第1章 绪论

本章学习目标

通过本章学习，学员应该能够：

(1) 了解什么是机器学习；

(2) 了解机器学习的类型和一般步骤；

(3) 了解机器学习在商务决策中的应用；

(4) 熟悉机器学习的发展历程。

引导案例："智慧小苏"

江苏银行作为一家积极拥抱金融科技的金融机构，面临着客户对金融服务个性化、便捷化需求日益增长的挑战。尤其在客户服务领域，传统的人工服务模式已难以满足日益复杂多变的客户需求，同时成本高、效率低的问题也日益突出。

为了应对上述挑战，江苏银行自主研发了"智慧小苏"——一个拥有1760亿参数的大语言模型平台(见图1-1)。该平台采用了先进的自然语言处理、机器学习和深度学习技术，能够处理包括中文对话、提纲写作、摘要生成、信息抽取、数理推理在内的多种任务。"智慧小苏"主要应用于客服场景，以"话务工单助理"身份融入人工客服流程，能够实现智能客服应答、个性化服务推荐、风险预警与处理等功能。

图1-1 "智慧小苏"

“智慧小苏”助力江苏银行实现24小时高效服务，提升客户满意度与忠诚度，同时减轻客服负担，加快处理速度，降低成本。作为大型自研AI模型，“智慧小苏”成为金融业智能化转型的典范，引领金融科技新进展。

资料来源：2023鑫智奖第四届中小金融机构数智化转型优秀案例评选。

1.1 什么是机器学习

机器学习(machine learning，ML)是人工智能领域及计算机科学领域中的一个重要分支，它是一门涉及概率论与数理统计、计算机科学等多学科相互交融的综合性学科。机器学习的基本概念是通过输入大量的训练数据对模型进行训练，从而使模型能够捕捉到数据内部的潜在规律，并据此对新的数据进行有效的分类或预测。机器学习的核心在于运用算法对数据进行解析和学习，进而对新数据做出决策或预测。这一过程类似于人类通过积累经验来对新情境进行判断和预测的学习方式，如图1-2所示。机器学习的过程体现了从数据中学习和模式识别的能力，是现代信息技术发展的重要成果之一。

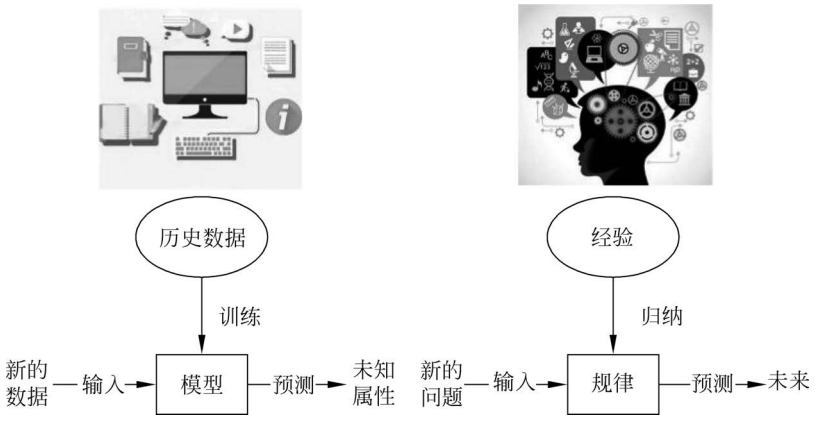

图1-2 机器学习的过程与人的学习过程

以支付宝春节期间的“集五福”活动为例，该活动允许用户通过智能手机扫描包含“福”字的图像，以识别并收集相应的福卡。这一过程正是基于机器学习原理实现的。为了使计算机能够有效地识别“福”字，首先需要向其提供充足的不同字体、不同大小等的“福”字图像数据。通过这些丰富的样本，机器学习模型能够学习并捕捉到“福”字的特征。然后通过算法模型进行训练，模型不断更新学习。当用户输入一张新的“福”字照片，机器自动识别这张照片上是否有“福”字。

