数据预处理

- 数据集是由数据对象组成的

脏数据问题

- 有噪声：包含错误或者孤立点
    - 例如：Salary = -10；
- 不一致：
    - 例如：等级：1、2、3； A、B、C；甲、乙、丙

数据质量

- 数据完整性、一致性、相关性、时效性、可信性、可解释性

主要步骤

- 描述性数据归总
- 数据清洗
    - 填写空缺的值
    - 平滑噪声数据
    - 识别、删除孤立点
    - 解决不一致性
- 数据集成
    - 数据库
    - 数据立方体
    - 文件
- 数据规约
    - 将数据压缩，但可以得到相同或相近的结果
- 数据变换

- 规范化和聚集，提高涉及距离度量的挖掘算法的准确性和有效性

属性

分类	示意
标称属性	标称属性的值是一些符号或事物的名称
二元属性	是一种标称属性，只有两个状态：0 或 1
序数属性	值之间具有有意义的序或秩评定（ranking），但是相继值之间的差是未知的
数据属性	是定量的可度量的量，用整数或实数表示，可以是区间标度的或比率标度的
离散属性	具有有限个或无限个可数个数，可以用或不用整数表示
连续属性	如果属性不是离散的，则它是连续的

示例

学生ID	姓名	性别	选修科目	成绩	评定
20201203	Tom	男	语文	95.0	优
20201204	Jerry	女	英语	80.5	良
20201205	Kevin	男	语文	77.2	中
20201206	Mary	女	数学	85.3	良
20201207	Jim	男	数学	61.2	差

（离散） （标称） (二元) (标称) (连续) (序数)

统计学规律

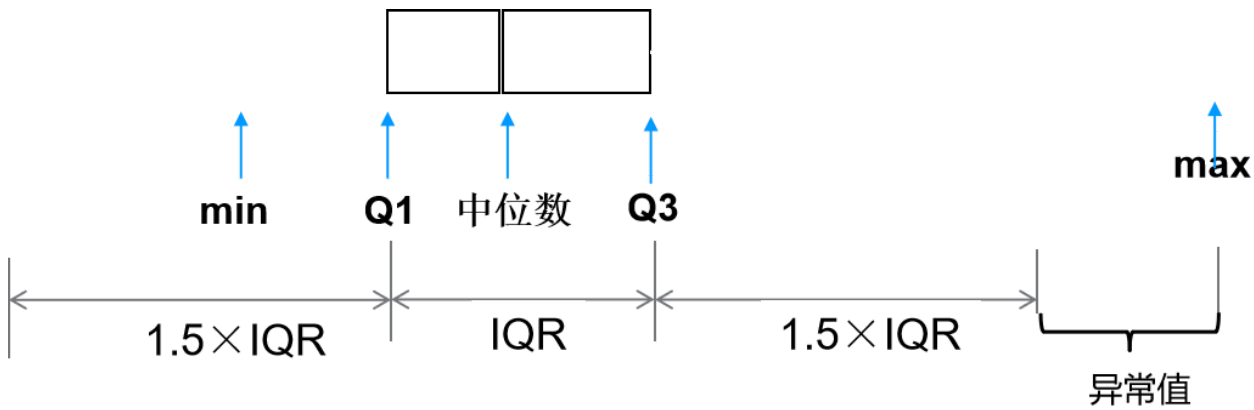
- 本福特定律：描述的是自然随机变量首位数字“1-9”的使用频率相对稳定
    - 概率公式：$P(n)=\log_{10}(1+\frac{1}{n})$
    - 数据必须跨度足够大，样本数量足够多
    - 数据不能认为规则和修饰，比如电话号码、发票编号，身份证号码
- 小概率原理：一个事件发生的概率很小的
    - 小概率：$P\leq 0.05$或者$P\leq 0.01$

