第3章: 数据预处理

- 为什么预处理数据?
- 数据清理
- 数据集成
- 数据归约
- 离散化和概念分层产生
- 小结

为什么数据预处理?

- 现实世界中的数据是脏的
  - 不完整的: 缺少属性值, 缺少某些感兴趣的属性, 或仅包含聚集数据
    - 例, occupation=“”
  - 含噪声的: 包含错误或偏离期望的离群值
    - 例, Salary=“-10”
  - 不一致的: 编码或名字存在差异
    - 例, Age=“42” Birthday=“03/07/2010”
    - 例, 以前的等级 “1,2,3”, 现在的等级 “A, B, C”
    - 例, 重复记录间的差异

数据为什么脏?

- 不完整数据源于
  - 数据收集时未包含
  - 数据收集和数据分析时的不同考虑.
  - 人/硬件/软件问题
- 噪音数据源于
  - 收集工具
  - 录入
  - 变换
- 不一致数据源于
  - 不同的数据源
  - 违反函数依赖

为什么数据预处理是重要的?

- 没有高质量的数据, 就没有高质量的数据挖掘结果!
  - 高质量的决策必然依赖高质量的数据
    - 例如, 重复或遗漏的数据可能导致不正确或误导性的统计.
  - 数据仓库需要对高质量数据进行一致地集成

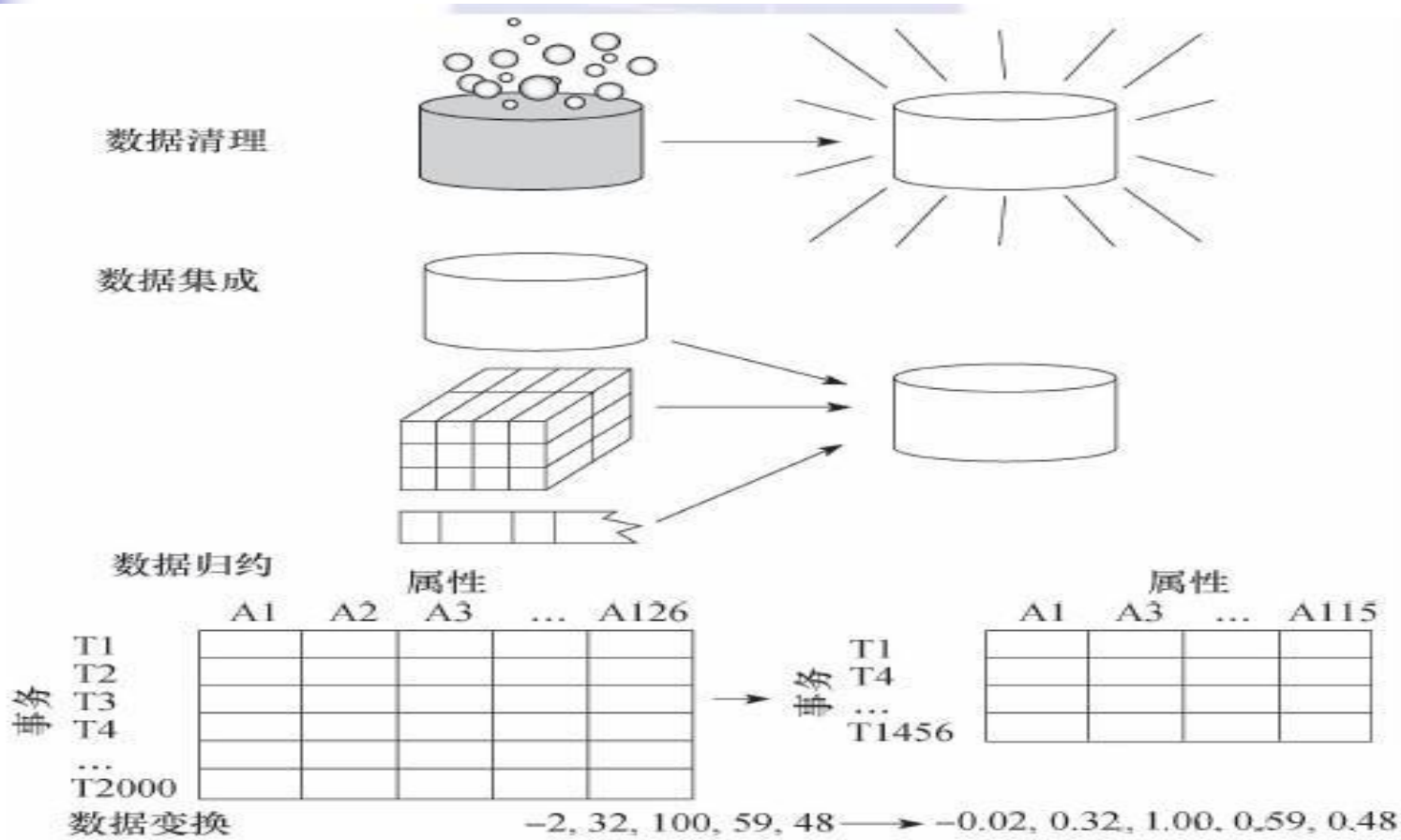
数据质量：一个多维视角

- 一种广泛接受的多角度:
  - 精确度(Accuracy)
  - 完整性(Completeness)
  - 一致性(Consistency)
  - 合时(Timeliness)：timely update?
  - 可信性(Believability)
  - 可解释性(Interpretability)
  - 可存取性(Accessibility)

数据预处理的主要任务

- 数据清理
  - 填充缺失值, 识别/去除离群点, 光滑噪音数据, 纠正数据中的不一致
- 数据集成
  - 多个数据库, 数据立方体, 或文件的集成
- 数据变换
  - 规范化和聚集
- 数据归约
  - 得到数据的归约(压缩)表示, 它小得多, 但产生相同或类似的分析结果：维度规约、数值规约、数据压缩
- 数据离散化和概念分层

数据预处理的形式


数据清理
数据集成
数据归约
属性
A1 A2 A3 … A126
事务
T1
T2
T3
T4
…
T2000
属性
A1 A3 … A115
事务
T1
T4
…
T1456
数据变换
−2, 32, 100, 59, 48 → −0.02, 0.32, 1.00, 0.59, 0.48

第3章: 数据预处理

- 为什么预处理数据?
- 数据清理
- 数据集成
- 数据归约
- 离散化和概念分层产生
- 小结

数据清理

- 现实世界的数据是脏：很多潜在的不正确的数据，比如，仪器故障，人为或计算机错误，许多传输错误
  - **incomplete**:缺少属性值, 缺少某些有趣的属性, 或仅包含聚集数据
    - e.g., *职业*=“ ” (missing data)
  - **noisy**:包含错误或孤立点
    - e.g., *Salary*=“−10” (an error)
  - **inconsistent**:编码或名字存在差异, e.g.,
    - 以前的等级 “1, 2, 3”, 现在等级 “A, B, C”
    - 重复记录间的差异
  - 有意的(e.g.,变相丢失的数据)
    - 如系统中默认生日为1月1号

如何处理缺失数据?

- 忽略元组: 缺少类别标签时常用(假定涉及分类)—不是很有效，当每个属性的缺失百分比变化大时
- 手工填写缺失数据: 乏味+费时+不可行 ?
- 自动填充
  - 一个全局常量 : e.g., “unknown”, a new class?!
  - 使用属性均值
  - 与目标元组同一类的所有样本的属性均值: 更巧妙
  - 使用最可能的值填充空缺值: 使用基于推理的方法，如贝叶斯公式或决策树

噪音数据

- 噪音: 被测量的变量的随机误差或方差
- 不正确的属性值可能由于
  - 错误的数据收集工具
  - 数据录入问题 data entry problems
  - 数据传输问题data transmission problems
  - 技术限制 technology limitation
  - 不一致的命名惯例 inconsistency in naming convention
- 其他需要数据清理的问题
  - 重复记录 duplicate records
  - 数据不完整 incomplete data
  - 不一致的数据 inconsistent data