机器学习和模式识别、统计学习、数据挖掘、计算机视觉、语音识别、自然语言处理等领域有着很深的联系。从范围上来说，机器学习和模式识别、统计学习、数据挖掘是类似的，同时，机器学习与其他领域的处理技术的结合，形成了计算机视觉、语音识别、自然语言处理等交叉学科，如图1-3所示。因此，一般说数据挖掘时，可以等同于说机器学习。同时，我们平常所说的机器学习应用是通用的，不应仅仅局限在结构化数据，还涵盖图像、音频等应用。

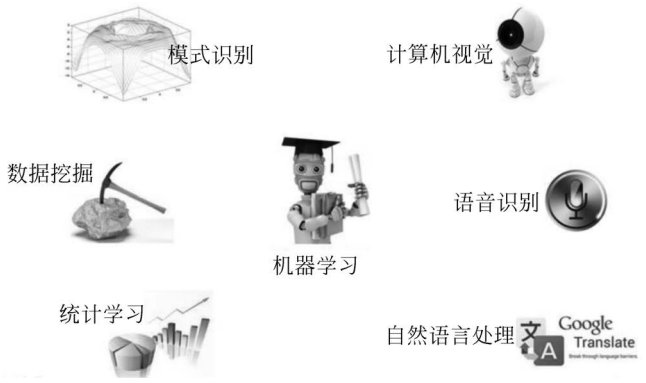

图 1-3　机器学习的范围

1.2　机器学习类型

根据训练数据对人工参与类别划分或标签标识的需求程度，可将机器学习划分为三种主要类型：监督学习、无监督学习和强化学习。机器学习的分类和主要算法见图 1-4。

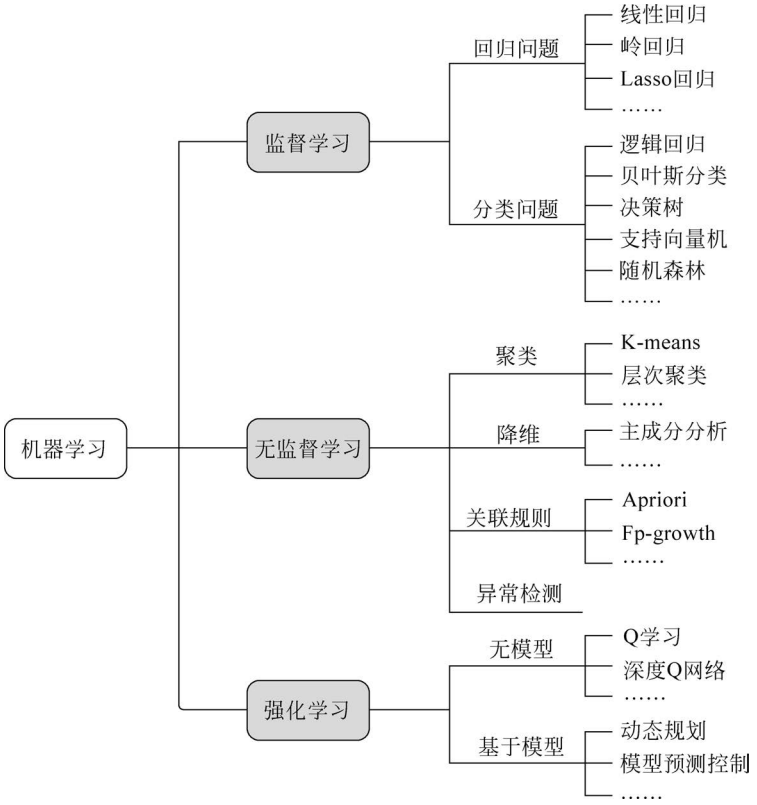

图 1-4　机器学习的分类

在**监督学习**(**supervised learning**)算法中,提供给算法用于模型训练的数据,其实例的类别或标签是需要人工进行标注的。监督学习可用于垃圾邮件识别、文本情感分析、图像内容识别、股价预测等方面。

