思想引领

第3章 机器学习

机器学习是人工智能的核心技术，旨在让计算机通过数据自动学习规律并具备预测与决策能力，无须显式编程即可完成复杂任务。在数据爆炸式增长的今天，机器学习已成为从海量信息中提取价值、实现智能化处理的关键支撑，广泛应用于生活、生产、科研与社会治理等各个领域。根据数据是否带有标注、学习目标是否明确，机器学习主要分为有监督学习与无监督学习两大分支。有监督学习依托标注数据构建预测模型，实现分类与回归；无监督学习从无标签数据中挖掘内在结构，完成聚类、降维与分布估计。二者互为补充，共同构成机器学习的基础理论体系。本章将系统介绍机器学习的基本概念、核心方法、典型算法与适用场景，帮助读者理解算法原理、优缺点与实际应用，为进一步学习深度学习与人工智能技术奠定坚实基础。

3.1 机器学习概述

3.1.1 机器学习的定义与核心思想

学习是人类与动物适应环境、获取技能、提升能力的基本方式，指通过观察、实践、积累经验，使自身行为与认知发生相对持久改变的过程。人们在生活中不断接触事物、总结规律，用已有的经验判断新的问题，例如，儿童见过多种植物后，便能识别从未见过的新植物；医生积累大量病例后，可依据症状快速判断病情；驾驶员经过反复练习，能在复杂路况中安全行驶。这种“从经验中归纳规律，用规律应对新情况”的能力，是智能最核心的体现。传统计算机程序则完全依赖人工编程，由开发者预先设定清晰、固定的执行规定，逐条告诉计算机在何种条件下执行何种操作，计算机只负责机械地按照指令运行，不具备自主总结、自主调整的能力。当场景复杂、规则多变或数据量巨大时，人工编写规则不仅效率低下，更难以覆盖全部可能性，无法适应现实世界的不确定性。正是为了突破这一局限，人们开始研究如何让计算机也具备类似人类的学习能力，机器学习由此应运而生。

机器学习（Machine Learning，ML）是人工智能的核心分支，也是计算机科学的重要研究领域。它致力于研究如何通过计算手段与算法设计，使计算机系统在不依赖人工显式编程的前提下，自动从数据与经验中学习模式、优化性能，并利用学到的规律对未知数据进行预测、分类与决策。学术界对机器学习的经典描述为：针对某一类任务和对应的性能度量，计算机程序能够随着经验的不断积累，持续提升在该任务上的执行效果，即可认为它具备了

机器学习能力。从20世纪50年代初以Arthur Samuel发明的西洋棋自我学习程序为代表的早期模式识别阶段，到20世纪80年代反向传播算法推动人工神经网络发展，再到21世纪初Google Brain项目、AlexNet在ImageNet竞赛中取得突破性成绩后席卷全球的深度学习阶段，机器学习始终沿着“数据更丰富、算法更高效、模型更强大”的方向发展，逐步从实验室走向工业界、融入日常生活，成为支撑语音识别、图像分类、推荐系统、自动驾驶等技术落地的关键基础。

机器学习的整个过程围绕数据、算法、模型三大核心要素展开，三者紧密关联、相互支撑，共同决定系统的学习效果与应用能力。数据是机器学习的基础与前提，是经验的载体，没有高质量、足够规模的数据，算法便失去学习对象，模型也无法被有效训练。在机器学习中，数据以样本的形式存在，每个样本包含描述对象的特征，部分样本还带有用于指导学习的标签，数据的完整性、多样性、准确性直接影响模型能否学到稳定、通用的规律。算法是机器学习的核心驱动力，是实现学习过程的规则与步骤，它决定了如何读取数据、提取特征、度量误差、更新参数，不同的算法适用于不同的任务类型与数据结构，算法的设计与优化，直接决定模型能否高效收敛、准确捕捉数据中的潜在模式。模型是算法在数据上学习后得到的最终产物，是数据规律的数学抽象与表示，它封装了从特征到输出的映射关系，能够直接用于预测与决策。模型的能力取决于数据的质量与算法的选择，同时，模型在使用中产生的新数据与反馈，又可以反过来优化算法、补充数据集，形成持续迭代的闭环。简单来说，数据为学习提供原料，算法为学习提供方法，模型是学习的最终成果，三者协同工作，让计算机从被动执行指令的工具，转变为能够自主学习、自主判断的智能系统。

机器学习算法与基于规则的推理方法的核心区别在于，基于规则的推理方法由人根据经验提前判断、预设固定逻辑与判断条件，机器仅按照给定规则执行匹配与推理，而机器学习算法直接从数据出发，由算法自动归纳、提取并生成判断规则，无须人工逐条定义处理逻辑。以生活中常见的自动驾驶避让行人为例，基于规则的方法只能人工设定“行人在前方、距离小于多少米就刹车”等固定条件，当遇到行人弯腰、跑步、推婴儿车等复杂姿态或雨天、逆光、遮挡等场景时原规划就容易失效，必须不断手动补充规则，这属于无人驾驶技术里的基于规则的路径规划模式。而机器学习属于数据驱动的规划模式，只需输入大量路况与行人行为数据，结合深度学习和计算机视觉算法，系统便能自主学习行人动作、距离、光线、环境等关联规律，面对行人突然横穿等紧急情况，或是雨天这类恶劣环境，都能灵活判断并安全避让，不用人工频繁修改判断条件，适应性与泛化能力更强。基于规则的推理方法和机器学习算法的区别如图3-1所示。

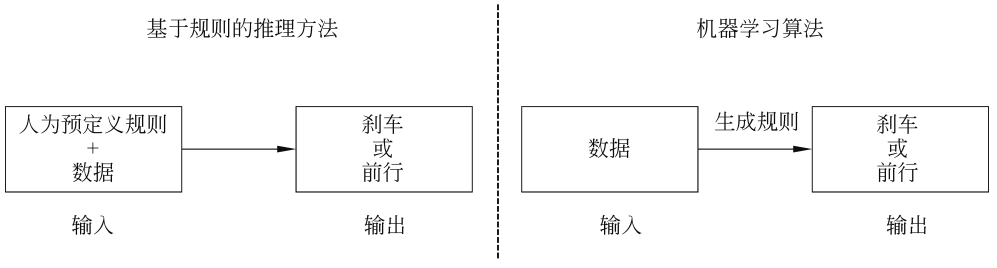


图3-1　基于规则的推理方法和机器学习算法的区别

机器学习的核心思想正是模拟人类“从经验中归纳规律,用规律泛化到新场景”的认知过程,以数据驱动替代规则驱动,以归纳推理替代固定逻辑,让系统在不断接触数据的过程中自主优化,最终具备对未知场景的适应与判断能力。这种思想彻底改变了计算机处理复杂问题的方式,使得人工智能核心技术之一的机器学习能够在医疗、金融、制造等众多领域落地实用,成为推动数字社会与智能产业发展的核心技术。

3.1.2 机器学习的基本概念

在学习机器学习的过程中,会频繁遇到一系列专用术语与核心概念,这些概念贯穿数据处理、模型训练、算法选择与效果评估的全过程,是理解机器学习工作机制、阅读技术资料、开展实践操作的基础。熟练掌握这些基本概念,能够建立清晰的认知框架,为深入学习有监督学习、无监督学习、深度学习等内容奠定扎实的认知基础。本节将按照从数据到模型、从训练到评估、从任务类型到常见问题的顺序,逐一介绍机器学习中最常用、最重要的基础概念。

1. 数据与样本的相关概念

(1)数据集与样本:数据集是机器学习的起点,指由多条数据记录按一定结构组成的集合,是机器用于学习的“原材料”。数据集中每条独立的记录,对应一个被描述的对象或事件,被称为样本。例如,在“房价预测”任务中,包含上千条房屋面积、城区、房龄、楼层、朝向及真实售价的表格,就是一个数据集;其中每行对应一套房屋的全部信息,这套房屋即为一个样本。数据集的规模、质量与多样性,直接决定机器学习模型能否学到稳定可靠的规律。

(2)特征:特征也被称为属性,是数据集中用来描述样本特点的各项信息,是模型进行学习与判断的依据,相当于给样本贴上的“可识别标签”。特征可以是数值、文字、类别等多种形式。在房价预测任务中,房屋面积、所在城区、房龄、楼层、朝向、与地铁站的距离等,均为典型特征。特征质量直接影响模型精度,优质特征能显著简化学习过程、提升预测效果。

