新时代·技术新未来

深入浅出 Python 人工智能

段小手　著

清華大學出版社
北　京

内 容 简 介

本书以 Python 为工具、以实战为核心，构建从入门到进阶的 AI 完整学习体系，适合零基础学习者、转型从业者及 AI 爱好者。全书共 18 章，涵盖传统机器学习算法、数据工程化能力、深度学习及生成式 AI 等核心内容，前言解析职业前景，结束语提供进阶建议。

本书贴合初学者认知规律，详略安排得当、实用性强，以幽默的语言和贴近生活的实例引导读者入门，配套可直接落地的 Python 代码示例，避免了纯理论堆砌。

读者对象包括零基础入门者、跨岗转型人员、院校学生、创业者及各类 AI 兴趣爱好者，既可作为入门引路书，也可作为提升实战能力的工具书。

图书在版编目（CIP）数据

深入浅出 Python 人工智能 / 段小手著 . -- 北京 : 清华大学出版社 , 2026. 7.
(新时代 • 技术新未来). -- ISBN 978-7-302-72096-6

Ⅰ. TP312.8；TP18

中国国家版本馆 CIP 数据核字第 20264EX091 号

责任编辑：刘　洋
封面设计：徐　超
版式设计：方加青
责任校对：宋玉莲
责任印制：刘海龙

出版发行：清华大学出版社
网　　址：https://www.tup.com.cn，https://www.wqxuetang.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-83470000　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：大厂回族自治县彩虹印刷有限公司
经　　销：全国新华书店
开　　本：187mm×235mm　　**印　　张：**20.75　　**字　　数：**513 千字
版　　次：2026 年 8 月第 1 版　　**印　　次：**2026 年 8 月第 1 次印刷
定　　价：99.00 元

产品编号：114310-01

前言 PREFACE

人工智能，更火了

几年前，笔者曾写过《深入浅出 Python 机器学习》一书，彼时人工智能还只是科技领域大佬们口中充满前瞻性的概念，仿佛离我们的日常生活很遥远。然而，时光飞逝，如今人工智能已无处不在：使用手机时，AI 助手随叫随到；出行路上，智能驾驶能自动规避障碍、平稳跟车；走进医院，AI 算法可辅助医生识别病灶。它早已从“遥远概念”变成“日常工具”，像一位无形的助手，悄然融入生活的方方面面，为我们带来便捷与高效。

当下，人工智能领域呈现出蓬勃发展的态势，诸多新技术、新应用层出不穷。例如，由中国知名量化投资机构幻方量化孵化的 DeepSeek，便是通用人工智能（AGI）领域的先锋探索者。它专注于大语言模型与多模态技术的底层创新，其 DeepSeek-V3 模型仅以 557 万美元的训练成本，达到了 GPT-4 级别的性能表现，重新定义了大模型的经济可行性。此外，DeepSeek 推出的智能助手 DeepSeek-R1 还曾登顶美区 AppStore 免费榜，单日活跃用户突破 2200 万名，展现出强大的市场竞争力。

在生图 AI 方面，字节跳动发布的豆包 • 图像创作模型 Seedream 4.0 表现亮眼。它在文生图和图像编辑两大榜单登顶，将两个功能集成为一个单体模型，是国内 AIGC 领域的重大突破。公开资料显示，Seedream 4.0 支持 4K 超高清输出，生成 2K 图像仅需 1.8 秒，速度较前代提升 10 倍以上；同时，它能够精准理解中文指令，生成符合要求的高质量图像，非常适配电商、短剧等企业级批量生产需求。

生成视频的 AI 技术也取得了显著进展，如 OpenAI Sora 和阿里通义万相实现了从文本描述到视频内容的端到端生成，支持复杂场景下的动作逻辑呈现与光影渲染。这些技术在数字内容产业中应用广泛，可用于生成个性化广告、专属游戏视频等，能为相关产业创造数万亿的增量空间。

置身事外，还是投身其中

人工智能究竟会替代人类，还是让人类生活更美好，这是一个老生常谈的话题。事实上，人工智能的初衷并非取代人类，而是作为人类的得力助手，助力人类创造更美好的生活。

诚然，人工智能能够高效处理许多重复性、规律性的任务，如工厂中的零件分拣、银行的数据录入等，这些工作在一定程度上会被人工智能替代。但人类拥有独特的创造力、情感和同理心，在艺术创作、教育、医疗关怀等领域，人类的价值无可替代。例如，医生不仅需要诊断病情，还需给予患者情感上的支持与鼓励，这是人工智能难以做到的。