监督学习算法主要分为两类:回归和分类。回归算法通常根据数据实例的各种属性值,去预测一个目标数值。比如,给出一辆二手车的一些属性值,像车龄、里程数、品牌、型号等,模型预测出该二手车所对应的估值交易价格。分类算法通常用于执行分类操作。比如,给一邮件,模型判断该邮件是不是垃圾邮件。

在**无监督学习**(**unsupervised learning**)算法中,不需要对训练数据进行类别或标签的标注。无监督学习算法主要分为四类:聚类算法、降维算法、关联规则学习和异常检测算法。

聚类算法主要用于对数据实例进行聚合分组。例如,聚类算法可以让我们对网购的消费者进行分类,而不需要人工为算法提供消费者的预定义类型数目。降维算法应用于高维数据的可视化展示、降维优化。例如,我们对二手车价格进行评估的时候,其车龄和里程通常是高度正相关的,通过降维算法就可以将这两个特征进行合并,以简化问题的复杂性。特别是当处理大量数据的时候,降维算法可以有效节约内存空间、存储空间、计算时间等。关联规则学习应用于挖掘实例数据不同属性之间的关联关系。例如,针对网购消费者,通过关联规则学习可以发现,买了羊蹄和羊肉串的顾客,通常也会买些孜然粉。异常检测算法主要应用于对异常实例的检测发现。例如,挖掘交易事务流中的异常交易。

强化学习(**reinforcement learning**),主要基于"行动+赏罚"机制,用于描述和解决智能体在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题。例如走路机器人、DeepMind 的 AlphaGo 等。按是否建立环境模型可以分为无模型的强化学习和基于模型的强化学习。相比于前述两种学习机制,强化学习无疑另辟蹊径,自成一体。本书不深入讲解强化学习,感兴趣的读者可自行探索学习。

1.3 机器学习一般步骤

机器学习从有限的观测数据中学习出具有一般性的规律,并使用这些规律对未知数据进行预测,整个过程主要包括数据采集、数据预处理、特征工程和数据建模 4 个步骤。

数据采集:由于机器学习是从数据中进行学习的方法,所以首先要针对想要解决的问题进行数据的采集。数据的采集主要有两种途径,一种是自己采集,另一种就是去网上找公开的数据集。数据采集完成后,就得到了原始的数据。

数据预处理:从数据中检测,纠正或删除损坏、不准确或不适用于模型的数据的过程。可能面临的问题有:数据类型不同,比如有的是文字,有的是数字,有的连续,有的间断。也可能,数据的质量不行,有噪声,有异常,有缺失,数据出错,量纲不一等。

特征工程:从原始数据提取、选择、转换或创建特征的过程,以提高机器学习模型的性能和准确性。它是连接原始数据与机器学习模型之间的桥梁,通过一系列工程化的手段,将原始数据转换为模型能够有效利用的形式。

数据建模:包括机器学习模型选择、参数调整、在测试集上评估最佳模型、解释模型结果、得出结论等。

数据建模中，**模型、学习准则与优化算法**是机器学习的三大要素。

模型的作用是根据输入的特征给出输出的结果（针对具体的问题），也可以将模型理解为函数。不同的机器学习模型实质上是不同的待选择函数簇。当模型的类型确定后，函数的大体框架就确定了，剩下的就是对函数中的参数的学习。所以，机器学习的本质就是在一堆由不同的参数所决定的函数里面，选出最好的那个（一个优化问题）。

学习准则的作用是针对想要解决的问题，评价某一个模型的好坏程度。在监督学习中，一般是看模型的输出与数据集中的真值的差异，差异越小，一般就代表模型越好。

优化算法的作用是对选出最好的模型这个优化问题进行求解。

这三大要素确定好之后，将数据集带入其中，即可训练出一个在当前的数据集情况下的最优模型。

1.4 机器学习在商务决策中的应用

1.4.1 商务决策管理

研究表明，经理们在平均一周的工作时间里可能需要做出数百个决定。然而，有些决定是如此复杂，以至于一个决定可能需要几周、几个月，甚至更长时间才能最终确定。无论你的工作场所的决策有多少或多复杂，重要的是要知道决策是管理者的核心管理能力，并有机会审查他们的方法，在必要时做出调整。有效的管理决策往往包括选择和应用适当的工具，并意识到每一种工具在哪里可以发挥作用。

