

第2章课后习题参考答案

1. 对于有缺失值的元组，当前有6种处理的方法：

(1) 忽略元组：当缺少类标号时通常这么做(假定挖掘任务涉及分类)。除非元组有多个属性缺失值，否则该方法不是很有效。当每个属性缺失值的百分比变化很大时，它的性能特别差。采用忽略元组，你不能使用该元组的剩余属性值。这些数据可能对手头的任务是有利的。

(2) 人工填写缺失值：一般来说，该方法很费时，并且当数据集很大、缺失值很多时，该方法可能行不通。

(3) 使用一个全局常量填充缺失值：将缺失的属性值用同一个常量(如“unknown”或-)替换。如果缺失值都用“unknown”替换，则挖掘程序可能误以为它们形成了一个有趣的概念，因为它们都具有相同的值——“unknown”。因此，尽管该方法简单，但是并不十分可靠。

(4) 使用属性的中心度量(如均值或中位数)填充缺失值：第2章讨论了中心趋势度量，它们指示数据分布的“中间”值。对于正常的(对称的)数据分布，可以使用均值，而倾斜分布的数据则应使用中位数。

(5) 使用与给定元组属同一类的所有样本的属性均值或中位数。

(6) 使用最可能的值填充缺失值：可以用回归、使用贝叶斯形式化方法的基于推理的工具或决策树归纳确定。

2. 数据采集系统由传感器、模拟信号调理电路、数据采集电路组成，可分为集中式数据采集系统和分布式数据采集系统。

(1) 集中采集式。分时采集型特点：多路信号共同使用一个S/H和A/D电路，简化了电路结构，降低了成本。(但对信号的采集式由模拟多路切换器即多路转换开关分时切换、轮流选通的，因而相邻两路信号在时间上是依次被采集的，不能获得同一时刻的数据，这样就产生了时间偏斜误差。)

同步采集型特点：在多路转换开关之前给每路信号通路各加一个采样保持器，使多路信号的采样在同一时刻进行，即同步采样。(然后由各自的保持器保持着采样信号的幅值，等待多路转换开关分时切换进入公用的A/D电路将保持的采样幅值转换成数据输入主机。这样可以消除分时采集型结构的时间偏斜误差，这种结构既能满足同步采集的要求，又比较简单。不足之处：在被测信号路数较多的情况下，同步采得的信号在保持器中保持的时间会加长，而保持器会有一些泄露，使信号有所衰减。)

(2) 分散采集式：每一路信号一般都有一个S/H和A/D，不再需要模拟多路转换器MNX。(每一个S/H和A/D之对本路信号进行模数转换即数据采集，采集的数据按一定的顺序或随机地输入计算。)

3. 通常有以下4个基本流程：

(1) 发起请求。

通过HTTP库向目标站点发起请求，也就是发送一个Request，请求可以包含额外的

header 等信息，等待服务器响应。

(2) 获取响应内容。

如果服务器能正常响应，会得到一个 Response，Response 的内容便是所要获取的页面内容，类型可能是 HTML、Json 字符串、二进制数据(图片或者视频)等类型。

(3) 解析内容。

得到的内容可能是 HTML，可以用正则表达式，页面解析库进行解析；可能是 Json，可以直接转换为 Json 对象解析；可能是二进制数据，可以做保存或者进一步的处理。

(3) 保存数据。

保存形式多样，可以存为文本，也可以保存到数据库，或者保存特定格式的文件。

4. 经过可视化之后的数据能够加深人对数据的理解和记忆，增加信息的可传播性。数据可视化是关于数据视觉表现形式的科学技术研究，它为大数据分析提供了一种更加直观的挖掘、分析、解释和展示手段，从而让大数据更有意义，更容易解释和贴近大多数人，因此大数据可视化是艺术与技术的结合。数据可视化将各种数据用图形化的方式展示给人们，是人们理解数据、诠释数据的重要手段和途径，因此从本质上讲，数据可视化是帮助用户通过认知数据，进而发现这些数据所反映的实质。

流程：主要包括数据采集、数据处理和变换、可视化映射、用户交互和用户感知。

5. 数据模型是一组数字或符号的组合，它包含数据的定义、类型等，可以进行各类数学操作等。概念模型描述的是事物的语义或状态行为等。现实世界可以用概念模型来描述，而概念模型又可以用数据模型来描述。经过两层抽象，数据便可以描述我们的现实生活中的方方面面。

6. 类别型数据、有序型数据和数值型数据(包括区间型数据和比值型数据)。

(1) 类别型数据：用于区分物体。例如，根据性别可以将人分为男性或者女性；商品可按用途、原材料、生产方法、化学成分、使用状态等进行不同的分类。这些类别可以用来区分一组对象。

