



中国科学技术大学
University of Science and Technology of China

数据挖掘与数据仓库

第三讲 数据预处理

October 8, 2019



① 概述

② Summary

学号	姓名	性别	出生日期	电话	地址
SA13026003	张三		19700101	2312	地址1
SA13026003	李四	M	19700101	13805341123	地址2
...

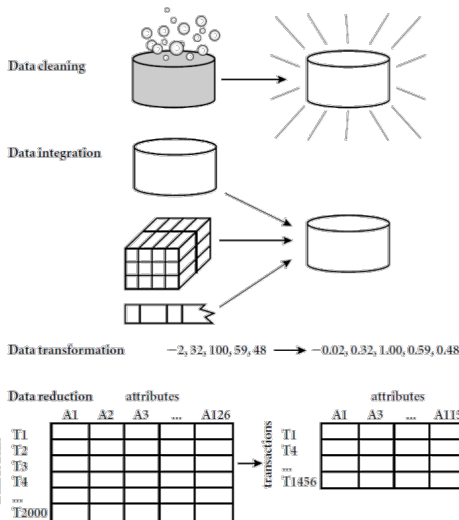
表中的问题

- 数据缺失: 空白没填写/使用了缺省值
- 数据不一致: 同一个数据在某些属性上表现出不同的值
- 错误的的数据: 明显错误的的数据 (如第一个数据的电话号码)

数据质量

- 准确性
- 完整性
- 一致性
- 时效性
- 可信性
- 可解释性





数据预处理的四项工作

● 数据清理

- 填充缺失数据
- 平滑噪声数据
- 识别/删除离群点
- 解决不一致性

● 数据集成

- 融合来自多个数据库/数据文件/数据立方的数据

● 数据归约

- 维归约
- 数值规约
- 数据压缩

● 数据变换及离散化

- 归一化
- 产生分层概念

存在哪些问题?

学号	姓名	性别	出生日期	电话	地址
SA13026003	张三		19700101	2312	地址1
SA13026003	李四	M	19700101	13805341123	地址2
...

脏数据：可能不正确的数据

- 张三的“性别”是什么？缺失值，不完备的数据
- 张三的“电话”是四位，错误，噪声数据，离群点
- 学号“SA13026003”究竟是张三还是李四？不一致
- 张三和李四都是 19700101 出生的？故意/伪造数据，使用缺省数据

学号	姓名	性别	出生日期	电话	地址
SA13026003	张三		19700101	2312	地址1
SA13026003	李四	M	19700101	13805341123	地址2
...

缺失值定义及产生原因

- 某些元组在部分属性上没有被记录下来的值
- 可能原因:
 - 数据产生设备故障
 - 数据录入时因为觉得不重要或者理解错误, 暂时放弃录入
 - 因为和其它数据不一致而被删除

场景	处理方法	使用条件
一批产品记录在表，标注了质量等级，其中部分产品质量等级缺失	忽略元组	任何属性的缺失率不能太高，若任何元组常同时缺多个属性则适用，会丢失潜在有用的数据
学生成绩单中某个学生缺少了2门课的成绩	手工填写真实值	费时，当缺失值多，数据集大时；故适用于重要的较小的数据集
一批产品记录在表，标注了质量等级，其中部分产品质量等级缺失	标注缺失质量等级为“unknown”	形成新的质量等级“unknown”？ “自动”使用全局常数代替缺失值，可能会影响将来的挖掘结果。
一批产品记录在表，其中部分产品重量值缺失	自动标注为属性的中心度量值	可以是均值或中位数，一般用于数值型数据。标称变量的均值？
一批产品记录在表，标注了质量等级，其中部分产品质量等级缺失	自动标注为产品所属类的中心度量值	如果没有类属性，怎么办？基于某种推理算法或机制，如决策树，Bayes推理等，复杂
张三的性别未填写	使用最有可能的值“M”	基于某种推理算法或机制，如决策树，Bayes推理等，复杂

Notes: 后四种方法，自动填写数据，会造成数据“有偏”。
数据库定义时，可控缺失数据；但是非关系数据库中的数据呢？

学号	姓名	性别	出生日期	电话	地址
SA13026003	张三		19700101	2312	地址1
SA13026003	李四	M	19700101	13805341123	地址2
...