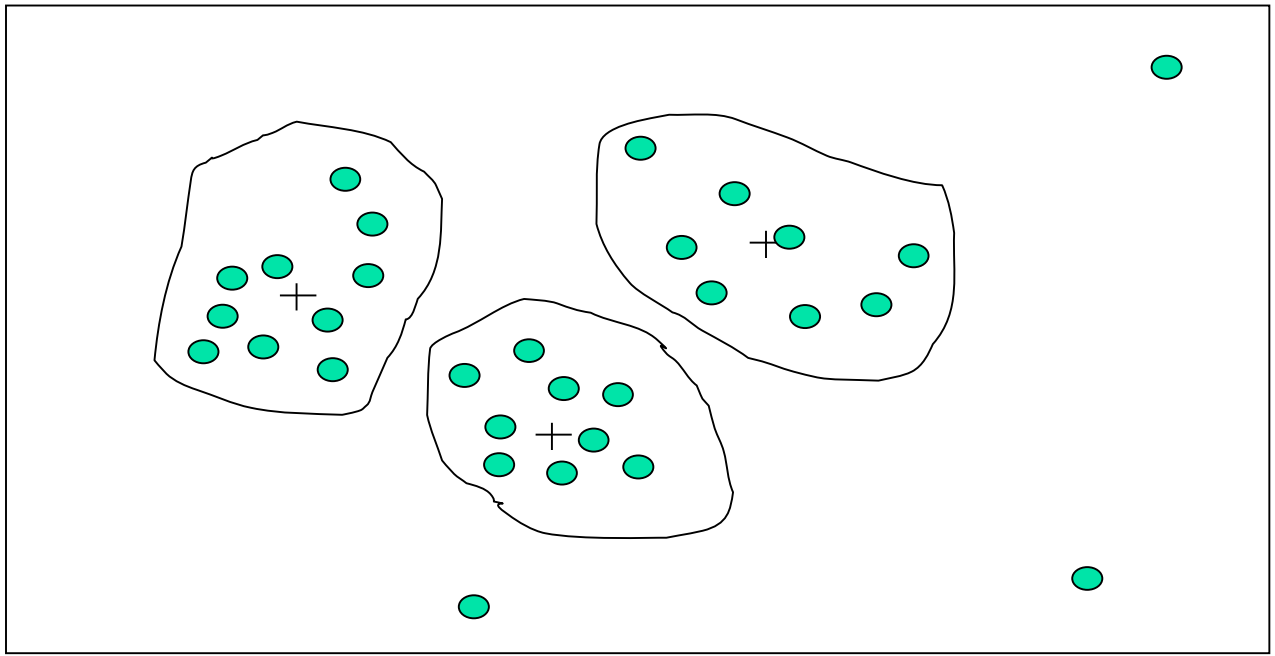
如何处理噪音数据?

- 分箱Binning method:
  - 排序数据，分布到等频/等宽的箱/桶中
  - 箱均值光滑、箱中位数光滑、箱边界光滑, 等.
- 聚类Clustering
  - 检测和去除离群点/孤立点 outliers
- 计算机和人工检查相结合
  - 计算机检测可疑数据，然后对它们进行人工判断 (e.g., deal with possible outliers)
- 回归 Regression
  - 回归函数拟合数据

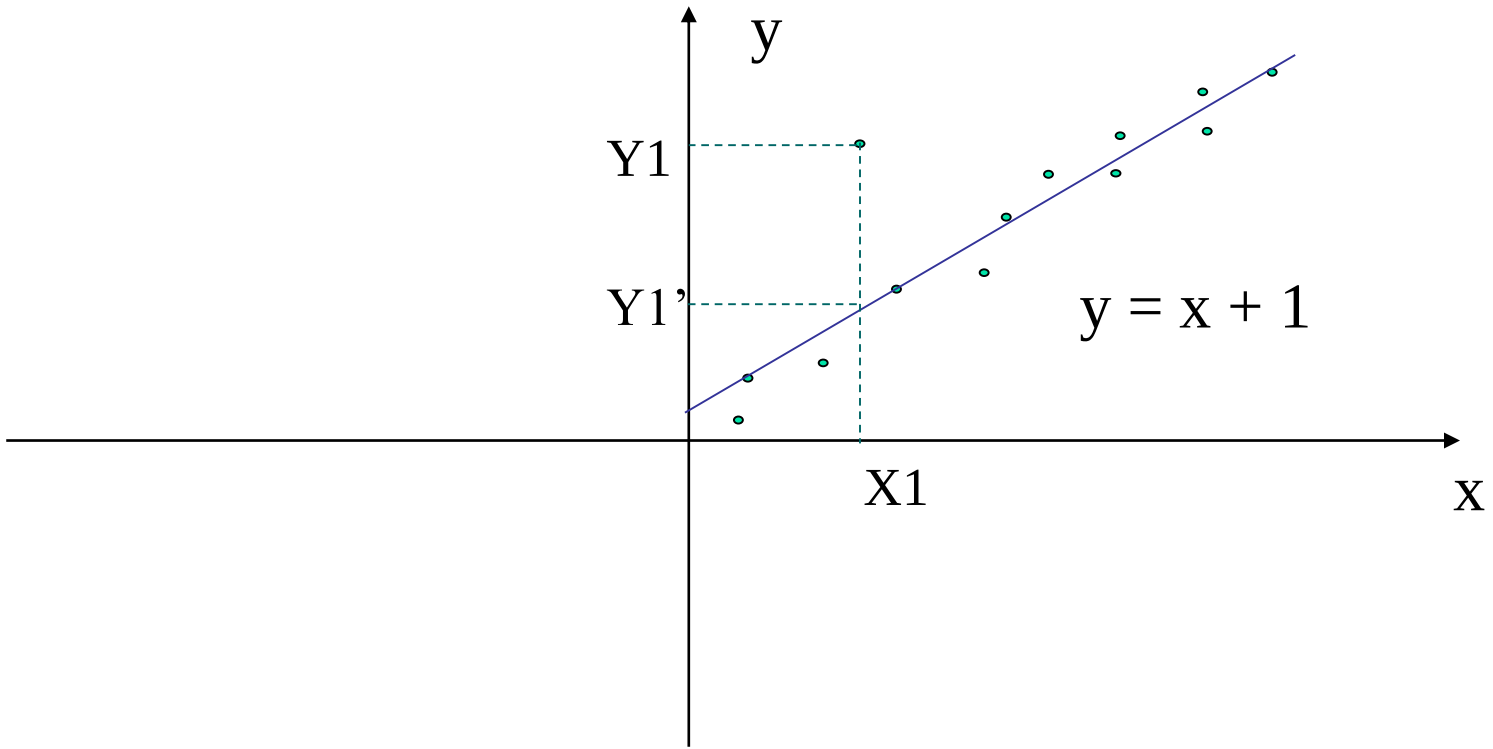
数据平滑的分箱方法

* **price的排序后数据(单位：美元): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34**
* **划分为（等深的）箱:**
  - **箱1: 4, 8, 9, 15**
  - **箱 2: 21, 21, 24, 25**
  - **箱3: 26, 28, 29, 34**
* **用箱平均值平滑:**
  - **箱1: 9, 9, 9, 9**
  - **箱2: 23, 23, 23, 23**
  - **箱3: 29, 29, 29, 29**
* **用箱边界光滑:**
  - **箱1: 4, 4, 4, 15**
  - **箱2: 21, 21, 25, 25**
  - **箱3: 26, 26, 26, 34**

聚类分析



回归



数据清理作为一个过程

- 数据偏差检测
  - 使用元数据(数据性质的知识)(e.g.,领域, 长度范围,从属, 分布)
  - 检查字段过载 field overloading
  - 检查唯一性规则, 连续性规则,空值规则
  - 使用商业工具
    - 数据清洗: 使用简单的领域知识(邮编, 拼写检查) 检查并纠正错误
    - 数据审计 : 通过分析数据发现规则和联系发现违规者(孤立点)
- 数据迁移和集成
  - 数据迁移工具:允许指定转换,如将串“gender”用“sex”替换
  - 提取/变换/装入工具ETL (Extraction/Transformation/Loading) tools: 允许用户通过图形用户界面指定变换
- 整合两个过程
  - 两个过程迭代和交互执行

第3章: 数据预处理

- 为什么预处理数据?
- 数据清理
- 数据集成
- 数据归约
- 离散化和概念分层产生
- 小结

数据集成

- 数据集成 **Data integration**:
  - 合并多个数据源中的数据，存在一个一致的数据存储中
  - 涉及3个主要问题：模式集成、冲突数据值、冗余数据
- 模式集成 **Schema integration**
  - 实体识别问题：多个数据源的真实世界的等价实体的识别。例如 **A.cust-id = B.customer-no, Bill Clinton = William Clinton**
  - 集成不同来源的元数据
- 冲突数据值的检测和解决
  - 对真实世界的实体，其不同来源的属性值可能不同
  - 原因:不同的表示,不同尺度,公制 **vs.** 英制

