普通高中教科书

信息技术

必修 1 数据与计算

华东师范大学出版社

普通高中教科书

信息技术

必修 1

数据与计算

华东师范大学出版社
·上海·

总 主 编：李晓明

副总主编：赵　健

本册主编：冯　忻

编写人员(按姓氏笔画排序)：

毛黎莉　冯　忻　张逸中　欧阳元新　周永麒

责任编辑：程　滨

美术设计：储　平

普通高中教科书　信息技术　必修 1　数据与计算

上海市中小学(幼儿园)课程改革委员会组织编写

出版发行　华东师范大学出版社(上海市中山北路 3663 号)

印　　刷　上海昌鑫龙印务有限公司

版　　次　2020 年 6 月第 1 版

印　　次　2020 年 6 月第 1 次

开　　本　890 毫米×1240 毫米　1/16

印　　张　9.25

字　　数　166 千字

书　　号　ISBN 978-7-5760-0547-9

定　　价　11.60 元

本册教材图片提供信息：

本册教材中的部分图片由全景网、视觉中国等图片网站提供。

致同学们

亲爱的同学们：

当今，信息技术的发展日新月异，物联网、大数据、人工智能等新技术、新工具扑面而来，显著地改变着人们的生活、学习和工作模式。生存于信息社会中，我们每一个人都不可避免地会接触信息技术、应用信息技术，甚至去创造新的信息技术。在具备了基本信息技术应用能力的基础上，高中阶段我们要进一步学习信息技术的知识与技能，能够利用信息技术负责任地解决生活与学习中的问题，全面提升信息素养，迎接信息社会的挑战。

“数据与计算”作为高中信息技术学科的必修模块，是学习高中信息技术学科其他模块的基础。本教科书采用“项目活动”方式组织学习内容，通过“信息技术伴我学”“编程应用助健康”“交通数据利抉择”“智能工具好帮手”项目，将数据与大数据、算法与编程实现、数据处理与应用、人工智能等基础知识与技能融入到学习活动中。教科书的每章围绕“信息意识”“计算思维”“数字化学习与创新”“信息社会责任”四个学科核心素养提出本章的学习目标，利用“本章知识结构”图示呈现本章知识脉络，帮助同学们从总体上了解本章学习内容。

在学习过程中，同学们可以通过“体验思考”栏目，将现实问题、个人经验与知识技能相关联，带着问题开始学习；通过“探究活动”和“项目实践”栏目，将“做中学”与“学中做”的学习方法相互融合，把知识技能应用于解决实际问题中；通过“技术支持”栏目，将新技术与新工具适时应用于作品制作中，提高合理选用技术工具创造性完成作品制作的能力；按照个人的学习需求，学习“知识延伸”栏目中的内容，拓展个人学习视野。

提升信息素养，要求我们在掌握基本信息技术知识和常用信息技术工具的同时，能够用计算思维来分析问题；要求我们在体验信息技

术给我们带来的更高效率的同时，积极运用技术来创造性地解决问题和创作作品；要求我们在享受信息技术提供的便利的同时，关注信息安全，参与和促进信息社会的伦理与道德建设。同学们可以通过本教科书与配套资源学习信息技术，负责任地应用信息技术，逐步成长为新时代合格的社会主义建设者。

编　者

目　录

第一章

数据与大数据

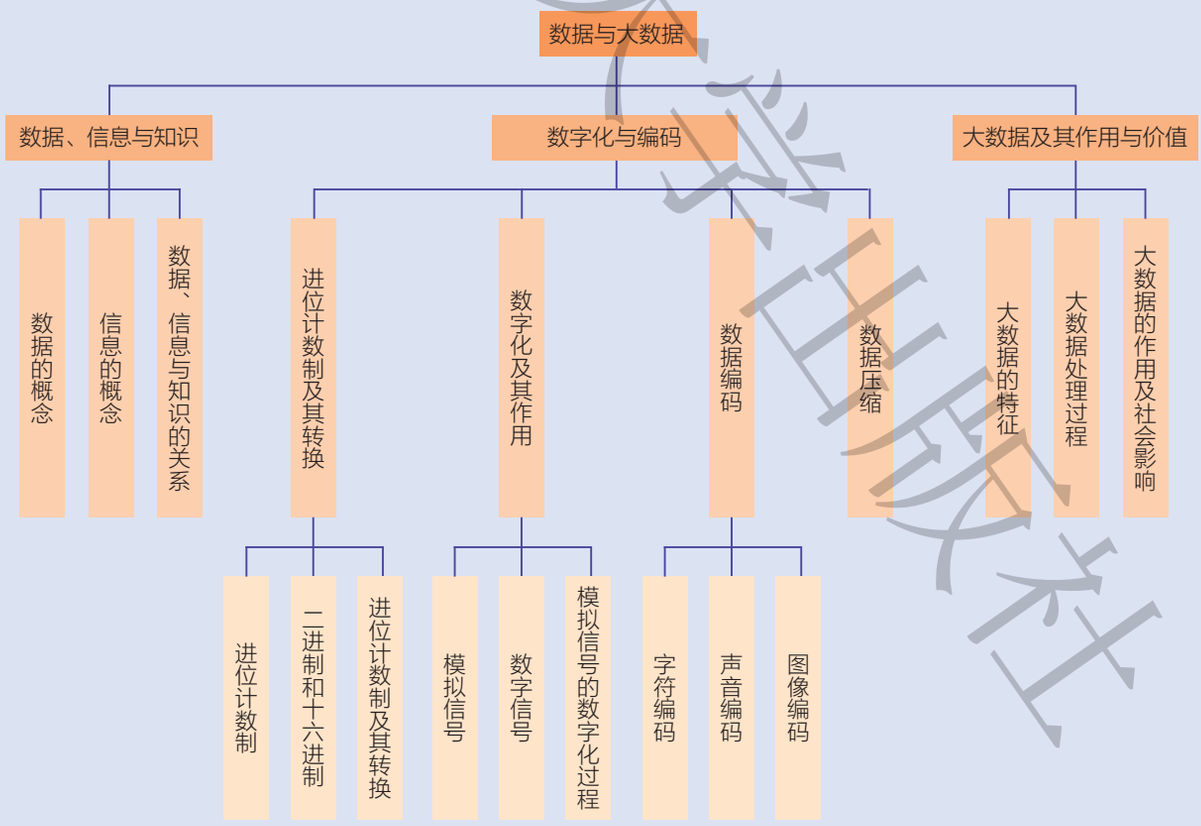
本章学习目标

- 在实际生活与学习中感知数据与信息，知道数据与信息的特征，理解数据、信息与知识的区别和联系，认识数据与信息对社会发展和个人成长的影响。
- 掌握二进制数与十进制数、二进制数与十六进制数相互转换的方法，了解数字化的过程与意义，知道字符、声音、图像编码的基本方式。
- 针对学习任务，选择数字化学习工具和资源，感受利用它们进行自主学习和知识分享的优势。

信息技术的发展与普及改变着我们的生活与学习。我们的晨起，可能始于查看手腕上带有睡眠监测功能的智能手环；我们的休闲，可能始于点击手机中具有歌曲智能推荐功能的音乐软件；我们的出行，可能始于扫描小区外停放的共享单车的二维码；我们的学习，也可能不再是从步入教室的一刻开始，而是始于打开一款能够时时、处处伴随着我们的在线学习软件。

目前，越来越多的高中生使用慕课学习平台、数字图书馆、数字实验系统开展学习。登录慕课学习平台后，可以实现跨校学习，共享优质学习资源；访问数字图书馆，能够快速查阅信息，自助完成电子图书的借阅和归还；应用数字实验系统，能够体验实验数据的生成过程，以可视化方式感知实验规律。信息技术在改变着我们的学习环境，也改变着我们的学习方式。掌握信息技术知识，运用信息技术促进学习，是新时代每位中学生都应具备的能力。

本章知识结构



项目主题

信息技术伴我学

项·目·情·境

随着技术的发展，人们阅读的书籍已经从纸质图书拓展为既能看也能听的电子图书，人们传统的阅读方式也随之发生变化。

一年一度的学校诗词大会即将举行，学校图书馆收到了一些读者需求：有的同学希望图书馆能够增购一些电子图书，供同学们借阅；有的同学希望图书馆能够将馆藏的纸质校刊制作成电子校刊，方便查阅往年学校诗词大会的征文；有的同学希望在面对浩瀚书海时，学校图书馆能够根据同学们的阅读习惯，提供个性化的图书推荐。

小申是学校图书馆的志愿者，他能为学校图书馆提供一些建议吗？

项·目·任·务

任务 1

搜索电子图书网站，记录电子图书选择过程中的参考数据，了解这些数据所反映的信息，描述数据在选择电子图书过程中所起到的作用。

任务 2

将学校馆藏的纸质校刊制作成电子校刊，感受数字化的过程，小组合作完成电子校刊的制作。

任务 3

分析电子图书网站向读者推送电子图书的方法与策略，举例说明大数据在其中的作用，为学校图书馆设计一份电子图书推荐方案。

第一节　数据、信息与知识

人类对于数据的应用由来已久，早在春秋战国时期，齐国国相管仲就通过对农业生产数据的统计分析来制定相关的农业生产政策，在《汉书·地理志》《史记·平准书》等众多史籍中都留下了有关农业生产、天文历法、地理山川的大量数据，这些数据的应用一定程度上提高了人类的生产效率。如今，信息技术的发展赋予了人们采集和分析数据的新工具与新方法，通过这些工具和方法，人们可以更高效地处理数据，解决问题。

体　验　思　考

随着学校诗词大会举办日期的临近，图书馆老师希望小申能够根据同学们的需求，列出与诗词相关的电子图书购买清单。小申访问了一些电子图书网站，通过查找和比较，列出了电子图书的购买清单。

思考：

1. 如图 1.1 所示，网站上提供了哪些数据来帮助人们选择要购买的图书？
2. 这些数据对人们选择图书有什么帮助？

图 1.1　在电子图书网站上搜索电子图书

一、感知数据

数据无处不在。数字图书馆中，人们输入的账号、密码，读者对于图书的评论等都是数据；公交站台电子屏幕上显示的行车线路号、预计到达时间等也是数据；天气预报播报的气温、湿度、风级等同样是数据。数据已广泛应用于我们的生活与学习。

1. 数据的概念

数据是对事物描述的记录。例如，描述一个学生的基本特征，可以通过姓名、性别、年龄等方面的数据来记录；确定某一地理位置，可以通过经度和纬度的数据来记录；表示城市空气质量检测中细颗粒物（$PM_{2.5}$）随时间变化的情况，可以用一个时间序列数据来记录。数据可以帮助人们有效地描述事物。

数据的表现形式多种多样，可以有数字、文字、图形、图像、声音等形式。对同一事物的描述记录也可以有不同的数据表现形式，例如，导航仪行车线路中表示车辆左转时，可以用文字“左转”来表示，也可以用图形  来表示，还可以通过语音来播报。

同一数据也可能描述不同的事物。例如，数字“60”可以表示一个人的年龄、一次考试的成绩、一件物品的长度，或者是某个路段的机动车的最高限速值等。因此，脱离具体的情境和形式，无法确定数据的意义。

数据是可加工、可处理的。从已知数据出发，参照相关数据进行加工计算，生成一些新的数据，从中可以得到新的结论，从而作为人们决策的依据。例如，在线学习网站会记录学习者的访问数据，通过学习者浏览某一页面的起始时间和结束时间，计算得到这一页面的学习时长，并将该学习时长和系统设定的有效学习时长进行比较，从而判断学习者的该次学习是否有效。

在人类文明的历史长河中，人们发明了很多处理数据的工具，从古人发明的算盘到故宫馆藏的计算尺，从十六世纪帕斯卡发明的加法器到今天功能强大的计算机，人们处理数据的能力越来越强大，数据的含义也越来越丰富。在计算机科学中，数据是计算机识别、存储和加工的对象。例如，我们常用的演示文稿文件、电子表格文件、图像文件、音频文件和视频文件等都是计算机处理的数据，如图 1.2 所示。

图 1.2　用计算机处理的数据

2. 数据的价值

在信息社会，随着数据处理技术的迅速发展，数据被广泛地应用于社会的方方面面，给人们的学习、生活与工作带来了巨大的变化。

在数字图书馆中，图书管理人员利用采集到的借阅数据，调整管理方式，提供个性化服务；读者借助网络平台中的图书数据，足不出户就可以有针对性地选择和借阅图书，享受读书的乐趣；图书作者还可以根据读者对图书的阅读和评价数据，进一步完善图书内容。

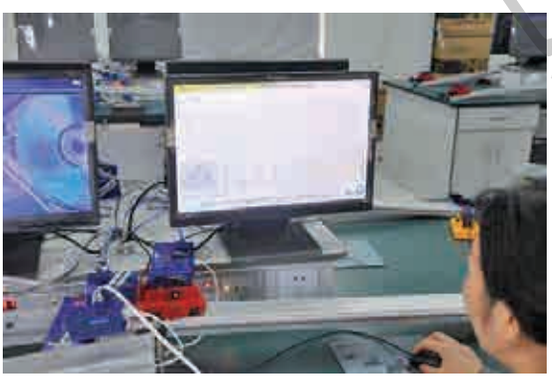


图 1.3　数字化实验系统

在学校餐饮管理中，通过食堂管理系统，可以快速获取和分析学生的用餐数据，根据不同菜品的销售数据，食堂管理员可以适时地调整菜品种类，合理安排每种菜品的数量，提高服务质量。

在教学实验中，通过数字化实验系统，可以采集需要测量的物理量，如温度、电压、压强等，将其转换成计算机可以处理的数据。计算机处理后，能够直观地呈现实验结果，提高学生的探究能力，如图 1.3 所示。

图 1.4　道路检测车

在道路检修过程中，道路检测车可以自动采集路面的损坏状况、道路平稳度等各项数据，如图 1.4 所示。通过数据分析，车载计算机可以判断道路的安全情况，甚至还可以估算出维修费用，避免了由于人工目测而导致的误差，为道路养护提供准确、有效的数据支撑。

随着信息技术与人们生产生活的交汇融合以及互联网的快速普及，全球数据呈现出爆发式增长、海量集聚的特点。数据对改善人民生活、促进经济发展、推动社会进步等，起着越来越重要的作用，它已成为像水、电、煤气一样重要的资源。数据作为一种资源，同样需要通过各种各样的“管道”输送到社会的各个领域中去，将数据转化为用户决策或行动的依据，促进社会的发展。

二、认识信息

自古以来，人类的生存和发展就与信息有着不解之缘。我国古代利用烽火台传递示警信息，通过活字印刷术促进知识和文化的广泛传播。今天，移动通信设备和网络成为我们获取信息的重要途径。信息的获取与应用影响着我们分析问题与解决问题的方式。

项 目 实 践

请尝试读取图 1.5中的数据，填写表 1.1，并回答问题。

图 1.5　电子图书搜索页面

表 1.1　电子图书信息分析表

	《唐诗百话》	《北山楼词话》
作者		
出版时间		
字数		
在读人数		
评分		

对于上述两本电子图书，如果希望购买其中一本出版日期较新的，应该选择_____，选择的依据是_____；如果希望购买其中一本较为热门的，应该选择_____，选择的依据是_____。

1. 信息的概念

生活中，人们总是自觉或不自觉地获取和应用信息。在教室听到上课铃声，学生保持安静，开始上课。在交通路口，行人根据交通信号灯的变化决定“行”还是“停”。在数字图书馆，读者通过比较电子图书数据，了解图书信息，确定需要选择的图书。人们时时刻刻接触信息、应用信息。对于什么是信息，人们从不同的角度给出了不同的定义。

唐代诗人岑参写道：“马上相逢无纸笔，凭君传语报平安。”南宋诗人杨万里则写下了“花落六回疏信息，月明千里两相思”。这些都表达出古人对传送和得知信息的渴望。在通信并不发达的古代，信息的意思更多地还是指消息。

信息学奠基人克劳德·艾尔伍德·香农（Claude Elwood Shannon）认为“信息是能够用来消除不确定性的东西”。对某一不确定的情况，当获得信息之后，这种“不确定性”就可以减少或消除。香农从信息的产生、传播、接收等通信系统角度来考虑信息的涵义，这也推动了信息学的发展，而这一定义也常被人们看作经典定义而加以引用。

信息管理专家 F. W. 霍顿（F. W. Horton）将信息定义为：“信息是

为了满足用户决策的需要而经过加工处理的数据。”简单地说，信息是经过加工的数据，或者说，信息是数据处理的结果。

综合上述信息的定义可以看出，信息表示的是事物之间的相互关系，它可以通过数字、字符、图像、声音和视频等载体进行传播。人们借助信息可以了解情况、形成判断、做出决策、指导行动。在信息社会里，有效获取和合理应用信息已成为人们需要具备的一项重要的信息素养。

2. 信息的特征

信息在人际交流、生产管理、知识传播和科学研究等方面都发挥着巨大的作用。了解信息的特征有助于我们加深对信息的认识和理解。

（1）信息可以传播和存储

信息的传播和存储需要依附于一定的载体。承载信息的数字、字符、图像、声音和视频等可称为信息的载体。在信息处理中，如果存储信息的载体遭到破坏，其承载的信息就会丢失。例如，古书中文字的缺失导致了它所传达信息的丢失。通信信号受到强烈干扰，也会破坏其所传递的信息。

（2）信息的价值是相对的

同一条信息对于不同的持有者具有不同的价值，例如，有一条信息“我国发现了兵马俑”，新闻记者需要将这条信息传播给更多的人，历史学家需要这条信息帮助其更好地进行历史研究等。同样，信息的价值也取决于信息接收者对信息的理解、认知和应用能力。人们在信息的应用过程中，经过对原有信息的加工后可能会产生新的信息，进而产生新的价值，从而使原来的信息增值。

（3）信息可以被共享

人们可以将一条信息传播出去，让其他人也能接收并反复利用，如图 1.6 所示。英国著名戏剧家萧伯纳曾经说过：“如果你有一个苹果，我有一个苹果，彼此交换，我们每个人仍然只有一个苹果；如果你有一种思想，我有一种思想，彼此交换，我们每个人就有了两种思想，甚至多于两种思想。”这在一定程度上体现了信息是可以被共享的。

图 1.6　信息可以被共享

（4） 信息具有时效性

信息往往反映了事物在某个特定时间的状态，信息的时效会随着时间的推移而变化。例如，用户可以通过使用手机扫描二维码登录一些网站或邮箱，提供给用户扫描的二维码每隔一定时间便会刷新，重新生成，它所传递的信息只在一定时间内有效。在信息社会中，信息的变化越来越快，信息价值的实现取决于对其及时的把握和运用。如果不能及时利用最新信息，信息的价值就可能会贬值甚至会变得毫无价值，这就是信息的时效性。

3. 合理应用信息

图 1.7 航班实时数据示例

社会信息总量的快速增长为人们应用信息解决问题带来了便利条件。但是，过于庞杂的信息量，各种各样的干扰信息，持续更新的技术工具，都对人们感知与获取信息、甄别信息的真伪和合理应用信息带来了挑战。

（1） 敏锐感知周围世界，正确获取信息

在信息社会中，信息的变化日益频繁，生存于其中的社会成员要能敏锐地感知到变化的信息，依据信息的变化做出相应决策。例如，旅行社利用移动应用软件（App）发布的实时航班数据，了解航班运营时间等信息，调整接送旅客的出车计划，提高工作效率，如图 1.7 所示。事实上，缺少对信息变化的敏锐感知，有可能造成不必要的损失。例如，当城市地铁出现故障时，交通管理部门通过电台、微信公众号等途径发布故障信息，如果用户不能及时获取这些信息，依然采用坐地铁的方式出行，就会影响出行计划。

信息技术为我们感知和获取信息提供了便利的条件。官方微信公众号、官方微博实时推送和发布的信息可以帮助我们做出判断；借助移动应用软件中的数据，可以了解事物变化的情况，做出相应的行动调整。合理利用信息技术获取数据，应用其中的信息指导所要采取的行动，可以让我们自信、从容地生活在信息社会中。

（2） 具备信息辨别能力，有效甄别信息

日常生活与学习中，大家可能收到过一些虚假信息，例如中奖短信、诈骗电话、虚假照片合成、微信朋友圈和公众号中骇人听闻的假新闻等等。这些信息会影响我们的正常生活，甚至会给我们带来不必要的损失。在纷繁复杂的信息环境下，人们需要具备有效甄别信息、判断信息真伪的能力。

图 1.8　移动应用安全软件

生活中，人们可以通过多种方式和渠道来辨别信息的真伪，例如通过主流媒体对所获得的信息进行核对，与所获信息的相关人员进行实时沟通确认，或者借助技术工具对所获信息进行分析辨别。人们在手机中安装移动应用安全软件来识别和标记诈骗电话等(如图 1.8 所示)，就是一种常见的辨别信息真伪的防护方式。

(3) 遵守信息安全法规，负责任地使用信息

信息技术拓展了人们的生存时空，创造出人们新的生存环境。在新的环境中，人们也要遵守其中的新秩序。为维护信息社会的秩序，我国先后出台了一系列旨在推动信息化建设的法律法规，这就要求每位社会成员都要担负起相应的责任。2017 年 6 月，我国正式施行《中华人民共和国网络安全法》。其中，第十二条要求“任何个人和组织不得利用网络从事编造、传播虚假信息扰乱经济秩序和社会秩序，以及侵害他人名誉、隐私、知识产权和其他合法权益等活动”。

网络是继陆地、海洋、天空、太空之外，又一个人类活动空间。人们在网络空间中的各项活动要遵守信息社会的法律法规，如有违反法律法规的行为，必将会受到法律法规的惩罚。例如，不法分子通过网络散布病毒程序，用以盗取他人手机通信录、短信、银行卡账号等信息，危害社会信息安全。经公安部门查实、认定其违法行为后，根据相应法律法规，对其进行了相应处罚。

三、学习知识

“知识就是力量”这句名言一直流传至今，知识可以推动社会的进步。在人类文明的历史长河中，人们可以从种类繁多的资料记录中获取知识，也可以在生活实践中通过分析数据和信息来发现知识。如今，信息技术的发展为我们学习知识创造了新的条件。结合具体问题情境，应用这些知识可以帮助我们解决问题。

探 究 活 动

在电子图书网站上，小申发现许多位于热门榜单上的电子图书都拥有大量的读者评论。小申想通过这些评论，了解某一电子图书的读者最突出的阅读感受。老师推荐他借助语义分析工具，对电子图书的相关读者评论进行研究。使用语义分析工具分析《唐诗百话》的读者评论，结果如图 1.9所示。

1 根据图 1.9，分析读者对于《唐诗百话》最突出的阅读感受是什么。

2 尝试利用语义分析工具，分析某一电子图书的读者评论，并思考在此过程中，信息技术工具起到了怎样的作用。

图 1.9 《唐诗百话》读者评论词频统计图

在前面的活动中，我们尝试使用语义分析工具来辅助分析读者对电子图书的评论。在实际应用中，数字图书馆可以对读者的图书评论加以分析，并以此为依据向读者提供阅读推荐。信息技术为人们处理数据提供了强有力的工具，人们可以利用信息技术对数据进行分析，找出其中的相互关系，形成规律，获得知识并加以运用。

1. 数据、信息与知识的关系

数据是描述事物的记录，它能够承载信息，因此人们可以在处理数据的过程中获得信息。随着人类的进步以及处理数据和信息的能力不断提升，人类从数据中获取有用信息的能力越来越强。

信息表示的是事物间的相互关系，通过分析数据可以发现其中包含的关系。例如，分析某电子图书借阅人数和读者评价，可以发现该图书基本内容和写作特点，为从同类图书中选择合适的图书提供依据。

知识分为一般知识和科学知识。在日常生活和工作中，人们所获得的认识和经验的总和，通常称为一般知识。科学知识是人们对信息的科学组织，它是经过严谨的验证，获得学界一致认可的内容，如物理学科中的牛顿三大定律、化学学科中的元素周期律和生物学科中的遗传基本定律等。随着认知工具和方法的发展，人们对世界的认识也会不断深入，人类的知识也在不断地发展。

今天，人们运用各种信息技术工具来认识事物、表达思想、分享知识，让学习和工作更加高效。例如，通过数字实验设备，可以便捷地采

集数据、获取信息，从而发现新知识；借助网络平台加快信息的传播速度，可以快速分享知识。合理地应用信息技术，人们就能更好地认识世界、发现知识，推动人类文明的进步。

2. 体验数字化学习

今天，学习者处于全新的数字化学习环境中，需要不断提升个人信息素养，选择合适的学习资源和学习方式开展学习。网络的发展拓展了学习时空，学习者足不出户就可以获得优质的学习资源。例如，不同地域的学校可以通过网络进行跨校研讨与交流，推动远程合作学习，实现优质教育资源的效益最大化，如图 1.10 所示。

图 1.10　互联网环境下的远程学习

虚拟现实技术的应用可以模拟真实情境，帮助学习者开展探究学习。例如，在学习海洋生物的相关知识时，学习者很难到深海中去体验深海生物的生存环境。但是，在虚拟现实技术的支持下，学习者可以身临其境地感受深海生物的生存环境，更好地学习、理解相关知识，如图 1.11 所示。

图 1.11　虚拟现实技术支持深海生物知识的学习

大数据与人工智能技术在教育中的应用可以记录学习者的学习行为数据，针对学习者的学习需求，依据数据分析结果提供精准学习支持。目前，一些在线学习平台应用大数据和人工智能技术分析学生的学习过程，实时采集学生的学习数据，依据数据发现学生学习的不足，针对学习中存在的问题提供相应的学习资源与指导，让学生感觉到“教师”时时刻刻都在自己身边。

技　术　支　持　　**语义分析工具**

借助语义分析工具，我们可以用量化的方式分析文本内容，获取文本所表达的深层次信息。例如，通过语义分析工具，可以快速分析读者对图书的评价，了解读者对图书的关注要点等。

如图 1.12所示，在应用语义分析工具时，将要分析的文本复制到分析工具中。通过计算，系统可以完成文本内容的实体抽取，识别文本中出现的人名、地名等关键词；实现词频统计，对不同词性的词语进行分类呈现，分析每类词语出现的频度；绘制文本的词云，突出文本中出现频率较高的关键词；判断相关词语，分析语义关联情况，对文本内容进行精简提炼，从长篇文章中提取关键句和关键段落，编辑文本的摘要，等等。

图 1.12　应用语义分析工具分析文本

作业练习

小申与学习同伴在数字化信息系统（DIS）实验室做“测定位移和速度”实验。实验过程中，他们通过 DIS 中的位移传感器采集实验小车的位移数据，然后将采集的数据传入计算机中，利用实验软件系统描绘出 $s-t$ 图像。通过 $s-t$ 图像，可以确定实验小车从起始时刻到任意时刻的位移和任意一段时间内的平均速度。实验环境与实验数据如图 1.13所示。

图 1.13　“测定位移和速度”的 DIS 环境与实验数据

1 分析“测定位移和速度”的实验环境与实验数据，填写表 1.2。

表 1.2 “测定位移和速度”的实验环境与实验数据分析

分析内容	描　　述
运用 DIS 采集了哪些实验数据	
从实验数据中可以发现哪些信息	
通过实验可总结出哪些物理知识	
列举实验中所使用的数字化工具	

2 列举你在日常学习中所使用的数字化学习工具，描述它们的功能，分析它们在学习中的优势和局限性。

知 识 延 伸　香农与信息论

1948年，香农在定义信息时，借用了物理学中的“熵”一词，解决了对“信息”一词的量化和度量问题。从此，“信息”一词就有了一次数学化的提炼，信息就有了定量计算的单位，这是一个划时代的进步，推动了通信技术的发展，也推动了整个信息技术的发展。信息论在信息与不确定性、信息与信息熵之间架起了桥梁。

香农认为“信息是能够用来消除不确定性的东西”，他用概率分布来衡量信息的“不确定性”，同时引入了“比特”作为计量单位。“信息熵”一词是从“热力熵”而来，即求整个系统事件中平均信息量的大小，香农给出了信息熵的数学公式：

$$H(x)=-\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)\ (i=1, 2, \cdots, n)$$

其中 $p(x_i)$：某一随机事件的概率；$\log_2 p(x_i)$：某一随机事件的信息量

这些概念奠定了信息论的基础，并且为信息技术领域的进步开辟了新的道路。

第二节　数字化与编码

信息技术的发展创造出一个全新的数字化环境，生活在其中的每个人都能感受到数字化带来的变化。人们利用数字化设备可实时获取自己的心率、血压等身体健康数据，通过分析这些数据，可以主动管理自己的健康；乘客可以通过移动智能终端查询车辆到站的实时信息，避免了以往久等公交车而不知车何时到达的尴尬。移动通信、移动智能终端等新技术的广泛使用，使全球正在成为一个互联互通的数字化世界。

体　验　思　考

同学们希望查阅学校图书馆馆藏校刊上刊载的往年诗词大会征文。由于馆藏的校刊数量较少，因此图书馆只能满足少数同学的借阅需求。同时，同学们在借阅校刊的过程中，也令校刊产生了不同程度的污损，影响了校刊的收藏。因此，学校图书馆希望能够将历年的纸质校刊制作成电子校刊，供同学们借阅。

思考：

1. 纸质校刊的内容承载于墨迹和纸张之中，那么电子校刊的内容是以怎样的形式存储在计算机中的？

2. 分析纸质校刊和电子校刊在借阅过程中各自的优势和不足。

一、进位计数制及其转换

目前，计算机的硬件组成通常可以呈现两种状态，如电路的导通和断开。这样就决定了计算机内部采用二进制，即以“0”和“1”的组合来表示信息，用“1”来表示一种状态（如电路的导通），用“0”来表示相反的另一种状态（如电路的断开）。由于计算机采用二进制数进行运算和存储，因此要使用计算机进行信息处理，首先要把待处理的信息用二进制数来表示。

1. 进位计数制

进位计数制，是按进位方式实现计数的一种规则。进位计数制包含数码、基数和位权三个要素。我们将用来表示某种进位计数制的一

组符号称为数码，所使用的数码个数称为基数，数码在不同数位上的倍率值称为位权。

十进制是人们生活中常用的进位计数制，它的基数为 10，由 0，1，2，…，9 共 10 个数码组成，整数位的位权从右向左依次为 10^0，10^1，10^2，…。例如，十进制数 463 各个数位上的数字所代表的数值分别为 4×10^2、6×10^1、3×10^0。

二进制是一种常用于计算机中的进位计数制，它的基数为 2，只有 0、1 两个数码，整数位的位权从右向左依次为 2^0，2^1，2^2，…。例如：二进制数 $(110)_2$ 中，各个数位上的数字所代表的数值分别为 1×2^2、1×2^1、0×2^0。

在计算机科学中，除了二进制之外，为了便于使用，常用的进位计数制还有十六进制。由于采用二进制数描述信息的位数较多，不便于记忆、交流和阅读，因此为了方便书写和表达，人们常常将二进制数转换为十六进制数。十六进制的基数是 16，包含 0，1，2，3，…，9，A，B，C，D，E，F，共 16 个数码。

不同的进位计数制用 $(S)_R$ 表示，其中 S 是具体的数码，下标 R 是该进位计数制的基数，例如 $(102A)_{16}$ 和 $(1011)_2$。有时，也用特定的字母标在末尾来标识进位计数制，例如 1011B，这里的“B”是二进制的特定字母，十进制和十六进制则分别用“D”和“H”来表示。一般情况下，十进制是默认进位计数制，因此字母“D”通常被省略。

图 1.14　十进制数转换为二进制数

2. 不同进位计数制的相互转换

将十进制整数转换为二进制数的方法是除以 2 反向取余。例如，将十进制数 37 转换为二进制数，即：$(37)_{10}=(100101)_2$，如图 1.14 所示。

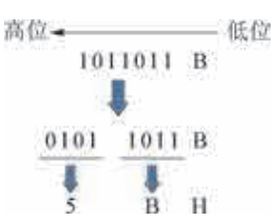


图 1.15　二进制数转换为十六进制数

二进制数转换为十进制数，一般可以将每位二进制数和该位的位权相乘再求和，这种方法称为按权展开。例如：

$$(1011)_2=1\times2^3+0\times2^2+1\times2^1+1\times2^0=8+0+2+1=(11)_{10}$$

二进制数转换为十六进制数时，把二进制数从低位到高位按 4 位一组划分，每组用一位十六进制数表示，不足 4 位二进制数，高位用“0”补齐。例如，$(1011011)_2=(5B)_{16}$，如图 1.15 所示。

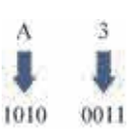


图 1.16　十六进制数转换为二进制数

十六进制数转换为二进制数时，将每一位十六进制数转换为 4 位二进制数，不足 4 位二进制数，高位用“0”补齐。例如，$(A3)_{16}=(10100011)_2$，如图 1.16 所示。

3. 数据的存储单位

比特(bit)是计算机中最小的数据存储单位，即一个二进制位，一位的取值只能是 0 或 1。

字节(Byte)是计算机中信息组织和存储的基本数据存储单位，1 字节就是 8 比特。字节常用 B 表示，描述存储容量的常用单位还有 KB、MB、GB、TB、PB、EB 等，其换算规则如表 1.3 所示。

表 1.3　常用存储单位换算表

存储单位	换算规则	存储单位	换算规则
KB，千字节	1 KB= 1024 B= 2^{10} B	TB，太字节	1 TB= 1024 GB= 2^{40} B
MB，兆字节	1 MB= 1024 KB= 2^{20} B	PB，拍字节	1 PB= 1024 TB= 2^{50} B
GB，吉字节	1 GB= 1024 MB= 2^{30} B	EB，艾字节	1 EB= 1024 PB= 2^{60} B

二、数字化

今天，数字技术向人类生活各个领域全面推进，迅速改变着我们的学习和生活：网上购物可以让消费者足不出户购买商品，电子地图能及时规划出最优的出行路线，在线政务让市民办事更高效，数字博物馆让人们跨时空浏览馆藏珍品。在丰富多彩的信息社会里，数字化是计算机处理信息的基础，将现实世界中各种各样的信息用二进制数来表示的过程就是信息的数字化。

1. 模拟信号和数字信号

现实世界中，我们将连续变化的物理量称为“模拟量”，如温度、速度等。数字化可将模拟量转换成数字量，数字量的变化在时间或数值上都是离散的。模拟量和数字量都是对某一个物理量的反映或表达。两者的主要区别是：模拟量是连续的，数字量是离散的。例如，水银温度计中的水银汞柱伸缩是连续变化的，反映的是模拟量；数字温度计显示的数字是离散的，反映的是数字量。

在电子设备中，模拟量通常以模拟信号的形式进行传递，数字量则以数字信号的形式进行传递。在一定条件下，模拟信号和数字信号可以相互转换，如图 1.17 所示。以声音的数字化为例，麦

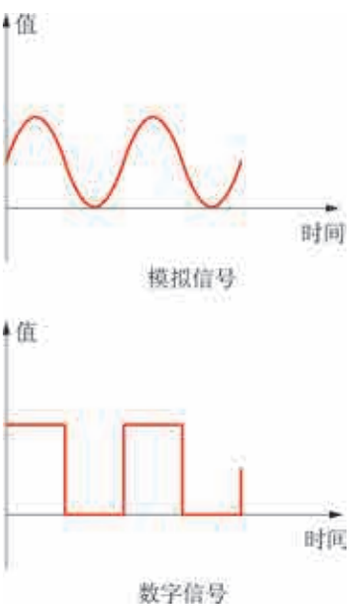


图 1.17　模拟信号与数字信号

克风能够将声波的振动转化为电信号，这是一种模拟信号，再经过模数转换设备（如声卡等）的处理后，可以转换成计算机内部能够处理的数字信号。

2. 模拟信号的数字化过程

在计算机领域，数字化是指将复杂多样的信息表示为计算机可以处理的二进制代码的过程。通常，使用电子设备（如话筒等）采集的信号是模拟信号，为了能让数字设备进行存储和处理，就需要将模拟信号转换为数字信号，这种转换过程主要包括采样、量化和编码。

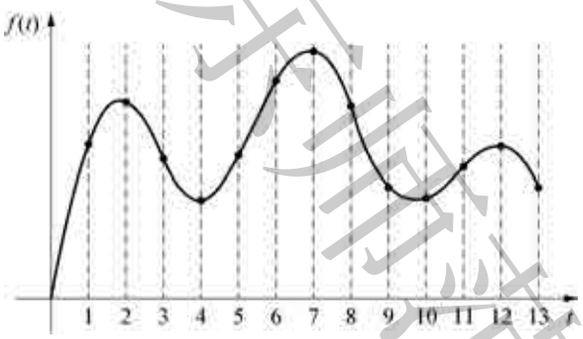


图 1.18　采样

采样是在连续的模拟信号中，每隔一定时间（或空间）取一个值的过程，如图 1.18 所示。对于同一模拟信号，采样的时间间隔设置越小，单位时间内采集的样本数量越多。每秒的采样次数称作采样频率，单位用赫兹（Hz）表示。

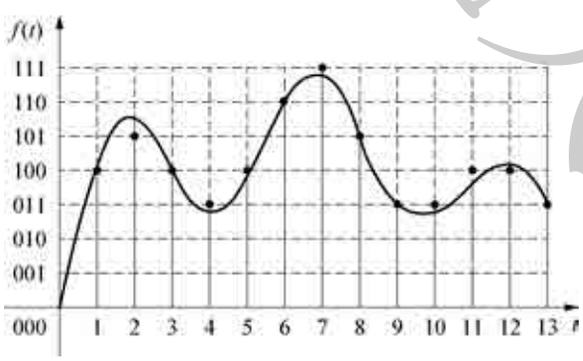


图 1.19　量化

量化是把采样的值用二进制数值表示出来。其过程是按模拟信号变化的幅度将其划分为几个区段，把落在某个区段的采样样本值归成一类，并赋予相应的二进制数值来表示量化值，如图 1.19 所示。

