胜。

目前，以机器学习和深度学习为核心分析技术的人工智能，已经具备接近人类水平的图像识别和分类能力、手写文字转录能力、语音识别和对自然语言提问的回答能力，以及多国语言的翻译能力等。正因如此，人工智能技术得以广泛应用。例如，智能家电等物联网(Internet of Things，IoT)设备比以往任何时候都更加聪明和智能；Slack 等聊天机器人正在提供比人类更快、更高效的虚拟客户服务；自动驾驶汽车能够识别和解析交通标志，自动实现导航和维护；等等。人工智能正改变着今天人们的日常生活方式，未来人工智能还将继续探索机器感知和自然语言理解之外的各种应用问题，协助人类开展科学研究，自动进行软件开发等。

1.2 机器学习的核心：数据和数据建模

1.1.2 节谈到，从计算机程序设计角度看，机器学习是一种新的编程范式。实现这种范式的核心任务是发现隐藏在“数据”和“答案”中的“规则”。其理论可行性最早可追溯到 1783 年托马斯·贝叶斯(Thomas Bayes)提出的贝叶斯定理，其证明了存在一种能够从历史经验，即数据集中的“数据”和“答案”中，学习两者之间关联性“规则”的数学方法。

若将“数据”和“答案”视为一种广义数据，则借助数学方法学习“规则”的本质便可看作一种基于数据的建模。从这个角度看，机器学习是一种基于大数据集，以发现其中隐藏的、有效的、可理解的规则为核心目标的数据建模过程，旨在辅助解决各行业领域的实际应用问题。

本书将聚焦机器学习解决实际应用问题的主流算法和 Python 应用实践。以下将首先围绕大众熟知的空气质量监测数据分析问题，讨论机器学习的对象：数据集，以及与数据集相关的基本概念。然后论述机器学习基于数据集的具体任务。

1.2.1 机器学习的对象：数据集

机器学习的对象是数据集合，简称数据集(也称样本集)。常规的数据集一般以二维表(也称扁平表)形式组织，由多个行和列组成。表 1.1 所示为北京市空气质量监测数据集。

电子工业出版社版权所有 盗版必究

表 1.1　北京市空气质量监测数据集[①]

日期	AQI	空气质量等级	污染物浓度					
			PM2.5/(μg/m^3)	PM10/(μg/m^3)	SO_2/(μg/m^3)	CO/(mg/m^3)	NO_2/(μg/m^3)	O_3/(μg/m^3)
2019/1/1	45	优	28	45	8	0.7	34	47
2019/1/2	78	良	57	75	12	1	56	28
2019/1/3	162	中度污染	123	136	21	1.9	85	12
2019/1/4	40	优	18	40	5	0.5	26	61
2019/1/5	47	优	17	34	7	0.5	37	49
2019/1/6	88	良	64	95	12	1.4	70	13
2019/1/7	55	良	34	54	9	0.9	44	52
2019/1/8	35	良	10	59	4	0.5	28	62
2019/1/9	74	优	41	66	12	0.9	59	26
2019/1/10	100	良	75	113	14	1.5	79	17
2019/1/11	135	轻度污染	103	130	15	1.7	80	21
2019/1/12	267	重度污染	217	212	14	2.7	101	14
2019/1/13	169	中度污染	128	183	6	1.7	64	58
2019/1/14	137	轻度污染	104	143	8	1.7	72	46
2019/1/15	54	良	9	57	4	0.3	18	61
2019/1/16	73	良	38	74	10	0.9	58	37
2019/1/17	78	良	45	77	13	1.2	62	34
2019/1/18	94	良	64	105	15	1.4	75	18
2019/1/19	33	优	11	33	5	0.5	24	59
2019/1/20	33	优	9	29	3	0.3	19	66
2019/1/21	57	良	24	63	6	0.6	44	62
2019/1/22	64	良	28	70	7	0.9	51	54
2019/1/23	60	良	23	53	6	0.7	48	56
2019/1/24	79	良	58	75	12	1.2	53	41
2019/1/25	32	优	10	32	4	0.5	21	61
2019/1/26	44	优	25	40	5	0.6	35	50
2019/1/27	76	良	49	102	9	1	45	67
2019/1/28	56	良	35	61	6	0.6	33	59
2019/1/29	139	轻度污染	106	140	17	1.5	76	24
2019/1/30	63	良	28	75	5	0.6	22	61
2019/1/31	30	优	13	27	5	0.5	20	59
20191/2/1	68	良	42	85	7	0.9	49	54
20191/2/2	142	轻度污染	108	137	8	1.4	55	29

数据集中的一行通常称为一个样本观测。例如，表 1.1 中第一行是 2019 年 1 月 1 日的北京市空气质量监测数据，这就是一个样本观测。若数据集由 N 个样本观测组成，则称该数据集的样本容量或样本量为 N。当利用机器学习解决复杂问题时一般要求具有样本量较大的数据集(也称大数据集，是相对小数据集而言的)。

