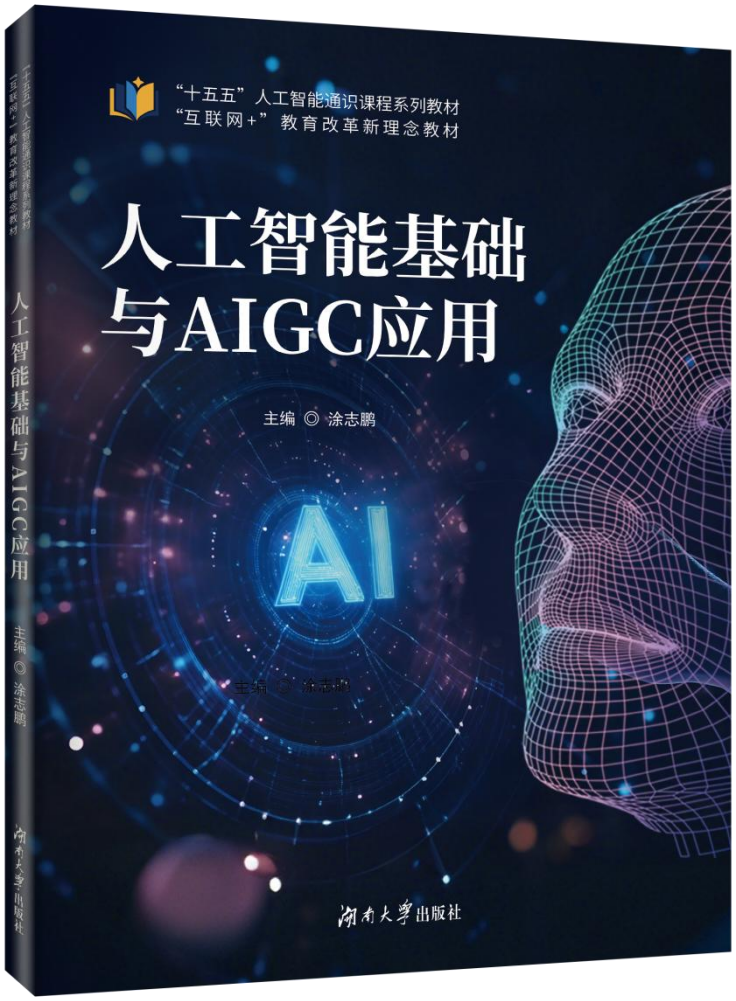
人工智能基础与 AIGC 应用



类目：人工智能类
书名：人工智能基础与 AIGC 应用
主编：涂志鹏
出版社：湖南大学出版社
开本：大 16 开
书号：978-7-5667-4902-4
使用层次：通用
出版时间：2026 年 9 月
定价：59.80 元
印刷方式：四色
是否有资源：有

人工智能基础与AIGC应用

责任编辑：周繁燕
封面设计：顺语书装

ISBN 978-7-5667-4007-6
9 787566 740024 >
定价：59.80元

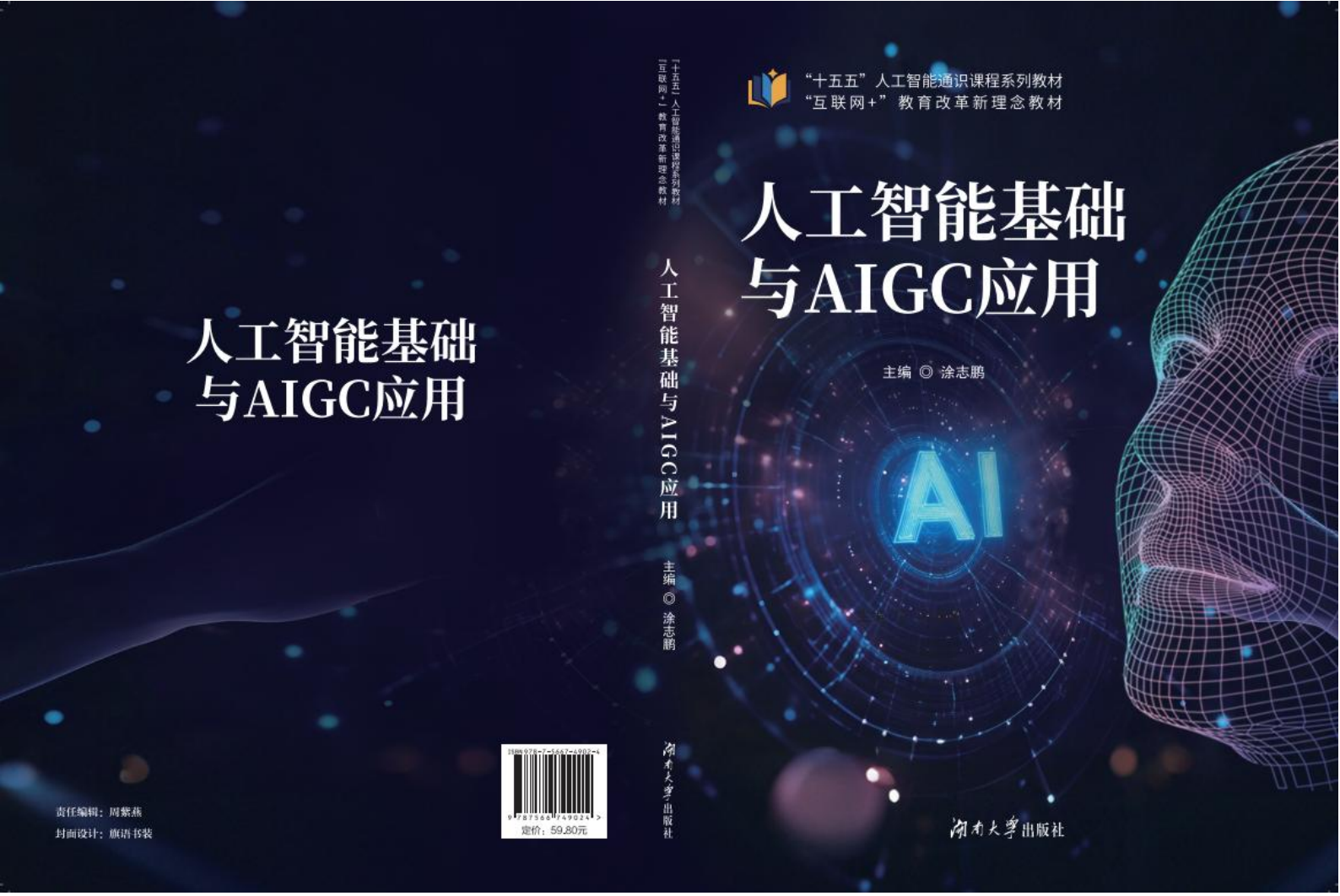
【十五五】人工智能通识课程系列教材
【互联网+】教育改革新理念教材
人工智能基础与AIGC应用
主编 ◎ 涂志鹏

湖南大学出版社

“十五五”人工智能通识课程系列教材
“互联网+”教育改革新理念教材

人工智能基础与AIGC应用

主编 ◎ 涂志鹏



湖南大学出版社

“十五五”人工智能通识课程系列教材
“互联网 +”教育改革新理念教材

人工智能基础与 AIGC 应用

主　编 ◎ 涂志鹏
副主编 ◎ 蒋庆文　蓝金丽　张丹烘



湖南大学出版社
·长 沙·

图书在版编目（CIP）数据

人工智能基础与 AIGC 应用 / 涂志鹏主编 . -- 长沙：湖南大学出版社，2026. 9. -- ISBN 978-7-5667-4902-4

Ⅰ. TP18

中国国家版本馆 CIP 数据核字第 2026AN5290 号



人工智能基础与 AIGC 应用

RENGONG ZHINENG JICHU YU AIGC YINGYONG

主　　编：涂志鹏
责任编辑：周紫燕
印　　装：河北昌联印刷有限公司
开　　本：889 mm×1194 mm　1/16　　**印　　张：**14　**字　　数：**376 千字
版　　次：2026 年 9 月第 1 版　　**印　　次：**2026 年 9 月第 1 次印刷
书　　号：ISBN 978-7-5667-4902-4
定　　价：59.80 元

出 版 人：李文邦
出版发行：湖南大学出版社
社　　址：湖南·长沙·岳麓山　　**邮　　编：**410082
电　　话：0731-88822559（营销部）　88821174（编辑室）　88821006（出版部）
传　　真：0731-88822264（总编室）
网　　址：http://press.hnu.edu.cn
电子邮箱：xiaoshulianwenhua@163.com

版权所有，盗版必究
图书凡有印装差错，请与营销部联系



前 言

PREFACE

教育部于 2024 年启动了教育系统人工智能大模型应用示范行动，旨在打造人工智能通识课程体系，赋能理工医农文等各类人才培养。“人工智能 +”时代来临，高等院校亟需搭建系统完备的人工智能通识教学体系，面向社会输送大批具备良好人工智能素养的复合型技术技能人才。

党的二十大报告提出要实施科教兴国战略，强化现代化建设人才支撑。强调要深化教育领域综合改革，加强教材建设和管理。为了响应党中央的号召，深入贯彻全国教育大会和《教育强国建设规划纲要（2024—2035 年）》精神，更好构建适应社会需求的人工智能通识课程教育体系，我们在充分进行调研和论证的基础上，精心编写了这本《人工智能基础与 AIGC 应用》教材。

以大模型为代表的生成式人工智能快速普及，ChatGPT、DeepSeek 等产品的出现，让人工智能技术走进大众视野，展现出强大的变革力量。人工智能正在深刻改变人们的思维方式、生活模式与生产形态，对经济社会发展产生广泛而深远的影响。为帮助学习者系统掌握人工智能基础理论，熟练开展 AI 工具实操应用，树立正确的人工智能伦理观念，我们结合教学实际，完成本教材的编撰工作。

与同类书相比，本书具备以下特点：

- **循序渐进，由浅入深**

本书属于人工智能入门类教材，遵循认知规律由浅入深组织教学内容。从人工智能基础概念出发，依次讲解机器学习、神经网络、AI 大模型、提示工程、多模态 AIGC 应用、AI 智能体以及前沿技术等内容，帮助学习者搭建完整的人工智能知识框架，为后续进一步深入学习打下扎实基础。

● 理实并重，突出实操

教材兼顾基础理论讲解与实践能力培养，弱化复杂晦涩的公式推导，侧重技术原理理解与工具落地使用。全书以项目为载体，每个项目配套案例实训与思考练习，引导学生学以致用，提升运用人工智能工具解决真实问题的实践能力。

● 课程思政，润物无声

将课程思政有机融入专业教学全过程，在讲授人工智能技术知识的同时，结合 AI 伦理安全、技术规范、国产 AIGC 工具实践、科技发展责任等内容，引导学生树立科技向善的理念，培养正确的科技价值观、社会责任感与职业素养。

● 项目引领，学做合一

全书采用项目式教学的编写模式，将知识要点拆解融入 8 个学习项目之中。每个项目将理论学习、工具实操和综合案例相结合，贴合课堂教学改革需求，便于开展课堂教学、实训作业与自主学习，着力提升学生实践应用能力与自主探究能力。

由于编者能力有限，书中难免存在不妥之处，恳请读者朋友提出宝贵意见，不胜感激。

编　者

2026 年 6 月



目 录
CONTENTS

项目 1　人工智能入门

1.1　认识人工智能 / 2
1.1.1　人工智能的定义 / 2
1.1.2　人工智能的分类 / 4
1.1.3　人工智能的能力分级 / 6
1.1.4　人工智能的关键特征 / 7
1.1.5　人工智能发展简史 / 9
1.2　人工智能的技术架构 / 11
1.2.1　基础层 / 12
1.2.2　技术层 / 14
1.2.3　应用层 / 15
1.3　人工智能伦理与安全 / 16
1.3.1　关注 AI 伦理的意义 / 16
1.3.2　使用 AI 的基本准则 / 17
1.3.3　AI 安全风险与应对策略 / 18

项目 2　机器学习与神经网络

2.1　机器学习 / 22
2.1.1　机器学习的基础 / 22
2.1.2　监督学习 / 24
2.1.3　无监督学习 / 28
2.1.4　强化学习 / 32
2.2　神经网络 / 32
2.2.1　神经元 / 33
2.2.2　人工神经网络 / 33
2.3　深度学习典型网络 / 34
2.3.1　卷积神经网络 / 34
2.3.2　循环神经网络 / 37
2.3.3　生成对抗网络 / 39

项目 3　AI 大模型与提示工程实践

3.1　认识 AI 大模型 / 43
3.1.1　大模型概述 / 43
3.1.2　大模型技术演进背景 / 45
3.1.3　大模型分类 / 46
3.1.4　主流大模型介绍 / 48
3.2　大模型关键技术 / 53
3.2.1　Transformer 架构与自注意力机制 / 53
3.2.2　预训练与微调策略 / 54
3.3　提示工程入门 / 56
3.3.1　提示词的定义与重要性 / 56
3.3.2　基础提示词构成要素与原则 / 57
3.4　大模型安全与规范使用 / 58
3.4.1　数据安全与风险 / 58
3.4.2　提示词注入攻击原理与防范 / 59
3.4.3　API 安全漏洞和应对策略 / 60
3.4.4　算法偏见的形成与预防 / 61

项目 4　自然语言处理与 AIGC 应用

4.1　自然语言处理基础 / 65
4.1.1　语言学基础知识 / 65
4.1.2　文本的数字化表示 / 66
4.1.3　自然语言处理主要应用领域 / 67
4.2　AI 文本生成技术 / 70
4.2.1　文本生成原理 / 70
4.2.2　文本提示词设计策略 / 71
4.2.3　主流 AIGC 文本工具 / 72
4.3　文本生成工具：Deepseek / 73
4.3.1　原创文案 / 75
4.3.2　文案的改写和润色 / 76
4.4　文本生成工具：Kimi / 79
4.4.1　文案创作 / 80
4.4.2　Kimi+ 助手 / 82
4.5　文本生成工具：豆包 / 92
4.5.1　帮我写作 / 93
4.5.2　AI 阅读 / 96
4.5.3　编辑文本 / 98

项目 5　视觉智能与 AIGC 应用

5.1　计算机视觉基础 / 104
5.1.1　视觉感知原理 / 104
5.1.2　图像的数字化表示 / 105
5.1.3　计算机视觉应用领域 / 106
5.2　AI 图像生成技术 / 108
5.2.1　AI 图像生成技术概览 / 108
5.2.2　图像生成提示词策略 / 109
5.2.3　主流 AIGC 图像工具 / 111
5.3　图像生成工具：通义万相 / 112
5.3.1　基本功能 / 113
5.3.2　智能扩写 / 113
5.3.3　创意模板 / 114
5.4　图像生成工具：文心一言 / 116
5.4.1　文本生成图像 / 117
5.4.2　图像生成图像 / 119
5.4.3　智慧绘图 / 119