(3)维度:维度指一个样本所包含的独立特征数量,用来衡量数据的复杂程度与信息丰富度。一个样本包含多少个不同的特征,就称其为多少维数据。例如,只用“房屋面积”一个特征预测房价,就是一维数据;同时使用“房屋面积”“所在城区”两个特征,就是二维数据;若进一步加入房龄、楼层、朝向等特征,数据维度便会随之升高。在图像识别、自然语言处理等任务中,一张图片可对应数万甚至数百万维的像素数据,属于高维数据,高维数据会增加计算复杂度,因此需要进行降维处理。

2. 模型与训练的核心概念

(1)模型:模型是机器学习算法在数据上完成训练后得到的最终产物,是对数据内在规律的数学抽象与表达,可理解为封装了学习结果的“预测工具”。模型的核心作用是建立输入特征到输出结果的映射关系,用于预测、分类或决策。例如,基于历史房价数据训练出的模型,可依据房屋特征预测其市场售价;基于海量邮件数据训练出的分类模型,可判别新接收邮件是否为垃圾邮件。

(2)训练集:训练集是用于让模型学习数据规律、调整内部参数的数据集,是模型“积累经验”的核心来源。例如,在垃圾邮件识别任务中,训练集包含数千封已标注“垃圾邮件”“正常邮件”的邮件数据,模型通过学习这些邮件的文本特征,形成判别规则。

（3）验证集：验证集在训练过程中使用，不参与模型参数的直接优化，主要用于调整超参数及选择最优模型结构，并实时监测模型是否出现过拟合。验证集的作用相当于“训练中的监考员”，帮助开发者在训练过程中及时调整配置，提升模型泛化能力。

（4）测试集：测试集完全独立于训练与验证过程，其中的数据在模型训练期间从未被模型接触过，专门用于模型训练完成后的最终性能评估，衡量模型在真实场景中的可用程度。测试集的评估结果最具客观性，能够真实反映模型的泛化能力，避免模型在训练集或验证集上“背答案”导致的效果虚高。万级以下的小规模数据集通常按 6∶2∶2的比例划分训练集、验证集、测试集；百万级以上的大规模数据集则可采用 98∶1∶1或 99.5∶0.3∶0.2 的比例，只需保证验证集与测试集样本数量达到 1 万个左右即可，无须严格遵守小规模数据集的划分比例。

（5）训练：训练也称为学习，指利用机器学习算法在训练集上迭代运行，让模型从数据中捕捉规律、缩小预测误差、提升精度的全过程。训练的本质是让模型从“一无所知”到“掌握规律”，从无法预测到稳定输出结果。在训练过程中，算法会不断计算模型在训练集上的预测值与真实值之间的经验误差，并依据经验误差调整参数，使经验误差逐步缩小，同时兼顾模型的泛化能力，直到模型达到预设的性能要求。

3. 机器学习的典型任务

（1）分类：分类的预测目标是离散的类别标签，模型需要将样本划分到预定义的类别中。分类任务的输出是固定的选项，而非连续数值。例如，判断邮件是垃圾邮件还是正常邮件、判断图片中的动物是猫还是狗、判断学生成绩是否及格、判断信用卡交易是否为欺诈，都是典型的分类任务。其中，垃圾邮件识别、图像分类（猫狗识别）属于有监督学习范畴的分类任务，而信用卡交易欺诈检测常作为无监督学习中的异常检测类任务，也归属于广义的分类任务范畴。分类又可分为二分类与多分类，二分类只有两个类别，多分类则包含三个或三个以上类别。

（2）回归：回归的预测目标是连续的数值，模型需要输出一个具体的数字，而非固定类别。例如，根据房屋特征预测房价、根据历史气温预测明日温度、根据用户信息预测消费金额、根据学习数据预测考试分数，都是典型的回归任务。回归任务关注数值的精确程度，误差越小说明模型效果越好。

（3）降维：降维指在尽可能保留数据核心信息的前提下，减少特征的数量，降低数据维度。高维数据会带来计算量大、训练慢、模型易过拟合等问题，降维能够简化数据、提升效率，同时便于数据可视化。例如，将包含 20 个特征的学生学习数据降维为 2～3 个核心特征，既不丢失关键信息，又能快速训练、直观展示数据分布。

（4）聚类：聚类指按照样本之间的相似程度，自动将数据划分为多个组别（簇），使同一个组内的样本高度相似，不同组之间差异明显。聚类不需要提前定义类别，由算法自主发现分组规则，典型案例包括电商平台根据用户行为将用户分为高消费人群、价格敏感人群、低频消费人群，社交网络根据互动关系发现用户社区，异常检测中将明显偏离正常数据的样本归为异常簇等。

4. 模型训练相关术语

（1）泛化能力：泛化能力是机器学习模型最重要的评价指标之一，指模型在从未见过的新数据上做出准确预测的能力。机器学习的目标不是让模型完美记住训练数据，而是能

够将学到的规律应用到真实场景中,泛化能力差的模型,只能在训练数据上表现良好,无法应对真实环境。

(2) 欠拟合:欠拟合指模型过于简单,学习能力不足,没有充分捕捉到训练数据中的核心规律,在训练集与测试集上的表现都很差,预测误差较大。欠拟合的原因通常是模型结构过于简单、特征数量不足、训练轮数过少等。例如,用直线去拟合曲线分布的房价数据,由于直线模型过于简单,无法捕捉数据的真实模式,因此无论怎么训练都无法提升精度。

(3) 过拟合:过拟合指模型过于复杂,不仅学到了数据的真实规律,还记住了训练集中的噪声、偶然特例与随机误差,导致在训练集上的表现近乎完美,但在测试集与真实场景中的效果急剧下降。过拟合多由训练数据量不足、数据噪声过多、模型过度复杂导致。例如,一个图像分类模型在训练集上100%正确,但遇到轻微模糊、光线变化的新图片就识别错误。这类模型因为学习能力过强,过于精确地匹配了训练集的特定数据,甚至捕捉到了训练样本的局部特征或噪声,导致泛化能力极弱,无法良好地拟合新的测试数据。

过拟合和欠拟合是机器学习模型训练中的两种常见的问题,反映了模型复杂度和泛化能力之间的平衡。过拟合和欠拟合的示意图如图3-2所示。在机器学习中,随着模型越来越复杂,模型对训练数据的拟合能力不断增强,训练误差会单调减小。但泛化误差则是先减小后增大:当模型复杂度较低时,拟合能力不足,存在欠拟合问题,泛化误差较高;随着复杂度提升,模型逐渐能捕捉数据的潜在规律,泛化误差随之降低;可当模型复杂度过高时,会过度拟合训练数据中的噪声,引发过拟合,此时泛化误差又会上升。在表示最佳容量的虚线左侧,训练误差和泛化误差都较大,模型出现欠拟合,说明在训练集上表现很差,未能充分学习数据中的规律,此时可以采取增加模型复杂度(如决策树或人工神经网络)、补充有效特征(加入特征组合、高阶特征等)、延长训练时间等方法改善;在表示最佳容量的虚线右侧,当训练误差较小,而泛化误差开始增大时,出现过拟合,说明模型过于复杂,捕捉到了训练数据中不能推广到新数据的噪声和偶然模式,为了解决过拟合问题可以采取扩充训练数据、降低模型复杂度、使用正则化约束、提前停止训练、采用集成模型等方法。过拟合和欠拟合都是需要避免的,理想的效果是将训练误差和泛化误差都控制在合理的范围内,训练误差较小,而泛化误差略大于训练误差。

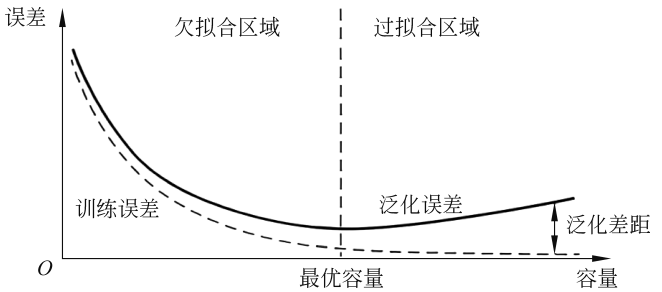


图3-2 过拟合和欠拟合的示意图

3.1.3 机器学习的分类

根据训练样本所提供的监督信息、标签形式以及模型与环境的交互方式不同,机器学习可划分为有监督学习、无监督学习、半监督学习、强化学习四大基本类型。这四类学习方法

是机器学习领域最核心、最常用的分类方式，它们的主要区别在于训练过程中是否提供标签、标签数量多少，以及是否通过与环境交互获得奖惩反馈，分别适用于不同的数据条件、任务场景与应用目标，共同构成了机器学习完整的方法体系。