噪声数据定义及产生原因

- 被测变量的随机误差或方差
- 产生原因：
 - 产生数据的仪器设备精度不够
 - 数据录入错误
 - 数据传输误差
 -(重复/不一致数据)

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28,

Partition into (equal-frequency) bins:

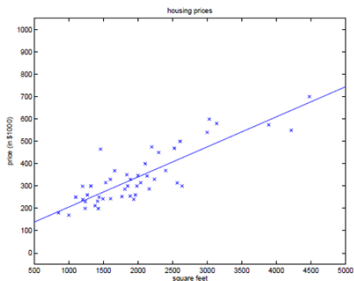
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24

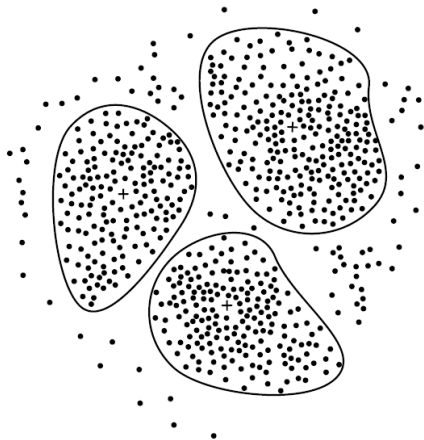


分箱法-binning

- 首先将数值型数据排序，然后分割成若干等份，保证每等份中数据个数一样多/或者将数值等份（等宽），每份中的数据个数可能不一样多
- 每个等份中的数据构成一个“箱子（bin）”
- 对每个箱子中的数据进行“平滑”处理，用中心度量来替换同一箱子中的所有数据
- 也可以不用中心度量，用每个箱子的最大值和最小值替换箱子中的数据，规则：数据离最大值近就用最大值替换，离最小值近就用最小值替换

回归法-regression

- 用一个函数，例如线性函数，拟合已有样本数据
- 在二元属性时，当确定其中一个属性时，另一个属性的值由拟合函数给出
- 多元回归将数据拟合到一个高维空间的曲面



离群点分析

- 右图，三个类，类中心用“+”标注
- 三个类之外的点被认为是“离群点” / “孤立点”

噪声数据处理的核心任务

进行数据平滑/光滑, smooth

思考：标称数据中的噪声？



经验整理

表头

- 充分利用“元数据”，扩展为利用描述数据的任何领域知识，将这些知识在数据清理中发挥作用
- 属性尽量不用复合编码，不同数据源的同一属性的识别非常重要
- 关系数据库的完整性约束要充分利用；
- 合理、充分利用现有的软件工具帮助进行数据清理，数据清理是枯燥乏味的体力活
- 不要害怕编写数据清理的小工具

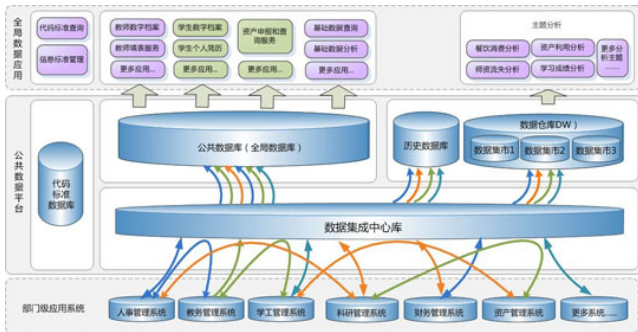


数据集成的含义:

将来源于多个数据源/库的数据合并到一个 coherent 存储中

思考: 数据集成的问题

来自不同数据库的表, 逻辑模式的合并? 元组的合并? 属性相等的识别? 行/列冗余? 不一致?





数据库表中的一行表示一个实体

实体及其识别

- 相同/不同属性名，可能是同一或者不同的属性，例如：bh（编号），id 表示两张不同源的表的相同含义的东西
- 不同表中的同一属性（不同名或同名），具有不同的元数据，例如：bh 是字母数字混合，id 是纯数字
- 不同源的参照完整性/函数依赖如何保持？
- 实体识别是数据集成的基础！几乎无法靠计算机自动实现！（手工）
- 非关系数据库中的数据



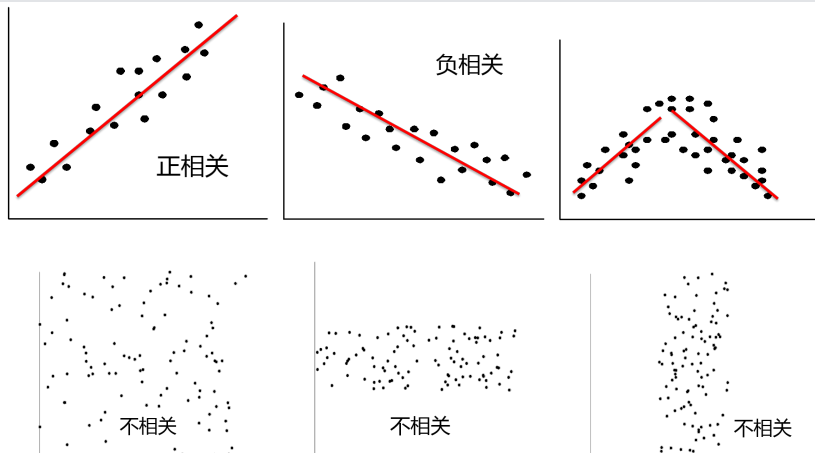
同一个数据出现两次即有冗余: 行冗余

- 单个数据源（库）不会出现，关系的完整性约束保证
- 不同源的数据行可能部分/完全重叠，例如，财务部门和人事部门关于雇员的基本资料可能存在冗余
- 可能有些数据库为了提高查询性能，减少表的连接操作而造成了冗余（去规范化）
- 冗余是造成不一致的主要原因之一，（其它不一致的原因包括数据格式/单位/编码方式/概念层次不一样，造成冗余）
- 最简单的冗余：关系中重复的行