正如我们现在所感受到的，人工智能已为我们带来了诸多便利。它让疾病诊断更精准、生活更便捷、工作更高效，还能助力环境保护，提升社会的包容性。未来，人工智能将继续与人类携手共进，人类负责把控方向、发挥创意，人工智能则专注于数据处理与任务执行，共同开创更加美好的未来。

政策的支持也为人工智能发展营造了良好环境。中国《新一代人工智能发展规划》将人工智能列为核心战略，地方政府通过建设智算中心、提供人才补贴等措施，加速产业落地。2025 年 8 月印发的《关于深入实施“人工智能 +”行动的意见》更是明确提出，到 2027 年要率先实现人工智能与六大重点领域的广泛深度融合。

与此同时，市场对人工智能的需求也在急剧增长。企业端，制造业借助人工智能质检提升生产效率，金融业依靠人工智能风控降低坏账率；消费端，人工智能创作工具让普通用户也能生成高质量内容，智能硬件更是不断普及，人工智能正从“可选技术”变为“基础设施”。

此外，人工智能领域人才缺口巨大，人工智能训练师、算法工程师等职业缺口超 500 万人，高薪与良好的职业前景吸引着大量人才涌入。这不仅为从业者提供了丰富的就业机会，也为技术进一步迭代注入动力，形成行业发展的良性循环。因此，无论你是想追求职业发展，还是期望在技术浪潮中有所创新，现在都是投身人工智能领域的绝佳时机。

前途光明，马上开始

相关人才已成为市场稀缺资源，其价值水涨船高，薪酬水平更是令人瞩目。相关数据显示，2025 年人工智能领域五大核心方向的主力岗位月薪普遍突破 2 万元，部分稀缺人才的薪资甚至达到传统 IT 岗位的 2 ～ 3 倍。如果大家打开“猎聘”网，搜索“大模型训练工程师”，就能看到这个岗位的年薪可达 30 万～ 100 万元，如图 1 所示。