① 数据来源：中国空气质量在线监测分析平台。

电子工业出版社版权所有 盗版必究

数据集中的一列通常称为一个变量(也称特征)，用于描述数据的某种属性或状态。例如，表 1.1 中包括日期，AQI(Air Quality Index，空气质量指数)，空气质量等级，PM2.5、PM10、SO_2、CO、NO_2、O_3 浓度共 9 个变量。其中，AQI 作为空气质量状况的无量纲指数，值越大表明空气中污染物浓度越高，空气质量越差。影响 AQI 的主要污染物包括 PM2.5、PM10、SO_2、CO、NO_2、O_3。空气质量等级是 AQI 的分组结果。一般 AQI 在 0～50、51～100、101～150、151～200、201～300、大于 300 时，空气质量等级依次为一级优、二级良、三级轻度污染、四级中度污染、五级重度污染、六级严重污染。

进一步，依各变量的取值类型可将变量细分为数值型、顺序型和类别型三类，后两类统称为分类型。数值型变量是连续或非连续的数值(在计算机中以整型或浮点型等存储类型存储)，可以进行算术运算，如表 1.1 中 AQI、PM2.5、PM10、SO_2、CO、NO_2、O_3 列的数值。分类型变量一般以数字(如 1, 2, 3 等)或字符(如 A, B, C 等)标签(在计算机中以字符串型或布尔值等存储类型存储)表示，进行算术运算没有意义。顺序型变量的标签存在高低、大小、强弱等顺序关系，如表 1.1 中的空气质量等级。此外，诸如学历、年龄段等变量也属于顺序型变量。类别型变量的标签没有顺序关系，如表 1.1 中的日期(可视为一个样本观测的标签)。此外，诸如性别、籍贯等变量也属于类别型变量。

机器学习以变量为基本数据单元，旨在发现不同变量取值之间的数量关系。不同机器学习算法适用于不同类型的变量，因此明确变量类型是极为重要的。

表 1.1 中的数据是结构化数据的一种具体体现。结构化数据是计算机关系数据库的专业术语，还包括实体和属性等概念。其中，实体对应这里的样本观测，属性对应这里的变量。结构化数据通常有以下两个方面的特征：第一，属性(变量)值通常是可定长的，如可定长的数值型变量、可定长的数字或字符标签；第二，各实体都具有共同的、确定性的属性，如每天的空气质量实体都通过 AQI、PM2.5 浓度等共同的属性度量。当然日期不同，AQI、PM2.5 浓度的具体值不尽相同。对此采用二维表形式组织数据不仅直观而且存储效率高。

机器学习的数据对象并不局限于结构化数据，还可以是半结构化数据和非结构化数据。半结构化数据的重要特点是包含可定长的属性和部分非定长的属性，且无法确保每个实体都具有共同的、确定性的属性。例如，员工简历数据就是典型的半结构化数据，其中可定长的属性包括性别、年龄、学历等，非定长的属性包括工作履历(如职场小白的工作履历是空白的，职场精英会有曾在多家公司任职的经历)等。此外，未婚员工配偶信息空缺，已婚员工的子女信息多样化等，不同实体并非均具有共同的、确定性的属性。对此需要经过一定的格式转换方可将其组织成二维表形式，且表格中会存在一些冗余数据，存储效率不高。半结构化数据往往可采用 JSON 文档格式组织。

非结构化数据一般不方便直接采用二维表形式组织。常见的非结构化数据主要有文本、图像、音频和视频数据等。这些数据往往是非定长的，且很难直接确定属性，需要进行必要的数据转换处理。

1.2.2 机器学习的任务：数据建模

基于数据集，机器学习通过数据建模主要完成以下两大任务：第一，数据预测；第二，数据聚类。

1. 数据预测

以下将基于两个实际应用场景，讨论数据预测的内涵。

【场景 1】基于空气质量监测数据集，我们希望得到以下两个问题的答案：

- SO_2、CO、NO_2、O_3 浓度中哪些是影响 PM2.5 浓度的重要因素，可否用于对 PM2.5 浓度的预测？
- PM2.5、PM10、SO_2、CO、NO_2、O_3 浓度对空气质量等级大小的贡献不尽相同，哪些污染物浓度的降低将有效降低空气质量等级，可否用于对空气质量等级进行预测？

上述两个问题就是典型的数据预测问题。

PM2.5 的主要来源是 PM2.5 的直接排放，以及某些气体污染物在空气中的转化。PM2.5 的直接排放主要源自石化燃料（煤、汽油、柴油）的燃烧、生物质（秸秆、木柴）的燃烧、垃圾焚烧等。可在空气中转化成 PM2.5 的气体污染物主要有 SO_2、氮氧化物、NH_3、挥发性有机物等。基于各种污染物浓度监测数据，如果能从量化角度准确发现 PM2.5 的主要来源和影响因素，度量各污染物浓度对 PM2.5 浓度的数量影响，则一方面有助于制定有针对性的控制策略，通过控制 SO_2 等污染物的排放降低 PM2.5 浓度；另一方面可基于其他污染物浓度对 PM2.5 浓度进行预测。该问题就是一个数据预测问题。

空气质量好坏的测度，即 AQI 计算或空气质量等级评定也是一个复杂的问题。空气污染是一个复杂的现象，在特定时间和地点，污染物浓度会受许多因素的影响，均会影响空气质量。例如，车辆、船舶、飞机的尾气，工业企业生产排放，居民生活和取暖，垃圾焚烧等，均会影响空气质量。此外，城市发展密度、地形地貌和气象等也是影响空气质量的重要因素。目前，参与空气质量等级评定的主要污染物包括 PM2.5、PM10、SO_2、CO、NO_2、O_3。基于各种污染物浓度监测数据和空气质量等级数据，如果能从量化角度准确找到导致空气质量等级敏感变化的污染物，则不仅能通过对其进行控制有效降低空气质量等级，还可基于污染物浓度对空气质量等级进行预测。该问题同样是一个数据预测问题。