数据集成中冗余数据处理

- 冗余数据Redundant data （集成多个数据库时出现）
  - 目标识别：同一个属性在不同的数据库中有不同的名称
  - 衍生数据：一个属性值可由其他表的属性推导出, 例如“年收入”
- 小心的集成多个来源的数据可以帮助降低和避免结果数据集中的冗余和不一致，提高数据挖掘的速度和质量

数据变换

- 光滑: 去掉噪音，技术：分箱、回归、聚类
- 聚集Aggregation:汇总, 数据立方体构造
- 数据泛化Generalization:概念分层
- 规范化Normalization:按比例缩放到一个具体区间
  - 最小-最大规范化
  - z-score 规范化
  - 小数定标规范化
- 属性/特征构造
  - 从给定的属性构造新属性
  - 机器学习中称为：特征构造

} 数据规约

规范化数据的方法

- 最小-最大规范化

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

  - 新数据可能“越界”
- z-score 规范化

$$v' = \frac{v - 均值_A}{标准差_A}$$

- 小数定标规范化
  - 移动属性A的小数点位置(移动位数依赖于属性A的最大值)

$$v' = \frac{v}{10^j}$$ J为使得 Max($|v'|$)<1的整数中最小的那个

第3章: 数据预处理

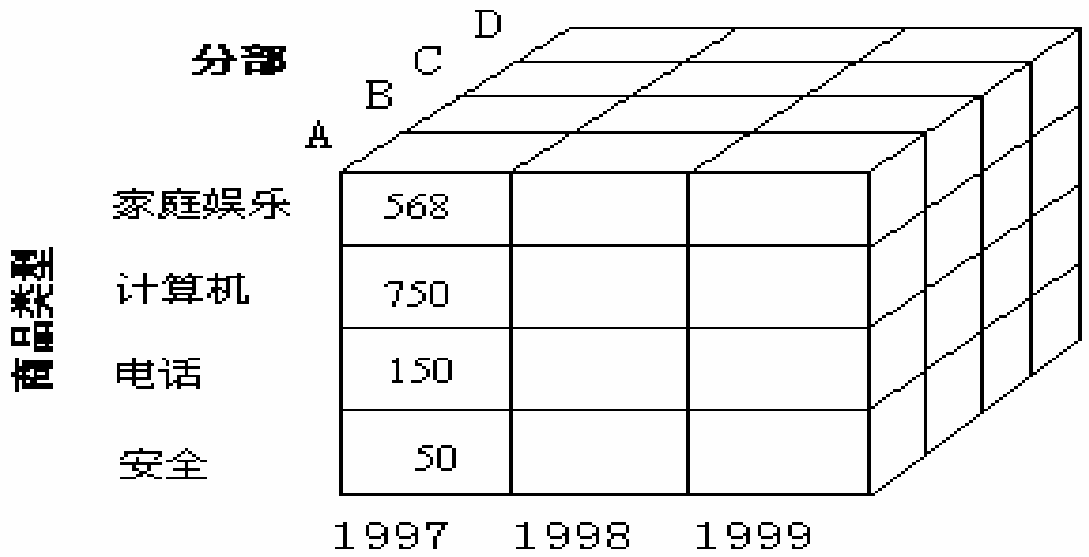
- 为什么预处理数据?
- 数据清理
- 数据集成
- 数据归约
- 离散化和概念分层产生
- 小结

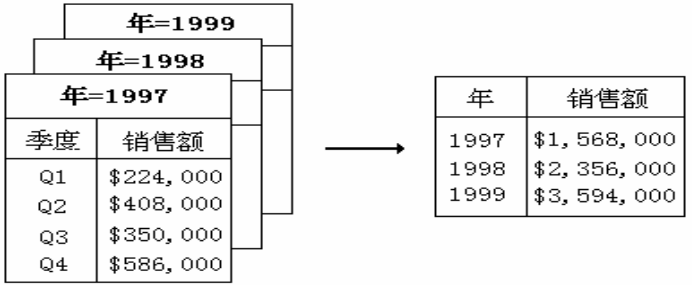
数据规约策略

- 在完整数据上的分析/挖掘耗时太长，以至于不现实
- **Data reduction** 获得数据集的一个规约表示，小很多，接近保持原数据的完整性，使得可得到相同/几乎相同的分析结果
- 数据规约策略如下；
  - 数据立方体聚集：聚集数据立方体结构的数据
  - 维度规约—去除不重要的属性
    - 主成份分析**Principal Components Analysis (PCA)**
    - 特征子集选择**Feature subset selection,**
    - 属性产生
  - 数据压缩 **Data Compression**
    - 基于离散小波变换的数据压缩：图像压缩
  - 数值规约，用其种表示方式替换/估计原数据