5.5 图像生成工具：爱设计 / 125
5.5.1 字生画 / 127
5.5.2 图生画 / 128

项目 6 视频智能与 AIGC 应用

6.1 AI 视频处理与应用概述 / 132
6.1.1 AI 视频处理的核心能力 / 132
6.1.2 AI 视频处理的典型应用场景 / 133
6.1.3 AI 视频处理技术的优势与局限 / 134
6.2 AI 视频生成技术 / 134
6.2.1 视频生成原理与方式 / 134
6.2.2 视频生成提示词 / 136
6.2.3 主流 AIGC 视频工具 / 137
6.3 视频生成工具：可灵 / 138
6.3.1 文本生成视频 / 139
6.3.2 图像生成视频 / 142
6.3.3 多模态编辑 / 144
6.3.4 音效设置 / 147
6.3.5 对口型 / 148
6.4 视频生成工具：即梦 / 152
6.4.1 视频生成 / 152
6.4.2 数字人 / 154
6.4.3 动作模仿 / 155
6.4.4 视频编辑 / 157
6.5 视频生成工具：闪剪 / 160
6.5.1 直播快剪 / 161
6.5.2 口播视频 / 161
6.5.3 定制数字人 / 163
6.5.4 链接成片 / 165

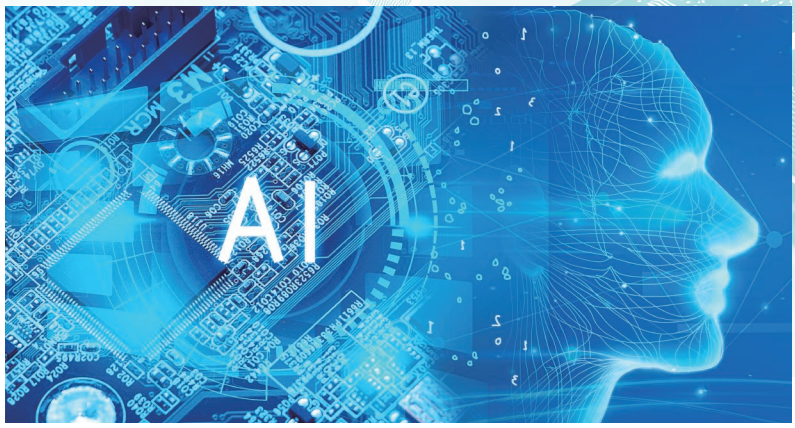
项目 7 AI 智能体应用

7.1 AI 智能体基础 / 173
7.1.1 AI 智能体的定义与特征 / 173
7.1.2 AI 智能体的构建与工作原理 / 174
7.1.3 AI 智能体的主要应用场景 / 175
7.1.4 AI 智能体的技术挑战与展望 / 176
7.2 AI 智能体：Cherry Studio / 176
7.2.1 核心功能 / 177
7.2.2 应用特点 / 177
7.2.3 安装和配置模型 / 177
7.2.4 创建知识库 / 181
7.2.5 搭建智能体 / 183
7.3 AI 智能体：扣子 / 187
7.3.1 创建知识库 / 187
7.3.2 创建智能体 / 191

项目 8　人工智能前沿与行业赋能

8.1　机器人与具身智能 / 201
8.1.1　机器人技术基础 / 201
8.1.2　具身智能 / 203
8.1.3　具身智能的应用 / 204
8.2　类脑计算与混合智能 / 206
8.2.1　类脑计算原理 / 206
8.2.2　混合智能与脑机接口（BCI）简介 / 207
8.3　AI 未来趋势与伦理责任 / 208
8.3.1　通用人工智能（AGI）展望 / 208
8.3.2　AI 技术发展趋势 / 209
8.3.3　负责任 AI 的原理与框架 / 209
8.3.4　智能时代的学习与适应 / 210

参考文献



项目 1
人工智能入门

项目导读

人工智能（artificial intelligence，AI）已深度融入我们的生产生活，从日常的智能音箱、推荐算法，到复杂的自动驾驶、医疗影像诊断，其影响无处不在。本项目将带你走进人工智能的世界，从认识人工智能的定义、分类、能力分级和关键特征入手，回顾其发展历程，了解其技术架构，同时探讨人工智能伦理与安全问题。通过学习，你将对人工智能有全面且系统的认识，为后续深入学习相关技术和应用打下坚实基础。

学习目标

理解人工智能的四种主流定义，掌握人工智能的分类方式。

熟悉人工智能的能力层级，了解人工智能的关键特征。

知晓人工智能发展的关键阶段，从图灵测试、学科诞生到不同时期的发展浪潮。

掌握人工智能技术架构的构成，包括基础层、技术层和应用层的具体内容。

认识人工智能伦理与安全的重要性，了解使用 AI 的基本准则以及常见的安全风险与应对策略。

技能目标

能够结合具体案例，判断人工智能系统属于哪种定义视角和分类类型。

能识别不同人工智能应用所体现的能力层级，并分析其在各层级的具体表现。

能够简要描述人工智能技术架构中各层的关键组成，并举例说明其在实际应用中的作用。

面对具体的人工智能应用场景，能初步分析其中可能涉及的伦理问题和安全风险，并提出相应的应对思路。

思政目标

树立正确的技术伦理观，认知人工智能技术的发展必须符合人类的价值观和道德准则，增强对技术伦理问题的敏感性和责任感。

培养安全意识和风险防范意识，明白人工智能在带来便利的同时也存在安全隐患，自觉遵守相关法规和准则，推动技术的安全、健康发展。

激发对人工智能技术的探索热情和创新精神，同时了解技术发展的复杂性和系统性，培养严谨、务实的科学态度。

增强社会责任感，理解人工智能技术对社会各领域的深远影响，树立利用技术服务社会、造福人类的信念。

1.1 认识人工智能

1.1.1 人工智能的定义

人工智能作为一项深刻影响当代社会的技术，其概念与应用已广泛渗透到我们生产生活的各个层面。然而，当人们探究“人工智能究竟是什么”这一根本性问题时，却发现难以得出一个公认的、统一的答案。不同背景的个体基于其经验、知识结构与关注焦点的差异，往往会形成多元化甚至显著不同的认知。这种现象恰如中国古代寓言“盲人摸象”所描述的情景：每位观察者都只接触到 AI 这个复杂系统的某一特定方面——有的关注其模仿人类对话的能力，有的惊叹其超越人类的计算水平，还有的则聚焦于其自主决策与学习的机制——并据此构建对 AI 的片面理解。因此，为了全面、准确地把握人工智能的本质，我们有必要对其主流定义进行系统性的梳理与辨析。

人工智能的定义

为了系统地研究人工智能，通常可将其定义划分为两个维度、四种视角。第一个维度是“思维过程与推理”与“行为与表现”的对比；第二个维度是“以人类为基准”与“以理性为基准”的对比。“理性”在这里指基于已有信息，做出最大化实现某个目标的“正确”决策。将这两个维度组合，便得到了四种定义人工智能的主流视角。

1.1.1.1　像人一样思考

这一视角的核心目标是让机器模拟人类的思维过程，其关键问题是：“计算机程序如何模拟人脑的思考方式？”要实现这一目标，研究者需要通过心理学实验、脑科学扫描等多种方法，深入探究人类大脑处理信息、形成决策的内部机制，并将这些机制转化为计算机可以执行的模型。认知科学是这一领域的代表性学科。例如，通过研究人类解数学题时的思维步骤，设计出能够模拟该过程的计算机程序，使其在解题时展现出与人类相似的思考路径。

1.1.1.2　像人一样行动

与“像人一样思考”不同，这一视角不关注机器内部的思考过程是否与人类一致，只注重其外部行为表现是否能与人类无法区分。著名的“图灵测试”是该视角的经典判据。如图 1-1 所示，在测试中，一名人类裁判通过文本界面同时与一台机器和一个真实的人类进行对话。如果在足够长的时间后，裁判无法可靠地分辨出哪一方是机器，那么就可以认为这台机器通过了测试，具备了与人相当的智能行为。例如，我们日常使用的智能客服能够像人类客服一样回答用户的咨询，聊天机器人能与用户进行自然的日常对话，这些都是在这一路径上的具体应用。

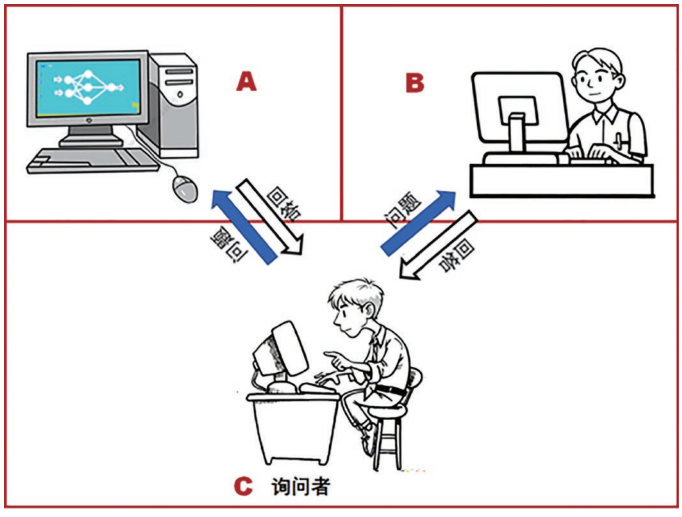


图 1–1　图灵测试示意图

1.1.1.3　理性地思考

该视角致力于让机器遵循严谨的数理逻辑进行推理，追求一种理想化的、无懈可击的“正确”思考。古希腊哲学家亚里士多德等人提出的三段论等逻辑法则，是其思想源头。例如：

◆ 前提 1：凡是人都会死。

◆ 前提 2：苏格拉底是人。

◆ 结论：因此，苏格拉底会死。

基于这一视角的 AI 系统，被期望能像数学家一样，从一系列公理出发，通过逻辑推演，证明新的定理。早期的许多专家系统，如用于医疗诊断的 MYCIN 系统，就采用了这种基于逻辑规则

的推理方法，根据输入的症状和相关规则，推导出可能的疾病。

1.1.1.4 理性地行动

这是目前人工智能领域最主流、最被广泛接受的视角。其将智能定义为理性主体的行为。所谓“主体”，是指任何能够感知其环境（通过传感器）并对其环境采取行动（通过执行器）的事物。所谓“理性”，是指该主体总是在其感知和知识的基础上，采取能使其性能度量（即目标完成度）最大化的行动。这个视角有两个显著优点：

（1）更具通用性：它不局限于模仿人类，因为在许多任务（如下棋、运算）中，超越人类表现的“超人”方法是存在的，也是我们所追求的。例如，AlphaGo 在围棋对弈中，采用了许多人类从未使用过的棋路，却能战胜人类顶尖棋手。

（2）更易于科学验证：逻辑推理虽然理想，但现实世界充满不确定性，难以用纯粹的逻辑表达。而“理性行动”可以通过数学方法进行清晰的定义和优化，使其成为一门可度量、可工程实现的科学。

1.1.2 人工智能的分类

对人工智能系统进行分类，可以依据不同的标准。当前，业界与学界最常采用的分类维度有两个，一是根据 AI 的应用范围与能力强度进行划分，二是从其底层的技术学派与实现路径进行划分。

1.1.2.1 按应用范围与能力强度划分

这是最广为人知的一种分类方式，它根据 AI 系统所能处理任务的广度及其智能水平，将其划分为三个递进的类别，如图 1-2 所示。

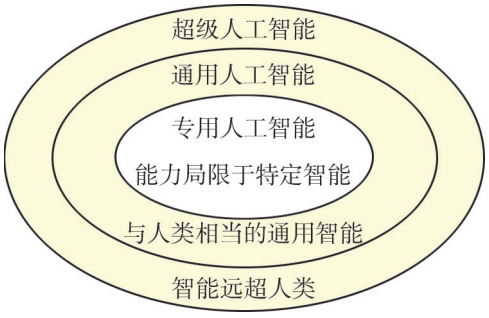


图 1-2 人工智能能力强度分类

（1）专用人工智能（artificial narrow intelligence，ANI）。

专用人工智能，也常被称为“弱人工智能”，是目前我们能够实现并广泛应用的 AI 形态。其核心特征是“专能”，即被设计用于解决某一个特定领域或单一类型的任务。虽然它在特定任务上可以达到甚至远超人类的水平，但其能力不具备通用性。例如，在围棋领域战胜人类顶尖棋手的 AlphaGo，只能专注于围棋对弈，无法去执行一项机器翻译任务；同样，用于人脸识别的安防系统，能够精准识别出人脸信息，但无法用来撰写一篇新闻稿。大家日常接触到的几乎所有 AI 产品，如智能音箱（主要处理语音交互和信息查询）、推荐算法（为用户推荐感兴趣的商品或内容）、自动驾驶系统（专注于驾驶相关的决策和操作）等，均属于专用人工智能的范畴。

（2）通用人工智能（artificial general intelligence，AGI）。

通用人工智能，或称“强人工智能”，指具备与人类同等智慧，能够理解、学习并应用其智

能来解决任何问题的 AI 系统。一个 AGI 系统将拥有跨领域的思考、推理、学习和规划能力，就像人类一样可以胜任从驾驶、烹饪到进行科学研究、艺术创作等一系列广泛的活动。如科幻电影《流浪地球》中的“MOSS”，能够统筹管理地球的各项事务，做出关乎人类命运的决策；《钢铁侠》中的“贾维斯”，既能帮助主人处理日常事务，也能参与到高科技装备的研发中，这些都是对通用人工智能的形象化构想。当前，尽管大型语言模型（large language model，LLM）展现出了一定的通用能力雏形，能够处理文本生成、问答、翻译等多种任务，但距离真正意义上的 AGI 仍有巨大差距。

（3）超级人工智能（artificial superintelligence，ASI）。

超级人工智能是一个假想中的阶段，它指的是在几乎所有领域都远远超越最聪明人类的智慧的 AI。这种智能不仅在计算速度上，更在创造力、洞察力和社会交往等质化层面全面超越人类。ASI 的出现可能会对人类文明产生颠覆性的影响，例如在科学研究上，它可能在短时间内攻克人类难以解决的难题；在社会结构上，可能会改变人类的生产生活方式，因此它也是科技哲学与伦理学领域持续探讨的重要议题。