复杂的冗余：列/属性的多次存储

- 同一属性，不同属性名
- 一个属性能被另一个/组属性“导出”，或“部分导出”
- 一个属性的取值/取值范围可以由另一个/组属性确定（或缩小）
- 如何发现列之间这种“隐蔽的冗余”？



数值型数据的线性相关性

如何定量描述上述线性相关性，而非绘图用人眼观察？

数值数据的相关系数 Pearson's product moment coefficient

设两个属性 A 和 B ，其线性相关系数为：

$$r_{A,B} = \frac{\sum_i (a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_i (a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

其中 n 是元组个数， \bar{A}, \bar{B} 分别是属性 A, B 的均值， σ_A, σ_B 分别是属性 A, B 的标准差； a_i, b_i 是元组 i 在属性 A, B 上的取值， $\sum_i (a_i b_i)$ 是 A, B 两个列向量的“点积”

进一步解释和说明：

- $-1 \leq r_{A,B} \leq 1$, $r_{A,B} > 0$ 表示 A, B 正（线性）相关， $r_{A,B} = 0$ 表示 A, B 不（线性）相关， $r_{A,B} < 0$ 表示 A, B 负（线性）相关
- $|r_{A,B}|$ 越接近 1，表示线性相关性越强，一个属性蕴涵另一个属性的可能性越大，或者说从一个属性能推断出关于另一个属性越多的信息
- 负相关表明一个属性值增加时，另一个属性值减少，正相关表明两个属性值同时增加或减少
- 相关并不意味蕴涵因果关系（一个属性的取值导致了另一个属性的取值）

数值数据的协方差, Covariance 给定两个属性 A, B , 及其上的 n 个样本 $(a_1, b_1), (a_2, b_2), \dots, (a_n, b_n)$, A, B 的协方差为:

$$\text{Cov}(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_i (a_i - \bar{A})(b_i - \bar{B})}{n} = E(A \cdot B) - \bar{A}\bar{B}$$

其中 $E()$ 表示求期望

$$\frac{\sum_i (a_i b_i) - n\bar{A}\bar{B}}{n}$$

进一步解释和说明:

- 注意到 $r_{A,B} = \frac{\text{Cov}(A,B)}{\sigma_A \sigma_B}$
- 当两个属性“同向”改变时 (同时变大或变小), 也就是说如果其中一个大于自身的期望值时另外一个也大于自身的期望值, 那么两个变量之间的协方差就是正值
- 当两个属性“反向”改变时 (一个变大, 一个变小), 即其中一个变量大于自身的期望值时另外一个却小于自身的期望值, 那么两个变量之间的协方差就是负值
- 所谓“变大/变小”, 相对于中心度量期望而言
- 若两个属性统计独立的, 那么二者之间的协方差就是 0; 反过来并不成立, 即如果两个属性的协方差为 0, 二者并不一定是统计独立的



Time point	<i>AllElectronics</i>	<i>HighTech</i>
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

- 左表表示两个公司（属于同一个行业）在 5 个不同时间点的股价
- 请问，他们的股价会一起涨跌吗？

协方差分析 $E(A) = (6 + 5 + 4 + 3 + 2)/5 = 4$

$E(H) = (20 + 10 + 14 + 5 + 5)/5 = 10.8$

$Cov(A, H) = (6 * 20 + 5 * 10 + 4 * 14 + 3 * 5 + 2 * 5)/5 - 4 * 10.8 = 7$

即协方差为正，两个公司的股票一起涨跌。

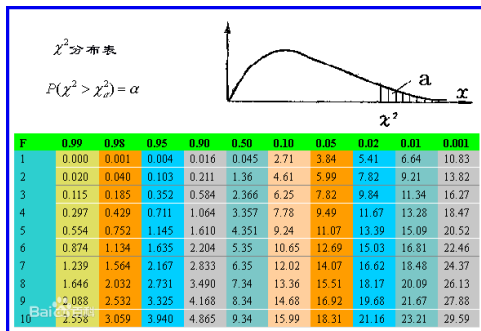
	Play chess	Not play chess	Sum (row)
fiction	250(90)	200(360)	450
Non-fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