对这些二进制数值进行编码，就可以形成一系列二进制代码。这样，计算机就可以对其进行识别、存储和加工了。

数字化的应用很广泛，例如，图像的数字化就是用离散的量来表示连续的空间，视频的数字化就是用离散的量来表示连续的时间和空间。

项　目　实　践

电子校刊方便借阅和存放，而且还可以提供多媒体形式的内容，丰富了阅读资源。

1　选用数字化工具，将纸质校刊中的文字制作成电子文本或音频文件。

2　选用数字化工具，将纸质校刊中的插图制作成图像文件。

3　与同学合作，尝试利用数字化工具制作电子校刊，体验电子校刊的制作过程。

三、编码

为了有效处理信息，人们常常通过编码的方式来表示信息。例如，公交车线路号就是一种编码，人们通过线路号来选择和区分公交线路。学校里，教学楼的楼号和教室号也是编码，老师和同学们根据楼号和教室号就能确定上课和活动的地点。生活中，身份证号、银行卡号、学籍号、车牌号等都是编码。

编码是为了方便信息的存储、检索和使用而规定的符号系统。编码的过程是将信息按照一定的规则进行变换。图书馆中，管理员根据图书类目、种次等信息对图书进行编码，形成索书号，然后按索书号将图书放在书架相应的位置上，便于读者顺利地找到图书。

要使用计算机处理各种各样的信息，需要通过编码的方式将信息转换成用“0”和“1”表示的二进制代码。

1. 字符编码

(1) ASCII码

目前，国际上广泛使用的英文字符编码是 ASCII 码（American standard code for information interchange，美国信息交换标准码）。ASCII 码诞生于 1963 年，用于计算机内部字符的存储和计算机与外部设备的通信。标准的 ASCII 码为 7 位二进制编码（即 D6～D0 位），存储时占用一个字节，最高位为 0。例如，字符“A”的 ASCII 码编码为 1000001，存储时如图 1.20 所示。标准的 ASCII 字符集定义了 128 个字符，其中包括 10 个阿拉伯数字（“0”～“9”）、26 个大写英文字母（“A”～“Z”）、26 个小写英文字母（“a”～“z”）和 33 个符号共 95 个可打印字符，以及 33 个控制字符。

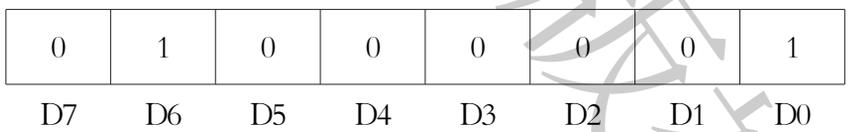


图 1.20 字符“A”的 ASCII码的存储

(2) 汉字编码

汉字也是字符。用计算机处理汉字时也要采用二进制表示的编码。目前，我国主要使用的汉字编码标准是 GB 18030—2005，它支持多种字节的汉字编码，如单字节、双字节和四字节编码等。在 GB

18030—2005 中，大部分常用汉字采用双字节编码。

利用键盘输入汉字时，还需要通过另外设计的汉字输入码来实现。汉字输入码可以使用字母、数字或符号来对汉字进行编码。汉字输入码有多种形式，例如以汉字的字音为主的音码和汉字的字形为主的五笔字型码等。

此外，输出汉字时，还会使用汉字字形码。汉字字形码是字库中存储的汉字字形的数字化信息，用于汉字的显示和打印输出。目前，汉字字形既可以用点阵方式表示，也可以用矢量方式表示。

（3）Unicode字符集和编码方案

由于不同语言的编码各不相同，为了统一所有文字的编码，Unicode 应运而生。Unicode 是计算机科学领域里的一项业界标准。它对世界上大部分的文字系统进行了整理、编码，避免由于编码冲突而产生的乱码问题，使得计算机可以用更为简单的方式来处理和呈现文字。

Unicode 字符集分为 17 组（平面），每组含有 65536 个码位，共 1114112 个。它就像一本“大字典”，每一个码位都唯一对应一个字符。其中，汉字位于 0 号平面和 2 号平面。

要将这本“大字典”里的 Unicode 字符转换成可用于传输、存储的二进制代码，则需要使用字符编码方案。目前，主要使用的是 UTF－8、UTF－16、UTF－32 三种编码方案。

2. 声音编码

声音是振动产生的波，由不同频率的正弦波合成。声波的振幅反映了声音的强弱，声波的频率反映了音调的高低。它是一种连续变化的模拟信号，需要通过采样、量化和编码后实现数字化。

声音的采样是指每隔一段时间在声音的模拟信号上采集一个样本数值。采样间隔时间越短，采样频率就越高，那么单位时间内得到的样本数据就越多，对声音的模拟信号的表示就越精确，声音的保真度就越高。

声音的量化是用二进制数值表示采样所得到的幅度值的过程。首先将幅度值范围划分为 2^n 个等级，每个等级对应一个幅度值，然后将采样得到的各个幅度值按一定的规则近似到某个等级，并用 n 位二进制数表示这些值。这里的 n 是量化位数。划分的等级越多，量化的位数就越多，量化精度也就越高，采样结果近似到某个等级时产生的误差就越小，音质就越有保证，如图 1.21 所示。

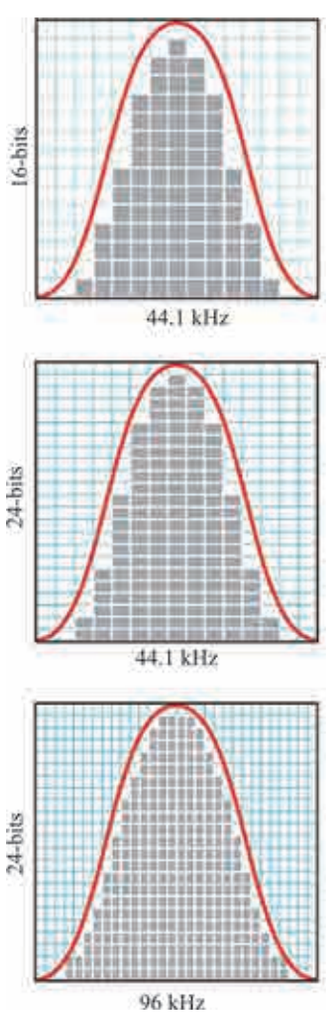


图 1.21　声音量化的级别

通过采样和量化，对获得的二进制数进行编码后，就可以将声音的模拟信号转换成二进制代码表示的数据。

一般情况下，未经压缩的音频文件的数据存储量可以按如下方法进行计算：

数据存储量 = 采样频率 × 量化位数 × 声道数 ÷ 8 × 持续秒数（字节）

例如，一组 1 小时的数字音乐（未经压缩）的采样频率为 44.1 kHz，量化位数为 16 位，声道数为双声道。则其数据存储量可按以下方法进行计算：

数据存储量 = 44100 × 16 × 2 ÷ 8 × 3600（字节）

通常，未经压缩的数字音乐会被保存为 WAV 文件格式。

3. 图像编码

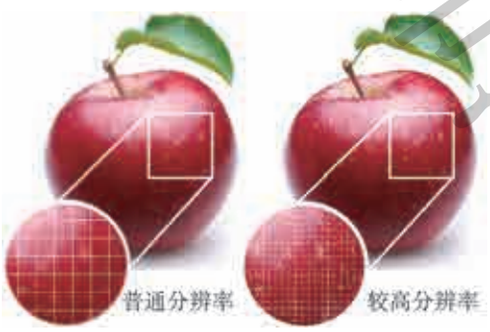


图 1.22　同一图像的不同分辨率对比

图像的采样是按一定的空间间隔从左到右、自上而下提取画面信息，将图像在空间上转换成若干个像素点，每个像素点呈现不同的颜色。

水平方向上的像素数量乘以垂直方向上的像素数量称为图像分辨率。例如，一幅图像的分辨率为 640 × 480 像素，表示该图像由水平方向上 640 个像素点、垂直方向上 480 个像素点，共 640 × 480 = 307200 个像素点组成。如果不考虑其他因素的影响，图像分辨率越高，采样的精度就越高，数字化后的图像越清晰，同时图像所占的存储空间也越大，如图 1.22 所示。

图像的量化是用若干位二进制数表示采样得到的每个像素点的颜色。首先确定颜色的取值范围，然后将近似的颜色划分成同一种颜色，每种颜色用一个二进制数来表示。

记录每个像素点的颜色所需的二进制数的位数，称为颜色深度（位深度）。对于一幅图像来说，颜色深度决定了该图像中的像素可以使用的最多颜色数量。例如，颜色深度为 8 比特时，可以表示 256 种颜色；颜色深度为 16 比特时，则可以表示 65536 种颜色。颜色深度越大，显示的图像色彩越丰富，画面越自然、逼真，如图 1.23 所示。在图像分辨率相同的情况下，颜色深度越大，图像所占的存储空间也越大。

图 1.23　同一图像的不同颜色深度对比

通常，计算机中可以采用 RGB 颜色模型来描述颜色。RGB 颜色模型（RGB color model）又称为三原色光模式，将红（red）、绿（green）、

蓝(blue)三原色以不同的比例相加，可以产生不同的颜色。每一个像素可以用 24 比特表示，所以三种原色各分 8 比特，即每一种原色的强度可以用 0～255 之间的整数来表示，用这种方法可以组合成 16777216 种颜色。例如，纯红色用 RGB 颜色模型表示为(255，0，0)，纯绿色用 RGB 颜色模型表示为(0，255，0)。

这种由纵横排列的像素点组成的图像称为位图。位图的质量主要由图像分辨率和颜色深度决定。未经压缩的位图图像的数据存储量可以按如下方法进行计算：

未经压缩的位图图像的数据存储量＝图像分辨率×颜色深度÷8(字节)

例如，一幅分辨率为 1920×1080 像素的图像，保存格式为“24 位位图”，则其数据存储量可按以下方法进行计算：

数据存储量＝1920×1080×24÷8(字节)

通常，未经压缩的位图图像会被保存为 BMP 文件格式。

四、数据压缩

很多信息，包括视频、音频等多媒体信息，经数字化转换后生成的二进制数据是以文件形式保存的，它们占用的存储空间相对较大。为了更少地占用存储空间，就需要采用合适的方式对数据进行压缩。

探 究 活 动

小申和同学们在将纸质校刊制作成电子校刊的过程中，需要将校刊中的插图制作成计算机中的图像文件。

1 尝试将图像文件保存为不同的图像文件格式。

2 从文件大小、图像质量等方面比较不同图像文件格式的差异。

实际上，数据压缩就是采用特殊的编码方式处理数据，使数据占用的存储空间相对减少，以便存储和传输。

数据之所以能够被压缩并保证压缩后可用，主要是因为数据存在如下几种现象：

数据中存在冗余。例如，在一个计算机文件中，某些符号会重复出现，某些符号会比其他符号出现得更频繁，某些字符总是在各数据

块中可预见的位置上出现等。这些冗余数据便可在数据编码的过程中去除或减少。

相邻数据之间经常存在相关性。例如，图片中常常有色彩均匀的背景，电视信号的相邻两帧之间可能只有少量的变化，声音信号有时具有一定的规律性和周期性等。因此，利用某些变换来尽可能地去除这些相关性数据，也可以实现压缩的目的。

由于人的耳、目对信号的时间变化和幅度变化的感受能力都有一定的极限，如人眼对视频有视觉暂留效应，因而将信号中这部分感觉不出的地方“掩蔽掉”，也可以实现压缩的目的。

数据压缩的方法比较多。常用的压缩方法分为无损压缩和有损压缩两种。

无损压缩是指对压缩后的数据进行还原后，得到的数据与压缩前完全相同。例如，一幅分辨率为 24×24 像素的图像，其中一行像素色彩值的排列为“红红红红红红红红红红红红红红红红绿绿绿绿绿绿绿绿”，经过某种压缩后，可以表示为“红 16 绿 8”，这种压缩就称为无损压缩。我们可使用常见的无损压缩软件对文件进行压缩，压缩后生成的压缩文件存储容量可能只有原来的几分之一甚至更小。压缩文件中的数据需要用压缩软件解压缩后才能还原使用。对于一些程序数据和文档数据而言，只能采用无损压缩，以确保数据的准确性，否则一旦还原的数据有误，就无法使用了。

图 1.24　图像文件的有损压缩示例

相对于无损压缩，有损压缩通常应用于图像、声音等数字化后存在大量冗余信息的文件。有损压缩过程中会损失一定的信息，压缩后的数据无法还原到与压缩前一致，但不会导致人们对原始数据表达的信息产生误解。以图像的有损压缩为例，图像的有损压缩是在较小地损失图像质量的情况下，对图像文件中相同或相似的数据进行大量压缩，使得生成的文件更小，如图 1.24 所示。这种技术在一定程度上损害了图像的原始质量，也就是丢掉了一些数据的信息。同一张图像，保存为不同的格式时，其数据量的差别可能会非常大。例如，分别使用 TIF 和 JPEG 格式生成的文件，大小有时会相差几十倍。常见的图像、音频、视频有损压缩格式分别是 JPEG（图像数据压缩格式）、MP3（音频数据压缩格式）、MPEG（视频数据压缩格式）等。

作业练习

1 一个 7位二进制数，如果其最高位和最低位都为 1，则用十进制表示该二进制数，其最大值可能是_____，最小值可能是_____。

2 小申同学录制了一个时长 1分钟的音频文件，他准备将这个文件传输至个人在线学习空间。该文件采用的是 WAV格式，文件采样频率是 44.1 kHz，量化位数为 16位，双声道。请计算该文件在个人在线学习空间中占用的存储容量。

3 小申和同学们在小组活动中设计了本小组的 LOGO，如图 1.25所示。他们用 2位二进制数来表示图中的颜色，并对图中的颜色进行了编码，从而将本小组的 LOGO表示为一串二进制数，其中第 4行的编码为 0101010110000010。

（1）用 2位二进制数表示颜色，最多可以表示_____种颜色。根据上述编码，图中蓝色色块的二进制编码应表示为_____；红色色块的二进制编码应表示为_____；第 6行的编码为_____H。

（2）请在图 1.26中设计自己的 LOGO，并涂上相应的颜色。然后对使用的颜色进行二进制编码，并说明编码规则。

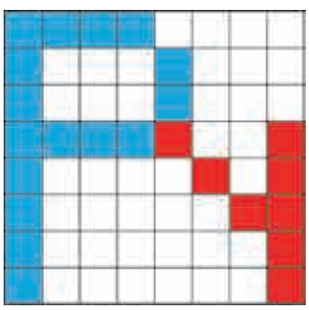


图 1.25 小组的 LOGO

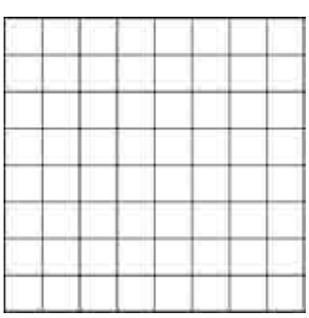


图 1.26 我的 LOGO

知 识 延 伸　图像隐写术

图像作为信息的一种表现方式，其版权问题一直备受关注。通常，很多图像都会被加上水印来标明版权，防止盗版。这种方式的确能够在一定程度上减少图像被盗版滥用的情况，但会对图像的美观造成一定的影响，也不利于对图像进行再处理。目前除了显式的添加水印的方式，还可以采用数字图像隐写的方式，将版权信息隐藏在图像中，使其能够被检测。隐写术是一门关于信息隐藏的技巧与科学，所谓信息隐藏是指不让除预期的接收者之外的任何人知晓信息的传递事件或信息的内容。数字图像隐写术的应用十分广泛。

例如：一个 24位位图图像中的每个像素的三原色（红、绿和蓝）各使用 8比特来表示。仅考虑红色的情况下，就可以用 2^8 个不同的数值来表示深浅不同的红色。以 11111111和 11111110这两个数值所表示的红色为例，它们对人眼来说几乎是无法进行区分的。因此，所有的最低有效位（通常指右起最低位）就可以被用来存储颜色以外的某些信息，如图 1.27所示。如果对绿色和蓝色进行同样处理的话，则可以在三个像素中存储一个字节的信息。

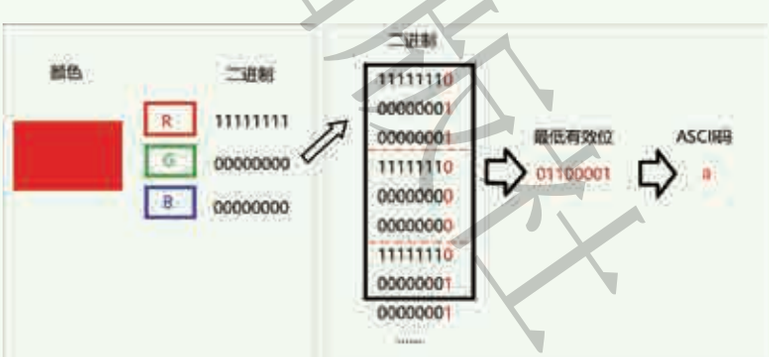


图 1.27 图像隐写术

第三节　大数据及其作用与价值

信息技术的快速发展与广泛应用加速了数据总量的增长。公共场所的安保监测系统在实时监测过程中会产生数据；人们在使用城市公共交通的刷卡付费系统时也会产生数据；人们在利用各种网络社交工具进行互动交流时同样会产生数据。这些数据的规模呈现指数级数增长，人类进入了大数据时代。大数据已成为社会的一项重要资源。

体　验　思　考

小申在电子图书网站上查看电子图书的信息时，发现网站上会立即显示“搜索此图书的人还在搜索的图书”或“给您推荐的图书”等信息，这些信息还会根据小申搜索图书情况的变化而持续更新，如图 1.28 所示。

图 1.28　推荐的图书信息

思考：

1. 电子图书网站向小申推荐图书的依据是什么？
2. 举例说明，日常生活中还有哪些场景具有类似的智能推荐功能。

一、理解大数据

信息技术的快速发展改变着人们对数据的采集、分析与使用方式。依托移动互联网，人们可以更便捷地访问网络，从而大量地产生和传输数据；通过传感技术，人们可以不间断地采集数据。这些数据规模巨大、格式多样，已经很难用传统的方式进行处理。于是，大数据技术应运而生，人们通过对这些数据进行分析、挖掘，从而发现和应用其蕴藏的价值。

图 1.29　大数据的特征

大数据也是数据。通常，大数据具有海量的数据规模、多样的数据类型、快速的数据流转和价值密度低四大特征，如图 1.29 所示。

1. 数据规模大

借助可穿戴设备、物联网和云计算等技术，人们的行为和物体的运行轨迹可以被采集和保存，形成大规模数据。以出租车定位系统中的数据为例，某市出租车公司的出租车通常每隔 10 秒钟会自动向总部的服务器发送一条数据，记录自己所在的经纬度、车速、车上是否有乘客、行驶方向等信息。那么，按照该市大约有 5 万余辆在运营出租车计算，一天产生的定位数据就大概有 4 亿余条。今天的数据已经从 TB 级别跃升为 PB 级别。

2. 数据类型多

网络时代，数据的种类越来越多样，从二维表格表示的关系数据，扩展为文本数据、网络日志、音频、视频、图片、地理位置信息等等。大数据技术的发展不仅为快速处理海量数据提供了支持，也为处理不同来源、不同格式的多元化、多维度数据提供了可能。

3. 处理速度快

大数据的“快”表现在多个层面上，它既包括数据产生得快，也要求数据处理得快。海量、多样的数据快速处理对数据处理的速度提出了更高的要求。过去，需要几天或更多时间处理的数据，现在可能要

在几分钟之内完成处理。以衡量大数据领域计算实力的 100TB 数据的排序速度为例，目前人们已经能够在百秒内完成对 100TB 数据的排序。

4. 价值密度低

大数据是有价值的，价值密度的高低与数据总量的大小几乎成反比。大量的不相关数据，不经过处理则价值较低，属于价值密度低的数据。以安全监控视频为例，一部连续 1 小时不间断的监控视频中，有用的视频数据可能仅有一两秒。如何对大数据进行分析，获得有价值的信息，已成为目前大数据背景下亟待解决的问题。

与传统数据相比，大数据在数据规模、采集方式、分析方法、价值利用等方面都有了很大的发展，影响着我们每个人的生活与学习。例如，人们对网络用户的搜索请求及交互数据进行分析，建立用户行为模型，为用户提供个性化智能搜索和内容推荐。

二、大数据处理过程

网络时代，每天来自商业、社会、科学、工程、医学以及日常生活等方方面面的数据，不断存储到我们的计算机和各种存储设备中。面对海量数据，人们利用技术方法，经过整合、归纳与评估，提取出有价值的信息，为用户的决策提供依据。

计算机与网络技术的快速发展使得数据处理方式发生了巨大的变化，数据的处理效率得到了极大的提高。借助信息技术，可以对人的在线行为进行记录，也可以对社会中的各种事物进行记录，通过大数据分析，更好地做出预测和决策。一般而言，大数据处理可分为四个步骤：数据采集、数据预处理、数据分析和数据挖掘应用。

项　目　实　践

如图 1.30所示，使用不同的电子图书阅读平台（网站或移动应用程序），观察这些电子图书阅读平台是否具有智能推荐功能。尝试查询图书、阅读图书、收藏图书，观察平台推荐的图书是否随之发生变化，判断平台是依据读者的哪些行为推荐图书的，并填写表 1.4。

图 1.30　某电子图书阅读平台的图书推荐

表 1.4　电子图书阅读平台智能推荐功能分析表

电子图书平台	图书推荐	读者行为及其影响	图书推荐依据
××阅读平台	读过还读	读者收藏电子图书后，平台推荐图书发生变化	读者收藏的图书

1. 数据采集

大数据的采集是大数据处理过程中的最初环节，它可以通过射频识别技术（RFID）、传感器技术、社交网络媒体等方式，获得各种类型的海量数据。例如，随处可见的共享单车就是通过智能锁中的通信模块和用户身份识别卡（SIM），将单车的通信连接状态、车锁状态、使用记录等数据，通过网络上传到共享单车的服务平台上。而有“掌上实验室”之称的手持实验技术（如图 1.31 所示）应用中，通过采集与传感设备把外界环境中的某个物理量的变化以模拟信号输出，再经过模拟数字转换装置进行转换，得到实验数据后，可上传至云端存储起来，以备实验分析和处理用。

图 1.31　部分手持实验技术仪器

2. 数据预处理

大数据采集过程中通常有一个或多个数据源，所采集的数据易受到噪声数据、数据缺失、数据冲突等影响。因此，需要对采集的数据进行预处理，以保证大数据分析与预测结果的准确性与价值。例如，出

租车在运营过程中，可能会由于高层建筑物遮挡、驶入地下隧道以及出租车本身的定位系统装置故障等原因，导致出租车定位数据缺失。此外，由于数据采集设备的不稳定或者司机的误操作，会导致两条或多条记录重复。对于这样的重复数据，就需要进行合并或删除操作。预处理后的出租车轨迹数据如图 1.32 所示。

Index（索引号）	number（车牌号）	longitude（经度）	lantitude（纬度）	carstatue（运营状态）	time（时间）
2739	沪AT[illegible]1	108.94698	34.232581	4	2016-01-07 23:58:59
2740	沪AT[illegible]1	108.94702	34.232578	4	2016-01-07 23:59:29
2741	沪AT[illegible]1	108.94704	34.232578	4	2016-01-07 23:59:59
2742	沪AT[illegible]64	108.98535	34.27746	4	2016-01-07 04:02:22
2743	沪AT[illegible]64	108.98534	34.277455	4	2016-01-07 04:03:38
2744	沪AT[illegible]64	109.01538	34.269248	4	2016-01-07 12:52:51
2745	沪AT[illegible]64	109.01538	34.269248	4	2016-01-07 12:53:21
2746	沪AT[illegible]64	109.01537	34.269235	4	2016-01-07 12:53:50
2747	沪AT[illegible]64	109.01538	34.269238	4	2016-01-07 12:54:20
2748	沪AT[illegible]64	109.01539	34.26924	4	2016-01-07 12:54:50
2749	沪AT[illegible]64	109.01537	34.269226	4	2016-01-07 12:56:51
2750	沪AT[illegible]64	109.01538	34.269247	4	2016-01-07 13:00:50
2751	沪AT[illegible]64	109.01538	34.269245	4	2016-01-07 13:01:20
2752	沪AT[illegible]64	109.01536	34.269225	4	2016-01-07 13:01:51

图 1.32　预处理后的出租车轨迹数据示例

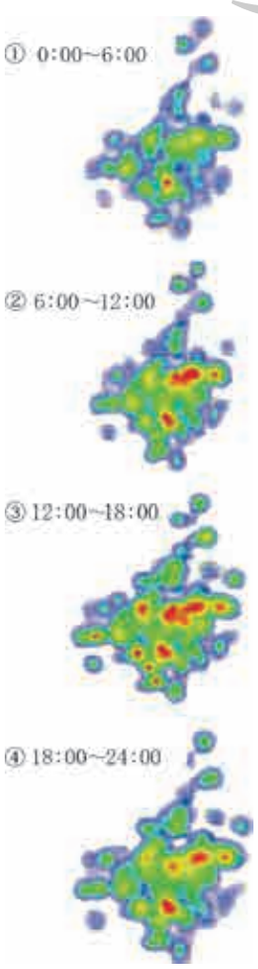


图 1.33　乘客出行时段与地区热力图

3. 数据分析

数据分析是大数据处理中的关键环节。根据大数据应用情境与决策需求，选择合适的数据分析技术，提高大数据分析结果的可用性、准确性和价值。这个过程的主要特点就是目的清晰，按照一定的规则进行处理，才能得到有效分析。例如，在获得出租车的运行数据后，可以进一步得到出租车每一次载客运营的起点和终点以及相应的时间，进而得到人们在工作日各个时间段的出租车用车情况。如图 1.33 所示，在 0:00～6:00，出租车用车数量较少；在 12:00～18:00，红色区域表示出租车用车数量较多。

4. 数据挖掘应用

经过数据分析，可以描述出事物的变化状况，找出其中的规律，将分析结果运用到实践中。例如，根据时段与地区的出行热力图，实时分析旅游景点、居民区、学校、商业区等地区的人流量变化，调整社会

安保人员，保持良好的社会秩序。

大数据处理一般需要经历上述四个步骤。在具体实施时，其中的细节、工具的使用、数据的完整性等还需要结合具体需求、行业特点和整个时代的变化而不断变化更新，才能符合大数据时代的特点。

三、大数据的作用及社会影响

互联网产生的海量数据汇同强大的计算技术，使大数据发挥着越来越重要的作用。人们利用算法对收集到的庞大数据进行分析处理，找到数据之间的关联性，并匹配出某些结果或现象，从而找寻到某种相关性，用于调整和制订后续的各种策略。大数据与人工智能技术的紧密结合，帮助人们从数据中获取更准确、更深层次的信息，挖掘出数据背后的价值，催生出新业态和新模式。

探 究 活 动

小申在学校图书馆的志愿服务活动中，常常会遇到同学提出这样的要求："能否推荐一本适合自己阅读的书籍？"

1 如果你是小申，在推荐书籍的时候，会考虑哪些因素，需要哪些数据？

2 学校准备在不久的将来，建设数字图书馆。请参考志愿者向读者推荐图书的过程，帮助设计数字图书馆的电子图书推荐方案。

1. 大数据的作用

大数据可以反映社会现象。借助大数据，能够反映出人们的意图、情感、观点和需求，这些情感因素会决定人们在决策或行动时所采取的方式和所选择的方法。例如，通过搜索引擎中的检索数据，可以分析和了解人们的浏览习惯；通过购物网站中的数据，可以了解人们的购物喜好；通过微博等平台中的数据，可以了解人们对某一问题的评判；通过信息技术实时分析数据，可以反映出社会现象。

大数据可以预测发展趋势。在大数据分析过程中，通过分析不同类型数据的相互关联，描述数据的动态变化，就可以比较清晰地展示事件的脉络关系，预测其发展趋势。例如，基于用户和车辆的 LBS（基

图 1.34　大数据与交通

于位置的服务）定位数据，分析人车出行的个体和群体特征，进行交通行为的预测，如图 1.34 所示。交通部门可以预测不同时间不同道路的车流量，进行车辆智能调度，用户则可以根据预测结果选择拥堵概率更低的道路。

大数据可以指导决策的制订。通过大数据技术，人们可以获得所希望的数据，并能得到与之相关联的分析结果，从而能更全面地认识事物的特征及发展趋势，为行动决策奠定基础。例如，在社会治安方面，警方通过“案件数据分析和趋势预测系统”中的各类数据，预测未来某时段、某区域可能发生治安问题的概率及类型，作为警力布置和安全防范的决策，从过去“被动围着案件转”发展为“前瞻性地巡逻防控”。

2. 大数据对社会发展的影响

随着互联网的发展，大数据已经渗透到很多行业，成为重要的生产要素，并通过各行各业的不断创新，逐步为人类创造更多的价值和财富。


图 1.35　大数据技术提升城市公交运行效率

大数据技术优化社会管理模式。政府部门通过分析社会中各领域的大数据，可以改善城市生活，提升城市管理水平，促进智慧城市的建设。例如，由于出租车和公交车都服务于中长距离的居民出行，在人们的出行选择中互相替代的可能性较高，因此，科学家利用出租车运营系统来获取城市居民出行的起点和终点、经过的路线等数据，分析城市居民出行的交通需求，对城市公交线路网进行优化，提升城市公交运行效率，如图 1.35 所示。

大数据技术创新提升服务质量。通过分析用户的消费数据，商家可以有指向性地向用户推送商品。例如，商业网站记录用户在网站中的搜索、浏览、购买、点评等在线行为数据，通过分析这些数据，了解用户的购物习惯，判断用户的购物喜好，以此为依据进行个性化商品推送，实现个性化服务。

图 1.36　大数据技术助力科学研究

大数据技术成为科学研究的新途径。借助对大数据的分析研究，能够发现医学、物理、经济和社会等领域的新现象，揭示自然与社会中的新规律，并预测未来趋势，如图 1.36 所示。正在兴起的环境应用科学、基于全球数据共享的天文观测、下一代传感器网络与地球科学等，都是正在快速成长和发展的交叉学科方向，也是大数据在科学研究和发现中的新应用。

随着大数据技术的广泛应用，大数据技术已经渗透社会的很多领域。许多国家先后将大数据研究提升到国家战略层面。我国也充分认识到大数据时代带来的重大机遇，部署了一系列与大数据研究紧密相关的科研计划。大数据已经成为关系国家经济发展、社会安全和科技进步的重要战略资源。

图 1.37　指纹识别技术

通过对大数据的挖掘以及对分析结果的应用，在给生活带来便利的同时，也可能会引发一些新的社会问题。随着指纹识别、人脸识别等技术的应用，人们开始关注指纹、虹膜、面容等个人生物信息的所有权问题，如图 1.37 所示。如何避免因生物信息被搜集可能带来的个人隐私泄露、数据窃取、网络欺诈等问题？甚至还曾有网络不法分子通过收集电子邮件、微博、电子商务等数据，利用大数据技术向所搜索的目标发起精准攻击。因此，我们在学习利用大数据预测并做出决策的方法时，也要掌握特定的防护措施，加强数据安全意识。

作业练习

1　在日常出行中，越来越多的人开始使用电子地图进行导航。电子地图在 GPS导航系统的支持下，能够较为准确地显示行进的路线和路况，并能实时地进行调整。请分析为什么 GPS导航系统能根据路况实时调整行进的路线？它又是如何知道路况变化情况的呢？

2　结合个人网络学习体验（如慕课学习、在线课程等），从学习诊断、资源推送、过程管理等方面，思考大数据是如何支持个性化学习的。

知识延伸 用户画像

网络技术的发展使得越来越多的电商网站开始通过用户画像的方式改进服务质量，提高服务效率。

用户画像通常是根据用户社会属性、生活习惯和消费行为等信息而抽象出的一个具有特征标识的用户模型。构建用户画像的核心工作是通过分析用户信息得到高度精炼的特征标识。例如，一个用户最近经常购买一些玩具，电商网站即可根据玩具购买的情况将用户特征标识为“有孩子”，甚至还可以判断出孩子的大概年龄，贴上“有 5~10岁的孩子”这样更为具体的标识，将所有用户特征标识整合起来，就成为了用户画像。

构建用户画像时，一般可以从以下几个维度加以分析：用户静态属性、用户动态属性和用户消费属性等。用户静态属性是用户画像建立的基础，主要从用户的基本信息进行划分，如性别、年龄、学历、地域、婚姻等，依据不同的产品，有针对性地提取相关信息，并将这些信息进行不同程度的权重划分。用户动态属性指用户在互联网环境下的上网行为，如娱乐偏好、社交习惯、出行方式、学习手段等，反映出一个用户可能会对某类产品感兴趣。消费水平、消费嗜好等是用户消费属性，这些在一定程度上能够反映用户的消费观念。

数据量的爆发式增长和大数据分析技术的成熟使用户可被捕捉的数据越来越多，用户画像在各行业应用的价值也在不断提升。例如，在零售业，精准服务是用户画像最直接和有价值的应用，如图 1.38所示。

图 1.38 用户画像与精准服务

第二章

算法与程序实现

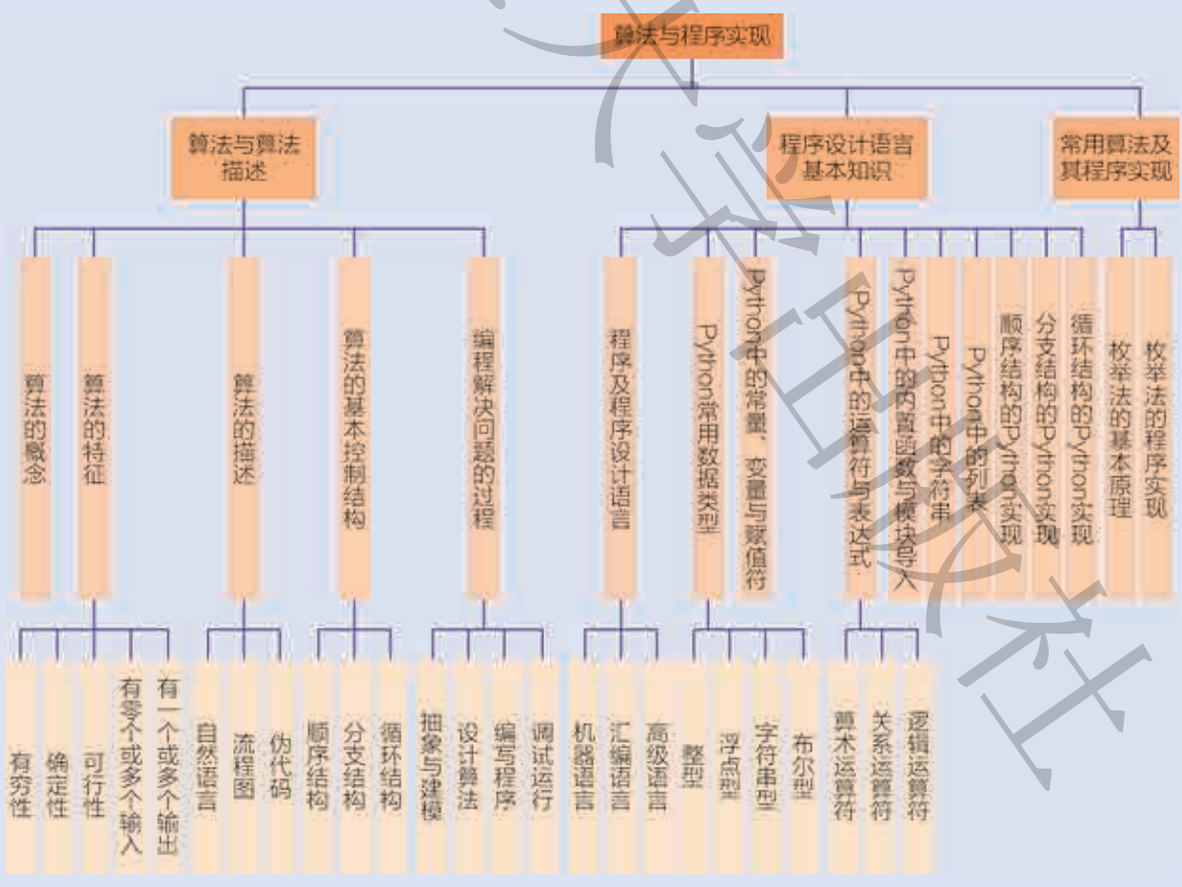
本章学习目标

- 理解算法的概念和特征，运用恰当的描述方法和控制结构表示简单算法，认识算法在问题解决中所起的作用。
- 掌握一种程序设计语言的基本知识，并使用程序设计语言编写程序解决简单问题，掌握运行和调试程序的方法。
- 体验编程解决问题的一般过程，认识问题解决过程中不同算法的效率，学会选择恰当的算法进行求解。