简而言之，数据预测就是基于已有数据集，归纳出输入变量和输出变量之间的数量关系的过程。基于这种数量关系，一方面，可发现对输出变量产生重要影响的输入变量；另一方面，在数量关系具有普适性和未来不变的假设下，可对新数据输出变量取值进行预测。

进一步，数据预测可细分为回归预测和分类预测。对数值型输出变量的预测（对数值的预测）统称为回归预测。对分类型输出变量的预测（对类别的预测）统称为分类预测，如果输出变量仅有两个类别，则称其为二分类预测；如果输出变量有两个以上的类别，则称其为多分类预测。

参照 1.1.2 节中机器学习编程范式，这里的输入变量对应其中的“数据”，输出变量对应“答案”。数据集中输入变量和输出变量的取值均是已知的，可为数值型变量、顺序型变量或类别型变量。第一个问题中的 SO_2、CO、NO_2、O_3 浓度是数值型输入变量，PM2.5 浓度是数值型输出变量，该问题属于回归预测问题；第二个问题中的各种污染物浓度为数值型输入变量，空气质量等级为分类型输出变量，该问题属于多分类预测问题。

参照 1.1.2 节中机器学习编程范式，发现“规则”就是寻找输入变量和输出变量取值规律及应用规律的过程。这些规律不是显性的，而是隐藏在数据集中的，寻找这些规律需要基于对数据集的归纳学习。回归和分类正是这样旨在发现规律的归纳学习策略。

【场景 2】有关于顾客特征和其近 24 个月的消费记录数据集，其中包含顾客的性别、年龄、职业、年收入等属性特征，以及顾客购买的商品、消费金额等消费记录数据。基于这些数据，我们希望得到以下两个问题的答案：

- 具有哪些特征(如年龄和年收入等)的新顾客会购买某种商品？
- 具有某些特征(如年龄等)和消费行为(购买或不购买)的顾客，其平均年收入是多少？

上述两个问题均属数据预测的范畴。

第一个问题的答案无非是买或者不买，显然属于分类预测问题。其中，输入变量为性别、年龄、职业、年收入等，输出变量为是否购买。通过数据建模，找到顾客特征(输入变量)与其消费行为(输出变量)之间的规律。进一步，依据该规律对具有某些特征的新顾客的消费行为(购买或不购买)进行预测。

第二个问题要求对顾客的平均年收入进行预测，属于回归预测问题。其中，输入变量为性别、年龄、职业、是否购买等，输出变量为平均年收入。通过数据建模，找到顾客特征(输入变量)与其平均年收入(输出变量)之间的规律。进一步，依据该规律对具有某些特征和消费行为的新顾客的平均年收入进行预测。

2. 数据聚类

数据集中蕴含着非常多的信息，其中较为典型的是，数据集可能由若干个小的数据子集组成。例如，对于【场景 2】中的顾客特征和消费记录数据集，依据经验来看，通常具有相同特征(如相同性别、年龄、年收入等)的顾客群消费偏好较为相似，具有不同特征(如男性和女性、教师和 IT 人员等)的顾客群消费偏好可能差异明显。客观上存在着特征和消费偏好等差异较大的若干个顾客群。发现不同顾客群是实施精细化营销的前提。

各个顾客群将对应到数据集的各个数据子集上。机器学习称这些数据子集为子类或小类或簇等。数据聚类的目的是发现数据中可能存在的小类，并通过小类刻画和揭示数据的内在组织结构。数据聚类的最终结果是，给每个样本观测指派一个属于哪个小类的标签，称为聚类解。聚类解将保存在一个新生成的分类型变量中。

数据聚类和数据预测中的分类预测有联系也有区别。联系在于，数据聚类是给每个样本观测一个小类标签，分类预测是给输出变量一个分类值，本质上也是给每个样本观测一个标签。区别在于，分类预测中变量有输入变量和输出变量之分，且分类标签(保存在输出变量中，如空气质量等级、购买或不购买等)的真实值是已知的，但数据聚类中变量没有输入变量和输出变量之分，所有变量都将被视为聚类变量参与数据分析，且小类标签(保存在聚类解变量中)的真实值是未知的。正因如此，数据聚类有不同于分类预测的算法策略。

3. 其他方面的应用

除数据预测和数据聚类之外，机器学习还可解决其他应用场景下各种类型的问题，如关联分析和模式诊断。

1)关联分析

世间万物都是有联系的，这种联系让这个世界变得丰富多彩而又生动有趣。关联分析的目的就是寻找事物之间的联系规律，发现它们之间的关联性。

【场景 3】有一段时间内某超市会员的购物小票数据集，其中每张购物小票均记录了哪

电子工业出版社版权所有

个会员在哪个时间购买了哪些商品及购买的数量等。基于该数据集，我们希望得到以下两个问题的答案：

- 购买面包的会员中同时购买牛奶的可能性大，还是同时购买香肠的可能性大？
- 购买电水壶的会员未来一个月内购买除垢剂的可能性有多大？

显而易见，找到上述问题的答案对超市的货架布置、进货计划制订、有针对性地营销等有重要帮助。发现关联性的目的就是找出这些问题的答案。

发现关联性的关键是找到变量取值的内在规律。这里可将会员的购买行为视为一个变量，则该变量的所有可能取值为该超市销售的所有商品的名称。发现关联性就是要找到变量(购买行为)的不同取值(该超市销售的所有商品的名称)之间是否存在某些一般性的规律。