（1）问题分析与决策制定的区别

虽然问题分析和决策制定是相互关联的过程，但它们在本质上是不同的活动。决策制定通常针对一个具体的问题或挑战，涉及在多个备选方案中进行选择。在做出决策之前，决策者需要收集和评估相关数据，以确保选择的最优性。问题分析包括通过定义问题的边界来确定问题的框架，建立从备选方案中选择的标准，以及根据现有信息得出结论。分析问题虽然是决策过程中的关键环节，但并不直接导致决策的产生。尽管如此，分析结果仍是所有决策制定中不可或缺的组成部分。

（2）决策制定的步骤

管理决策可以被视为有两个关键组成部分：内容和过程。内容指的是决策所依据的数据、信息和知识。而过程指的是做出决策所经历的步骤。虽然内容对每个决策都是独一无二的，但无论决策是简单还是复杂，都应该包含以下步骤。

- 确定决策问题；
- 认识到在决策时应该咨询谁以及为什么；
- 从适当的来源收集正确的决策内容，包括酌情咨询；
- 通过使用机器学习方法分析内容和生成选项；
- 批判性地评估备选方案；
- 选择最好的方案，做出决策；
- 传达决策；

• 执行决策并审查这一决策的影响。

尤其是当管理者处于压力之下时，这些步骤中的一个或多个往往会被忽略或妥协。那么，本可以是一个有效的决定就会变成一个产生不利后果的决定。例如，花时间正确地定义问题为接下来的步骤提供了坚实的基础。当管理者将症状误认为正确的问题定义时，他们可能会冒险收集和处理错误的数据。当管理者避免咨询其他可能会受到决策影响的人时，他们可能会疏远那些原本可能会致力于决策的人。

1.4.2 机器学习的应用领域

机器学习帮助企业识别通常不被注意或隐藏的重要事实、趋势、模式、关系和异常。如今，由于机器学习在各种行业和学科的分析工作中发挥着核心作用，它被广泛应用于不同的领域。表 1-1 列出了机器学习的应用领域及其使用情况。通过分析顾客的购买模式，零售商可以想出更聪明的营销促销活动，增加销售额。通过市场细分，零售商可以识别购买相同产品的顾客。因此，他们可以通过分析客户的兴趣和人口统计数据，在正确的时间推出新产品。机器学习还可以用来预测客户的需求，谁最有可能从这些市场竞争对手转移购买。

表 1-1 机器学习在一些代表性领域的应用和用途

应　　用	用　　途
通信行业	机器学习技术用于通信领域，以预测客户行为，提供高度针对性和相关性的活动。
保险业	机器学习帮助保险公司为其产品定价，使其有利可图，并向其新客户或现有客户推广新产品。
教育	机器学习帮助教育工作者获取学生数据，预测成绩水平，找到需要额外关注的学生或学生群体。例如，数学成绩差的学生。
制造业	通过机器学习技术，制造商可以预测生产资产的磨损情况，可以预期维护，这有助于减少停机时间。
银行业	机器学习帮助金融部门了解市场风险和管理合规。它帮助银行识别可能的违约者，以决定是否发放信用卡、贷款等。
零售行业	机器学习技术帮助零售商场、杂货店分析顾客的购物篮数据(即顾客在一次购物中购买的商品集合)，识别出潜在的商品关联规则，从而优化货架布局，将相关商品摆放得更近，促进连带销售。
服务提供商	手机和公用事业行业等服务提供商使用机器学习来预测客户离开公司的原因。服务提供商分析账单细节、客户服务互动、公司收到的投诉，然后给每个客户打分并提供奖励。
电子商务	电子商务网站利用机器学习提供交叉销售和向上销售。使用机器学习算法让更多的客户进入电子商务商店。
犯罪调查	机器学习帮助预测犯罪最可能发生的地点和时间，有助于警局合理部署警力。
生物信息学	机器学习有助于从生物学和医学中收集的海量数据集中挖掘生物数据。

