有关Github仓库，欢迎来Star

1. 导论: 大数据&数据挖掘

1 大数据

1. 含义：数据量巨大的数据，以至于合理时间内人类无法整理出可用信息
2. 特性：Volume(规模大)+Variety(多样)+Velocity(数据产生/处理极快)+Veracity (真实但低质)

2 数据挖掘

1. 含义：从大数据中挖掘有价值的知识/规律
2. 任务：分析(关联性/聚类)+预测(分类/回归)+关联规则等

3 其它

1. 大数据的应用：进人工智能(算力驱动/神经符号协同/记忆启发)+促进教育
2. 面临的挑战：相关性≠因果，可解释性，群智涌现(群体智力远超个体)，隐私，可视化

2. 数据预处理

2.1. 数据及其描述

1 数据对象及其属性

1. 对象：数据集的组成单元，代表一个实体
2. 属性：对实体(对象)的描述

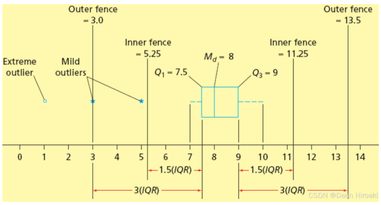
属性类型	含义	举例	描述
二元	属性值域只有$\mathrm{True}/\mathrm{False}$	诊断结果	$\mathrm{N/A}$
枚举	属性值域由无序/不定量符号组成	职业类型	众数
序数	属性值间的序有意义，但前后序是定性的	军衔级别	众数/中位数
数值	可用整数或实数度量	好多	众数/中位数/平均数

2 数据基本统计描述

1. 传统的：算术/加权平均，中位数，众数(模)，极差，标准差/方差
2. 百分位：第k个百分位数x_k表示$k\%$的数据低于x_k，如Q_1/中位数/Q_3(即$25/50/75$百分位数)

3 数据基本图形描述

1. 传统的：直方图，分位数图，散点图
2. $\mathrm{Box\ Plot}$:



   - 四分位极差：$\mathrm{IQR}=Q_3-Q_1$
   - 孤立点(Outlier)：在$Q_1-1.5\mathrm{IQR}$之下或者$Q_1+1.5\mathrm{IQR}$之上
   - 盒图要素：上下端在Q_1/Q_3上，中位数处划线，胡须延伸到最大最小观测值

4 数据相关性描述：$\mathrm{Pearson}$相关系数$r=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

2.2. 数据预处理

1 概述

1. 为何要预处理：数据不完整/有噪声/不一致(比如年龄可用汉字或数字表示)
2. 数据预处理任务：数据清理/集成/变换/归约(压缩)/离散化.....

2 数据清洗

1. 填补空缺值：人工补全，全局(千篇一律)补全，平均值补全，基于概率(如$\mathrm{Bayesian}$)补全
2. 噪声处理：用自适应回归来平滑，通过聚类检测并去除孤立点，排序后分箱

3 数据集成和变换

1. 数据/模式集成：
   - 含义：将多个数据源中的数据/元数据合并到一个一致的存储
   - 难题：解决数值/属性的冲突(如去掉强相关属性中的一个)，实体识别，检测并去除冗余数据
2. 数据变换：将数据统一成适合挖掘的形式
   - 归一化：将数据缩放到特定区间，如最值归一$v'=\dfrac{v-\min}{\max-\min}$/Z-Score归一 $v'=\dfrac{v-\mu}{\sqrt{\sigma}}$
   - 属性构造：通过现有属性构造新的属性
   - 数据泛化：沿概念分层向上汇总

4 数据规约

1. 含义：大大压缩数据的存储空间，但是保证数据分析的质量
2. 策略：堆规约(移除不重要元素/属性)，数据压缩(有损/无损)，数值规约(用较小的数据表示替代)