1.1.2.2　按技术学派与实现路径划分

这一分类方式更侧重于 AI 实现的底层技术哲学与方法论，主要分为三大主流学派。

（1）符号主义。

符号主义是 AI 领域的开创性学派，其核心思想是“智能源于物理符号系统的操作”。它认为，人类的认知过程本质上是信息的处理，而信息可以用符号来表示，认知过程就是对这些符号进行逻辑运算和推理。因此，构建 AI 的关键在于建立一个完善的知识库（由符号构成）和一套强大的推理引擎。早期的专家系统，如用于地质勘探的 PROSPECTOR 系统，通过将地质知识以符号和规则的形式存储起来，再利用推理引擎进行分析和判断，从而得出勘探建议；逻辑编程（如 Prolog 语言）也是符号主义的典型代表，通过定义事实和规则，进行逻辑推理来解决问题。

（2）连接主义。

连接主义的灵感来源于对人脑神经网络结构的研究。它认为智能并非源于中央处理器的逻辑推理，而是从大量简单的、相互连接的单元（人工神经元）的集体活动中涌现出来的。该学派不强调预先编写规则，而是通过“训练”（从大量数据中学习）来让网络自动调整连接权重，从而获得某种能力。在当前所处的深度学习浪潮中，其核心技术如卷积神经网络（在图像识别领域表现出色，能识别图像中的物体、场景等）、循环神经网络（适用于处理序列数据，如文本、语音）以及 Transformer 模型（在自然语言处理领域取得重大突破，如 GPT 系列模型）等，都是连接主义思想的直接成果，也是当前 AI 发展的主导力量。

（3）行为主义。

行为主义不关注智能的内部实现是符号逻辑还是神经网络，它主张智能体现在与环境的交互行为之中。该学派认为，智能的核心是在“感知—行动”的循环中，通过不断试错和获取反馈（奖励或惩罚）来学习最优的行为策略。强化学习是行为主义的典型技术范式，在机器人控制（如机器人通过不断尝试学会行走、抓取物体）、自主决策（如游戏 AI 通过多次游戏试错，学习到最优的游戏策略）、游戏 AI（如 AlphaGo 通过与自己对弈，不断优化棋路）等领域取得了巨大成功。

这三大技术学派并非完全独立，现代复杂的 AI 系统常常融合了多个学派的思想。例如，一个高级的 AI 智能体可能会用连接主义模型来感知世界（处理图像、语音等数据），用符号主义方

法进行长远规划（基于逻辑规则制订整体方案），并通过行为主义的强化学习来优化其具体行动（根据环境反馈调整每一步操作）。

1.1.3 人工智能的能力分级

能力层级描述了 AI 系统从信息输入到最终输出的内部功能架构，它构成了一个完整任务处理的逻辑链条。人工智能的能力层级可以被视为一个信息处理的“能力栈”，它自下而上依次为创造层、决策层、认知层与感知层，如图 1-3 示。这四个层级协同工作，共同构成了智能系统的核心能力。

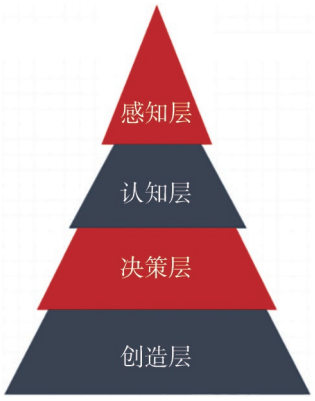


图 1–3 人工智能的能力层级关系图

1.1.3.1 感知层

感知层是智能系统的输入端，负责从外部物理世界或数字世界中获取原始数据，并将其转换为机器可处理的结构化信息。它模拟了生物体的感官功能。其核心功能包括：

（1）计算机视觉。处理和理解图像、视频信息，实现目标检测（如在监控视频中识别出行人、车辆）、图像分类（如区分猫和狗的图片）、场景理解（如判断图片中的场景是街道还是公园）等。

（2）自然语言处理。理解和分析人类语言文本，实现文本分类（如将邮件分为垃圾邮件和正常邮件）、情感分析（如判断一段评论是正面还是负面）、实体识别（如从文本中识别出人名、地名）等。

（3）语音识别。将人类的语音信号转换为文本或命令，如手机的语音助手能够将用户说的“打电话给张三”转换为相应的操作指令。

（4）多源数据感知。处理来自各类传感器（如激光雷达、温度传感器）的数据，如自动驾驶汽车通过激光雷达感知周围环境的距离信息，通过温度传感器了解车内温度。

1.1.3.2 认知层

认知层是智能系统的“信息处理中枢”。它对来自感知层的结构化信息进行分析、理解、推理和学习，形成对世界的内在表示和知识。其核心功能包括：

（1）知识表示与推理：使用知识图谱、逻辑规则等形式来组织知识，并在此基础上进行演绎或归纳推理。例如，知识图谱可以表示“北京是中国的首都”“中国是一个国家”等知识，通过推理可以得出“北京是一个国家的首都”这一结论。

（2）学习与记忆：通过机器学习算法（尤其是深度学习）从数据中发现模式，并将其固化为

模型参数，形成长期记忆。例如，推荐系统通过学习用户的历史购买数据，发现用户的购物偏好，并记住这些偏好，以便后续为用户推荐合适的商品。

（3）抽象与概括：从具体实例中提炼出通用概念或规律。例如，从多个“苹果”“香蕉”“橘子”的实例中，抽象出“水果”这一通用概念。

认知层连接了感知与决策，是实现“智能”的核心环节。一个系统的“智慧”程度主要体现在其认知层的深度和广度上。

1.1.3.3　决策层

决策层基于认知层的分析结果，根据预设的目标或价值函数，进行规划并选择最优的行动方案。它体现了智能系统的目标导向性。其核心功能包括：

（1）规划：为达成目标制定一系列有序的步骤或行动序列。例如，导航系统根据用户的起点和终点，规划出一条包含“直行”“左转”“右转”等步骤的路线。

（2）策略选择：在不确定的环境中，根据当前状态选择最有可能获得最大回报的行动。强化学习是实现该功能的核心技术，如游戏 AI 在面对不同的游戏局面时，会选择能够增加获胜概率的操作。

（3）预测与评估：对不同行动方案可能产生的后果进行预测和价值评估。例如，在投资决策中，AI 系统会预测不同投资组合的收益和风险，并评估其价值，为投资者提供参考。

决策层是智能从“思考”到“行动意图”的转换器。它直接输出控制指令，指导物理实体或数字程序的行为。例如，在自动驾驶系统中，决策层负责输出“加速”“刹车”或“向左转向 30 度”等具体指令。

1.1.3.4　创造层

创造层是 AI 能力的高级表现形式。它不仅限于分析和决策，而是能够依据学习到的模式和知识，生成全新的、具有特定风格或功能的、在数据集中未曾出现过的内容。其核心功能包括：

（1）文本生成：撰写报告、编写代码、创作文学作品。例如，AI 可以根据用户的需求生成一份市场分析报告，或者根据简单的指令编写出一段计算机代码，甚至创作出一首诗歌、一篇小说。

（2）图像与视频生成：依据文本描述或其他输入，创造出新的视觉内容。如用户输入“一片开满鲜花的草地，天空中有彩虹”，AI 可以生成符合该描述的图像；也可以根据文本脚本生成一段动画视频。

（3）音频生成：作曲、生成语音。AI 可以创作出不同风格的音乐作品，也可以模仿特定人的声音生成语音，如将文本转换为某个人的说话声。

创造层代表了 AI 从信息处理者到内容生产者的角色转变。它由生成式人工智能技术驱动，极大地拓展了 AI 的应用边界，尤其是在内容创作、艺术设计和科学发现等领域。

1.1.4　人工智能的关键特征

关键技术特征并非对基础能力的简单罗列，而是对塑造当前人工智能发展方向的核心技术范式与趋势的总结。这些特征相互关联，共同推动着 AI 能力层级的整体跃升。下面介绍几个核心技术特征。

1.1.4.1　多模态融合

传统的人工智能系统通常是单模态的，即只能处理单一类型的数据，如仅处理文本或仅处理

图像。而现代 AI 的一个显著特征是其多模态融合能力，即系统能够同时接收、理解、关联和生成多种不同类型的数据（模态），如文本、图像、音频、视频和传感器信号等。这使得 AI 能够像人类一样，通过多种感官通道来感知和理解世界。

其技术内涵在于将不同模态的信息映射到一个统一的语义空间中，实现信息的对齐与关联。例如，OpenAI 的 GPT-4o 和 Google 的 Gemini 模型，都能在一次对话中无缝地理解用户上传的图片（如用户上传一张汽车的图片）、输入的文本（如“介绍一下这辆车”）和实时的语音（如用户用语音补充“它的油耗怎么样”），并生成相应的多模态回复（如用文字和语音结合的方式回答）。该特征是实现更鲁棒、更通用的人工智能的关键一步，在跨媒体内容检索（如根据一段文字描述检索相关的图片和视频）、人机交互（如用户可以同时用语音和手势与 AI 交互）、自动驾驶环境感知（同时处理摄像头的图像、激光雷达的数据等）等领域具有重要应用价值。

1.1.4.2 自主智能体

在大型语言模型出现之前，AI 更多地扮演着被动工具的角色，执行用户给出的具体指令。而自主智能体标志着一种新的 AI 应用范式的诞生。它是一个能够围绕一个高级目标，自主进行任务分解、规划、调用工具，并进行自我反思和修正的 AI 系统。

其核心是利用大型语言模型作为“大脑”或“中央控制器”。当接收到一个复杂任务时（如“帮我规划一次为期三天的北京旅游并预订机票和酒店”），智能体可以自主地将其分解为“查询航班”“比较酒店”“搜索景点”“规划行程”等子任务，并决定调用外部接口（如航班预订网站）或执行代码来完成这些子任务。这个过程是动态和循环的，智能体会根据工具返回的结果不断调整后续计划。例如，如果查询到的航班时间不合适，会重新搜索其他航班。这是 AI 从“工具”向“助手”甚至“代理人”演进的关键体现。例如，一些智能办公助手能够根据用户“整理本周会议纪要并安排下周重点工作”的指令，自主调取日历，查看会议记录、提取关键信息形成纪要，再结合项目进度规划下周任务，全程无需用户过多干预。

1.1.4.3 生成式突破

这是本轮 AI 浪潮中最具颠覆性的技术特征，其核心是以 Transformer 等大型模型架构为基础，实现了高质量、高多样性内容的生成，即所谓的人工智能生成内容（artificial intelligence generated content，AIGC）。这一突破使得 AI 不再仅仅是信息的分析者和决策者，更成了内容的创造者。

AIGC 技术范式摆脱了传统依赖判别式模型进行分类或预测的思路，转而通过学习海量数据的内在分布规律，来生成全新的、符合该规律的数据。典型的应用包括：ChatGPT 等对话系统能够生成流畅、连贯且富有逻辑的文本，无论是撰写邮件、创作故事还是解答专业问题，都能应对自如；Midjourney、可灵等文生图模型能够根据用户输入的文字描述，创造出具有惊人艺术表现力的图像，风格涵盖写实、抽象、动漫等多种类型；还有 AI 视频生成工具，能根据文本脚本生成动态的视频内容。这一突破正在深刻地改变软件开发（如 AI 辅助编写代码）、媒体（如自动生成新闻稿件）、娱乐（如游戏场景和角色设计）等众多行业的内容生产方式。

1.1.4.4 高效轻量化

随着模型规模和能力的急剧增长，其高昂的训练成本和推理延迟成为制约 AI 技术广泛应用的瓶颈。因此，高效轻量化成了学术界和工业界共同关注的关键技术趋势。它旨在保证或小幅牺牲模型性能的前提下，显著降低模型的参数量、计算量和存储需求。实现这一目标的技术路径主要

包括：

（1）模型压缩：通过剪枝（去除神经网络中冗余的连接或神经元）、量化（将高精度的模型参数转换为低精度，如从 32 位浮点数转为 8 位整数）等手段减少模型的冗余，使其更小巧高效。

（2）知识蒸馏：用一个大型的“教师模型”（性能强但复杂）去指导一个轻量的“学生模型”（简单但高效）进行学习，让“学生模型”继承“教师模型”的知识和能力，同时保持较小的规模。

（3）高效模型架构设计：设计本身计算效率就更高的网络结构，如 MobileNet 系列模型专门为移动设备设计，通过特殊的卷积方式减少计算量。

高效轻量化的技术趋势的最终目标是推动强大的 AI 能力从云端走向终端设备（如手机、智能汽车、物联网设备），实现低延迟、低成本、高隐私保护的本地化部署。例如，手机上的 AI 美颜算法、离线语音识别功能，都是轻量化模型的应用，它们无须依赖云端计算，就能快速响应用户需求，同时避免了数据上传带来的隐私泄露风险。

1.1.4.5　具身智能

具身智能的目标是让 AI 拥有与物理世界进行交互的“身体”，它强调将 AI 模型（大脑）与物理实体如机器人、无人机等（身体）相结合，使 AI 能够在真实环境中通过感知、行动、与环境交互来进行学习和完成任务。

这需要 AI 不仅能处理文本和图像，还要能理解三维空间、物理规律和因果关系。例如，一个具身智能的家务机器人，需要能够“看到”桌子上的杯子（感知），理解“把它拿到厨房”的指令（认知），规划出避开障碍物的路径（决策），并精确地控制机械臂完成抓取和放置的动作（行动），在这个过程中，它还能通过多次尝试学习如何更稳地抓取不同形状的杯子。具身智能是人工智能与机器人学的深度融合，被认为是通往通用人工智能（AGI）的一条重要路径，在智能制造（如工厂中的协作机器人）、物流（如自主搬运机器人）、家庭服务（如陪伴机器人）等领域有巨大的应用潜力。

1.1.5　人工智能发展简史

人工智能的探索之旅跨越了半个多世纪，其发展并非一条平坦的直线，而是呈现出螺旋式上升的态势，其间交织着理论的突破、范式的演进以及对模仿乃至超越人类智能的持续追求。理解其发展历程中的关键节点与思想变迁，是把握人工智能技术本质及其在各领域应用潜力的基础。本节将人工智能的发展历程划分为以下几个关键阶段，如图 1-4 所示。