数据立方体聚

- 数据立方体存储多维聚
  - 某抽象层上建的数据
  - 最底层建的方体称为
  - 最高层的立方体称为
- 每个更高层的抽象将减少数据的

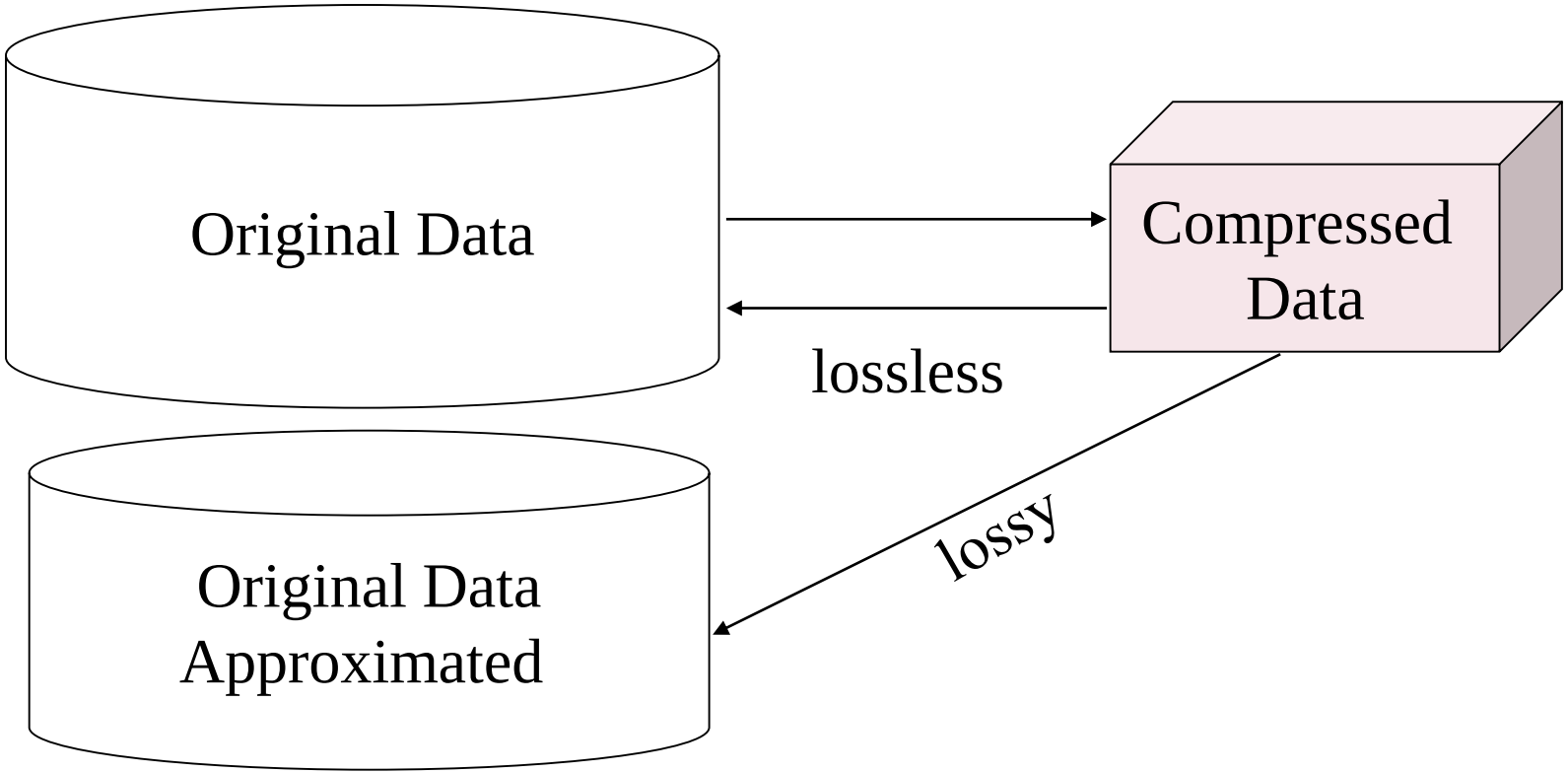




数据压缩 Data Compression

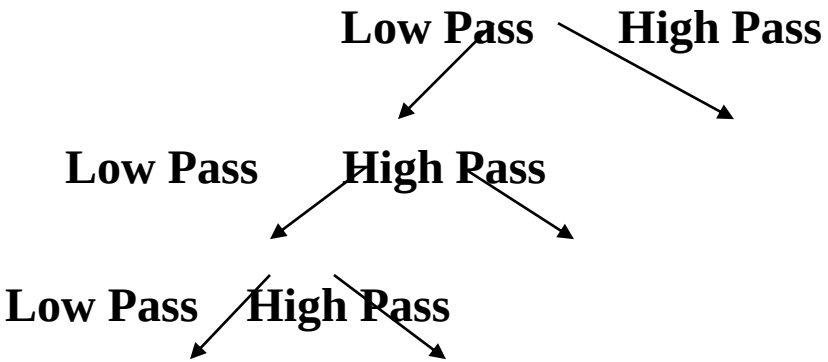
- 字符串压缩
  - 有丰富的理论和调优的算法
  - 典型的是有损压缩；
  - 但只有有限的操作是可可行的
- 音频/视频压缩
  - 通常有损压缩，逐步细化
  - 有时小片段的信号可重构，而不需要重建整个信号
- 时间序列不是音频
  - 通常短，随时间缓慢变化

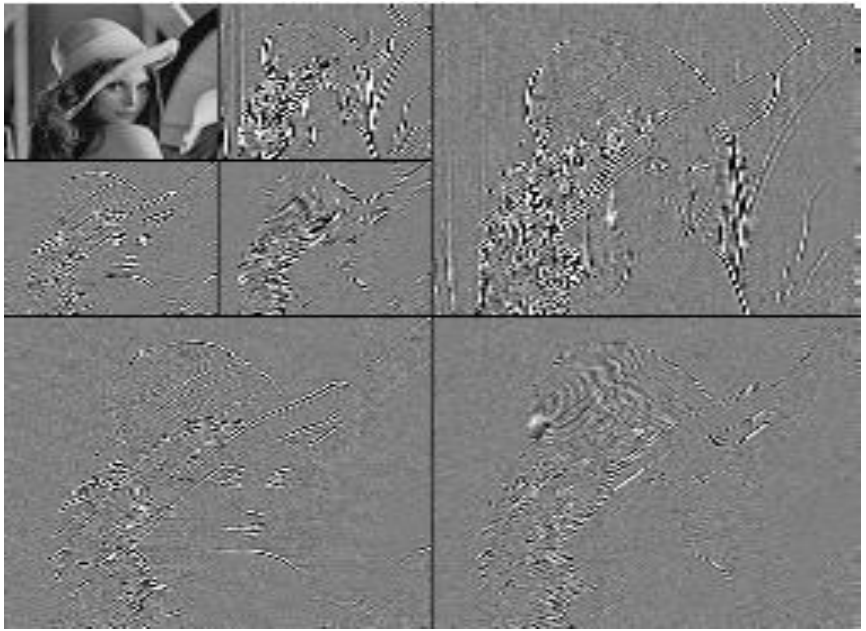
数据压缩


Original Data
Compressed Data
lossless
lossy
Original Data Approximated

DWT for Image Compression

- Image





- Discrete wavelet transform(DWT):

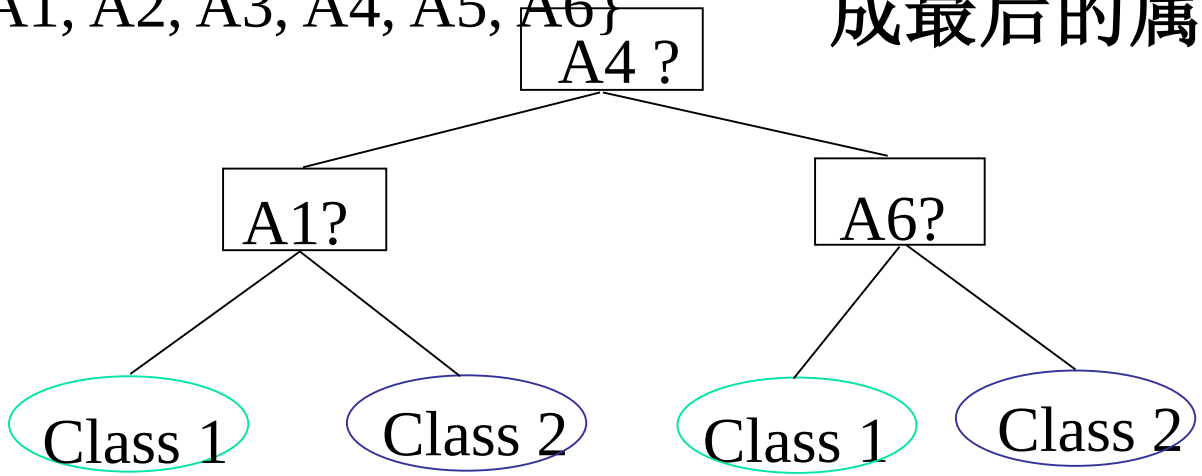
维度规约-特征选择

- 特征选择Feature selection (i.e., 属性子集选择):
  - 删除不相关/冗余属性，减少数据集维度
  - 找出最小属性集，类别的数据分布尽可能接近 使用全部属性值的原分布
  - 减少了发现的模式数目, 容易理解
- d个属性，有2^d个可能的属性子集
- 启发式方法Heuristic methods (因为指数级的可能性):
  - 局部最优选择，期望获得全局最优解
  - 逐步向前选择
  - 逐步向后删除 step-wise backward elimination

维度规约-决策树规约

最初的属性集合:
{A1, A2, A3, A4, A5, A6}

出现在决策树中的属性构成最后的属性子集



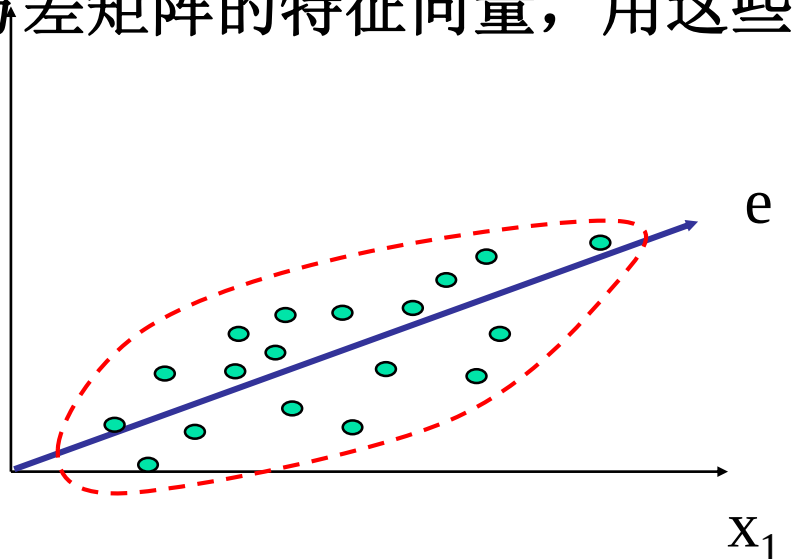
-------> 最后的集合: {A1, A4, A6}

维度规约-属性/特征产生

- **Feature Generation 产生新的属性，其可以比原始属性更有效地表示数据的重要信息。**
- **三个一般方法:**
  - **属性提取 Attribute extraction**
    - **特定领域的**
  - **映射数据到新空间**
    - **E.g., 傅立叶变换, wavelet transformation, 流形方法 (manifold approaches)**
  - **属性构造**