算法非常古老，它的诞生早于计算机数千年，但它的神奇之处就在于它存在于我们生活中的各个角落，并使得我们的生活更加轻松、美好。比如，最大公约数算法可用于铺地砖时的砖块规格选择；最短路径算法可用于公共交通网络规划；人脸识别算法广泛用于数码相机拍照时的人脸捕捉；启发式算法可用于机场决定飞机的起飞顺序；匹配算法可用于器官移植配对……算法无处不在！

随着信息技术的发展，人们使用程序设计语言对各种算法进行了程序实现，并将程序安装在不同的数字化设备上，当我们使用这些设备来解决生活中的各类问题时，就更加便捷了。面对同一问题，往往可以使用多种不同的算法，算法的效率则又会影响到程序的效率和问题的解决。因此，如何选择合适的算法来解决实际问题，也是解决问题过程中至关重要的一步。

本章知识结构



项目主题

编程应用助健康

项·目·情·境

要保持健康的身体，就离不开科学、规律的运动。进入智能时代后，以物联网、云计算、大数据为特征的智能运动环境正改变着人们的运动方式。

小申是一名运动爱好者，这学期学校健身中心更新了一批跑步机，他和同学就经常在体育活动课上去锻炼。他们发现在跑步机上可以选择不同的跑步预设模式，在不同的模式下，跑步机会动态调节运行速度和坡度；跑步机上还有很多传感设备，可以在运动过程中实时监测运动者的各种身体数据。他们都很疑惑：跑步机是如何实现这些功能的呢？

项·目·任·务

任务 1

学习智能跑步机中预设跑步模式的算法，理解算法的特征，设计并完成跑步机其他预设模式的算法描述。

任务 2

学习使用 Python 程序实现身体质量指数的计算、显示和简单统计，查阅资料并完成卡路里的计算、显示和统计等各项任务。

任务 3

使用常用算法，设计跑步训练课程报表，并描述完成这些信息统计所选择的算法和理由。

第一节　算法与算法描述

在实际生活中，人们一直都在寻求有效的问题解决方法。例如，做饭时，如何在做完一桌饭菜后，还能保证饭、菜、汤都有一个合适的温度；旅游时，如何规划旅行路线，以确保在有限的时间和预算内使行程的性价比最高；如何设置有效的电梯调度方案，以确保乘客等待的总时间最短……对问题解决的思考在生活中比比皆是，当这些解决问题的步骤被人们描述并记录下来之后，就成为了可以重复执行的、用来解决一类问题的算法。

体　验　思　考

由于城市中的人口密集度高，在有限的空间内进行锻炼成为了大部分人的不二选择，所以智能跑步机逐渐成为了人们家庭中常备的运动器材。为了满足不同人群的锻炼需求，提高锻炼效果，常见的家用智能跑步机可以提供多种预设模式选择的功能，例如“心率跑”“坡度跑”等，如图 2.1所示。

图 2.1　某智能跑步机

思考：

1. 智能跑步机是如何根据用户选择的跑步模式，控制用户的具体跑步过程的？

2. 当设定为某种跑步模式时，跑步机又是如何根据不同的人和实时运动的情况进行调节，从而使人获得最佳运动效果的？

一、认识算法

信息时代，日益先进的采集设备和存储技术记录着人们每天产生的大量数据，人们对于各种应用需求的类型和难度也逐渐增加，无论是三维图形生成、海量数据处理、机器学习，还是图像识别等，都需要

靠算法来解决。

探 究 活 动

智能跑步机能够为用户提供多种预设的跑步模式，会根据用户的选择和用户输入的跑步参数（包括年龄、体重、跑步时长等），控制跑步机的机电设备运转。以某款智能跑步机为例，当用户开机并选择“心率跑”（即用心率来指导跑步训练，在特定的心率下进行训练来提高心肺能力）模式后，跑步机运转过程描述如下：

① 直接选择预设值，或是等待用户输入个人体重、跑步时长、年龄、跑步时速等；

② 计算并显示目标心率；

③ 倒计时 3 秒，然后提示用户开始跑步；

④ 给电机发送信号，启动跑带，运转至设定的跑步时速；

⑤ 在跑步过程中监测当前心率，如果当前心率不在目标心率的浮动范围内，则调节跑带坡度，直至当前心率稳定在目标心率的浮动范围内；

⑥ 判定是否达到设定的跑步时长，如未达到，则继续监测当前心率，否则给电机发送信号，逐渐降低跑带运转速度至停止；

⑦ 结束本次跑步。

请仔细阅读以上关于“心率跑”模式的说明，思考以下问题：

1 上述描述是否存在不够明确的地方？请罗列出来。

2 描述中有一项为“选择预设值”，请解释一下此处“预设值”的含义和作用。

1. 算法的概念

算法在生活中是普遍存在的，广义地讲，算法是在有限步骤内求解某一问题所使用的步骤和方法。例如，在炒菜时，先放什么，后放什么，这也有一定的顺序和方法，这种顺序和方法我们称之为炒菜的算法；在做数学题时，每一道题都有对应的具体计算方法和步骤，可以称之为这道题的解题算法；使用跑步机跑步时，跑步机会根据用户的选择执行不同的跑步模式，每种跑步模式对应一种算法。

在计算机科学领域中，算法是一系列的计算步骤，用来将输入的数据转换成输出的结果。借助于计算机处理的高效、自动化计算能力，人们的很多算法思想已经变成现实。例如，将设计人员设计好的三维模型交给计算机来渲染，可以实现三维虚拟场景生成；将下棋的规则和方法借助计算机来实现，可以实现人与计算机对弈；将人对图片的识别和认识过程通过模型设计让计算机进行模拟，可以实现图像

的自动识别等等。

2. 算法的特征

算法是解决问题过程中“做什么”和“怎么做”的步骤的描述，一个算法必须满足有穷性、确定性、可行性、有零个或多个输入、有一个或多个输出这五个特征。

（1）有穷性

算法必须是由有限个步骤组成，即算法一定要能够结束。例如，智能跑步机中预设的各种跑步模式的算法都可以完成并结束一次跑步训练。

（2）确定性

算法中的每一个步骤都应该是确定的、没有歧义的。模糊不清、模棱两可或带有二义性的描述都会影响算法的确定性。例如，智能跑步机检测用户当前心率是否在目标心率浮动范围内时，这个范围就必须是明确的值，或者是可以通过输入值计算后得到的明确的值。

（3）可行性

算法的可行性就是指每一个步骤都可以被计算机执行，可以方便地用来解决某一类问题。

（4）有零个或多个输入

输入就是算法在执行时要从外部获取的数据。输入可以是多个也可以是零个，零个输入并不代表这个算法没有输入数据，所需数据一般已包含在算法中，只是这个输入的数据没有直观地显现出来。例如，智能跑步机在提醒用户开始跑步前会有 3 秒的倒计时，这个 3 秒就是在算法中预设的值。

（5）有一个或多个输出

输出就是算法实现所得到的结果，是算法对输入的数据加工处理后得到的。输出可以有一个或多个，没有输出的算法是没有意义的。例如，用户在智能跑步机上的实时心率显示就是一种输出；对跑带坡度的调节也是一种输出。

探 究 活 动

之前的“探究活动”中描述的智能跑步机“心率跑”模式的操作还不能够直接被跑步机执行，它并没有满足算法的特征，如“计算并显示目标心率”中并没有说明如何计算目标心率。

请结合算法的特征，针对已经找出的描述不够明确的方面，提出修改建议，并记录在表 2.1中。

表 2.1　智能跑步机“心率跑”模式说明分析表

序号	不符合算法特征的方面	修改建议
1		
2		

二、算法的描述

算法的描述就是把解决问题的方法和步骤用规范的方式描述出来。这种描述既可以作为程序设计人员编写代码的依据，又可以供算法研究、学习和交流之用，并不依赖于任何一种语言。

探　究　活　动

之前的“探究活动”中描述的“心率跑”模式存在着一些不符合算法特征的地方，经过调整后的算法如下：

① 初始化：个人体重(千克)$weight = 60$，跑步时长(分)$time = 60$，年龄(岁)$age = 30$，跑步速度(千米/时)$speed = 6$，计时(分)$count = 0$；

② 显示 $weight$、$time$、age、$speed$ 的预设值，并等待用户修改和确认；

③ 根据公式 $TargetHR = 0.7 * (220 - age)$，计算并显示目标心率 $TargetHR$；

④ 倒计时 3 秒，显示“开始跑步”；

⑤ 给电机发送“启动跑带运转”信号，启动跑带，运转速度达到 $speed$；

⑥ 监测当前心率 $CurrentHR$，如果 $CurrentHR < TargetHR - 5$，则增大跑带坡度 $Angle$；如果 $CurrentHR > TargetHR + 5$，则减小跑带坡度 $Angle$；

⑦ 计时 $count$ 加 1 分钟；

⑧ 如果 $count \geqslant time$，则执行步骤⑨，否则执行步骤⑥；

⑨ 给电机发送“逐渐降低跑带运转速度”信号，直至跑带运转速度降至停止(0 千米/时)；

⑩ 结束本次跑步。

经过以上算法的学习后，请思考一下，以上算法描述有什么优缺点？除了这种描述方法之外，通常还有哪些其他的描述方法？

算法的描述方法很多，其中主要有自然语言、流程图和伪代码三种。

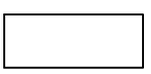
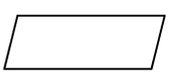
1. 自然语言

自然语言就是人们日常生活中使用的语言。用自然语言描述的算法通俗易懂，但也有明显的不足：

用自然语言描述比较复杂的算法时，会显得很冗长，表述不够直观、清晰。自然语言在描述上容易出现歧义，容易引起算法步骤的不确定性，尤其是在算法中存在较复杂的逻辑时，不易清晰地表示出来。

2. 流程图

流程图是由一些简单的图形符号组成，用来表示问题解决的步骤及顺序的方法。根据我国颁布的信息文件编制符号及约定标准（GB1526－89），常用的流程图符号如下：

	起止框	表示算法的开始或结束
	处理框	表示要处理的内容
	输入/输出框	表示数据的输入或结果的输出
	判断框	表示条件判断的情况：满足条件，执行一条路径；不满足条件，则执行另外一条路径
	连接符	用于连接因页面不够而断开的流程线
	流程线	指出流程控制方向，即执行的次序

用流程图表示算法，整个流程直观、清晰。但流程图表示算法时占用的篇幅比较大，也不易于修改。例如，“心率跑”模式算法的流程图如图 2.2 所示。

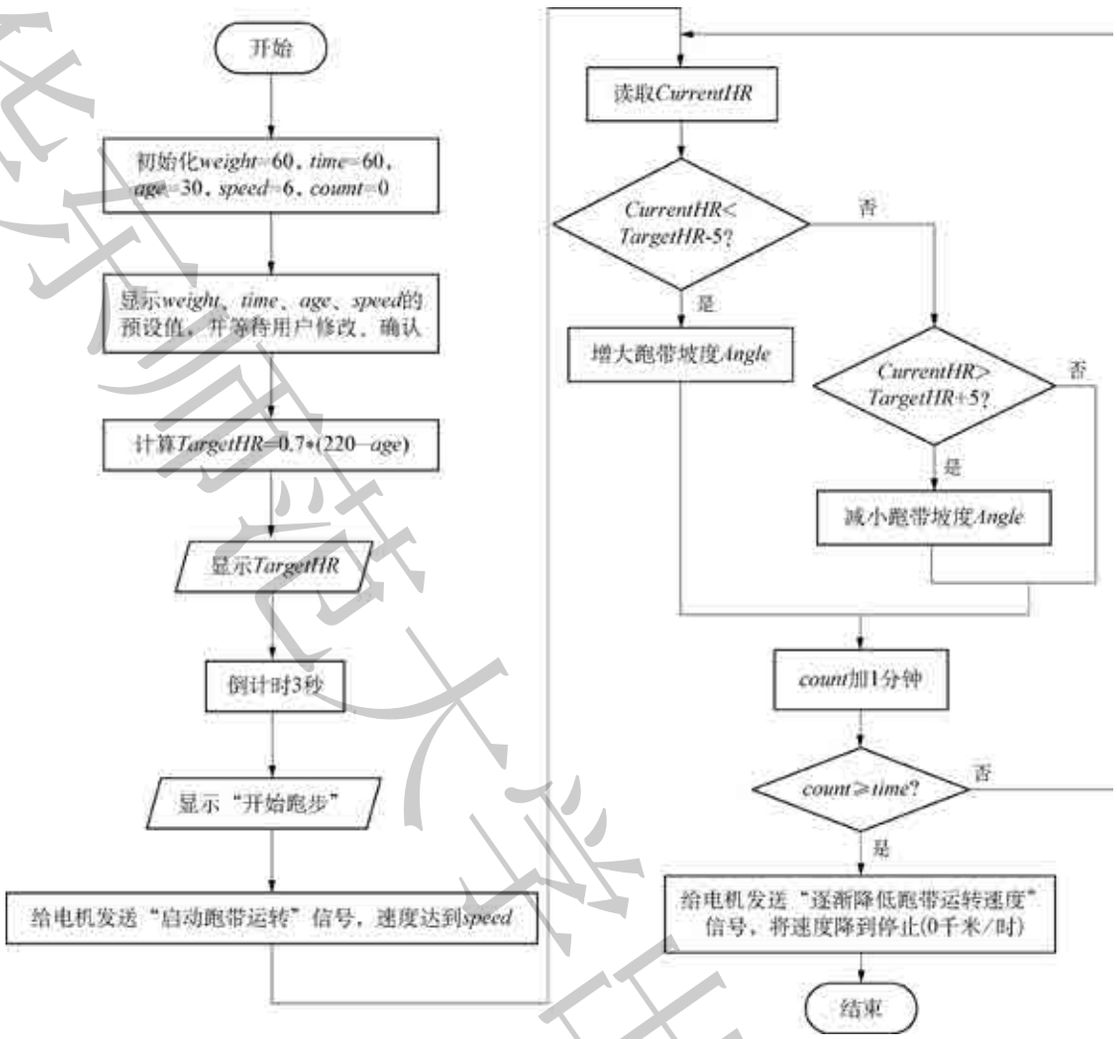


图 2.2 “心率跑”模式算法的流程图

3. 伪代码

伪代码是一种介于自然语言和计算机程序设计语言之间的算法描述语言。伪代码能够较容易地被转换成程序设计语言。虽然流程图描述算法要比自然语言描述算法清晰直观，但如果需要能够快速转换成计算机可以执行的语言，一般会采用伪代码的方式进行描述。例如：

```
weight = 60, time = 60, age = 30, speed = 6, count = 0
Output "weight = 60, time = 60, age = 30, speed = 6"
Input weight, time, age, speed
```

```
TargetHR = 0.7 * (220 - age)
Output TargetHR
倒计时 3 秒
Output "开始跑步"
给电机发送“启动跑带运转”信号，直至跑带运转速度达到 speed
Repeat
    读取当前心率值 CurrentHR
    If CurrentHR<TargetHR - 5
        增大跑带坡度 Angle
    ElseIf CurrentHR>TargetHR + 5
        减小跑带坡度 Angle
    End If
    count 加 1 分钟
Until count>= time
给电机发送“逐渐降低跑带运转速度”信号，直至跑带运转速度降到停止(0 千米/时)
```

图 2.3　顺序结构示意图

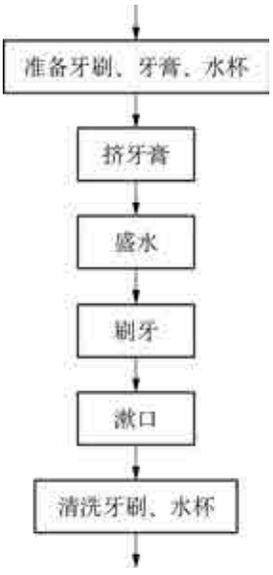


图 2.4　刷牙过程的顺序结构示意图

三、算法的基本控制结构

顺序结构、分支结构和循环结构是用来描述算法的三种基本控制结构。由这三种基本结构组成的算法结构清晰，易于正确性验证和纠错。

1. 顺序结构

顺序结构是一种自上而下，按先后顺序依次执行算法中各个步骤的结构，如图 2.3 所示。例如，某人的刷牙过程就可以用顺序结构表示，如图 2.4 所示。

2. 分支结构

分支结构也称为选择结构，是根据给定的条件进行判断，再依据判断结果的不同而执行不同操作的一种结构。分支结构流程图中一定会有判断框，当满足条件时执行一个分支，不满足条件时执行另一个分支。分支结构的流程图表示方法如图 2.5 所示。例如，挤牙膏时会判断是否有足够的牙膏，就可以使用分支结构来表示，如图 2.6 所示。

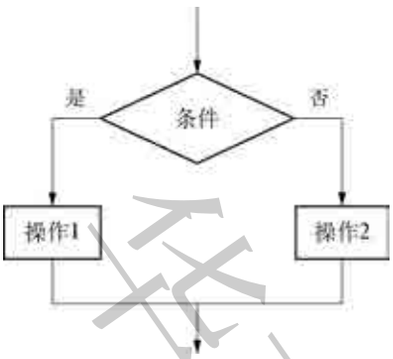


图 2.5　分支结构示意图

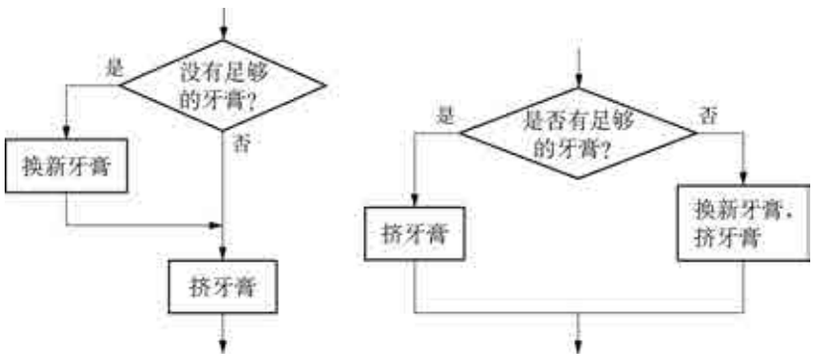


图 2.6　挤牙膏过程的两种分支结构示意图

3. 循环结构

循环结构也称为重复结构，是一种根据条件重复执行某一部分操作的结构，其中重复执行的这部分操作也称为循环体。有两种典型的循环结构：当循环（如图 2.7 甲所示）和直到循环（如图 2.7 乙所示）。

当循环先判断循环条件，后执行循环体。当判断框中的条件为“是”时，执行循环结构中的循环体，再根据条件判断是否需要继续执行循环体，直到条件为“否”，结束循环。直到循环先执行循环体，再判断循环条件是否成立。先执行一次循环体，然后判断条件，当条件为“否”时，返回重新执行循环体，再判断条件，直到条件为“是”时结束循环。

例如，在刷牙时需要判断是否刷了 100 下，来决定是否要继续在口腔中移动牙刷，这就需要使用循环结构来表示，如图 2.8 所示。

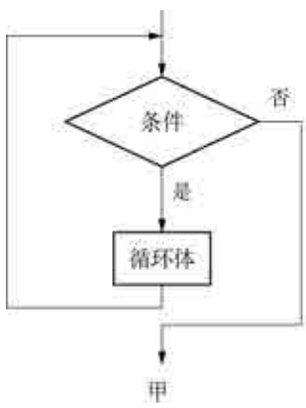


图 2.7　循环结构示意图

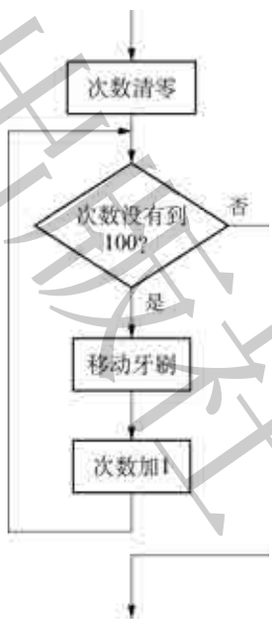


图 2.8　刷牙过程的循环结构示意图

项 目 实 践

很多智能跑步机都提供了“坡度跑”模式供选择，“坡度跑”也就是指跑步机会自动调节跑带的坡度，为用户模拟爬坡的情况。为了能够让用户得到科学、安全的跑步体验，在“坡度跑”模式中也需要考虑合理的热身过程、锻炼过程和调整过程。

1 通过查询了解“坡度跑”模式，分析智能跑步机“坡度跑”模式的实现过程。

2 如果请你来实现智能跑步机的“坡度跑”模式，该算法应如何设计？

3 请采用流程图的方式来描述“坡度跑”模式，想一想该算法需要包含哪些基本控制结构？

图 2.9 编程解决问题的一般过程

四、编程解决问题的过程

在生活中，我们经常会遇到各式各样的问题需要解决，而随着计算机技术的发展，计算机在问题解决中已经成为越来越重要的角色。人们从一开始依赖计算机的高速运算能力，到目前无处不在的计算机应用，计算机也在各类问题解决的过程中，从单一的计算发展成了全过程的参与。

当面对特定的问题时，往往需要根据设计的算法编写特定的程序来解决问题。编程解决问题的一般过程包括：抽象与建模、设计算法、编写程序、调试运行这四个步骤，如图 2.9 所示。根据问题的需要，可以反复修正和执行这四个步骤，直到问题得到有效解决。

1. 抽象与建模

解决问题前，需要对问题进行深入分析，明确问题的需求。然后分析问题的求解目标、约束条件等，将问题抽象化、模型化。抽象与建模是指从现实问题出发，忽略非本质的细节，提炼出核心要素，将具体的问题抽象化，并将其描述成为一个明确已知条件、约束条件和求解目标的问题，再用数学符号来描述计算模型。人们每天都在不自觉地使用抽象和建模，例如地铁线路图就是一种抽象，在图中并没有显示所有的细节，而仅仅是提炼了路线、站点和换乘的信息。当人们在计算地铁乘车费用时，则是通过已建立的计算模型，针对不同起点和终点进行计算，获得对应的票价信息。

以计算中国农历年份为例，在中国古代的历法中，甲、乙、丙、丁、戊、己、庚、辛、壬、癸被称为“十天干”，子、丑、寅、卯、辰、巳、午、未、申、

酉、戌、亥叫作“十二地支”。十天干和十二地支依次相配，组成六十个基本单位，两者按固定的顺序互相配合，组成了干支纪年法。基于天干地支序列，如果已知 2000 年是庚辰年，则对于输入的公历年份 2019 年，可以通过以下方法来推算其农历年份（天干地支列表为循环推算，即“癸”的后一个天干为“甲”，“亥”的后一个地支为“子”）：

天干：用 2019 减去 2000，得到的差除以 10 取余后得到 9，然后天干从“庚”向后推 9 位为“己”；

地支：用 2019 减去 2000，得到的差除以 12 取余后得到 7，然后地支从“辰”向后推 7 位为“亥”。

最后，可以推算出公历 2019 年是农历己亥年。

因此，本问题可以抽象为已知天干、地支序列和对照的年份 2000 庚辰年 *basey*，对于输入的公历年份 *year*，求出其对应的农历年份。该问题的计算模型如下：

天干：用 *year* 减去 *basey*，得到的差除以 10 取余后得到余数，然后根据余数的数值，在已知天干列表中从“庚”向后推算相应位数；

地支：用 *year* 减去 *basey*，得到的差除以 12 取余后得到余数，然后根据余数的数值，在已知地支列表中从“辰”向后推算相应位数。

2. 设计算法

在中国农历年份的计算中，我们可以选择现有的软件进行查询，也可以自己设计算法、编写程序来完成对用户输入年份的天干地支显示。针对问题分析的结果，设计一个对应求解的算法，其关键步骤如下：

① 设定十天干、十二地支序列；
② 设定对照年份 2000 年及其所对应的天干和地支；
③ 输入公历年份 *year*；
④ 根据计算模型计算该公历年份所对应的天干；
⑤ 根据计算模型计算该公历年份所对应的地支；
⑥ 输出该公历年份对应的农历年份。

3. 编写程序

确定算法后，就可以使用计算机编程实现了。编写程序就是选择

合适的计算机程序设计语言按照算法来实现问题求解。程序是一组计算机能识别和运行的指令，是计算机执行算法的具体步骤的实现，计算机通过运行这些指令来完成预期的任务。根据中国农历年份的求解算法，可以使用多种不同的程序设计语言来编程实现，以下为使用 Python 语言实现的程序。

```
tian = ["甲","乙","丙","丁","戊","己","庚","辛","壬","癸"]
di = ["子","丑","寅","卯","辰","巳","午","未","申","酉","戌","亥"]

basey = 2000
basetian = 6
basedi = 4
year = int(input("请输入年份："))
print(tian[((basetian + (year - basey))%10)%10], sep = "", end = "")
print(di[((basedi + (year - basey))%12)%12])
```

4. 调试运行

编写完成的程序需要进行调试运行，以验证所编写的程序是否正确。在这个阶段，不仅在发现错误时需进行修改，还要对运行结果进行分析和验证。根据调试结果不同，可能还需要重复前面的几个阶段，进行问题的分析、算法的优化和程序的重新编写。如图 2.10 所示是计算中国农历年份程序的运行结果。

```
请输入年份：2019
己亥
>>>
```

图 2.10　计算中国农历年份程序的运行结果

作业练习

智能服药系统也是家庭健康系统中重要的一部分，可以针对“忘记服药”“不按时服药”“重复服药”等多种问题进行监测和提醒，现在请你设计一款“智能药盒”。

1　请描述你设计的“智能药盒”的功能。

2　针对某个功能进行算法设计，并用流程图方式进行描述。

3　智能药盒和大数据、物联网结合，又可以增加哪些功能，为生活提供哪些便利？

知 识 延 伸

算法的效率

解决问题的算法往往不止一个，通常可以从时间和空间两个角度来对算法的效率进行评价。使用的指标分别为时间复杂度和空间复杂度。

1.时间复杂度：用于描述算法运行所需要的时间开销，一般采用算法中基本语句的执行次数进行度量。例如，要将一张纸等分成 16个大小相等的格子，可以有多种算法。一种算法是以每次画一个格子的方式，画 16个格子将纸等分。如果画一个格子记为一次操作的话，那就需要 16次。另一种算法是将纸折起来、再折、再折、再折，经过 4次操作后，打开就可以得到 16个格子了。显然，第二种算法所需要的执行时间比第一种算法要少，也称为第二种算法的时间复杂度较小。

2.空间复杂度：用于描述算法运行所需要的存储空间，一般主要考虑临时占用的空间大小。

第二节　程序设计语言基本知识

当人们完成问题的抽象与建模，并通过各种方法和设备采集了大量的数据、设计了解决问题的有效算法后，还需要相应的计算机程序来实现这些算法。只有通过编写程序，给计算机下达指令，才能处理数据，得到有价值的信息。因此，就需要选择合适的程序设计语言，根据其语法规则编写程序，最终在计算机上实现自动运行。

体　验　思　考

很多智能跑步机会内置称重传感器，可以快速获取用户的体重数据。体重数据被传送到远程服务器上之后，用户可以使用配套的移动应用程序再次读取该数据。移动应用程序除了能够显示体重数据之外，还能够同时显示身体质量指数 (body mass index，BMI)和体型描述，如图 2.11所示。

思考：

1. 程序是如何实现 BMI 指数的计算的？
2. 程序是如何根据 BMI 指数显示用户的胖瘦程度的？

图 2.11　移动应用程序上显示的体重等数据

一、Python语言基础

人们要想借助计算机快速准确地得到结果，完成某些特定的功能，就要为计算机编写相应的程序。

程序设计语言是人与计算机进行交互的语言。为了使用计算机解决问题，人们用程序设计语言编写程序，让计算机运行后完成预期的任务。程序是一组操作指令或语句序列，是计算机执行算法的操作步骤。

1. 程序设计语言

程序设计语言经历了从机器语言到高级语言的发展过程。

（1） 机器语言

机器语言是一种用二进制代码标识的、计算机能够直接识别和执行的机器指令的集合。机器语言具有灵活、直接执行和速度快等特点。以完成“9 + 11”的计算为例，用某种类型计算机适用的机器语言编写的程序如表 2.2 所示。

表 2.2 机器语言示例

语　句	说　明
10110000 00001001 00101100 00001011 11110100	将 9放入累加器 A 11与累加器 A中的值相加，结果仍放在累加器 A中 停止操作

一般，一条指令就是机器语言的一条语句。指令包括操作码和地址码，其中操作码指明了指令的操作性质及功能，地址码则给出了操作数或操作数的地址。

（2） 汇编语言

用机器语言编写程序非常困难，因此产生了汇编语言，也称为符号语言。在汇编语言中，用类似英语缩略词的语言代替机器指令的操作码，用地址符号或标号代替指令或操作数的地址，运行时再转换为机器语言。以完成“9 + 11”的计算为例，用汇编语言编写的程序如表 2.3 所示。

表 2.3 汇编语言示例

语　句	说　明

MOV AL,9
ADD AL,11
HLT
	将 9放入累加器 A，其中 AL表示累加器 A
	11与累加器 A中的值相加，结果仍放在累加器 A中
	停止操作

（3） 高级语言

由于汇编语言依赖于硬件体系，且助记符量大难记，于是人们又

发明了更加易用的高级语言。高级语言是以人类的日常语言为基础的一种编程语言，使用一般人易于接受的文字来表示，从而使程序编写更容易，有较高的可读性。目前，常用的高级语言有 C、C++、Java、Python 等。以完成“9 + 11”的计算为例，用 Python 语言编写的程序如表 2.4 所示。

表 2.4　Python语言示例

语　　句	说　　明
print(9+11)	计算 9+11并输出

高级语言和汇编语言一样，编写的程序也不能直接被计算机执行，必须经过转换后才能被执行。

2. Python 常用数据类型

为了能够处理日常生活中各式各样的数据，程序设计语言提供了多种数据类型。常见的 Python 数据类型如表 2.5 所示。

表 2.5　常见 Python数据类型

数据类型	类型标识符	类型说明
整型	int	Python 中的整数可以是任意大小，如 51、-67、0等
浮点型	fbat	由整数部分与小数部分组成，也可以使用科学记数法表示，如 3.076、-2e3等
字符串型	str	用单引号 (' ') 或双引号 (" ") 括起来的一串字符，如'Hello'或"上海""\n"等
布尔型	bool	只存在两种值：真 (True) 或假 (False)，常用于逻辑判断

在程序设计过程中，可以通过强制类型转换操作，把数据从一种类型强制转换成另一种类型。Python 语言中可用于数据类型转换的内置函数如表 2.6 所示。

表 2.6　Python语言的数据类型转换函数

数据类型转换函数	说　　明
float(3)	将整型数据 3转换为一个浮点型数据，为 3.0
int(3.6)	将浮点型数据 3.6转换为一个整型数据，为 3
str(3)	将整型数据 3转换为字符串型数据，为'3'或"3"

3. Python 中的常量、变量与赋值符

常量是直接给定的，指在程序运行过程中不变的量，如常用的数学常数 π 就是一个常量。

变量指程序运行过程中可以被改变的量。在程序运行过程中，变量被存储在内存中，可以通过变量名进行访问。变量命名时，需要遵守命名规则：由大小写英文字母、数字或下划线组成，以英文字母或下划线为首字符，长度不限，不能与 Python 保留字同名，大小写敏感。变量的数据类型由被赋值的数据对象的类型决定。

“=”为 Python 中的赋值符，其作用是把赋值号右边表达式的计算结果存储到赋值号左边指定的变量中。例如：c = 3，就是将 3 赋值给变量 *c*。

4. Python 中的运算符与表达式

Python 中的表达式是操作数（参与运算的数据）、变量和运算符的组合，是用来描述数据的计算过程，或描述对于某种情况下所遇到的条件判断，单独一个操作数或变量都可以看作是表达式。常用的运算符有算术运算符、关系运算符、逻辑运算符等。

（1）算术运算符

算术运算符主要用于算术运算，运算的结果为整型或浮点型。常见的算术运算符如表 2.7 所示，运算符有优先级，最高级别表示为 1，数字越大，优先级越低。

表 2.7　Python语言的常见算术运算符

运算符	描述	优先级	实　　例
+	加法	3	23+45，运算结果为 68
-	减法	3	78.4-34.2，运算结果为 44.2

（续 表）

运算符	描述	优先级	实　例
*	乘法	2	5.3*20，运算结果为 106.0
/	除法	2	2/5，运算结果为 0.4
%	取模	2	17% 6，运算结果为 5
**	幂	1	3**3，运算结果为 27
//	整除	2	25//7，运算结果为 3

（2） 关系运算符

关系运算符也称为比较运算符，用于比较两个值的大小，其运算结果为布尔值真或假，常见的关系运算符如表 2.8 所示。

表 2.8　Python语言的常见关系运算符

运算符	描述	实例(a=10 ,b=20)
==	等于	a==b，返回假
>	大于	a>b，返回假
<	小于	a<b，返回真
>=	大于等于	a>=b，返回假
<=	小于等于	a<=b，返回真
!=	不等于	a!=b，返回真

（3） 逻辑运算符

逻辑运算符用于对关系表达式或布尔值进行逻辑运算，运算结果为布尔值真或假。在实际问题中，对一些逻辑复杂的条件，需要用多个关系表达式组合起来表示。常见的逻辑运算符如表 2.9 所示。

表 2.9　Python语言的常见逻辑运算符

运算符	描述	优先级	实　例
not	非	1	not(a==b)，如果 a和 b相等，则返回假，否则返回真
and	与	2	a and b，只有当 a和 b都为真时，结果才为真
or	或	3	a or b，只要 a和 b中有一个为真时，结果就为真

如果多个运算符出现在同一个表达式中，则需要按照优先级确定运算顺序。优先级高的运算符先运算，优先级相同的从左向右依次运

算。括号运算的优先级最高，应先计算括号内的表达式；三种运算符的优先级为：算术运算符＞关系运算符＞逻辑运算符。

5. Python 中的内置函数与模块导入

内置函数是已经预定义并且已经实现的、可以供用户直接调用的函数，很多高级语言都有内置函数。函数可以直接通过“函数名(参数列表)”的方式调用，多个参数值之间一般以逗号分隔。例如，abs(x)为 Python 提供的求取绝对值的内置函数，abs(-1)的返回值为 1；round(a,b)为求取指定位数的小数的内置函数，round(3.1415926,2)的返回值为 3.14。

Python 语言中的模块是一个程序文件，在使用之前通过“import 模块名”的方式导入。例如，通过“import math”导入数学模块后，在程序中就可以直接调用该模块中定义的函数了，使用 factorial()函数输出阶乘的程序代码如下：

```
import math
print(math.factorial(6))
```

6. Python 中的字符串

字符串主要用于存储和表示文本，是 Python 中最常用的数据类型之一。计算机中文本的最基本单位是字符，包括可见字符和不可见字符，其中可见字符有英文大小写字母、数字字符、标点符号和一些常见符号；不可见字符包括回车、空格等。

Python 语言提供了对字符串类型数据的一些通用操作，包括连接、复制等，如表 2.10 所示。

表 2.10　Python语言中字符串类型数据的通用操作

操作	描述	实例 (str1= "Hello",str2= "Python")
x+y	连接两个字符串 x和 y	str1+" "+str2,返回 "Hello Python"
x*n	复制 n次字符串 x	str1*3,返回 "HelloHelloHello"
len(x)	返回字符串 x的长度	len(str1),返回 5

7. Python 中的列表

列表是 Python 中常见的一种数据形式，它可以把大量的数据放在一起，对其进行集中处理。列表是以“[]”包围的数据集合，不同成员间以“,”分隔。列表中可以包含任何数据类型，也可以包含另一个列表。我们可以通过序号来访问列表中的成员，例如有列表：tian = ["甲","乙","丙","丁","戊","己","庚","辛","壬","癸"]，其中 *tian*[0]为“甲”，*tian*[2]为“丙”。

Python 语言对列表提供了一些与字符串相似的通用操作。此外，还提供了一些常用的列表方法，如表 2.11 所示。

表 2.11　Python语言中的常用列表方法

方法	描　　述
list.append(x)	在列表尾部追加成员 x
list.insert(i,x)	向列表中指定位置 i 插入 x
list.remove(x)	删除列表中的指定成员（有多个则只删除第一个，指定成员不存在则报错）

在交互环境下对列表进行操作的示例代码如下：

```
>>>di = ["子","丑","寅","卯","辰","巳","午","未","申","酉","戌"]
>>>di
['子','丑','寅','卯','辰','巳','午','未','申','酉','戌']
>>>di.append("亥")
>>>di
['子','丑','寅','卯','辰','巳','午','未','申','酉','戌','亥']
>>>di.insert(0,"亥")
>>>di
['亥','子','丑','寅','卯','辰','巳','午','未','申','酉','戌','亥']
>>>di.remove("亥")
>>>di
['子','丑','寅','卯','辰','巳','午','未','申','酉','戌','亥']
```

二、顺序结构的 Python 实现

探 究 活 动

当手机和智能跑步机通过无线网络连接后，移动应用程序上即可获得跑步过程中的各类数据。例如，移动应用程序上除了能显示体重值，还可以显示 BMI 指数。BMI 指数是用体重千克数除以身高米数的平方得出的数值，是目前国际上常用的衡量人体胖瘦程度以及评定身体是否健康的参考标准之一。BMI 指数可以通过以下数学公式计算得出：

$$BMI=\frac{体重(千克)}{身高^2(米^2)}$$

例如，一个人的身高为 1.75米，体重为 68千克，他的 BMI 指数计算如下：

$$BMI=\frac{68}{1.75^2}=22.2$$

根据这个公式，我们可以使用纸笔、计算器等工具进行 BMI 指数的计算，但这些都需要根据不同的身高体重值来重复进行手动计算，而程序只需要根据传感器实时测得的用户数据，就可以进行实时、自动计算并输出结果。

1 讨论一下，要通过编程解决 BMI 指数的计算，需要哪些步骤？

2 要编写程序，可能需要用到哪些数据类型、运算符和函数等？

顺序结构的程序设计简单，只要按照解决问题的顺序写出相应的语句即可，它的执行顺序是自上而下，依次执行，直到结束。常见的顺序结构语句有输入语句、输出语句和赋值语句。

例如，要根据不同的身高和体重值计算 BMI 指数，就可以使用顺序结构来编程实现。

1. 抽象与建模

计算 BMI 指数，需要用到身高和体重值，因此程序需要通过用户输入的方式获取不同的身高、体重值，计算后输出 BMI 指数。其中，用 $height$ 表示身高，$weight$ 表示体重，bmi 表示 BMI 指数。

输入：$height$、$weight$；

输出：bmi；

计算模型：对于不同的输入，可以通过 $bmi=\frac{weight(千克)}{height^2(米^2)}$ 公式进行计算。

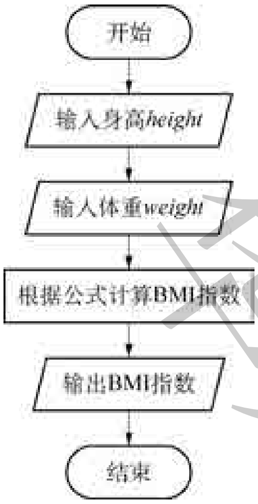


图 2.12　计算 BM I指数算法流程图

2. 设计算法

经过分析，需要先输入身高和体重值，然后根据公式计算并输出 BMI 指数的算法描述如下，其流程图如图 2. 12 所示。

① 输入身高 *height*；
② 输入体重 *weight*；
③ 根据公式计算 BMI 指数；
④ 输出 BMI 指数.