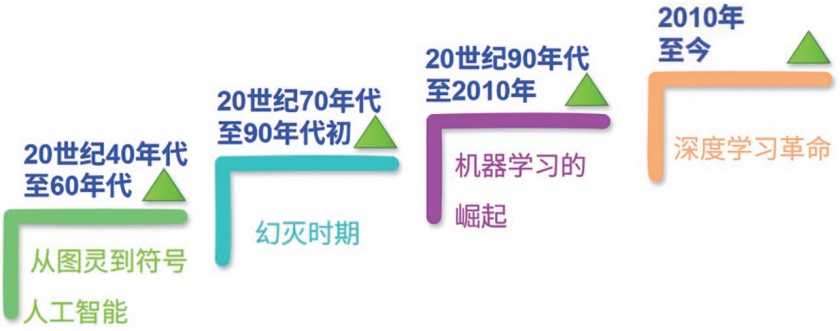


图 1–4　人工智能发展历程

1.1.5.1 从图灵到符号人工智能（20 世纪 40 年代至 60 年代）

人工智能的理论萌芽可以追溯到 20 世纪中叶计算机科学的诞生。

（1）理论基础。

英国数学家艾伦·图灵（Alan Turing）为此奠定了坚实的理论基石。他提出的“图灵机”是一种抽象计算模型，理论上能够模拟任何可计算的过程，为计算机的诞生提供了重要的理论支撑。更为深远的是，图灵于 1950 年发表论文《计算机器与智能》，提出了著名的“图灵测试”，旨在为判断机器是否具备与人类无法区分的智能行为提供一个可操作的标准。尽管历经争议与修正，图灵测试至今仍是人工智能领域一个重要的哲学参照。

（2）学科诞生。

1956 年，由约翰·麦卡锡（John McCarthy）、马文·李·明斯基（Marvin Lee Minsky）等人组织的“达特茅斯夏季人工智能研究计划”被公认为人工智能作为一门独立学科正式诞生的标志。在这次会议上，“人工智能”一词被首次提出，并确立了其核心研究目标：用机器模拟人类的学习、推理、感知和语言等智能行为。

（3）主流范式。

这一时期的主流研究范式是符号主义（symbolism），也被称为“基于规则的人工智能”。其核心思想是：智能源于对符号的操作和逻辑推理。研究者试图通过构建符号表示系统和形式化的规则库来复现人类的认知过程。在这一思想指导下，赫伯特·西蒙（Herbert Simon）和艾伦·纽厄尔（Allen Newell）开发了人工智能程序“逻辑理论家”（Logic Theorist），它能够证明数学定理，是第一个能够模拟人类推理过程的程序；随后他们又开发了“通用问题解决器”（General Problem Solver，GPS），旨在解决各种不同类型的逻辑问题。这些早期程序在证明数学定理、解决逻辑谜题和下棋等任务上取得了初步的成功。

1.1.5.2 幻灭时期（20 世纪 70 年代至 90 年代初）

早期符号主义的成功激发了研究者们极大的乐观情绪，但这很快被现实的挑战所取代，导致了两次被称为“AI 寒冬”的低谷期。

（1）第一次 AI 寒冬（20 世纪 70 年代中期至 80 年代初）。

符号主义的局限性逐渐暴露。首先，基于规则的系统难以处理现实世界中的不确定性、模糊性和不完整信息，例如在自然语言理解中，一词多义、语境依赖等问题难以用固定规则描述。其次，“组合爆炸”问题使得其难以扩展到大规模、复杂的现实场景，当规则数量过多时，系统的推理效率会急剧下降。最后，早期系统对常识知识的匮乏使其显得脆弱而缺乏变通。由于技术瓶颈迟迟无法突破，政府和机构的研究资助大幅削减，人工智能研究进入低谷。

（2）专家系统的短暂繁荣（20 世纪 80 年代）。

在第一次寒冬之后，专家系统的商业化成功带来了短暂的复苏。专家系统通过构建特定领域的知识库和推理引擎，模拟人类专家进行决策（如医疗诊断、地质勘探等）。例如，MYCIN 系统能够帮助医生诊断细菌感染疾病，并推荐合适的抗生素。然而，其知识获取过程烦琐昂贵（需要专家逐条整理知识），且系统维护困难（规则之间易冲突，难以更新），适用范围狭窄（只能解决特定领域问题），最终未能实现持续发展。

（3）第二次 AI 寒冬（20 世纪 80 年代末至 90 年代初）。

专家系统泡沫的破裂，以及专为 AI 编程设计的 Lisp 机器市场未能普及（价格昂贵且兼容性差），共同导致了第二次研究低潮期。尽管处于低谷，但神经网络、遗传算法等非符号主义方法的研究仍在部分学者中延续，为后续的复兴埋下了伏笔。

1.1.5.3　机器学习的崛起（20 世纪 90 年代至 2010 年）

从 20 世纪 90 年代开始，人工智能的主流范式从符号主义逐渐转向机器学习。其核心思想不再是为机器编写显式的规则，而是让机器从数据中自动学习模式和规律。以统计学为基础的机器学习方法成为研究热点。例如，支持向量机（在分类任务中表现优异）、决策树（直观易懂，可用于数据分类和预测）、贝叶斯网络（擅长处理不确定性推理）等。这些算法在文本分类、垃圾邮件过滤、信用评分等商业应用中取得了显著成效。

这一转变的背后有两大关键驱动力。一是互联网的普及带来了前所未有的海量数据，为数据驱动的模型训练提供了充足的“燃料”，如搜索引擎积累的大量网页数据、电商平台的交易数据等。二是计算能力的持续提升（遵循摩尔定律），使得处理大规模数据和复杂模型成为可能，普通计算机也能运行较为复杂的机器学习算法。

1.1.5.4　深度学习革命（2010 年至今）

进入 21 世纪的第二个十年，深度学习的出现，标志着人工智能进入了一个全新的高速发展阶段。深度学习是机器学习的一个分支，它利用包含多个处理层的深度神经网络（deep neural networks，DNN）来学习数据从低级到高级的层次化特征表示。

2012 年，AlexNet 在 ImageNet 图像识别大赛中以巨大优势夺冠，其错误率远低于传统方法，证明了深度学习在图像领域的巨大潜力；2016 年，AlphaGo 战胜世界围棋冠军李世石，展现了强化学习在复杂决策任务中的强大能力，让世界看到了 AI 在战略思维方面的突破；2018 年，Transformer 架构的提出革新了自然语言处理领域，催生了 GPT 系列、BERT 等预训练大模型，使得机器在语言理解和生成方面的能力大幅提升。当前，生成式 AI（如 Stable Diffusion、Midjourney）、多模态大模型（如 GPT-4V、Gemini）不断突破，推动 AI 向通用智能迈进。

深度学习技术已成为当今主流 AI 应用的核心引擎，从智能手机的语音助手、人脸解锁，到自动驾驶的环境感知（识别行人、车辆、交通信号灯）和医疗影像的辅助诊断（检测肿瘤、病变），其影响已渗透到社会生产和生活的方方面面。

1.2 人工智能的技术架构

人工智能的技术架构通常由基础层、技术层和应用层构成，如图 1-5 所示。每一层都有其特定的功能和任务，共同构成了人工智能系统的整体架构。

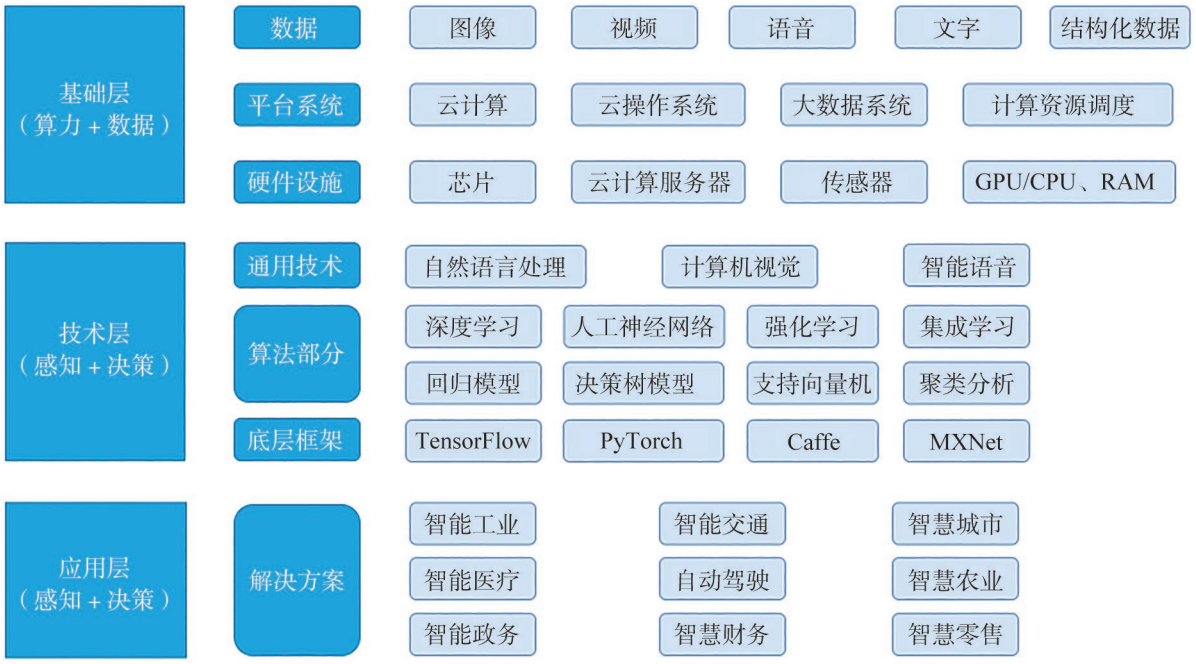


图 1–5　人工智能的技术构架

1.2.1　基础层

基础层是人工智能技术架构的最底层，它为 AI 的学习、决策和执行提供了坚实的支持。这个层次主要涉及数据的收集与处理、平台系统的支持和硬件设施的提供。基础层的作用是为技术层和应用层提供必要的资源和工具，确保 AI 系统能够高效、稳定地运行。

1.2.1.1　数据

数据是人工智能的核心资源之一，AI 系统的学习和决策过程通常依赖于大量的高质量数据。数据可以分为以下几种类型：

（1）图像数据：用于计算机视觉任务，如物体识别、人脸识别、图像分类等。图像数据通常以像素的形式呈现，每个像素包含颜色信息，通过图像处理和深度学习算法可以分析图像中的内容。例如，安防摄像头拍摄的图像可用于识别可疑人员。

（2）视频数据：用于视频分析、行为识别、视频监控等任务。视频数据是一系列连续的图像帧，通过对视频数据的帧间变化进行分析，AI 能够理解动态场景中的信息，如识别视频中行人的行走轨迹、车辆的行驶状态。

（3）语音数据：语音数据用于语音识别、语音生成、语音翻译等任务。AI 通过将语音信号（声波）转化为文本或指令进行处理，为语音助手（如 Siri、小爱同学）、智能客服等应用提供支持。

（4）文字数据：广泛应用于自然语言处理（natural language processing，NLP），如文本分类、情感分析、机器翻译、问答系统等。文字数据可从书籍、文章、网页、社交媒体等渠道获取，经过分词、词性标注等处理后，帮助 AI 理解和生成人类语言。

（5）结构化数据：这类数据通常以表格或数据库的形式存在，具有明确的格式和字段，如用户信息表（包含姓名、年龄、地址等字段）、销售数据表（包含日期、产品、销量等字段）。结构化数据适用于传统的数据分析、预测建模等任务，通过标注数据进行学习和决策，如信用卡欺诈检测系统利用用户的交易数据（结构化）来识别异常交易。

这些数据为人工智能提供了学习的基础和决策的依据，通过数据的分析与处理，AI 能够识别模式、做出预测，并完成复杂的任务。

1.2.1.2　平台系统

平台系统为基础层提供了数据的存储、处理和计算资源，确保人工智能能够高效地运行。以下是基础层中重要的几个平台系统：

（1）云计算平台：云计算为人工智能提供了强大的计算资源支持。AI 应用需要大量的计算能力，尤其是深度学习模型的训练，通常需要高性能的计算资源进行并行处理。通过云计算平台，AI 可以动态获取计算资源，按需进行大规模数据处理和存储，无需用户自行搭建昂贵的硬件设施。常见的云计算平台有 Amazon Web Services（AWS）、Google Cloud、Microsoft Azure、阿里云等。

（2）云操作系统：云操作系统为云计算平台提供操作系统级别的管理和控制功能。它负责管理云平台上的虚拟化资源、计算资源、存储资源以及网络连接，实现资源的高效分配和调度。通过云操作系统，人工智能应用可以在云端灵活部署，确保数据处理的高效性和稳定性。

（3）大数据系统：大数据系统处理和分析海量的、复杂的非结构化或半结构化数据，为人工智能提供丰富的训练数据和实时分析能力。例如，Hadoop、Spark 等大数据处理框架，能够分布式处理大规模数据，进行数据清洗、转换、聚合等操作，为 AI 模型训练提供高质量的数据。通过大数据系统，AI 可以快速处理不同来源的数据，进行数据挖掘和知识发现。

（4）计算资源调度系统：计算资源调度系统负责高效地分配和管理计算资源，确保 AI 任务在高效、合理的计算环境下进行。尤其在进行深度学习训练时，计算资源调度系统能够根据任务的优先级和需求（如 CPU、GPU 的数量）动态调整资源分配，避免资源浪费，从而提高系统整体的运行效率。

1.2.1.3　硬件设施

硬件设施为人工智能提供了必要的物理基础。它包括芯片、云计算服务器以及各类传感器等，为 AI 系统提供算力支撑和环境数据输入。

（1）芯片：AI 芯片是专为加速 AI 运算而设计的硬件。它们优化了 AI 计算中的特定任务，如深度学习中的矩阵计算（这是神经网络运算的核心），相比通用 CPU，AI 芯片（如 GPU、TPU、NPU）能显著提高计算效率。例如，NVIDIA 的 GPU 广泛应用于深度学习模型的训练和推理。

（2）云计算服务器：云计算服务器提供了大量的计算资源和存储空间，支持大规模数据的处理和模型训练。这些服务器通常采用集群架构，通过网络连接形成强大的计算集群，AI 能够利用大量的计算力进行数据分析和训练，而无须依赖本地硬件。

（3）传感器：传感器用于感知环境并收集数据，是 AI 系统获取外部信息的“眼睛”和“耳朵”。常见的传感器包括图像传感器（摄像头，用于采集图像信息）、激光雷达（用于测量距离和构建三维环境地图，广泛应用于自动驾驶）、温度传感器（采集温度数据）、压力传感器（感知压力变化）等，这些传感器提供的数据可以被 AI 系统用于实时决策和控制。

（4）GPU/CPU 和 RAM：图形处理器（graphics processing unit，GPU）和中央处理器（central processing unit，CPU）是计算机系统的核心处理单元，GPU 具有强大的并行处理能力，适合处理