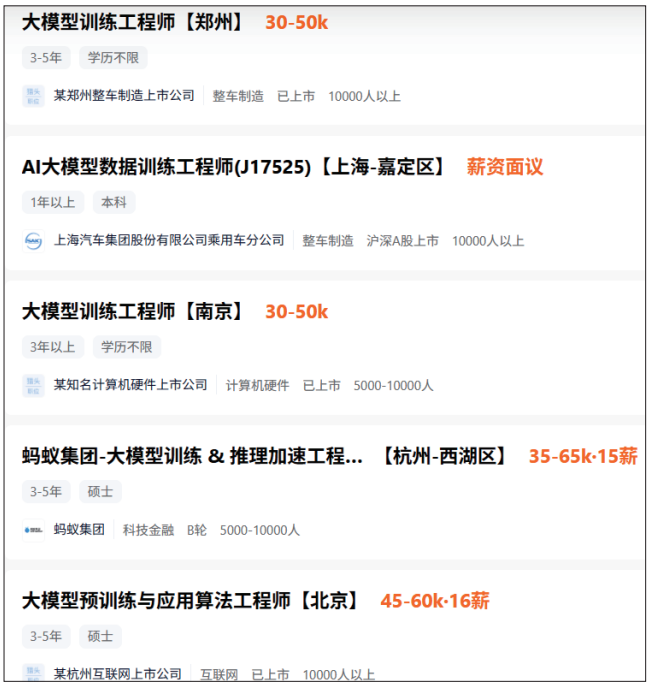

图 1　大模型训练工程师的薪酬情况

与此同时，多模态算法工程师年薪在 30 万～ 200 万元，如图 2 所示。

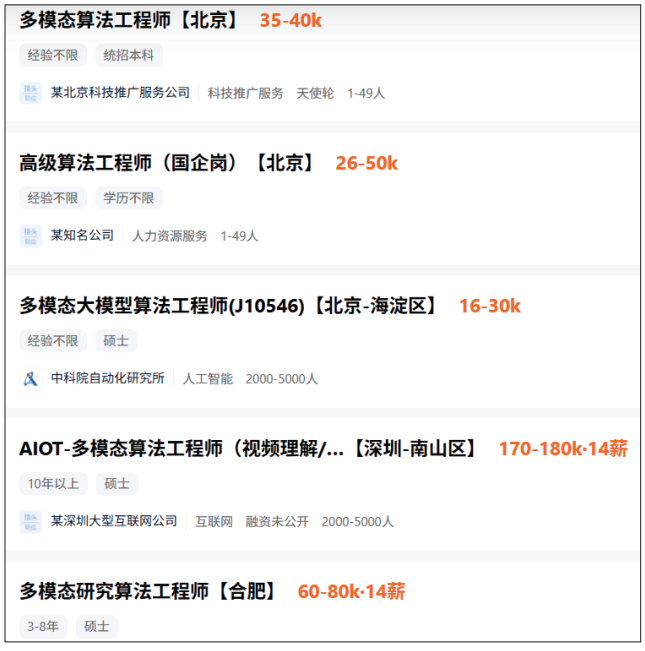

图 2　多模态算法工程师的薪酬情况

AI 芯片设计师年薪也是高达 50 万～ 100 万元，如图 3 所示。

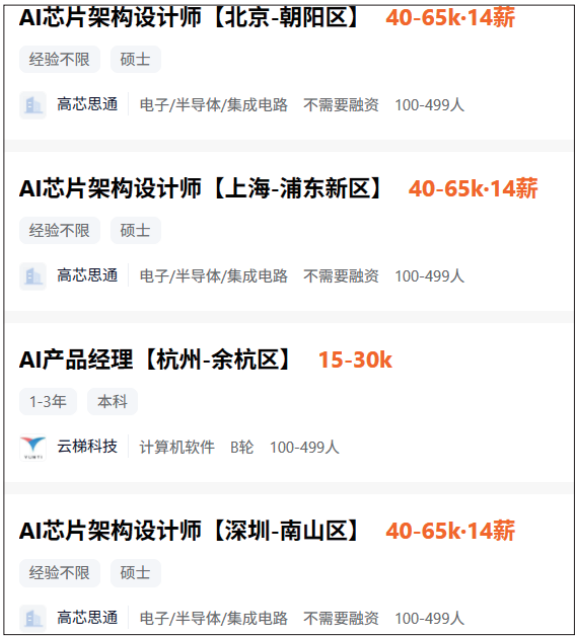

图 3　AI 芯片设计师薪酬情况

不仅如此，人工智能与生物科技、碳中和等跨学科领域融合的人才也备受青睐，年薪同样可达 40 万～ 100 万元。由于岗位众多，这里就不一一给大家截图展示了。在这样的背景下，掌握人工智能技术，无疑是握住了开启高薪职业大门的钥匙。

相信大家现在也知道，学习人工智能相关技术有着非常光明的职业发展前景。但初学者该如何开始呢？关键是不要被“技术复杂”“门槛高”的印象吓退，要“找对路径、循序渐进”，从基础能力搭建到实战落地，一步步建立信心。以下是几个清晰可落地的起步方向，帮你平稳开启人工智能学习之旅。

首先，打好“编程与数学”两大基础。在编程方面，优先掌握 Python——它是人工智能领域的“通用语言”，语法简洁且有丰富的 AI 库（如 NumPy、Pandas、Scikit-learn）。建议从基础语法（变量、循环、函数）学起，再通过“处理表格数据”（用 Pandas 做数据清洗）、“画可视化图表”（用 Matplotlib/Seaborn）等小任务练手，确保能熟练操控数据。在数学方面，不用一开始就啃高深理论，重点掌握“高中数学 + 大学基础数学”，如概率论（理解“概率分布”“期望”，能看懂模型评估指标如准确率、召回率）、线性代数（知道“向量、矩阵”是数据的基本形态，明白“矩阵乘法”在模型计算中的作用）、微积分（了解“导数”是梯度下降优化的核心），这些知识能帮你看懂模型原理，避免“只会调参、不懂逻辑”。

其次，从“机器学习”入手，而非直接冲刺深度学习。机器学习是人工智能的基础，涵盖“分类、回归、聚类”等任务，是理解人工智能“从数据中学习规律”的关键。建议先

学经典算法（如线性回归、逻辑回归、决策树、随机森林），不用死记硬背公式，而是通过“场景对应”理解用途，如用线性回归预测房价，用决策树做客户流失分类，用 K-Means 对用户做分群。可以借助 Scikit-learn 库，用几行代码实现算法，观察“调整参数对结果的影响”（如决策树的深度如何影响准确率），在实践中感受算法逻辑。这个阶段的核心是建立“数据思维”——知道“什么样的数据适合什么样的模型”“如何评估模型好不好”，为后续学习深度学习打牢地基。

最后，用“小而具体的项目”积累实战经验，避免“只学理论不练手”。入门阶段不用追求复杂项目，从“贴近生活的需求”出发即可。例如，用 Python 爬取某平台的商品数据，用机器学习分析“价格与销量的关系”；或用 OpenCV 库处理自己的照片，实现“人脸检测”“图像灰度化”；甚至可以用简单的文本数据（如自己的日记、电影评论），做“情感分析”（判断文本是正面还是负面）。这些项目所需数据少、代码量不大，却能让你完整经历“数据采集→预处理→模型训练→结果验证”的全流程。当遇到“数据格式错误”“模型过拟合”等实际问题时，通过查文档、搜社区（如 Stack Overflow、GitHub）解决，既能巩固知识，又能培养“独立解决问题的能力”——这是人工智能学习中比理论更重要的技能。