为便于直观理解，可通过二维特征空间示意图对四类学习方法的差异进行描述：假设样本由两个特征构成，在平面坐标系中分布，监督学习的样本带有明确标签（正类或反类），模型在“标准答案”指导下学习输入到输出的映射关系，目标是预测与分类；无监督学习的样本无任何标签，模型自主探索数据内部结构、相似性与分布规律，目标是发现隐藏模式；半监督学习仅少量样本有标签，大量样本无标签，机器以从有标签样本中获得的知识为基础，结合无标签样本的分布情况逐步修正已有的知识，并判断测试样本的类别；强化学习则不依赖静态样本标记，而是通过与环境持续交互获得奖励或惩罚反馈，学习最优决策策略。机器学习的分类如图 3-3 所示。

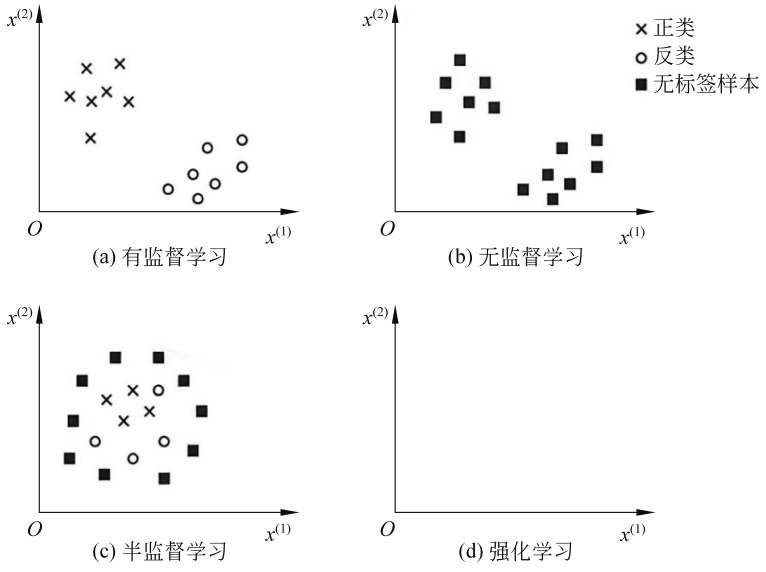


图 3-3　机器学习的分类

1. 有监督学习

有监督学习也被称为有导师学习，这里的“导师”就是训练数据中的标签。它的核心目标是通过标记好的训练集学习输入特征到输出标签之间的统计映射规律，从而构建出可对测试集中未标记数据进行预测的模型。在监督学习中，每组训练数据有一个明确的类别标签或结果，模型通过将模型的预测结果与训练集的真实结果进行比较，计算误差，不断调整优化模型，直到模型的预测结果达到一个预期的准确率。有监督学习凭借训练目标清晰、预测精度较高的优势，成为目前工业界与日常生活中应用最广泛的机器学习方法。从任务类型来看，它可分为回归与分类两大类别：回归任务可实现房价预测、股票价格预测等连续数值预测；分类任务则覆盖图像分类、人脸识别、物体检测、疾病诊断、垃圾邮件识别、机器翻译等场景。除此之外，有监督学习在语音识别、医学影像分析、金融风险评估等领域也发挥着关键作用，助力各行业提升效率与准确性。根据输出结果类型，有监督学习主要分为分类任务（输出离散类别）和回归任务（输出连续数值）。

2. 无监督学习

无监督学习也被称为无导师学习,它在完全不提供标签的训练数据上,自动学习数据的内在结构、统计规律与分布特征。由于没有人为标注的标准答案作为指导,模型只能依据样本在特征空间中的位置、距离、密度、相关性等信息,自主发现数据中的潜在模式与分组。无监督学习更接近人类探索未知事物的认知过程,适合标签难以获取、标注成本过高或需要先对数据进行探索分析的场景。其典型任务包括聚类与降维,如根据消费行为将用户划分为不同群体、根据画作风格分成不同流派、根据电影主题分为不同类型、根据社交关系发现用户社区、对高维数据进行降维以便可视化等。

3. 半监督学习

半监督学习是介于有监督学习与无监督学习之间的一种学习范式,它在训练时同时使用少量有标签样本与大量无标签样本。在现实任务中,高质量标注数据需由专业人员完成,耗时费力、成本高昂,而无标签数据可轻松大量获取,半监督学习正是为解决这一痛点而生。该方法以少量标签数据确定基本学习方向,再利用大量无标签数据补充数据整体分布信息,从而提升模型的泛化能力与稳定性,避免因标注样本过少导致模型效果差。半监督学习广泛应用于标注难度大的数据场景,如网页分类、文本情感分析、语音识别、图像识别、医学影像分析等领域。

4. 强化学习

强化学习是一类通过智能体与环境的连续互动来学习最优行动策略的机器学习算法。与前三种依赖静态数据集的学习方式不同,强化学习没有预先给定的有标签样本,而是通过"试错"方式进行学习:智能体在某一状态下选择动作,环境根据动作执行结果返回奖励或惩罚,智能体根据反馈不断调整策略,最终实现长期累积回报最大化。强化学习适合无法直接获取标签、只能通过环境反馈进行优化的序贯决策问题,强调动态过程与长期收益。其典型应用包括游戏 AI、机器人路径规划、迷宫导航、自动驾驶决策、智能调度、博弈对抗等场景。

机器学习方法丰富多样,随着大模型时代的到来,机器学习的任务更加多样化,出现了迁移学习、元学习、自监督学习、演示学习、组合泛化等,每种方法都有其独特的特点和适用范围。在实际应用中,可以根据问题的性质、数据是否标记等选择合适的机器学习方法。

【例 3-1】 判断以下任务分别适合采用有监督学习、无监督学习、半监督学习、强化学习中的哪一类,并简要说明理由:

(1) 根据患者的体检指标数据以及已确诊的疾病类型,训练模型对新患者进行疾病类型诊断。

(2) 在大量无标签的商品文本评论中,自动将评论划分成正面、负面、中性等不同类别。

(3) 利用少量有标签语音数据和大量无标签语音数据,训练语音自动转文字模型。

(4) 让机器人在未知仓库环境中不断尝试行走,通过碰壁惩罚、到达目标奖励,学习最优导航路线。

(5) 根据社交平台用户的关注、互动关系,自动发现具有相同兴趣爱好的用户群组。

(6) 根据历史房屋特征与真实成交价格,训练模型预测新建房屋的市场售价。

解:(1) 有监督学习:数据带有明确的疾病类型标签,目标是对新的样本进行分类判断。

（2）无监督学习：评论数据无标签，由模型自主聚类发现情感类别。

（3）半监督学习：使用少量有标签语音数据和大量无标签语音数据训练，符合标注成本高的场景。

（4）强化学习：通过与环境交互获得奖惩反馈，学习连续决策策略。

（5）无监督学习：无标签数据，自主挖掘用户群体结构。

（6）有监督学习：房屋数据带有真实成交价格标签，属于预测连续值的回归任务。

3.1.4　机器学习的基本流程

机器学习并非一次性执行的简单步骤，而是一套标准化、可复用、可迭代的完整工程流程。典型的机器学习流程主要包括以下 6 个步骤。

1. 问题定义与任务建模

机器学习的第一步是明确问题定义与任务建模，即把现实问题转换为计算机可处理的学习任务。这一步骤需要确定任务目标，判断属于分类、回归、聚类还是异常检测，并明确评价指标与预期效果。例如，将“判断用户是否会流失”定义为二分类任务，将“预测下月销售额”定义为回归任务。

2. 数据采集与数据源确定

数据是机器学习的基础，需要根据任务需求收集相关、可靠、具有代表性的数据，并确定数据来源，如业务系统、传感器、日志、数据库、公开数据集、图片、文本、音视频等。采集过程需保证数据覆盖足够多的场景，具备一定的规模与多样性，避免因数据偏斜导致模型效果不佳。

3. 数据预处理

原始数据通常存在缺失、噪声、异常、重复、格式不统一等问题，无法直接用于训练，因此必须进行数据预处理。该步骤包括数据清洗、缺失值填充、异常值剔除、重复值删除、数据归一化与标准化、类别型数据编码等。预处理的目的是把杂乱的原始数据变为规范、干净、可训练的数据，提升模型训练的稳定性与精度。

4. 特征工程

数据规范后，需要进行特征工程，即从原始数据中提取、构造、筛选对任务最有价值的特征。特征是模型学习的依据，好的特征能显著提升模型效果。特征工程包括特征提取、特征组合、特征转换与特征选择，目的是保留关键信息、剔除冗余与无关特征，让模型更高效、更准确地学习数据规律。

