数据预处理

为什么要进行预处理

现实世界的数据很“脏”

- 不完整的: 缺少属性值, 感兴趣的属性缺少属性值, 或仅包含聚集数据
  如, occupation=“ ”
  来源于：收集数据时，在不同的阶段具有不同的考虑；人/硬件/软件的问题等。
- 含噪声的: 包含错误或存在孤立点
  如, Salary=“-10”
  来源于：收集阶段；数据传输阶段等。
- 不一致的: 在名称或代码之间存在着差异
  如, Age=“42” Birthday=“03/07/1997”
  如, 原来排序 “1,2,3”, 现在排序 “A, B, C”
  来源于：不同的数据源；功能依赖冲突。

数据预处理的主要任务

- 数据清洗（Data cleaning）
  填充遗失的数据, 平滑噪声数据, 辨识或删除孤立点, 解决不一致性问题
- 数据集成（Data integration）
  对多个数据库，数据立方或文件进行集成
- 数据变换（Data transformation）
  规范化与聚集（Normalization and aggregation）
- 数据约简（Data reduction）
  得到数据集的压缩表示，它小的多，但能够产生同样的（或几乎同样的）分析结果
- 数据离散化（Data discretization）
  特别对数字值而言非常重要

数据清洗

清洗的主要任务：

- 填充遗失数据
- 辨识孤立点、平滑噪声数据
- 修正不一致性数据
- 解决数据集成时带来的数据冗余问题

怎样处理遗失的数据：

- 忽略元组: 除非元组有多个属性缺少值，否则该方法不是很有效

generated by haroopad

- 人工填充: 费时费力
- 自动填充:
  - 使用一个全局常量填充: 如, “unknown”, 会误认为是一个新的、有意义的类?!
  - 该属性的平均值
  - 使用最可能的值: 使用基于推导的方法，如Bayesian公式或决策树

怎样处理噪声数据:

- 分箱方法:
  先对数据进行排序，然后把它们划分到箱
  然后通过箱平均值，箱中值等进行平滑.

  - 等宽 (距离)划分:
    根据属性值的范围划分成N等宽的区间
    如果A和B 属性值的最大与最小值，则区间宽度为: W = (B –A)/N.
    很直接，但孤立点将会对此方法有很大的影响

  - 等深 (频率) 划分:
    划分成N个区间，每个区间含有大约相等地样本数。具有较好的数据扩展性。

实例:

```
*  价格排序: 4, 8, 9, 11, 15, 21, 21, 22, 24, 25, 26, 28, 29, 30, 40
*  划分成箱 (等深) :
      - Bin 1: 4, 8, 9, 11, 15
      - Bin 2: 21, 21, 22, 24, 25
      - Bin 3: 26, 28, 29, 30, 40
*  用箱平均值平滑数据:
      - Bin 1: 9.4, 9.4, 9.4, 9.4, 9.4
      - Bin 2: 22.6, 22.6, 22.6, 22.6, 22.6
      - Bin 3: 30.6, 30.6, 30.6, 30.6, 30.6
*  用箱中值平滑数据:
      - Bin 1: 9, 9, 9, 9, 9

      - Bin 2: 22, 22, 22, 22, 22
      - Bin 3: 29, 29, 29, 29, 29
```

- 聚类
  探测并去除孤立点

- 回归分析（Regression）
  让数据适合一个函数（如回归函数）来平滑数据

数据集成

数据集成: 将多个数据源中的数据结合起来存放在一个一致的数据存储中(如数据仓库)。

存在以下问题:

1. 实体识别问题 (EI): 从不同的数据源辨识实体，如, A.cust-id □ B.cust-#。
2. 检测与解决值冲突问题
   对客观世界的同一实体, 不同数据源可能具有不同的值
   可能原因: 不同的表示方式, 不同的刻度，如公制与英制（metric vs. British units）等

3. 数据冗余问题
   冗余属性可以通过相关分析检测出来

数据变换

规范化（Normalization）: 刻度变换

- 最小－最大规范化

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

- z-score规范化

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- 小数定标规范化

$$v' = \frac{v}{10^j} \quad \text{其中}j\text{是使得}\mathrm{Max}(|v'|) < 1\text{的最小整数}$$

数据约简

数据约简: 获得原有数据的约简表示，它的规模小的多，但具有与原有规模的数据相同或近似的分析结果

策略:

- 降维（Dimension Reduction）： 主成分分析PCA、线性判别分析LDA、决策树归纳等
- 参数化或非参数化方法

1. 参数化方法:假设数据适合某个模型，然后估计模型参数，仅仅存储这些模型参数，而不再存储原有数据(除了可能的孤立点).如线性回归、多元线性回归、对数线性模型等
2. 非参数化方法
   不假设模型,有: 直方图(histograms)、聚类, 采样(sampling)等.

- 采样
  选择原有数据集的具有代表性的一个子集，自适应采样方法，分层采样。

关于主成分分析PCA:

是一种统计方法。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的新变量（转换后的这组变量叫主成分），同时根据实际需要选取几个较少的新变量以尽可能多地反映原来变量的信息。

generated by haroopad

- 方差（Variance）与协方差（Covariance）

$$\mathrm{var}(X)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(X_i-\overline{X})}{(n-1)}$$

$$\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{(n-1)}$$

- $\mathrm{Cov}(X,Y)>0$，正相关(同升同降)；$\mathrm{Cov}(X,Y)=0$，二者独立；$\mathrm{Cov}(X,Y)<0$，负相关(一升一降)；

协方差矩阵（3个属性）

$$C=\begin{bmatrix}\mathrm{cov}(X,X) & \mathrm{cov}(X,Y) & \mathrm{cov}(X,Z)\\ \mathrm{cov}(Y,X) & \mathrm{cov}(Y,Y) & \mathrm{cov}(Y,Z)\\ \mathrm{cov}(Z,X) & \mathrm{cov}(Z,Y) & \mathrm{cov}(Z,Z)\end{bmatrix}$$

- X：$m\times n$矩阵，其中有m个特征，n条数据
- 先对X中心化：即把X_i的每个元素减去特特征i的平均值
- 令

$$S_X=\frac{1}{n-1}XX^T$$

- 则S_X的第i个对角元素表示特征i的方差，非对角元素表示相应两个特征的协方差。

generated by haroopad

- 思想：找到一个变换P，使得$Y=PX$的协方差矩阵C满足：
  - C的对角元素越大越好（即一个属性的方差越大越好）
  - 属性之间相关性越低越好：C的非对角元素越小越好（最好是一个对角阵）
- 具体：找到一个标准正交矩阵P，使得S_Y对角化，其中，

$$S_Y = \frac{1}{n-1} YY^T$$

  - P的行就是X的主成分

离散化

属性的三种类型

1. 标称性的 — 取自于无序集合(unordered set)的值
2. 有序的(Ordinal) —取自于有序集合(ordered set)的值
3. 连续的 — 实数

把连续型属性的取值范围划分成区间
通过离散化减少数据集大小
为进一步分析做好准备

离散化方法

- 分箱
- 直方图分析
- 聚类分析
- 基于熵的离散化

generated by haroopad

- 给定一个样本集合 S, 如果用边界值T把 S划分成2个区间 S_1 与S_2，则划分后的熵为：

$$I(S,T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

- 选择某一边界T的准则是：它使其后划分得到的信息增益(Information Gain，见上式）最大.
- 上述过程递归地用于所得到的划分，直到满足某个终止条件。
- 实验表明这种划分方法能够约简数据集并提高分类精度。

generated by haroopad