第四章 数据预处理（占整个过程的60%的工作量）

2023年7月30日 18:44

- 数据清洗、数据集成、转换、归约

1. 数据清洗
   a. 缺失值：删除记录、数据插补（平均数、中位数、众数、固定的常量、最邻近的值、回归拟合预测的值、已有函数算出的y值、拉格朗日插值法、牛顿插值法）、不处理
   b. 异常值：删除、视为缺失值处理、平均值修正、不处理
2. 数据集成：将多个数据源合并在一个一致的数据存储位置（比如数据仓库）
   a. 实体识别问题
      i. 从不同数据源识别出现实世界的实体，他的任务是统一不用源数据的矛盾之处
         1）同名异义：数据源A、B的同名属性ID描述的不同的实体
         2）异名同义：属性名称不一样，但是描述的是同样的东西
         3）单位不统一：秒速一个实体时，用的分别是国际单位和中国的传统计量单位
   b. 冗余属性识别
      i. 同一属性多次出现
      ii. 同一属性命名不一致导致重复
         1）先分析冗余属性，检测后再将其删除。
         2）用相关系数度量一个属性在多大程度上蕴含另一个属性（数值型数据）
   c. 数据变换（规范化处理，转成“适当的”形式）
   d. 简单函数变换（不具有正态分布—>正态分布的数据）
      i. 平方、开放、取对数、差分运算（非平稳序列-->平稳序列）
   e. 规范化
      i. 目的：消除指标之间的量纲和取值范围差异的影响--将数据按照比列进行缩放

ii. 对于基于距离的挖掘算法很重要

1）最小-最大规范化（离差标准化）

2）零-均值规范化（标准差规范化）--均值为0，标准差=1

3）小数定标规范化

f. 连续属性离散化（连续属性—>分类属性）

i. 应用的算法：分类算法（ID3、apriori）

ii. 离散化的过程：确定分类数，如何将连续属性值映射到这些分类值

iii. 常用的离散化方法:等宽法、等频法（人工）、（一维）聚类

g. 属性构造（构造新的指标）

h. 小波变换：新型的信号分析手段

3. 数据规约

a. 属性归约

i. 合并属性

ii. 逐步向前选择

iii. 逐步向后删除

} 直接删除不相关属性

iv. 决策树归纳

v. 主成分分析（PCA）（连续属性的数据降维）

b. 数值归约：选择替代的、较小的数据来减少数据量

i. 直方图（归到相应的值域）

ii. 聚类（将对象划分为簇）

iii. 抽样（用随机样本代表数据集-无放回/有放回简单随机抽样、聚类抽样、分层抽样），可以用中心极限定理来计算样本的大小

c. 参数回归（简单线性模型/对数线性模型）

4. python主要数据预处理函数