从发现关联性角度解决第一个问题的思路是，依据大量的购物篮数据(一张购物小票对应一个购物篮)，计算不同商品被同时购买的概率，如购买面包的同时购买牛奶概率等。这里的概率计算较为简单，如只需清点所有购买面包的购物小票中有多少张同时购买了牛奶，并计算百分比即可。这些概率可揭示商品购买间的简单关联关系，是商品推荐的基础。

为回答第二个问题，需要依时间连续跟踪每个会员的购物行为，即清点在指定时间段内购买电水壶的会员中有多少人在一个月内又购买了除垢剂，并计算百分比。该问题涉及时间因素，可揭示商品购买间的时序关联关系。

进一步，若将商品间的关联性和关联性强弱绘制成图，则可得到如图 1.2 所示的网状图。

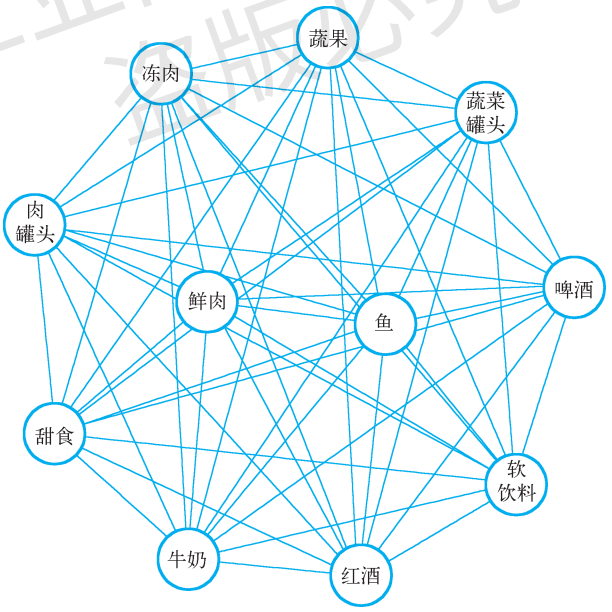

图 1.2　网状图

图 1.2 中的圆圈通常被称为网状图的节点，这里代表各个商品；节点之间的连线被称为节点连接，其粗细表示连接权重，这里表示商品间关联性强弱。事实上，关联性的研究可推广到许多应用中。例如，若网状图中的节点代表微信好友，则节点连接及连接权重可表示好友间的私聊频率，如果两个好友间从未私聊过，则相应节点间可以没有连接；若网状图中的

电子工业出版社版权所有 侵权必究

节点代表各个国家，则节点连接及连接权重可以表示各个国家间的贸易状况；若网状图中的节点代表股票，则节点连接及连接权重可表示各只股票价格的相互影响关系；若网状图中的节点代表学术论文，则节点连接及连接权重可表示学术论文间的相互引用关系；若网络图中的节点代表立交桥，则节点连接及连接权重可表示立交桥间的车流量；等等。

2) 模式诊断

模式(Pattern)是一个数据集合，由分散在数据集中的零星数据组成。模式通常具有其他众多数据所没有的某种局部的、非随机的、非常规的特殊结构或相关性。模式诊断就是指从不同角度采用不同方法发现数据中可能存在的模式。

例如，在工业生产过程中，数据采集系统或集散控制系统通过在线方式收集大量的可反映生产过程中设备运行状况的数据，如电压、电流、气压、温度、流量、电机转速等。在常规生产条件下，若设备运行正常，则这些数据的取值变化很小，基本维持在一个稳定水平上。若一小段时间内数据忽然剧烈变化，但很快又回归原有水平，且类似情况多次重复出现，即显现出局部的、非随机的、超出正常范围的变化，则意味着生产设备可能发生了间歇性异常。这里少量的变动数据所组成的集合就是模式，如图 1.3(a)中椭圆内的数据。

模式具有局部性、非随机性和非常规性的特点，很可能是某些重要因素所导致的必然结果，所以模式诊断[①]是极为必要的。例如，图 1.3(b)中椭圆内的会员构成了模式，表现出不同于绝大多数会员的特征，找到它们并探究其原因是有意义的。