总之，人工智能入门的核心不是“一开始就追求高深”，而是“让自己每一步都能看到成果”——从写好第一行 Python 代码到跑通第一个机器学习模型，再到完成第一个小项目，每一次小进步都会积累你的信心。当你能靠自己的能力用人工智能解决一个小问题时，你就已经走在了正确的路上，后续再逐步向深度学习、大模型等方向进阶，自然会水到渠成。

本书内容及体系结构

基于上述背景，本书便与大家见面了。本书在前作的基础上，增加了生成式人工智能的内容。当然，本书也延续了前作的风格——写作力求语言生动，并且以实例来进行讲解。

本书第 1 章先用小 C 追求女神的故事阐述了什么是机器学习，然后用蝙蝠公司的业务单元展示了机器学习的应用场景，力图让读者对机器学习产生兴趣，并且参考本章给出的建议开始学习。

本章后面的内容会向读者阐述机器学习的基本概念，包括有监督学习和无监督学习、分类与回归的基本概念、模型的泛化，以及过拟合、欠拟合的具体含义。

本书第 2 章将详细地指导读者配置机器学习的环境。在本章中，读者将能够掌握 Python 的下载和安装，以及相关工具的安装和使用。

本书第 3 章到第 8 章将详细介绍机器学习中常见的一些算法，在这些章节中，读者将会掌握常见算法的原理和使用方法。

在第 9 章和第 10 章中，读者将可以学习到如何对数据进行预处理和聚类，以便让复杂数据集中的关键点一目了然。

在第 11、12、13 章中，读者将可以学习如何让算法有更良好的表现、如何找到模型的最优参数，以及如何通过建立流水线实现模型协同工作，以便达到最好的效果。

本书第 14 章介绍了如何编写一个简单的爬虫来进行数据的获取，并且介绍了如何使用潜在狄利克雷分布进行文本数据的话题提取。

从第 15 章开始，本书将带领大家了解稍进阶的深度学习知识。在这一章中，大家会理解深度学习的概念，并学习如何搭建卷积神经网络进行图像识别。

本书的第 16、17 章，分别向大家介绍了生成式人工智能的基本概念，让大家了解如何搭建一个扩散模型来生成图像，以及如何训练一个 GPT 模型来生成文本。

本书的最后，向读者简单介绍了未来的学习方向，以及如何提升自己的技能。

注意：限于篇幅，本书不会详细介绍 Python 语言的基本语法和其在机器学习之外的应用。若读者对此感兴趣，可以搜索 Python 基础教程学习。不过，本书的学习并不要求读者对 Python 达到精通的水平。

扫码可获取本书赠送的学习视频 + 源程序 +PPT 课件。

本书学习视频 +
源程序 +PPT 课件

目录 CONTENTS

第 1 章　概述

这几年，人工智能浪潮席卷全球，而大语言模型更是其中最耀眼的“明星”——从日常聊天的智能助手到自动化的内容创作，从代码生成的高效辅助到跨语言沟通的实时翻译，它们的身影无处不在，相关话题频繁登上热搜，成为街头巷尾的热议焦点。面对这股势不可挡的技术热潮，不少人既感到好奇又有些困惑：这些能说会道的模型究竟是什么？它们和人工智能之间又是什么关系？为了帮大家拨开迷雾，真正理解这一领域的核心逻辑，本章将从最基础的概念说起，带你一步步走进人工智能的世界。

本章主要涉及的知识点如下：

- 什么是人工智能；
- 人工智能的一些应用场景；
- 人工智能应该如何入门；
- 有监督学习和无监督学习的概念；
- 分类、回归、泛化、过拟合与欠拟合等概念。

1.1　什么是人工智能——从一个小故事开始

要搞清楚什么是人工智能（artificial intelligence，AI），我们可以从一个小故事开始。

小 C 是一个即将大学毕业的单身小伙子，他一直暗地里喜欢一位女生，却苦于没有机会接近她，于是在很长一段时间里，小 C 只能保持这种暗恋的状态。

直到有一天，小 C 偶然得到了这位女生的微信号并添加了她。之后他开始密切关注她的朋友圈，留意她的一举一动。

不久，小 C 就有了重大发现：女生在朋友圈经常发三种类型的内容——书籍、电影和旅游。这对小 C 来说可是个了不起的发现，机会来了。

接下来，小 C 把女生喜欢的书名和特征（features）保存在电脑上，做成一个数据集（dataset），然后根据这个数据集用“算法（algorithm）”建立了一个“模型（model）”，并通过这个模型预测出女生可能会喜欢的新书。之后，小 C 买下了模型预测的书，作为礼物送给了女生。