主成分分析 (PCA)

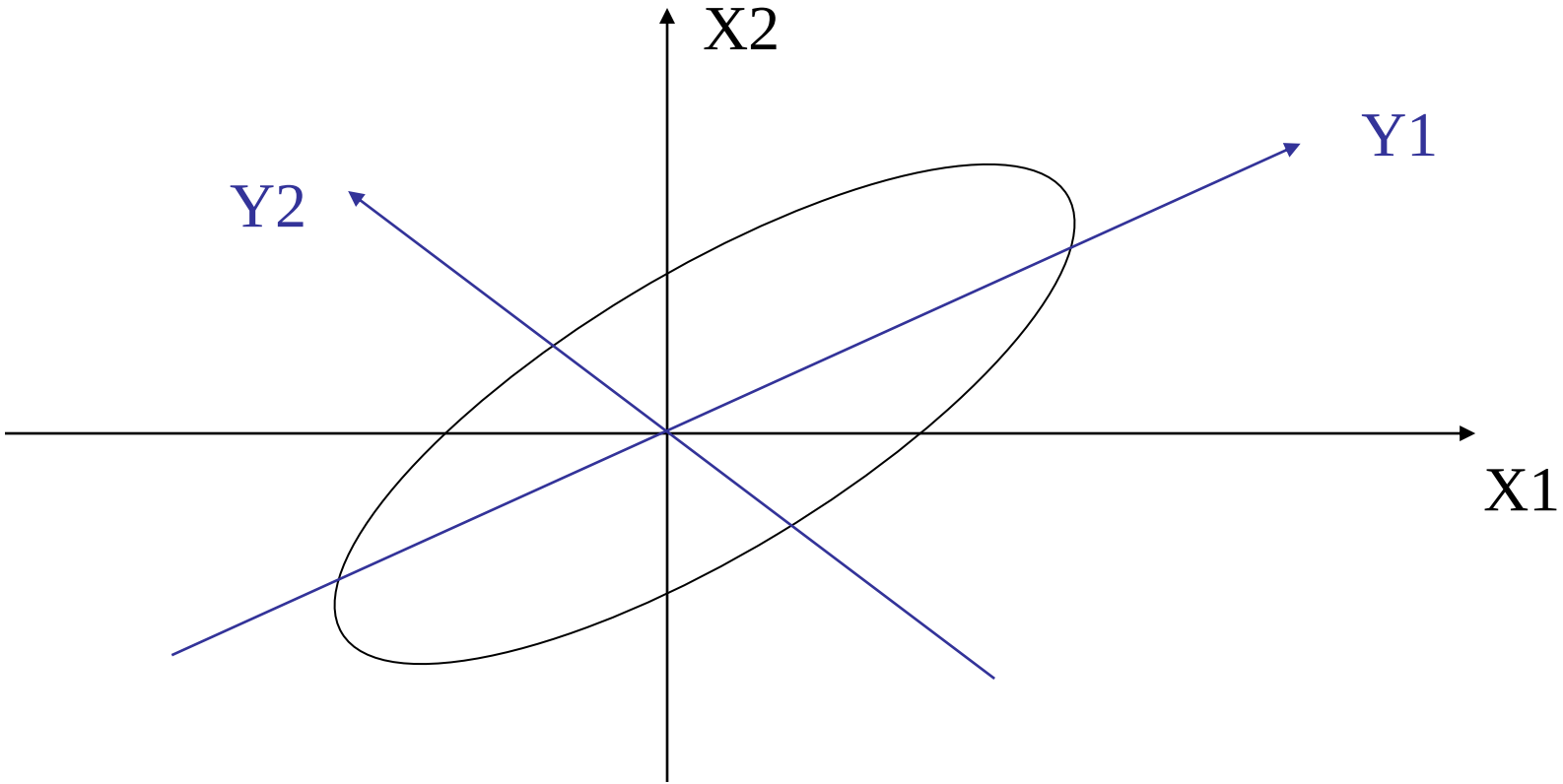
- **principal component analysis，K-L变换**
- 找到一个投影，其能表示数据的最大变化
- 原始数据投影到一个更小的空间中，导致维度减少.
  - 发现的协方差矩阵的特征向量，用这些特征向量定义新的空间



主成分分析 (Steps)

- 给定 ***p*维**空间中的N各点, 找到 $k \leq p$ ***个正交向量*** **(*principal components*)** 可以很好表示原始数据的
  - 归范化输入数据: 每个属性值位于相同的区间内
  - 计算 ***k*** 个标准正交向量, **i.e., *principal components***
    - 每个输入的点是这***k*** 个主成分的线性组合
  - **The principal components are sorted in order of decreasing "significance" or strength**
  - **Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*(i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)**

Principal Component Analysis



数值规约

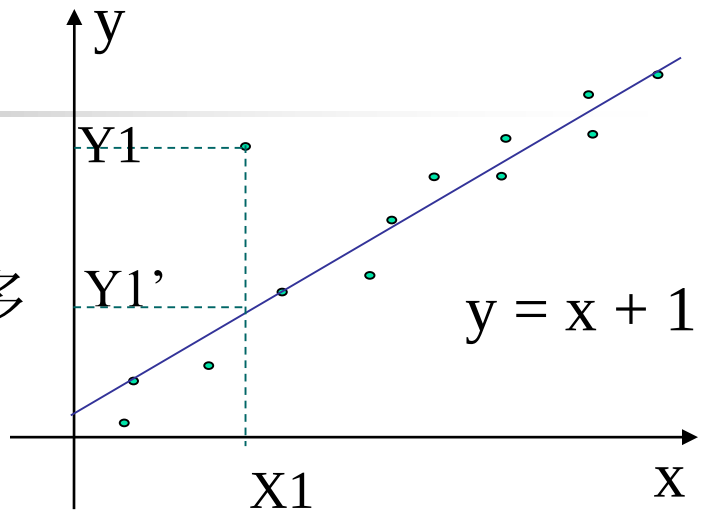
- 选择替代的、“较小的”数据表示形式
- 参数方法
  - 假设数据适合某个模型，估计模型参数，仅存储的参数，并丢弃数据（孤立点除外）
  - 对数线性模型:
    - 基于一个较小的维组合的子集来估计 离散属性的多维空间中每个点的概率
- 非参数方法
  - 不假定模型

回归和对数线性模型

- 线性回归: 数据拟合到一条直线上
  - 通常使用最小二乘法拟合
- 多元线性回归
  - 允许响应变量Y表示为多个预测变量的函数
- 对数线性模型:

回归分析

- 研究**因变量/响应变量Y(dependent variable/response variable)** 对个或多个**自变量/解释变量(*independent variable / explanatory variable*)**的相依关系的方法的统称
  - **参数需要估计以最好的拟合给定的数据**
- 绝大多数情况“最好的拟合”是由最小二乘法(*least squares method*)实