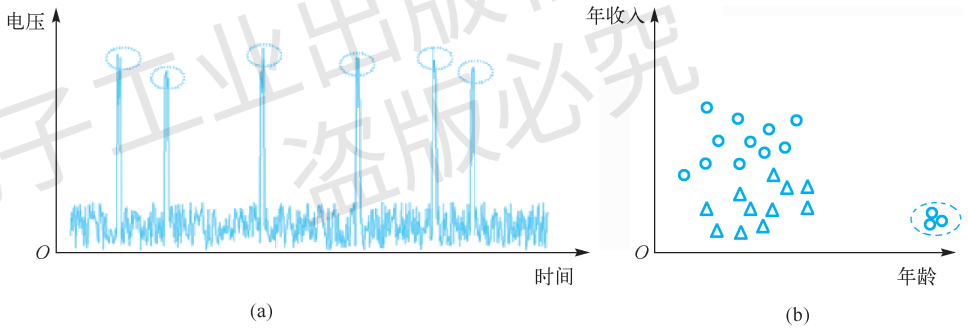

图 1.3　模式的示意图

模式诊断不仅可用于设备故障诊断，还可广泛用于众多主观故意行为的发现，如计算机网络入侵行为(网络流量或访问次数出现非随机性突变等)、恶意欺诈行为(信用卡刷卡金额、手机通话量出现非常规增加等)、虚报瞒报行为(商品销售额的非常规变化等)。进行模式诊断并探究模式的成因，能为技术更新、流程优化、防范升级等方案的制订提供重要依据。

需要注意的是，这里的模式与统计学中从概率角度界定的离群点有一定差别。例如，统计学中经典的 3σ 准则认为，若某随机变量服从正态分布，则绝对值大于 3σ 的变量值因出现的概率很小(小于或等于 0.3%)而被界定为离群点。尽管这些离群点与模式的数量都较少，且均表现出严重偏离数据全体的特征，但离群点通常由随机因素所致，模式则不然，它具有非随机性和潜在的形成机制。找到离群点的目的是剔除它们以消除对数据分析的影响，但模式很多时候就是人们关注的焦点，是不能剔除的。

尽管模式并非以统计学中的概率标准来界定，但从概率角度诊断模式仍是有意义的。

① 这里的模式诊断的含义与模式识别不尽相同。模式识别一般是指识别图像等数据阵中的某些形状等。

应注意的是，小概率既可能是模式的表现，也可能是随机性离群点的表现。因此，究竟是否为“真正”的模式，需要行业专家定夺。如果能够找到相应的常识、合理的行业逻辑或有说服力的解释，则可认定为“真正”的模式。否则，可能是数据记录错误而导致的“虚假”的模式或没有意义的随机性结果。从统计角度诊断模式需要已知或假定概率分布。当概率分布未知或无法做出假定时就需要从其他角度分析。对此，机器学习多采用数据聚类的方式实现模式诊断。

综上，作为人工智能的重要组成部分，机器学习的核心任务是数据预测和数据聚类，同时也正朝着智能数据分析的方向发展。随着大数据时代数据产生速度持续加快，数据体量空前增长，各种半结构化和非结构化的数据不断涌现，机器学习及深度学习在文本分类、文本摘要提取、文本情感分析，以及图像识别和图像分类等智能化应用中发挥着越来越重要的作用。

1.3　机器学习的典型应用

目前机器学习在电子商务、金融、电信等行业得到了极为广泛的应用。

1.3.1　机器学习的典型行业应用

机器学习的应用极为广泛。从应用成熟度和市场吸引力两个维度看，当前机器学习有如下几个典型行业应用，其应用成熟度和市场吸引力分布图[①]如图 1.4 所示。

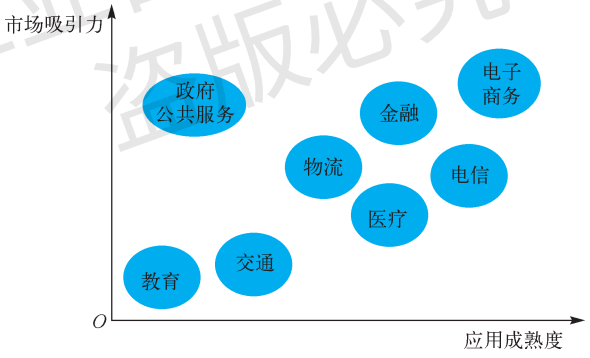

图 1.4　机器学习的典型行业应用的应用成熟度和市场吸引力分布图

图 1.4 表明，机器学习在电子商务行业中的应用是最成熟和最具市场吸引力的，金融和电信行业紧随其后。机器学习在政府公共服务行业中有较大的发展潜力，其未来的应用成熟度有巨大的提升空间。

机器学习在电子商务行业中的应用价值主要体现在市场营销和个性化推荐等方面，目的是有效实现用户消费行为规律的分析，制订有针对性的商品推荐方案，根据用户特征研究广告投放策略并进行广告效果的跟踪和优化；在金融行业中，机器学习主要应用于客户金融行为分析及金融信用风险评估等方面；机器学习在电信行业中的应用主要集中在客户消费感受分析等方面，目的是通过洞察客户需求，有针对性地提升电信服务的质量和安全；

① 资料来源：易观智库。

电子工业出版社版权所有 盗版必究

在政府公共服务行业中，机器学习可在智慧交通和智慧安防等方面发挥重要作用，实现以数据驱动的政府公共服务；机器学习在医疗行业中的应用价值集中在药品研发、公共卫生管理、居民健康管理及健康危险因素分析等方面。

尽管上述机器学习的典型行业应用所解决的问题不同，但其研究思路具有一定的共性。本节仅对机器学习在金融、电子商务、电信行业中的典型商业应用的共性问题进行梳理并进行详尽讨论，主要包括客户细分、客户流失分析、营销响应分析、交叉销售、欺诈甄别等方面，如图 1.5 所示。