5. 模型选择、训练与调参

根据任务类型选择合适的模型，如决策树、支持向量机、神经网络等，并使用训练集数据让模型学习数据模式。在训练过程中，通过不断调整超参数，如学习率、训练轮数、树深度等，优化模型表现，避免欠拟合与过拟合，使模型在复杂度与泛化能力之间达到平衡。

6. 模型评估与性能验证

模型训练完成后，需要在独立的测试集上进行模型评估与性能验证，使用准确率、精确率、召回率、F1 值、均方误差等指标客观评价模型在未知数据上的预测能力。若评估结果不满足要求，则需要回到前面阶段重新优化数据、特征或模型。

常用的评估方法有留出法、交叉验证法和自助法等。留出法直接将数据集按比例划分

为训练集、验证集和测试集,数据互不重叠,简单高效,但结果易受单次划分影响,稳定性一般。交叉验证法将数据均分为 k 组,轮流取一组为验证集,其余为训练集,循环 k 次取平均结果,评估结果稳定可靠,是最常用方法。自助法采用有放回抽样生成训练集,未抽到样本作测试集,适合小数据场景,但会改变数据分布、引入偏差。在实际应用中,中等数据量优先用交叉验证法,大数据可用留出法,极小样本可选用自助法。

【例 3-2】 在模型性能评估阶段需要选择与训练集相互独立的验证集来评估模型效果,请思考:

(1) 训练集和验证集可以交叠吗?

(2) 为什么需要交叉验证?

解:(1) 不可以交叠,如果两者存在交集,模型在训练时已经见过验证集数据,会直接记住样本特征与标签,导致评估结果虚高,无法反映真实泛化能力,严重影响模型选择与参数调优的可靠性。

(2) 使用交叉验证是为了更稳定、客观地评估模型性能。单次随机划分验证集容易受抽样偏差影响,评估结果不稳定。交叉验证将数据分成多组并多次轮换验证,能充分利用样本信息,有效降低单次划分带来的偶然性,得到更可靠、更具代表性的平均性能,从而更准确地选择模型与调整超参数。

综上所述,机器学习的整个流程不是单向线性过程,而是不断迭代、反复优化的闭环工程,通过持续迭代数据、特征、算法与参数,逐步提升模型的泛化能力与稳定性,最终得到可落地、可信赖的智能模型。

3.1.5 机器学习的应用场景

随着数据资源日益丰富、计算能力不断提升以及算法持续创新,机器学习已从实验室研究走向规模化产业应用,成为驱动各行各业数字化、智能化转型的核心技术。凭借从数据中自动学习规律、高效完成预测与决策的独特优势,机器学习的应用场景几乎覆盖社会生产与日常生活的各个领域,从医疗诊断、金融风控到电商推荐、工业质检,再到交通出行、公共服务,机器学习正以低成本、高效率、高适应性的方式,解决大量传统方法难以处理的复杂问题,显著提升行业效率,改善用户体验,创造新的产业价值。

1. 医疗健康领域

机器学习在医疗健康领域的应用,有效提升了诊断精度、治疗效率与药物研发速度。在疾病诊断方面,机器学习可对 CT、X 光、病理切片等医学影像进行自动分析,快速识别病灶、肿瘤与异常区域,辅助医生实现早筛早诊,降低漏诊率。在个性化治疗中,系统可根据患者基因信息、病史、检查指标等数据,推荐适配的治疗方案与用药剂量,实现精准医疗。此外,机器学习还能通过电子病历数据预测疾病风险、监测重症患者恶化趋势,并在药物研发中快速筛选有效分子结构,大幅缩短研发周期、降低成本,为公共卫生事件监测与健康管理提供强大支撑。

2. 金融服务领域

金融服务是机器学习应用最成熟、最广泛的领域之一,主要集中在风险控制、信用评估、智能投顾与反欺诈等方向。在信用评估中,银行与信贷机构通过用户收入、消费、还款、社交行为等多维度数据,构建机器学习模型,精准评估用户还款能力与违约概率,提升审批效率

并降低坏账风险。在反欺诈场景中，模型可实时监测交易行为，自动识别盗刷、虚假交易、异常转账等风险操作，毫秒级做出拦截判断。同时，机器学习还可用于市场趋势分析、股价预测、智能客服、保险理赔自动审核等，提升金融服务的安全性与智能化水平。

3. 电商领域

机器学习已成为电商平台的核心技术支撑，最典型的应用是个性化推荐系统。平台根据用户的浏览历史、购买记录、收藏偏好、停留时长等行为数据，自动学习用户兴趣，精准推送商品、内容与优惠活动，提升转化率与用户体验。此外，机器学习在价格智能优化、销量预测、库存管理、智能客服、评论情感分析、虚假订单识别等环节同样发挥重要作用。通过预测市场需求，平台可动态调整库存与补货策略，降低滞销与缺货损失；通过分析用户评论，自动识别好评、差评与投诉热点，帮助商家快速优化商品与服务。

4. 工业制造领域

在工业制造领域，机器学习推动生产过程向自动化、可视化、预防性方向升级。在质量检测环节，机器学习可对产品外观、尺寸、缺陷进行高速视觉检测，替代人工实现全自动化质检，提升精度与效率。在设备运维方面，通过分析设备运行电流、温度、振动等数据，模型可提前预测故障发生概率，实现预测性维护，避免突发停机造成巨大损失。同时，机器学习还可优化生产排程、能耗管理、供应链调度，提升生产线整体效率，降低生产成本，助力工厂实现数字化与智能化改造。

5. 智能决策与公共服务领域

机器学习广泛应用于交通、城市管理、安防、教育等公共服务领域，提供高效智能决策支持。在智能交通领域，模型可实时分析车流量、路况、事故数据，优化红绿灯控制、预测拥堵路段、规划导航路线。在安防领域，通过视频分析实现异常行为检测、人员追踪、危险区域预警，提升公共安全保障能力。在教育领域，机器学习可根据学生学习行为与测评结果，推送个性化学习内容、预测薄弱知识点，实现因材施教。此外，在气象预测、环境监测、灾害预警、政务服务等领域，机器学习也能从海量数据中快速提取规律，为科学决策提供可靠依据。

综上，机器学习以数据驱动、自动学习、泛化能力强的特点，深度融入各行各业，不仅提升了效率、降低了成本，更催生了新模式、新业态，成为人工智能落地最广泛、价值最突出的核心技术。

3.2　有监督学习

在实际机器学习应用中，许多任务需要根据已知输入预测明确的输出结果，如分类、回归、识别、预测等，且具备可直接用于模型训练的带标签数据。面对这类有明确学习目标、需要精准输出的实际问题，需要一种依靠已知标注信息、由数据驱动并带有明确学习目标的学习方式——有监督学习。它以高质量标记数据集为输入，依赖预先给定的分类标签或连续数值目标，学习输入特征与输出结果之间的映射关系，自动构建出可泛化的预测模型，最终实现对未知新的样本的精准判断、数值预测与决策支持。有监督学习凭借目标清晰、效果可控、易于评估等优势，成为人工智能在金融风控、图像识别、自然语言处理、医疗诊断等真实场景中最常用、最可靠的核心技术手段。

3.2.1 有监督学习概述

在有监督学习框架下,根据标签的取值类型与任务目标的差异,可分为分类问题与回归问题两大类典型任务。这两类问题都基于带标签数据进行建模与预测,但在输出形式、任务目标、评价方式上存在本质区别:回归问题用于预测连续型数值,输出可以在一个区间内取任意实数;分类问题用于预测离散型类别,输出只能是有限个固定类别。

1. 回归问题的基本概念

回归问题是有监督学习中用于预测连续数值的一类任务。在回归问题中,模型的输出是连续取值,理论上可以在一定范围内取无限多个值,输出结果之间存在明确的大小关系与数值差异。回归建模的目标是找到一个最优拟合函数,使模型预测值尽可能接近真实值,从而实现对未知样本的量化预测。以外卖订单配送时长预测为例,平台根据历史配送数据,包括配送距离、天气状况、时段、订单数量、骑手位置、路况等特征,以及每笔订单对应的实际送达时长,训练预测模型,对新订单的配送时间进行预测,就是典型的回归问题。配送时长可以是 25 分钟、25.5 分钟、31.2 分钟等任意数值,属于连续型输出。在日常生活中类似的回归任务十分常见,例如,根据历史客流数据预测商场下一时刻的到访人数、根据设备运行参数预测下一小时的耗电量、根据气象特征预测次日降水量、根据用户信息预测月度消费金额等。这类任务的共同特点是关注数值精度,预测结果越接近真实值,模型效果越好。例如,已知某汽车的历史行驶里程和实际耗油量数据,想要预测该汽车驶往 450 km 外目的地所需的耗油量,就是一个回归问题。汽车耗油量预测示意图如图 3-4 所示。