(2) 有序型数据：用来表示对象间的顺序关系，如成绩排名、身高排序等。

(3) 区间型数据：用于得到对象间的定量比较。相对于有序型数据，区间型数据提供了详细的定量信息。例如，身高 160 cm 与身高 170 cm 相差 10 cm，而 170 cm 与 180 cm 也相差 10 cm，它们的差值是相等的。由此可见，区间型数据基于任意的起始点，只能衡量对象间的相对差别。

(4) 比值型数据：用于比较数值间的比例关系，可以精确地定义比例。比如，2 班的学生数量是 1 班的 2 倍(2 : 1)。

7. 用于观测、跟踪数据；用于数据分析；辅助理解数据；增加数据吸引力。

8. 10 世纪，数据可视化起源；

14—17 世纪，数据可视化拉开帷幕；

18 世纪，最早的地图和图表产业；

19 世纪，数据绘图广泛应用；

20 世纪，可视化绘图步入低谷期；

21 世纪，大数据可视化日新月异。

目前,大数据可视化已经注定成为可视化历史中的新的里程碑,VR、AR、MR、全息投影等当下最热门的数据可视化技术已应用到教育、游戏、房地产等各行各业。

9. 数据可视化是指将大型数据集中的数据以图形图像形式表示,并利用数据分析和开发工具发现其中未知信息的处理过程,以便借此分析或报告数据的特征或属性之间的关系。可视化的目标是形成可视化信息的人工解释和信息的意境模型。

10. (1) 政府机构。

我国“十三五”规划纲要中明确指出,要实施国家大数据战略,这充分体现了政府对大数据的重视。该战略不仅推动了大数据的发展,而且政府推动数据、数字治理,以提高政府实现科学决策和高效治理。而数据可视化更能直观为政府在短时间内,制定高效、准确的治理手段和管理决策。

(2) 金融业。

在当今互联网金融的激烈竞争下,市场形势瞬息万变,金融行业面临诸多挑战。通过引入数据可视化可以对企业各地的日常业务动态实时掌控,对客户数量和借贷金额等数据进行有效监管,帮助企业实现数据实时监控和管理;通过对核心数据多维度的分析和对比,指导企业科学的运营策略,制定发展方向,不断提高企业风控管理能力和竞争力。

(3) 工业生产。

数据可视化在工业生产中有着重要的应用,例如可视化智能硬件的生产与使用。可视化智能硬件通过软硬件结合的方式使设备具有智能化的功能,并对硬件采集来的数据进行可视化的呈现,因此在智能化后的硬件具备了大数据等附加值。随着相关技术的发展,可视化技术出现在智能电视、智能家居、智能汽车、医疗健康、智能机器人、智能教育等各个应用领域。

(4) 现代农业。

随着计算机技术、物联网技术以及人工智能技术的不断发展,农业生也向智能化方向发展。将智能物联网设备应用到现代农业的整个生产过程,采集现代农业现场的空气温湿度、天气、土壤温湿度、光照数据、二氧化碳以及物流等一系列数据并公开,让消费者用得放心,吃得放心。同时还充分利用电商平台为农户销售农业产品,并大力发展智慧农业数据可视化。

(5) 医学。

数据可视化可以辅助医生对病人诊断,如应用于诊断和外科手术中的精准建模;还可以对重大疫情数据进行可视化,以还原患者的行为轨迹;将可穿戴设备采集人体心率、血压等身体指标的数据可视化。同时,还可以对医院之前分散、凌乱的数据加以整合,构建全新的医疗管理体系模型,以帮助领导快速解决关注的问题。

(6) 教育科研。

在计算机技术快速发展的今天,可视化教学逐渐部署到有人教学模式。可视化教学是在计算机辅助下,将被感知、被认知、被想象、被推理的事物及其发展变化的形式和过程用仿真化、模拟化、形象化及现实化的方式在教学过程中尽量呈现出来,有助于学生较好的获取、存储、重组知识,并能将知识迁移应用,以促进多元思维的养成。

(7)电子商务。

大数据可视化技术在电子商务中有着重要的作用。对于电商企业，对商品开展数字化的运营分析，是商家日常必要的工作，通过可视化技术可以很直观地提供销售和交易的数据，为供商家提供决策。同时，还可以根据用户对商品的关注度或个人喜好，呈现出可视化数据，供商家为其推荐比较感兴趣的商品，以提高销售业绩等。

(8)其他领域。

数据可视化技术还可以在气候、天体、股票交易、汽车运行、物流管理、卫星运行监测、城市交通管理、城市轨道交通监控、现代旅游等多方面加以应用。如气候可视化可以集地理信息、视频监控、警力警情数据于一体，帮助实现城市的社会治安管理等。

11. 数据可视化分为 3 个特征，即功能特征、使用人群特征、应用场景特征。