标称属性的冗余：相关性检验

- 左表，行表示两种阅读材料（小说/非小说），列表示是否下棋
- 左表记录了 1500 个人的问卷调查，调查记录是否下棋和其喜爱的阅读材料是否是小说，结果如左表
- 能否断言或者说明，下棋决定/蕴涵喜爱阅读小说？或反之？

引入 χ^2 统计检验

- 假设两个属性 A, B ，给定 n 个观测样本 (a_i, b_j)
- 令 A 的值域为 (a_1, a_2, \dots, a_c) ， B 的值域为 (b_1, b_2, \dots, b_r)
- 绘制相依表， A 的 c 个值构成列， B 的 r 个值构成行，如上表
- 从观测样本中统计联合事件 $(A = a_i, B = b_j)$ 发生的概率/频度/出现次数，记作 o_{ij}
- 计算联合事件 $(A = a_i, B = b_j)$ 发生的期望概率/期望频度/期望出现次数，记作 $e_{ij} = (\text{count}(a_i) * \text{count}(b_j)) / n$ ，例如： $e_{11} = (\text{count}(\text{chess}) * \text{count}(\text{fiction})) / n = (300 * 450) / 1500 = 90$ ，上述相依表中括号内的值就是期望频度 e_{ij} 。
- e_{ij} 的计算蕴涵假设 “ A, B ” 是独立的
- 我们注意观察 o_{ij}, e_{ij} 的差异，二者差异越小，表明我们的假设 “ A, B ” 独立越接近于成立
- 故有统计量： $\chi^2 = \sum_i \sum_j (o_{ij} - e_{ij})^2 / e_{ij}$ ，该值越小， A, B 统计独立的假设越有可能被接受。
- 例子中 $\chi^2 = 507.93$ ，对于自由度 1，在 0.001 的置信水平下，拒绝假设的阈值是 10.828（查表可知），拒绝下棋和喜爱阅读材料二者之间独立的假设。二者不独立的概率超过 $1 - 0.001 = 0.999$



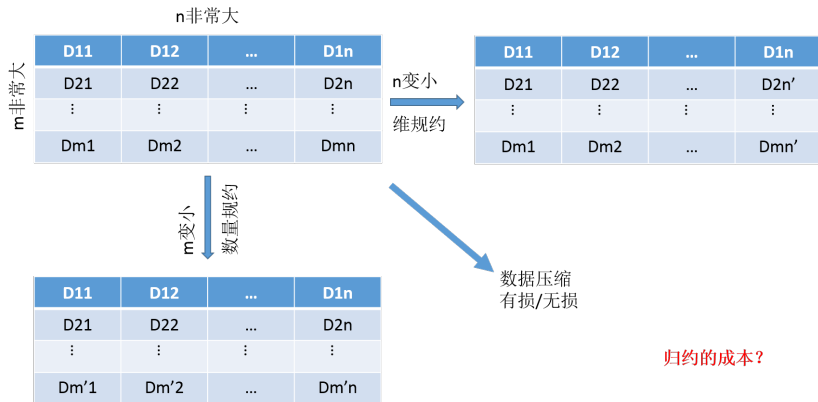
- 自由度 $F = (c - 1) * (r - 1)$
- 置信水平，相信统计独立的可能性

混合属性的相关性

思考：如果标称属性和数值属性混在一起，如何计算相关性？

数据规约的含义

数据集“量”被减少，但是保证数据“几乎”是完整的，使得最终挖掘结果与不规约的结果”几乎“一致。



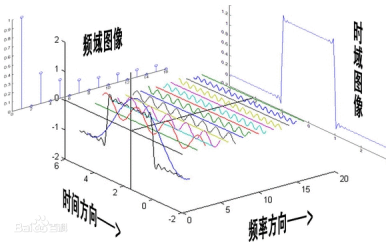
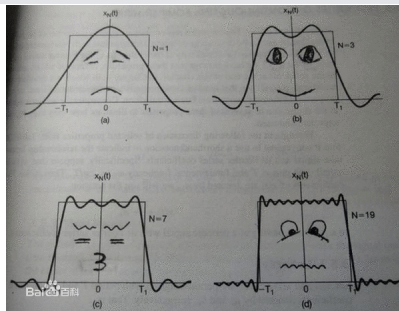
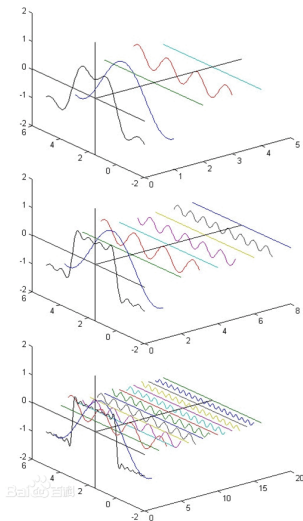
归约的成本？