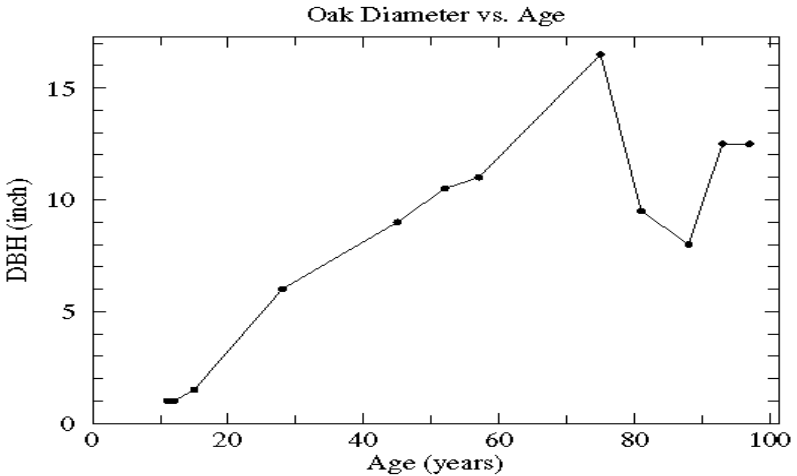


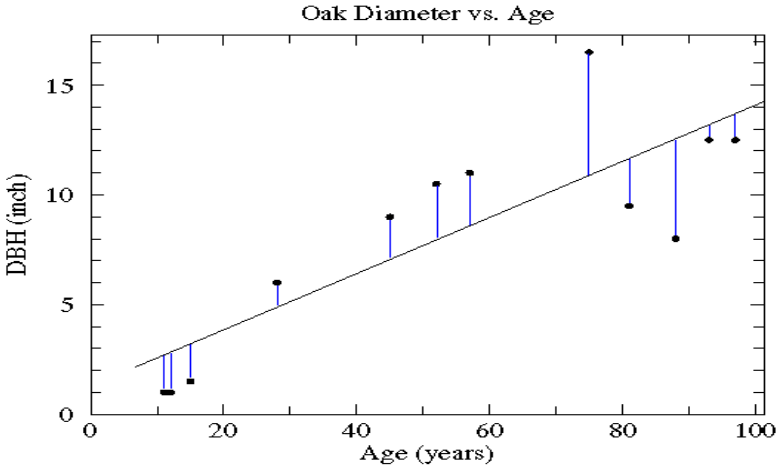
- **用于预测（包括时间序列数据的预测），推断，假设检验和因果关系的建模**

线性回归-用于预测

Y: --diameter at breast height(*DBH*) ←→ X: -- Age

	0	1	2	3	4	5	6	7	8	9	10	11	12
Y	?	1.0	1.0	1.5	6.0	9.0	10.5	11	16.5	9.5	8.0	12.5	12.5
X	34	11	12	15	28	45	52	57	75	81	88	93	97





线性回归(cont.)

- **Given x, construct the linear regression model for y against x as:**

$$y = \alpha + \beta x + e$$

- **L**

of α and β is $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ and

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}}, \qquad \text{where} \qquad s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$$

is the empirical covariance between x and y,

$$s_{xx} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

$$\hat{y} = \bar{y} + \frac{s_{xy}}{s_{yy}}(x - \bar{x}).$$

多元线性回归

- 响应变量: w，自变量: A1,A2,…,Ak.
- “5” 样本数目

$$w \quad A_1 \quad A_2 \quad \dots \quad A_k \qquad (1)$$

$$\min_{\mathbf{x}} \|A^{\mathrm{T}}\mathbf{x} - \mathbf{w}\|_2.$$

value α is estimated as a li
genes

$$\alpha = \mathbf{b}^{\mathrm{T}}\mathbf{x} = \mathbf{b}^{\mathrm{T}}(A^{\mathrm{T}})^{\dagger}\mathbf{w},$$

$$\begin{pmatrix} \mathbf{g}_1^{\mathrm{T}} \\ \mathbf{g}_{s_1}^{\mathrm{T}} \\ \vdots \\ \mathbf{g}_{s_k}^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} \alpha & \mathbf{w}^{\mathrm{T}} \\ \mathbf{b} & A \end{pmatrix}$$

$$= \begin{pmatrix} \alpha & \mathbf{w}_1 & \mathbf{w}_2 & \mathbf{w}_3 & \mathbf{w}_4 & \mathbf{w}_5 \\ \mathbf{b}_1 & A_{1,1} & A_{1,2} & A_{1,3} & A_{1,4} & A_{1,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{b}_k & A_{k,1} & A_{k,2} & A_{k,3} & A_{k,4} & A_{k,5} \end{pmatrix}$$

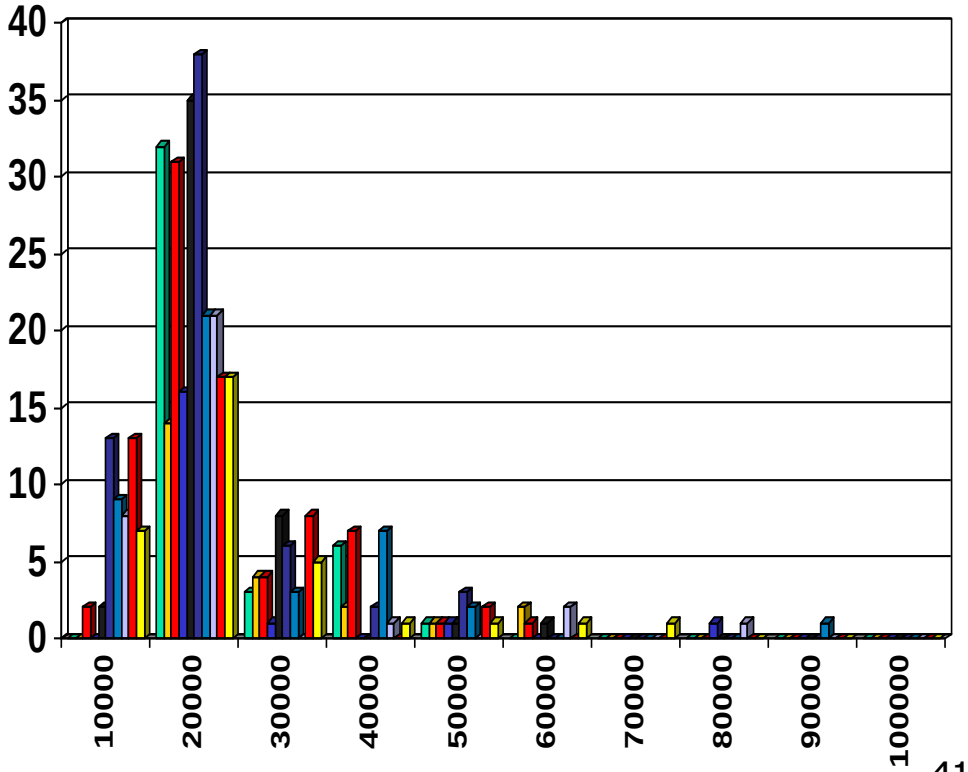
$$\mathbf{w} \simeq \mathbf{x}_1\mathbf{a}_1 + \mathbf{x}_2\mathbf{a}_2 + \cdots + \mathbf{x}_k\mathbf{a}_k,$$

are the coefficients of the linear combinat
least squares formulation (2). Accord
alue α in $\mathbf{g}_1$ can be estimated by

$$\alpha = \mathbf{b}^{\mathrm{T}}\mathbf{x} = \mathbf{b}_1\mathbf{x}_1 + \mathbf{b}_2\mathbf{x}_2 + \cdots + \mathbf{b}_k\mathbf{x}_k.$$