收到新书的女生很开心，也对小 C 产生了好感。后来，小 C 又用同样的方法预测出了女生喜欢的电影，并买票请女生去看。不出所料，每次女生的观影体验都棒极了，两个人的关系也越来

越近。再后来，小 C 又预测了女生会喜欢的旅游地点，提前订好机票和酒店，向她发出了邀请。当然，女生没有拒绝，因为这次旅行的目的地正是她一直想去的地方！整个旅途愉快极了，小 C 总能精准地聊到女生最感兴趣的话题上。女生觉得太不可思议了，她问小 C：“为什么你会这么了解我呢？”小 C 按捺住内心的喜悦，故作镇定地说道：“这是人工智能的力量。”“什么是人工智能啊？”女生不解。是时候让小 C 展现扎实的知识储备了，他笑着解释道：

“人工智能这个词，是 1956 年由约翰·麦卡锡（John McCarthy，见图 1-1）在达特茅斯会议上首次提出的，其本质上是让机器通过数据学习规律，模拟人类的判断、分析甚至决策能力。就像我这个模型，通过学习你的朋友圈数据，‘学会’了预测你的偏好——虽然这个模型很简陋，但核心逻辑和人工智能是一样的：让机器从数据中总结经验，完成原本需要人来做的思考类任务。”

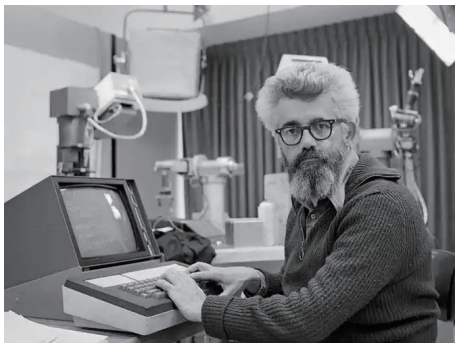

图 1-1　提出人工智能概念的约翰·麦卡锡

毫无悬念，女生对小 C 产生了深深的崇拜感，芳心暗许。从此，两个人走到了一起，过上了幸福的生活。

对于一部童话来说，故事到这里就可以结束了。但对于一本人工智能入门书来说，我们的知识讲解才刚刚开始。

有了女朋友的小 C 也开始承担起自己的责任了，他需要一份工作来为两个人的生活提供经济来源。幸运的是，他通过校园招聘进入了国内最大的互联网公司——蝙蝠公司，成为一名人工智能算法工程师，从此开始了他的职业生涯。

1.2　人工智能的一些应用场景——蝙蝠公司的业务单元

小 C 入职的蝙蝠公司，作为国内互联网行业的龙头企业，其业务覆盖面十分广泛，涵盖电子商务、社交网络、互联网金融、新闻资讯等。每一个方向在内部都被称为一个 BU（业务单元）。每个 BU 相对独立运作，拥有自己完善的体系。而人工智能技术，在每个 BU 都有非常深入的应用。下面我们来大致了解一下。

电子商务中的智能推荐

蝙蝠公司的电子商务 BU，是国内最大的在线零售平台。其用户接近 5 亿名，每天在线商品数超过 8 亿个，平均每分钟会售出 4.8 万件商品。正因为此，电子商务 BU 拥有海量的用户和商品数据。当然，为了让平台的成交总额（gross merchandise volume，GMV）不断提高，电子商务 BU 必须精确地为用户提供商品优惠信息。和小 C 预测女生的喜好类似，电子商务 BU 要通过机器学习来预测用户的行为。但在如此海量的数据下，模型要比小 C 的模型复杂得多。

例如，某个男性用户的浏览记录和购买记录中有大量的数码产品，而且系统识别出该用户访问平台时使用的设备是 iPhone 14，则算法很有可能会给该用户推荐 iPhone 17 的购买链接。而另一位女性用户浏览和购买最多的是化妆品和奢侈品，那么机器就会把最新款的爱马仕（Hermès）或者香奈儿（Chanel）产品推荐给她。

社交网络中的效果广告

蝙蝠公司旗下的社交网络平台目前有超过 9 亿名的月活跃用户（monthly active users，MAU）。其主要盈利模式是通过在社交网络中投放效果广告，为商家提供精准营销的服务。在这种盈利模式下，该 BU 需要保证广告的投放尽可能精准地到达目标受众，并转化为销售。因此，需要人工智能算法来预测用户可能感兴趣的广告内容，并且将符合要求的内容展示给用户。