一个向量经过线性变换，成为另一个线性空间中的向量

- 给定向量 x 和变换矩阵 W ，则变换后的结果 $y = xW$
- 假定原始未归约的数据为 x ，归约后的数据为 y ，则要求 y 的维数小于 x

其他变换

- 傅里叶变换
- 小波变换
- PCA 分析
- 特征选择/属性子集选择



变换公式

给定 N 个点的序列 $x[n]$, 其离散傅里叶变换后得到新的长度为 N 的序列 $\hat{x}[k] = \sum_{n=0}^{N-1} e^{-i\frac{2\pi}{N}nk} x[n]$

- 原始序列可以是时域/空域上的信号, 一个向量, 将其代入上述公式进行离散傅里叶变换, 得到等长新序列
- 取新序列的前若干项 (小于 N), 后面的高频项丢弃, 用缩短的序列进行傅里叶逆变换, 得到原始序列降维后的序列
- 令 $\omega = e^{-i\frac{2\pi}{N}}$, 则有离散傅里叶变换的变换矩阵:

$$W = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \omega^3 & \cdots & \omega^{N-1} \\ 1 & \omega^2 & \omega^4 & \omega^6 & \cdots & \omega^{2(N-1)} \\ 1 & \omega^3 & \omega^6 & \omega^9 & \cdots & \omega^{3(N-1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{N-1} & \omega^{2(N-1)} & \omega^{3(N-1)} & \cdots & \omega^{(N-1)(N-1)} \end{bmatrix}$$

线性变换的核心就是确定变换矩阵 W 的每一行 (基向量, 所有的基向量一起确定了一个线性空间), 不同的基向量组, 得到不同的线性变换。

步骤：例示

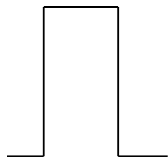
- ① 给定数据向量 $x = [45 \ 65 \ 5 \ 23 \ 48 \ 51 \ 78 \ 54]$
- ② 生成新的数据向量 (方法：顺序两两分组，计算均值和半差，并重排)
$$x' = [(45 + 65)/2 \ (5 + 23)/2 \ (48 + 51)/2 \ (78 + 54)/2 \ (45 - 65)/2 \ (5 - 23)/2 \ (48 - 51)/2 \ (78 - 54)/2] = [55 \ 14 \ 49.5 \ 66 \ -10 \ -9 \ -1.5 \ 12]$$
- ③ 对向量 x' 的前一半数据，重复步骤 2 的操作得到
$$x'' = [34.5 \ 57.75 \ 20.5 \ -7.25 \ -10 \ -9 \ -1.5 \ 12]$$
- ④ 对向量 x'' 的前一半数据，重复步骤 2 的操作，直到某个新得到的向量前一半数据只剩 1 个，算法停止。 $x''' = [46.125 \ -11.125 \ 20.5 \ -7.25 \ -10 \ -9 \ -1.5 \ 12]$

解释说明：

- 计算均值，是一种平滑技术，半差是记录细节差异，这个计算过程可逆！
- 一般把重排后的结果，前面一半称为“低频部分”，表示数据的平滑版本，描述整体“模样”；后面一半称为“高频部分”，描述细节特征
- 最终得到的向量就是“小波系数”，不同于离散傅里叶变换，小波系数的丢弃方法：将绝对值小于某个阈值 t 的所有系数置为 0
- 执行小波变换的逆过程，重建原始数据；或者存储/传输/分析有大量 0 的小波系数，实现数据压缩/归约

进一步讨论

- 傅里叶变换，用不同正弦信号的加权和来表示原始信号
- 小波变换，用不同缩放和平移的“母小波”的线性和来表示原始信号
- 上一个例子的线性变换矩阵如右图所示

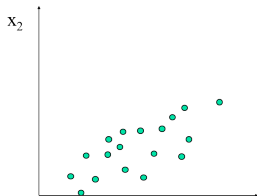


Haar2

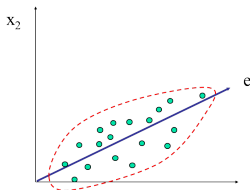


Daubechie4

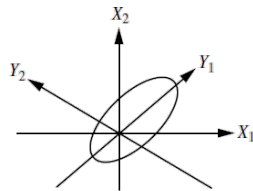
$$= \begin{bmatrix} \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & 0 & -\frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{8} & \frac{1}{8} & -\frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{8} & \frac{1}{8} & -\frac{1}{4} & 0 & 0 & -\frac{1}{2} & 0 & 0 \\ \frac{1}{8} & -\frac{1}{8} & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{8} & -\frac{1}{8} & 0 & \frac{1}{4} & 0 & 0 & -\frac{1}{2} & 0 \\ \frac{1}{8} & -\frac{1}{8} & 0 & -\frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{8} & -\frac{1}{8} & 0 & -\frac{1}{4} & 0 & 0 & 0 & -\frac{1}{2} \end{bmatrix}$$