数据的分布特征

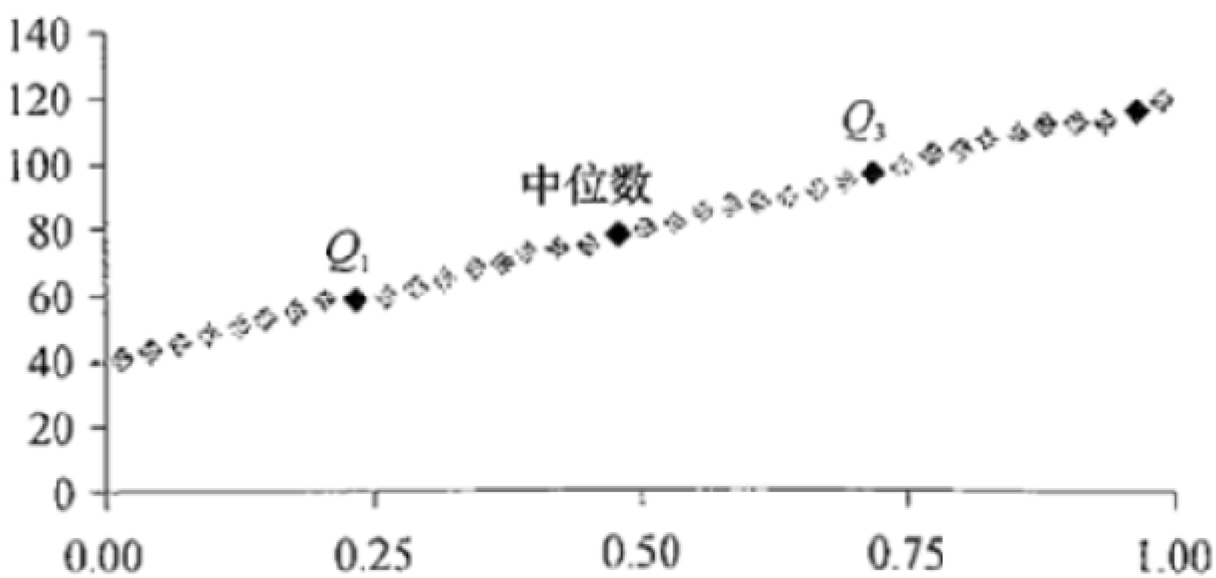
- 度量数据的中心趋势： 均值、中位数、众数
- 度量数据的离散程度：方差，四分位数、四分位数极差
- 基本公式：
    - 算数平均值：$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$
    - 加权平均值：$\overline{x}=\frac{\sum_{i=1}^{n}w_ix_i}{\sum_{i=1}^{n}w_i}$
    - 截断均值：去掉高、低极端值得到的均值
    - 中位数：有序集的中间值或者中间两个值平均
    - 众数：集合中出现频率最高的数
        - 适度倾斜（非对称的）：$mean-mode=3\times(mean-median)$
    - 方差S^2:$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2=\frac{1}{n}[\sum_{i=1}^{n}x_i^2-\frac{1}{n}(\overline{x}^2)]$
    - 标准差S：关于平均值的离散的度量，仅当选平均值做中心度量时使用有观测值相同则 s = 0，否则 s>0
    - 正态分布函数曲线：(μ,σ)
        - 3σ 原则:如果数据服从正态分布，异常值为测定值中超过3倍标准差的值。
    - 极差：数据集的最大值和最小值之差
    - 百分位数：k%的数据项位于或低于x
        - 四分位数：Q_1 (25^{th} percentile),Q_3 (75^{th} percentile)

- 中间四分位数极差(IQR)： IQR =$Q_3 - Q_1$

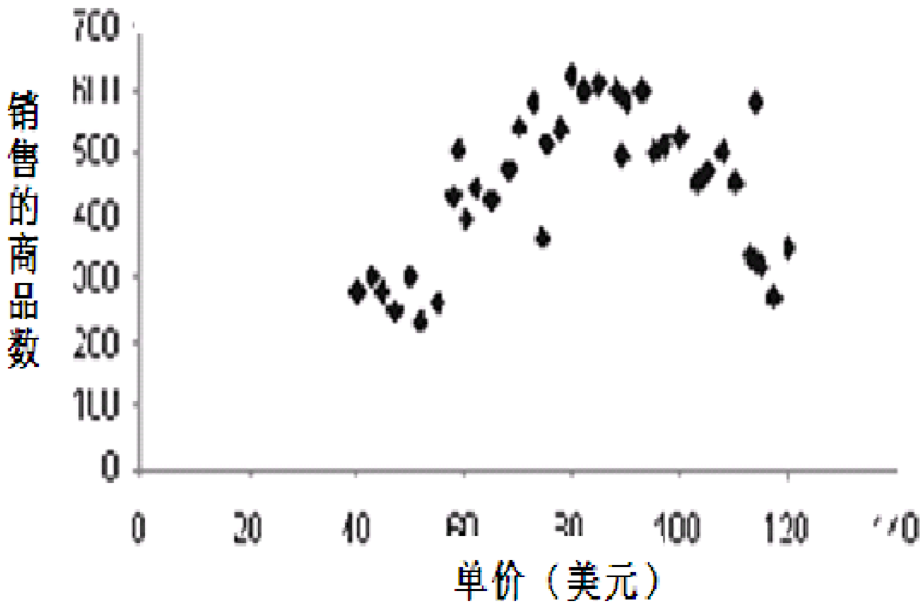
- 盒图：



- 直方图：略
- 分位数图：观察单变量数据分布的简单有效方法



- 散布图：确定两个量化的变量之间看上去是否有联系、模式或者趋势的最有效的图形方法之一



- 局部回归曲线:Loess（Local regression）曲线为散布图添加一条平滑的曲线
    - 平滑参数α
    - 被回归拟合的多项式的阶 λ

数据清洗

缺失值的处理

#Imputer类

- 删除带有缺失值的样本或特征
    - 删除样本：适合某些样本有多个特征存在缺失值，当存在缺失样本数量过大时，不能使用
    - 删除特征：当某个特征缺失值较多，且该特征对数据分析的目标影响不大时，可以将该特征删除.
- 采用某种方法对缺失值进行填补
    - 均值填补：计算非缺失值的平均值或者众数
        - 对于连续型特征，通常使用平均值进行填补；
        - 对于离散型特征，则使用众数进行填补.
        - 缺陷：