1.5 机器学习发展

1.5.1 机器学习的发展历程

机器学习实际上已经存在了几十年，或者也可以认为存在了几个世纪。追溯到 17 世纪，贝叶斯(Bayes)、拉普拉斯(Laplace)关于最小二乘法的推导和马尔科夫链，这些构成了

机器学习广泛使用的工具和基础。

自1950年阿兰·图灵(Alan Turing)提出图灵测试机,到21世纪有深度学习的实际应用,机器学习有了很大的进展。从20世纪50年代研究机器学习以来,不同时期的研究途径和目标并不相同,可以划分为四个阶段。

(1) 知识推理期

知识推理期始于20世纪50年代中期,当时普遍认为,一旦机器被赋予逻辑推理的能力,它就能够展现出智能行为。这一阶段的代表性工作有赫伯特·西蒙(Herbert Alexander Simon)和艾伦·纽厄尔(Allen Newell)共同开发了"逻辑理论家"(Logic Theorist)程序,证明了著名数学家伯特兰·罗素(Bertrand Russell)和怀特海(Alfred North Whitehead)的经典著作《数学原理》中的全部52条定理,并且其中一条定理甚至比罗素和怀特海证明得更巧妙。然而随着研究的深入,人们逐渐意识到,单纯的逻辑推理能力并不足以实现真正意义上的人工智能,要使机器具有智能,就必须设法使机器具有知识。

(2) 知识工程期

从20世纪70年代中期开始,人工智能进入知识工程期。这一时期大量专家系统问世,在很多应用领域取得了大量成果,费根鲍姆作为知识工程之父在1994年获得了图灵奖。由于人工无法将所有知识都总结出来教给计算机系统,所以这一阶段的人工智能面临知识获取的瓶颈。

在这个时期,研究的重点是将各领域的知识植入到系统里,目的是通过机器模拟人类的学习过程。研究者们采用了图结构和逻辑结构来详细描述系统,并使用各种符号来表示机器语言。研究内容从学习单一概念扩展到掌握多个概念,探索不同的学习策略和方法。同时,学习系统开始与实际应用相结合,并取得了很大的成就。专家系统对知识获取的高需求,极大地促进了机器学习领域的深入研究和快速发展。

(3) 归纳学习期

1980年夏,在美国卡耐基梅隆大学举行了第一届机器学习研讨会(IWML);1983年Tioga出版社出版了里夏德·S. 米哈尔斯基(Ryszard S. Michalski)、海梅·吉列尔莫·卡博内尔(Jaime Guillermo Carbonell)和汤姆·米切尔(Tom R. Mitchell)主编的《机器学习:一种人工智能途径》,对当时的机器学习研究工作进行了总结;1986年,第一本机器学习专业专刊 *Machine Learning* 创刊;1989年,人工智能领域的权威期刊 *Artificial Intelligence* 出版机器学习专辑,刊发了当时一些比较活跃的研究工作。总的来看,20世纪80年代是机器学习成为一个独立的学科领域、各种机器学习技术百花初绽的时期。

20世纪80年代以来,被研究最多、应用最广的是"从样例中学习",即从训练样例中归纳出学习结果,也就是广义的归纳学习,它涵盖了监督学习和无监督学习等。在20世纪80年代,"从样例中学习"的一大主流是符号主义学习,其代表包括决策树和基于逻辑的学习。典型的决策树学习以信息论为基础,以信息熵的最小化为目标,直接模拟了人类对概念进行判定的树形流程;基于逻辑的学习的著名代表是归纳逻辑程序设计,可以看做机器学习与逻辑程序设计的交叉,它使用一阶逻辑(即谓词逻辑)来进行知识表示,通过修改和扩充逻辑表达式(例如Prolog表达式)来完成对数据的归纳。符号主义学习占据主流地位与整个人工智能领域的发展历程是分不开的。