例如，用户经常转发或点赞和汽车有关的信息，那么系统就会把某品牌新车上市的广告展示给用户；而如果用户经常转发或点赞时尚类信息，那么系统推荐的广告就会是新一季的服装搭配潮流等。

互联网金融中的风控系统

蝙蝠公司旗下另一个业务单元——互联网金融事业部，主要为用户提供小额贷款和投资理财服务。目前，该 BU 拥有 4.5 亿名用户，具有每秒处理近 9 万笔支付的能力，其资产损失率仅有 0.001%，这是一个非常惊人的数字。要知道，全球最老牌的在线支付工具资产损失率也要 0.2%。要达到如此低的资产损失率，必须以强大的风控系统作为支撑，而风控系统背后，就是人工智能算法的应用。

例如，在风控场景中，系统要能够收集已知的用户欺诈行为，对欺诈者的行为数据进行分析，再基于这些分析建立模型，在类似的欺诈行为再度发生之前就把它们扼杀在摇篮里，从而降低平台的资产损失比率。

新闻资讯中的内容审查

蝙蝠公司旗下的新闻资讯业务单元的表现也同样让人眼前一亮。这个 BU 的产品主要是新闻客户端 App，据称其激活用户数已经超过 6 亿名，而平均每个用户使用时长达到了 76 分钟。而令人咋舌的是，这个业务单元下据说没有编辑人员，所有的内容处理都是通过人工智能算法完成的。

例如，该产品的“精准辟谣”功能，主要就是依赖人工智能技术对内容进行识别。如果判断某条内容为虚假信息，则会提交给审核团队，待审核确认属实之后，虚假信息就会被系统屏蔽，不再向用户推送。

当然，蝙蝠公司的业务单元远远不止上述这几个，人工智能在这些业务单元中的应用也远远不止上述这几个方面。限于篇幅，本书就不再一一罗列了。读者如果感兴趣，可以进行更深入的研究。

人工智能在蝙蝠公司之外的应用

蝙蝠公司代表的是互联网行业，而在互联网行业之外，人工智能也被广泛地应用。例如，在医疗行业中所使用的专家系统，典型的案例就是诞生于 20 世纪 70 年代的 MYCIN 系统，该系统由斯坦福大学研制，它可以用患者的病史、症状和化验结果等作为原始数据，运用医疗专家的知识进行逆向推理，找出导致感染的细菌种类。若存在多种细菌，则用 0 到 1 之间的数字给出每种细菌的可能性，并在上述基础上，给出针对这些可能的细菌的药方。

此外，还有诸如智能物流、智能家居、无人驾驶，以及近几年爆火的大语言模型等领域。可以说，人工智能已经非常深切地融入我们的生活与工作之中。

一些炫酷的“黑科技”

除了上述我们提到的已经广泛应用的领域，还有一些代表着未来发展趋势的案例，如多模态

人工智能能够融合文本、图像、音频、视频等多种数据类型，实现更自然的人机交互和复杂任务执行。例如，人工智能驱动的虚拟数字人已在教育、娱乐、客服等领域广泛应用，成为人机协作的新典范；而具身智能（如人形机器人、自动驾驶等）正加速落地，推动物理世界与数字世界的深度融合。可以预见，人工智能将在更多领域发挥重要作用。

人工智能作为一个涵盖范围极广的交叉学科，包含了多个分支领域和技术方向，其中机器学习、深度学习是最核心且应用最广泛的部分。

机器学习是人工智能的核心子领域，它让机器通过“数据学习规律”，而非依赖人工编写的固定规则来完成任务。而深度学习是机器学习的“进阶版本”，其核心是模拟人脑神经元网络的结构，通过“深度神经网络”处理更复杂的数据。

随着时代的发展，当前芯片的计算能力越来越强，同时用户的数据量也越来越大，这为人工智能的进一步发展提供了必要的先决条件。在此背景下，机器学习、深度学习、神经网络等概念也随之火爆起来。在人才市场上，提示词工程师、算法工程师、数据分析师等岗位呈现出供不应求的态势，因此有更多的人开始投身到人工智能领域的研究当中。

1.3 人工智能应该如何入门——世上无难事

相信在看了上面的内容之后，一些读者朋友已经动了心，想要加入人工智能领域。说不定能像故事中的小 C 一样，既能收获爱情，又能找到心仪的工作。但与此同时，也有人会担心自己基础薄弱，不知该从何入手。

