Data Preprocessing

- 关于数据
- 为什么要预处理数据？
- 描述性数据汇总
- 数据清理
- 数据集成和变换
- 数据规约
- 小结

数据集成

- 数据集成:
  - ✓ 把多个数据源的数据合并到一个一致的数据存储(如数据仓库)中
- 模式集成
  - ✓ 整合来自不同来源的元数据
  - ✓ 实体识别问题:识别来自多个数据来源的真实世界实体, 例如, **A.cust-id ≡ B.cust-#**
- 检测和解决数值冲突
  - ✓ 对于同一个真实世界实体，来自不同来源的属性值不同
  - ✓ 可能的原因: 不同的表示, 不同的范围, **e.g.,** 十进制 **vs.** 英式单位

处理数据整合中的冗余问题

- 当整合多个数据库的数据时，多余的数据经常出现
  - ✓ 对象识别: 同一个属性或者对象在不同的数据库或许有不同的名字
  - ✓ 导出性数据: 一个属性可能是另一个表中的派生属性, **e.g., annual revenue**（年度税收）
- 多余的属性可以通过相关性分析检测
- 仔细的对多个来源的数据进行整合，可以帮助减少冗余和不一致，提高挖掘速度和质量

相关性分析(数字数据)

- 相关系数（也叫皮尔逊积矩系数）

$$r_{A,B} = \frac{\sum(A-\overline{A})(B-\overline{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum(AB)-n\overline{A}\overline{B}}{(n-1)\sigma_A\sigma_B}$$

其中， **n** 是元组的个数, $\overline{A}$ $\overline{B}$ 分别是**A**和**B**的均值, σ_A 和 σ_B e分别是**A**和**B**的标准差，
Σ(AB) 是 **AB** 叉积的和.

- $r_{A,B} > 0$, **A** 和 **B**正相关 (**A**的值随**B**的值增加而增加 。值越大， 相关性越强
- $r_{A,B} = 0$: 独立;
- $r_{A,B} < 0$: 负相关

数据转换

- **Smoothing**（平滑）：除去数据的噪音
- **Aggregation**（聚合）:概括, 构造数据立方体
- **Generalization**（泛化）：概念层次攀升
- **Normalization**（规范化）：设定在一个特定的小的范围内
  - ✓ 最小值最大值规范化
  - ✓ **z-score**规范化
  - ✓ 小数定标规范化
- 属性构造
  - ✓ 对指定的对象构造属性

数据转换: 规范化

- 最小值最大值规范化

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- **z-score规范化 (μ: 均值, σ: 标准差)**

$$v' = \frac{v - \mu_A}{\sigma_A}$$

- 小数定标规范化

$$v' = \frac{v}{10^j}$$ 其中，j 是最小的整数，因此Max(|v'|) < 1