深度学习中大量的矩阵运算和大数据分析任务；CPU 则适合处理一般的计算任务和系统控制。随机存取内存（random access memory，RAM）则用于存储数据和程序，为 AI 应用提供快速的存取能力，确保数据处理的高效性。

这些硬件设施共同构成了人工智能基础层的物理基础，它们为数据处理、模型训练和推理提供了强有力的支持。

1.2.2 技术层

技术层是人工智能系统的核心部分，涉及 AI 算法、技术和底层框架的实现。该层主要包括通用技术、算法和底层框架，它们共同为人工智能系统提供强大的智能功能。通过这些技术和算法，AI 能够处理各种复杂的任务，如语音识别、图像处理、决策制定等。

1.2.2.1 通用技术

通用技术是人工智能实现智能功能的基础能力，它们使得 AI 能够处理各种类型的数据并进行复杂的决策。

（1）自然语言处理：致力于使计算机能够理解、解释和生成人类语言。它包括文本分析、语义理解、机器翻译、情感分析等任务，使得 AI 能够与人类进行自然的语言交互，如智能客服与用户的对话、机器翻译软件将中文翻译成英文等。

（2）计算机视觉：使计算机能够“看懂”图像和视频，即从图像或视频中提取有意义的信息。它包括目标检测、图像分类、图像分割、场景理解等技术，应用于人脸识别、自动驾驶的环境感知、医疗影像诊断等领域。

（3）智能语音：涵盖语音识别（将语音转换为文本）、语音合成（将文本转换为语音）、语音唤醒等技术，使得 AI 能够“听懂”人类说话并“开口”回应，如手机的语音助手能够响应“你好，小艺”的唤醒指令，并根据语音指令执行操作。

1.2.2.2 算法部分

人工智能的核心能力来自各类算法。不同的 AI 任务通常需要不同的算法来进行处理和解决。

（1）深度学习：基于深度神经网络的算法，能够自动学习数据的特征表示，在图像识别、自然语言处理等领域取得了巨大成功，如卷积神经网络、循环神经网络、Transformer 等。

（2）人工神经网络：由大量人工神经元连接组成的网络结构，是深度学习的基础，通过模拟人脑神经元的连接方式进行信息处理。

（3）强化学习：通过与环境的交互，智能体从试错中学习最优行为策略，以最大化累积奖励。例如，机器人通过不断尝试行走并根据是否摔倒获得反馈（奖励或惩罚），逐渐学会稳定行走。

（4）集成学习：将多个模型的预测结果结合起来，以提高整体性能和稳定性。常见的集成学习方法包括随机森林（由多个决策树组成，通过投票得出最终预测结果）、梯度提升机等，在分类和回归任务中表现出色。

（5）回归模型：用于预测连续数值型变量，如预测房价、股票价格等。通过建立输入特征与输出变量之间的函数关系，根据已知数据预测未知结果，如线性回归、多项式回归等。

（6）决策树模型：以树状结构表示决策规则，通过对数据的一系列判断（如“年龄是否大于30岁”“收入是否高于5000元”）来进行分类或回归。决策树直观易懂，易于解释，广泛应用于信用评估、风险分析等领域。

（7）支持向量机：在高维空间中寻找一个最优超平面，用于分类或回归任务。它通过最大化分类间隔来提高模型的泛化能力，在文本分类、图像识别等领域有较好的应用效果。

（8）聚类分析：属于无监督学习，用于将数据集中的样本按照相似性划分为不同的簇，使得同一簇内的样本相似度高，不同簇间的样本相似度低。例如，客户分群（将具有相似消费习惯的客户归为一类）、异常检测等。

1.2.2.3　底层框架

为了实现上述算法，人工智能需要依赖底层框架提供支持。底层框架为AI开发人员提供了高效的计算工具、数据处理能力和算法实现平台。常见的人工智能底层框架包括：

（1）TensorFlow。

由Google开发的开源机器学习框架，支持多种编程语言（如Python、C++），提供了丰富的API和工具库，适用于深度学习模型的构建、训练和部署。它支持分布式计算，能够处理大规模数据和复杂模型，在图像识别、自然语言处理等领域应用广泛。

（2）PyTorch。

由Facebook开发的开源深度学习框架，以其动态计算图特性和灵活性受到学术界和工业界的青睐。动态计算图允许开发者在运行时修改模型结构，便于调试和实验。PyTorch在快速原型开发、科研实验以及计算机视觉、自然语言处理等领域有广泛应用。

（3）Caffe。

专注于计算机视觉领域的深度学习框架，其设计强调速度和模块化。Caffe提供了丰富的预训练模型和工具，能够快速实现图像分类、目标检测等任务，适合工业界的快速部署和应用。

（4）MXNet。

由Apache开发的开源深度学习框架，支持多语言接口和分布式计算，具有高效的内存使用和计算性能。MXNet适合大规模深度学习模型的训练和推理，在自动驾驶、语音识别等领域有一定的应用。

1.2.3　应用层

应用层是人工智能技术架构中的最上层，它是AI技术在实际行业和领域中的具体实现。应用层将技术层的算法和模型转化为可操作的解决方案，广泛地应用于各个行业和领域，推动各行业的智能化发展。

（1）智能工业：通过AI技术优化生产流程、提高生产效率和产品质量。例如，在制造业中，利用计算机视觉进行产品缺陷检测，相比人工检测更快速、更精准；通过AI算法预测设备故障，提前进行维护，减少停机时间。

（2）智能交通：借助AI实现交通管理的智能化，如智能交通信号灯根据实时车流量调整信号灯时长，缓解交通拥堵；车牌识别系统自动识别车辆信息，用于停车场收费、交通违章监控等。

（3）智慧城市：整合AI、物联网等技术，对城市的交通、能源、安防等进行智能化管理。例如，智能安防系统通过摄像头和AI分析，实时监测异常行为并报警；智能能源系统根据居民用电

习惯优化电力分配，节约能源。

（4）智能医疗：AI 在医疗领域的应用包括医学影像辅助诊断（如通过 AI 分析 CT、MRI 图像，帮助医生快速发现肿瘤等病变）、疾病预测（基于患者的健康数据预测患病风险）、药物研发（利用 AI 加速药物分子的筛选和设计）等，提高医疗服务的效率和准确性。

（5）自动驾驶：通过融合计算机视觉、激光雷达、GPS 等多种技术，AI 系统能够感知周围环境、识别交通标志和障碍物，并做出加速、刹车、转向等决策，实现车辆的自主行驶，是智能交通领域的重要应用。

（6）智慧农业：利用 AI 技术实现农业生产的精准化管理，如通过传感器收集土壤湿度、光照等数据，AI 系统根据这些数据指导灌溉和施肥，提高农作物产量；利用计算机视觉识别农作物病虫害，及时采取防治措施。

（7）智能政务：将 AI 应用于政府服务和管理中，如智能政务机器人为市民提供政策咨询服务，提高服务效率；通过 AI 分析政务数据，辅助政府进行决策，提升政务管理水平。

（8）智慧财务：AI 在财务领域的应用包括智能记账（自动识别和分类财务票据）、风险控制（通过分析企业财务数据和市场信息，预测信用风险）、智能投顾（根据用户的风险偏好和财务状况，提供个性化的投资建议）等。

（9）智慧零售：通过 AI 技术提升零售行业的运营效率和用户体验，如智能推荐系统根据用户的购物历史推荐商品；无人超市利用计算机视觉和传感器实现自动结算，以减少人工成本。

1.3 人工智能伦理与安全

1.3.1 关注 AI 伦理的意义

随着人工智能技术以前所未有的速度融入我们的生活与工作，从智能语音助手到复杂的医疗诊断系统，从精准的推荐算法到自动驾驶汽车，AI 对社会的影响日益深远。在享受 AI 带来的便利与效率提升的同时，与之相关的伦理问题也逐渐浮出水面，引起了广泛关注。

AI 伦理的核心在于确保人工智能技术的设计、开发与应用符合人类的价值观和道德准则。这一领域的兴起，源于人们对技术快速发展可能带来的潜在风险的担忧。AI 决策的公正性、隐私保护、责任归属等问题，已不再是科幻作品中的假想，而是切实影响着人们生活的现实挑战。

例如，在招聘场景中，AI 算法可能会无意识地强化性别、种族等偏见，导致某些群体在求职过程中受到不公平对待。在医疗领域，AI 辅助诊断系统的决策可能会影响患者的生命健康，其准确性与可靠性至关重要。此外，随着自动驾驶技术的发展，一旦发生事故，责任如何界定成为难题。这些问题的出现，凸显了关注 AI 伦理的紧迫性与重要性。

关注 AI 伦理，不仅是为了规避潜在风险，更是为了充分发挥 AI 技术的潜力，确保其造福人类。只有在伦理框架的约束下，AI 才能在推动社会进步的同时，避免对个人权益、社会公平和人类价值观造成损害。

1.3.2　使用 AI 的基本准则

为了确保人工智能的健康发展与合理应用，一系列使用 AI 的基本准则逐渐形成。这些准则旨在为 AI 开发者、使用者以及相关政策制定者提供指导，帮助他们在技术创新与伦理考量之间找到平衡。

1.3.2.1　公正性与无歧视

AI 系统应避免产生或强化偏见与歧视，确保决策过程对所有个体公平公正。开发者在设计和训练模型时，需仔细审查数据来源，防止数据中存在的历史偏见（如性别、种族、地域等方面的偏见）导致不公平的结果。例如，在信用评估系统中，应避免因数据中存在的性别偏见而对女性用户给予不公正的信用评分；在招聘算法中，应消除对特定群体的歧视，确保所有符合条件的求职者都有平等的机会。

1.3.2.2　透明度与可解释性

AI 系统的决策过程应具备一定程度的透明度，尤其是在涉及重大影响的决策中（如医疗诊断、司法判决辅助等）。可解释性 AI 技术的发展，使得人们能够理解 AI 模型为何做出特定决策，从而增强对 AI 系统的信任。例如，在医疗诊断中，医生和患者需要理解 AI 辅助诊断系统得出某一诊断结果的依据（如哪些特征在决策中起了关键作用），以便制订更合适的治疗方案；在贷款审批中，申请人有权知道 AI 系统拒绝其贷款申请的原因。

1.3.2.3　隐私保护与数据安全

随着 AI 对数据的高度依赖，保护用户隐私和数据安全变得尤为重要。开发者必须遵循严格的数据保护法规（如欧盟的《通用数据保护条例》），采取加密、匿名化、去标识化等技术手段，确保用户数据在收集、存储和使用过程中的安全性。例如，智能设备在收集用户的位置、健康等敏感数据时，应明确告知用户数据的用途、存储期限，并获得用户的明确同意；在数据传输过程中，采用加密技术防止数据被窃取。

1.3.2.4　责任归属明确

在 AI 系统的开发、部署与使用过程中，明确责任归属至关重要。当 AI 决策导致不良后果时，需要确定开发者、使用者或其他相关方的责任，避免责任“真空”。例如，在自动驾驶事故中，需要明确汽车制造商（若因硬件缺陷导致事故）、软件开发者（若因算法漏洞导致决策失误）以及驾驶员（若因未及时接管车辆导致事故）各自的责任范围；在医疗 AI 误诊案例中，需明确医院、AI 系统开发者以及医生的责任。

1.3.2.5　人类监督与控制

尽管 AI 技术具有强大的自主决策能力，但人类的监督与控制不可或缺。在关键领域，如医疗、金融和交通，人类应保留最终决策权，确保 AI 决策符合人类价值观和社会利益。例如，在金融交易中，AI 算法可提供投资建议，但最终的交易决策权由人类投资者掌握；在手术机器人辅助手术时，医生应全程监督并在必要时干预机器人的操作，确保手术安全。

这些基本准则为构建一个安全、可靠、符合伦理道德的 AI 生态系统提供了基础。在实际应用中，需要各方共同努力，将这些准则融入 AI 的全生命周期中，推动人工智能技术朝着对人类有益的方向发展。

1.3.3 AI 安全风险与应对策略

人工智能技术在带来巨大机遇的同时，也伴随着一系列安全风险。这些风险涉及多个层面，对个人隐私、社会稳定乃至国家安全都构成潜在威胁。了解这些风险并制定相应的应对策略，是确保 AI 安全发展的关键。

1.3.3.1 数据安全风险

AI 系统依赖大量数据进行训练，数据的安全性直接影响到 AI 的性能与可靠性。数据泄露（如用户的个人信息、企业的商业数据被非法获取）、篡改（如训练数据被恶意修改，导致模型学习到错误信息）或被恶意利用，可能导致 AI 模型做出错误决策。例如，黑客攻击医疗数据系统，篡改患者的病历数据，可能使 AI 辅助诊断系统基于错误的数据做出错误的诊断，危及患者生命健康；企业的客户数据被泄露，可能导致客户遭受诈骗等损失。

应对策略：采用加密技术（如对称加密、非对称加密）保护数据在传输和存储过程中的安全；建立严格的数据访问控制机制，对数据的访问进行身份认证和权限管理，确保只有授权人员能够访问敏感数据；定期进行数据备份，并对备份数据进行加密存储，防止数据丢失；加强数据安全监测，及时发现和处理数据泄露、篡改等安全事件。

1.3.3.2 算法安全风险

AI 算法本身可能存在漏洞或被恶意攻击，导致算法失效或被操控。例如，对抗样本攻击通过向 AI 模型输入精心构造的微小扰动数据（人眼难以察觉），使模型做出错误预测。在图像识别系统中，攻击者可以对交通标志图片进行微小修改，使 AI 将“停止”标志误识别为“直行”标志，导致自动驾驶车辆发生事故；在垃圾邮件过滤系统中，通过对垃圾邮件进行特定修改，可使 AI 将其误判为正常邮件。

应对策略：加强算法安全性测试，通过模拟各种攻击场景（如对抗样本攻击、数据投毒攻击），发现并修复算法漏洞；采用鲁棒性更强的算法设计，提高模型对异常输入的抵抗能力（如在模型训练过程中加入对抗样本进行增强训练）；建立算法安全监测机制，实时监控算法的运行状态和输出结果，及时发现并应对异常情况。

1.3.3.3 隐私侵犯风险

AI 在收集和分析用户数据的过程中，可能会侵犯用户隐私。例如，一些智能设备在未经用户充分授权的情况下，过度收集用户的日常行为数据（如浏览记录、通话记录），并用于商业目的（如定向广告推送）。此外，AI 技术还可能通过对非敏感数据的分析，推断出用户的敏感信息（如通过分析用户的购物记录、社交网络数据，推断出其政治倾向、健康状况等），这种“隐私推断”也构成了对用户隐私的侵犯。