3. 编写程序

程序设计是能够将算法转换成计算机可以处理的指令的重要步骤。为了提高程序代码的可读性，方便维护，在程序设计的过程中，需要加入注释。Python 中的注释从“＃”字符开始，该行代码“＃”字符之后所有的内容都是注释。用来实现根据不同的输入值，计算相应的 BMI 指数的 Python 程序如下：

```
height = float(input("请输入身高(m):"))      #输入身高 height
weight = float(input("请输入体重(kg):"))     #输入体重 weight
bmi = weight/height ** 2                     #根据公示计算 BMI 指数
print("BMI = ", bmi)                         #输出 BMI 指数
```

该段 Python 程序中使用了输入输出语句、变量定义、运算符、表达式和类型转换。其中，输入输出是用户与程序进行交互的主要途径。通过输入语句，程序能够获取运行所需要的原始数据；通过输出语句，程序能够将数据处理结果告知用户。

Python 中使用 input()函数接收用户的输入，当用户输入程序所需的数据时，会以字符串形式返回。在输入时也可给出提示信息，例如：height = float(input("请输入身高(m)："))，程序将在屏幕上显示“请输入身高(m)：”信息，并等待用户输入；程序接收用户输入的数据后，将输入的数据转换为浮点型数据，保存在变量 *height* 中。

Python 中使用 print()函数显示一项或多项程序处理的结果，中间用逗号分隔。例如：print("Hello，world!")，程序将在屏幕上输出“Hello，world!”信息。再如：print("BMI = ",bmi)，程序将在屏幕上输出“BMI = 22. 20408163265306”信息。

4. 调试运行

在 Python 运行环境中对程序进行调试运行，程序会根据不同的身高、体重值，相应地输出计算后的 BMI 指数。

```
请输入身高(m)：1.75
请输入体重(kg)：68
BMI= 22.20408163265306
```

项 目 实 践

许多移动应用程序除了可以显示 BM 指数外，还能够显示运动过程中消耗的卡路里数，请查询卡路里消耗的计算方式，完成计算卡路里消耗的程序编写，并填写表 2.12。

表 2.12 “计算卡路里消耗”的分析与实现

<table>
<tr><td rowspan="3">抽象与建模</td><td>输入</td><td colspan="3"></td></tr>
<tr><td>输出</td><td colspan="3"></td></tr>
<tr><td>计算模型</td><td colspan="3"></td></tr>
<tr><td rowspan="2">设计算法</td><td colspan="2">流程图</td><td rowspan="2">编写程序</td><td colspan="2">Python程序</td></tr>
<tr><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>调试运行情况记录</td><td colspan="5"></td></tr>
</table>

三、分支结构的 Python 实现

顺序结构的程序能严格按照语句出现的先后顺序解决计算、输

入、输出等问题，但对于要做判断和选择的问题而言，就需要使用分支结构，依据一定的条件选择执行路径。

探 究 活 动

虽然 BMI 指数可以直接通过身高、体重值计算得出，但这个指数代表的具体含义并不是大多数人所熟悉的。例如，身高为 1.75米，体重为 68千克的人，他的 BMI 指数为 22.2到底代表什么，大家并不清楚。世界卫生组织（WHO）的判断标准如表 2.13所示。

表 2.13 BMI 指数参考标准表

描述	WHO 标准	描述	WHO 标准
偏瘦	<18.5	肥胖	30~34.9
正常	18.5~24.9	重度肥胖	35~39.9
偏胖	25~29.9	极重度肥胖	≥40

请参考表 2.13，设计算法，实现对于计算得出的任意 BMI 指数显示相应的人体胖瘦程度的描述。

在分支结构的程序设计中，程序要能根据是否满足条件来执行不同的语句。Python 语言中使用 if 语句实现分支结构，主要包括三种基本格式，如表 2.14 所示。

表 2.14 Python语言的三种分支结构基本格式

分支类型	基本格式
单分支语句	if条件表达式： 　语句块
双分支语句	if条件表达式： 　语句块 1 else： 　语句块 2
多分支语句	if条件表达式 1： 　语句块 1 elif条件表达式 2： 　语句块 2 …… else： 　语句块 *n*

单分支语句中，if 语句首先判断条件表达式，结果为真，则执行语句块中的语句序列；结果为假，则不执行任何语句。例如，求 x 绝对值的语句如下：

```
if x<0:
    x = -x
```

Python 程序要求代码全部使用缩进来分层，否则将导致程序错误，无法运行。Python 编程规范中指出：缩进最好采用空格形式，每一层向右缩进 4 个空格，在同一段代码中不能混用 Tab 键和空格键，如表 2.15 所示。

表 2.15　缩进对 Python程序设计语言的影响

	Python语言	
程序	if a>b: 　print(">")	if a>b: print(">")
说明	正确，可读性好	错误

双分支语句中，用 if…else 语句实现，如果条件表达式结果为真，则执行语句块 1 中的语句序列；如果结果为假，则执行语句块 2 中的语句序列。例如，判断 x 的奇偶性的语句如下：

```
if x%2==0:
    print("x 为偶数")
else:
    print("x 为奇数")
```

多分支语句中，用 if…elif…else 语句实现。程序首先判断 if 语句的条件表达式，如果结果为真，则执行语句块 1 中的语句序列；如果结果为假，则继续判断 elif 语句的条件表达式，如果条件表达式结果为真，则执行这个 elif 对应的语句块中的语句序列，如果结果为假，则继续判断下一个 elif 语句的条件表达式。依此类推，如果所有的条件表达式结果均为假，则执行 else 后的语句块。例如，根据气温判断高温预警信号级别的语句如下，其对应的流程图如图 2.13 所示。

```
if temperature>=40:
    print("高温红色预警")
elif temperature>=37:
    print("高温橙色预警")
elif temperature>=35:
    print("高温黄色预警")
else:
    print("请注意防暑降温")
```

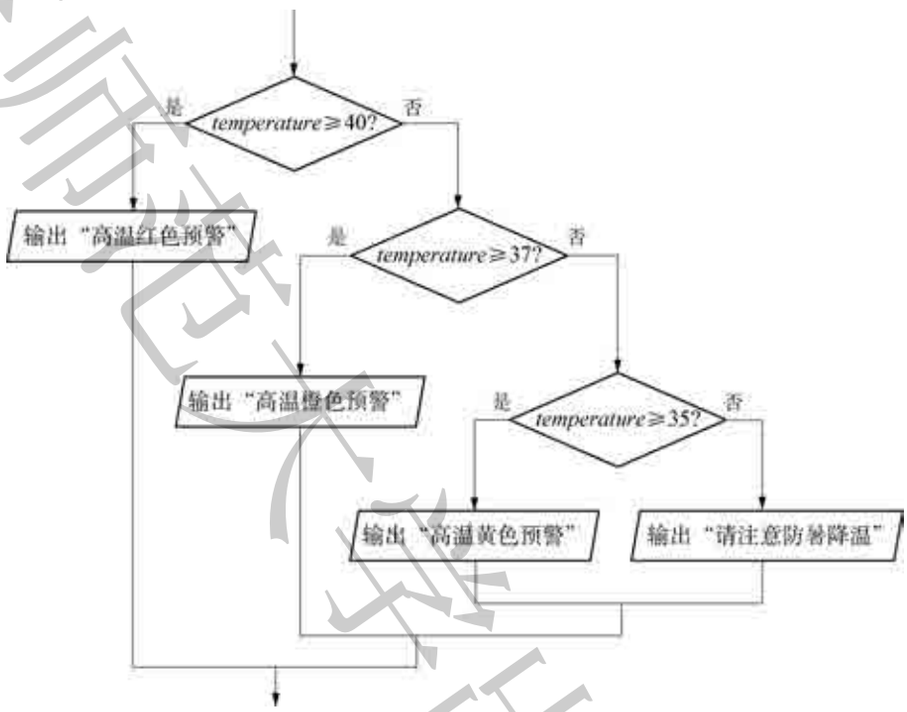


图 2.13　多分支流程图

例如，在计算和显示 BMI 指数时，一般会选择文字、颜色或图形等不同的表现形式给用户更直观的提示。要实现这个功能，需要使用分支结构对 BMI 指数进行范围判定，并相应地显示“正常”或“需注意”的提示。

1. 抽象与建模

要通过使用文字描述的方式显示 BMI 指数代表的含义，首先要计算出 BMI 指数，然后与 BMI 指数标准范围值进行比较，并根据结果显示相应的文字描述，如表 2.16 所示。

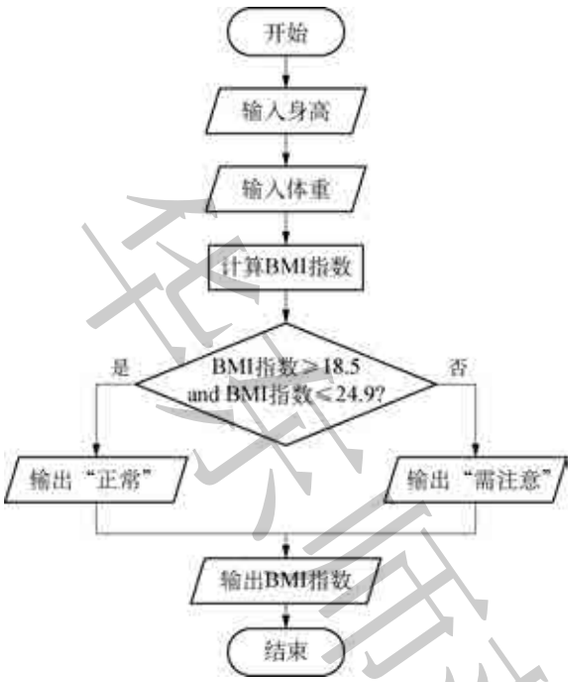


图 2.14　BMI 指数判定算法流程图

表 2.16　BMI 指数标准对照

描述	WHO 标准
需注意	<18.5
正常	18.5~24.9
需注意	>25

2. 设计算法

由于需要将计算得出的 BMI 指数与标准范围值进行比较，如果比较结果为真，则输出“正常”；如果比较结果为假，则输出“需注意”。因此，该算法就需要采用分支结构，具体算法描述如下，其流程图如图 2.14 所示。

① 输入身高 *height*；
② 输入体重 *weight*；
③ 计算 BMI 指数，保留一位小数；
④ 判定 BMI 指数是否大于等于 18.5 且小于等于 24.9，如果为真，则输出“正常”；如果为假，则输出“需注意”；
⑤ 输出 BMI 指数。

3. 编写程序

根据算法，用 Python 程序设计语言编写程序如下：

```
import math                                    #导入 math 库

height = float(input("请输入身高(m):"))
weight = float(input("请输入体重(kg):"))
bmi = round(weight/(math.pow(height,2)),1)
#使用内置函数 round 保留一位小数

if bmi>=18.5 and bmi<=24.9:                    #正常
    print("正常")
```

```
else:                                    #需注意
    print("需注意")

print("BMI=",bmi)                        #输出 BMI 指数
```

```
请输入身高(m):1.75
请输入体重(kg):75
正常
BMI =  24.5
>>>
请输入身高(m):1.70
请输入体重(kg):50
需注意
BMI =  17.3
>>>
```

图 2.15　程序运行结果

4. 调试运行

将程序在 Python 环境中进行调试运行，程序会更根据不同的身高、体重值，相应地给出不同的文字描述，并输出计算后的 BMI 指数，如图 2.15 所示。

探　究　活　动

上述程序中用两种文字描述来表示 BMI 指数的两个范围，请你使用 turtle 库再设计一个图形显示方案来表示这两种 BMI 指数范围（例如可使用两种不同颜色的形状等），并填写表 2.17。

表 2.17　“图形化显示 BMI 指数范围”的分析与实现

<table>
<tr><td rowspan="3">抽象与建模</td><td>输入</td><td colspan="3"></td></tr>
<tr><td>输出</td><td colspan="3"></td></tr>
<tr><td>计算模型</td><td colspan="3"></td></tr>
<tr><td rowspan="2">设计算法</td><td colspan="2">流程图</td><td rowspan="2">编写程序</td><td>Python程序</td></tr>
<tr><td colspan="2"></td><td></td></tr>
<tr><td>调试运行情况记录</td><td colspan="4"></td></tr>
</table>

技 术 支 持 turtle库介绍

turtle库是 Python语言中一个较常用的绘制图像函数库，类似一个小乌龟，从一个横轴为 x、纵轴为 y 的坐标系原点（该原点位于画布中心位置）(0,0)位置开始，根据一组函数指令的控制，在这个平面坐标系中移动，从而在它爬行的路径上绘制图形。常用的 turtle绘图基础知识包括画布和画笔两部分。

画布用于表示绘图区域，画布大小和颜色可以设置。例如：turtle.screensize(400,300,"black")，表示宽 400像素、长 300像素的黑色画布。画笔为画布上一个默认的箭头，常用的属性和命令如表 2.18所示。

表 2.18 画笔常用命令表

命 令	描 述
turtle.goto(x,y)	将画笔移动到坐标为 (x,y)的位置
turtle.pendown()	放下笔，移动时绘制图像
turtle.penup()	提起笔，移动时不绘制图像
turtle.circle(radius)	画圆，radius为正（负），表示画圆时圆心在画笔的左边（右边）
turtle.color(color1,color2)	同时设置 pencolor为 color1，fillcolor为 color2
turtle.begin_fill()	准备开始填充图形，与 turtle.end_fill()命令配合使用
turtle.end_fill()	填充完成，与 turtle.begin_fill()命令配合使用

可以通过 help()函数查看详细说明和用途，如 help(turtle.circle)。

四、循环结构的 Python实现

在不少实际问题中，有许多具有规律性的重复操作，因此在程序中就需要重复执行某些语句，如果仅使用顺序结构和分支结构是无法完全实现的，此时就需要使用循环结构了。

探 究 活 动

智能跑步机每次测量的实时体重数据都会被自动上传到远程服务器上进行存储。当配套的移动应用程序启动并连接到远程服务器上时，就能够自动读取并显示多日的体重数据和 BMI指数变化。

1 分析多日的体重数据分别用变量和列表来储存的不同之处。

2 设计算法，完成连续一周的体重数据和 BMI指数显示，并用流程图进行描述。

在循环结构中，通过判定条件，来确定部分语句是否需要被重复

执行，当条件结果为真时，循环体被重复执行，直到条件结果为假为止。Python 语言中，循环结构主要通过 while 语句和 for 语句实现。

while 语句的基本格式为：

```
while 条件表达式：
    语句块
```

while 语句中的条件表达式为循环条件，语句块为循环体，表达式后的冒号不能省略。例如，要输出给定列表中数字的平方数，可以用 while 语句实现，程序代码及运行结果如下：

```
alst = [1,2,3,4,5]
total = len(alst)
i = 0
while i<total:
    print(alst[i],"的平方是：",alst[i] * alst[i])
    i = i + 1
```

运行结果：

```
1 的平方是： 1
2 的平方是： 4
3 的平方是： 9
4 的平方是： 16
5 的平方是： 25
```

用 while 语句来实现循环时，要注意避免出现死循环。一般，在循环体中要有能够使循环条件表达式的值发生改变的语句，使循环条件表达式的值从真变为假，从而结束循环。while 语句既可以实现不确定次数的循环，也可以实现确定次数的循环。

for 语句的基本格式为：

```
for 循环变量 in 序列：
    语句块
```

for 循环语句每次从序列中取出一个元素赋值给循环变量，循环变量的初值为序列中的第一个元素值，依次访问完序列中所有元素后循环结束，序列后面的冒号不能省略。例如，要输出给定列表中数字的平方数，可以用 for 语句实现，程序代码及运行结果如下：

```
alst = [1,2,3,4,5]
for i in alst:
    print(i,"的平方是：",i * i)
```

运行结果：

```
1 的平方是： 1
2 的平方是： 4
3 的平方是： 9
4 的平方是： 16
5 的平方是： 25
```

for 语句中的循环次数由序列中元素的个数决定，通常用于确定循环次数的问题求解。对于不能确定循环次数的情况，则需要使用 while 语句。

序列还可以是 range()函数产生的列表或字符串。其中，range()函数返回一个等差整数序列，格式为：range(起始值，终值，步长)。该函数生成一个包含起始值但不包含终值的序列，起始值和步长可以省略，默认为 0 和 1，如表 2.19 所示。

表 2.19　range()函数示例

range()函数示例	描　　述
range(1,5,3)	生成起始值为 1，终值为 4，步长为 3的序列：[1,4]
range(1,5)	生成起始值为 1，终值为 4，步长为默认值 1的序列：[1,2,3,4]
range(5)	生成起始值为 0(没有设定起始值，默认为 0)，终值为 4，步长为默认值 1的序列：[0,1,2,3,4]
range(5,1,-1)	生成起始值为 5，终值为 2，步长为 -1的序列：[5,4,3,2]

上例中输出数字 1～5 的平方，我们还可以用如下程序实现同样的运行结果：

```
for i in range(1,6):
    print(i,"的平方是:",i*i)
```

```
运行结果:
1 的平方是: 1
2 的平方是: 4
3 的平方是: 9
4 的平方是: 16
5 的平方是: 25
```

如果要在移动应用程序中显示多日的体重和 BMI 指数变化，就可以使用列表和循环结构来实现。

1. 抽象与建模

要完成一周的体重和 BMI 指数对应显示，体重为每日测定获得，BMI 指数根据身高、体重值计算得出，此外还需要显示对应的星期。

输入：身高 $height$ 和一周的体重；

输出：一周的体重和对应的 BMI 指数 bmi(显示对应的星期)；

计算模型：一周 BMI 指数列表 = $\bigcup_{i=1}^{7} item_i$，$item_i = \frac{weight_i}{height^2}$。

其中，$weight_i$ 代表一周中每日的体重，$\bigcup$ 为并集符号。

2. 设计算法

要在移动应用程序中显示多日的体重变化，就需要重新读取这些数据，通过计算后，对应显示 BMI 指数和星期。具体算法如下：

① 输入身高 *height*；
② 读取体重 *weight*；
③ 根据公式计算 BMI 指数；
④ 重复执行步骤②，直到所有体重数据被读入并计算，循环结束；
⑤ 显示对应的星期、体重、BMI 指数。

3. 编写程序

由于需要重复读入体重数据来计算 BMI 指数，因此可以考虑使用一个循环结构通过读取每日的体重数据来计算 BMI 指数，然后保存在列表中。

```
import math

weekday = ["SUN","MON","TUE","WED","THU","FRI","SAT"]
height = float(input("请输入身高(m):"))
weight = [68.4,67.7,66.9,67.7,68.5,69.2,68.4]

bmi = []
for i in weight:
    bmi.append(round(i/height ** 2,1))

print(weekday)
print(weight)
print(bmi)
```

4. 调试运行

在 Python 环境中调试运行上述程序，结果显示如下：

```
请输入身高(m)：1.75
['SUN','MON','TUE','WED','THU','FRI','SAT']
[68.4,67.7,66.9,67.7,68.5,69.2,68.4]
[22.3,22.1,21.8,22.1,22.4,22.6,22.3]
```

项 目 实 践

请设计程序，实现同时显示消耗的卡路里数，并填写表 2.20。

表 2.20 “显示多日消耗的卡路里数”的分析与实现

抽象与建模	输入	
	输出	
	计算模型	
设计算法	流程图	
编写程序	Python程序	
调试运行情况记录		

作业练习

请根据学习的内容为用户设计一张反映一周体重变化的、更清晰的基本数据统计表。

如果要给出如图 2.16所示的“较上日比较结果”（如“比上次变化了 0.3斤”）的显示，应该如何编程实现？（1斤 = 0.5千克）

图 2.16　移动应用程序显示体重测量结果

知 识 延 伸　Python程序设计语言的发展

自 1989年诞生以来，Python的技术不断更迭，生态逐渐完善，加上互联网、大数据、人工智能浪潮的推波助澜，Python逐渐受到大众的青睐。

Python的诞生和崛起：1991年，Python的第一个版本正式诞生。2000年发布的 Python 2.0标志着其框架的基本确定。Python提供了丰富的 API和工具，方便用户能够轻松地使用其他语言来编写扩充模块，而且用 Python编写的功能模块也可以嵌入其他语言编写的程序中。Python有强大的标准库，同时支持第三方库的扩展应用。

在网络爬虫上的应用：Python自带的一些标准库对于实现基于网页的分析和操作比较简单，从而使得用户在实现网络爬虫应用时，只需要开发少量的程序，便能较快地完成任务。

在数据科学领域中的广泛使用：2008年发布的 Numpy、Scipy和 2009年发布的 Pandas库是广泛应用于数据分析与科学计算领域的重要工具，在数据计算、数据预处理和数据分析方面提供了强大的功能。

Python和人工智能的结合：Python有很多与人工智能相关的库和框架。其中，较常用的是 sklearn、PyTorch和 TensorFlow等，可用于机器学习、神经网络、深度学习的应用开发。

第三节　常用算法及其程序实现

当今社会，人们可以借助计算机来解决的问题越来越多，但由于问题类型和复杂程度的不同，很难有解决所有问题的统一算法。因此，只能针对某个或某类问题，通过分析研究，寻找解决问题的适当算法。

体　验　思　考

某些智能跑步机的配套移动应用程序除了能记录和显示运动过程中人体的各项身体数据，还可以提供丰富的跑步训练课程，让用户可以在科学的课程指导下，达到更好的锻炼效果，如图 2.17所示。

思考：

如果要在移动应用程序上显示用户跑步训练课程的报表信息，例如显示用户在一周内已经完成的训练项目和未完成的训练项目等信息，应该如何实现？如果时间跨度延长至一个月，又应该如何实现？

图 2.17　移动应用程序提供的跑步训练课程

一、枚举法

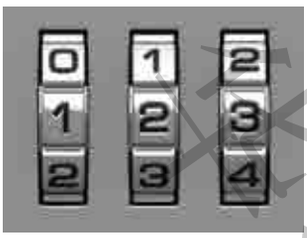


图 2.18　数字密码锁

在生活中，人们有时会遇到数字密码锁打不开的情况。遇到这种情况，大部分人会采用类似的解决方法，那就是逐个数字进行尝试，直到打开为止，如图 2.18 所示。这种逐一尝试的方法在算法中被称之为枚举法，它是处理问题常用的算法思想之一。

枚举法的基本原理是根据已知条件，在给定的范围内对所有可能的答案按某种顺序进行逐一列举和检验，从中找出那些符合要求的答案。即列举出所有可能的情况，然后逐个判断有哪些符合问题所要求的条件，从而得到问题的解答。其一般模式可以总结如下：

① 确定范围：问题所涉及的情况有哪些，情况的种数是否可以确定；

② 验证条件：这些情况需要满足什么条件才能成为问题的解答。

在生活中，我们经常会使用到枚举法。例如，在果农挑选苹果的时候，“待选苹果”为枚举范围，“苹果是否符合规格”则为验证条件，其算法的流程图描述如图 2.19 所示。

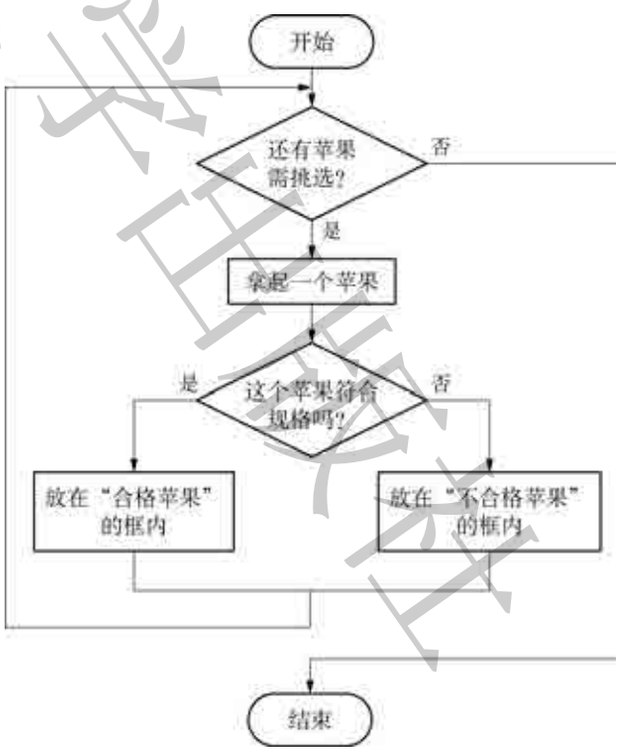


图 2.19　使用枚举法“挑选苹果”的流程图

在枚举过程中，应当尽可能缩小枚举范围，减少枚举次数，缩短求解时间，提高解决问题的效率。

在实际解决问题的过程中，如果需要枚举的范围较小，采用枚举法会比较直观；但当枚举范围比较大的时候，则会显得繁琐，需要花费大量的时间去逐一尝试。而能用计算机实现枚举法解决问题的关键是依赖于计算机运算快和自动化的特点。

通常对于解决同一个问题，可以有多个不同的算法，不同算法的效率也可能不同。例如，需要求解给定的两个正整数 m 和 n 的最大公约数。我们可以使用枚举法来求解，即逐一检验从 2 到 m 和 n 中较小数的范围中，最大的能够同时被 m 和 n 整除的数，则该数就是 m 和 n 的最大公约数；此外，也可以使用辗转相除法来求解。表 2.21 中提供了分别使用枚举法和辗转相除法来实现“求解最大公约数”的程序和运行结果。假设输入的两个数分别为 12345 和 54433，则从程序的运行结果可以看出使用枚举法需要执行“循环体”12344 次，而使用辗转相除法只需要执行循环体 11 次。因此，针对问题解决而选择合适的算法，对程序的效率乃至问题解决的效率都有很大的影响。

表 2.21　使用枚举法和辗转相除法实现“求解最大公约数”的程序

枚举法	辗转相除法

m=int(input("输入 m:"))
n=int(input("输入 n:"))
 	
if(m>n):　　# 如果 m 比 n 大
　m,n=n,m　　# 则交换两个变量的值
gcd=1
times=0　　# 设定循环计数变量初值
for i in range(2,m+1):　　# 枚举法求最大公约数
　　if(m% i==0 and n% i==0):
　　　gcd=i
　　times=times+1　　# 循环计数增加
 	
print(gcd)
print(times)
	m=int(input("输入 m:"))
	n=int(input("输入 n:"))
	
	if(m>n):　　# 如果 m 比 n 大
	　m,n=n,m　　# 则交换两个变量的值
	times=0　　# 设定循环计数变量初值
	while (n!=0):　　# 辗转相除法求最大公约数
	　　t=m% n
	　　m=n
	　　n=t
	　　times=times+1　　# 循环计数增加
	print(m)
	print(times)

输入 m:12345 输入 n:54433 1 12344	输入 m:12345 输入 n:54433 1 11

二、枚举法的程序实现

用程序设计语言实现枚举法时，需要列举出所有可能的情况，逐个判断有哪些情况符合问题所要求的条件，可以采用循环结构实现列举的过程，而其中判断有哪些情况符合问题所要求的条件则可以采用分支结构来实现。

探　究　活　动

一般，智能跑步机配套的移动应用程序会通过统计报表的形式给用户提供明确的跑步训练课程完成度信息，用户也可以选择日报表、周报表或月报表进行查看。假设用户设定的某跑步训练课程需要持续一周，每天需要完成固定的训练内容和时长，完成后的数据会上传到远程服务器上。

1　请设计算法，显示某用户在训练课程中完成了哪几项训练内容，未完成哪几项训练内容。

2　选用适当的方法描述算法。

根据跑步课程的设定，用户每天的跑步数据都会被实时传输到远程服务器上。当用户打开移动应用程序时，会通过网络对服务器上的数据进行读取，并且显示该用户的课程完成情况。

1. 抽象与建模

要显示用户的跑步课程完成情况，就需要对课程每日的训练内容以及完成后自动上传的已消耗卡路里数进行统计。

输入：跑步课程中的每日训练内容和对应训练内容消耗的卡路里数，这些信息由程序自动读取数据库中存储的数据，不需要用户输入；

输出：完成的项目数和未完成的项目名称；

计算模型：已完成项目数 $=\sum_{i=1}^{n} data_i$，

$$data_i=\begin{cases}0(\text{消耗的卡路里数}=0),\\1(\text{消耗的卡路里数}\neq 0);\end{cases}$$

未完成项目数 $=\bigcup_{i=1}^{n} item_i$，

$$item_i=\begin{cases}\text{空字符串}(\text{消耗的卡路里数}=0),\\\text{项目名称}(\text{消耗的卡路里数}\neq 0)。\end{cases}$$

其中，n 为该跑步课程共有多少项训练内容，消耗的卡路里数为大于等于零的整数，$\sum$ 为求和符号。

2. 设计算法

在进行统计时，可以使用枚举法来逐一列举并检测。根据枚举法的一般模式，确定范围和验证条件如下：

确定范围：用户一周的跑步课程训练内容；

验证条件：检测某训练内容消耗的卡路里数是否为 0，如不为 0，则表示已完成训练，并计数；如为 0，则表示该训练内容未完成，需要记录该训练内容。

算法描述如下：

① 读入一周的跑步课程训练内容；
② 读入对应课程训练内容消耗的卡路里数；
③ 初始设定已完成项目数 *finished* 为 0，未完成项目名称 *unfinished* 为空字符串；
④ 逐一列举一周的对应训练内容消耗的卡路里数；
⑤ 如果当前枚举的卡路里数为 0，则将对应的训练内容名称加入 *unfinished*，否则 *finished* 的计数加 1；
⑥ 输出完成项数目及未完成项目名称。

3. 编写程序

移动应用程序读取服务器上的数据后，可将用户在跑步课程中消耗的卡路里数存储在列表中，例如某用户一周训练课程对应消耗的卡路里数的初始列表值为[0,0,0,0,0,0,0]，训练内容则固定存储在另一个列表中，并且与消耗的卡路里数存储位置一一对应。以下显示的是某用户一周训练内容安排和每天对应消耗的卡路里数在列表中的存储情况。可以看出，该用户在“变速练习”训练中消耗了 600 千卡，总共完成了 6 项训练，但没有完成“快走”训练。

["低速低强度","变速练习","低速低强度","快走","低速低强度","坡度练习","低速低强度"]
[400,600,380,0,420,620,397]

根据设计的算法，用 Python 程序实现如下：

```
itemlist = ["低速低强度","变速练习","低速低强度","快走","低速低强度","坡度练习","低速低强度"]
datalist = [400,600,380,0,420,620,397]
finished = 0                    #记录完成了多少项训练内容
unfinished = ""                 #记录未完成的训练内容
for i in range(7):
    if datalist[i] == 0:
        unfinished = unfinished + itemlist[i] + " "
    else:
        finished = finished + 1
print("完成了",finished,"项")
print("未完成项目:",unfinished)
```

4. 调试运行

在 Python 环境中调试运行以上程序，结果显示如下：

```
完成了 6 项
未完成项目:快走
```

项 目 实 践

假设每项训练内容都有要求的卡路里消耗最低值，请参考以上程序进行编程，实现统计“该用户哪些训练项目未达标”，请将你的分析和程序填入表 2.2.2中。

例如，每项训练内容要求的卡路里消耗最低值存储在 mindata[380,580,380,450,380,650,380]中，从中可以看出“坡度练习”训练要求的卡路里消耗最低值为 650千卡。

请同学们分析一下本题使用枚举法实现的 Python程序，从程序结构的角度上总结枚举法一般需要包含哪几种基本控制结构。

表 2.22　算法分析及程序实现表

<table>
<tr><td rowspan="3">抽象与建模</td><td>输入</td><td colspan="3"></td></tr>
<tr><td>输出</td><td colspan="3"></td></tr>
<tr><td>计算模型</td><td colspan="3"></td></tr>
<tr><td rowspan="2">设计
算法</td><td colspan="2">流程图</td><td rowspan="2">编写
程序</td><td>Python程序</td></tr>
<tr><td colspan="2"></td><td></td></tr>
<tr><td>调试运行
情况记录</td><td colspan="4"></td></tr>
</table>

知　识　延　伸　排序和查找

排序算法，就是如何使得记录按照要求（升序或降序）排列的方法。与枚举法一样，排序算法早在计算机出现之前就已经被人们在实际生活中使用了。人们可以通过“看一眼”“扫一遍”的方式对两个数据快速分辨大小，因此可以在较短的时间内对有限的数据快速地进行排序。而计算机只有通过比较才能够分清两个数（在计算机中不仅仅指数值数据）的大小，从而在排序时要进行反复的比较和交换，但由于计算机具有高速运算的能力，因此面对大数据量时就比人显得更有优势。

选择排序的基本思想非常直接，每一次从序列的所有元素中先找到最小的，然后放到第一个位置。之后再看剩余元素中最小的，放到第二个位置……依此类推，就可以完成整个排序工作了。选择排序的关键是帮助固定位置找到合适的元素。

例如，待排序数据为“45 67 12 78 39 23”，使用选择排序算法的排序过程如表 2.23所示。

表 2.23 选择排序算法过程记录表

步骤	待排序数据	最小值	备注
第 1步	45 67 12 78 39 23	12	45和 12位置交换
第 2步	12 67 45 78 39 23	23	67和 23位置交换
第 3步	12 23 45 78 39 67	39	45和 39位置交换
第 4步	12 23 39 78 45 67	45	78和 45位置交换
第 5步	12 23 39 45 78 67	67	78和 67位置交换
第 6步	12 23 39 45 67 78		结束排序

除了选择排序，还有冒泡排序、插入排序、归并排序、快速排序、堆排序等许多排序算法，每一种排序算法都能按照一定的基本原理，完成对数据的排序。

查找也是生活中最常用的算法之一，指通过一定的方法找出与给定关键字相同的数据元素的过程。常用的查找算法有顺序查找和二分查找。顺序查找是一种最基本、最简单的查找算法。例如，使用顺序查找在手机通信录里查找名字叫“孙老师”的联系人信息，则需要从通信录的第一个联系人开始，逐个往下比较，直到找到名字叫“孙老师”的联系人为止。如果所有联系人都查找完了，仍旧找不到这个联系人，则表示查找不成功。二分查找也称为折半查找，其基本思想是当待查找数据有序时，首先找到待查找数据中的中间元素，将其值与关键字进行比较，若不等，则根据该值与关键字的比较结果来决定继续查找待查数据的前半部分或后半部分。例如，要在手机通信录中查找“孙老师”的联系信息是不会有人愿意从头开始一一比较的。手机中的联系人信息一般都会按拼音首字母进行排序，因此要查找“孙老师”的信息，就可以初步判定“ S”应该处于通信录的后部，点击通信录的相应位置，如果运气好，一次就能找到；即使没找到，也可以根据相对位置往前或往后翻阅通信录。这种算法和顺序查找相比，比较次数少，查找效率要高得多，但实施二分查找的先决条件就是待查范围的数据必须是有序排列的。

查找算法的效率取决于查找的次数，顺序查找的次数由所查找的元素在序列中的位置所决定。而二分查找的效率要高得多，但二分查找必须基于有序排列。因此，针对查找规模较小的无序数列，顺序查找也是一种常用的有效方法。