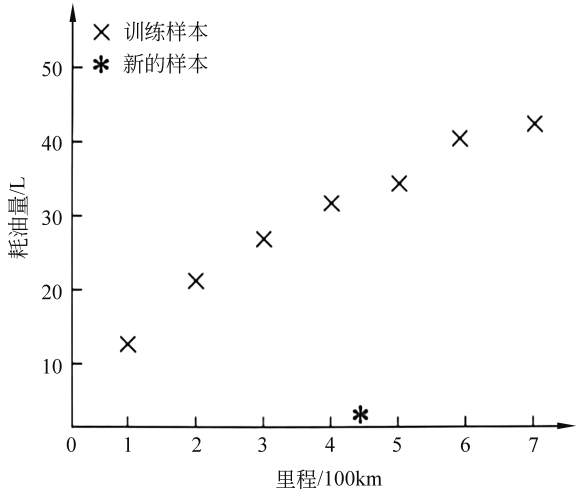


图 3-4 汽车耗油量预测示意图

2. 分类问题的基本概念

分类问题是有监督学习中用于判断样本所属类别的一类任务。在分类问题中,模型的输出是有限个离散类别,类别之间无序、不存在数值大小关系,建模目标是学习分类决策边界,将不同类别的样本区分开,并对新的样本做出正确的类别判断。以短视频内容违规检测为例,平台根据视频的文本标题、画面特征、音频信息、播放行为、用户举报记录等特征,判断视频是否属于违规内容,就是典型的分类问题,模型输出只有“合规”或“违规”两个选项,属

于离散类别，只需要判断正确与否，不涉及数值计算。常见的分类任务还包括：根据交易信息判断银行卡交易是否为欺诈、根据用户行为判断是否为恶意账号、根据学生学习数据判断是否为学业风险学生、根据医学检验指标判断是否为阳性等。当输出类别数为 2 时，称为二分类问题；当类别数大于或等于 3 时，称为多分类问题。例如，已知车辆行驶的历史数据对道路上的车辆进行形式意图预测，假定已知某汽车的速度和转角数据，想要预测汽车是直行还是变道，这就是一个分类问题。车辆行驶意图预测示意图如图 3-5 所示。

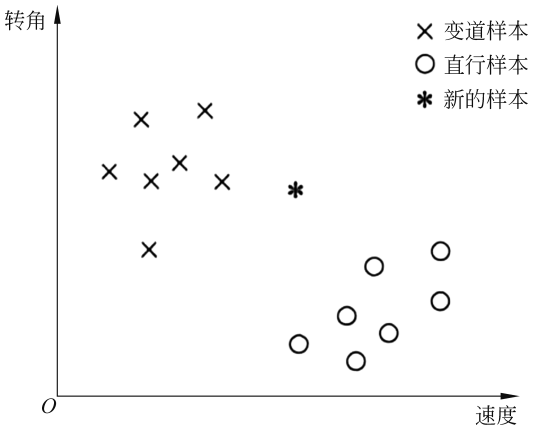


图 3-5　车辆行驶意图预测示意图

3. 回归问题和分类问题的区别与联系

1）回归问题和分类问题的主要区别

（1）输出类型与取值空间不同：回归问题的输出为连续数值，取值无限、有序、可比较大小，如时长、销量、温度、金额等；分类问题的输出为离散类别，取值有限、无序、无量化关系，如合规/违规、正常/不正常、是/否等。这是二者最根本、最直观的区别。

（2）核心建模目标不同：回归问题的目标是拟合数据分布，最小化预测值与真实值之间的误差，追求“预测得更准”；分类问题的目标是划分决策边界，最大化分类正确的样本比例，追求“分类得更对”。

（3）损失函数与评价指标不同：回归问题使用连续型损失函数，如平方损失、绝对损失，常用评价指标包括均方误差（MSE）、均方根误差（RMSE）、平均绝对误差（MAE）等，用于衡量数值偏差；分类问题使用概率型损失函数，如交叉熵损失、合页损失，常用评价指标包括准确率、精确率、召回率、F1 分数等，用于统计分类正确率。

（4）应用场景与问题性质不同：回归问题适用于预测、估算、趋势分析等量化场景，强调数值结果；分类问题适用于判断、识别、甄别、过滤等定性场景，强调类别归属。在工程实践中，任务类型直接决定算法选择、模型结构与评估方式。

2）回归问题和分类问题的内在联系

回归问题和分类问题虽然存在明显区别，但二者同属有监督学习，底层逻辑高度相通，具有紧密的内在联系。

（1）学习流程完全一致：二者都遵循数据采集、预处理、特征工程、模型训练、验证、评估的完整流程，都依赖带标签数据驱动学习，都以提升泛化能力为最终目标。

（2）算法体系高度通用：绝大多数机器学习算法同时支持分类与回归任务，仅需少量

改造即可切换。如决策树、随机森林、支持向量机、神经网络、线性模型等,既可以用于回归问题预测连续值,也可以用于分类问题输出离散标签。典型的如逻辑回归,以线性回归为基础,通过激活函数将连续输出映射为概率,再通过阈值判定转换为分类结果,是连接两类问题的典型模型。

(3) 问题之间可以相互转换:分类模型通常先输出属于某一类别的概率(连续值),再判定类别,本质是“先回归、后分类”;回归问题也可以根据业务需求转换为分类问题,如将预测的连续销售额划分为“低、中、高”三个等级,将数值预测变为等级分类。这种灵活转换在实际项目中广泛使用,体现了两类问题的互补性与统一性。

(4) 底层理论基础同源:分类与回归都基于统计学习、函数拟合与泛化理论,都面临过拟合、欠拟合、数据不平衡、特征选择等共性问题,优化思路与工程策略高度一致。

回归与分类是有监督学习的两大核心问题,共同构成了机器学习应用的主体。回归问题预测连续值,分类问题预测离散类别,二者在输出、目标、指标上存在明确区分,但在学习流程、算法适配、理论基础上高度统一。在实际工程项目中,准确识别问题属于分类还是回归,是开展数据处理、模型选择、参数调优和效果评估的第一步,也是机器学习模型能够落地应用的关键前提。

【例 3-3】 回归与分类问题辨析:某学校计划开发一套校园智能管理系统,需要使用机器学习算法处理以下两个实际问题,请分别判断每个问题属于回归问题还是分类问题,并说明理由。

(1) 根据学生的出勤次数、作业完成率、课堂表现、周测成绩,预测该学生本学期期末考试的具体分数。

(2) 根据学生近一年的成绩波动、缺勤记录、课堂参与度,判断该学生是否属于学业预警对象。

解:(1)回归问题。预测目标是期末考试的具体分数,属于连续型数值,因此是回归问题。(2)分类问题。判断结果为“是/否”学业预警对象,输出为离散类别,因此是分类问题。

3.2.2 分类问题

分类问题是有监督学习中应用最广泛、研究最深入的核心任务。分类问题的目标是:基于一组带有类别标签的训练数据,学习一个从样本特征到离散类别标签的映射模型,使该模型能够对未知类别的新的样本做出准确的类别判断。在医疗诊断、金融风控、文字识别、图像分类、故障检测、舆情分析等众多实际场景中,分类问题都占据主导地位。针对分类问题,机器学习领域提出了多种经典求解方法,它们在数学原理、模型结构、适用数据、计算效率与可解释性上存在显著差异。最具代表性的分类算法有线性判别分析、支持向量机、k 近邻法、决策树。

1. 线性判别分析

线性判别分析(Linear Discriminant Analysis,LDA)是有监督学习中经典的线性分类方法,同时兼具线性降维功能,由统计学家 Fisher 于 1936 年提出,因此也常被称为 Fisher 线性判别分析。线性判别分析的核心思想非常直观:将数据投影到一条直线(或低维空间)上,让同类样本尽可能紧凑、不同类样本尽可能分开,通过最小化类内离散度、最大化类间离