(a) 给定一个 2-D 数据集



(b) 计算数据主要散布方向 e



(c) 计算数据集正交的两个主要散步方向 Y_1, Y_2

Figure: PCA 分析 (仅数值数据有效)

PCA 分析 (主成分分析)

- m 个 n 维数据, 转化为 m 个 k 维数据, 其中 $k \leq n$, 这 k 维数据最能“代表”原始数据
- 离散傅里叶变换和小波变换, 通过 $n \times n$ 的方阵, 把原始数据变换到“系数空间”, 通过对系数的丢弃/修改, 实现数据规约
- PCA 分析, 通过 $n \times k$ 的矩阵, 在原始数据空间中进行“投影”, 实现数据规约
- PCA 分析中, 投影的目标维就是所谓“主成分”, 数据最重要的维度/最具代表性的维度, 本质上就是方差最大的维度

主成分分析的进一步解释

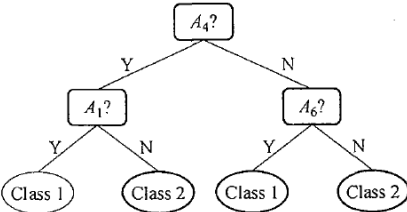
- 规范化每一维数据，使得所有维数据值域相同（仅对数值型数据有效）
- 计算 k 个正交的“主成分”或者“最重要成分”或者“散布最大的方向”
- 去掉不是“主成分”的方向/维度，实现降维
- 数学上的做法：首先获得协方差矩阵，然后对协方差矩阵进行特征分解，以得出数据的主成分（即特征向量）与它们的权值（即特征值），所谓主成分就是大特征值对应的特征向量/方向
- 从信息论角度考虑主成分分析在降维上是最优的

属性子集选择

- 去掉冗余的列，比如“单价/数量/总金额”三者中，任意一个属性都可看成冗余
- 去掉和挖掘任务无关的列，比如“学号/成绩”，二者相关吗？（假设挖掘任务是预测成绩）
- n 维原始数据，具有的属性子集个数为 2^n 。主成分分析，很可能获得的不是原始数据的维度，而这里必然是保留原始数据的维度！
- 选择最好的 k 个属性，何为“最好”？

以“分类挖掘”任务为例，属性子集的（贪婪的）选择方法

- 逐步添加：初始选择出的特征子集为空，每次选择一个“最好”的属性：将所有属性分别和类属性进行 χ^2 相关检验，将 χ^2 值最大（最相关）的属性选择出来；
- 逐步删除：初始选择出的特征子集包括所有属性，每次选择一个“最坏”的属性删除掉，所谓最坏，就是 χ^2 检验和类属性最不相关/独立的属性
- 逐步添加/逐步删除的混合方法
- 决策树方法

向前选择	向后删除	决策树归纳
<p>初始属性集: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>初始化归约集: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow 归约后的属性集: $\{A_1, A_4, A_6\}$</p>	<p>初始属性集: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>$\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow 归约后的属性集: $\{A_1, A_4, A_6\}$</p>	<p>初始属性集: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p>  <pre> graph TD A4["A4?"] -- Y --> A1["A1?"] A4 -- N --> A6["A6?"] A1 -- Y --> C1_1((Class 1)) A1 -- N --> C2_1((Class 2)) A6 -- Y --> C1_2((Class 1)) A6 -- N --> C2_2((Class 2)) </pre> <p>\Rightarrow 归约后的属性集: $\{A_1, A_4, A_6\}$</p>

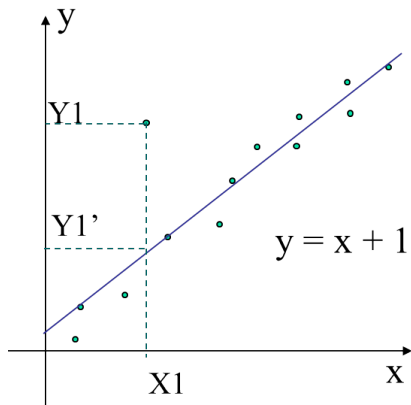


数量规约

用可替代的，较小的数据表示形式来数据，实现“元组”的减少

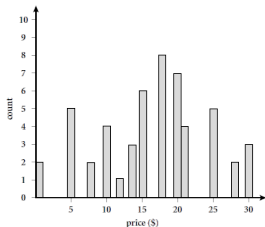
数量规约的方法

- 参数方法：linear regression, multiple linear regression, log-linear model
- 非参数方法：没有假定模型，如直方图，聚类，采样等
- 数据立方聚集