不用担心！只要你肯多动脑、勤动手，相信很快就可以入门。首先，机器学习作为人工智能的核心基石，是入门人工智能的绝佳起点。它不仅能帮助你理解人工智能“如何通过数据学习规律”的底层逻辑，还能为后续学习深度学习、大语言模型等更复杂的技术打下坚实基础。下面是给大家的一些学习建议。

1.3.1 从一种编程语言开始

如果你之前完全没有编程的基础，那么我们建议先从一门编程语言开始。目前市面上常用的编程语言有很多种，如 C++、Java、Python、R 等。那么该选择哪一种呢？不必纠结，编程语言本身没有绝对的“好”和“不好”的区别，它们各自有各自的特点。而且一旦掌握了其中的一种，后续再学习其他编程语言时，上手会快得多。

本书使用的编程语言是 Python，主要原因如下：在数据科学领域，Python 已经成为通用编程语言。它既有通用编程语言的强大能力，又具有诸如 MATLAB 或者 R 之类针对某个专门领域语言的易用性；同时，丰富和强大的库，让 Python 在数据挖掘、数据可视化、图像处理、自然语言处理等领域都有非常不俗的表现。

Python 还被称为“胶水语言”，因为它能够轻松整合其他语言编写的各类模块。此外，其简洁清晰的语法和强制缩进的特点，都让 Python 对初学者非常友好；而且它完全开源，用户完全不需要支付任何费用。

由于 Python 语言的简洁性、易读性及可扩展性，在国外用 Python 做科学计算的研究机构日

益增多，一些知名大学已经采用 Python 来教授程序设计课程。众多开源的科学计算软件包都提供了 Python 的调用接口，如著名的计算机视觉库 OpenCV、三维可视化库 VTK、医学图像处理库 ITK。谷歌（Google）的深度学习框架 TensorFlow，其兼容性最好的语言之一，便是 Python。

截至 2024 年 10 月，IEEE Spectrum 发布的编程语言排行榜显示，Python 凭借在人工智能领域的广泛应用，已连续 9 年稳居榜首。

还有一个重要的原因，即对于用户来说，Python 的学习成本是非常低的。哪怕是完全零基础的读者，在一个月左右的努力学习之后，也能大致掌握其基本语法和主要的功能模块。

因此，我们推荐读者使用 Python 进行机器学习方面的研究与开发。在后续章节中，我们会带大家配置基于 Python 的开发环境。

1.3.2 熟悉机器学习中的基本概念

在对编程语言有了基本的掌握之后，读者朋友需要熟悉机器学习中的一些基本概念。例如，什么是“有监督学习”，什么是“无监督学习”，它们之间的区别是什么，在应用方面有什么不同。此外，要对“分类”和“回归”有基本的认知，清楚在什么场景下使用分类算法，在什么场景下使用回归算法。最后，要理解模型的“泛化”，明白在什么情况下模型会出现“过拟合”现象，在什么情况下会出现“欠拟合”现象。

1.3.3 了解机器学习中最常见的算法

在了解了基本概念之后，读者朋友就可以开始了解机器学习中最常用的一些算法了。例如，*K* 最近邻算法、线性模型、朴素贝叶斯、决策树、随机森林、SVM、神经网络等。

在这个过程中，读者需要了解每种算法的基本原理和用途，清楚它们的特性分别是什么，在不同的数据集中表现如何，以及如何使用它们进行建模、模型的参数如何调整，等等。

1.3.4 掌握对数据进行处理的技巧

经过上面的几个步骤，读者朋友将能够对小数据集进行建模并且做出一些预测。但是在真实世界中，数据往往比我们拿来实验的小数据集复杂很多倍。它们的特征变量会大很多，也就是说数据的维度会高很多，同时可能完全没有训练数据集供你使用，这时候读者就需要掌握一些数据处理的技能，如如何对数据进行降维或聚类，从而让数据更容易被理解，并从中提炼关键信息，为建模奠定基础。

1.3.5 学会让模型更好地工作

在学会用算法建模和对数据进行处理后，读者朋友要做的是如何让模型更好地工作。例如，如何提升算法运行效率，如何找到最适合的模型，如何确定模型的最优参数，以及如何搭建模型流水线，让多个模型协同工作以解决复杂问题等。

1.3.6 动手，一定要动手操作

学习一门知识最好的办法是实践，因此建议读者朋友一定要自己动手实操。不要嫌麻烦，尽可能把本书中的代码全部自己敲一下，只有这样，才能对内容有更加深刻的理解。如果觉得基础