散度，实现最优分类。之所以要在最大化投影点类间距离的同时最小化投影点类内距离，是为了使两类样本的投影点各自集中，有利于样本的正确分类，且避免不同类样本投影点出现数据混叠，导致样本分类错误。在分类问题中，线性判别分析以模型简单、计算高效、可解释性强著称，是理解线性分类的基础算法，广泛应用于文本分类、人脸识别、医学诊断、特征降维等场景。线性判别分析的示意图如图 3-6 所示，其中，“×”表示变道样本，“○”表示直行样本，将高维空间的点向低维空间投射，可以使投影后不同类样本间的区别更加明显。当对新的样本进行分类时，同样将其投射到这条直线上，投射后距离哪一类样本的投射中心更近，则新的样本就属于哪个类别。根据以上原则分析，则新的样本属于“直行”类别。

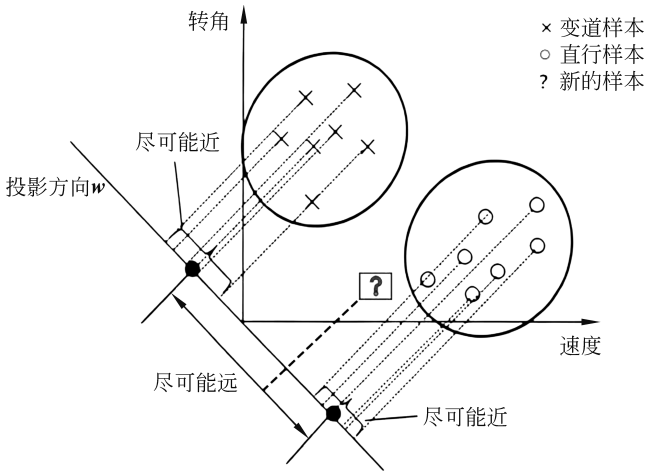


图 3-6　线性判别分析的示意图

从数学原理来看，线性判别分析围绕以下两个关键矩阵展开。

（1）类内散度矩阵 $\boldsymbol{S}_w$：衡量同类样本内部的离散程度，值越小表示同类越集中。

（2）类间散度矩阵 $\boldsymbol{S}_b$：衡量不同类别中心之间的距离，值越大表示类别区分越明显。

设投影直线 L 的方向向量为 $\boldsymbol{w}$，样本为 x，L 的方程为 $z=\boldsymbol{w}^{\mathrm{T}}x$，则样本 x 在 L 上的投影为 $\boldsymbol{w}^{\mathrm{T}}x$。第 i 类样本的协方差矩阵为 $\boldsymbol{S}_i$，投影点的协方差为 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_i\boldsymbol{w}$，线性判别分析的优化目标是最大化类间散度矩阵 $\boldsymbol{S}_b$ 与类内散度矩阵 $\boldsymbol{S}_w$ 的比值：

$$J(\boldsymbol{w})=\frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{w}} \tag{3-1}$$

其中，使 $J(\boldsymbol{w})$ 最大的方向向量 $\boldsymbol{w}$ 就是最优投影方向。

$\boldsymbol{w}$ 还有另外一种求解方法，为了求得目标函数 $J(\boldsymbol{w})$ 的极值，令 $J(\boldsymbol{w})$ 对 $\boldsymbol{w}$ 的导数等于 0，从而可以得出：

$$\boldsymbol{S}_w^{-1}\boldsymbol{S}_b\boldsymbol{w}-J(\boldsymbol{w})\boldsymbol{w}=0 \tag{3-2}$$

令 $\lambda=J(\boldsymbol{w})$，得到

$$\boldsymbol{S}_b\boldsymbol{w}=\lambda\boldsymbol{S}_w\boldsymbol{w} \tag{3-3}$$

因此，求最大化 $J(\boldsymbol{w})$ 对应的 $\boldsymbol{w}$ 等价于求解广义特征值 λ 的问题，最大特征值对应的特征向量 $\boldsymbol{w}$ 即为最优投影方向。

线性判别分析的实现过程清晰规范，可分为以下 5 个标准步骤。

（1）数据标准化与预处理：对输入特征进行归一化或标准化，消除量纲差异，保证不同

特征在距离计算中权重公平。

(2) 计算每个类别的均值向量：分别计算每个类别在各个维度上的均值，用于描述类别中心位置。

(3) 计算类内散度矩阵与类间散度矩阵矩阵：类内散度矩阵矩阵衡量同类样本的分散程度，类间散度矩阵衡量不同类别中心之间的距离，两者是线性判别分析的核心数学基础。

(4) 求解最优投影方向：通过最大化类间散度矩阵与类内散度矩阵的比值求解投影向量。在二分类中得到一个一维方向；在多分类中可以得到多个投影方向，构成低维空间。

(5) 投影与分类决策：将训练集与测试集样本投影到低维空间，使用距离判别、阈值判别等方法完成分类。通常将测试样本投影后，判断其离哪个类别中心更近，即可确定类别。

投影后，样本在新空间中同类靠近、异类远离，只需用最近邻类别中心即可完成分类。由于投影是线性变换，线性判别分析属于线性分类模型，只能处理线性可分问题，但计算高效、可解释性强。

线性判别分析作为经典的线性分类与降维方法，具有多项突出优势。首先，线性判别分析模型简单、计算高效，仅依赖均值、方差与矩阵运算即可完成训练与预测，运行速度远快于支持向量机、神经网络等复杂模型，适合实时性要求较高的场景，如文本分类、医疗指标快速筛查等。其次，线性判别分析可解释性强，投影方向与特征权重清晰直观，能够明确展示分类依据，满足金融风控、医疗诊断等对可解释性有严格要求的领域。同时，在高维数据，如图像、文本特征中，线性判别分析可同时完成降维和分类，能将高维空间压缩到低维空间，大幅降低计算量并提升分类效果，尤其适合样本数量不多但特征维度很高的场景，如人脸识别、疾病预测等。而且，得益于线性判别分析基于全局统计信息构建的基础，其对一般噪声具有较强的鲁棒性，模型稳定性好，不容易出现过拟合，在小样本数据集上依然能保持可靠的性能。

此外，线性判别分析也存在一定局限性。它仅适用于线性可分或近似线性可分的数据，面对非线性可分数据时效果有限。该方法对数据分布有较强假设，要求数据近似服从高斯分布且各类别协方差矩阵相同，不符合条件时性能会明显下降。同时，它对异常值较为敏感，极端点会影响投影方向，降低分类精度，在类别重叠度高、复杂多分类任务中，线性划分能力也难以满足需求。因此，在线性场景下线性判别分析高效实用，在非线性复杂任务中，则通常需要结合更灵活的模型以达到更佳效果。

2. 支持向量机

支持向量机(Support Vector Machine，SVM)是一类以最大化分类间隔为核心的有监督学习分类模型，其基本思想是在特征空间中寻找一个最优分类超平面，使得不同类别的样本被正确分开，且距离超平面最近的样本点与平面之间的间隔达到最大。所谓的支持向量，就是指在样本集中距离分类超平面最近且直接决定分类间隔大小的少量关键训练样本，是决定决策边界位置的关键数据。凭借坚实的数学理论基础、优秀的泛化能力以及处理高维数据的突出优势，支持向量机既可以解决线性可分问题，也能借助核技巧高效处理非线性可分问题，在小样本、高维、非线性场景下表现尤为稳定，是机器学习中应用最广泛、鲁棒性最强的经典算法之一。

在理解支持向量机之前，首先需要明确线性可分的概念。如果在样本的特征空间中，存在一个超平面能够将不同类别的样本完全无误地划分到两侧，则称这批数据是线性可分的。

在二维空间中，这个超平面表现为一条直线；在三维空间中表现为一个平面；在更高维空间中则被称为超平面。然而在实际任务中，大量数据并非天然线性可分，或存在噪声干扰，此时简单的线性划分不再适用。同时，在样本量有限、特征维度很高、模型泛化要求严格的场景下，传统方法容易出现过拟合或分类边界不稳定的问题。正是在这些情况下，支持向量机凭借最大间隔原则、软间隔容错能力以及核函数映射能力，成为最理想的选择。以车辆行驶意图预测问题为例，分类线 Z_1 和 Z_2 都能将两类样本分开，但是对于一个新的样本，基于两条分类线的预测结果是不一样的，基于分类线 Z_1，新的样本属于“变道”；基于分类线 Z_2，新的样本属于“直行”，那么选择哪条分类线更好呢？两条分类线示意图如图 3-7 所示。