      - 均值填补法会使得数据过分集中在平均值或众数上，导致特征的方差被低估;
        - 由于完全忽略特征之间的相关性，均值填补法会大大弱化特征之间的相关性;
    - 解决方案：根据一定的辅助特征，将数据集分成多组，然后在每一组数据上分别使用均值插补.
  - 随机填补
    - [贝叶斯Bootstrap方法]
    - [近似贝叶斯Bootstrap方法]
  - 基于模型的填补
  - 哑变量方法
    - 对于离散型特征，如果存在缺失值，可以将缺失值用同一个常量进行处理，这种方法称为哑变量方法

贝叶斯Bootstrap方法

- 假设数据集有 n 个样本，某特征 f 存在 k 个非缺失值和 (n−k) 个缺失值，贝叶斯Bootstrap方法填补共有两步：
  - 第一步：从均匀分布 U(0, 1)中随机抽取 k−1个随机数，并进行升序排序记为{0, a_{1},a_{2}, ⋯,a_{k−1}, 1}；
  - 第二步：对 (n−k) 个缺失值，分别从非缺失值$f_1, f_2, ..., f_k$中以概率 a_{1},a_{2}−a_{1}, ⋯,1−a_{k−1}采样一个值进行填补.
- 示例：
- 假设我们有一个包含 (n) 个观测值的数据集，其中有 (k) 个非缺失值，(n - k) 个缺失值。我们将使用均匀分布 (U(0, 1)) 来生成随机数，并根据这些随机数填补缺失值。
- 步骤 1：从均匀分布中抽取随机数
  - 随机抽取：
    - 从均匀分布 (U(0, 1)) 中随机抽取 (k - 1) 个随机数。假设我们抽取到的随机数为： 0.2, 0.5, 0.8
    - 这里 (k = 4)，所以我们抽取了 (4 - 1 = 3) 个随机数。

- 升序排序：
  - 将这些随机数进行升序排序，并加上边界值 0 和 1。结果为：
    0, 0.2, 0.5, 0.8, 1
  - 记这些值为
    $(0, a_1, a_2, a_3, 1)$，其中$(a_1 = 0.2)$，$(a_2 = 0.5)$，$(a_3 = 0.8)$。
- 步骤 2：填补缺失值
  - 确定非缺失值：
  - 假设我们有 (k = 4) 个非缺失值，记为
    (f_1, f_2, f_3, f_4)。例如：$f_1 = 10, f_2 = 20, f_3 = 30, f_4 = 40$
  - 计算概率：
    - 根据升序排序的随机数，计算每个区间的概率：
      - $(a_1 = 0.2)$代表从 0 到 0.2 的区间；
      - $(a_2 - a_1 = 0.5 - 0.2 = 0.3)$ 代表从 0.2 到 0.5 的区间；
      - $(a_3 - a_2 = 0.8 - 0.5 = 0.3)$ 代表从 0.5 到 0.8 的区间；
      - $(1 - a_3 = 1 - 0.8 = 0.2)$ 代表从 0.8 到 1 的区间。
  - 概率分配为：
    {0.2, 0.3, 0.3, 0.2}
- 填补缺失值：
  - 对于 (n - k) 个缺失值（假设有 2 个缺失值），我们可以根据上述概率从非缺失值中进行填补。例如，假设我们需要填补两个缺失值：
  - 第一个缺失值：
    以概率(0.2)选择(f_1)，以概率(0.3)选择(f_2)，以概率(0.3)选择(f_3)，以概
    。
  - 第二个缺失值：同样的选择方式。
  - 假设我们分别选择了 (f_2) 和 (f_3) 来填补缺失值，最终的数据集将变为：
    {10, 20, 30, 40, 20, 30}

近似贝叶斯Bootstrap方法

- 首先从 k 个非缺失值$f_1, f_2, ..., f_k$中有放回地抽取k个值建立一个新的大小为k的集合F. 然后对于 (n− k)个缺失值，分别从 F 中随机抽取一个值进行填补.

噪声平滑

- 噪声: 在可测度变量中的随机错误或偏差。
- 分箱法:
  - 分箱方法通过考察“邻居”（即周围的值）来平滑存储数据的值，首先将数据排序并将其分割到一些相等深度的“桶”（bucket or bin）中，然后可根据桶均值，桶中间值，桶边界值等进行平滑。

价格按升序排序后的数据：4,8,15,21,21,24,25,28,34

划分为等频的箱：
箱1：4，8，15
箱2：21，21，24
箱3：25，28，34

用箱均值光滑：
箱1：9，9，9
箱2：22，22，22
箱3：29，29，29

用箱边界值光滑：
箱1：4，4，15
箱2：21，21，24
箱3：25，25，34

- 回归法:
  - 通过让数据适配一个函数（如线性回归函数）来平滑数据

异常值检测与处理

- 检测方法:
  - 简单统计
  - 3σ原则
  - 基于模型检测
  - 基于距离
    - 通常可以在对象之间定义邻近性度量，异常对象是那些远离其他对象的对象
  - 基于密度
    - 当一个点的局部密度显著低于它的大部分近邻时才将其分类为离群点。适合非均匀分布的数据
  - 离群点分析
    - 一个对象是基于聚类的离群点，如果该对象不强属于任何簇

