第五节　数据科学与大数据

数据科学
大数据
数据挖掘

数据科学是一门通过系统性研究获取与数据相关的知识体系的学科。数据科学一方面研究数据本身的特性和变化规律，另一方面通过对数据的研究为自然科学和社会科学提供一种新的方法，从而揭示自然界和人类行为的现象和规律。


数据科学
太极生两仪，两仪生四象，四象生八卦，八卦演万物
道生一、一生二、二生三、三生万物
自然之数，逐渐演化到万物之象的道理

考点 1 数据科学

数据科学研究的是从“数据”整合成“信息”进而组织成“知识”的整个过程，包含对数据进行采集、存储、处理、分析、表现等一系列活动。

| 研究对象 | 数据 |
| --- | --- |
| 研究目标 | 获得洞察力和理解力 |
| 范围 | 统计学、机器学习、计算机科学、可视化、人工智能、领域知识等 |

考点 2 大数据

大数据指无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

大数据具有“4V”特性

| 数据量大 | 大数据的起始计量单位是 PB.EB 或 ZB。未来可能会达到：YB 或 BB。 |
| --- | --- |
| 数据多样性 | 包括网络日志、音频、视频、图片、地理位置等各种结构化、半结构化和非结构化的数据。 |
| 价值密度低 | 大数据价值密度的高低与数据总量的大小成反比。 |
| 数据的产生和处理速度快 | 大数据的处理要符合“一秒定律”。 |

大数据具有“4V”特性，具体包括（　）。
A.数据量大
B.数据多样化
C.价值密度低
D.数据的产生和处理速度快
E.数据服务范围广
【答案】ABCD

考点 3 数据挖掘（教材变化调整）

数据挖掘就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐藏在其中但又有潜在价值的信息和知识的过程。
该定义包含以下几层含义：
1）数据源必须是真实的、大量的、有噪声的；
2）发现的是用户感兴趣的知识；
3）发现的知识是可接受、可理解 、可运用的；
4）并不要求发现放之四海而皆准的知识，仅支持特定的发现问题。

| | |
|---|---|
| 出发点 | 解决实际问题 |
| 核心任务 | 对数据关系和特征进行探索 |
| 分类 | ①监督学习<br>②无监督学习<br>③半监督学习（新增） |

| | |
|---|---|
| 监督学习 | 监督学习的数据集中，每个观测单位既有自变量(特征 xi)，又有因变量（标签 yi）。根据已有的数据集，训练出模型可以根据自变量数据得到因变量预测结果的过程称为监督学习。<br>监督学习中有两大类典型任务：分类和回归。 |

| | |
|---|---|
| 分类 | 是通过特征变量确定观测单位所属类别，因变量为分类变量。<br>例如：用户的满意度，财务信息判断客户是否到期后续约，根据发件人，主题，内容，判断垃圾邮件。<br>常用的分类方法：逻辑斯特回归、决策树、随机森林和支持向量机等。 |
| 回归 | 回归是通过特征变量确定观测单位因变量取值，因变量是定量变量。<br>例如，根据钻石克拉数，颜色，工艺预测钻石价格。根据楼房面积，位置信息判断价格。<br>常用的回归方法：线性回归、非线性回归和分位数回归等。 |

| | |
|---|---|
| 无监督学习 | 无监督学习的数据集中，每个观测单位只有自变量(特征 xi)，没有因变量（标签 yi）。<br>无监督学习的主要任务是探索数据之间的内在联系和结构。<br>无监督学习中有两大类典型任务：聚类和降维。 |

| | |
|---|---|
| 聚类 | 把一组数据按照差异性和相似性分为几个类别使得同类数据相似性尽可能大，不同类数据相似性尽可能小，跨类的数据关联性尽可能低。<br>聚类分析常用于客户细分，文本归类，结构分组，行为跟踪。<br>常用的聚类方法包括：基于划分的方法（例如 k-均值算法）、基于分层的方法、基于密度的方法、 |

| | |
|---|---|
| | 基于网格的方法和基于模型的方法。 |
| 降维 | 在不损失过多信息的前提下将 N 个相关特征降为 K 个不相关特征，使其具有更好的解释性，也称为特征提取。<br>例如：根据客户的能力，品格，担保，资本，环境等特征评价客户的信用等级。<br>常用的降维方法包括：主成分分析法，因子分析法。 |

| | |
|---|---|
| 半监督学习 | 半监督学习是监督学习与无监督学习相结合的一种学习方法。<br>半监督学习的数据集中，一部分观测单位既有自变量(特征 xi)，又有因变量（标签 yi），另一部分观测单位只有自变量(特征 xi)，没有因变量（标签 yi），而且没有标签的观测单位数量远大于有标签的观测单位数量。 |
| 常见的半监督学习有半监督分类、半监督回归、半监督聚类。 | |

【多选题】数据挖掘包含了几种不同的含义，具体有（　）。
A.数据源必须是真实的、大量的、有噪声的
B.挖掘完成后必须能获得反馈
C.发现的是用户感兴趣的知识
D.发现的知识是可接受、可理解、可运用的
E.并不要求发现放之四海而皆准的知识
【答案】ACDE

【多选题】下列属于聚类分析的是（　）。
A.结构分组
B.客户细分
C.行为跟踪
D.主成分析法
E.因子分析法
【答案】ABC

【单选题】下列关于数据挖掘说法正确的是（　）。
A.每个观测单位只有自变量(特征 xi)，没有因变量（标签 yi）属于监督学习
B.每个观测单位既有自变量(特征 xi)，又有因变量（标签 yi）属于无监督学习
C.根据客户的能力，品格，担保，资本，环境等特征评价客户的信用等级属于聚类技术
D.把一组数据按照差异性和相似性分为几个类别使得同类数据相似性尽可能大，不同类数据相似性尽可能小，跨类的数据关联性尽可能低
【答案】D

（2022）下列属于无监督学习的是（　）。
A.决策树
B.线性回归
C.因子分析
D.随机森林

【答案】 C
【解析】C.因子分析是无监督学习的降维方法； B.线性回归是监督学习的回归方法； A 决策树和 D.随机森林是监督学习的分类方法。

（2023 补）下列数据挖掘方法中，属于监督学习的有（　）。

A.聚类分析

B.因子分析

C.支持向量机

D.逻辑斯特回归

E.主成分分析

【答案】CD

【解析】监督学习中有两大类典型任务：分类和回归。常用的分类方法有逻辑斯特回归、决策树、随机森林和支持向量机等。常用的回归方法有线性回归、非线性回归和分位数回归等。