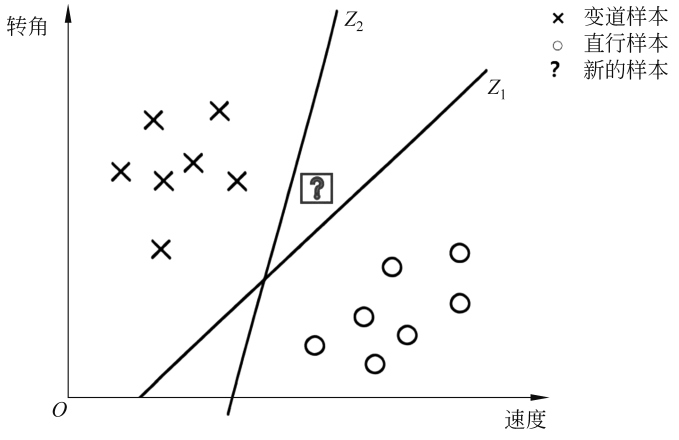


图 3-7　两条分类线示意图

面对这种多种分类线的问题，支持向量机的解决方式是寻找一个最优的决策边界，即能够把两类数据分开的最大间隔超平面。当前这个问题 $\rho_1 > \rho_2$，说明 Z_1 相比 Z_2 在不同类别样本间的分类间隔更大，意味着分类线或超平面对于噪声的抗干扰能力越强，泛化能力越好，所以选择距离分类边界两侧样本间隔最大的分类线或超平面对训练集数据的噪声或局限性有最大的“容忍”能力。因此，选择 Z_1 作为分类线更好，新的样本属于“变道”类型。不同分类线对应的分类间隔示意图如图 3-8 所示。

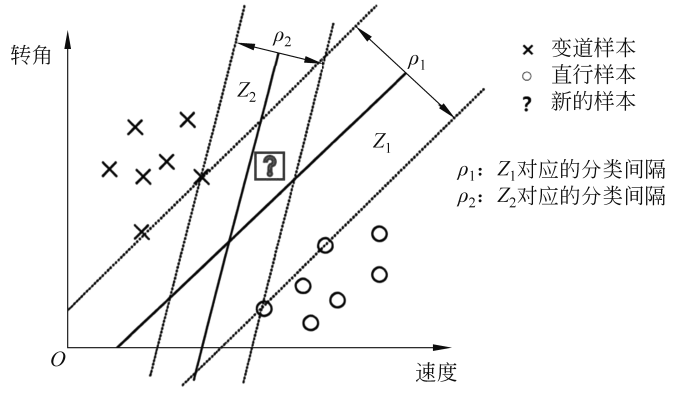


图 3-8　不同分类线对应的分类间隔示意图

寻找最大间隔超平面是一个有约束条件的优化问题，通过优化算法可以得到问题的解，

按照所使用的分类模型是否为线性,支持向量机分为线性支持向量机和非线性支持向量机。

1) 线性支持向量机

线性支持向量机是针对线性可分或近似线性可分数据设计的基础分类模型,其核心目标是在特征空间中找到一个具有最大分类间隔的线性超平面,从而实现最优且泛化能力最强的分类效果。线性支持向量机假设数据可以被一条直线(二维)或超平面(高维)准确区分,并在此基础上追求间隔最大化,因此具有稳定性高、泛化能力强的特点。

在线性可分情况下,$\boldsymbol{w}^{\mathrm{T}}$ 为法向量,b 为偏置项,最优分类超平面可表示为

$$\boldsymbol{w}^{\mathrm{T}}x+b=0 \tag{3-4}$$

线性支持向量机不只是寻找能够划分样本的超平面,还是选取让两类最近样本到平面距离最大的最优超平面。间隔超平面穿过支持向量,并与最优分类超平面相互平行,分列其两侧,是界定、计算分类间隔的基准平面。据此,正、负间隔超平面表达式分别为

$$\boldsymbol{w}^{\mathrm{T}}x+b=1 \tag{3-5}$$

$$\boldsymbol{w}^{\mathrm{T}}x+b=-1 \tag{3-6}$$

这两个超平面之间的距离即为分类间隔,其宽度由支持向量唯一决定,是支持向量机追求最大化的核心目标。只有距离最近、决定间隔宽度的样本点也就是支持向量落在间隔超平面上,其余样本均位于间隔区域之外不参与决策边界的构建,这也是支持向量机高效稳定的重要原因。支持向量机分类间隔示意图如图 3-9 所示。图中展示了一个用($\boldsymbol{w}$,b)表示的具有最大分类间隔的标准超平面,同时给出了间隔超平面。

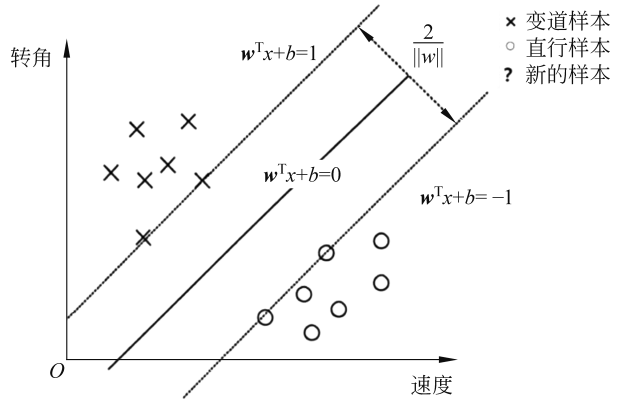
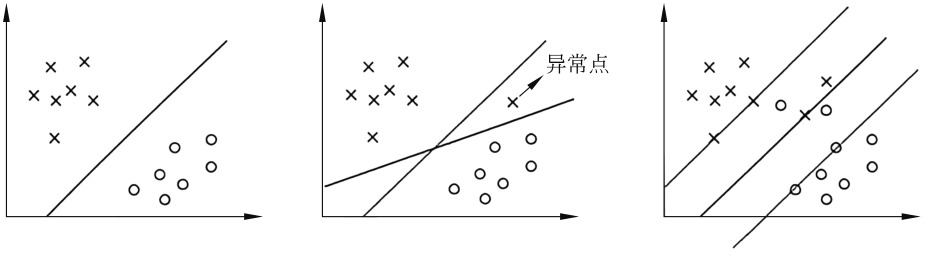


图 3-9 支持向量机分类间隔示意图

以上给出的线性支持向量机是硬间隔分类器,它是线性可分支持向量机的基础形式,要求必须存在一个超平面能将所有训练样本完全正确分类,不允许任何样本出现在间隔内部或被错分,并在此前提下最大化分类间隔。这种严格约束使得硬间隔分类器的决策边界完全由支持向量决定,数学性质优良、解唯一、泛化能力较强,但同时存在明显局限:一是仅适用于完全线性可分数据,对现实中普遍存在的噪声、异常值和轻微重叠数据无法适应;二是对异常点极度敏感,少量异常样本就会大幅改变间隔与超平面位置,导致模型过拟合,在新数据上表现显著下降;三是鲁棒性差,无法处理实际任务中无法严格线性可分的场景,应用范围十分有限。

在实际问题中,为克服硬间隔分类器的缺陷,支持向量机引入软间隔策略。软间隔不再

追求完全无错分，而是允许部分样本违反分类约束，即允许少量样本落在间隔内部或被错误划分，同时通过惩罚因子控制错分数量与间隔大小，在分类精度和泛化能力之间取得平衡。其实现方式是在原优化目标中加入松弛变量与惩罚参数，松弛变量衡量样本违反约束的程度，惩罚参数用于控制对错分样本的惩罚力度。惩罚参数越大，对错误分类的惩罚越重，间隔越窄，模型越接近硬间隔；惩罚参数越小，允许的错误分类越多，间隔越宽，模型越平稳。软间隔使得线性支持向量机具备更强的实用性与鲁棒性，能够有效处理近似线性可分数据，抑制噪声与异常点影响，避免过拟合，成为实际工程中最常用的线性支持向量机形式。支持向量机的硬间隔与软间隔分类示例图如图 3-10 所示。