数据集成

实体识别问题

- 指从不同数据源识别出现实世界的实体，它的任务是统一不同源数据的矛盾之处
- 常见形式：同名异义，异名同义，单位不统一

检测和解决数据值冲突

- 对于现实世界的同一实体，来自不同数据源的属性值可能不同，这可能是因为表示、比例、或编码、数据类型、单位不统一、字段长度不同。

冗余数据与相关性分析

- 同一属性多次出现，同一属性命名不一致等也可能导致结果数据集中的冗余。
- 处理方式：
  - 对于标称数据，我们使用卡方检验。
    - 检测两个属性A 和B 之间的相关联系：$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij}-e_{ij})^2}{e_{ij}}$
    - $e_{ij} = \frac{count(A=a_i) \times count(B=b_j)}{n}$
    - O_{ij}是联合事件（A_i, B_j）的观测频度，即实际计数，e_{ij}是（A_i, B_j）的期望频度
  - 对于数值属性，我们使用相关系数(correlation coefficient)和协方差(covariance)，都评估一个属性的值如何随另一个变化
    - 计算属性A和B相关系数（又称Pearson积矩系数）估计这两个属性的相关度 $r_{A,B}$
      - $r_{A,B} = \frac{\sum_{i=1}^{n}(a_i-\overline{A})(b_i-\overline{B})}{n\sigma_A\sigma_B}$
      - 相关性并不蕴含因果关系，也就是说，如果A和B是相关的，并不意味着A 导致B 或者B 导致A .
    - 数值数据的协方差:
      - $Cov(A,B) = E((a_i - E(A))(b_i - E(B))) = \frac{\sum_i^n (a_i-E(A))(b_i-E(B))}{n}$
      - 评估两个属性如何一起变化
    - $r_{A,B} = \frac{Cov(A,B)}{\sigma_A\sigma_B}$

卡方检验

- 性别与投票意向的关系
- 我们想要研究性别（男性和女性）与投票意向（支持或反对某个候选人）之间的关系。假设我们进行了一个调查

	支持	反对	总计
男性	30	10	40

	支持	**反对**	**总计**
女性	20	40	60
总计	50	50	100

- 构建假设
  - **零假设 ((H_0))**：性别与投票意向无关。
  - **备择假设 ((H_1))**：性别与投票意向有关。
- 计算期望频数
  - 根据零假设，我们可以计算每个单元格的期望频数。期望频数的计算公式为：
  - E = $\frac{\text{行总计} \times \text{列总计}}{\text{总计}}$
- 计算结果如下：

	支持	**反对**	**总计**
男性	$(\frac{40 \times 50}{100} = 20)$	$(\frac{40 \times 50}{100} = 20)$	40
女性	$(\frac{60 \times 50}{100} = 30)$	$(\frac{60 \times 50}{100} = 30)$	60
总计	50	50	100

- 计算卡方统计量
  - 卡方统计量的计算公式为：$\chi^2 = \sum \frac{(O-E)^2}{E}$
  - 其中 (O) 是观察频数，(E) 是期望频数。
- 计算每个单元格的贡献：
  - 男性支持：$\frac{(30-20)^2}{20} = \frac{100}{20} = 5$
  - 男性反对：$\frac{(10-20)^2}{20} = \frac{100}{20} = 5$
  - 女性支持：$\frac{(20-30)^2}{30} = \frac{100}{30} \approx 3.33$
  - 女性反对：$\frac{(40-30)^2}{30} = \frac{100}{30} \approx 3.33$
  - 将所有贡献相加：$\chi^2 = 5 + 5 + 3.33 + 3.33 \approx 16.66$
  - 确定自由度和临界值

- 自由度 ((df)) 的计算公式为：
  $df = (行数-1) \times (列数-1) = (2-1) \times (2-1) = 1$
- 使用卡方分布表查找自由度为 1 的临界值（例如，显著性水平 ($\alpha = 0.05$) 时，临界值约为 3.841）。
- 做出结论:比较计算得到的卡方统计量和临界值：16.66 > 3.841

