

第五节

数据科学与大数据

(一) 数据科学★

(二) 大数据★★

(三) 数据挖掘★★

(一) 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。

数据科学涉及的范围非常广泛，例如统计学、计算机科学、可视化、人工智能领域知识等。

(二) 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。

大数据的特征包括（5V特征）【新增】

1. 数据量大（Volume）；

2. 数据多样化（Variety）：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。

3. 价值密度低（Value）：大数据价值密度的高低与数据总量的大小成反比。

4. 数据的产生和处理速度快（Velocity）。

5. 真实性（Veracity）：需通过数据清洗、数据治理、质量控制等一系列规范手段确保数据可靠

【练习·多选】大数据的特性包括（ ）。

A. 数据的产生速度快

B. 数据的价值密度低

C. 数据非结构化

D. 数据量大

E. 真实性

答案：ABDE

解析：大数据的特征包括（5V特征）【新增】

1. 数据量大（Volume）；

2. 数据多样化（Variety）：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。

3. 价值密度低（Value）：大数据价值密度的高低与数据总量的大小成反比。

4. 数据的产生和处理速度快（Velocity）。

5. 真实性 (Veracity)：需通过数据清洗、数据治理、质量控制等一系列规范手段确保数据可靠

(三) 数据挖掘★★

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。

方法

1. 监督学习 (xy)
2. 无监督学习 (x)
3. 半监督学习

1. 监督学习

每个观测单位既有自变量 (特征x)，又有因变量 (标签y)。

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。

(1) 分类：通过特征变量确定观测单位所属的类别，因变量是分类变量。

例如，根据发件人、主题、内容等信息判断邮件是否是垃圾邮件。

方法：逻辑斯特回归、决策树、随机森林和支持向量机等。【口诀】逻辑支持随机决策

(2) 回归：通过特征变量确定观测单位因变量的取值，因变量是定量变量。

例如，根据房屋面积、位置、楼层等信息预测房价。

方法：线性回归、非线性回归和分位数回归等。

【新增】集成学习、神经网络等方法在复杂高维数据的分类和回归中也被广泛应用

2. 无监督学习

每个观测单位只有自变量 (特征x)，没有因变量 (标签y)。

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。

(1) 聚类：把一组数据按照差异性和相似性分为几个类别。

聚类分析常用于客户细分、文本归类、结构分组、行为跟踪等。

与分类不同，聚类要划分的类是未知的，聚类是根据观察学习来确定数据之间的关系。

方法：基于划分的方法 (例如k均值聚类算法)、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。

(2) 降维：在不损失过多信息的前提下将N个相关的特征降为k个不相关的特征 (其中 $k < N$)，也称为特征提取。

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。

方法：主成分分析法、因子分析法等。

3. 半监督学习

监督学习与无监督学习相结合的一种学习方法。

（1）方式：半监督学习的数据集中，一部分观测单位既有自变量（特征x）又有因变量（标签y：），另一部分观测单位只有自变量（特征x）没有因变量（标签y：），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。

（2）方法：半监督分类、半监督回归、半监督聚类。

监督学习	分类	逻辑斯特回归、决策树、随机森林和支持向量机 【口诀】逻辑支持随机决策
	回归	线性回归、非线性回归和分位数回归等。
无监督学习	聚类	基于划分的方法（例如 k 均值聚类算法）、 基于分层的方法、基于密度的方法、 基于网格的方法、基于模型的方法
	降维	主成分分析法、因子分析法
半监督学习	半监督分类、半监督回归、半监督聚类	

【2023年·多选】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。回归是通过特征变量确定观测单位因变量的取值，因变量是定量变量。例如，根据钻石的克拉数、颜色、切割工艺等信息预测钻石的价格，根据房屋面积，位置，楼层等信息预测房价。常用的回归方法有线性回归、非线性回归和分位数回归等。

【2025年·多选】属于学习方式中的监督学习的有（ ）。

- A. 逻辑回归
- B. 线性回归
- C. 聚类
- D. 随机森林
- E. 决策树

答案：ABDE

解析：监督学习包括逻辑回归、线性回归、随机森林、决策树，故ABDE正确。C聚类属于无监督学习。

框架导航