第三章

数据处理与应用

本章学习目标

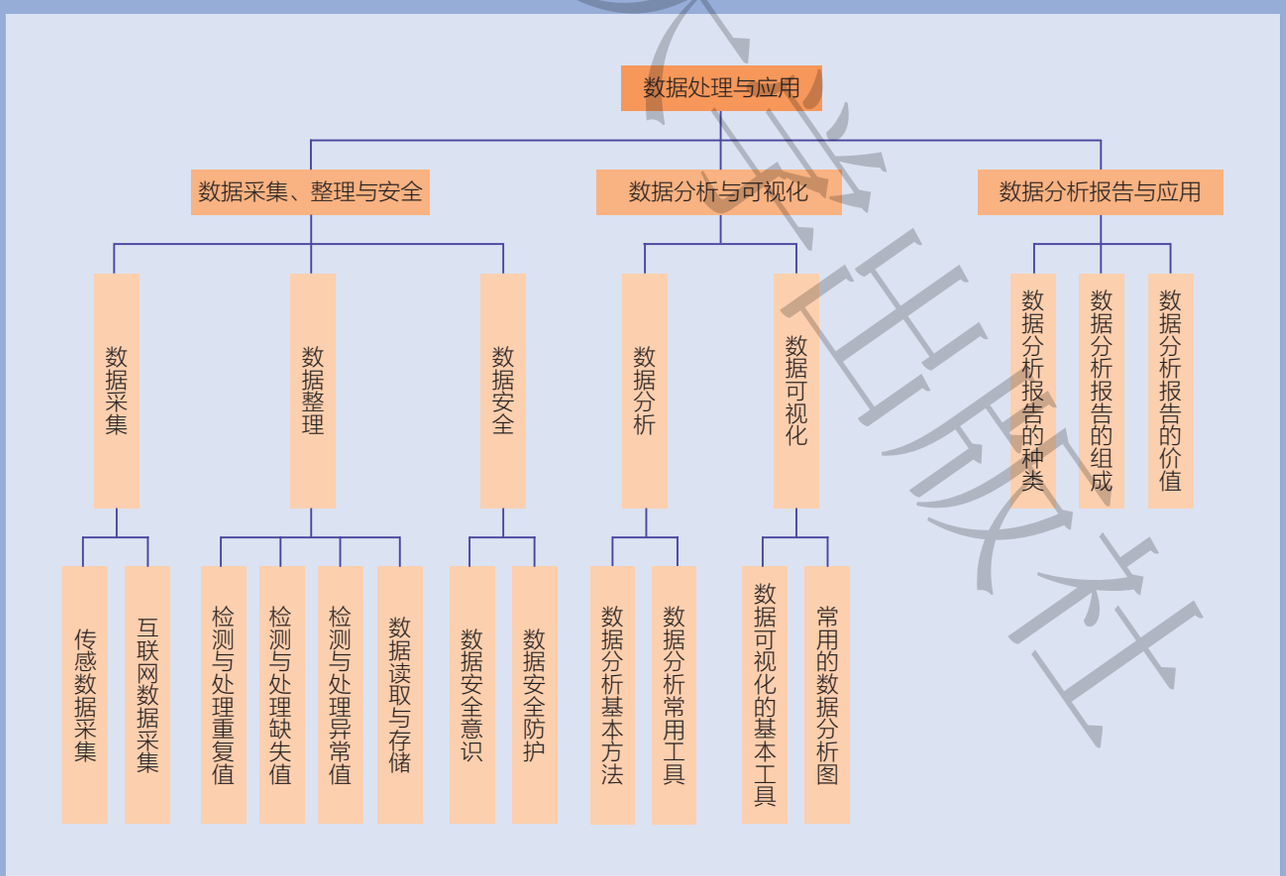
- 通过典型的应用实例，了解数据采集和整理的基本方法，理解数据安全的重要性。
- 选用合适的软件工具或平台处理数据，掌握数据可视化的基本方法。
- 了解数据分析报告的结构形式，完成解决问题的数据分析报告，感悟数据分析的价值。

数据处理是对数据的采集、整理、分析和可视化表达的过程。数据处理的基本目的是从大量的、可能杂乱无章或难以理解的数据中抽取并推导出有价值、有意义的数据，为人们的判断、决策、预测提供依据。

如今，数据处理已贯穿社会生产和社会生活的各个领域，数据处理技术的发展及其应用的广度和深度，极大地影响着人类社会发展的速度。人流如织的地铁站中，无论是刷卡进站，还是线路选择，数据处理都悄然相伴于人们的步履匆匆；繁忙的物流行业中，无论是物流效率的提高，还是物流成本的降低，都离不开数据处理对物流网络的优化；宏伟的城市群布局中，无论是城市规划，还是环保监测，数据处理都为城市的科学管理提供了强有力的支撑。

在数据处理的学习和应用中，我们可以感受数据的魅力，探索数据的价值，提升驾驭数据的能力，让数据成为我们学习、工作和生活的得力助手。

本章知识结构



项目主题

交通数据利抉择

项·目·情·境

共享单车的诞生，顺应了“绿色出行”的环保理念，解决了人们出行“最后一公里”的烦恼。但与此同时，又有新的问题浮出水面。

小申是一名“优秀志愿者”，他的服务岗位是学校附近的共享单车站点。因为学校周边还有地铁站、图书馆等，所以小申服务的站点的共享单车租放量很大。有时共享单车太多而挤占了人行道，有时人多而共享单车却供不应求，小申看在眼里急在心里。共享单车使用的“潮汐”难题如何破解呢？

项·目·任·务

任务 1

利用信息技术工具收集共享单车使用过程中的相关数据，形成数据集。

任务 2

学习数据处理的常用工具和方法，对数据集进行整理，用可视化方式呈现出来。

任务 3

应用项目活动中的数据处理结果，以小组为单位撰写数据分析报告，交流分享学习成果。

第一节　数据采集、整理与安全

当今社会，信息技术开始渗透至人类日常生活的方方面面，随之而产生的数据量也呈现指数级数增长的态势，例如物联网传感器、社交网络等每时每刻都产生着大量的数据。面对数据量的快速增长及变化、数据来源的多元化、数据呈现方式的多样化，我们在遵守相关法律法规、尊重知识产权的前提下，有效地采集与整理数据是进行数据处理的基础。

体　验　思　考

共享单车在使用时有解锁和闭锁环节。解锁和闭锁方式有多种，图 3.1和图 3.2所示是用户在使用共享单车时采用蓝牙方式借还车的场景。

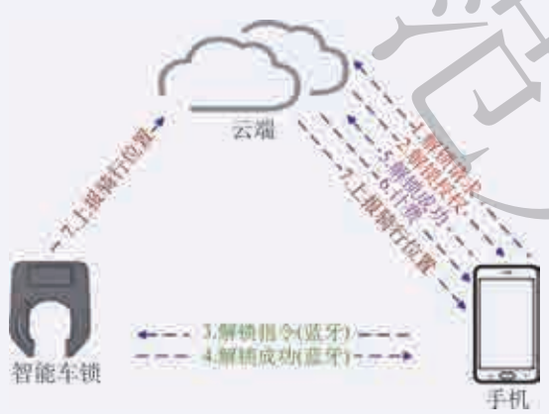


图 3.1　蓝牙模式解锁流程

图 3.2　蓝牙模式锁车、还车流程

思考：在借还车的过程中产生的大量数据有哪些类型？如何采集这些数据？我们可以将数据保存在哪里？把思考的结果填写在表 3.1 中。

表 3.1　共享单车的数据

	单 车 数 据	用 户 数 据
数据描述	通信连接状态、车锁状态、使用记录、_____、_____、_____等	用户基本信息、消费记录、骑行的路径、_____、_____、_____等
数据采集方法		
数据保存方法		

一、数据采集

数据采集一般需要经历明确数据要求、确定数据来源、选择采集方法、实施数据采集的过程。数据来源有多种渠道，如传感设备、互联网、问卷调查、企业内部数据库等途径。采集数据的方法有很多，目前较为广泛使用的是传感数据采集和互联网数据采集。

1. 传感数据采集

传感数据是由传感设备收集和测量的数据，传感设备可穿戴在用户身上，也可设置在现实环境中，如图 3.3 所示。传感数据涉及很多方面，如人体的传感数据、网络信号的传感数据和气象的传感数据等。通常情况下，传感设备以一定的频率采集数据并发送至相应的数据接收端，这些数据精准地记录着某个具体参数的实时变化情况。基于传感设备所采集的数据为后续的数据分析和数据可视化提供了重要的数据来源。

图 3.3　可穿戴设备示意图

2. 互联网数据采集

互联网数据采集是指利用互联网搜索引擎技术实现有针对性、行业性的数据抓取，并按照一定规则和筛选标准进行数据归类，最终形成数据库文件的一个过程。

通常，实现互联网数据采集的流程有三个步骤：获取网页、解析网页（提取数据）和保存数据。

（1）获取网页

获取网页的工作主要是获取网页的源代码。源代码里包含了网页的部分有用信息，只要得到源代码，就可以从中提取想要的信息了。获取源代码的关键就是构造一个请求并发送给服务器，然后在接收到服务器的响应后将其解析出来。

Python 提供了许多库来帮助我们实现这个操作，如 urllib、Requests 等。请求和响应都可以用库提供的数据结构来表示，得到响应之后只需要解析数据结构中的 body 部分，即可得到网页的源代码，这样我们就可以用程序来实现获取网页的过程了。

（2）解析网页

获得网页的源代码后，接下来就是要分析网页源代码，从中提取

我们想要的数据。由于网页的结构有一定的规则，所以可以利用一些用于提取网页信息的库（如 Beautiful Soup、PyQuery、lxml 等），高效快速地提取网页信息。

解析网页并从中提取信息，可以使杂乱的数据变得条理清晰，以便我们后续处理和分析数据。

（3）保存数据

提取数据后，我们一般会将其保存，以便后续使用。保存的形式多种多样，如文件存储、数据库存储或网络存储等。

项 目 实 践

上网搜索共享单车的运营和用户使用信息，尝试使用互联网数据采集的方法，采集我们生活区域内的共享单车数据，如站点数量、单车数量、开（闭）锁时间、骑行距离等，以 TXT文件格式进行保存。

技 术 支 持　网页爬取与解析

1 网页请求过程

要让网页展示在我们的面前，首先要经历的第一步，就是向服务器发送访问请求（requests）。服务器接收到请求后，会验证请求的有效性，然后向客户端发送响应的内容（responses），客户端接收服务器响应的内容，将内容展示出来，如图 3.4所示。

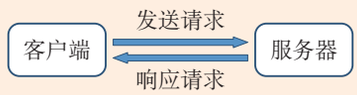


图 3.4 请求示意图

2 网页请求方式

Python有一个强大的 Requests库能够让我们很方便地发送 HTTP请求，这个库的功能比较完善，而且操作比较简单。导入 Requests库的语法为：

```
import requests
```

网页请求方式有多种，GET是最常见的方式，一般用于获取或者查询资源信息，响应速度较快。相比GET方式，POST方式多了以表单形式上传参数的功能，因此除了可以查询信息外，还可以修改信息。

在写程序前要先确定向谁发送请求，用什么方式发送。

以下是以 GET方式获取所需网页源代码的 Python程序代码。

```
import requests                          # 导入 Requests 库
url = 'http://www.xxxxxx.cn/'            # 确定请求对象
html = requests.get(url)                 # 用 GET 方式获取网页数据
html.encoding = "utf-8"                  # 用 UTF-8 文本编码
print(html.text)                         # 输出网页的源代码
```

“html=requests.get(url)”语句中的 html 是一个 URL对象，它代表整个网页，而 html.text才是网页源代码。很多情况下，如果直接使用 html.text会出现乱码的问题，而使用 html.encoding属性来设置文本编码（如 GBK、UTF-8等），就解决了通过 html.text直接返回会显示乱码的问题。

3 网页解析

通过 Requests库抓取到网页源代码后，我们要从源代码中找到并提取数据。BeautifulSoup库也是Python的一个功能很强的库，其主要功能是从网页中抓取数据。BeautifulSoup库目前已经被移植到bs4库中，所以导入 BeautifulSoup库的语法为：

```
from bs4 import BeautifulSoup
```

导入 BeautifulSoup库后，先用 Requests库中的 GET方法取得网页源代码，然后利用 Python内建的 html.parser解析器对源代码进行解析，解析的结果返回到 BeautifulSoup类对象 sp中。具体语法为：

```
sp = BeautifulSoup(源代码,'html.parser')
```

以下是获取网页 <title> 内容的 Python程序代码。

```
import requests                                  # 导入 Requests 库
from bs4 import BeautifulSoup                    # 导入 Beautiful Soup 库
url = 'http://www.xxxxxx.cn/ '                   # 确定请求对象
html = requests.get(url)                         # 用 GET 方式获取网页数据
html.encoding = "utf-8"                          # 用 UTF-8 文本编码
sp = BeautifulSoup(html.text,'html.parser')      # 解析源代码
title = sp.select("title")                       # 用 select 属性抓取 title 数据
print(title.text)                                # 输出 title 数据
```

BeautifulSoup库常用的属性和方法见表 3.2。

表 3.2 BeautifulSoup库常用的属性和方法

属性和方法	说　明
title	返回网页标题
text	返回去除所有 HTML标签后的内容
find()	返回第一个符合条件的标签，例如 sp.find("a")
find_all()	返回所有符合条件的标签，例如 sp.find_all("a")
select()	返回指定 CSS样式（如 id或 class）的内容，例如 sp.select "#id"可实现通过标签的 id抓取，sp.select(".classname"可实现通过标签的类名抓取

4 数据存储

将所抓取的数据存储到本地的 TXT文件中，只需再加上三行代码：

```
with open('title.txt', "a+")  as  f:
    f.write(title)
    f.close( )
```

二、数据整理

数据整理是数据分析过程中的重要环节，包括检查处理数据的重复值、缺失值和异常值等。数据的重复值会导致数据分布发生较大变化。数据的缺失值会导致样本信息减少，降低数据分析的准确性。数据的异常值不仅增加了数据分析的难度，而且会导致数据分析的结果产生偏差。

数据整理的过程是否科学、结果能否真实地反映客观实际，将直接影响数据处理的质量，影响整个数据分析的准确性。

探 究 活 动

图 3.5所示是采集到的共享单车位于某区各站点的部分数据，其中包含了共享单车编号（bike_id）、开锁时间（datetime）、工作日（workingday）、站点名（bcal_name）、气温（temp_value）、风速（wind_speed）等各种特征数据。仔细观察数据，查找、列举数据中存在的问题，讨论解决的方法。

index	bike_id	datetime	date	month	season	workingday	local_name	weather	isdaytime	temp_value	temp_unit	wind_speed	wind_unit
505	ar96862	9:50:22	6月21日	6	Summer	Yes	图书馆	Sunny	TRUE	20	C	3.7	km/h
506	ar96368	9:50:22	6月21日	6	Summer	Yes	上海金领谷科技产业园	Sunny	TRUE	22	C	3.7	km/h
507	ar96383	9:50:22	6月21日	6	Summer	Yes	上海电力工业学校	Sunny	TRUE	25	C	3.7	km/h
508	ar96396	9:50:22	6月21日	6	Summer	Yes	上海市第五人民医院	Sunny	TRUE	23	C	3.7	km/h
509	ar96413	9:50:22	6月21日	6	Summer	Yes	上海市第五人民医院	Sunny	TRUE	21	C	3.7	km/h
510	ar96701	9:50:22	6月21日	6	Summer	Yes	水生园	Sunny	TRUE	18	C	3.7	km/h
511	ar96727	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	19	C	3.7	km/h
512	ar96753	9:50:22	6月21日	6	Summer	Yes	春申创意园	Sunny	TRUE	18	C	3.7	km/h
513	ar96763	9:50:22	6月21日	6	Summer	Yes	图书馆	Sunny	TRUE	20	C	3.7	km/h
514	ar96942	9:50:22	6月21日	6	Summer	Yes	莘松中学	Sunny	TRUE	17	C	3.7	km/h
515	ar96983	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	20	C	3.7	km/h
516	ar97068	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	17	C	3.7	km/h
517	ar97070	9:50:22	6月21日	6	Summer	Yes	莘松中学	Sunny	TRUE	20	C	3.7	km/h
516	ar97068	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	18.7	C	3.7	km/h
518	ar97128	9:50:22	6月21日	6	Summer	Yes	莘庄镇政府	Sunny	TRUE	20	C	3.7	km/h
519	ar96448	9:50:22	6月21日	6	Summer	Yes	吉参生活广场	Sunny	TRUE	18	C	3.7	km/h
520	ar96478	9:50:22		6	Summer	Yes		Sunny	TRUE		C	3.7	km/h
521	ar96847	9:50:22	6月21日	6	Summer	Yes	莘松中学	Sunny	TRUE	20	C	3.7	km/h
522	ar96532	9:50:22	6月21日	6	Summer	Yes	上海电力工业学校	Sunny	TRUE	17.8	C	3.7	km/h
523	ar96254	9:50:22		6	Summer	Yes		Sunny	TRUE		C	5.6	km/h
524	ar96234	9:50:22	6月21日	6	Summer	Yes	鑫都商业广场	Sunny	TRUE	20	C	5.6	km/h
525	ar96397	9:50:22	6月21日	6	Summer	Yes	吉参生活广场	Sunny	TRUE	20	C	3.7	km/h
526	ar96445	9:50:22	6月21日	6	Summer	Yes	万达广场	Sunny	TRUE	17.8	C	3.7	km/h
527	ar96682	9:50:22	6月21日	6	Summer	Yes	图书馆	Sunny	TRUE	17.8	C	5.6	km/h
528	ar96854	9:50:22	6月21日	6	Summer	Yes	吴泾医院	Sunny	TRUE	17.8	C	5.6	km/h
529	ar97011	9:50:22	6月21日	6	Summer	Yes	吴泾医院	Sunny	TRUE	17.8	C	5.6	km/h

图 3.5　某区共享单车各站点部分数据

1. 检测与处理重复值

处理重复数据是数据分析经常要面对的问题之一。对重复数据进行处理前，需要分析重复数据产生的原因以及去除这部分数据后可能造成的不良影响。常见的数据重复情况分为两种：一种为记录重复，即某几条记录的一个或多个特征值完全相同；另一种为特征重复，即存在一个或者多个特征名称不同，但数据完全相同的情况。针对记录重复的处理，Python 的数据分析核心库 Pandas 提供了一个名为 drop_duplicates()的去重方法。该方法只对 DataFrame 或者 Series 类型有效。其基本语法如下：

```
pandas.DataFrame(Series).drop_duplicates(self, subset=None, keep='first', inplace=False)
```

该方法的常用参数及其说明如表 3.3 所示。

表 3.3　drop_duplicate()方法的常用参数及其说明

参数名称	说　明
subset	接收字符串或序列，表示进行去重的列，默认为 None，表示全部列
keep	接收特定字符串，表示重复时保留第几个数据：first 保留第一个，默认为 first，last 保留最后一个；False，只要有重复都不保留
inplace	接收 boo 型数据，表示是否在原表上进行操作，默认为 False

例如，对“某区共享单车数据”(test.csv)进行数据检测与去重处理的过程如下：

（1）分析数据

以图 3.5 所示的数据为例，编号为“mr97068”的共享单车记录不仅出现了编号重复的情况，而且开锁时间等也有重复的情况出现。

（2）确定方法

因为同一辆单车在同一时间不可能被开锁两次，所以删除其中一条记录不会对数据造成不良影响，可以用 drop_duplicates()方法对数据进行去重处理。

（3）编程与调试

编写程序，对 test.csv 中的数据进行去重处理，具体代码如下：

```
import pandas as pd
df = pd.read_csv('test.csv', encoding = "ANSI")        #读取 test.csv 文件
mydf = df.drop_duplicates(subset = ['bike_id','datetime'], keep = 'first', inplace = True)
#去除 bike_id 和 datetime 的重复数据
print(df)
```

2. 检测与处理缺失值

缺失值是指数据中的某个或多个特征的值是不完整的。Pandas 库提供了识别缺失值的方法 isnull()和识别非缺失值的方法 notnull()，这两种方法在使用时返回的都是布尔值，即 True 和 False。删除法是常用的缺失值处理方法，它通过减少样本量来换取信息完整度，是一种较简单的缺失值处理方法。Pandas 库中提供了简便的删除缺失值的方法 dropna()。通过参数控制，该方法既可以删除观测记录，也可以删除特征，其基本语法如下：

```
pandas.DataFrame.dropna (self, axis=0, how='any', inplace=False)
```

该方法的常用参数及其说明如表 3.4 所示。

表 3.4　dropna()方法主要参数及其说明

参数名称	说　明
axis	接收 0或 1,表示轴向；0为删除观测记录(行)，1为删除特征(列)，默认为 0
how	接收特定字符串，表示删除的形式：any表示只要有缺失值存在就执行删除操作，默认为 any；all表示当且仅当全部为缺失值才执行删除操作
inplace	接收 boo型数据，表示是否在原表上进行操作，默认为 False

例如，对“某区共享单车数据”(test.csv)进行数据检测与处理缺失值的过程如下：

(1) 分析数据

同样以图 3.5 所示数据为例，编号为“mr96478”和“mr96254”的共享单车记录中出现了缺失日期(date)和站点名(local_name)这两个特征值的情况。

（2） 确定方法

缺失日期和站点名两个特征值的记录对后续数据统计的意义不大，可以用 dropna()方法进行删除。

（3） 编程与调试

编写程序，对 test. csv 中的数据处理缺失值，具体代码如下：

```
import pandas as pd
df = pd.read_csv('test.csv', encoding = "ANSI")          #读取 test.csv 文件
mydf = df.dropna(axis = 0, inplace = True)               #处理缺失值，按行删除
print(df)
```

对图 3.5 所示数据进行去重和处理缺失值后，结果如图 3.6 所示。

index	bike_id	datetime	date	month	season	workingday	local_name	weather	isdaytime	temp_value	temp_unit	wind_speed	wind_unit
505	mr96862	9:50:22	6月21日	6	Summer	Yes	图书馆	Sunny	TRUE	20℃		3.7	km/h
506	mr96388	9:50:22	6月21日	6	Summer	Yes	上海金桥谷科技产业园	Sunny	TRUE	22℃		3.7	km/h
507	mr96383	9:50:22	6月21日	6	Summer	Yes	上海电力工业学校	Sunny	TRUE	25℃		3.7	km/h
508	mr96396	9:50:22	6月21日	6	Summer	Yes	上海市第五人民医院	Sunny	TRUE	23℃		3.7	km/h
509	mr96413	9:50:22	6月21日	6	Summer	Yes	上海市第五人民医院	Sunny	TRUE	21℃		3.7	km/h
510	mr96701	9:50:22	6月21日	6	Summer	Yes	水生园	Sunny	TRUE	18℃		3.7	km/h
511	mr96727	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	19℃		3.7	km/h
512	mr96753	9:50:22	6月21日	6	Summer	Yes	春申创意园	Sunny	TRUE	18℃		3.7	km/h
513	mr96763	9:50:22	6月21日	6	Summer	Yes	图书馆	Sunny	TRUE	20℃		3.7	km/h
514	mr96942	9:50:22	6月21日	6	Summer	Yes	莘松中学	Sunny	TRUE	17℃		3.7	km/h
515	mr96963	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	20℃		3.7	km/h
516	mr97068	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	37℃		3.7	km/h
517	mr97070	9:50:22	6月21日	6	Summer	Yes	莘松中学	Sunny	TRUE	20℃		3.7	km/h
518	mr97128	9:50:22	6月21日	6	Summer	Yes	莘庄镇政府	Sunny	TRUE	20℃		3.7	km/h
519	mr96448	9:50:22	6月21日	6	Summer	Yes	青少生活广场	Sunny	TRUE	18℃		3.7	km/h
521	mr96847	9:50:22	6月21日	6	Summer	Yes	莘松中学	Sunny	TRUE	20℃		3.7	km/h
522	mr96582	9:50:22	6月21日	6	Summer	Yes	上海电力工业学校	Sunny	TRUE	17.8℃		3.7	km/h
524	mr96234	9:50:22	6月21日	6	Summer	Yes	嘉都商业广场	Sunny	TRUE	20℃		5.6	km/h
525	mr96397	9:50:22	6月21日	6	Summer	Yes	青少生活广场	Sunny	TRUE	20℃		3.7	km/h
526	mr96445	9:50:22	6月21日	6	Summer	Yes	万达广场	Sunny	TRUE	17.8℃		3.7	km/h
527	mr96682	9:50:22	6月21日	6	Summer	Yes	图书馆	Sunny	TRUE	17.8℃		5.6	km/h
528	mr96954	9:50:22	6月21日	6	Summer	Yes	吴泾医院	Sunny	TRUE	17.8℃		5.6	km/h
529	mr97011	9:50:22	6月21日	6	Summer	Yes	吴泾医院	Sunny	TRUE	17.8℃		5.6	km/h

图 3.6 处理后的某区共享单车各站点部分数据

3. 检测与处理异常值

异常值是指数据中个别值的数值明显偏离其余的数值，有时也称为离群点。检测异常值就是检验数据中是否有输入错误以及是否含有不合理的数据。一般使用箱形图或散点图能较清晰地观察到异常值的存在，如图 3.7 所示。异常值的处理方法包括：直接将含有异常值的记录删除；用前后两个观测值的平均值修正该异常值；将异常值视为缺失值，利用处理缺失值的方法进行处理等。

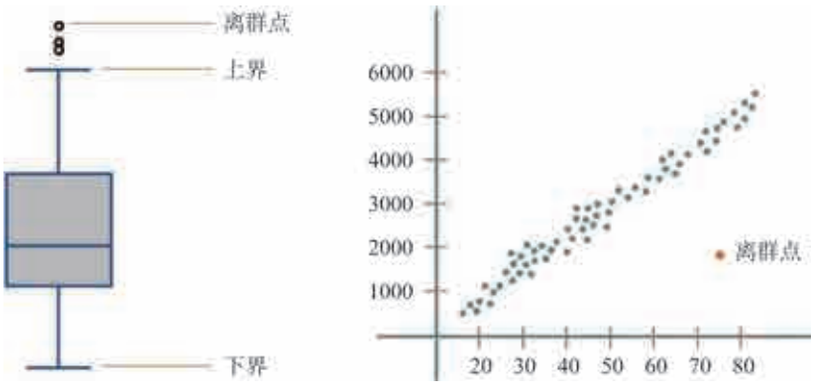


图 3.7　用箱形图和散点图分析异常值

技　术　支　持　Pandas库与 DataFrame()方法

Pandas库兼具 NumPy库的高性能数组计算功能以及电子表格和关系型数据库灵活的数据处理功能。Pandas库是本章中使用的主要工具。导入 Pandas库的语法为：

```
import pandas as pd
```

Pandas库主要的数据类型有两种：Series是一维数据结构，其用法与列表类似；DataFrame是二维数据结构，表格即为 DataFrame的典型结构。

创建 DataFrame的语法为：

```
数据变量 = pd.DataFrame(数据类型)
```

“数据类型”可以是多种类型：由包含相同数量的列表数据作为键值的字典创建的 DataFrame数据、自行设置的行及列标题创建的 DataFrame数据等。

例如，创建一个包含四种交通工具以及每种交通工具的日均客运量和占比的 DataFrame，数据变量名称为 df：

```
datas = [[969.2,53.9],[602.9,33.5],[208,11.6],[13.6,0.8]]
indexs = ["轨道交通","公共汽/电车","出租车","轮渡"]
columns = ["日均客运量(万人次)","占公共交通日均客运总量(%)"]
df = pd.DataFrame(datas, columns = columns, index = indexs)
```

生成的 DataFrame如下：

	日均客运量(万人次)	占公共交通日均客运总量(%)
轨道交通	969.2	53.9
公共汽/电车	602.9	33.5
出租车	208	11.6
轮渡	13.6	0.8

这种方式会按用户输入数据的顺序来生成 DataFrame，且具有行、列标题。如无特殊需求，我们通常

会以这种方式生成 DataFrame。

读取 DataFrame一个列数据的语法为：

```
df[列标题]
```

要读取两个以上列数据的语法为：

```
df[[列标题 1,列标题 2……]]
```

我们可以按行或列读取数据，也可以读取全部数据。读取 DataFrame数据的方法有很多种，如 df.values、df.loc、df.iloc、df.ix等，我们可以根据需要进行选用。

4. 数据读取与存储

数据读取与存储是进行数据处理与分析的前提。不同的数据源，需要使用不同的函数来读取。Pandas 内置了十余种数据源读取函数和对应的数据写入函数。常见的数据源有文本文件(包括一般文本文件和 CSV 文件)、电子表格文件等。

文本文件是一种由若干行字符构成的计算机文件，它是一种典型的顺序文件。CSV 是一种用分隔符分隔的文件格式，因为其分隔符不一定是逗号，因此又被称为字符分隔文件格式。CSV 文件以纯文本形式存储表格数据(数值和文本)。

(1) 文本文件的读取

Pandas 库提供了 read_csv()函数来读取 CSV 文件，其语法如下：

```
pandas.read_csv(filepath, sep = ',', header = 'infer', names = None, index_col = None, dtype = None, encoding = utf-8)
```

read_csv()函数的常用参数及其说明如表 3.5 所示。

表 3.5 read_csv()函数的常用参数及其说明

参数名称	说　明
filepath	接收字符串，表示文件路径，无默认值
sep	接收字符串，表示分隔符，默认为“,”
header	接收 int型数据，表示将某行数据作为列名，默认为 infer，表示自动识别
names	接收数组，表示列名，默认为 None
dtype	接收字典，表示写入的数据类型，默认为 None

（2） 文本文件的存储

文本文件的存储与读取类似，对于结构化数据，可以通过 Pandas 库中的 to_csv()函数实现以 CSV 文件格式进行存储。to_csv()函数的语法如下：

```
DataFrame.to_csv(path_or_buf = None, sep = ',', na_rep = '', columns = None, header = True, index = True, index_label = None, mode = 'w', encoding = None)
```

to_csv()函数的常用参数及其说明如表 3.6 所示。

表 3.6 to_csv()函数的常用参数及其说明

参数名称	说　明
path_or_buf	接收字符串，表示文件路径，无默认值
sep	接收字符串，表示分隔符，默认为“,”
na_rep	接收字符串，表示缺失值，默认为空
columns	接收列表，表示写出的列名，默认为 None
header	接收 boo 型数据，表示是否将列名写出，默认为 True
index	接收 boo 型数据，表示是否将行名（索引）写出，默认为 True
index_label	接收序列，表示索引名，默认为 None
mode	接收特定字符串，表示数据写入模式，默认为“w”
encoding	接收特定字符串，表示存储文件的编码格式，默认为 None

项　目　实　践

在本章第一节“一、数据采集”的项目实践中，我们尝试上网搜索了共享单车的运营和用户使用信息，并使用互联网数据采集的方法，采集了我们生活区域内的共享单车站点数量、单车数量、开（闭）锁时间、骑行距离等数据。用本节所学的数据整理方法，检测采集的数据是否存在重复、缺失和异常值，并编写程序加以处理，以 CSV文件格式保存结果。

三、数据安全

随着大数据、物联网、云计算等技术和应用的日渐兴起，大数据应用越来越被人们所重视。然而，数据在体现和创造价值的同时，也面临着严峻的安全风险。在复杂的应用环境下，保障国家重要数据、企

业机密数据和用户个人隐私数据等不发生外泄，是数据安全的首要任务。海量多源数据在大数据平台汇聚，强化数据隔离和访问控制，实现数据“可用不可见”，是大数据环境下数据安全的新要求。

探 究 活 动

目前，全国各地正在积极有序地推进新版社会保障卡的换发工作，如图 3.8所示。新版社会保障卡既有社会保障应用功能，也有金融应用功能，并采用先进的互联网安全技术手段，构建网络与人之间的可信链接，确保在互联网上实现“实人、实名、实卡”，使人们能够高效、安全地享受各项公共服务。

图 3.8　新版社会保障卡示意图

查阅资料并讨论：

1 新版社会保障卡增加了哪些功能？有什么用途？

2 新版社会保障卡应用了哪些新技术？这些新技术是如何保护我们的个人信息的？

将讨论结果填入表 3.7中。

表 3.7　新、旧版社会保障卡应用对照

	旧版社会保障卡	新版社会保障卡
功能用途		
主要技术		
安全保障		

1. 数据安全意识

数据安全问题越来越被个人、企业乃至国家所重视。对数据安全的威胁主要有计算机病毒、黑客攻击、数据存储介质的损坏、自身数据管理不善等方面。

计算机病毒能影响计算机软件、硬件的正常运行，破坏数据的正确性与完整性，甚至导致系统崩溃。黑客通常是先收集被攻击方的有关信息，分析可能存在的漏洞，然后实施攻击。黑客一旦入侵成功，就可以读取邮件、搜索和盗窃文件、毁坏重要数据、破坏系统信息，造成不堪设想的后果。数据存储介质的安全隐患包括物理损坏、设备故障

和电磁辐射影响等。自身数据管理不善主要是人为因素造成的，用户安全意识不强、口令选择不慎、用户将自己的账号随意转借他人或与别人共享等都会对数据安全带来威胁。

我们的姓名、学历、家庭地址、身份证号、手机号等都是个人隐私数据。2018 年 5 月 1 日正式颁布实施的《信息安全技术个人信息安全规范》中进一步明确，人的基因、指纹、声纹、耳廓、面部识别特征等都属于个人敏感信息。大数据时代的各种便利已经渗透进我们生活的每一个角落，我们在访问互联网享受各项服务时，应同时意识到自己会留下个人隐私信息、会被他人获取用户数据等情况的存在。我们要辨别和使用可信的网站、网络服务或其他联网的应用程序，谨慎提交个人隐私信息等，要定期清理历史信息及更换各类账户的密码，保护自身的数据安全，防止数据外泄。

数据安全是一项系统工程，需要政府机关、行业主管部门、组织和企业、个人等积极发挥多元主体的作用，依据《国家安全法》《网络安全法》等法律法规要求，共同参与到数据安全保障体系的建设中来，做到知法守法，认真履行有关数据安全风险控制的义务和职责，增强数据安全可控意识，共同维护国家安全秩序。

2. 数据安全防护

为了保护数据安全，我们还需要在提高数据安全意识的同时，于技术层面上提升数据安全的防护水平。数据安全一般有两方面的含义：一方面是数据本身的安全，主要采用现代密码算法对数据进行主动保护，如数据加密、数据脱敏、访问控制等；另一方面是数据防护的安全，主要是采用现代信息存储手段对数据进行主动防护，如通过数据备份、异地容灾等手段保证数据的安全。

（1） 数据加密

数据加密是计算机系统对数据进行保护的一种较为可靠的办法。对需要保护的数据（也称为明文）进行加密，即利用加密算法和加密密钥将需要保护的数据转化成另外一种数据（也称为密文），然后将密文进行存储或者传输给需要使用数据的人，使得窃取者在没有密钥和不了解加密算法的情况下无法识别密文，从而起到数据保密的作用。

（2） 数据脱敏

数据脱敏是在不影响数据分析结果准确性的前提下，对需要保护的数据进行一定的变换操作，如替换、过滤或删除等，从而降低数据的敏感性，保护用户的隐私不被泄露，如图 3.9 所示。

原始数据：

序号	姓名	性别	银行卡号	消费金额
1	申　华	男	6225210106311234	128.30

脱敏后数据：

序号	性别	银行卡号	消费金额
1	男	6225 ********1234	128.30

图 3.9　数据脱敏示例

（3）访问控制

在各种计算机系统中，涉及各类服务的使用、文件的访问、数据的存取时，需要规定特定的人对部分数据负责或获得管理权限，从而做到被授权的人允许使用特定信息。此时，就需要进行访问控制，这是确定用户身份及其所享有权限的一种技术。访问控制主要由身份验证与授权两个部分组成，身份验证是用于验证用户身份合法性的一种技术。身份验证本身并不足以防护数据，还需要授权技术来确定用户是否可以访问数据或执行其所尝试的操作。

（4）数据备份

图 3.10　数据备份

数据备份是指为了防止由于操作失误、系统故障等人为因素或意外原因导致数据丢失，而将整个系统的数据或者一部分关键数据通过一定的方法从主计算机系统的存储设备中复制到其他存储设备中的过程，如图 3.10 所示。一旦数据丢失，就可以从备份中恢复历史版本的数据。数据备份往往需要定期定时进行，从而使得备份的数据能够保持最新的状态。

（5）异地容灾

当某处的计算机系统因意外、不可抗力因素（如火灾、地震等）的原因导致停止工作并且无法提供计算机服务时，往往需要切换到另外一套备用系统上，使其能够继续提供相关计算机服务。如果两套或多套计算机系统都安放在同一处，一旦遭到不可抗力因素的影响时，将会是灾难性的。为了防止出现这种情况，人们采用了一种异地容灾的方式，在相隔较远的地方，建立两套或多套功能相同的计算机系统，相互进行数据备份或应急时提供备用计算机服务。例如，银行的数据中心都实现了异地容灾，从而可以保证用户的金融数据安全。

作业练习

目前,“停车难”几乎成为了掣肘城市发展、影响市民生活的顽疾。市中心周转腾挪的空间越来越有限,在众多停车问题中,路面停车问题尤为突出。图 3.11是道路停车标准化管理系统示意图,该系统通过视频自动检测和记录停车过程,无需人工干预,结合停车费线上支付和停车信息发布,实现路面停车管理无人值守,节省人力成本,提升泊位周转率,从而实现道路停车智能化的管理。

图 3.11　道路停车自动检测和缴费示意图

利用所学的数据采集方法,采集“停车点名称”“停车点地址”“停车泊位”“车辆出入记录”“收费记录”等相关特征数据,并用合适的方法进行整理,保存为“道路停车数据 .csv”文件。

知识延伸　云存储

云存储是一种新型的互联网存储技术,它采用集群应用、网格技术和分布式文件系统等,将网络中大量不同类型的存储设备通过应用软件集合起来协同工作,共同对外提供数据存储和业务访问功能,如图 3.12所示。

从用户使用的角度来看,云存储的优势主要表现为:

1.便捷存取文件

通过联网装置连接到云端后,用户就可以随时、随地存取文件,便于备份本地数据并可异地处理文件。

2.易于存储扩容

在需要扩大存储空间时,用户可以依靠云存储服务方提供的存储扩展功能,随时扩大存储空间,按需使用,而无需自身购置存储硬件和相关设施。

图 3.12　云存储示意图

3.节省存储成本

由于云存储的相关设施是由服务方来承担和负责维护的,用户采用云存储后,就不必操心设备的购置、升级与维护等方面的问题,这样就可以节省大量的设备投资经费。

尽管当前云存储还存在着存取速度受网络带宽限制、数据安全还需加强等方面的问题,但其便捷地存取文件、易于存储空间扩容、节省存储成本等优势已越来越受到用户的重视,并得到了广泛的应用。

第二节　数据分析与可视化

数据分析是指使用适当的分析方法对采集和整理后的数据加以详细研究，提取有用的信息和形成概括总结的过程。我们通常运用统计方法，对数据进行定性与定量的分析，然后借助可视化工具，直观清晰地呈现信息，并把信息的特征形象地传递给人们。

体　验　思　考

共享单车的出现方便了民众出行。但是随着车辆投放总量的增多，整体调控不足，也就出现了肆意挤占城市公共道路，局部公共空间饱和，偏远地区车辆不足的问题。依据骑车用户已有数据，合理调控车辆已成为管理共享单车的一种有效策略。图 3.13是某市共享单车骑行特征和时间分布数据图。

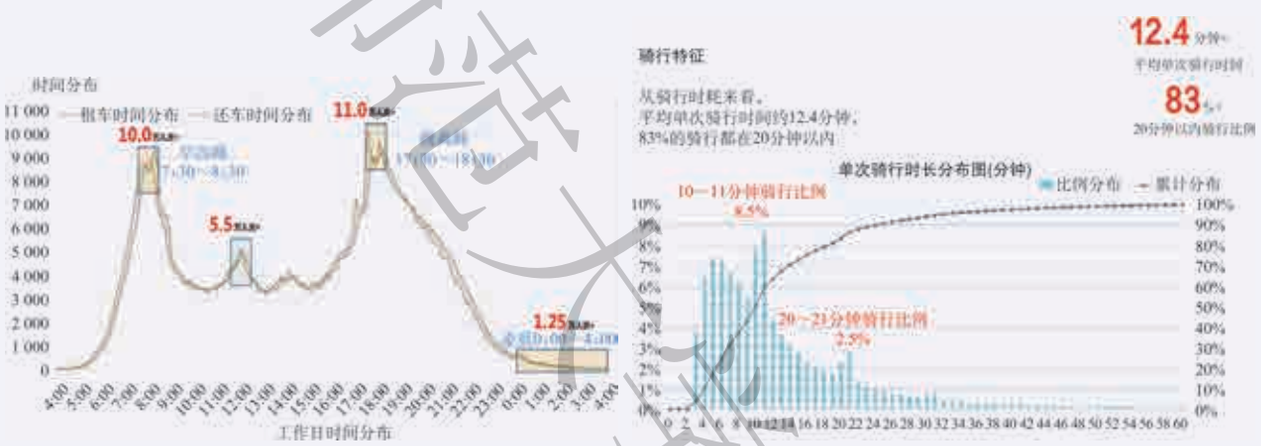


图 3.13　某市共享单车骑行特征和时间分布数据图

思考：

1. 分析图 3.13，指出图中所反映的用户使用共享单车的特征。
2. 针对分析得到的特征，尝试提出共享单车管理的建议。

一、数据分析

为了从数据中获取有价值的信息，对数据进行采集和整理后，还需要选用适当的方法与工具对数据进行分析。通过数据分析，可以描述事物的现状，发现相关要素的关系，并对事物的发展趋势做出相应的预测。

1. 数据分析基本方法

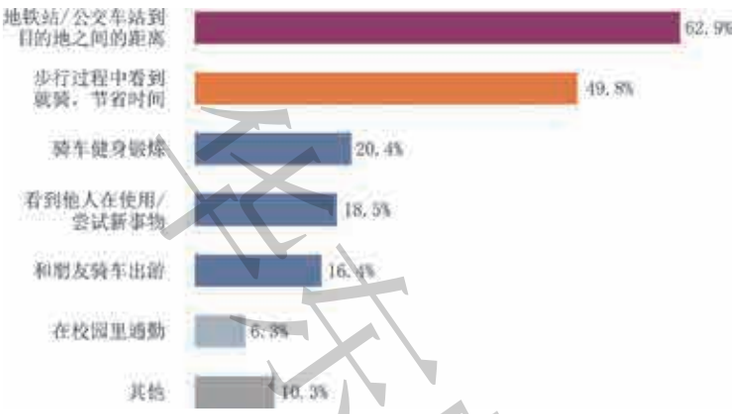


图 3.14　共享单车骑行场景分析

数据分析有很多种方法，其中基本的数据分析方法有对比分析法、平均分析法和结构分析法等。

（1）对比分析法

对比分析法是指将两个或两个以上的数据进行比较，分析它们的差异，从而揭示这些数据所隐含的事物发展变化或差距，并且可以准确、量化地表示出这种变化或差距。图 3.14 对不同场景下的共享单车使用情况进行了比较。

（2）平均分析法

平均分析法是利用平均数指标来反映某一特征数据总体在一定时间、地点条件下的一般水平。通过特征数据的平均数指标，呈现事物目前所处的位置和水平，进而对不同时期、不同类型单位的平均数指标进行对比，明示事物的发展趋势和变化规律。图 3.15 是对选择不同出行方式的平均出行距离所做的分析。

图 3.15　各种出行方式的平均出行距离分析

（3）结构分析法

结构分析法是通过计算各个部分占总体的比重，进而分析某一总体现象的内部结构特征、总体的性质、总体内部结构随时间推移而表现出的变化规律性。各个部分占总体的比重即为结构指标，总体中各结构指标的总和为 100%。图 3.16 反映了某中学学生选择的各种上学出行方式的占比。

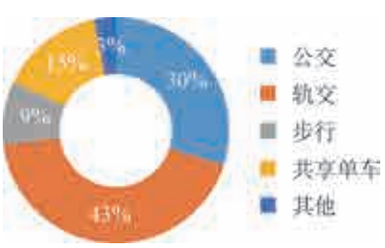


图 3.16　某中学学生上学出行方式分析

数据分析的每种方法都有各自的特点和适用范围，在实际应用中应根据解决问题的需求来选择合适的方法。

2. 数据分析常用工具

数据分析过程中使用较多的分析工具主要有三类，分别是电子表格软件、在线数据分析平台和数据分析语言。

（1）电子表格软件

电子表格软件是一种以表格形式组织、分析数据的计算机软件，

它以直观的表格形式进行数据分析，具有“所见即所得”的特点。电子表格软件通常具有表格制作、统计计算、图表处理等功能。图 3.17 所示是一款电子表格软件的应用界面。

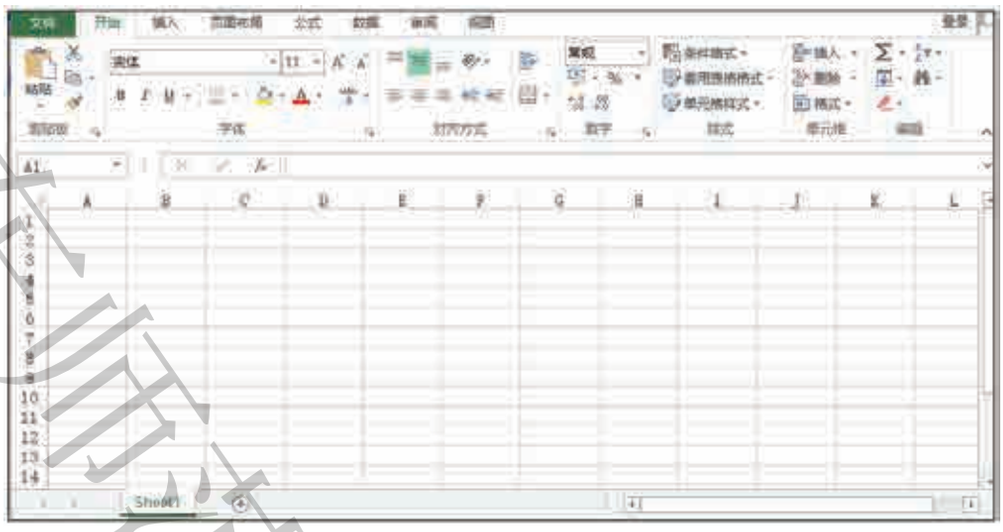


图 3.17　电子表格软件应用界面示例

（2）在线数据分析平台

在线数据分析平台是在网络环境下对所采集的数据进行实时分析的一种交互式数据服务平台。它可以通过与数据源的实时连接来获取最新数据，提供实时数据分析结果；也可以根据用户的数据分析计划实施数据分析，提供针对性的分析结果。在线数据分析平台提供了简单易用的交互界面，集成了多种数据分析功能，可以使用户更方便地获取数据分析结果。图 3.18 所示是我国国家统计局的一个在线

图 3.18　在线数据分析平台示例

数据分析平台。

（3） 数据分析语言

使用数据分析语言编写程序对数据进行分析时，可以按照实际需求，灵活、深入地进行数据分析和挖掘。在一定程度上，可以摆脱具体软件格式和平台专用功能的限制。目前，比较主流的数据分析语言有Python 语言、R 语言和 MATLAB 语言。

在对数据进行分析时，Python 语言具有较强的网络数据获取优势，还可调用丰富的工具库。例如，Numpy 库中的 sum()、mean()、min()、max()和 Pandas 库中的 value_counts()等都是可以用于统计的函数。R 语言和 MATLAB 语言依靠其独特的功能在相关专业领域使用得更为广泛。例如，R 语言在统计学领域使用较多，而MATLAB 语言则在工程计算等领域更受欢迎。

针对图 3.19 所示的数据，使用 Python 语言的 Numpy 和 Pandas 库编写程序，统计共享单车某个站点全天的开锁量，并计算其单位时间（小时）内最大的开锁量。其过程如下：

index	bike_id	datetime	date	month	season	workingday	local_name	weather	isdaytime	temp_value	temp_unit	wind_speed	wind_unit	year_month
505	mr96862	9:50:22	6月21日	6	Summer	Yes	图书馆	Sunny	TRUE	20	C	3.7	km/h	06-23-09
506	mr96388	9:50:22	6月21日	6	Summer	Yes	上海金领谷科技产业园	Sunny	TRUE	22	C	3.7	km/h	06-23-09
507	mr96389	9:50:22	6月21日	6	Summer	Yes	上海电力工业学校	Sunny	TRUE	25	C	3.7	km/h	06-23-09
508	mr96396	9:50:22	6月21日	6	Summer	Yes	上海市第五人民医院	Sunny	TRUE	23	C	3.7	km/h	06-23-09
509	mr96413	9:50:22	6月21日	6	Summer	Yes	上海市第五人民医院	Sunny	TRUE	21	C	3.7	km/h	06-23-09
510	mr96701	9:50:22	6月21日	6	Summer	Yes	水生园	Sunny	TRUE	18	C	3.7	km/h	06-23-09
511	mr96727	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	19	C	3.7	km/h	06-23-09
512	mr96753	9:50:22	6月21日	6	Summer	Yes	春申创意园	Sunny	TRUE	18	C	3.7	km/h	06-23-09
513	mr96763	9:50:22	6月21日	6	Summer	Yes	图书馆	Sunny	TRUE	20	C	3.7	km/h	06-23-09
514	mr96942	9:50:22	6月21日	6	Summer	Yes	莘松中学	Sunny	TRUE	17	C	3.7	km/h	06-23-09
515	mr96963	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	20	C	3.7	km/h	06-23-09
516	mr97068	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	37	C	3.7	km/h	06-23-09
517	mr97070	9:50:22	6月21日	6	Summer	Yes	莘松中学	Sunny	TRUE	20	C	3.7	km/h	06-23-09
518	mr97068	9:50:22	6月21日	6	Summer	Yes	青少年活动中心	Sunny	TRUE	18.7	C	3.7	km/h	06-23-09
518	mr97128	9:50:22	6月21日	6	Summer	Yes	莘庄镇政府	Sunny	TRUE	20	C	3.7	km/h	06-22-09
519	mr96448	9:50:22	6月21日	6	Summer	Yes	麦多生活广场	Sunny	TRUE	18	C	3.7	km/h	06-23-09
521	mr96847	9:50:22	6月21日	6	Summer	Yes	莘松中学	Sunny	TRUE	20	C	3.7	km/h	06-23-09
522	mr96582	9:50:22	6月21日	6	Summer	Yes	上海电力工业学校	Sunny	TRUE	17.9	C	3.7	km/h	06-23-09
524	mr96234	9:50:22	6月21日	6	Summer	Yes	嘉都商业广场	Sunny	TRUE	20	C	5.6	km/h	06-22-09
525	mr96397	9:50:22	6月21日	6	Summer	Yes	麦多生活广场	Sunny	TRUE	20	C	3.7	km/h	06-23-09
526	mr96445	9:50:22	6月21日	6	Summer	Yes	万达广场	Sunny	TRUE	17.8	C	3.7	km/h	06-23-09
527	mr96682	9:50:22	6月21日	6	Summer	Yes	图书馆	Sunny	TRUE	17.8	C	5.6	km/h	06-23-09
528	mr96954	9:50:22	6月21日	6	Summer	Yes	吴泾医院	Sunny	TRUE	17.8	C	5.6	km/h	06-23-09
529	mr97011	9:50:22	6月21日	6	Summer	Yes	吴泾医院	Sunny	TRUE	17.8	C	5.6	km/h	06-23-09

图 3.19　某区共享单车各站点部分数据

① 分析数据

图 3.19 所示的开锁时间（datetime）、站点名（local_name）和日期（year_month）这三个特征值为统计提供了数据支撑。

② 确定方法

要统计全天的开锁量，先要利用 value_counts()函数进行频数统计，得到单位时间（小时）内的开锁量，然后利用 sum()函数和 max()函数分别求得总开锁量和单位时间（小时）内的最大开锁量。

③ 编程与调试

具体代码如下：

```
import numpy as np
import pandas as pd
df = pd.read_csv('test.csv', encoding = "ANSI")          #读取 test.csv 文件
count = df[(df['year_month']>='06-21-00')&(df['year_month']<='06-21-23')
&(df['local_name']=='图书馆')]
#选出日期是 6 月 21 日且地点是“图书馆”的数据集
thiscount = count['year_month'].value_counts()          #按照时间进行频数统计
print(thiscount)
print(np.max(thiscount))   #查看 6 月 21 日一天中共享单车单位时间内的最大开锁量
print(np.sum(thiscount))   #查看 6 月 21 日一天中共享单车的开锁总量
```

程序运行结果如图 3.20 所示。

```
06-21-19    29
06-21-20    27
06-21-18    23
06-21-08    20
06-21-10    17
06-21-17    16
06-21-15    15
06-21-03     8
06-21-16     6
06-21-06     6
06-21-11     6
06-21-14     6
06-21-13     6
06-21-07     5
06-21-22     4
06-21-09     4
06-21-21     3
06-21-12     3
06-21-23     2
06-21-00     2
06-21-02     1
Name: year_month, dtype: int64
29
209
```

图 3.20　程序运行结果

项　目　实　践

编写程序，对本章第一节“一、数据采集”的项目实践中整理后的数据进行分析，统计共享单车某个站点全天的单位时间（小时）内平均开锁量，并计算单位时间（小时）内的最少开锁量。

二、数据可视化

数据可视化是将数据以图形化方式呈现，从而能够清晰、有效地传达与沟通信息。与文字和表格相比，用图形方式展示数据的特征，能够

更准确地表示数据的分布情况，便于人们有效地分析和理解数据。

1. 数据可视化的基本工具

目前，数据可视化的工具有很多，常用的数据分析软件一般都包含了创建可视化图表的功能。例如，电子表格软件中的图表功能可以基于选定的数据，用柱形图、折线图、饼图等方式呈现出来。创建图表后，可以通过修改数据标记、图例、标题、文字等来美化图表或强调某些信息，也可以用图案、颜色、对齐方式、字体及其他格式属性来对图表进行设置。电子表格软件的数据可视化过程直观、易用，但是对于大量数据可视化的实现就比较困难了。

当数据量较大时，可以使用编程语言对这些数据进行可视化。Python 语言中，Matplotlib 是一种应用较广的绘图工具包，使用其中的 pyplot 子库所提供的函数可以快速绘制图形，并能使用标签进行修饰，从而制作出高质量的数据分析图。

Python 语言中，引入 Matplotlib 的 pyplot 子库的语法为：

```
import matplotlib.pyplot as plt
```

pyplot 绘制图形有一个基本流程：创建画布与创建子图、添加画布内容、保存与显示图形。使用这个流程可以完成大部分图形的绘制。创建画布与创建子图的主要作用是构建一张空白的画布，并可以选择是否将整张画布划分为多个部分，以便在同一幅图上绘制多个图形。当只需要绘制一个简单的图形时，这部分内容可以省略。在 pyplot 中，创建画布以及创建并选中子图的函数如表 3.8 所示。

表 3.8　pyplot中创建画布以及创建并选中子图的常用函数

函数名称	函 数 作 用
plt.figure()	创建一张空白画布，可以指定画布大小、像素
figure.add_subplot()	创建并选中子图，可以指定子图的行数、列数和选中图片的编号

添加画布内容是绘图的主体部分。其中的添加标题、添加坐标轴名称、绘制图形等步骤是并列的，没有先后顺序，可以先绘制图形，也可以先添加各类标签。但是添加图例一定要在绘制图形之后。pyplot 中添加各类标签和图例的函数如表 3.9 所示。

表 3.9　pyplot中添加各类标签和图例的常用函数

函数名称	函 数 作 用
plt.title()	在当前图形中添加标题，可以指定标题的名称、位置、颜色、字体大小等参数
plt.xlabel()	在当前图形中添加 x轴名称，可以指定位置、颜色、字体大小等参数
plt.ylabel()	在当前图形中添加 y轴名称，可以指定位置、颜色、字体大小等参数
plt.xlim()	指定当前图形 x轴的范围，只能确定一个数值区间，而无法使用字符串标识
plt.ylim()	指定当前图形 y轴的范围，只能确定一个数值区间，而无法使用字符串标识
plt.xticks()	指定 x轴刻度的数目与取值
plt.yticks()	指定 y轴刻度的数目与取值
plt.legend()	显示当前图形的图例，可以指定图例的大小、位置、标签

保存和显示图形的常用函数只有两个，并且参数很少，如表 3.10 所示。

表 3.10　pyplot中保存和显示图形的常用函数

函数名称	函 数 作 用
plt.savefig()	保存绘制的图形，可以指定图形的分辨率、边缘的颜色等参数
plt.show()	在本机显示图形

例如：已知两条曲线 $y=x^2$ 和 $y=x^4$，当 $x\in[0,\ 1.1)$时，绘制一个最简单的不含子图的图形，如图 3.21 所示。具体代码如下：

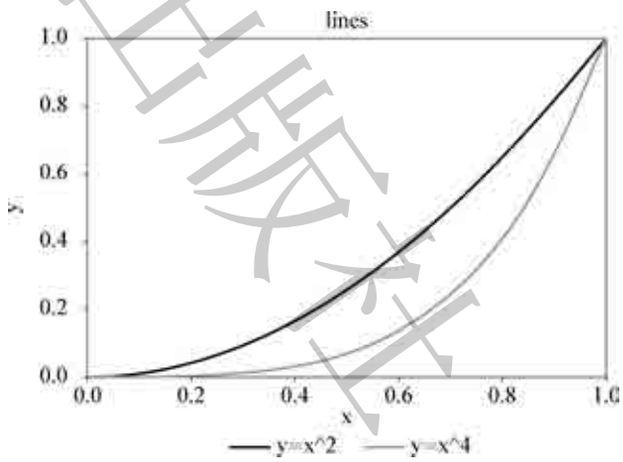


图 3.21　绘制不含子图的图形

```
import numpy as np
import matplotlib.pyplot as plt
data = np.arange(0,1.1,0.01)
#在[0,1.1)区间内,以0.01为间隔,创建一维数组
plt. title('lines')                          #添加标题
plt. xlabel('x')                             #添加x轴的名称
plt.ylabel('y')                              #添加y轴的名称
plt.xlim((0,1))                              #确定x轴范围
plt.ylim((0,1))                              #确定y轴范围
plt.xticks([0,0.2,0.4,0.6,0.8,1])            #确定x轴刻度
plt.yticks([0,0.2,0.4,0.6,0.8,1])            #确定y轴刻度
plt.plot(data, data ** 2)                    #添加y=x^2曲线
plt.plot(data, data ** 4)                    #添加y=x^4曲线
plt.legend(['y=x^2','y=x^4'])
plt.show( )
```

2. 常用的数据分析图

为了选择合适的图形实现数据的可视化,可以从分析特征间的关系、特征内部的数据分布与分散情况等方面,来选择和制作数据分析图。

(1) 分析特征间的关系

散点图和折线图是数据分析最常用的两种图形。这两种图形都能够分析不同数值型特征间的关系。其中,散点图主要用于分析特征间的相关关系,折线图则用于分析自变量特征和因变量特征之间的趋势关系。

散点图(scatter diagram)又称为散点分布图,是以一个特征为横坐标,以另一个特征为纵坐标,利用坐标点(散点)的分布形态反映特征间统计关系的一种图形。散点图中,值由点在图中的位置表示。

散点图可以提供两类关键信息:

① 特征之间是否存在数值或者数量的关联趋势,以及关联趋势是线性的,还是非线性的。

② 如果某一个点或者某几个点偏离大多数点，则这些点就是离群值，通过散点图可以一目了然，从而可以进一步分析这些离群值是否会在建模分析中产生很大的影响。

pyplot 中绘制散点图的函数为 scatter()，其语法如下：

```
matplotlib.pyplot.scattar(x, y, s = None, c = None, marker = None, alpha = None)
```

scatter()函数常用参数及其说明如表 3.11 所示。

表 3.11　scatter()函数常用参数及其说明

参数名称	说　明
x,y	接收数组，表示 x轴和 y轴对应的数据，无默认值
s	接收标量或一维数组，指定点的大小，若传入一维数组，则表示每个点的大小，默认为 None
c	接收颜色或一维数组，指定点的颜色，若传入一维数组，则表示每个点的颜色，默认为 None
marker	接收特定字符串，表示绘制的点的类型，默认为 None
alpha	接收 0~1的小数，表示点的透明度，默认为 None

例如：当 $x \in [0, 1.4)$时，绘制曲线 $y = x^2$ 和 $y = x^4$ 的简单散点图，如图 3.22 所示。具体代码如下：

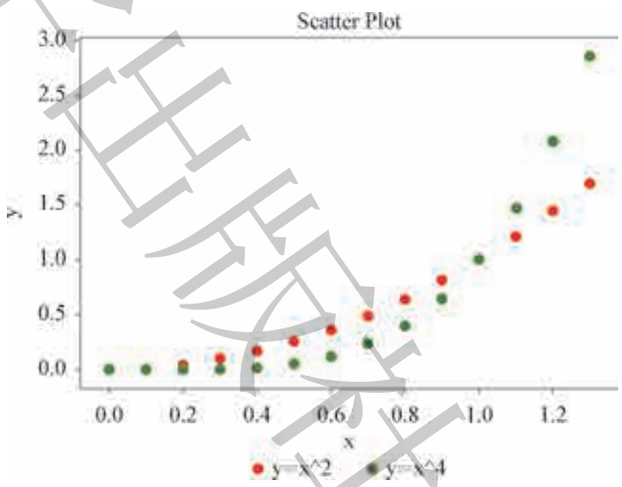


图 3.22　绘制简单的散点图

```
import numpy as np
import matplotlib.pyplot as plt
x = np.arange(0,1.4,0.1)
#在[0,1.4)区间内,以0.1为间隔,创建一维数组
plt.title('Scatter Plot')                    #设置标题
plt.xlabel('x')                              #设置x轴标签
plt.ylabel('y')                              #设置y轴标签
plt.scatter(x, x ** 2, c = 'r', marker = 'o')   #画散点图
plt.scatter(x, x ** 4, c = 'g', marker = 'o')
plt.legend(['y = x^2', 'y = x^4'])           #设置图标
plt.show()                                   #显示所画的图
```

折线图(line chart)是一种将数据点按照顺序连接起来的图形,可以看作是将散点图按照 x 轴坐标顺序连接起来。折线图的主要功能是查看因变量 y 随着自变量 x 改变的趋势,最适合用于显示随时间(根据常用比例设置)而变化的连续数据,同时还可以看出数量的差异和增长趋势的变化。

pyplot 中绘制折线图的函数为 plot(),其语法如下:

```
matplotlib.pyplot.plot(x, y, color = None, linestyle = '-', linewidth = 0.5)
```

plot()函数的常用参数及其说明如表 3.12 所示。

表 3.12 plot()函数的常用参数及其说明

参数名称	说　明
x,y	接收数组,表示 x轴和 y轴对应的数据,无默认值
color	接收特定字符串,指定线条颜色,默认为 None
linestyle	接收特定字符串,指定线条类型,默认为“-”
linewidth	接收 float型数据,指定线条宽度,默认为 0.5

以“某区共享单车数据”(test.csv)为例,绘制折线图,观察某区图书馆站点的共享单车在一天 24 小时中的使用情况,如图 3.23 所示。具体代码如下:

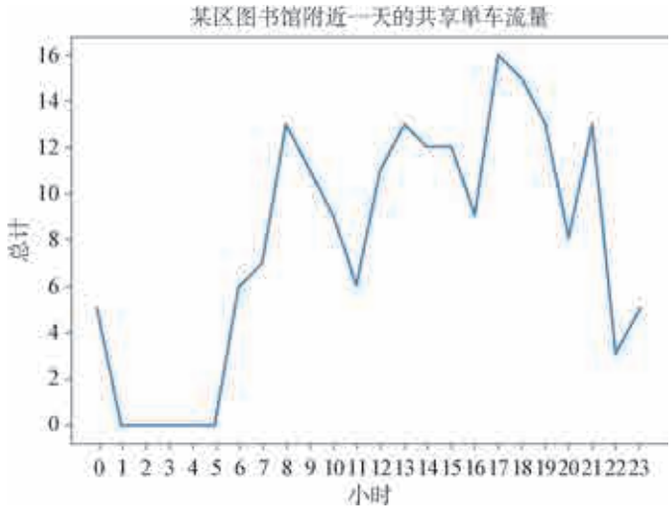


图 3.23　绘制折线图

```
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
#支持中文,用于正常显示中文标签
mydf = pd.read_csv('test.csv', encoding = "ANSI")
#读取 test.csv 文件
plt.title('某区图书馆附近一天的共享单车流量')
plt.xlabel('小时')
plt.ylabel('总计')
ax = plt.plot(mydf['index'], mydf['count'], linewidth = 2)
#绘制折线图
plt.xticks([0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,
17, 18, 19, 20, 21, 22, 23])
plt.show()
```

（2）分析特征内部数据分布与分散状态

柱状图、饼图和箱形图是数据分析常用的另外三种图形，主要用于分析数据内部的分布状态与分散状态。柱状图主要用于查看各分组数据的数量分布以及各分组数据之间的数量比较。饼图倾向于查看各分组数据在总数据中的占比。箱形图的主要作用是发现整体数据的分布、分散情况。

柱状图(bar chart)是由一系列高度不等的纵向条纹或线段表示数据分布的情况，一般用横轴表示数据所属类别，用纵轴表示数量或者占比。用柱状图可以比较直观地看出产品质量特性的分布状态，便

于判断其总体质量分布情况。

pyplot 中绘制柱状图的函数为 bar()，其语法如下：

```
matplotlib.pyplot.bar (left, height, width = 0.8, color = None)
```

bar()函数的常用参数及其说明如表 3.13 所示。

表 3.13　bar()函数的常用参数及其说明

参数名称	说　明
left	接收数组，表示 x轴数据，无默认值
height	接收数组，表示 y轴数据，无默认值
width	接收 0~1之间的 float型数据，指定柱状图宽度，默认为 0.8
color	接收特定字符串或者包含颜色字符串的数组，表示柱状图颜色，默认为 None

以“某区共享单车数据”(test.csv)为例，绘制柱状图，比较某区图书馆站点的共享单车在 7 月至 10 月的使用情况，如图 3.24 所示。具体代码如下：

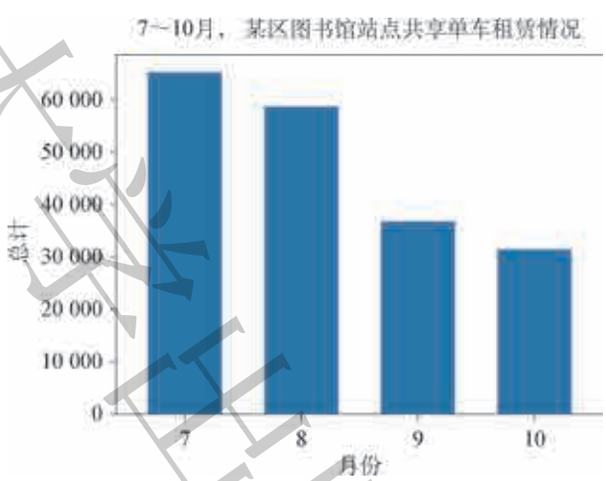


图 3.24　绘制柱状图

```
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
#支持中文，用于正常显示中文标签
mydf = pd.read_csv('test.csv', encoding = "ANSI")
#读取 test.csv 文件
plt.tick_params(labelsize = 10)
plt.title('7～10 月，某区图书馆站点共享单车租赁情况')
```

```
plt.xlabel('月份')
plt.ylabel('总计')
plt.bar(mydf['month'],mydf['count'])          #绘制柱状图
plt.xticks([7,8,9,10])
plt.show()
```

饼图(pie graph)是将各项的大小与各项总和的比例显示在一张“饼”中,以占“饼”的面积大小来确定每一项的占比。饼图可以比较清楚地反映出部分与部分、部分与整体之间的比例关系,易于显示每组数据相对于总数的大小,而且显示方式直观。

pyplot 中绘制饼图的函数为 pie(),其语法如下:

```
matplotlib.pyplot.pie (x,explode=None,labels=None,color=None,
autopct=None,pctdistance=0.6,labeldistance=1.1,radius=1)
```

pie()函数的常用参数及其说明如表 3.14 所示。

表 3.14 pie()函数的常用参数及其说明

函数名称	说 明
x	接收数组,表示用于绘制饼图的数据,无默认值
explode	接收数组,表示指定项距离饼图圆心为 n 个半径,默认为 None
labels	接收数组,指定每一项的名称,默认为 None
color	接收特定字符串或者包含颜色字符串的数组,表示饼图颜色,默认为 None
autopct	接收特定字符串,指定数值的显示方式,默认为 None
pctdistance	接收 float型数据,指定 autopct距离饼图圆心的位置相对于半径的比例,默认为 0.6
labeldistance	接收 float型数据,指定 labels距离饼图圆心的位置相对于半径的比例,默认为 1.1
radius	接收 float型数据,表示饼图的半径,默认为 1

图 3.25 绘制饼图

以“某区共享单车数据”(test.csv)为例,绘制饼图,观察某区内的共享单车在全年四个季节中的使用占比,如图 3.25 所示。具体代码如下:

```
import pandas as pd
import matplotlib.pyplot as plt
plt.rcParams['font.sans-serif'] = ['SimHei']
#支持中文,用于正常显示中文标签
```

```
mydf = pd.read_csv('test.csv', encoding = "ANSI")
# 读取 test.csv 文件
labels = '春季', '夏季', '秋季' ,'冬季'
plt.title('共享单车四季租赁情况')
plt.pie(mydf['count'], labels = labels, autopct = '%1.1f%%')
# 绘制饼图
plt.axis('equal')
plt.show()
```

绘制箱形图(boxplot)时需要使用常用的统计量,它能提供有关数据位置和分数情况的关键信息,尤其在比较不同特征时,更可表现其分散程度差异。图 3.26 标出了箱形图中每条线所表示的含义。

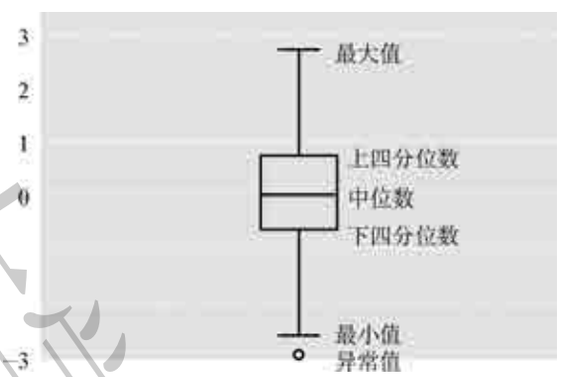


图 3.26 箱形图中每条线的含义

箱形图利用数据中的五个统计量(最小值、下四分位数、中位数、上四分位数和最大值)来描述数据。四分位数也称为四分位点,是把所有数据由小到大排列并四等分后,处于三个分割点位置的数值。处在 25%位置上的四分位数称为下四分位数,处在 75%位置上的四分位数称为上四分位数,处在中间的四分位数即为中位数。利用箱形图,可以粗略地看出数据是否具有对称性以及数据分布的分散程度等信息,也可以用于对数据进行异常值检测。

pyplot 中绘制箱形图的函数为 boxplot(),其语法如下:

```
matplotlib.pyplot.boxplot(x, notch = None, sym = None, vert = None, whis = 1.5, positions = None, widths = None, meanline = False, labels = None)
```

boxplot()函数的常用参数及其说明如表 3.15 所示。

表 3.15　boxplot()函数的常用参数及其说明

函数名称	说　　明
x	接收数组，表示用于绘制箱线的数据，无默认值
notch	接收 bool 型数据，表示中间箱体是否有缺口，默认为 None
sym	接收特定字符串，指定异常点形状，默认为 None
vert	接收 bool 型数据，表示图形是纵向或者横向，默认为 None
whis	接收 float 型数据，指定正常值范围，默认为 1.5
widths	接收标量或者数组，表示每个箱体的宽度，默认为 None
meanline	接收 bool 型数据，表示是否显示均值线，默认为 False
labels	接收数组，指定每一个箱形图的标签，默认为 None
positions	接收数组，表示图形位置，默认为 None

以“某区共享单车数据”(test.csv)为例，绘制箱形图，对某区共享单车各站点的部分数据进行异常值检测，如图 3.27 所示。具体代码如下：

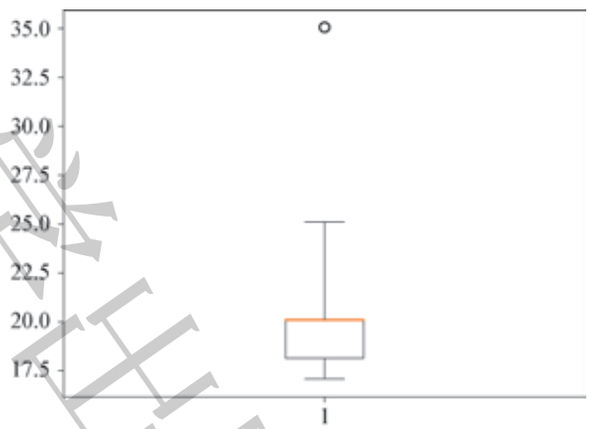
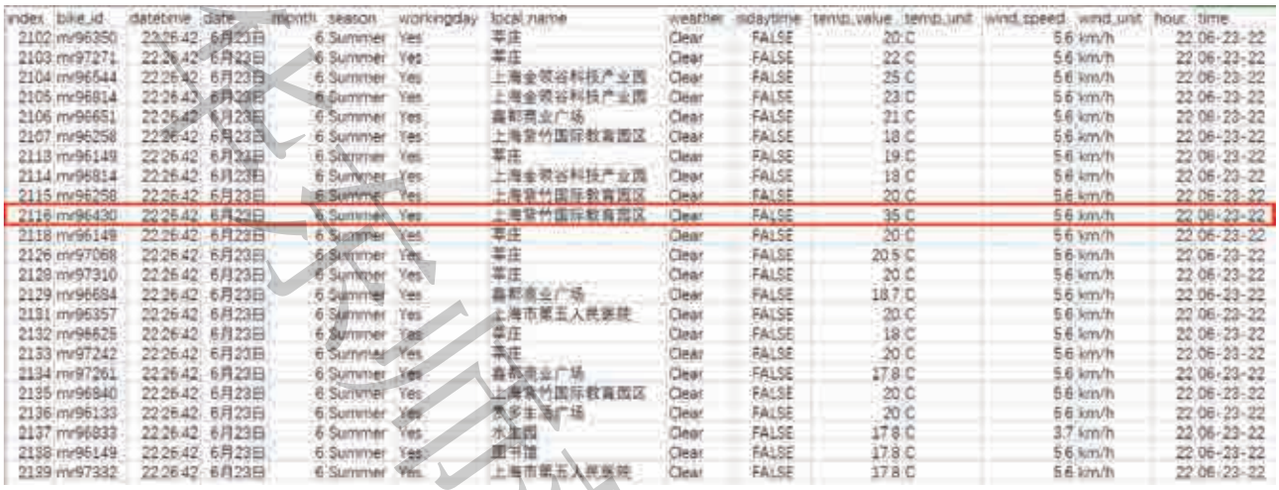