练习不够，还可以登录知名的数据科学竞赛平台 Kaggle，使用那些来自真实世界的数据来提升自己的技能。

当然，还有更好的方法，那就是去企业中寻找一个机器学习工程师或是算法工程师的岗位，在工作中学习，效果是最好的。

1.4 有监督学习与无监督学习

在机器学习领域，有监督学习和无监督学习是两种常用的方法。有监督学习（又称监督式学习）是通过现有训练数据集构建模型，再用模型对新的数据样本进行分类或者回归分析的机器学习方法。在监督式学习中，训练数据集一般包含样本特征变量及分类标签，机器运用不同的算法，从这些数据中学习并推断出分类的方法，并将其用于新的样本中。目前，有监督学习算法已经比较成熟，并且在很多领域都有很好的表现。

而无监督学习（又称非监督式学习）则是在没有训练数据集的情况下，对没有标签的数据进行分析并构建合适的模型，以提供问题解决方案的学习方法。在无监督学习当中，常见的两种任务类型是数据转换和聚类分析。

其中，数据转换的目的是通过非监督式学习算法，将本来非常复杂的数据集进行转换，使其更容易理解。常见的数据转换方法之一便是数据降维，即通过对特征变量较多的数据集进行分析，将无关紧要的特征变量去除，保留关键特征变量（如把数据集降至二维，方便进行数据可视化处理）。

而聚类算法是一种把样本划归到不同分组的算法，每个分组中的元素都具有比较接近的特征。目前，聚类算法主要应用在统计数据分析、图像分析、计算机视觉等领域。

1.5 机器学习中的分类与回归

分类和回归是有监督学习中两种最常见的任务。对于分类任务来说，机器学习的目标是对样本的类标签进行预测，判断样本属于哪一个分类，其结果是离散的数值。而回归任务的目标是要预测一个连续的数值或数值范围。

这样讲可能会有一点抽象，我们还是用小 C 的例子来理解这两个概念。

例如，小 C 在使用算法模型预测女生的电影喜好时，他可以将电影分为“女生喜欢的”和“女生不喜欢的”两种类型，这就是二元分类。如果他要把电影分为“女生特别喜欢的”“女生有点喜欢的”“女生不怎么喜欢的”及“女生讨厌的”四种类型，那么这就属于多元分类。

但如果小 C 要使用算法模型预测女生对某部电影的评分，如女生会给《速度与激情 8》打多少分，从 0 分到 100 分，分数越高说明女生越喜欢，最终模型预测女生会给这部电影打 88 分，这个过程就叫作回归。小 C 需要将女生给其他电影的评分和相对应的电影特征作为训练数据集，通过建立回归模型，来给《速度与激情 8》打分。

1.6 模型的泛化、过拟合与欠拟合

在有监督学习中，我们会在训练数据集上建立一个模型，之后会把这个模型用于新的、从未见过的数据中，这个过程称为模型的泛化（generalization）。当然，我们希望模型对于新数据的预测能够尽可能准确，这样才能说明模型的泛化准确度比较高。

那么我们用什么样的标准来判断一个模型的泛化能力是好还是差呢？

我们可以使用测试数据集对模型的表现进行评估。例如，你在训练数据集上使用了一个非常复杂的模型，以至于这个模型在拟合训练数据集时表现非常好，但是在测试数据集上表现非常差，这就说明模型出现了过拟合（overfitting）问题。

相反，如果模型过于简单，连训练数据集的特点都不能完全考虑到的话，那么这样的模型会在训练数据集和测试数据集上表现都非常差，这个时候我们说模型出现了欠拟合（underfitting）问题。

而只有当模型在训练数据集和测试数据集上的表现都比较好的情况下，我们才会认为模型对数据拟合的程度适中，同时泛化的表现也会更出色。

1.7 本章小结

现在我们来对本章内容进行总结。本章开篇通过一个小故事，帮助读者了解了人工智能的基本概念，随后对人工智能的部分应用场景展开了初步学习。

之所以说“部分”应用场景，是因为人工智能的应用范围实在太广，我们很难逐一穷举，相信读者朋友日后还会接触到更多元化的案例。

读者朋友对人工智能产生兴趣之后，还可以在本章中找到 AI 入门的步骤及相关建议。当然，每个人都有自己独特的学习方法，本章所列内容仅为参考，大家完全可以根据自身情况制订学习计划。

此外，为帮助读者朋友更加顺利地学习后续章节，本章还初步介绍了一些机器学习领域的术语，如有监督学习、无监督学习、过拟合和欠拟合等。不过，请读者朋友注意，这部分内容也并未涵盖所有术语，如半监督学习和强化学习的概念，本章尚未涉及。相信日后随着读者朋友学习进程的加深，还会解锁更多新的知识点。