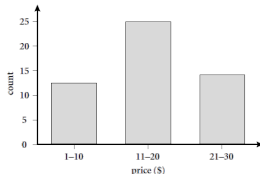


线性回归

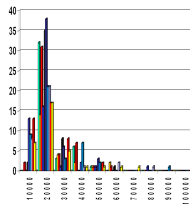
- 事先假定数据服从某个模型，比如线性模型 $y = wx + b$ ，其中 x 称为自变量， y 称为因变量，二者都是数据库中的数值属性/列
- 系数 w, b 称为回归系数，是需要寻找的最优参数，找到了最优参数，就是“最佳拟合”
- 寻找方法：最小二乘法（或其他）
- 所谓最小二乘：计算每个点到 w, b 确定的直线的距离平方和，找到一条直线（确定某组 w, b ），使得这个平方和最小。
- 多元线性回归：
 $y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$
- 对数线性模型：用于逼近离散的多维数据点的概率分布（教材很简略，阐述不清，课后自学）



(a) 单值桶/频率



(b) 等宽直方图



(c) 多维直方图

Figure: 各种直方图

解释说明

- 属性直方图，将数据在某个属性上的取值划分为不相交的子集（桶）
- 每个桶表示单个值及其出现次数（频率），称为单值桶
- 若每个桶值域宽度一样，称为等宽直方图
- 若每个桶内归入的数据个数基本相等，则称为等频/等深直方图
- 若用不同颜色表示不同维度的数据，则有多维直方图
- 用桶的均值或和来替换原始数据，实现数量规约



聚类用于数量规约

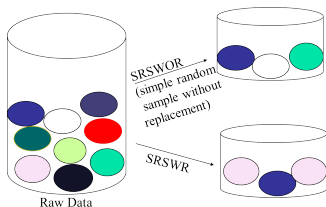
- 聚类：把元组视为数据对象，采用某种方法，把数据对象自动归为不同的“聚簇”/子集，使得簇内的对象尽可能相似，簇间数据对象尽可能相异
- 选用某种（比如重心，中心，特殊点）作为整个簇的代表
- 用簇来代表簇中所有的数据对象，即用簇的代表表示其它所有簇内数据，再消除重复元组，实现一个数据对象（簇代表）替换簇内所有对象，实现数量规约

抽样：

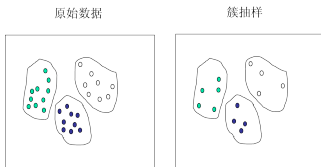
在一个大型数据集中，采用随机抽样的方式，选择一个小的样本来替换所有数据，简单说，就是选择能代表原始大型数据集的一个子集

抽样技术的种类

- 有放回简单随机抽样：假设每个元组被选择的概率是均匀的，被选择的样本仍在候选集中，可能被再次选中
- 无放回简单随机抽样：假设每个元组被选择的概率是均匀的，被选择的样本从候选集中删除
- 簇抽样：将数据分成若干个不相交的簇（簇的产生可以用聚类，也可以随机组合），从每个簇中抽样（比如，抽取簇中数据的 10% 为样本）
- 分层抽样：将数据对象分层（不相交），层内进行簇抽样（簇抽样的扩展）



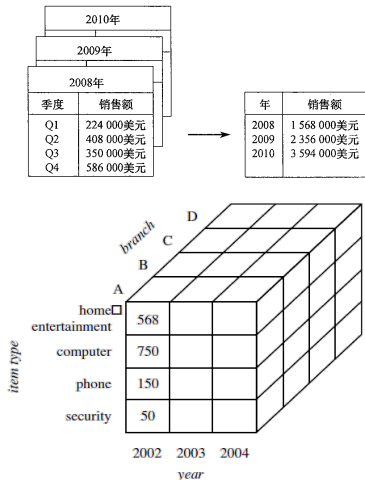
(a) 简单随机抽样

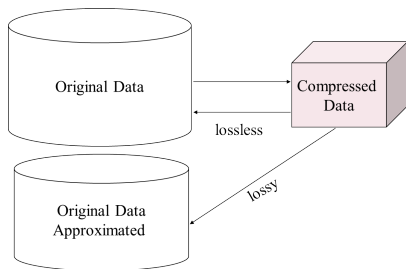


(b) 簇抽样/分层抽样

数据立方聚集实现数量规约

- 数据立方存储多维聚集信息，例如：利用关系数据库 SQL 中的聚集函数
- 右上图，数据挖掘任务和每年的销售总额相关，与每季度的销售额无关，因此“聚集”数据，不丢失影响任务的信息，但是数据量变小
- 右下图是一个数据立方体，每个单元格存放一个聚集，对应于多维空间的一个数据点
- 基本方体 (base cuboid): 面向分析任务的，最低层次的聚集 (抽象); 可以想象成在数据的某些维进行聚集 (求和/计数/求均值) 等等; 丢失不需要的细节信息;
- 顶点方体 (apex cuboid): 面向分析任务的，最高层次的聚集 (抽象), 数据没办法再使用聚集函数了
- 基本方体和顶点方体之间可能存在的不同聚集层次，越向上，数据量越小