图 3.27　绘制箱形图

```
import pandas as pd
import matplotlib.pyplot as plt
mydf = pd.read_csv('test.csv', encoding = "ANSI")
plt.boxplot(mydf['temp_value'], sym = 'o', whis = 3.0)
#绘制箱形图
plt.show()
```

从图 3.27 所示的箱形图中很容易就可以找出异常值。例如，气

温值为 35 摄氏度，该值超出了 25 摄氏度的上限，所以可以定义为异常值。进一步观察图 3.28 所示的数据，可以发现编号为“mr96430”的共享单车在 6 月 23 日 22:26:42 时的气温值(35 摄氏度)为异常值。



index	bike_id	datetime	date	month	season	workingday	local_name	weather	isdaytime	temp_value	temp_unit	wind_speed	wind_unit	hour	time
2102	mr96350	22:26:42	6月23日	6	Summer	Yes	莘庄	Clear	FALSE	20	C	5.6	km/h	22	06-23-22
2103	mr97271	22:26:42	6月23日	6	Summer	Yes	莘庄	Clear	FALSE	22	C	5.6	km/h	22	06-23-22
2104	mr96544	22:26:42	6月23日	6	Summer	Yes	上海金领谷科技产业园	Clear	FALSE	25	C	5.6	km/h	22	06-23-22
2105	mr96814	22:26:42	6月23日	6	Summer	Yes	上海金领谷科技产业园	Clear	FALSE	23	C	5.6	km/h	22	06-23-22
2106	mr96651	22:26:42	6月23日	6	Summer	Yes	鑫都商业广场	Clear	FALSE	21	C	5.6	km/h	22	06-23-22
2107	mr96258	22:26:42	6月23日	6	Summer	Yes	上海紫竹国际教育园区	Clear	FALSE	18	C	5.6	km/h	22	06-23-22
2113	mr96149	22:26:42	6月23日	6	Summer	Yes	莘庄	Clear	FALSE	19	C	5.6	km/h	22	06-23-22
2114	mr96814	22:26:42	6月23日	6	Summer	Yes	上海金领谷科技产业园	Clear	FALSE	18	C	5.6	km/h	22	06-23-22
2115	mr96258	22:26:42	6月23日	6	Summer	Yes	上海紫竹国际教育园区	Clear	FALSE	20	C	5.6	km/h	22	06-23-22
2116	mr96430	22:26:42	6月23日	6	Summer	Yes	上海紫竹国际教育园区	Clear	FALSE	35	C	5.6	km/h	22	06-23-22
2118	mr96149	22:26:42	6月23日	6	Summer	Yes	莘庄	Clear	FALSE	20	C	5.6	km/h	22	06-23-22
2126	mr97068	22:26:42	6月23日	6	Summer	Yes	莘庄	Clear	FALSE	20.5	C	5.6	km/h	22	06-23-22
2128	mr97310	22:26:42	6月23日	6	Summer	Yes	莘庄	Clear	FALSE	20	C	5.6	km/h	22	06-23-22
2129	mr96684	22:26:42	6月23日	6	Summer	Yes	鑫都商业广场	Clear	FALSE	18.7	C	5.6	km/h	22	06-23-22
2131	mr96357	22:26:42	6月23日	6	Summer	Yes	上海市第五人民医院	Clear	FALSE	20	C	5.6	km/h	22	06-23-22
2132	mr96625	22:26:42	6月23日	6	Summer	Yes	莘庄	Clear	FALSE	18	C	5.6	km/h	22	06-23-22
2133	mr97242	22:26:42	6月23日	6	Summer	Yes	莘庄	Clear	FALSE	20	C	5.6	km/h	22	06-23-22
2134	mr97261	22:26:42	6月23日	6	Summer	Yes	鑫都商业广场	Clear	FALSE	17.8	C	5.6	km/h	22	06-23-22
2135	mr96840	22:26:42	6月23日	6	Summer	Yes	上海紫竹国际教育园区	Clear	FALSE	20	C	5.6	km/h	22	06-23-22
2136	mr96133	22:26:42	6月23日	6	Summer	Yes	永多生活广场	Clear	FALSE	20	C	5.6	km/h	22	06-23-22
2137	mr96833	22:26:42	6月23日	6	Summer	Yes	水上园	Clear	FALSE	17.8	C	3.7	km/h	22	06-23-22
2138	mr96149	22:26:42	6月23日	6	Summer	Yes	图书馆	Clear	FALSE	17.8	C	5.6	km/h	22	06-23-22
2139	mr97332	22:26:42	6月23日	6	Summer	Yes	上海市第五人民医院	Clear	FALSE	17.8	C	5.6	km/h	22	06-23-22

图 3.28　某区共享单车各站点部分数据

充分利用各种可视化图形的优势，能够帮助我们更好地理解数据。例如，如图 3.23 所示，在某区图书馆站点，上午 8 时、下午 13 时和 17 时、晚上 21 时左右均出现了不同程度的骑行高峰，这可能是图书馆的客流与附近地铁 1 号线莘庄站的通勤客流交汇的缘故；如图 3.24 所示，7、8 月份某区图书馆站点的共享单车使用量明显高于 9、10 月份，这可能和学生们在暑假期间经常去图书馆借阅图书有关；如图 3.25 所示，从全年四个季节的共享单车使用量占比情况来看，春、秋两季的共享单车使用率较高，而夏、冬两季的使用率则较低，共享单车的使用情况与气温密切相关。

无论是分析特征间相关关系的散点图、特征间趋势关系的折线图，还是分析特征内部数据分布的柱状图、饼图以及特征内部数据分散情况的箱形图，数据可视化的目的就是直观、清晰地展现数据，所以选择合适的可视化图形尤为重要。

项　目　实　践

共享单车的精准投放和及时调配至今依然是共享单车经营企业所面临的难题。围绕“近三年共享单车租赁量变化”“用户骑行半径分布”“各站点用户租赁量比较”等特征，对本章第一节“一、数据采集”的项目实践中的数据进行可视化呈现，为有效管理共享单车提出合理建议。

1　选择合适的可视化方式，编写程序呈现数据分析结果，体会数据可视化的优势。

2　依据数据分析和可视化结果，针对上述问题，为有效管理共享单车提出合理建议，并说明理由。

作业练习

共享单车这种结合了移动支付、物联网、定位系统等多种技术，旨在满足城市出行“最后一公里”需求的新兴出行方式，受到了社会各界的关注。图 3.29是《年度共享单车行业发展分析报告》中的部分图示。

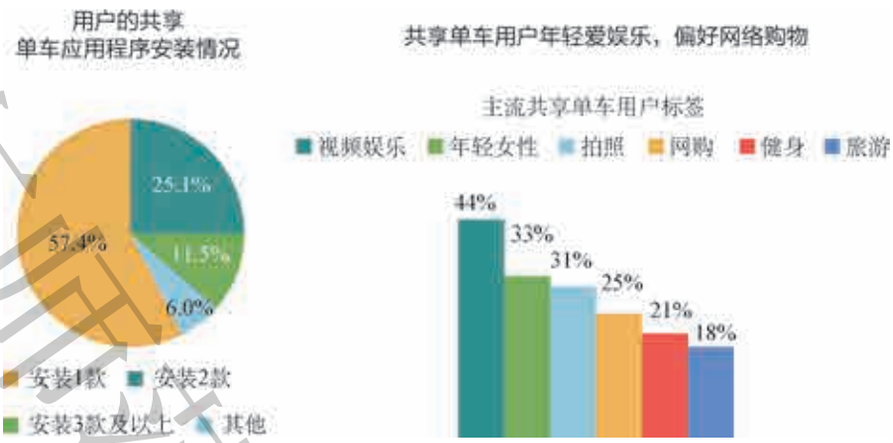


图 3.29 《年度共享单车行业发展分析报告》

思考分析报告中采集了什么数据，应用了哪些分析方法，选用了怎样的形式来呈现分析结果？把思考的结果填写在表 3.16中。

表 3.16 共享单车的数据分析

数据描述	数据处理		可视化呈现	
	方法	理由	形式	理由

第三节　数据分析报告与应用

运用数据来反映、研究和分析某项事物的现状、问题和原因，发现其本质和规律，得出分析的结论并给出解决方案，是数据分析过程和思路的最后呈现。一份完整的数据分析报告，应当围绕目标确定范围，遵循一定的架构，系统地反映事物数据分析的全貌，为决策者提供科学、严谨的依据。

体　验　思　考

任何新生事物都不可能是完美的，迅速火爆起来的共享单车也不例外。各种品牌的共享单车出现在人们眼前，逐渐融入人们的生活，它们被使用的频率越来越高，随之而带来的问题也越来越明显。图 3.30所示的词云图反映了当前消费者对共享单车关注的热点。

图 3.30　词云图

思考：

如何找准一个共享单车使用中的“痛点”问题，运用所学的数据分析方法，利用数据来反映、研究和分析问题的现状及原因，并提出解决问题的方法。

一、数据分析报告的种类

常用的数据分析报告有专题分析报告、综合分析报告和日常数据通报等。

专题分析报告是对社会现象的某一方面或某一个问题进行专门研究的一种数据分析报告，它的主要作用是为决策者制定某项政策、解决某个问题提供决策参考和依据。它具有两个特点：单一性和深入性，如《某年度共享单车新增移动应用程序注册用户专题分析报告》等。

综合分析报告是全面评价一个地区、单位、部门的业务或其他方面发展情况的一种数据分析报告。它具有两个特点：全面性和联系性，如《世界人口发展报告》《全国经济发展报告》《某年度共享单车运营分析报告》等。

日常数据通报是以定期数据分析报表为依据，反映计划执行情况，并分析其影响和形成原因的一种数据分析报告。它具有三个特点：进度性、规范性和时效性。

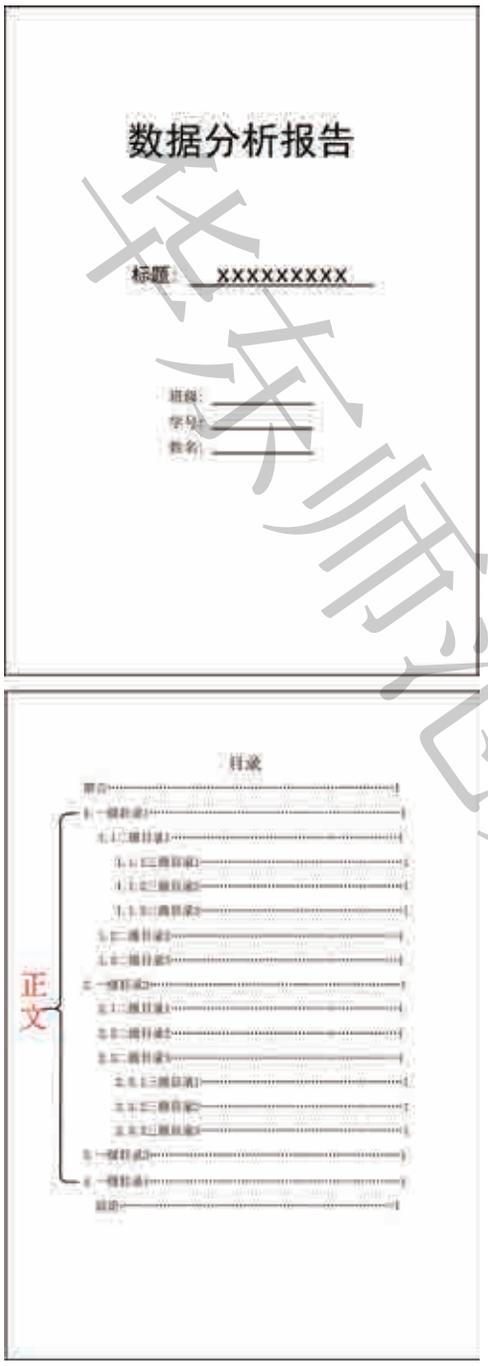


图 3.31 数据分析报告示例

二、数据分析报告的组成

数据分析报告依据不同的种类和不同的数据分析方法，其最后呈现的方式也可能会有所不同。数据分析报告通常由标题、目录、前言、正文、结论等组成，如图 3.31 所示。

标题是对数据分析报告的高度概括。标题不仅要体现数据分析的主题，并且能够激发读者的阅读兴趣。在命名标题时，可以直接在标题中呈现报告的结论，如《商业综合体类停车场日均使用率低，有较大优化空间》等；也可以在标题中提出分析研究的问题，如《商务写字楼停车场如何实施分时管理》等。

目录体现了数据分析报告的整体结构。读者通过目录可以快捷地找到所需的内容，所以在目录中要列出报告主要章节的名称，并附上对应的页码。

前言是数据分析报告的一个重要组成部分，主要阐述分析的背景和目的、需要解决的问题、运用的分析思路和方法、预期的效果或结论等。

正文是数据分析报告的核心部分。正文要系统地阐述数据分析的过程与结果，其中给出的事实、观点及分析论证必须严谨合理、逻辑性强。正文通常采用数据图表及文字相结合的方式，方便阅读和理解。

结论是对整个数据分析报告的总结，应包括依据数据分析结果得出的结论、建议和解决问题的方案等。结论要和正文相互衔接，与前言相互呼应。

完成一份数据分析报告通常需经历发现与界定问题、基于数据分析问题、基于数据分析给出解决问题的方案等过程。其中，发现与界定问题是关键，基于数据分析给出解决问题的方案是核心。在撰写数据分析报告时，首先需理清要解决什么问题，养成“先谋而后动”的习惯。

项 目 实 践

在完成本章第二节项目实践的基础上，小组成员明确分工，分步实施数据采集、整理、统计分析和制作可视化图示，体验数据处理的全过程；小组成员合作完成一份数据分析报告，并在班级中交流展示。

三、数据分析报告的价值

数据分析报告实质上是一种沟通与交流的形式，撰写数据分析报告的主要目的是呈现分析结果、可行性建议、问题解决方案等，其价值就在于使读者对结果做出正确的理解与判断，并可以根据其做出有针对性、操作性、战略性的决策。数据分析报告的价值如图 3.32 所示。

图 3.32　数据分析报告的价值

一份有效的数据分析报告能够为用户了解事物发展现状，有效判断所需解决问题的影响因素，有针对性地选择解决问题的方案，以及预判事物发展趋势提供数据支持和行动依据。例如，撰写《共享单车运营风险》数据分析报告，在数据分析的基础上描述城市共享单车的应用现状和存在的问题，找出引发问题的主要原因，针对问题和相关原因提出解决问题的建议和方案，为更好地管理城市共享单车提供依据与支持。

第四章

走近人工智能

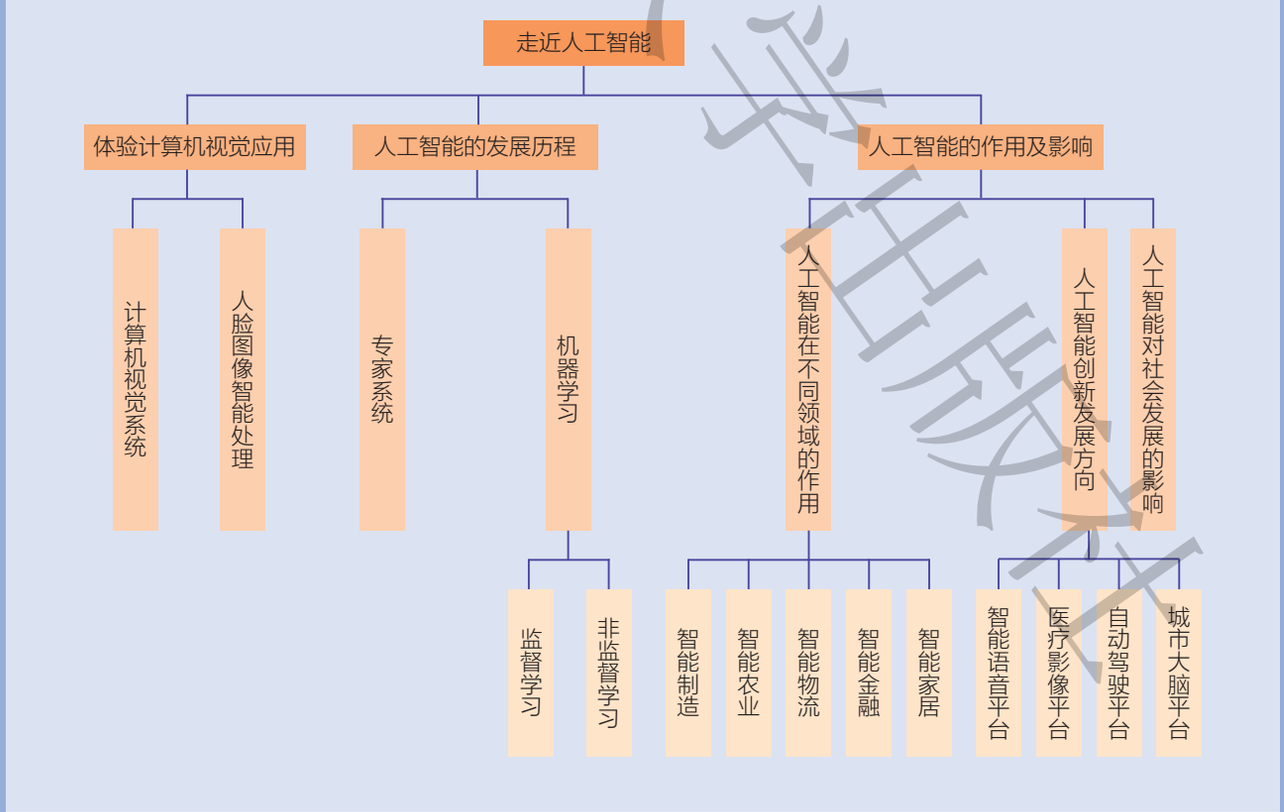
本章学习目标

- 体验借助人工智能平台实现人脸图像智能处理的过程，了解计算机视觉系统的作用和应用场景。
- 通过实例体验机器学习的基本过程，知道人工智能的发展历程。
- 认识人工智能在信息社会中的重要作用，感受人工智能产生的影响，了解人工智能的广泛应用可能会引发的社会问题及应对策略。

计算机真的能够具有“像人类一样思考”的能力吗？其实，人类对于机器智能的思考可以追溯到 1950年艾伦·麦席森·图灵（Alan Mathison Turing）在《智慧》（Mind）期刊发表的学术论文《计算机器与智能》（Computing Machinery and Intelligence）。1956年，约翰·麦卡锡（John McCarthy）等学者发起举行“人工智能夏季研讨会”，指出“人工智能”的研究目标是实现能模拟人类的机器，该机器能够使用语言，具有概念抽象和理解能力，能够完成人类才能完成的任务，并不断提高自身。

“人工智能”这一概念提出后，迅速发展成为一门广受关注的交叉和前沿学科。一方面，随着互联网的普及、物联网的渗透、大数据的形成、信息社区的崛起，数据和信息在人类社会、物理空间和信息空间之间逐渐交叉融合与相互作用；另一方面，新技术、新产业和新业态的不断涌现，也促使对人工智能基本理论和方法的研究出现新的变化。正是这些变化使得人工智能的新型应用呈现出勃勃生机。

本章知识结构



项目主题

智能工具好帮手

项·目·情·境

人工智能的发展深深地影响着人们的生活、学习和工作，它在给人们带来各种便利条件的同时，也对未来人类的思维方式、交流方式和工作方式的发展提出了挑战。

小申从小就对摄影感兴趣。近年来，小申发现不仅是数码相机，很多智能手机在拍摄人像的时候，都会用一个矩形框标记取景框中的人脸。也就是说，数码相机和智能手机已经能够和人类一样，通过“观察”和“思考”，判断出“有没有人”以及“人在哪里”。这是不是可以理解为一种由人创造的“类人智能”？小申经常在各种场合中听到“机器学习”这个概念。那么，机器如何学习？经过学习后，又能完成哪些任务呢？伴随着各种智能应用的普及，人工智能可能产生哪些社会问题，我们又应该如何应对呢？

项·目·任·务

任务 1

学习人工智能平台相关接口和图像处理模块的使用步骤，与同伴协作完成人脸识别功能简单应用的开发。

任务 2

体验使用监督学习方法实现鸢尾花分类的关键步骤，参考教材中的技术支持，与同伴协作完成鸢尾花识别程序的开发。

任务 3

采用小组活动方式，对“人工智能可能产生的社会问题及应对策略”进行研讨，记录每位小组同学的观点，完成一份研讨报告。

第一节　体验计算机视觉应用

视觉是人最重要的感觉，至少有80%的外界信息需要通过视觉来获取，如外界物体的大小、数量、颜色、动静等。在日常生活中，我们可以分辨出不同的动物，也可以识别出一间教室里有几个人，这主要是依靠我们的眼睛、视神经和大脑的视觉中枢实现的。眼睛用于成像，大脑对视网膜上的图像进行处理，最终得出反馈结果。

体　验　思　考

随着信息时代的发展，互联网上时时刻刻都在产生和传播着海量的图像，正是在这些海量数据和人工智能算法的帮助下，计算机的“视觉”正变得越来越“清晰”。因此，基于计算机视觉的应用被广泛推广，其中与我们生活最密切的恐怕要数人脸定位与人脸识别了。

当我们在使用数码相机或智能手机拍摄照片的时候，都会用到人脸检测功能，即使用矩形框将拟捕捉画面中的人脸标记出来，如图 4.1所示。有些应用软件还可以在成功定位人脸后，对人脸进行“美颜”处理，或者添加各种装饰。

图 4.1　拍照过程中的人脸自动检测

思考：对于计算机而言，无论是图像还是视频，都是一串由“0”和“1”构成的序列。那么计算机是如何在这些“0”和“1”中“找到”人脸的呢？

计算机视觉是一门研究如何使机器“看清”和“看懂”的学科，更进一步地说，就是指用图像采集设备和计算机代替人眼完成对目标的识别、跟踪和测量等工作。计算机视觉研究相关的理论和技术，从而构建能够从图像或多维数据中获取信息的人工智能系统。因为感知可以看作是从感官信号中提取信息，所以计算机视觉也可以看作是研究如何使人工智能系统能够从图像或多维数据中“感知”的科学。

人脸识别，是基于人的脸部特征信息进行身份识别的一种生物识别技术，有时也被称为人像识别、面部识别，也是计算机视觉重要的研究方向之一。广义的人脸识别包括构建人脸识别系统的一系列相关技术，如人脸图像采集、人脸检测（含定位）、身份确认以及身份查找等；而狭义的人脸识别特指通过人脸进行身份确认或者身份查找的技术或系统。

早在 20 世纪 50 年代，认知科学家就已经开展了人脸识别方面的研究。20 世纪 60 年代，人脸识别开始进入工程化应用。当时的方法主要利用了人脸的几何结构，通过分析人脸器官特征点及其之间的位置关系进行识别。这种方法简单直观，但是一旦人脸姿态、表情发生变化，则识别正确率严重下降。1991 年，著名的“特征脸”方法第一次将分析和统计特征引入人脸识别任务，在实用效果上取得了长足的进步。

进入 21 世纪，随着人工智能技术的发展，研究者开始关注在各种面部图像采集条件（不同的光照、不同的传感器以及是否进行了压缩等）或被拍摄者各种主观条件（面部的不同姿态、不同表情以及是否有遮挡等）下，是否都能成功进行识别。2014 年前后，随着大数据和深度学习的发展，神经网络技术在图像分类、手写体识别、语音识别等应用中获得了远超经典方法的成果，人脸识别应用也因此取得了突破性进展。

项 目 实 践

1.以小组为单位，围绕计算机视觉中的人脸检测与人脸识别开展调研，看看现在有哪些人工智能开发平台提供了这些功能？

2.通过查找资料、讨论交流，选择其中一个平台，了解该平台提供的人脸检测工具开发包（software development kit，SDK）的使用方法和操作步骤。

3.选择一张你喜欢的生活照，调用人工智能平台的人脸检测功能，定位照片中的人脸，使用技术支持中介绍的 Pillow 库，将照片中的人脸位置标记出来，如图 4.2所示。

图 4.2 人脸标记结果实例

技 术 支 持　使用 Pillow 库实现人脸标记

使用 Pillow 库，可以实现图像的缩放、切片、旋转以及画图等各种操作，而且操作方法非常简单。

```
# 导入 Pillow 库
from PIL import Image, ImageDraw
# 打开图像
im = Image.open(imageName)
```

```
# 生成一个可以用于画图的对象
draw = ImageDraw.Draw(im)
#使用人工智能平台返回的人脸坐标信息,在图像中人脸的位置画一个红色的矩形框
#用 left、top、width、height 分别表示人脸框左上角的横、纵坐标以及人脸框的宽度、高度
draw.rectangle((left, top, left + width, top + height), outline = (255,0,0))
# 将图像展示出来
im.show()
```

目前,有多个人工智能平台提供人脸检测和人脸识别应用的开发接口服务,不同平台返回的数据格式各不相同,但均能提供绘制人脸框所需的左上角位置横、纵坐标以及人脸框的宽度、高度信息。各平台的调用过程也大致相同,主要步骤如图 4.3所示。

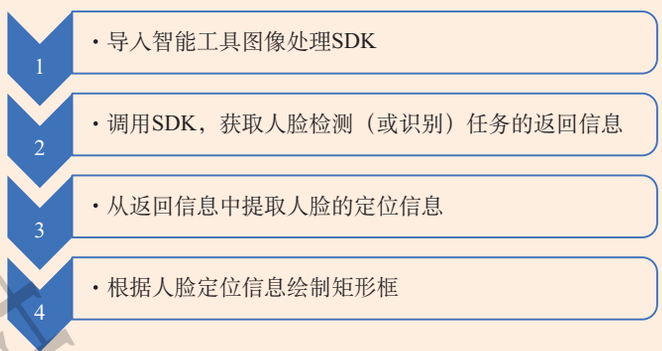


图 4.3 调用人工智能平台实现人脸标记主要步骤

作业练习

利用人工智能平台的人脸识别功能获得人脸位置后,请进一步获取人脸的性别,将不同性别的人脸用不同颜色的框标出。常见的性别表示方式如表 4.1所示。

表 4.1 人工智能平台返回的人脸性别信息

序号	返回值类型	说　明
1	字符串	“male”表示男性,“female”表示女性
2	整型数(0～99)	用整型数值表示性别的可能性,越接近 0,代表该人脸为女性的概率更大;越接近 99,代表该人脸为男性的概率更大
3	字符串+浮点数	除了用字符串“male”表示男性,“female”表示女性之外,还可以使用一个小于 1的正浮点数表示性别的可能性,数值越接近 1,表示该性别的可能性越大

例如,用绿色的框将女性的脸标出,用红色的框将男性的脸标出。在程序设计过程中,可以使用第一章中介绍过的 RGB颜色模型来描述想要的颜色。

知 识 延 伸　　人脸检测与人脸识别

人脸检测（face detection）的作用就是要检测出图像中人脸的所在位置，如图 4.4所示。人脸检测算法的输入是一张图像，输出是人脸框坐标序列，具体结果是 0个、1个或多个人脸框。输出的人脸框可以是正方形、矩形等。人脸检测算法的原理简单来说就是一个“扫描”加“判定”的过程，即首先在整个图像范围内扫描，再逐个判定候选区域是否是人脸。因此，人脸检测算法的计算速度会与图像尺寸大小及图像内容相关。在设计算法时，我们可以通过设置“输入图像尺寸”“最小脸尺寸限制”或“人脸数量上限”的方式来加速算法。

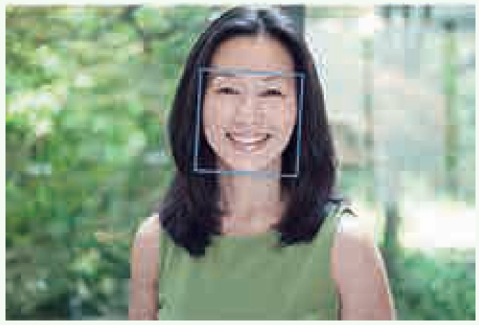


图 4.4　人脸检测

在人脸检测的基础上，可以实现人脸配准（也称为人脸关键点定位），如图 4.5所示。人脸配准算法的输入是“一张人脸图像”和“人脸坐标框”，输出是五官关键点的坐标序列。五官关键点的数量是预先设定好的一个固定数值，常见的有5点、68点、90点等。

当前效果比较好的人脸配准技术基本通过深度学习框架实现（关于机器学习与深度学习我们将在下一节中介绍）。这些方法都是基于人脸检测的结果，按某种事先设定的规则将人脸区域抠取出来，缩放到固定尺寸，然后进行关键点位置的计算。

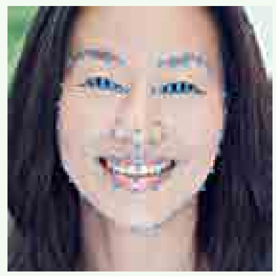


图 4.5　人脸关键点定位

人脸识别的目标是找出人脸图像所对应的身份。它的输入是一个人脸特征，通过和注册在库中的 N 个身份对应的特征进行逐个比对，找出“一个”与输入特征相似度较高的特征。将这个较高相似度值和预设的阈值进行比较，如果大于阈值，则返回该特征对应的身份，否则返回“不在库中”，如图 4.6所示。

图 4.6　1∶N 人脸识别

第二节　人工智能的发展历程

在不同的阶段，人类对“智能”的理解也不相同，因此历史上人工智能的定义也历经了多次转变。人工智能就是研究如何使计算机去做过去只有人类才能做的智能的工作，这是人工智能最通俗、简明的定义。这个定义反映了人工智能学科的基本思想和基本内容，即人工智能是研究人类智能活动规律，构造具有一定智能的人工系统，研究如何应用计算机的软、硬件来模拟人类某些智能行为的基本理论、方法和技术。另一种定义认为，人工智能是关于知识的学科——研究怎样表示知识、获得知识并使用知识的科学。这个定义突出了人工智能的知识基础，同时指出了计算机模拟智能行为的本质是模拟人类的知识行为能力，如图 4.7 所示。上述两种定义是人工智能领域中被广泛流传和认可的。本书中，我们采用下列定义：人工智能是指由人创造出来的，具有感知、认知、决策、学习、执行和社会协作能力，符合人类情感、伦理与道德观念的虚拟的或人工的系统。

图 4.7　人工智能的表现形式

体　验　思　考

2016年，阿尔法围棋（AlphaGo）在与围棋世界冠军李世石进行的围棋人机对弈中以 4∶1的总比分获胜，这是人工智能发展史上又一个新的里程碑。2017年，它又进化为阿尔法元（Alpha Zero），通过“自学成才”，仅用 3天就成为了围棋界的顶尖高手。

思考：

1. 请同学们分别了解互联网上不同的人机对弈平台，并体验人机对弈过程。
2. 以小组为单位进行讨论：在对弈过程中，机器是如何进行“思考”的。

一直以来，棋艺高超都被当作智力高超的象征，因此棋类博弈自古被视为一种代表人类智力的高级挑战。棋类博弈中的随机性和不可控因素要求对局双方的决策能更直接地控制整个局面的走势，进一步增强了智力的对抗性。这也是为什么，在每次有更好的人工智能程序面世时，被挑战的对象往往都是棋类。

1951 年，世界上诞生了第一个西洋跳棋程序。1962 年，西洋跳棋程序开始击败跳棋高手。这一时期的博弈程序基本上都是使用搜索的方式来求解问题，其中采用了多种算法技巧来提高搜索效率。

一、专家系统

专家系统最杰出的代表之一就是 1997 年战胜了国际象棋世界冠军的深蓝计算机(Deep Blue)。专家系统是早期人工智能的一个重要分支，它可以看作是一类具有专门知识和经验的计算机智能程序系统。这类系统就像在模仿人类专家做决定的过程，基于已经掌握的领域知识，根据推理规则得到相关结论，进而解决问题，因此专家系统也被称为基于知识的系统。专家系统适合于完成那些没有公认的理论和方法、数据不精确或信息不完整、人类专家短缺或专门知识十分昂贵的诊断、解释、监控、预测、规划和设计等任务。

医疗是专家系统的典型应用领域之一。中医药经过数千年的发展，积累了丰富的经验。我国在 2008 年研发出了基于知识的中医药对症开方专家系统。该系统能够在知识库的基础上，结合中药方剂理论及组方原则，为用户开出治疗特定病症的量化中药方剂，为医生及中药研发人员提供辅助决策支持，如图 4.8 所示。

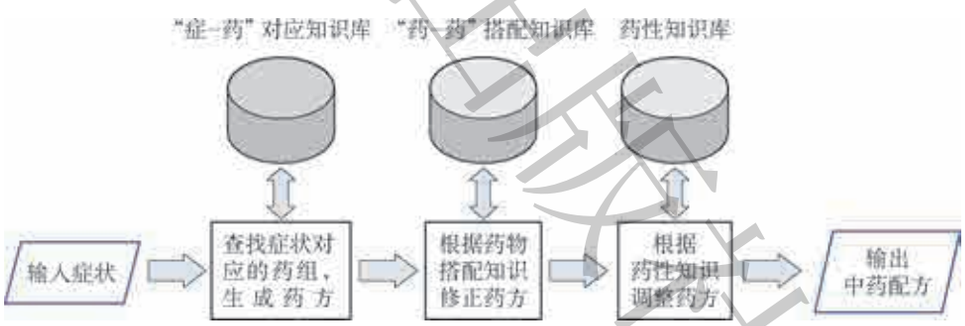


图 4.8　中医药对症开方专家系统工作流程示意图

二、机器学习

在人工智能战胜国际象棋世界冠军后，人们普遍认为围棋对局将是唯一一种计算机无法战胜人类的棋类博弈，因为围棋对局的变化最

图 4.9　围棋

为复杂，且博弈中还需要考虑到“形势判断”等模糊因素，如图 4.9 所示。然而在 2016 年至 2017 年间，阿尔法围棋（AlphaGo）和阿尔法元（Alpha Zero）先后战胜了代表国际最高水平的人类顶尖棋手，证明了机器学习技术发展带来的突破。

机器学习（machine learning）是人工智能的研究领域之一，其本质是基于互联网的海量数据以及计算机系统强大的运算能力，让机器自主模拟人类学习的过程，通过不断“学习”数据来做出智能决策行为。人工智能的研究历史有着一条从以“推理”为重点，到以“知识”为重点，再到以“学习”为重点的自然、清晰的脉络。显然，机器学习是实现人工智能的一个途径，即以机器学习为手段解决人工智能中的问题。

机器学习研究的主要目的是设计和分析一些让计算机可以自动“学习”的算法，使计算机从数据中自动分析获得规律，并利用规律对未知数据进行预测。机器学习的常见方法有监督学习和非监督学习等。

1. 监督学习

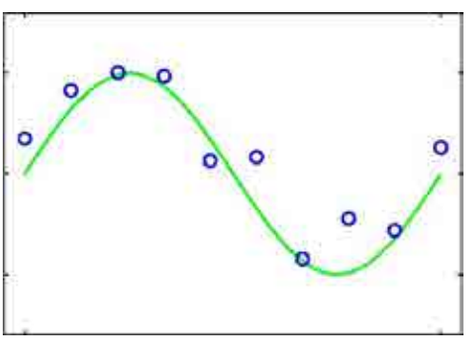


图 4.10　监督学习：回归

监督学习表示机器学习的数据是带标记的，这些标记可以包括数据类别、数据属性以及特征点位置等。以这些标记作为预期效果，不断地修正机器的预测结果。常见的监督学习有回归、分类。回归（regression）是将数据归到一条“线”上，即根据离散数据生成拟合曲线，因此其预测结果是连续的，如图 4.10 所示。分类（classification）是将一些实例数据分到合适的类别中，它的预测结果是离散的。图 4.11 给出了图像分类的一个简单应用示例。使用

图 4.11　图像分类应用示例

各种各样的猫和狗的图片构成监督学习训练集，其中的每张图片都已被正确标记为是猫还是狗。经过训练，可以得到基于监督学习的分类模型。当输入一张待分类图片时（如吐着舌头的猫），分类模型可以给出该图片的分类预测结果：该输入图片是猫的可能性为 0.923（92.3%），是狗的可能性为 0.231（23.1%）。

2. 非监督学习

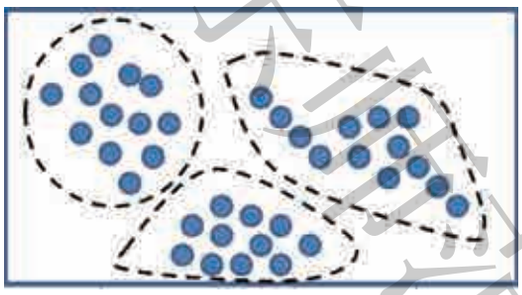


图 4.12　非监督学习：聚类

非监督学习表示机器学习的数据是没有标记的，机器需要从中探索并推断出潜在的联系。比如，在聚类（clustering）工作中，由于事先不知道数据类别，只能按照样本的某些属性将不同数据分开，把相似数据“聚合”成一类，即使得在同一类中的样本相似性尽可能大，不同类间的样本相似性尽可能小，如图 4.12 所示。例如，文本聚类根据文本的某种联系或相关性对文本集合进行有效的划分，方便人们从文档集中发现相关的信息。

项　目　实　践

鸢尾花又名蓝蝴蝶、紫蝴蝶，是生活中常见的一种观赏花。本任务的“原材料”是鸢尾花数据集，这是一个常用的分类实验数据集，其中包含了 150 条带标记（即标明花卉种类）的鸢尾花数据。数据集中的鸢尾花分为三类（如图 4.13所示），分别是山鸢尾、变色鸢尾和维吉尼亚鸢尾。每类 50条数据，每条数据包含四个属性：萼片长度、萼片宽度、花瓣长度、花瓣宽度，数据示例如表 4.2所示。