应对策略：制定严格的隐私保护法规，明确数据收集、使用和共享的边界，禁止过度收集和滥用用户数据；采用匿名化、差分隐私等技术手段，在不影响数据可用性的前提下，保护用户隐私（如差分隐私通过在数据中加入噪声，使攻击者无法准确识别个体信息）；加强用户隐私教育，提高用户对个人数据保护的意识，引导用户合理授权数据使用。

1.3.3.4 自主武器系统风险

随着 AI 技术在军事领域的应用，自主武器系统（能够在没有人类干预的情况下自主选择和攻

击目标的武器）的发展引发了广泛关注。这类武器系统可能导致战争行为的失控，因为它们缺乏人类的道德判断和情感考量，可能会误判目标或在不适当的情况下发动攻击，引发严重的人道主义危机。

应对策略：通过制定国际公约和法规，限制自主武器系统的研发与使用，推动建立全球性的监管框架；确保人类在军事决策中的最终控制权，明确规定自主武器系统必须在人类的监督和授权下才能使用，避免将生死攸关的决策权完全交给机器；开展国际合作，共同探讨自主武器系统的伦理与安全问题，促进各国在这一领域的共识与协作。

1.3.3.5　AI 滥用风险

AI 技术可能被滥用，用于恶意目的，如制造虚假信息（“深度伪造”技术可生成逼真的虚假图像、视频或音频，如伪造政治人物的讲话视频）、进行网络诈骗（利用 AI 生成逼真的语音模仿他人进行诈骗）或实施恐怖袭击（如利用 AI 控制无人机进行攻击）。这些滥用行为可能误导公众、破坏社会稳定、威胁公共安全。

应对策略：加强对 AI 技术应用的监管，建立审查机制，对高风险 AI 应用（如深度伪造技术）进行备案和审核，防止 AI 被用于非法或不道德的活动；提高公众对 AI 滥用风险的认识，通过媒体宣传、教育等方式，增强公众辨别虚假信息的能力；开发检测和防范 AI 滥用的技术工具，如虚假信息识别系统（通过分析内容的特征识别深度伪造内容）、AI 滥用监测平台等。

面对 AI 带来的安全风险，需要政府、企业、科研机构和社会各界共同努力，通过技术创新（如开发更安全的算法、隐私保护技术）、法规完善（如制定针对性的 AI 安全法规）、教育普及（如提高公众的 AI 安全意识）等多种手段，构建全方位的 AI 安全防护体系，确保人工智能技术在安全的轨道上发展。

思考与练习

一、选择题

1. 下列属于人工智能“像人一样行动”这一主流定义视角的经典判据是（　　）。

A. 认知科学实验　　B. 图灵测试

C. 三段论推理　　D. 理性决策模型

2. 我们日常使用的智能音箱、推荐算法等，属于人工智能按应用范围与能力强度划分中的（　　）。

A. 专用人工智能　　B. 通用人工智能

C. 超级人工智能　　D. 强人工智能

3. 人工智能能力层级中，负责从外部获取原始数据并转换为机器可处理的信息的是（　　）。

A. 感知层　　B. 认知层

C. 决策层　　D. 创造层

4. 下列属于人工智能技术架构中基础层硬件设施的是（　　）。

A. 自然语言处理　　B. TensorFlow 框架

C. GPU 芯片　　D. 智能医疗系统

5. 生成式 AI 技术属于人工智能关键特征中的（　　）。

A. 多模态融合　　B. 生成式突破

C. 自主智能体　　D. 高效轻量化

6. 下列属于人工智能应用层的是（　　）。

A. 云计算平台　　B. 决策树算法

C. 智能交通系统　　D. 语音识别技术

7. 使用 AI 的基本准则中，要求 AI 系统决策过程对所有个体公平公正的是（　　）。

A. 透明度与可解释性

B. 公正性与无歧视

C. 隐私保护与数据安全

D. 人类监督与控制

二、问答题

1. 请简述人工智能按技术学派与实现路径划分的三大主流学派及其核心思想。

2. 人工智能的能力层级从下到上依次为感知层、认知层、决策层与创造层，请说明认知层的核心功能。

3. 人工智能技术架构由基础层、技术层和应用层构成，简述技术层包含的主要内容。

4. 请列举人工智能安全风险中的两种，并分别说明对应的应对策略。

三、案例题

1. 某电商平台使用 AI 推荐算法为用户推荐商品，该算法通过分析用户的浏览记录、购买历史等数据，向用户推送其可能感兴趣的商品。

（1）该 AI 推荐算法属于按应用范围与能力强度划分中的哪一类人工智能?

（2）从人工智能能力层级来看，该算法主要体现了哪个层级的功能?

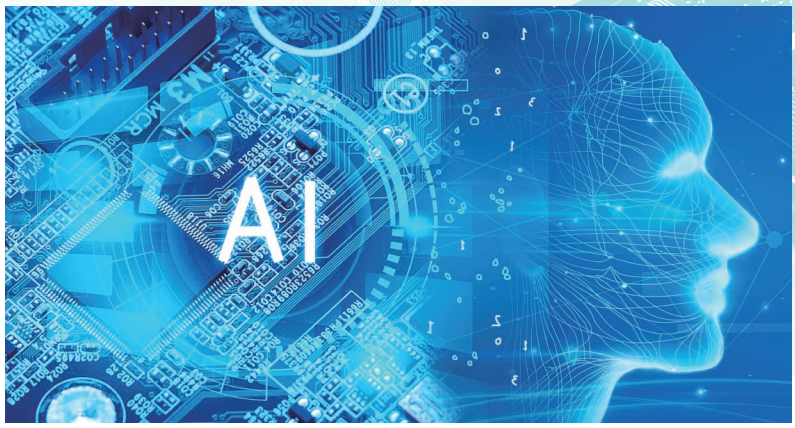
（3）分析该应用可能涉及的伦理问题，并提出相应的应对思路。

2. 某工厂引入 AI 系统对生产的产品进行缺陷检测，通过摄像头拍摄产品图像，由 AI 系统自动识别产品是否存在缺陷。

（1）该 AI 系统在技术架构中属于哪一层?

（2）该系统主要运用了技术层中的哪些通用技术和算法?

（3）若该系统出现误判，导致合格产品被判定为缺陷产品，从责任归属角度分析可能涉及哪些相关方的责任。



项目 2
机器学习与神经网络

项目导读

机器学习与神经网络是人工智能领域的核心技术，它们让计算机能够从数据中自主学习规律，实现预测、分类、决策等智能行为。本项目将带你深入探索机器学习的三大类型——监督学习、无监督学习和强化学习，了解它们的算法原理与应用场景。同时，你还会走进神经网络的世界，从单个神经元的工作机制，到人工神经网络的层级结构，再到卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network，RNN）、生成对抗网络（generative adversarial network，GAN）等典型网络的独特功能，感受深度学习如何攻克图像识别、自然语言处理等复杂任务。通过学习，你将掌握机器学习的完整流程，理解神经网络的学习奥秘，为运用这些技术解决实际问题打下基础。

学习目标

熟悉机器学习的完整工作流程，包括数据收集与预处理、特征工程、模型选择与训练、模型评估与优化以及模型部署与监控等各个环节及其作用。

掌握监督学习、无监督学习、强化学习的基本原理和适用场景。

掌握神经元的工作过程、在神经网络中的作用及与学习过程的关系，了解人工神经网络的基本构成和工作原理（前向传播与反向传播）。

熟悉 CNN、RNN 及 GAN 的基本结构、核心机制和典型应用。

技能目标

能够根据具体问题场景，判断适用的机器学习类型并选择相应的算法。

可以运用常见的监督学习算法完成简单的建模任务，并使用相关评估指标（如准确率、均方误差）评估模型性能。

能描述不同神经网络的结构特点，并举例说明其在实际领域的应用。

具备初步分析不同算法和网络模型优缺点的能力，能根据任务需求和数据特点选择合适的模型。

思政目标

培养严谨的科学态度和精益求精的工匠精神，认识到机器学习与神经网络流程中每个环节的重要性，任何一个步骤的疏忽都可能影响最终结果，从而在学习和实践中注重细节、追求精准。

树立创新意识和探索精神，了解机器学习与神经网络技术的不断发展，激发对新技术、新算法的探索热情，勇于尝试用新方法解决实际问题。

增强数据伦理意识，认识到机器学习依赖数据，在数据收集、使用过程中要遵守法律法规，保护个人隐私和数据安全，避免数据滥用带来的伦理问题。

培养合作与团队精神，理解复杂的机器学习项目往往需要团队协作完成数据处理、模型设计、部署应用等多个环节，学会在团队中发挥自身优势，共同达成目标。

2.1 机器学习

2.1.1 机器学习的基础

2.1.1.1 机器学习的定义

机器学习是一种通过数据驱动的方法，借助算法使计算机从数据中提取规律并进行预测或决策的技术。它与传统编程方式有着本质区别：传统编程依赖程序员编写固定规则实现功能，而机器学习则通过分析训练数据自主发现内在规律。其核心在于“学习”——让计算机通过积累“经验”（数据）持续改进决策能力。



机器学习的定义

例如，语音识别系统的开发无需人工编写所有语音转换规则，而是通过大量语音数据与对应

文本的训练，自主学习语音信号与文字的映射关系。当遇到新的语音输入时，系统能基于训练所得规律自动完成识别任务。这种数据驱动的特性，使机器学习在处理复杂、多变的现实问题时展现出显著优势。

2.1.1.2　机器学习的分类

机器学习的算法和技术可以根据任务类型、输入输出数据的性质以及学习过程中是否需要监督信息的提供，分为三大类：监督学习、无监督学习和强化学习。

（1）监督学习。

训练数据包含输入数据及对应标签（已知输出），目标是学习输入与输出的映射关系，构建能对新数据进行预测的模型。核心是通过“监督信号”（标签）指导学习过程，典型应用如房价预测（输入房屋特征，输出价格）、垃圾邮件识别（输入邮件内容，输出“垃圾 / 正常”标签）。

（2）无监督学习。

训练数据仅有输入部分，无标签信息。目标是自主探索数据中隐藏的模式或结构，如数据分组、特征关联等。它更接近人类对未知世界的认知方式，典型应用包括客户分群（将消费习惯相似的客户归类）、异常检测（识别与正常模式不符的数据）。

（3）强化学习。

智能体通过与环境交互，从“试错”中学习最优策略。智能体执行动作后会获得奖励或惩罚，目标是通过累积经验最大化长期奖励。与前两者不同，它没有固定的输入 - 输出对，而是通过动态交互优化决策，典型应用如自动驾驶（根据路况调整行驶策略）、游戏 AI（通过对战提升胜率）。

2.1.1.3　机器学习的工作流程

机器学习的完整流程涵盖从数据收集与预处理到模型部署与监控的全生命周期，每个环节都直接影响最终效果，如图 2-1 所示。

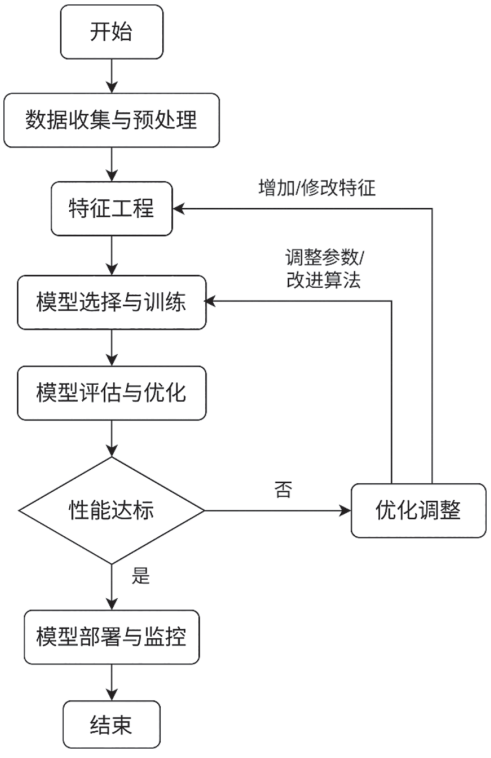


图 2-1　机器学习的工作流程

（1）数据收集与预处理。

数据是机器学习的基础，可通过公开数据集、传感器、用户日志等多种渠道获取，类型包括文本、图像、音频等。但原始数据往往存在噪声（如错误值）、缺失值（如信息不全）、异常值（如极端数据），需通过以下步骤处理：

①数据清洗：修正错误、填充缺失值、移除异常值。

②数据转换：将非数值数据（如文本、类别）转换为数值形式。

③规范化 / 标准化：统一数据尺度（如将数值缩放到 0 ～ 1 范围）。

④数据划分：将数据集分为训练集（模型学习）、验证集（参数调优）、测试集（性能评估）。

（2）特征工程。

从原始数据中提取有效信息，构建能提升模型性能的特征，是机器学习的关键环节。主要包括：

①特征选择：筛选与目标相关的特征（如用过滤法保留高相关性特征）。

②特征提取：从原始数据中提炼抽象特征（如从文本中提取 TF-IDF 值，从图像中提取边缘特征）。

③特征构建：结合业务需求创建新特征（如将 " 出生日期 " 转换为“年龄”，将“销售额”与“成本”组合为“利润”）。

特征工程高度依赖领域知识，优质特征能让简单模型达到出色效果，反之则可能导致复杂模型表现不佳。

（3）模型选择与训练。

根据任务类型（如分类、回归）和数据特点选择合适的算法（如线性回归适合数值预测，决策树适合分类）。训练过程通过优化算法（如梯度下降）调整模型参数，使预测结果尽可能接近真实值。例如，在训练线性回归模型时，需找到最优的权重参数，使预测值与实际值的误差最小。

（4）模型评估与优化。

通过验证集和测试集评估模型的泛化能力（对未知数据的预测效果），避免过拟合（训练集表现好，测试集表现差）。常见评估指标包括：

①分类问题：准确率（正确预测的比例）、精确率（预测为正例的样本中实际为正例的比例）、召回率（实际为正例的样本中被正确预测的比例）、F1 值（精确率与召回率的调和平均）。

②回归问题：均方误差（MSE，预测值与实际值差值的平方均值）、均方根误差（RMSE，MSE 的平方根）、R^2（模型解释数据变异的比例）。

③聚类问题：轮廓系数（衡量簇内相似度与簇间差异度）。

根据评估结果优化模型，可通过调整超参数（如决策树深度）、改进特征、更换算法（如用集成学习替代单一模型）等方式实现。

（5）模型部署与监控。

将优化后的模型应用于实际场景，如嵌入 APP、网站等系统中进行实时预测。部署后需持续监控模型表现，因为数据分布可能随时间变化（如用户行为改变）导致模型效果下降，此时需重新训练或调整模型。