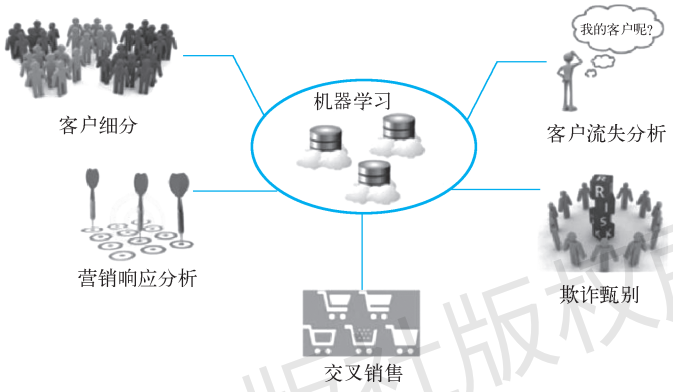

图 1.5　典型商业应用问题

1.3.2　机器学习在客户细分中的应用

客户细分(Customer Segmentation)的概念是 20 世纪 50 年代中期美国著名营销学家温德尔·史密斯(Wended R. Smith)提出的。客户细分是指经营者在明确其发展战略、业务模式和市场条件的情况下，依据客户价值、需求和偏好等诸多因素，将现有客户划分为不同的客户群，属于同一客户群的客户具有较强的相似性，不同细分客户群间存在明显的差异性。

在经营者缺乏足够的资源应对客户整体时，由于客户群的价值及需求存在异质性，因此进行有效的客户细分能够辅助经营者准确认识不同客户群的价值及需求，从而制定针对不同客户群的差异化经营策略，以资源效益最大化、客户收益最大化为目标，合理分配资源，实现持续发展新客户、保持老客户、不断提高客户忠诚度的总体目标。

客户细分的核心是选择恰当的细分变量，以及细分方法、细分结果的评价和应用。

1)选择恰当的细分变量

客户细分的核心之一是选择恰当的细分变量。根据不同的细分变量可能得到完全不同的客户细分结果。传统的客户细分是基于年龄、性别、婚姻状况、年收入、职业、地理位置等客户基本属性的。此外，还有基于各种主题，如客户价值、需求、偏好、消费行为等的客户细分。

不同行业的业务内容不同，其客户价值、需求、偏好及消费行为的具体定义也不同，需要选择有利于达成其分析目标的细分变量。例如，电信行业客户细分主要细分变量包括手机卡的使用月数、套餐金额、额外通话时长、额外流量、是否改变过套餐、是否签订过服务合约、是否办理过固定电话和宽带业务等。又如，商业银行为研发针对不同客户的金融产品和服务，对个人金融客户主要关注年龄、家庭规模、受教育程度、居住条件、收入来源、融资

电子工业出版社版权所有 盗版必究

记录等，对企业金融客户主要关注行业、企业组织形式、企业经营年限、雇员人数、总资产规模、月销售额、月利润等。同时，关注的贷款特征包括贷款期限、贷款用途、抵押物、保证人等。再如，电子商务行业的客户细分，除需要关注客户的收入、职业特点、行业地位、关系背景等基本属性之外，还需要关注客户的喜好风格、价格敏感、品牌倾向、消费方式等主观特征，以及交易记录、积分等级、退换货、投诉、好评传播等交易行为特征等。

能否选择恰当的细分变量取决于对于业务需求的认知程度。在不同行业的客户细分问题中，客户的“好坏”标准可能不同。随着业务的推进及外部环境的动态变化，这个标准也可能发生变化。因此，选择细分变量应建立在明确当前业务需求的基础之上。细分变量的个数应适中，以能否覆盖业务需求为准，同时各细分变量之间不应有较强的关联性。

2）细分方法、细分结果的评价和应用

机器学习实现客户细分的主要方法是聚类分析。有关聚类分析的原理和特点等将在第 12、13 章详细讨论。

客户细分的结果是多个客户群。只有在合理的客户群基础上制定有针对性的营销策略，才可能获得资源效益的最大化及客户收益的最大化。客户群划分是否合理，一方面依赖于细分变量的选择，另一方面依赖于所运用的细分方法。细分方法的核心是数据建模，而数据建模通常带有“纯粹和机械”的色彩。尽管它给出的客户群划分具有数理上的合理性，但并不一定都是符合业务需求的，所以还要从业务角度评价细分结果的实际适用性。例如，各个客户群的主要特征是否具有业务上的可理解性；客户群所包含的人数是否足够多，能否收回相应的营销成本；客户群的营销方案是否具有实施上的便利性；等等。

1.3.3　机器学习在客户流失分析中的应用

客户流失是指客户终止与经营者的服务合同或选择其他经营者提供的服务。通常客户流失有如下三种类型。

第一，企业内部的客户转移，即客户转移到本公司的其他业务上。例如，银行增加新业务或调整费率等引发客户的业务转移，如客户储蓄账户从活期存款转移至整存整取，客户理财账户从单一类信托产品转移至组合类信托产品等。企业内部的客户转移，就某个业务来看存在客户流失，可能对企业收入产生一定影响，但就企业整体而言客户并没有流失。