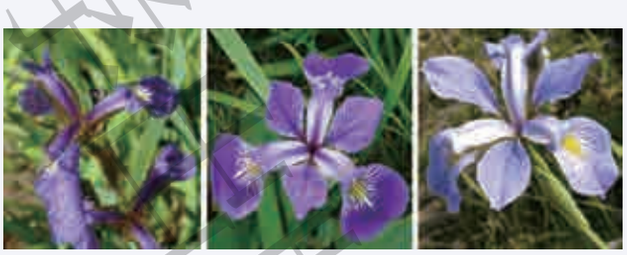


图 4.13　山鸢尾、变色鸢尾和维吉尼亚鸢尾

表 4.2　鸢尾花数据集示例

萼片长（厘米）	萼片宽（厘米）	花瓣长（厘米）	花瓣宽（厘米）	类别
5.1	3.5	1.4	0.2	山鸢尾
4.9	3	1.4	0.2	山鸢尾
4.7	3.2	1.3	0.2	山鸢尾
6.3	2.3	4.4	1.3	变色鸢尾

萼片长(厘米)	萼片宽(厘米)	花瓣长(厘米)	花瓣宽(厘米)	类别
5.6	3	4.1	1.3	变色鸢尾
5.5	2.5	4	1.3	变色鸢尾
6.4	3.1	5.5	1.8	维吉尼亚鸢尾
6	3	4.8	1.8	维吉尼亚鸢尾
6.9	3.1	5.4	2.1	维吉尼亚鸢尾

1 假设现在有一株鸢尾花，经测量后得知萼片长 6.7厘米，萼片宽 3厘米，花瓣长 5.1厘米，花瓣宽 1.8厘米，可以使用什么类型的机器学习方法，来判断它属于哪种鸢尾花？

2 请使用技术支持中介绍的方法，完成对鸢尾花类型的自动判别。

技 术 支 持　判断鸢尾花类型的方法

鸢尾花数据集中的每一条数据，都是一条带（分类）标记的数据，可以使用监督学习方法来尝试解决这个问题。如图 4.14所示，将 150条数据分成两个部分，其中 80%（三种各 40条）作为训练集，20%（三种各 10条）作为测试集，用来判断分类方法的准确率。在分类过程中，使用最简单的欧氏距离来判断，测试数据属于哪种类型的花。这里使用的欧氏距离是最容易、最直观的距离度量方法，我们在数学课中接触到的两

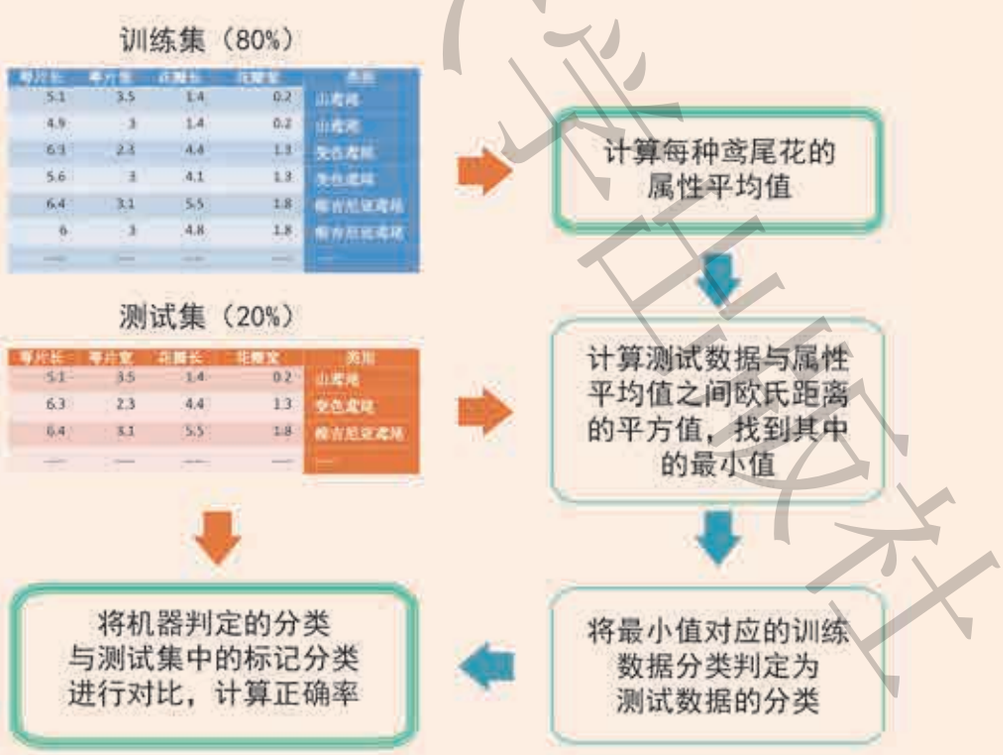


图 4.14　鸢尾花活动任务过程示意图

个点在空间中的距离一般都是指欧氏距离，例如二维平面上点 $a(x_1,y_1)$ 与 $b(x_2,y_2)$ 之间的欧氏距离为：$d_{12}=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$。具体到鸢尾花数据集，每条数据有四个属性，相当于四个维度，所采用的欧氏距离公式为：$d(x,y)=\sqrt{\sum_{i=1}^{4}(x_i-y_i)^2}$。在本程序中，由于仅需要比较大小（找到与测试数据距离最短的鸢尾花属性均值所对应的分类），并不需要输出具体距离值，因此只需要计算欧氏距离公式中的平方和，可以省略平方根的计算。

具体实现过程中的关键步骤及核心代码如下。

第一步：数据文件准备。将 120条训练数据和 30条测试数据分别存入 iris_training.csv和 iris_testing文件，两个文件的第一行为每一列数据项对应的标题，依次为 se_len、se_wid、pe_len、pe_wid和 classification，分别代表训练数据的萼片长度、萼片宽度、花瓣长度、花瓣宽度和实际分类；之后每一行为一条训练数据或测试数据。属性标题之间、属性数值之间均使用英文逗号分隔，如图 4.15所示。

```
se_len,se_wid,pe_len,pe_wid,classification
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
6.3,2.3,4.4,1.3,Iris-versicolor
5.6,3,4.1,1.3,Iris-versicolor
5.5,2.5,4,1.3,Iris-versicolor
6.4,3.1,5.5,1.8,Iris-virginica
6,3,4.8,1.8,Iris-virginica
6.9,3.1,5.4,2.1,Iris-virginica
.....
```

图 4.15　使用 CSV文件存储鸢尾花训练数据

第二步：初始化。导入 Pandas库，以实现对 CSV文件的读取，并对程序中用到的常量和变量进行初始化赋值，主要流程如图 4.16所示。

```
#将山鸢尾、变色鸢尾和维吉尼亚鸢尾的名称使用列表 iris_type 进行存储
iris_type = ['Iris-setosa', 'Iris-versicolor', 'Iris-virginica']

#初始化三个列表，分别存储三种鸢尾花每个属性的总值和训练样本数量
setosa_sum = [0,0,0,0,0]        #山鸢尾
versicolor_sum = [0,0,0,0,0]    #变色鸢尾
virginica_sum = [0,0,0,0,0]     #维吉尼亚鸢尾

#设置四个常量，分别代表三个列表的索引位所表示的含义
se_len = 0          #列表第 0 位代表萼片长度
se_wid = 1          #列表第 1 位代表萼片宽度
```

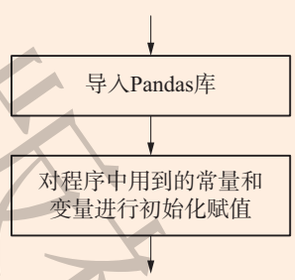


图 4.16　初始化阶段的主要流程

```
pe_len = 2      # 列表第 2 位代表花瓣长度
pe_wid = 3      # 列表第 3 位代表花瓣宽度
amount = 4      # 列表第 4 位代表鸢尾花训练样本的数量
```

第三步：训练。通过 Pandas库读入 iris_traing.csv文件中的训练数据并循环逐条处理，对四个属性分别进行累加求和，同时计算每种类型鸢尾花的训练样本数量，分别存储在 setosa_sum、versicolor_sum 和 virginica_sum 三个列表中，主要流程如图 4.17所示。

```
# 通过 Pandas 库的 read_csv 函数读入训练集数据文件 'iris_training.csv'
trainData = pd.read_csv('iris_training.csv')

# 循环处理训练集 CSV 文件中的每一条训练数据
# 分别累加计算每种类型鸢尾花的四个属性的总值以及训练样本数量
for index, row in trainData.iterrows():
    if row.classification == 'Iris-setosa':
        setosa_sum[se_len] += row.se_len          # 萼片长度
        setosa_sum[se_wid] += row.se_wid          # 萼片宽度
        setosa_sum[pe_len] += row.pe_len          # 花瓣长度
        setosa_sum[pe_wid] += row.pe_wid          # 花瓣宽度
        setosa_sum[amount] += 1                   # 样本数量
    elif row.classification == 'Iris-versicolor':
        versicolor_sum[se_len] += row.se_len
        versicolor_sum[se_wid] += row.se_wid
        versicolor_sum[pe_len] += row.pe_len
        versicolor_sum[pe_wid] += row.pe_wid
        versicolor_sum[amount] += 1
    elif row.classification == 'Iris-virginica':
        virginica_sum[se_len] += row.se_len
        virginica_sum[se_wid] += row.se_wid
        virginica_sum[pe_len] += row.pe_len
        virginica_sum[pe_wid] += row.pe_wid
        virginica_sum[amount] += 1
```

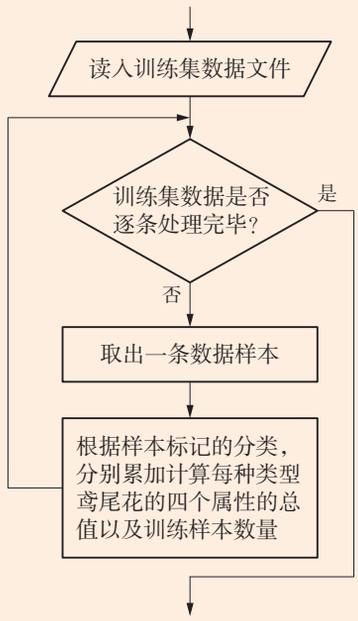


图 4.17　训练阶段的主要流程

其中，在 for循环中使用 Pandas库提供的 iterrows()方法，可以在循环中每次返回 CSV文件中的一行数据，并分别将该行的序号和内容存储到变量 index和 row 中。配合使用 CSV文件中第一行的标题名，即可分别访问每个数据项。例如，row.se_len表示训练数据对应的萼片长度，row.pe_wid表示训练数据对应的花瓣宽度等。

第四步：预测。通过 Pandas库读入 iris_testing.csv文件中的测试数据并循环逐条处理。对于每条测试数据，分别计算其与三种鸢尾花属性均值的欧氏距离的平方值，并存储在列表 distance 中。因为在训练阶段中，我们只计算了各属性的总值，所以在这里需要使用属性的总值除以样本数量，以获得相应的均值。之后，根据 distance列表中最小值元素的序号，在 iris_type列表中找到对应的分类信息，将其作为预测结果输出，主要流程如图 4.18 所示。

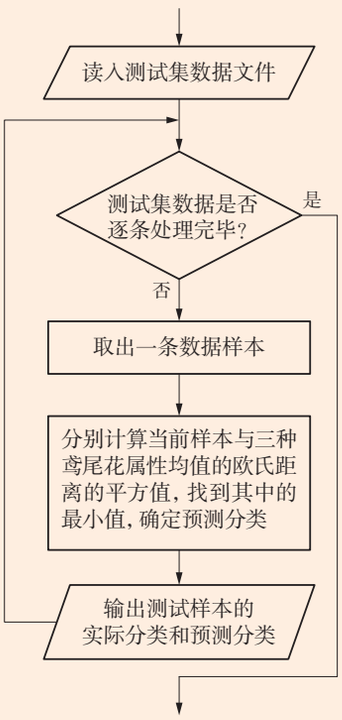


图 4.18　预测阶段的主要流程

```
#读入测试集数据文件'iris_testing.csv'
TestData = pd.read_csv('iris_testing.csv')

#循环读取测试集 CSV 文件中的每一条训练数据并进行处理
for index, row in TestData.iterrows():

    #分别计算输入数据与三种鸢尾花属性均值的欧氏距离平方值
    #存储在列表 distance 中
    distance = []
    distance.append((row.se_len-setosa_sum[se_len]/setosa_sum[amount]) ** 2\
                    + (row.se_wid-setosa_sum[se_wid]/setosa_sum[amount]) ** 2\
                    + (row.pe_len-setosa_sum[pe_len]/setosa_sum[amount]) ** 2\
                    + (row.pe_wid-setosa_sum[pe_wid]/setosa_sum[amount]) ** 2)

    distance.append((row.se_len-versicolor_sum[se_len]/versicolor_sum[amount]) ** 2\
                    + (row.se_wid-versicolor_sum[se_wid]/versicolor_sum[amount]) ** 2\
                    + (row.pe_len-versicolor_sum[pe_len]/versicolor_sum[amount]) ** 2\
                    + (row.pe_wid-versicolor_sum[pe_wid]/versicolor_sum[amount]) ** 2)

    distance.append((row.se_len-virginica_sum[se_len]/virginica_sum[amount]) ** 2\
                    + (row.se_wid-virginica_sum[se_wid]/virginica_sum[amount]) ** 2\
                    + (row.pe_len-virginica_sum[pe_len]/virginica_sum[amount]) ** 2\
                    + (row.pe_wid-virginica_sum[pe_wid]/virginica_sum[amount]) ** 2)

    #获取最小的欧氏距离的平方值
    min_distance = min(distance)
```

```
    # 获取最小值的序号
    idx = distance.index(min_distance)

    # 将 iris_type 列表中对应序号位置的分类信息作为预测分类结果
    # 打印当前样本的实际分类和预测分类
    print('实际分类:', row.classification, ';预测分类:', iris_type[idx])
```

输出结果如图 4.19所示。从输出结果中可以看出,有一条维吉尼亚鸢尾数据被错误地分类到了变色鸢尾中,因此共有 29条测试数据被正确分类,分类正确率为:29÷ 30× 100% = 96.67%。图 4.20给出了数据集中所有训练数据的分布,其中蓝色代表山鸢尾数据,橙色代表变色鸢尾,绿色代表维吉尼亚鸢尾,三种鸢尾数据的均值使用比数据点大的圆形标出。图中红色 X为未正确分类的测试数据,从图中可以看出,其距离橙色变色鸢尾的均值比距离绿色维吉尼亚鸢尾的均值要近,故被错判为变色鸢尾。

```
实际分类: Iris-setosa ; 预测分类: Iris-setosa
实际分类: Iris-setosa ; 预测分类: Iris-setosa
实际分类: Iris-setosa ; 预测分类: Iris-setosa
实际分类: Iris-setosa ; 预测分类: Iris-setosa
实际分类: Iris-setosa ; 预测分类: Iris-setosa
实际分类: Iris-setosa ; 预测分类: Iris-setosa
实际分类: Iris-setosa ; 预测分类: Iris-setosa
实际分类: Iris-setosa ; 预测分类: Iris-setosa
实际分类: Iris-setosa ; 预测分类: Iris-setosa
实际分类: Iris-setosa ; 预测分类: Iris-setosa
实际分类: Iris-versicolor ; 预测分类: Iris-versicolor
实际分类: Iris-versicolor ; 预测分类: Iris-versicolor
实际分类: Iris-versicolor ; 预测分类: Iris-versicolor
实际分类: Iris-versicolor ; 预测分类: Iris-versicolor
实际分类: Iris-versicolor ; 预测分类: Iris-versicolor
实际分类: Iris-versicolor ; 预测分类: Iris-versicolor
实际分类: Iris-versicolor ; 预测分类: Iris-versicolor
实际分类: Iris-versicolor ; 预测分类: Iris-versicolor
实际分类: Iris-versicolor ; 预测分类: Iris-versicolor
实际分类: Iris-versicolor ; 预测分类: Iris-versicolor
实际分类: Iris-virginica ; 预测分类: Iris-virginica
实际分类: Iris-virginica ; 预测分类: Iris-virginica
实际分类: Iris-virginica ; 预测分类: Iris-versicolor
实际分类: Iris-virginica ; 预测分类: Iris-virginica
实际分类: Iris-virginica ; 预测分类: Iris-virginica
实际分类: Iris-virginica ; 预测分类: Iris-virginica
实际分类: Iris-virginica ; 预测分类: Iris-virginica
实际分类: Iris-virginica ; 预测分类: Iris-virginica
实际分类: Iris-virginica ; 预测分类: Iris-virginica
实际分类: Iris-virginica ; 预测分类: Iris-virginica
```

图 4.19 输出结果

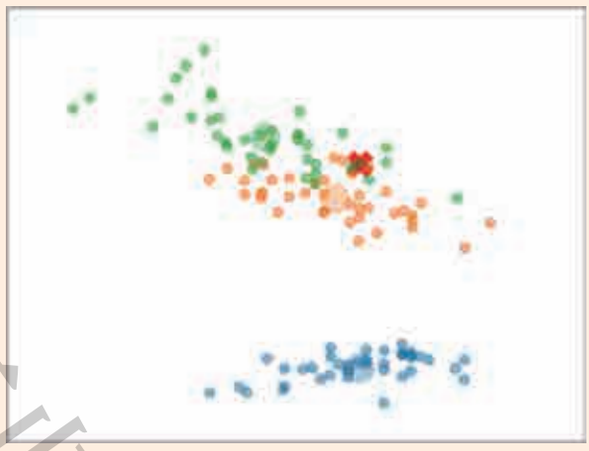


图 4.20　鸢尾花分类示意图

作业练习

K-近邻（KNN，k-nearest neighbor）算法是机器学习算法中比较基础和简单的算法之一，经常被用于分类任务。它的基本原理是：找到离测试样本最近的 K 条已标记训练数据，将其中最多的类别作为测试样本的类别。

如图 4.21所示，当 K 取 1时，测试样本（图中黄色三角形）周围最近的已标记训练数据点为其右上角的绿色正方形，则将其归为绿色正方形所属分类；当 K 取 2时，距离测试样本最近的两个训练数据点均为绿色正方形，则同样将其归为绿色正方形所属分类；当 K 取 3时，距离测试样本最近的三个训练数据点包括两个绿色正方形和一个红色圆形，则仍将其归为绿色正方形所属分类；当 K 取 5时，距离测试样本最近的五个训练数据点包括两个绿色正方形和三个红色圆形，则将其归为红色圆形所属分类。与其他监督学习分类方法相比，K-近邻方法并没有显式的训练过程，而是先把训练样本保存起来，待收到测试样本后再进行处理。

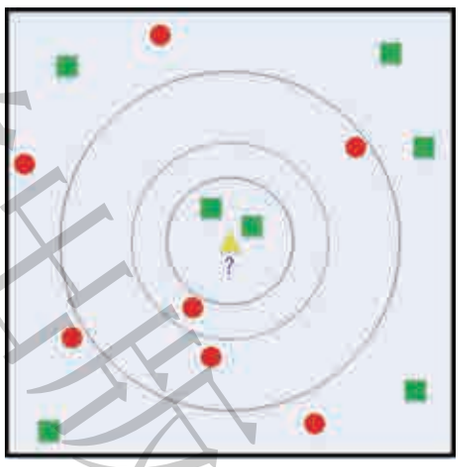


图 4.21　K-近邻算法示意图

请修改鸢尾花分类程序，利用 K-近邻的思想，取 K = 1，计算刚才未正确分类的那条测试样本与所有已标记训练数据之间的欧氏距离，找到距离最短的那条训练数据，将其分类作为测试样本的分类，看看是否能够正确分类。

知 识 延 伸

机器学习的其他方法

机器学习的其他常见方法还包括半监督学习和强化学习。

与监督学习相比，半监督学习的输入数据中仅有部分数据已被标记，即存在未被标记的输入数据。这种学习模型首先需要学习数据的内在结构，以便合理地组织数据来进行预测。其应用场景同样包括回归和分类，如图 4.22所示。

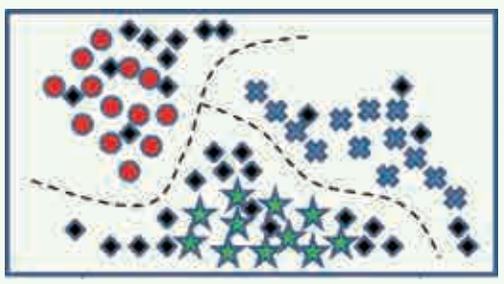


图 4.22 半监督学习：分类

如图 4.23所示，强化学习是带激励机制的。如果机器行动正确，将给予一定的“正激励”；如果行动错误，也同样会给出一个惩罚（也可称为“负激励”）。在这种情况下，机器将会考虑如何在一个环境中行动才能达到激励的最大化，具有一定的动态规划思想。强化学习最有代表性的一个应用便是阿尔法围棋的升级产品——阿尔法元。阿尔法元舍弃了先验知识，不再需要人为设计特征，直接将棋盘上黑白棋子的摆放情况作为原始数据输入到模型中，使用强化学习来自我博弈，不断提升自己，最终出色地完成了棋局。阿尔法元的成功证明了在没有人类的经验和指导下，深度强化学习依然能够出色地完成指定的任务。

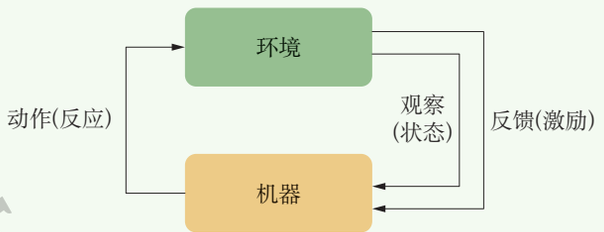


图 4.23 强化学习

近年来，机器学习方面取得了诸多突破性进展，深度学习也是其中之一。深度学习的概念最早于 2006年提出，它指的是基于样本数据，通过一定的训练方法，得到包含多个层级的深度网络结构的机器学习过程。深度学习的优势在于可以采用非监督式或半监督式的高效算法来替代人工方式获取特征。对于不同的观测值（如一幅图像），既可以用像素值来表示，也可以更抽象地用一系列边或特定形状的区域等来表示。使用某些特定的表示方法时，能够更容易通过样本数据实现学习任务（如人脸识别或面部表情识别）。

深度学习的“深度”是相对机器学习中浅层学习的方法而言的。浅层学习的方法通常采用只有一个输入层、一个隐藏层和一个输出层的神经网络，而在深度学习的模型中，可能包含多个隐藏层，如图 4.24所示。迄今为止，已经出现了多种深度学习框架，如深度神经网络、卷积神经网络、深度置信网络和递归神经网络等，这些方法已被成功应用于计算机视觉、语音识别、自然语言处理等领域。

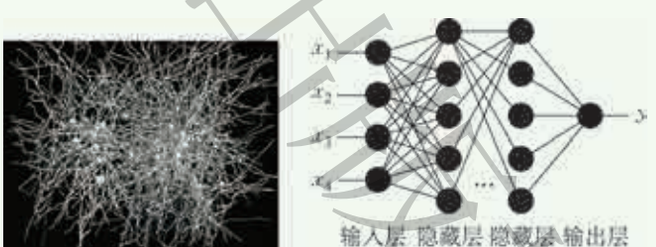


图 4.24 生物神经网络与深度学习神经网络

第三节　人工智能的作用及影响

人工智能是我国新一轮科技革命和产业变革的重要驱动力量。2017 年，国务院出台的《新一代人工智能发展规划》中提出："到 2020 年人工智能总体技术和应用与世界先进水平同步；到 2030 年人工智能理论、技术与应用总体达到世界领先水平，成为世界主要人工智能创新中心。"

体　验　思　考

2018年 4月 "博鳌亚洲论坛 2018年年会"在海南博鳌拉开帷幕，来自世界各国的两千多位嘉宾汇聚一堂。本次论坛官方首次使用人工智能翻译机（如图 4.25所示），实现了中文和英语、日语、韩语、法语、西班牙语等多种语言的实时互译，为现场嘉宾提供服务。可以预见，随着核心关键技术的突破，大多数人工翻译工作（包括笔译、口译甚至是同声传译）将会被机器全部或者部分取代。

图 4.25　人工智能翻译机

思考：

1. 人工智能应用在哪些方面具有人类所不具备的优势？随着人工智能技术的发展，哪些行业或工种会受到比较大的冲击？又会出现哪些新的行业？

2. 与人工智能应用相比，人类自身有哪些优势？人类与人工智能应该建立什么样的合作关系？

一、人工智能在不同领域发挥的作用

为了加快推进产业智能化升级，推动人工智能与各行业融合创新，我国将在制造、农业、物流、金融、家居等重点行业和领域开展人工智能应用试点示范，推动人工智能规模化应用，全面提升产业发展智能化水平。

1. 智能制造

智能制造旨在围绕建设制造强国重大需求，推进制造系统集成应

图 4.26　智能制造

用，研发制造服务平台，推广新型制造模式，建立智能制造标准体系，推进制造全生命周期活动智能化，如图 4.26 所示。

在国内某智能手机制造商的自动化生产线上，从送料开始到包装出货，每隔 28.5 秒就可以生产出一台智能手机。生产的全过程采用智能化管理，包括向生产基地运货、物料自动仓储入库、生产线自动提取配件、成品出库等各个环节。生产组装过程中，采用人机结合的方式，由机器完成其中的大部分工作。同样贯穿在全部生产过程中的还有生产数据的可视化管理，每一台生产设备、每一件物料，甚至是每一位员工，都可以转化为一个可视化的节点。

2. 智能农业

图 4.27　智能农业

智能农业建设的主要工作包括：研制农业智能传感与控制系统、智能化农业装备、农机田间作业自主系统等，并在此基础上建立典型农业大数据智能决策分析系统，开展智能农场、智能化植物工厂、智能牧场、智能渔场、智能果园、农产品加工智能车间、农产品绿色智能供应链等集成应用示范，如图 4.27 所示。

以往种植桃树的农民在收获后需要通过人力分拣桃子，不但所需工时长，而且只能凭手感和肉眼观察进行分拣，误差较大。除了重量，桃子的外观、色泽、是否有伤疤等品相指标同样重要，人工挑选时难免会出现看错的情况。2018 年，有些地区的果农用上了智能大桃分拣机，分拣机结合了开源平台提供的人工智能技术，能根据桃子的大小、颜色、形状等自动分拣，省时省工。有了这台机器，往年琐碎繁重的桃子分拣工作变得轻松省力了。

图 4.28　智能物流

3. 智能物流

智能物流通过加强智能化装卸搬运、分拣包装、加工配送等智能物流装备的研发和推广应用，建设深度感知智能仓储系统，提升仓储运营管理水平和效率，如图 4.28 所示。

在国内某公司正着力打造的智慧物流中心里，从入库、在库到拣货、分拣、装车，整个过程都无需人力

参与，使得仓储管理拥有极高的效率和出色的灵活性。这种以“无人仓”作为载体的全新一代智能物流技术，其核心特色体现为数据感知、机器人融入和算法指导生产，可以全面改变目前仓储的运营模式，极大提升效率并降低人力消耗。与传统的仓储模式相比，“无人仓”在运营效率、灵活性、吞吐量等方面跨上了一个新的台阶。

4. 智能金融

智能金融借助金融大数据系统，提升金融多媒体数据的处理与理解能力，创新智能金融产品和服务，发展金融新业态。通过在金融行业应用智能客服、智能监控等技术和装备，可以有效建立金融风险智能预警与防控系统。

图 4.29　智能金融

随着互联网的发展，一些不法分子利用商家规则和技术的漏洞来牟取暴利，严重影响了普通用户的权益和体验。从 2016 年至今，大数据风险控制相关产品已成功应用于电商、外卖、网约车、共享单车、医院、互联网金融等行业，实现对营销作弊、身份冒用等主要金融风险的自动识别。商家可以根据风险控制产品提供的分析，限制风险用户的购买、下单等行为，保障真正用户的权益，如图 4.29 所示。

图 4.30　智能家居

5. 智能家居

加强人工智能技术与家居建筑系统的融合应用，研发适应不同应用场景的家庭互联互通协议、接口标准，提升家电、耐用品等家居产品的感知和联通能力，能够有效提升建筑设备及家居产品的智能化水平，如图 4.30 所示。

如今的智能家居产品早已不再停留在概念阶段。智能语音助手的快速崛起已经成为连接智能家居设备的重要“入口”，可以实现包括家居设备控制在内的多种语音操作功能。而智能安防产品的不断发展，则明显提升了生活质量。

二、人工智能创新发展方向

2017 年 11 月 15 日，国家科技部召开了新一代人工智能发展规划暨

重大科技项目启动会，会上宣布了首批国家新一代人工智能开放创新平台，分别是智能语音平台、医疗影像平台、自动驾驶平台和城市大脑平台。

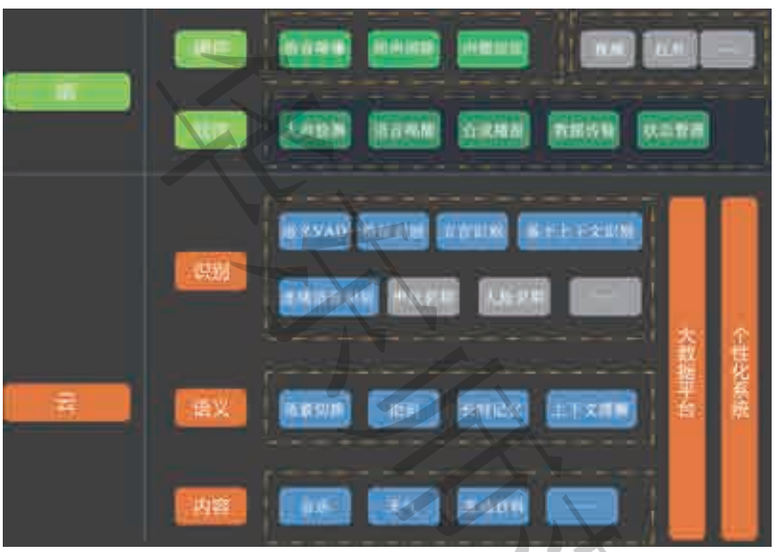


图 4.31　智能语音平台产品框架

1. 智能语音平台

智能语音旨在实现人机交互无障碍，使人与机器之间可以通过语音、图像、手势等自然交互方式，进行持续、双向、自然的沟通。智能语音平台并不是面向某个具体应用场景而设计的，它可以应用于诸多领域，如智能电视、可穿戴设备、智能车载设备、移动应用中的语音交互服务等，如图 4.31 所示。

图 4.32　人工智能辅助诊疗系统功能示意

2. 医疗影像平台

医疗影像平台是一个医疗机构、科研团体、器械厂商、高等院校、公益组织等多方参与的开放平台，是由图像识别、大数据处理、深度学习等领先技术与医学跨界融合研发而成，其目的是共同推进人工智能技术在医学影像、辅助诊断、医疗机器人等众多医疗环节的探索和应用。例如，基于人工智能技术的辅助诊疗系统主要功能如图 4.32 所示。

3. 自动驾驶平台

自动驾驶平台将帮助汽车行业及自动驾驶领域的合作伙伴，结合车辆和硬件设备，快速搭建一套自动驾驶系统。平台拥有海量数据仿真引擎和基于深度学习的自动驾驶算法，并且能够为用户提供覆盖广、高自动化的高精度地图服务。由此，用户可以更快地研发、测试和部署自动驾驶车辆，平台也会在用户的参与下变得越来越完善。

4. 城市大脑平台

城市大脑平台可以对整个城市进行全局实时分析，自动调配公共资源，修正城市运行中的漏洞，成为未来城市的基础设施。在已经部署城市大脑平台服务的地区，利用视频巡检替代人工巡检，日报警量多达 500 余次，识别准确率在 92%以上；高架车辆道路通行时间缩短 15%；采用信号灯自动配时的路段，平均道路通行速度提升 15%；应急车辆到达时间节省 50%。城市大脑平台最理想的状态就是帮助我们更高效地治理社会，如图 4.33 所示。除了交通治理之外，城市大脑平台还将能源、供水等基础设施做数据化处理，以节约更多的资源，实现城市的有效管理。

图 4.33　城市事件感知与智能处理

三、人工智能对社会发展的影响

人工智能已经开始逐渐与各领域紧密结合，渗透人们日常生活的方方面面，极大地提高了人们的工作效率和服务水平。在一些具有确定目标的任务中，人工智能已经表现出了更加优秀的性能。例如，某公安分局使用自动人脸识别系统后，在 40 个工作日内辨认出 69 名嫌疑人，相比人工识别的效率提升了近 200 倍。2018 年 6 月，在由我国举办的全球神经影像人工智能人机大赛总决赛中，首次与公众正式见面的神经影像人工智能辅助诊断系统以高出约 20%的准确率战胜了“人类战队”。比赛中的“人类战队”由 25 名全球神经影像领域专家、学者、优秀临床医生组成，而人工智能系统则基于某医院近十年来接诊的数万余神经系统相关疾病病例影像数据研发。人工智能技术产生的巨大推动力，促使人类社会的各方面都在发生着剧烈的变化。

人工智能应用的目的是将人类从部分脑力劳动中解放出来。在计算能力和数据储存方面，人工智能的优势毋庸置疑，一些原来由人工方式通过相对简单的重复性劳动来完成的任务，已经开始被人工智能应用所取代。但与此同时，新的工作岗位也在不断出现，并且对从业者提出了更高的素质要求。特别是一些非标准化、不确定性强的任务，需要从业者具有更加系统性、创造性和创新性的思维方式以及综合应用多方面知识解决问题的能力。

项 目 实 践

智慧城市，并不仅仅是为城市增加一些智能化系统，更是将人工智能和大数据技术应用于城市的发展和优化，形成城市交通、城市医疗、工业制造、农业生产等各个方面的智能化发展，如图 4.34所示。智慧城市中的许多人工智能应用都以大数据分析为基础，这其中涉及了大量的个人私密信息。如何在提供智能服务的同时，确保个人私密信息的安全，这是部署和实施人工智能应用必不可少的前提。

随着人工智能在各行各业中的普及与应用，包括个人隐私泄露在内，究竟会引发哪些社会问题？请以小组为单位，围绕人工智能可能产生的社会问题及应对策略，展开讨论并撰写报告，交流分享学习成果。

图 4.34 智慧城市

知 识 延 伸

新一代人工智能发展方向

为构建开放协同的人工智能科技创新体系，围绕增加人工智能创新的源头供给，从前沿基础理论、关键共性技术、基础平台、人才队伍等方面强化部署，促进开源共享，系统提升持续创新能力，确保我国人工智能科技水平跻身世界前列，为世界人工智能发展做出更多贡献，需要建立新一代人工智能基础理论体系。2017年初，中国工程院院刊信息与电子工程学部分刊《信息与电子工程前沿（英文）》发表了学术论文《人工智能 2.0》（*Special Issue on Artificial Intelligence 2.0*），对新一代人工智能中所涉及的大数据智能、群体智能、跨媒体智能、混合增强智能和自主智能系统等进行了阐述，如图 4.35所示。

新一代人工智能中五大智能方向		将推动人工智能发生的跃变
大数据智能	⇨	从人工知识表达到大数据驱动知识学习
群体智能	⇨	从处理类型单一的数据到跨媒体认知、学习和推理
跨媒体智能	⇨	从追求机器智能到迈向人机混合增强智能
混合增强智能	⇨	从聚焦研究个体智能到基于互联网络的群体智能
自主智能系统	⇨	从机器人到自主系统

图 4.35 新一代人工智能的五大智能方向

新一代的人工智能有以下几个跃变：一是从人工知识表达到大数据驱动的知识学习技术；二是从分类型处理的多媒体数据转向跨媒体的认知、学习、推理；三是从追求智能机器到高水平的人机、脑机相互协同和融合；四是从聚焦个体智能到基于互联网和大数据的群体智能，它可以把很多人的智能集聚融合起来变成群体智能；五是从拟人化的机器人转向更加广阔的自主智能系统，并不是一个单纯的机器人才叫人工智能。

后　记

本册教科书依据教育部《普通高中信息技术课程标准（2017 年版 2020 年修订）》编写，并经国家教材委员会专家委员会审核通过。全体编写人员认真领会国家基础教育改革精神，精心研究当代信息社会的人才培养要求，广泛调研上海及各地高中信息技术教育的现状和挑战，深入了解高中学生的学习需求，并汲取了上海市《普通高中信息科技（试用本）》的编写经验。

编写过程中，上海市中小学（幼儿园）课程改革委员会专家工作委员会，上海市教育委员会教学研究室，上海市课程方案教育教学研究基地、上海市心理教育教学研究基地、上海市基础教育教材建设研究基地、上海市信息技术教育教学研究基地（上海高校“立德树人”人文社会科学重点研究基地）及基地所在单位华东师范大学等单位给予了大力支持，李锋等老师作出了重要贡献。在此表示感谢！

本册教科书出版之前，我们已通过多种渠道与教科书选用作品（包括照片、画作）的作者进行了联系，得到了他们的大力支持。对此，我们衷心地表示感谢！恳请尚未联系到的作者与我们联系，以便出版社及时支付相关稿酬。

我们真诚地希望广大教师、学生及家长在使用本册教科书的过程中提出宝贵意见。我们将集思广益，不断修订，使教科书趋于完善。

编　者