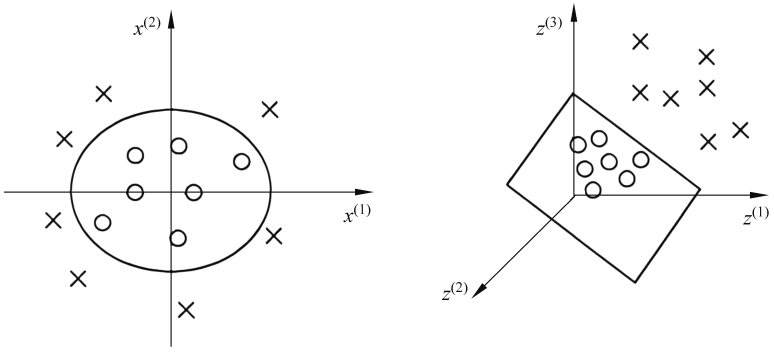


(a) 硬间隔分类　　(b) 一个异常点的硬间隔分类　　(c) 多个异常点的软间隔分类

图 3-10　支持向量机的硬间隔与软间隔分类示例图

2）非线性支持向量机

在实际工程与科研场景中，绝大多数分类任务并非简单的线性可分问题。受数据采集噪声、特征关系复杂、类别边界不规则等因素影响，样本在原始特征空间中往往呈现高度重叠、缠绕、非线性可分的特点，无法用直线、平面或超平面直接划分。此时，线性分类器包括线性支持向量机、逻辑回归、线性判别分析等难以取得理想效果，必须引入能够处理非线性关系的模型，即基于广义分类面的非线性分类方法。非线性支持向量机正是为解决此类问题而提出，它突破了线性模型的限制，能够学习复杂、弯曲、高维的决策边界，从而实现对非线性数据的精准分类。线性不可分问题原始数据与转换后数据分布图如图 3-11 所示。



(a) 椭圆形分类线数据分布图　　(b) 转换后的数据分布图

图 3-11　线性不可分问题原始数据与转换后数据分布图

非线性支持向量机的核心思想是将原始空间中线性不可分的数据，通过非线性映射投影到高维空间，使其线性可分，再利用线性支持向量机求解最优分类超平面。低维空间的复杂非线性问题在高维空间中往往可转换为线性问题。但直接显式映射会导致维度爆炸、计

算量剧增,因此非线性支持向量机采用核方法解决这一问题。核函数决定模型对复杂边界的拟合能力,是影响效果的关键。常用核函数主要有以下 4 种。

(1) 线性核:不进行非线性变换,等价于普通线性分类,计算快、稳定性高,适用于近似线性可分数据,如文本分类、高维稀疏特征分类。

(2) 多项式核:通过特征多项式组合构建非线性边界,可捕捉曲线趋势与高阶关联,调整阶数可控制拟合程度,但阶数过高易过拟合,适合有明显规律的数据与简单非线性分类。

(3) 高斯核:应用最广、通用性最强,可拟合任意复杂非线性可分,对不规则、重叠度高的数据效果优异,参数少、调参直观,是图像识别、故障检测、医疗诊断等场景的首选。

(4) Sigmoid 核:源于神经网络激活函数,可模拟多层神经元分类效果,但稳定性较弱,适用场景有限,多用于特定文本与序列数据。

总体而言,线性核速度最快,多项式核处理曲线问题,高斯核通用性强,Sigmoid 核接近神经网络方式。实际应用中可根据数据维度、线性可分性、样本量与非线性程度选择核函数,实现高效稳定分类。借助该方法,非线性支持向量机精度高、泛化能力强,能适应复杂数据分布,是处理图像、语音、文本等高维非线性数据的经典算法。

本节主要讲述分类方法,但是实际上支持向量机不仅支持分类,也适用于回归问题。支持向量机处理回归问题时也是从大量样本中筛选出对模型训练最有价值的部分样本作为支持向量,区别是回归器中标签变量是连续值。

支持向量机是基于最大间隔超平面与核技巧的经典有监督学习模型,整体性能稳健、泛化能力突出。该模型依靠间隔最大化提升泛化水平,决策边界仅由少数支持向量决定,结合核函数可同时处理线性与非线性分类问题。支持向量机的核心优势在于对小样本、高维数据具有良好适应性,即便特征维度高于样本数量也能稳定学习。最大间隔原则使其不易过拟合,对噪声和异常点具备一定鲁棒性,借助核函数可灵活应对各类非线性可分数据。其主要不足体现在大规模数据集上训练复杂度高、运算速度较慢,模型性能对核函数与惩罚参数较为敏感,需要细致调参。同时,模型可解释性较弱,原生方法更适用于二分类任务,处理多分类任务需额外扩展策略。因此,支持向量机适用于样本量适中、特征维度较高、分类边界复杂的场景,广泛用于文本分类、基因分析、医疗诊断、图像识别及工业故障检测等领域,尤其在标注成本高、样本有限但对精度要求严格的任务中表现优异。

3. k 近邻法

k 近邻法是有监督学习中最直观、最简单的分类与回归方法之一,属于基于实例的惰性学习算法。它的核心思想是:一个样本的类别,由与它距离最近的 k 个邻居样本“投票决定”,即在特征空间中,如果一个样本周围的大多数邻居都属于某一类别,那么该样本也属于这一类别。k 近邻法不需要显式训练模型,仅在预测时计算距离并决策,因此具有实现简单、无须假设数据分布、可处理非线性问题等特点,既适用于分类任务,也可用于连续值的回归预测,是机器学习入门与快速原型验证的基础算法。

k 近邻法的算法步骤简洁直观,核心围绕“距离计算—近邻选择—决策判断”展开,无须复杂的模型训练过程,具体可细化为以下 4 步。

(1) 数据预处理:由于 k 近邻法基于距离度量实现决策,不同特征的量纲差异会严重干扰距离计算结果,因此需先对训练集和测试样本进行标准化或归一化处理。预处理后使特征服从均值为 0、方差为 1 的正态分布,确保每个特征在距离计算中拥有公平权重,避免

因某一特征数值过大而主导决策。

(2) 距离计算：针对预处理后的测试样本，逐一计算其与训练集中所有样本的距离，距离越小代表两个样本的相似度越高。根据数据类型选择合适的距离度量方式，同时对所有距离结果进行排序，便于后续近邻选取。

(3) 近邻选取：根据预设的近邻数 k，从排序后的距离结果中选取距离最小的 k 个训练样本作为测试样本的“邻居”。这 k 个样本是决定测试样本类别的核心依据，选取过程需严格遵循距离由小到大的顺序，不遗漏、不重复。

(4) 决策判断：采用多数投票法，统计 k 个近邻样本中出现次数最多的类别，将其作为测试样本的预测类别。

在以上算法中，训练集是 k 近邻法的基础，其质量直接决定分类效果。训练集应尽可能覆盖各类别、各分布区域，样本均衡、噪声少、代表性强。若样本分布不均衡，会导致多数类压制少数类，使分类结果偏向数量更多的类别。由于 k 近邻法基于距离计算，不同特征的量纲与数值范围会严重影响距离大小，因此必须进行预处理。常用方法包括“最小-最大归一化”和“Z-Score 标准化”，使所有特征处于相近范围，保证各维度对距离的贡献均衡。

距离度量决定样本之间相似度的计算方式，是 k 近邻法的核心。连续数值数据常用距离包括以下三种。

(1) 欧氏距离：最常用、最直观的距离，适用于连续、数值型特征，默认用于大多数场景。

(2) 曼哈顿距离：对异常值更稳健，计算简单，适合噪声较多的数据。

(3) 切比雪夫距离：关注最大差异维度，适用于某些特殊工程数据。

此外，还有汉明距离、余弦相似度属于其他拓展距离度量，分别适配类别型数据和高维稀疏数据。在实际使用中，欧氏距离足以满足绝大多数分类任务。

k 值是 k 近邻法中唯一关键超参数，直接决定模型性能。k 值过小，模型会过度关注局部噪声，容易过拟合；k 值过大，模型过于平滑，会忽略局部结构，导致欠拟合。选择 k 值的基本原则包括以下几种。

(1) 通常取奇数，避免投票平局。如选择偶数，当某个样本的 k 个近邻中两类近邻点数量相同，则需要根据其他信息来判断，如随机选择或者选择最近邻点所属类。

(2) 从小值开始尝试，如 3、5、7、9 等。

(3) 使用交叉验证选择最优 k，以验证集准确率最高为标准。

(4) 样本量大时可适当增大 k；样本量小时 k 宜较小。

其中，k 值的选择体现了模型在“局部”与“全局”之间的权衡，也是 k 近邻法调参的核心。

k 近邻法原理示意图如图 3-12 所示。假设有一个二分类任务，训练样本分布如图 3-12(a) 所示，对于一个新输入的样本需考查其 k 个近邻样本，并依据这些近邻的类别来推断该新的样本的所属类别。通过分析可以发现，当 k 取值为 5 时，新的样本的预测结果与 k 取 1 或 3 时存在差异，因此合理选择 k 值成为关键问题。在实际应用中，k 值的确定往往需要经过多次尝试。当 k 值较小时，算法会基于较小邻域内的训练样本进行预测，此时分类模型复杂度较高，训练误差相对较小，但容易出现过拟合现象；当 k 值较大时，算法会利用更大邻域内的训练样本来决策，分类模型随之变得简单，然而距离新的样本较远的训练样本也会参