第二，客户被动流失，即经营者主动与客户终止服务关系。例如，金融服务商由于客户欺诈等行为而主动终止与客户的服务关系。

第三，客户主动流失，包括两种情况：一种情况是客户因各种原因不再接受相关服务；另一种情况是客户终止当前服务合同而选择其他经营者提供的服务，如手机用户从中国联通转到中国移动。通常客户主动流失的主要原因是，客户认为当前经营者无法提供其所期望的服务，或客户希望尝试其他经营者提供的新服务。

机器学习应用于客户流失分析主要针对上述第三种类型，以客户基本属性和历史消费行为数据为基础，重点围绕客户流失原因分析及客户流失预测两个目标进行数据建模。

1）客户流失原因分析

客户流失原因分析是指找到与客户流失高度相关的因素，如哪些因素是导致客户流失的主要因素，具有哪些属性值或消费行为的客户容易流失等。例如，抵押贷款公司需要了解具有哪些特征的客户会因为竞争对手采用低息和较宽松条款而流失；保险公司需要了解取消保单的客户通常有怎样的特征或行为。只有找到客户流失原因，才可能依此评估流失

电子工业出版社版权所有　盗版必究

客户对经营者的价值，分析哪类客户流失会给企业收入造成严重影响，哪类客户流失会影响企业的业务拓展，哪类客户流失会给企业带来人际关系上的损失，等等。客户流失原因分析的核心目的是为制订客户保留方案提供依据。

机器学习中的分类预测可应用于客户流失原因分析。分类预测的原理和特点等将在后续章节详细讨论。

2）客户流失预测

客户流失预测主要有以下两个方面。

第一，预测现有客户中哪些客户流失的可能性较高，给出一个流失概率由高到低排序的列表。对所有客户实施保留措施的成本很高，只对高流失概率的客户开展维系，将大大降低维系成本。对于流失概率较高的客户，还需要进一步关注其财务特征，分析可能导致其流失的主要原因是财务的还是非财务的。通常非财务原因流失的客户是高价值客户，这类人群一般会正常支付服务费用并对市场活动做出响应，是经营者真正需要保留的客户。给出流失概率列表的核心目的是为测算避免客户流失所付出的维系成本提供依据。

客户流失概率的研究也可通过机器学习中的数据预测建模实现。这些方法的原理和特点等将在后续章节详细讨论。

第二，预测客户可能在多长时间内流失。如果说上述第一方面的分析是为了预测客户在怎样的情况下将会流失，那么第二方面的分析是为了预测客户在什么时候将会流失。

统计学中的生存分析可有效解决上述问题。生存分析以客户流失时间为响应变量建模，以客户的人口统计学特征和行为特征为解释变量，计算每个客户的初始生存率。客户生存率会随时间和客户行为的变化而变化，当生存率达到一定的阈值后，客户就可能流失。

1.3.4 机器学习在营销响应分析中的应用

为发展新客户和推广新产品，企业经营者通常需要针对潜在客户开展有效的营销活动。在有效控制营销成本的前提下，了解哪些客户会对某种产品或服务宣传做出响应等，是提高营销活动投资回报率的关键，也是营销响应分析的核心内容。

营销响应分析的首要目标是确定目标客户，即营销对象。对正确的目标客户进行营销，是获得较高客户响应概率的前提。因为营销通常涉及发展新客户和推广新产品两个方面，所以营销响应分析中的关注点也略有差异。

1）发展新客户

在发展新客户的过程中，可以根据现有客户的数据分析其属性特征。通常具有相同或类似属性特征的客户很可能是企业的潜在客户，应将其视为本次营销的目标客户。

2）推广新产品

在推广新产品的过程中，若新产品是老产品的更新换代产品，或者与老产品较为相似，则可通过分析购买老产品的客户数据发现其属性特征。通常可视这些现有客户为本次营销的目标客户，同时具有相同或类似属性特征的潜在客户也可视为本次营销的目标客户，因为他们很可能对新产品感兴趣。

若新产品是全新的，尚无可供参考的市场和营销数据，则可先依据经验和主观判断确定目标客户的范围，并随机对其进行小规模的试验性营销，然后依据所获得的营销数据找到对营销做出响应的客户的属性特征。具有相同或类似属性特征的现有客户和潜在客户，通常可视为本次营销的目标客户。

电子工业出版社版权所有 盗版必究

确定目标客户之后还需要进一步确定恰当的营销活动。所谓恰当的营销活动，主要是指恰当的营销时间、营销渠道、营销频率，它们与目标客户共同构成营销活动的四要素。对于具有不同特征的目标客户，优化营销渠道和事件触发点，实施有针对性的个性化营销，实现获得客户和营销成本的最优结合，可进一步提高客户响应概率，取得更理想的营销活动投资回报率。

机器学习中的数据预测是营销响应分析的有效工具。这些方法的原理和特点等将在后续章节详细讨论。

1.3.5　机器学习在交叉销售中的应用

交叉销售是在分析客户属性特征及历史消费行为的基础上，发现现有客户的多种需求，向客户销售多种相关产品或服务的营销方式。

例如，保险公司在了解投保人需求的基础上，尽可能为现有客户提供其本人及家庭所需要的其他保险产品，其在为客户介绍某款意外险产品的同时，可以了解客户的其他保险需求并进行推荐，如了解客户的房产状况，为其介绍适合的家庭财险产品；了解客户的家庭成员情况，为其推荐少儿保险；了解客户的支付能力，为其推荐寿险产品；等等。在传统管理和营销模式下，这样的交叉销售被视为一种销售渠道的拓展方式。例如，寿险公司以寿险业务发展成熟为前提条件，通过寿险渠道代理财险业务等。但这种认识正在被慢慢弱化。