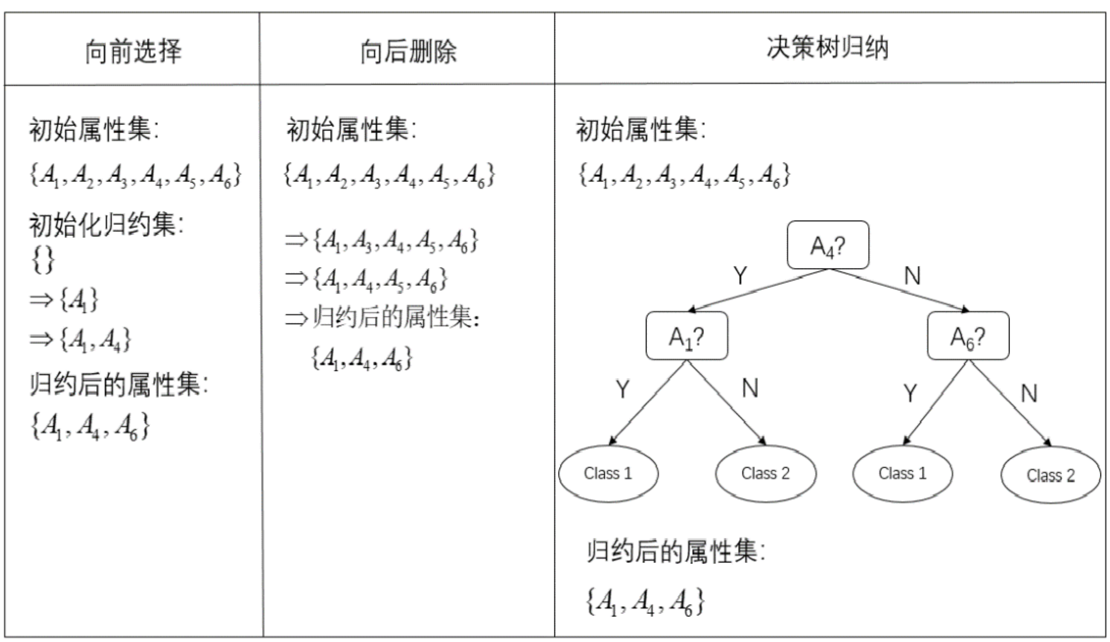
元组重复

- 除了检测属性间的冗余外，还应当在元组级检测重复（对于给定的唯一数据实体，存在两个或多个相同的元组）

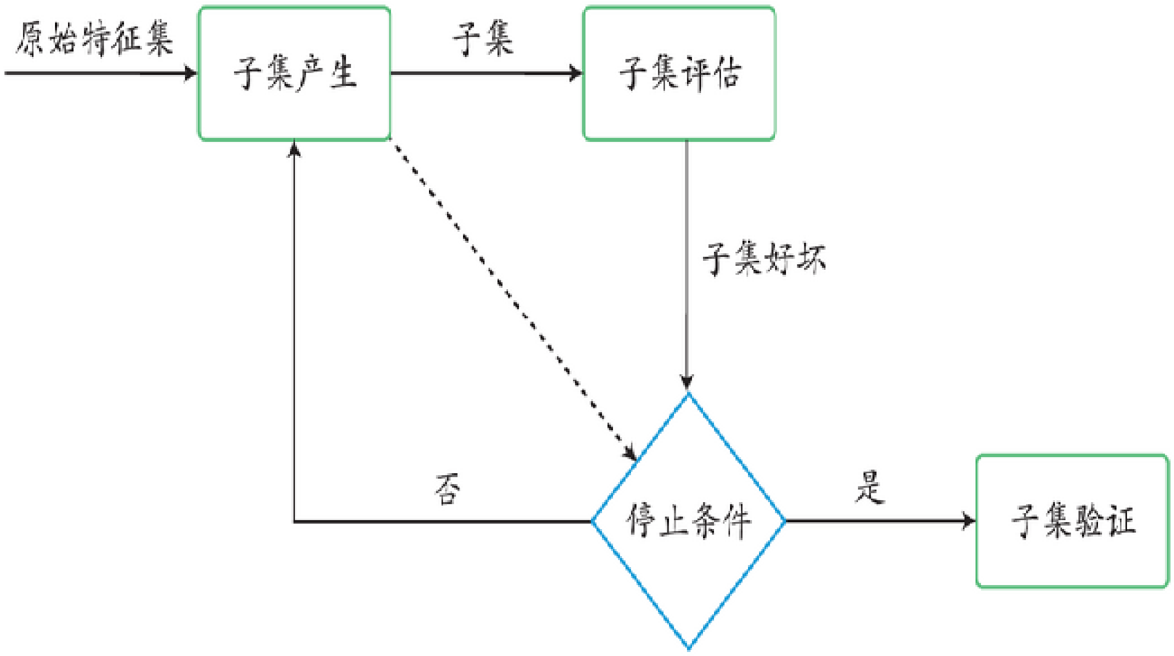
数据规约

- 获得数据集的一个简约表示，使得在容量上大大减小

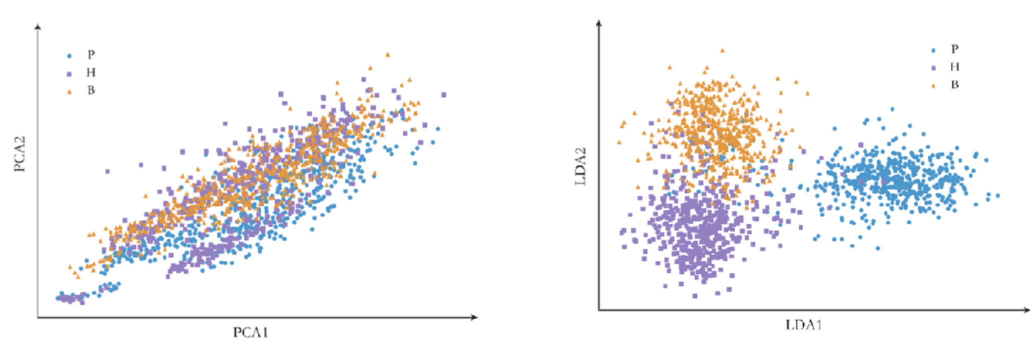
维规约

- 维规约能剔除不相关(irrelevant)或冗余(redundant)的特征，从而达到减少特征个数，简化模型，提高模型精确度，减少运行时间的目的
- 特征选择
  - 