20 世纪 90 年代中期之前,“从样例中学习”的另一主流技术是基于神经网络的连接主义学习。连接主义学习在 20 世纪 50 年代取得了大发展,但因为早期的很多人工智能研究者对符号表示有特别偏爱,所以当时连接主义的研究未被纳入人工智能主流研究范畴。1983 年,霍普菲尔德利用神经网络求解“流动推销员问题”这个著名的 NP 难题取得重大进展,使得连接主义重新受到人们关注。1986 年,著名的 BP 算法诞生,产生了深远的影响。

20 世纪 90 年代中期,统计学习出现并迅速占据主流舞台,代表性技术是支持向量机(SVM)以及更一般的“核方法”。这方面的研究早在 20 世纪 60 年代就已经开始,统计学习理论在那个时期也已打下了基础,但直到 90 年代中期统计学习才开始成为机器学习的主流。一方面是由于有效的支持向量机算法在 90 年代初才被提出,其优越性能到 90 年代中期在文本分类应用中才得以显现;另一方面,正是在连接主义学习技术的局限性凸显之后,人们才把目光转向了以统计学习理论为直接支撑的统计学习技术。在支持向量机被普遍接受后,核方法被广泛应用于机器学习的诸多领域,并逐渐成为机器学习的基础性内容之一。

(4) 深度学习

21 世纪初,连接主义学习又卷土重来,掀起了以“深度学习”为名的热潮。2006 年,深度学习概念被提出。2007 年,希尔顿(Hinton)发表了深度信念网络论文,约书亚·本吉奥(Yoshua Bengio)等人发表了逐层训练方法的论文 *Greedy Lay—Wise Training of Deep Networks*,扬·勒丘恩(YannLeCun)团队发表了 *Efficient Learning of Sparse Representations with an Energy-Based Model* 论文,这些研究成果标志着人工智能正式进入了深层神经网络的实践阶段。同时,云计算和 GPU 并行计算为深度学习的发展提供了基础保障,特别是最近几年,机器学习在各个领域都取得了突飞猛进的发展。

1.5.2 机器学习中的问题

新的机器学习算法面临的主要问题更加复杂,机器学习的应用领域从广度向深度发展,这对模型训练和应用都提出了更高的要求。随着人工智能的发展,冯·诺依曼式的有限状态机和理论基础越来越难以应对目前神经网络中层数的要求,这些都对机器学习提出了挑战。在机器学习发展成为一种成熟的和可信的学科之前,还有许多悬而未决的问题需要解决。下面将讨论其中一些问题。

安全和社会问题:对于任何共享或打算用于决策分析的数据,安全都是一个重要问题。此外,在收集数据进行客户分析、用户行为理解、个人数据与其他信息的关联等时,大量个人或组织的敏感隐私信息被收集和存储。考虑到这些数据的机密性和对信息的潜在非法访问,机器学习变得有争议。此外,机器学习可能会披露关于个人或团体的新的隐性知识,这些知识可能违反隐私政策,特别是在发现的信息可能被传播的情况下。由此引起的另一个问题是机器学习的适当使用。由于数据的价值,各种内容的数据库经常被出售,发现隐性知识可以获得竞争优势,因此,一些重要的信息可以被保留,而其他信息可以不受控制地广泛传播和使用。

偏见问题:人类会有偏见,有些人规避风险,有些人则是冒险家。有些人天生关心他人,有些人则不敏感。人们可能会认为,机器的一个优点是它们能做出合乎逻辑的决定,而

且根本不受偏见的影响。然而事实并非如此。机器学习算法表现出许多偏见。这种偏见与收集到的数据有关。它可能不具有代表性。这里有一个经典的例子(早在机器学习出现之前),那就是《文学文摘》试图预测 1936 年美国总统大选的结果。该杂志调查了 1000 万人(样本很大),收到了 240 万份回复。它预测兰登(共和党)将以 57.1%比 42.9%的优势击败罗斯福(民主党)。事实上,罗斯福赢了。到底是哪里出了错?答案是,《文学文摘》使用了一个有偏见的样本,包括《文学文摘》的读者、电话用户和那些有汽车登记的人。这些人主要是共和党的支持者。