直方图Histograms

- 把数据划分成不相交的子集或桶
- 一维时可用动态规划优化构建
- 涉及量化问题



聚类Clustering

- 将对象划分成集/簇, 用簇的表示替换实际数据
  - 技术的有效性依赖于数据的质量
- 使用层次聚类，并多维索引树结构存放
- 非常多

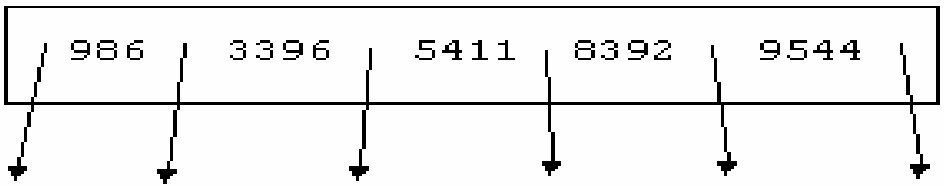


图 3.12　给定数据集的 B+树的根

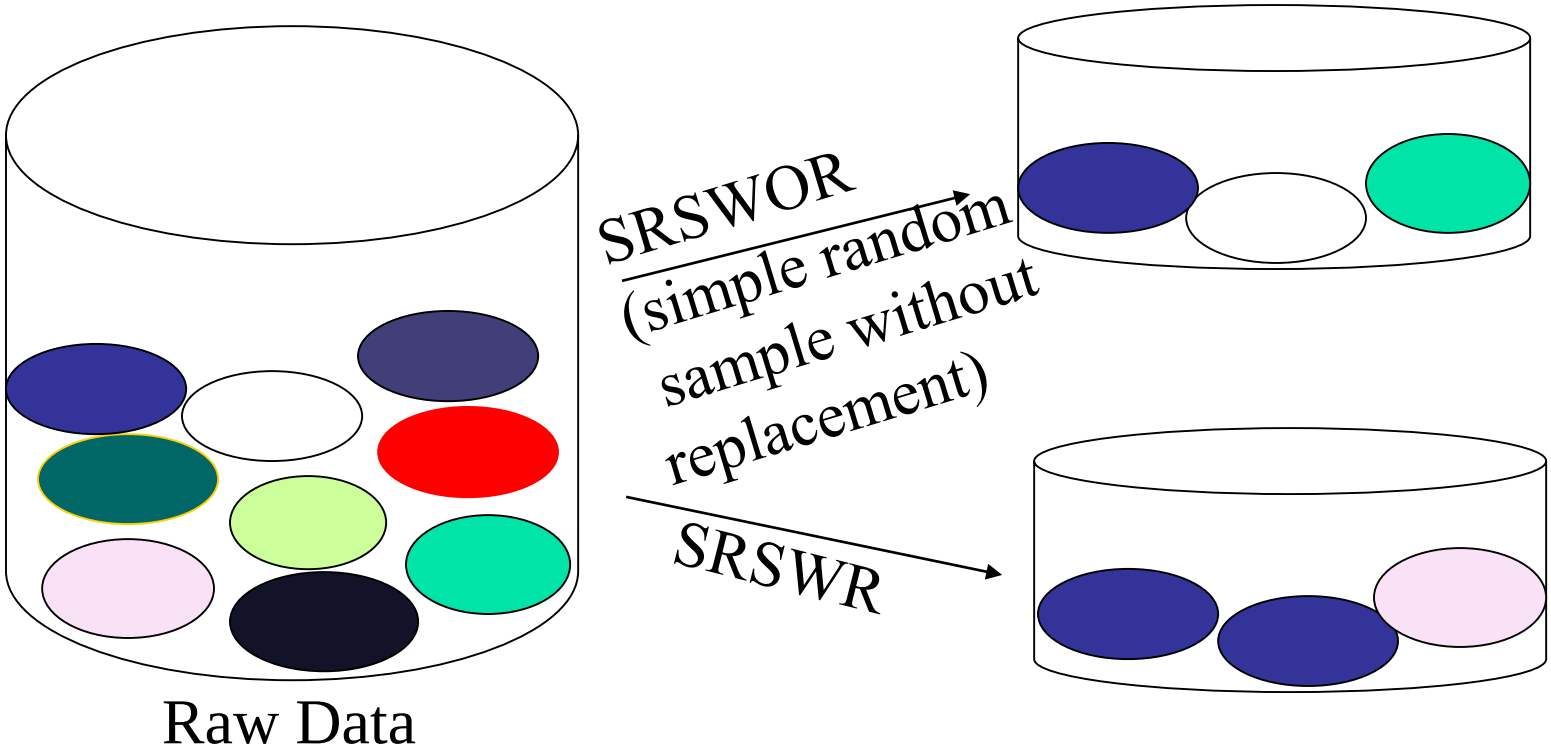
抽样Sampling

- 抽样: 获得一个小的样本集s来表示整个数据集 N
- 允许一个挖据算法运行复杂度子线性于样本大小
- 关键原则:选择一个有代表性的数据子集
  - 数据偏斜时简单随机抽样的性能很差
  - 发展适应抽样方法：分层抽样
- **Note: Sampling may not reduce database I/Os (page at a time)**

抽样类型 Types of Sampling

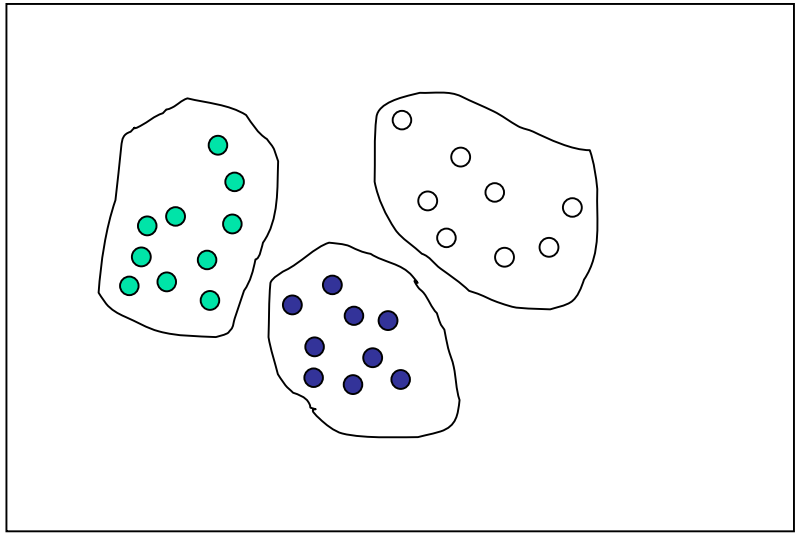
- 简单随机抽样 **Simple random sampling**
  - 相同的概率选择任何特定项目
- 无放回抽样 **Sampling without replacement**
  - **Once an object is selected, it is removed from the population**
- 放回抽样**Sampling with replacement**
  - 一个被抽中的目标不从总体中去除
- 分层抽样 **Stratified sampling:**
  - 把数据分成不相交部分(层), 然后从每个层抽样(按比例/大约相同比例的数据)
  - 偏斜数据

Sampling: With or without Replacement

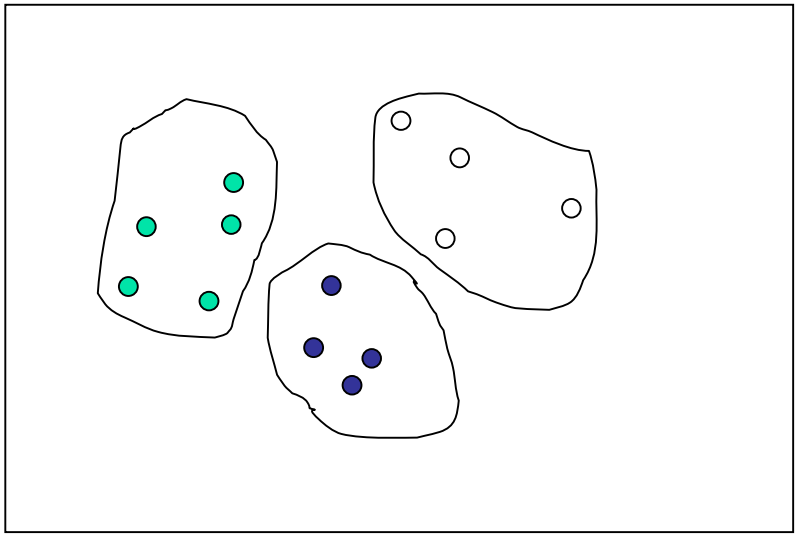


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



第3章: 数据预处理

- 为什么预处理数据?
- 数据清理
- 数据集成
- 数据归约
- 离散化和概念分层产生
- 小结

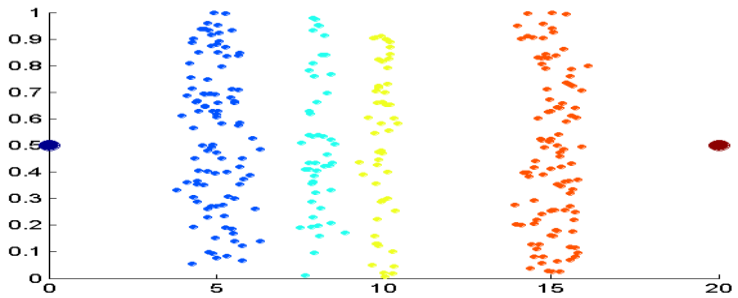
离散化 Discretization和概念分成

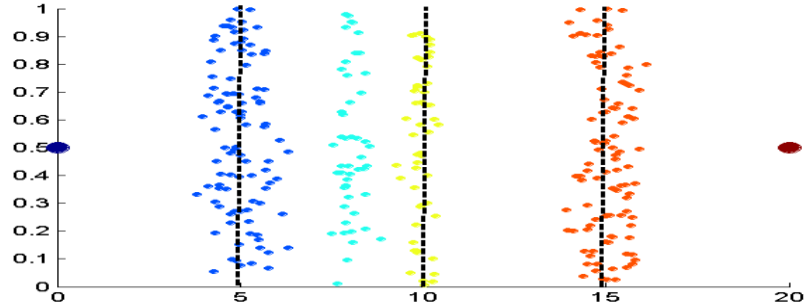
- 三种类型属性:
  - 名义 — values from an unordered set， color, profession
  - 顺序数 — values from an ordered set , e.g., military or academic rank
  - 连续 — real numbers
- 离散化 Discretization: 把连续属性的区域分成区间
  - 区间标号可以代替实际数据值
  - 利用离散化减少数据量
  - 有监督 vs. 无监督：是否使用类的信息
  - 某个属性上可以递归离散化
  - 分裂 Split (top-down) vs. 合并merge (bottom-up)