2.1.2 监督学习

监督学习（supervised learning）是任务驱动的学习范式，其核心在于利用带有明确标签的训

练数据来构建预测模型。根据输出标签类型的不同，监督学习主要分为回归与分类两大任务。

2.1.2.1　回归算法

回归算法是一种机器学习算法，用于根据输入数据预测数值。回归算法试图通过对数据拟合数学模型来找到输入变量和输出变量之间的关系。回归的目标是找到输入特征和目标变量之间的数学关系，可用于对新的、看不见的数据进行准确预测。

（1）线性回归算法。

线性回归是目前机器学习算法中最流行的一种，线性回归算法就是要找到一条直线，并且让这条直线尽可能地拟合散点图中的数据点。它试图通过将直线方程与该数据拟合来表示自变量（x）和数值结果（y）。然后就可以用这条直线来预测未来的值，如图 2-2 所示。

这种算法最常用的技术是最小二乘法。利用这个方法计算出最佳拟合线，使其与直线上的每个数据点垂直距离最小。总距离是所有数据点的垂直距离（绿线）的平方和。其思想是通过最小化这个平方误差或距离来拟合模型。

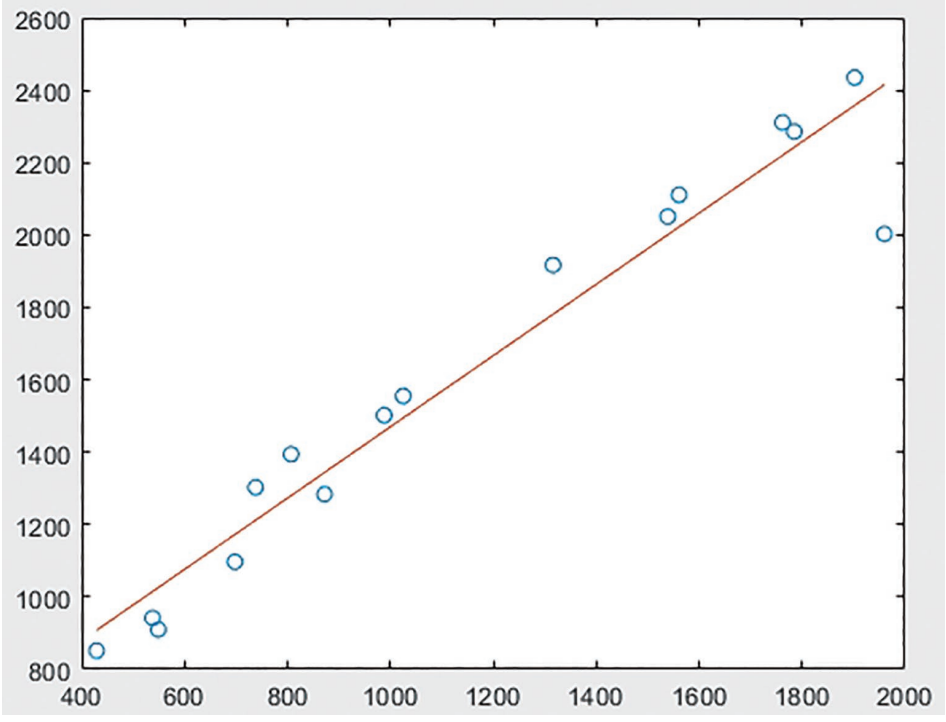


图 2–2　线性回归拟合直线

（2）多项式回归算法。

多项式回归是线性回归的延伸，其中变量之间的关系使用多项式方程建模。这允许在捕捉输入特征和目标变量之间的非线性关系方面具有更大的灵活性。它包括在线性方程中添加多项式项，如 x^2 或 x^3。。当数据呈现曲线模式时，多使用多项式回归是有用的。例如：分析商品价格与市场需求量之间的非线性关系，随着商品价格升高，商品需求量逐步下降，呈现非线性递减曲线，如图 2-3 所示。

回归算法在数据分析和预测中的应用非常广泛，从简单的线性回归到复杂的集成学习方法，每种算法都有其独特的优势和局限性。了解和掌握这些算法的原理和使用方法，对于解决实际问题具有重要意义。在选择具体的回归算法时，需要综合考虑数据的特点、问题的复杂性以及计算资源等因素。

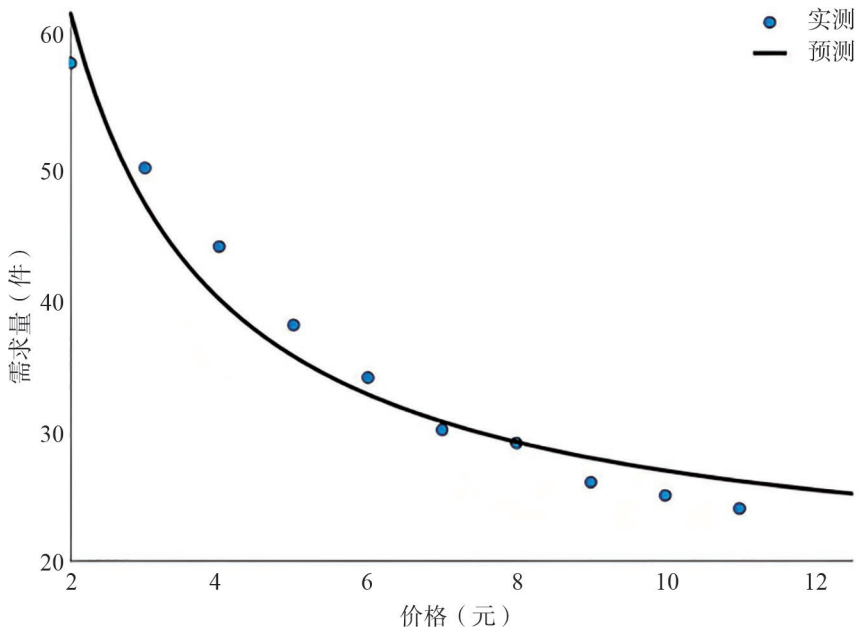


图 2-3　多项式回归拟合曲线

2.1.2.2　分类算法

机器学习算法中的分类算法是将分类规则应用于记录的目标映射，将它们划分为不同的类别，并建立具有泛化能力的算法模型，即构建映射规则以预测未知样本的类别。

机器学习的分类算法包括逻辑回归、决策树、支持向量机、最近邻算法、贝叶斯分类器、神经网络等。

（1）逻辑回归算法。

逻辑回归（logistic regression，LR）虽然借鉴了线性回归的思想，即先对输入特征进行加权求和，但最终是解决分类问题，尤其是二分类问题的基石算法。尽管其名称中包含“回归”，但它本质上是一种分类模型，而非回归模型。

LR 分类问题主要在于通过 h 函数得到分类的权值，求出一个样本的分类概率。LR 的 h 函数是一个 Sigmoid 函数，Sigmoid 函数是一个“S”形，如图 2-4 所示。Sigmoid 函数将中间结果映射到结果变量 Y，其值范围是从 0 到 1，该值可以解释为 Y 出现的概率。S 型 Sigmoid 函数的性质使得逻辑回归更适合用于分类任务。

（2）决策树算法。

决策树算法是一种流行的机器学习方法，用于分类和回归任务。它通过学习数据特征之间的简单决策规则来预测目标变量。决策树是一种树形结构的算法，其中每个内部节点代表一个属性上的测试，每个分支代表一个测试结果，每个叶节点代表一个类或值，其结构如图 2-5 所示。

（3）支持向量机算法。

支持向量机（support vector machine，SVM）是一种高效的二分类模型，以寻找最大间隔超平面为特点，通过凸二次规划问题求解来实现分类。

SVM 的主要目标是找到一个超平面，该平面可以明确地划分不同的类别，并且使得两个类别之间的间隔最大化，如图 2-6 所示。在这个超平面中，距离最近的样本点称为支持向量，这些支持向量是构建决策边界的关键。为了确保模型具有较好的泛化能力，SVM 采用间隔最大化

原则。这意味着在所有的超平面中，SVM 寻找的是那种可以使不同类别之间的间隔最大的超平面。

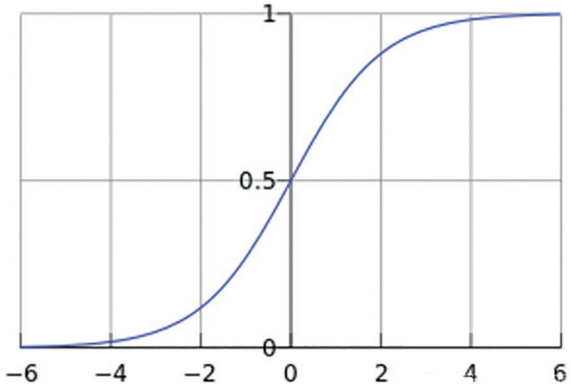


图 2–4　S 型函数

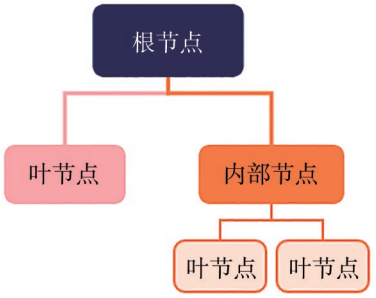


图 2–5　决策树的结构

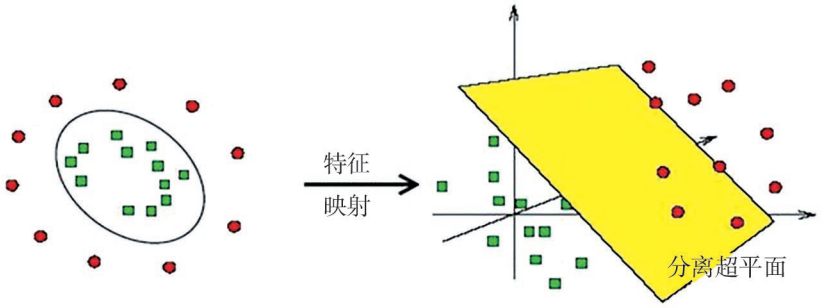


图 2–6　支持向量机算法示意图

（4）最近邻算法

最近邻（K-nearest neighbor，KNN）算法是机器学习中一种基本且易于理解的分类和回归方法。KNN 算法简单直观，不需要进行显式训练，是一种懒惰学习算法，因为它实际上并不构造一个全局的分类模型，而是等到需要对新样本进行预测时才执行计算。

KNN 算法首先需要计算新样本与数据集中每个样本之间的距离，如图 2-7 所示。常用的距离度量方法包括欧氏距离、曼哈顿距离等。确定 K 的值是 KNN 算法的关键步骤之一。K 值的选择对模型的性能有重要影响。通常，K 值的选择通过交叉验证来得到最优解。在找到最近的 K 个邻居之后，KNN 算法通常会基于这 K 个邻居中的多数类别来预测新样本的类别。在回归问题中，则可能取这 K 个邻居的目标值的平均值作为预测结果。

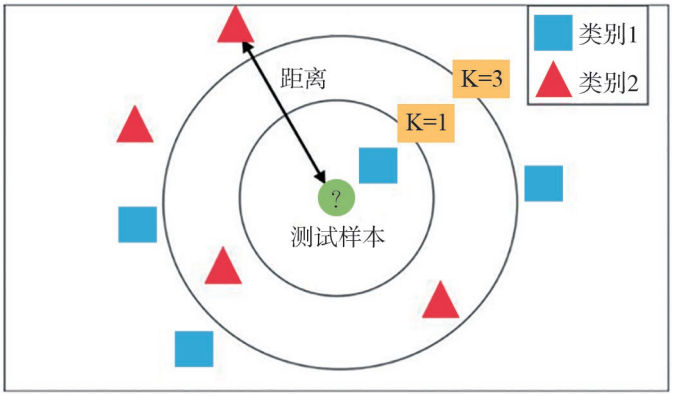


图 2–7　KNN 算法示意图

2.1.3 无监督学习

与监督学习不同，无监督学习（unsupervised learning）处理的是不带任何标签的原始数据。其核心目标不是预测一个已知的结果，而是自主地探索和发现数据内部隐藏的结构、模式与关联。这使得无监督学习在探索性数据分析、特征工程等领域扮演着至关重要的角色。无监督学习最主要的两类任务是聚类与降维。

2.1.3.1 聚类算法

聚类（clustering）算法是机器学习领域中的一个重要分支，属于无监督学习。与监督学习不同，聚类不需要标记原始数据，只需要根据数据固有的结构特征进行聚集。从而形成簇群，并实现数据的分离。

聚类是将相似的对象分组到一起的过程，目的是使得同一组内的对象相似度较高，而不同组之间的相似度较低。聚类和分类的主要区别在于，聚类不关心数据属于哪种类别，而是把具有相类似特征的数据聚集起来形成某一类别的簇。

常见的聚类算法包括 K-Means 算法、层次聚类、谱聚类。

（1）K-means 算法。

将数据集划分为 K 个簇，通过最小化簇内平方和（每个样本到簇中心的距离平方和）优化聚类结果。步骤如下：

①随机选择 K 个初始簇中心。

②计算每个样本到 K 个中心的距离，分配给最近的簇。

③重新计算每个簇的中心（簇内样本的平均值）。

④重复步骤②～③，直至簇分配稳定或达到迭代次数。

如图 2-8 所示，经过迭代后，无标签数据被清晰划分为不同簇，每个簇围绕其中心分布。K 值需预先指定（可通过肘部法选择使误差骤降的 K 值）。

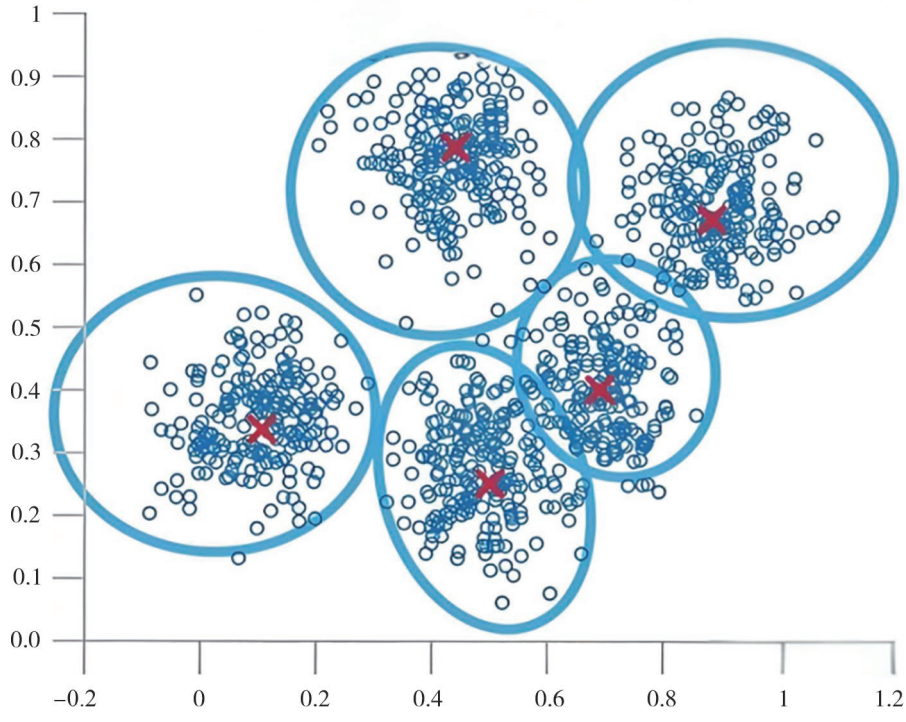


图 2–8 K–Means 算法迭代过程示意图

（2）层次聚类算法。

层次聚类算法是另一类重要的聚类方法，它旨在构建一个嵌套的簇的层次结构。与 K-means 算法需要预先指定簇数 K 不同，层次聚类能够生成一个完整的、从单个数据点到包含所有点的簇的多层级结构，其结果通常可以用一个称为谱系图的树状图来直观地展示，如图 2-9 所示。