交叉销售的深层意义在于主动创造更多客户接触企业的机会，一方面使企业有更多机会深入理解客户需求，提供更适时的个性化服务；另一方面加深客户对企业的信任和依赖程度，从而形成一种基于互动的、双赢的良性循环。交叉销售是提高客户忠诚度，以及提高客户生命周期价值的重要手段，也是一种通过低成本运作(如研究表明交叉销售的成本远低于发展新客户的成本)提高企业利润的有效途径。

交叉销售一般包括产品交叉销售、客户细分交叉销售等主要方面。

1)产品交叉销售

产品交叉销售是指通过分析客户消费行为的共同规律，从产品关联性和消费连带性角度观察，找出最有可能捆绑在一起销售的产品或服务，通过迎合客户需求的产品或服务的组合销售方式提高客户生命周期价值。产品交叉销售并不局限于对同次消费的产品绑定，还包括基于产品使用周期的客户未来时间段消费的预判，并由此在恰当的时间点向客户提供相关产品或服务等。

2)客户细分交叉销售

客户细分交叉销售是对产品交叉销售的拓展。不同特征客户群的消费规律很可能是不同的。客户细分交叉销售强调在客户细分的基础上，依据客户自身的属性特征找到其所属客户群的消费规律，并依此确定交叉销售产品或服务。这种交叉销售关注客户偏好，更有助于提高交叉销售的精准性和个性化程度。

目前产品交叉销售和客户细分交叉销售较常见于电子商务领域的个性化推荐系统中。个性化推荐系统是一个高级商务智能平台，它根据客户的性别、年龄、所在城市等属性特征和相应的消费规律，最热卖的商品，以及高概率的连带销售商品等数据，适时地向不同客户推荐其最可能感兴趣的商品。个性化推荐系统不仅有效缩短了客户浏览和挑选商品花费的时间，而且通过个性化服务创造了更多客户与企业接触的机会。

交叉销售的核心是发现关联性。

1.3.6 机器学习在欺诈甄别中的应用

新技术发展对各行业的欺诈防御带来了新的挑战。高性能的欺诈诊断程序不间断地执行，在欺诈防御失效的第一时间准确甄别出欺诈行为，是有效应对信用卡欺诈、电信欺诈、计算机入侵、洗钱、医药和科学欺诈等的重要手段。

欺诈甄别依据海量历史数据进行分析，涉及两种情况：第一，甄别历史上曾经出现过的欺诈行为；第二，甄别历史上尚未出现过的欺诈行为。

1)甄别历史上曾经出现过的欺诈行为

历史上曾经出现过的欺诈行为在数据上表现为带有明确的是否为欺诈的标签。例如，对已知的银行信用卡恶意透支行为，各账户上均有明确的变量取值，如 1 表示欺诈，0 表示正常。可借助机器学习中的分类预测发现欺诈行为与账户特征之间的一般性规律，并为甄别某个账户是否存在较高的欺诈风险提供依据。

由于欺诈行为的账户特征通常会因防范措施的不断改进而变化，所以欺诈甄别的模型分析结论一般只能作为参考，是否确为欺诈还需要人工判断。为此，欺诈甄别的分类预测不仅要给出判断结果，还要给出一个欺诈风险评分。评分越高，欺诈的可能性越大。按欺诈风险评分从高到低的顺序给出最有可能出现欺诈行为的账户列表供行业专家判断。

2)甄别历史上尚未出现过的欺诈行为

历史上尚未出现过的欺诈行为在数据上表现为没有明确的是否为欺诈的标签。在这种情况下，欺诈可定义为前文提及的模式，其通常具有其他众多数据所没有的某种局部的、非随机的、非常规的特殊结构或相关性。对此，通过数据聚类实现模式诊断是甄别欺诈的主要方法，并且还需要给出相关的欺诈风险评分或概率。

依据按欺诈风险评分排序的账户列表进行人工再甄别的成本通常是较高的。模型的错判损失(包括原本欺诈错判为正常的损失和原本正常错判为欺诈的损失，如因质疑清白账户对客户关系带来的负面影响等)会因行业不同而有高有低。因此，对于上述两种情况，实际欺诈甄别均需要依据行业特点，核算欺诈甄别成本和成功甄别所能挽回的损失，找到两者的平衡点，并最终确定一个欺诈风险评分的最低分数线。高于该分数线的账户需要进行人工再甄别。

本 章 总 结

本章回顾了人工智能的发展历程，指出机器学习是人工智能不断发展的必然产物。从不同视角看，机器学习既是一种新的编程范式，也是一套完整的数据建模方法论。机器学习通过数据建模实现数据预测和数据聚类，有着极高的实际应用价值。本章最后列举了机器学习的典型应用。

本 章 习 题

1. 请举例说明什么是数据集中的变量，并指出变量有哪些类型。
2. 请举例说明什么是机器学习中的数据预测，并指出其中的输入变量和输出变量。
3. 请举例说明机器学习中数据聚类的目的是什么，并指出它与数据预测中的分类预测有怎样的联系和不同。