数据压缩用于数据规约

- 字符串压缩：研究最多，一般无损，算法成熟
- 音频/视频压缩：研究非常多，一般有损
- 维规约和数量规约都可以看成是数据压缩
- 数据压缩的核心思想：发现冗余（相关性），消除冗余！

数据变换的含义

把给定属性的值域映射到一个新的值域，要求保证这个属性的某种特性，比如：保序，或保持任意两个值之间的相对距离

数据变换的方法

- 平滑: ? 去噪声的方法?
- 属性构造: 从已有的属性中构造新属性并添加到属性集中
- 对数据进行聚集/汇总: 构造新的属性?
- 规范化/归一化/标准化: 将值域按比例缩放，通常缩放到-1.0 1.0 或 0 1
- 离散化: 数值数据用区间标签或概念标签表示，如年龄是数值，可以转换成“婴幼儿/儿童/少年/青年/中年/老年”等年龄段的概念
- 标称数据的概念分层: 比如，年龄（数值）-> 年龄段（标称数据）-> “未成年人/成年人”-> “人”，这样的多层概念
- 各种方法/技术手段，可能达到多个不同的目的：如平滑可以去噪（清理数据），也可以当成数据变换；离散化和概念分层也可以看成是数据规约等等

- 最小-最大规范化: 对属性 A , 从 $[v_{min}, v_{max}]$ 映射到 $[v'_{min}, v'_{max}]$, 任意值 v 变换为 $v' = \frac{v-v_{min}}{v_{max}-v_{min}}(v'_{max} - v'_{min}) + v'_{min}$
- 例如, 收入 $[12000, 98000]$ 规范化到 $[0, 1]$, 则 73600 映射到 0.716
- z 分数规范化 (0 均值规范化): 设数据集在属性 A 上的均值为 μ_A , 标准差为 σ_A , 则对任意值 v 映射为 $v' = \frac{v-\mu_A}{\sigma_A}$
- 例如, 上例中 $\mu_A = 54000, \sigma_A = 16000$, 则有 $v' = 1.225$
- 小数定标规范化: $v' = \frac{v}{10^j}$, 其中 j 是使得 $\text{Max}(|v'|) < 1$ 的最小 j
- 例如, 上例中, $j = 5$, $v' = 73600 / (10^5) = 0.736$

理解数据离散化

- 将数值属性的连续值域分割成若干区间
- 每个区间添加一个标签，用标签代替原始数据
- 可以降低数据量
- 如果分割区间的过程利用了“类信息”，某个表示类的标签，则称为有监督的离散化，否则称为无监督的离散化
- 自顶向下离散化，从完整的值域（一个区间），逐步分裂成多个区间的方法
- 自底向上离散化，初始时刻每个值都是一个区间，依据某种规则依次合并邻近的区间获得更大的区间
- 离散化是一种预处理，为进一步的数据挖掘任务做准备



数据离散化的方法

- 分箱法，用箱的中心度量替换原始数据值，自顶向下，无监督
- 直方图，自顶向下，无监督
- 聚类分析，自顶向下/自底向上，无监督
- 决策树分析，自顶向下，有监督
- 相关性分析，如 chiMerge，自底向上，无监督

概念分层：一种数据变换方法

- 分类，有时也称为“概念学习”，分类器学习就是“概念学习”过程，每个类就是一个“概念”
- 概念分层，就是对标称数据（数值离散化之后也可以）形成多个不同层次/粒度组织形式
- 概念分层在数据立方中应用较多，较高层次的概念，来自低层次概念的“聚集/汇总”
- 手工/自动实现概念分层，有助于发现高层知识/抽象知识模式，能够在多个不同的抽象层次进行数据挖掘



- 数据质量：准确性/一致性/完整性/时效性/可信性/可解释性
- 数据清理：填补缺失/平滑噪声/识别离群点/纠正不一致
- 数据集成：将来自多个数据源的数据整合成一致的数据存储
- 数据规约：保证信心损失尽可能小（或不影响挖掘任务）的条件下，尽量减少数据的规模；维归约/数量规约/数据压缩
- 数据变换和离散化：规范化/离散化/概念分层



- ① 查阅对数线性模型，写不超过 10 pages 的读书笔记 (slides)
- ② 总结各种数据预处理技术（清理/集成/规约/变换）适用场景/条件