- 特征选择流程图



- 特征选择在本质上是一个组合优化问题，看可以遍历所有的特征组合，寻找最优特征子集
- 主成分分析（PCA）
  - 构造原始特征的一系列线性组合形成低维的特征，以去除数据的相关性，并使降维后的数据最大程度地保持原始高维数据的方差信息
  - 最大化保留：
    - 直接去掉某一维度
    - 需要做一定的旋转
  - 算法步骤：
    - 对X中的每一个样本x_i进行中心化处理：$x_i = x_i - m$,其中m为样本均值
    - 计算协方差矩阵：$\sum = \frac{1}{n-1}X^TX$
    - 对协方差矩阵$\sum$做特征值分解，并降序排列：$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$
    - 取最大的前l个特征值相对应的特征向量$w_1, w_2, \cdots, w_l$组成转换矩阵W
    - 通过矩阵W^T从 $\mathbf{Y}$ 可以得到重构数据为 XWW^T
- LDA算法：
  - 算法步骤：

        - 计算数据集的均值m和每一类的数据均值m_c:$m=\frac{1}{n}\sum_{i=1}^{n}x_i$ $m_c=\frac{1}{n_c}\sum_{i=1}^{n_c}x_i$
        - 计算类内离散度矩阵$S_w=\sum_{c=1}^{C}\frac{n_c}{n}S_c$
        - 计算类间离散度矩阵S_b:$S_b=\sum_{c=1}^{C}\frac{n_c}{n}(m_c-m)(m_c-m)^T$
        - 计算$S_w^{-1}S_b$，并做特征值分解，并降序排列
        - 选取前l个特征值相对应的特征向量$w_1,w_2,\cdots,w_l$组成转换矩阵W
- LDA与PCA区别
    - 基本思想不同
        - PCA选择样本投影具有最大方差的方向，最大化保留了数据的内部信息
        - LDA则考虑标签信息，使得投影后不同类之间的样本距离最大化以及同类样本距离最小化
    - 学习模式不同
        - PCA属于无监督式学习，适用范围更广，但并不能保证数据降维后数据易于分析
        - LDA属于有监督学习，同时具有分类和降维的能力
    - 结果对比:
        - 

数量规约

- 线性回归模型: $Y=\alpha+\beta X$
    - 采用最小二乘法
- 多元回归模型：$Y=b_0+b_1X_1+b_2X_2+\cdots$

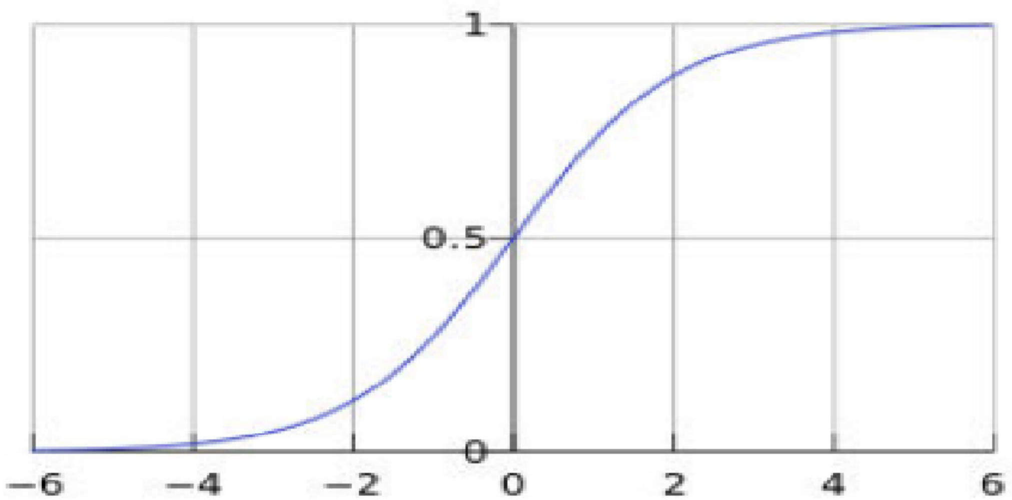
- 对数回归模型：$Y=\log_n X$
- 直方图：略
- 聚类：聚类技术把数据元组看做对象，将对象划分为群或簇，使得在一个簇中的对象相互相似，而与其他簇中的对象相异
- 抽样
    - s个样本的无放回简单随机抽样（SRSWOR）
    - s个样本的有放回简单随机抽样（SRSWR）
    - 簇抽样
    - 分层抽样