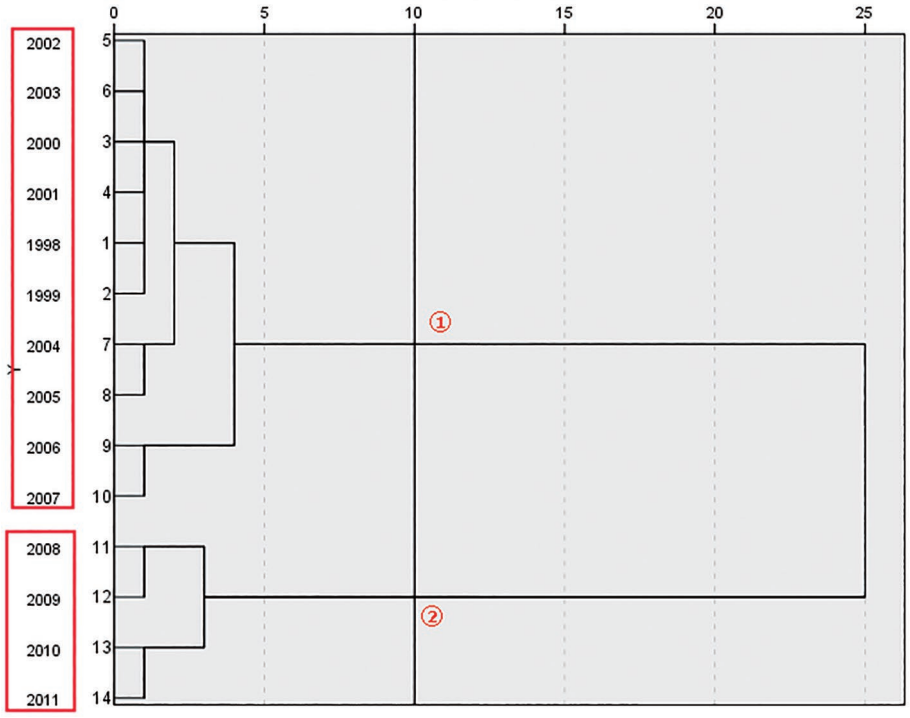


图 2–9　谱系图的树状图

层次聚类主要分为两种策略：

①凝聚式（自底向上）。初始时，每个数据点自成一簇。在每一步迭代中，算法会合并最相似（距离最近）的两个簇，直到所有点都合并到同一个簇中为止。

②分裂式（自顶向下）。初始时，所有数据点同属于一个簇。在每一步迭代中，算法会将当前最大的簇分裂成两个子簇，直到每个点都自成一簇或达到预设的终止条件。

在实际应用中，凝聚式层次聚类更为常见。通过在不同高度“切割”谱系图，用户可以根据需要获得任意数量的簇。

（3）谱聚类算法。

谱聚类算法是一种基于图论的聚类方法，它并不直接在原始数据空间中度量距离，而是将聚类问题转化为对数据关系图的划分问题，因此在处理非凸形状（如环形、月牙形）的簇时表现尤为出色，如图 2-10 所示。对于月牙形的数据分布，谱聚类则能准确地识别出其内在结构。

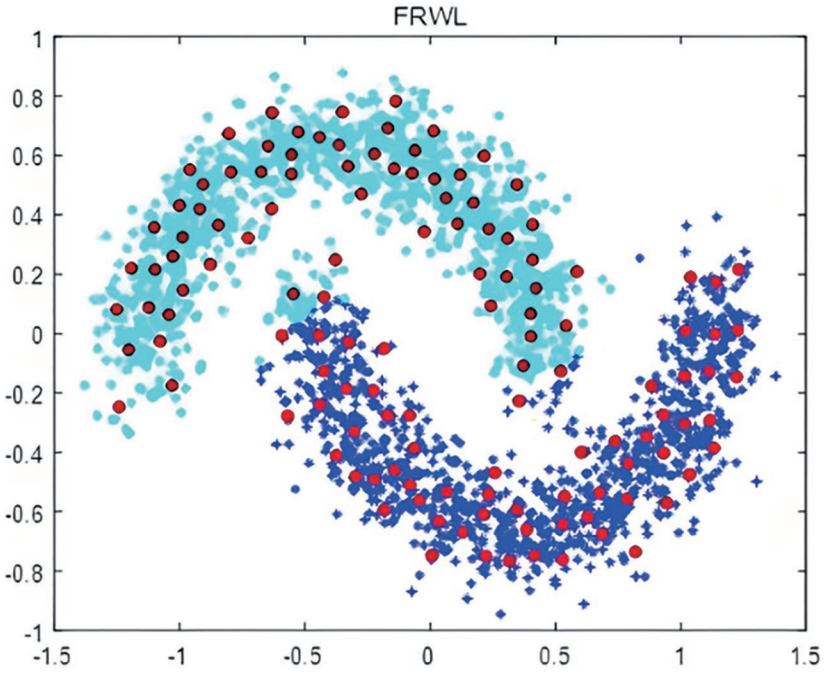


图 2–10　谱聚类月牙形数据识别

其核心思想是：首先，将所有数据点视为图的顶点，点与点之间的相似度（如高斯核函数计算的相似度）作为连接边的权重，从而构建一个加权无向图。然后，聚类任务就变成了如何切割这个图，使得被切断的边的总权重尽可能低（即不同簇之间的相似度低），同时保证每个子图的大小相对均衡。这一图划分问题可以通过对图的拉普拉斯矩阵进行特征分解来求解。算法最终在由拉普拉斯矩阵的前 K 个特征向量构成的低维空间中，再利用 K-means 等传统方法进行聚类，如图 2-11 所示。

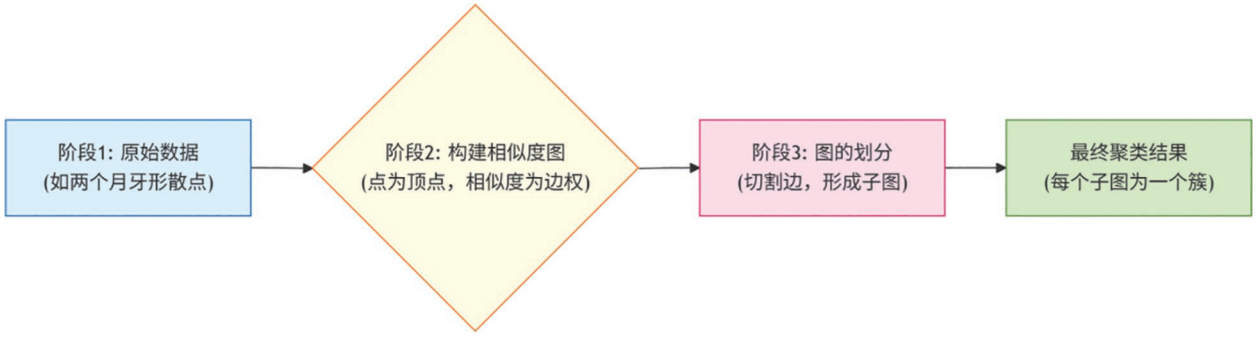


图 2–11　谱聚类核心流程示意图

在选择适合的聚类算法时，需要考虑数据的特性、簇的形状、算法的可扩展性以及是否需要预先指定簇的数量等因素。不同的聚类算法适用于不同类型的数据和应用场景，因此理解每种算法的特点和适用条件对于实际应用至关重要。

2.1.3.2　降维算法

降维旨在保留数据关键信息的前提下，将高维数据映射到低维空间，以解决“维度灾难”（高维数据导致的计算量大、模型性能下降的问题）。

（1）主成分分析（principal component analysis，PCA）。

最常用的线性降维算法，通过寻找正交的主成分（新坐标轴），使数据在主成分上的方差最大化（如图 2-12 所示）。主成分由数据协方差矩阵的特征向量确定，第一个主成分对应最大特征值，代表数据最主要的变化方向。保留前 K 个主成分可实现降维，同时去除噪声，常用于高维数据可视化。

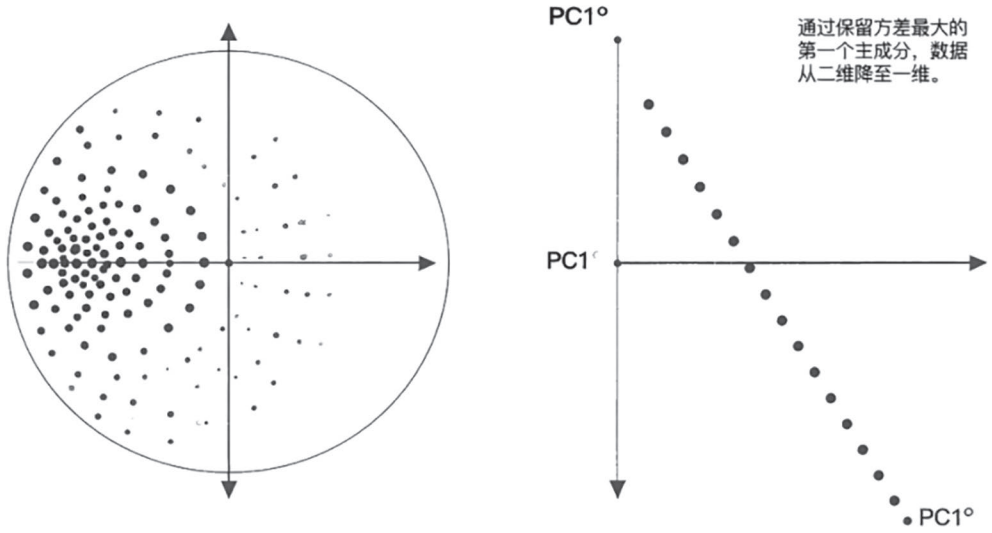


图 2–12　主成分分析（PCA）降维过程示意图

（2）t-分布随机邻域嵌入（t-SNE）。

该算法为非线性降维算法，擅长保留数据的局部邻域结构，可视化效果优异（如图 2-13 所示）。通过将高维空间的相似度转化为概率分布，在低维空间中优化概率分布使其匹配，从而清晰揭示数据的簇状结构。适用于高维数据（如文本、图像）的可视化，但计算复杂度高，不适合大规模数据降维。



（a）PCA 降维效果　　（b）t-SNE 降维效果

图 2–13　PCA 与 t–SNE 对高维数据降维可视化的效果对比

2.1.4 强化学习

强化学习（reinforcement learning，RL）是受行为主义心理学启发的学习范式，智能体通过与环境交互，在“试错”中学习最优决策策略。

强化学习的整个过程可以被模型化为一个持续的“感知—行动—反馈”循环，其基本框架如图 2-14 所示。这个框架由以下几个核心要素构成：

（1）智能体（agent）：学习的主体，也是决策的执行者。它可以是一个机器人、一个游戏角色，或者一套自动驾驶系统。

（2）环境（environment）：智能体所处的外部世界，它定义了智能体可以存在的状态以及可以执行的行动。

（3）状态（state，S）：对环境在某一特定时刻的完整描述。例如，在棋类游戏中，状态就是当前棋盘的布局；在自动驾驶中，状态可能包括车辆的速度、位置以及周围车辆和障碍物的信息。

（4）行动（action，A）：智能体在感知到某个状态后，可以采取的一组行为。例如，棋手可以走的下一步棋，或者自动驾驶汽车可以执行的转向、加速、刹车等操作。

（5）奖励（reward，R）：一个标量反馈信号，是环境对智能体在某个状态下采取某个行动后所产生的即时结果的评价。奖励是引导智能体学习的关键：正奖励表示鼓励当前行为，负奖励（或称惩罚）则表示抑制当前行为。例如，赢得棋局获得正奖励，发生碰撞则获得负奖励。

学习过程为：在 t 时刻，智能体感知状态 S_t，根据策略 π 选择行动 A_t；环境接收行动后转移到新状态 S_{t+1}，并反馈奖励 R_{t+1}；智能体通过不断循环优化策略，最终找到能最大化长期累积奖励的最优策略。

例如，训练机器人走迷宫时，机器人（智能体）在迷宫（环境）的某一位置（状态）选择“前进”“左转”等行动，到达终点获得正奖励，撞到墙壁获得负奖励，通过反复尝试学习最短路径策略。

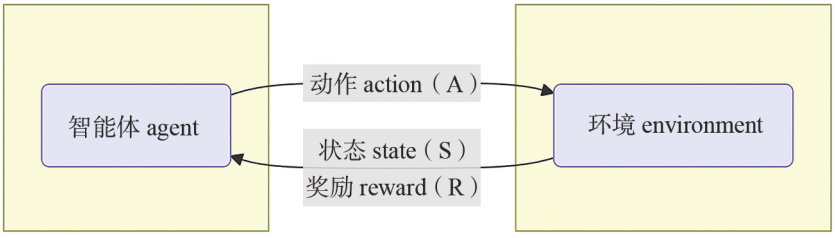


图 2-14　强化学习基本框架示意图

2.2 神经网络

神经网络（neural networks）是人工智能中的一种重要方法，受生物神经系统的启发，模仿大脑神经元之间的连接和信息传递方式。神经网络是一种用于处理复杂问题的机器学习算法，能够通过学习大量数据中的模式，从而进行预测或分类。随着深度学习的发展，神经网络在许多领域（如