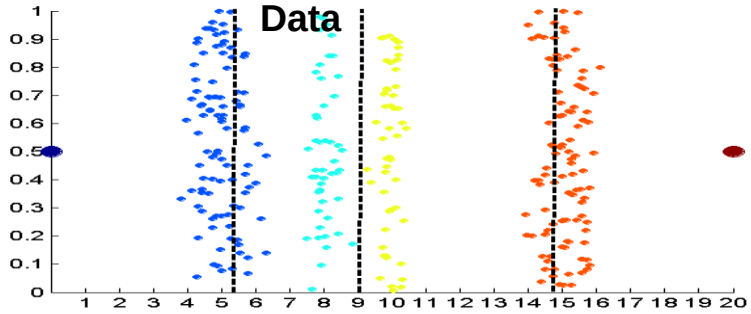
数值数据离散化/概念分层

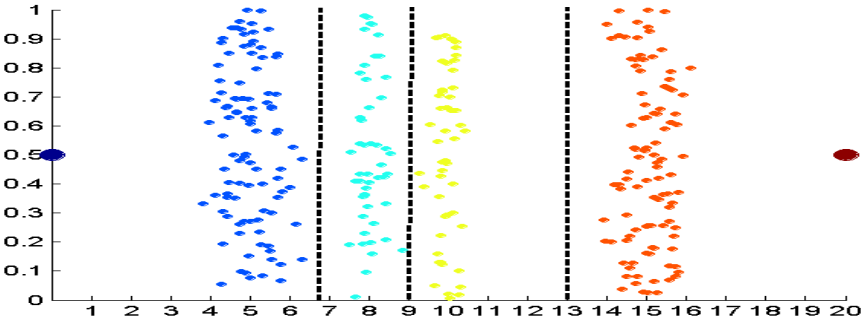
- 分箱 **Binning(Top-down split, unsupervised)**
- 直方图（**Top-down split, unsupervised**）
- 聚类 **(unsupervised, top-down split or bottom-up merge)**
- 基于χ^2 分析的区间合并**(unsupervised, bottom-up merge)**
- 基于熵 **Entropy-based discretization**
- 根据自然划分

不用类别(Binning vs. Clustering)









Equal frequency (binning)

K-means clustering leads to better results

基于熵Entropy的离散化

给定一个数据元组的集合S，基于熵对A离散化的方法如下：

1、A的每个值可以认为是一个潜在的区间边界

2、选择使

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

其中，S1和S2分别对应于S中满足条件A<T和A>=T的样本

$$Ent(S_1)=-\sum_{i=1}^{} p_i \log_2(p_i)$$

其中，Pi是类i在Si中的概率，等于S1中类i的样本数除以S1中的样本总数。

3、直到满足某个终止条件

Chi-merge离散化

- **Chi-merge: χ^2-based discretization**
  - **有监督: use class information**
  - **自低向上: find the best neighboring intervals (具有相似的类别分布, i.e., low χ^2 values) to merge**
  - **递归地合并, until a predefined stopping condition**

由自然划分离散化

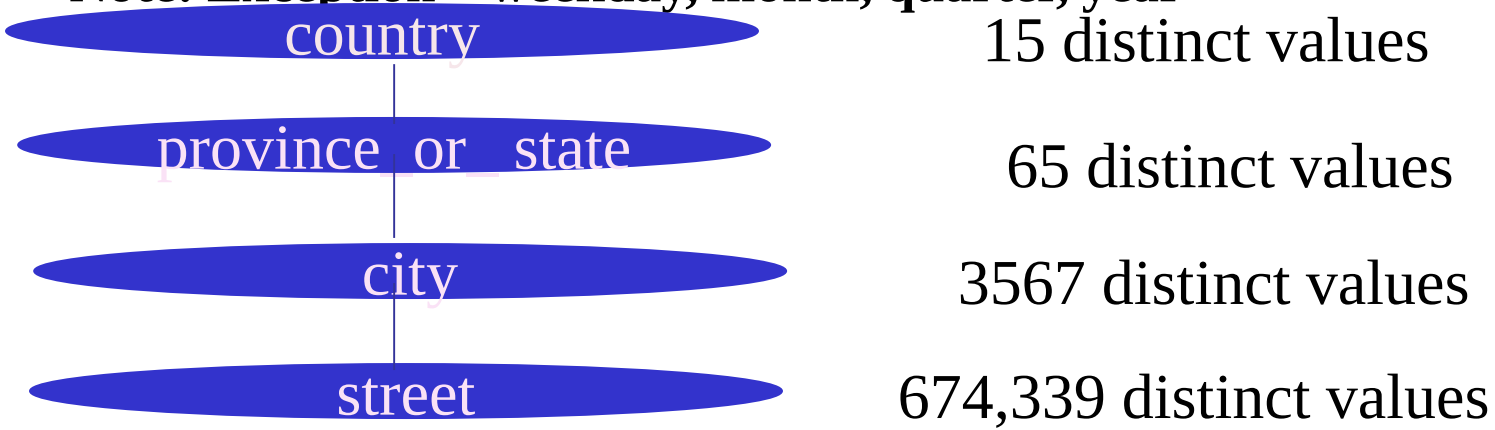
- 3-4-5 规则
  - 如果最高有效位包含 3, 6, 7 or 9 个不同的值, partition the range into 3 个等宽区间（7：2-3-2分成3个区间）
  - 2, 4, or 8 不同的值, 区域分成 4 个等宽区间
  - 1, 5, or 10 不同的值, 区域分成5 个等宽区间
  - 类似地，逐层使用此规则

分类数据的概念分层 Categorical Data

- 用户/专家在模式级显式地指定属性的偏序
  - **street<city<state<country**
- 通过显式数据分组说明分层
  - **{厄巴纳，香槟，芝加哥}<Illinois**
- 只说明属性集
  - 系统自动产生属性偏序，根据 每个属性下不同值的数据
  - 启发式规则：相比低层，高层概念的属性通常有较少取值
  - **E.g., street < city <state < country**
- 只说明部分属性值

自动产生概念分层

- **Some concept hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the given data set**
  - 含不同值最多的属性放在层次的最低层
  - **Note: Exception—weekday, month, quarter, year**



Summary

- Data preparation is a big issue for both warehousing and mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot a methods have been developed but still an active area of research

Data Reduction, Transformation, Integration

- Data Quality
- Major Tasks in Data Preprocessing
- Data Cleaning and Data Integration
  - Data Cleaning
    - i. Missing Data and Misguided Missing Data
    - ii. Noisy Data
    - iii. Data Cleaning as a Process
  - Data Integration Methods
- Data Reduction
  - Data Reduction Strategies
  - Dimensionality Reduction
    - i. Principal Component analysis
    - ii. Feature Subset Selection
    - iii. Feature Creation
  - Numerosity Reduction
    - iii. Data Cube aggregation
    - iv. Data Compression
    - v. Histogram analysis
    - vi. Clustering
    - vii. Sampling: Sampling without Replacement, Stratified Sampling
- Data Transformation and Data Discretization
  - Data Transformation: Normalization
  - Data Discretization Methods
    - i. Binning
    - ii. Cluster Analysis
    - iii. Discretization Using Class Labels: Entropy-Based Discretization
    - iv. Discretization Without Using Class Labels: Interval Merge by *Â*2 Analysis
  - Concept Hierarchy and Its Formation
    - i. Concept Hierarchy Generation

References

- E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Bulletin of the Technical Committee on Data Engineering. Vol.23, No.4*
- D. P. Ballou and G. K. Tayi. Enhancing data quality in data warehouse environments. Communications of ACM, 42:73-78, 1999.
- H.V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4), December 1997.
- A. Maydanchik, Challenges of Efficient Data Cleansing (DM Review - Data Quality resource portal)
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999.
- D. Quass. A Framework for research in Data Cleaning. (Draft 1999)
- V. Raman and J. Hellerstein. Potters Wheel: An Interactive Framework for Data Cleaning and Transformation, VLDB'2001.
- T. Redman. Data Quality: Management and Technology. Bantam Books, New York, 1992.