数据变换

数据标准化

- 0-1标准化：
    - 对数据进行线性变换，使其落在[0,1]的区间内
        - 函数：$x_i^*=\frac{x_i-min}{max-min}$
    - 若希望标准化后的数据以0为中心落在[-1,1]区间内
        - 函数：$x_i^*=\frac{x_i-\frac{max+min}{2}}{\frac{max-min}{2}}$
    - 0-1标准化适用于需要将数据简单地变换映射到某 一区间中,但其不足之处在于当有新数据加入时，可能会导致数据系列中的最大值或最小值发生变化，此时便需要重新定义最大值、最小值
- Z-Score标准化
    - 假设原取值集合为 {f_1, f_2,⋯f_n}，则f_i 经过 Z−Score 标准化后:
        - $f_i^{\backprime}=\frac{f_i-\mu}{\sigma}$
        - 其中$\mu=\frac{1}{n}\sum_{i=1}^{n}f_i$, $\sigma=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i-\mu)^2}$
    - Z-Score标准化是最常用的标准化方法，使得处理后的数据具有固定均值和标准差

        - 适用范围：Z-Score的标准化方法适用于数据系列中最大值或最小值未知、数据分布非常离散的情况。
        - 当数据中存在离群点时，为了降低离群值的影响，可以将标准差替换成平均绝对差:
            - $s = \frac{1}{n}\sum_{i=1}^{n}|f_i - \mu|$
- Logistic标准化
    - Sigmoid函数（又称Logistic函数）：
        - S形曲线
        - $S(x) = \frac{1}{1+e^{-x}}$



        - Logistic标准化方法适用于数据系列分布相对比较集中地分布于零点两侧

数据转换

- 转换原因：原始数据多为非数字型的数据
- 数字编码：
    - 创建单个数字特征来表示这个非数字特征
        - “收入水平”={贫困，低收入，小康，中等收入，富有}
        - 数字编码后转换成“收入水平”={0，1，2，3，4}
- One-hot编码
    - 独热编码（一位有效编码）

    - 使用N位状态寄存器来对N个状态进行编码，每个状态都有他独立的寄存器位，并且在任意时候，其中只有一位有效（为1），其余全为0
        - “国籍”={美国，英国，法国}
        - One-hot编码：美国（1,0,0）、英国（0,1,0）、法国（0,0,1）
        - 根据距离进行计算的模型得到的结论即为任意两国之间的距离均是$\sqrt{2}$
    - 优点：
        - 不会给名义型特征的取值人为地引入次序关系
        - 不同的原始特征取值之间拥有相同的距离
        - 线性回归模型中，对名义型特征，One-Hot编码通常优于数字编码
        - 对包含离散型特征的分类模型的效果有很好的提升
    - 缺点：
        - 特征维度会显著增多
        - 它会增加特征之间的相关性
- 哑变量编码
    - 对于一个包含K个取值的离散型特征，将其转换成K-1个二元特征
        - “国籍”={美国，英国，法国}
        - One-hot编码：美国（1,0,0）、英国（0,1,0）、法国（0,0,0）

离散化

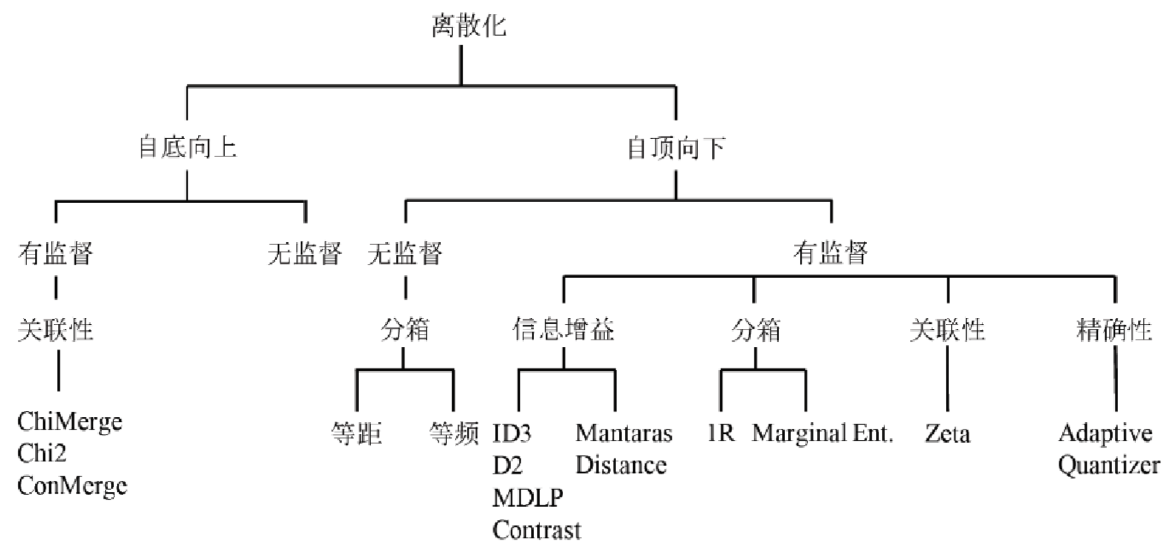
- 关联规则算法只能处理布尔类型的数据
- 决策树算法只能处理特征为离散型的数据
- 将连续性特征转换成为离散型特征的过程称为特征离散化（data discretization）
- 方法：
    - 将连续性特征的取值范围划分为若干区间段（bin），用区间段代替落在该区间段的特征取值
    - 区间段之间的分割点称为切分点（cut point）