机器学习有一种自然的倾向,即使用现成的可用数据,并倾向于支持现有的实践。未来用于贷款决策的数据很可能是过去实际发放的贷款数据。如果能知道过去没有发放的贷款是如何计算出来的就好了,但从本质上讲,这一数据是无法获得的。在开发机器学习算法时,分析师还可以通过许多其他方式(有意识或无意识地)表现出偏见。例如,数据清理的方式、模型的选择以及解释和使用算法结果的方式可能会受到偏见的影响。

机器学习方法问题:机器学习方法的通用性、可用数据的多样性、领域的维数、广泛的分析需求(已知时)、所发现知识的评估、背景知识和元数据的利用、数据中噪声的控制和处理等都是影响机器学习技术选择的问题。例如,通常认为可以使用不同的机器学习算法,因为不同的方法可能会根据手头的数据执行不同的操作。此外,不同的方法可能是合适的,满足不同的客户需求。然而,大多数算法都假设数据是无噪声的,这显然是一个很强的假设。许多数据集包含异常、无效或不完整的信息,或异常数据等,这些可能会使分析过程变得复杂,甚至模糊,在许多情况下会损害结果的准确性。因此,数据预处理(数据清洗和转换)变得非常关键。虽然数据清洗通常被认为是耗时和令人沮丧的,但它是数据挖掘和知识发现过程中最重要的阶段之一。机器学习技术应该能够处理数据中的噪声或不完整信息。

值得注意的是,对于数据挖掘技术来说,搜索空间的大小甚至比数据大小更具有决定性意义,这通常取决于域空间中的维数。当维数增加时,它通常呈指数增长。这就是众所周知的“维度诅咒”。这一“诅咒”严重影响了一些机器学习算法的性能,因此成为最迫切需要解决的问题之一。

过拟合和欠拟合:当与给定数据集关联的模型生成时,希望该模型也适合其他数据集。通常,学习算法是使用一组已知期望输出的“训练数据”来训练的。我们的目标是,当输入训练过程中没有遇到的“验证数据”时,算法也将在预测输出方面表现良好。当模型在训练数据上学习得太好,以至于学到了训练数据中的噪声和细节,导致模型泛化能力差(模型在新的、未见过的数据上表现不佳),就会发生过拟合。表现为模型偏差低但方差高,即模型对训练数据中的随机噪声反应敏感。例如拥有过多的参数,或者训练时间过长,使得模型对训练数据中的噪声也进行了学习。当模型在训练数据上没有获得足够的学习,以至于无法捕捉到数据的基本结构,既不能在训练数据上表现良好,也不能在新的数据上做出准确的预测,就会出现欠拟合。表现为模型方差低但偏差高,即模型预测的准确度低。例如使用线性模型来处理非线性问题,或者训练时间不足,数据特征提取不够等。过拟合和欠拟合都是模型泛化能力不足的表现,过拟合是由于模型过于复杂,而欠拟合则是由于模型过于简单。两者都会影响模型在新数据上的表现,因此,寻找合适的模型复杂度是提高预测性能的关键。

1.5.3 机器学习的未来

机器学习技术极大地转变了商业运营的方式。在众多行业中，每时每刻都在产生海量的数据，这就要求我们掌握相应的工具和技术来有效处理这些数据。在当前这个数据激增、信息泛滥、竞争激烈的时代，任何组织或企业的决策都不应仅凭经验行事。机器学习已经成为商务决策不可或缺的部分，如揭示产品间的购买关联，助力企业在商品布局和促销活动中提升交叉销售和增销效果。随着机器学习算法和软件的不断进步，管理者能够更准确地做出决策，以实现利润的最大化。尽管机器学习仍面临诸多挑战，但我们有理由相信，机器学习算法和技术将会不断优化，以应对未来数据的复杂性。

练 习 题

1. 简述机器学习的定义及其类型。
2. 简述机器学习的一般步骤。
3. 请找到一些已经在管理决策中使用的机器学习应用的例子。

即测即练题