    - 分割出来的区间段的个数称为元数（arity）
        - 特征排序
        - 切分点选择
        - 区间段分割或者合并
        - 重复直到满足终止条件
- 无监督的离散化方法
    - 等距离散化
    - 等频离散化
    - 基于聚类分析的离散化方法
- 有监督的离散化方法
    - 基于信息增益的离散化方法
    - 基于卡方的离散化方法
- 自顶向下的离散化方法
    - 等距离散化
    - 等频离散化
    - 基于信息增益的离散化方法
- 自底向上的离散化方法
    - 基于卡方的离散化方法
- 等距离散化:
    - 将区间均匀地划分成 k 个区间，每个区间的宽度相等，区间宽度ω：
        - $\omega = \frac{f_{max} - f_{min}}{k}$
    - 对数据质量要求高
    - 对离群值敏感
- 等频离散化
    - 根据连续性取值的总是 N,仍然将其划分为 k 个区段，每个区段包含的数据个数为：$\frac{N}{k}$
    - 示例:
        - {1,2,3,4,5,6,7,8,9,10,41,42,43,44,45,46,47,48, 49,50}

样本	区间	宽度
1, 2, 3, 4	$[1, 4]$	4
5, 6, 7, 8	$[5, 8]$	4
9, 10, 41, 42	$[9, 42]$	34
43, 44, 45, 46	$[43, 46]$	4
47, 48, 49, 50	$[47, 50]$	4
    - 取值相近的样本会被划分到不同区间，如8,9
    - 保证了每个区间段有相同的样本数
- 聚类离散化
  - 将相似的样本能落到相同的区间段内
  - 步骤：
    - 采用聚类算法（K-means，EM），把样本依据该特征划分成相应的簇或者类
    - 判断是否对簇进行进一步的分裂或合并（自顶向下继续运行聚类算法，或者自底向上对相邻的簇进行合并）
    - 确定切分点及区间的个数
- 信息增益离散化
  - 选择熵最小（信息增益最大）的特征作为正式分裂节点
  - 采用自顶向下的分裂策略
  - 步骤：
    - 对连续型特征进行排序
    - 计算出每一个取值相应的熵，选择熵最小的取值作为正式的切分点
    - 递归处理第二步中得到的两个新区间段，直到每个区间段内特征的类别一样为止
    - 合并相邻的，类的熵值为0且特征类别相同的区段，重新计算新区间段类的熵值
    - 重复第四步到满足终止条件（决策树的深度或叶子数）
- 卡方离散化

- 采用自底向上的合并策略
- 将特征的取值看作单独的区间
- 逐一递归进行区间合并逐一递归进行区间合并
- 比较两个总体之间是否存在显著性差异的方法：
  - $\chi^2=\sum_i^k \frac{(A_i-E_i)^2}{E_i}$
  - A_i 为落入区间段的样本个数(观察频数)
  - E_i 为对应的期望频数
- ChiMerge方法:
  - 判断相邻区间是否需要合并
  - 步骤：
    - 将连续型特征的每一个取值看作是一个单独的区间段，并进行排序
    - 针对每对相邻的区间段，计算卡方统计量。卡方值最小或者低于设定阈值的相邻区间段合并
      - $\chi^2=\sum_{i=1}^2 \sum_{j=1}^C \frac{(A_{ij}-E_{ij})^2}{E_{ij}}$
      - $E_{ij}=\sum_{j=1}^C A_{ij}\cdot \frac{\sum_{i=1}^k A_{ij}}{n}$
    - 对于新的区间段，递归进行第1,2步，只到满足终止条件
- 类别属性依赖最大化（CAIM）离散化：
  - 一种基于熵的特征离散化方法
  - CAIM二维表：
类别	$[d_0,d_1]$	$(d_1,d_2]$	$\cdots$	$(d_{k-1},d_k]$	类别样本数
1	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1\cdot}$
2	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C	n_{C1}	n_{C2}	$\cdots$	n_{Ck}	$n_{C\cdot}$
区间样本数	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot K}$	n
    - 评价离散化的好坏：
      - $CAIM=\frac{1}{N}\sum_{j=1}^K \frac{M_j^2}{n_{\cdot}^j}$

- 其中：$M_j = max(n_{ij}, n_{2j}, \cdots, n_{Cj})$
- 其中d_0和d_k分别为特征的最小值和最大值
- $n_i\cdot$表示属于类别 i 的样本个数
- $n._j$表示落在区间段(d_j−1,d_j]的样本个数
- n_{ij}表示在区间内$(d_j - 1, d_j]$的且属于类别 i 的样本个数
- CAIM的值越大,类和离散区间的相互依赖程度越大,离散化效果越好

- 总结:
  - 

数据脱敏

- 原则:
  - 单向性
  - 无残留
  - 易于实现

总结

- 数据预处理工作往往有一定代价的

- 导致数据损失，甚至可能对数据产生曲解。