



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有问题的
 - **不完全**: 缺少属性值, 缺少某些有趣的属性, 或仅包含聚集数据
 - 例, occupation=""
 - **噪音**: 包含错误或孤立点
 - 例, Salary="-10"
 - **不一致**: 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?

- 不完全数据源于
 - 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
 - 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的统计.
 - 比如, 数据仓库需要高质量数据的一致集成



数据质量：一个多维视角

- 一种广泛接受的多角度：
 - 正确性(Accuracy)
 - 完全性(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, 并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

数据清理 Data Cleaning

- 现实世界de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.,
 - *Age*=“42”, *Birthday*=“03/07/2010”
 - 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)
 - Jan. 1 as everyone’s birthday?



Background

- **Containing missing values**
 - **Insufficient resolution**
 - **Image corruption**
 - **Dust**
 - **Scratches on the chip**
 - **Experiment error**

	sample1	sample2	sample3	sample4
Gene1	2.23	1.32	0.62	6.1
Gene 2	3.12	?	3.12	0.89
Gene 3	-0.65	2.15	?	0.23
Gene 4	?	2.001	-3.10	?
...	1.23	5.14	2.88	4.31
...				
Gene p



Background(cont.2)

- **Some algorithms and analyses can not be used for data with missing values**
 - Clustering algorithm
 - Support vector machines
 - Principal component analysis
 - Singular value decomposition ...
- **One solution: repeat experiments**
 - Economic reason
 - Limitation of available biological material



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，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
 - **最可能的值：基于推理的方法，如贝叶斯公式或决策树**
 - 其它方法.



处理方法

- 删除带有缺失数据的基因或样本
 - ?
- 重新进行实验
 - ?
- 对缺失数据进行估计



Problem of Missing Values Estimation

- **Input: Given matrix (A) with missing values.**
- **Output: Complete Matrix, after estimating the missing values as accurate as possible.**



Exist Solutions

- **Imputing Zero**
- **Row/Column Average**
- **KNNimpute (2001)**
- **SVDimpute (2001)**
- **BPCA (2003)**
- **LSimpute (2004)**
- **LLSimpute(2005)**
- **Others**



简单的方法

- **Imputing Zero**
- **Row/Column Average**
 - 例子



KNImpute

- Missing value estimation methods for DNA microarrays
- **BIOINFORMATICS, 2001, Vol. 17(6): 520–525**

$$d_{nm}^2 = \frac{1}{M} \sum_{i=1}^M (x_{ni} - x_{mi})^2, \quad (7)$$

$$\hat{x}_{re} = \frac{\sum_{i=1}^K \frac{x_{ie}}{d_{ri}}}{\sum_{i=1}^K \frac{1}{d_{ri}}}, \quad (8)$$

Value of k

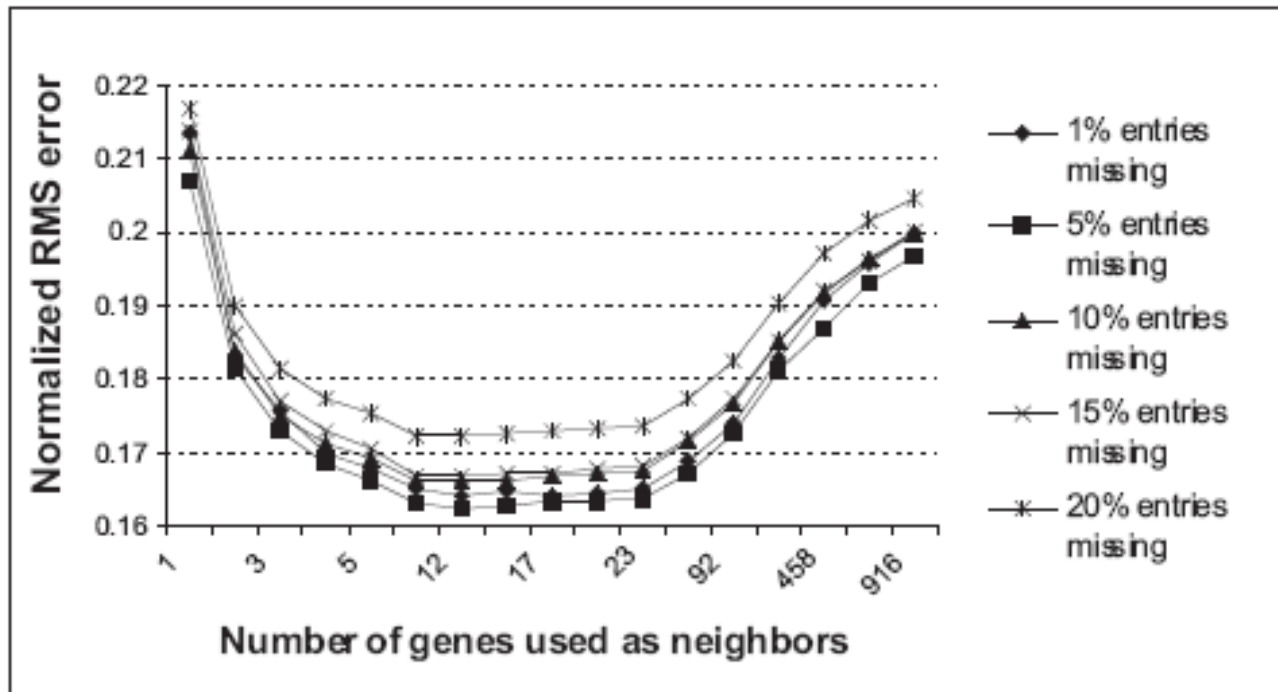


Fig. 1. Effect of number of nearest neighbors used for KNN-based estimation on noisy time series data. Different curves correspond to experiments performed for data sets with different percent of entries missing.



Lsimpute: accurate estimation of missing values in microarray data with least squares methods

Nucleic Acids Research, 2004, 32(3): e34



Six Lsimputes

- **Two basic methods**
 - **LSimpute_gene—the most basic**
 - **LSimpute_array**
- **Two weighted combinations**
 - **LSimpute_combined**
 - **LSimpute_adaptive**
- **EMimpute_gene & EMimpute_array**
 - **Based on Expectation-Maximization algorithm**



Lsimpute-gene

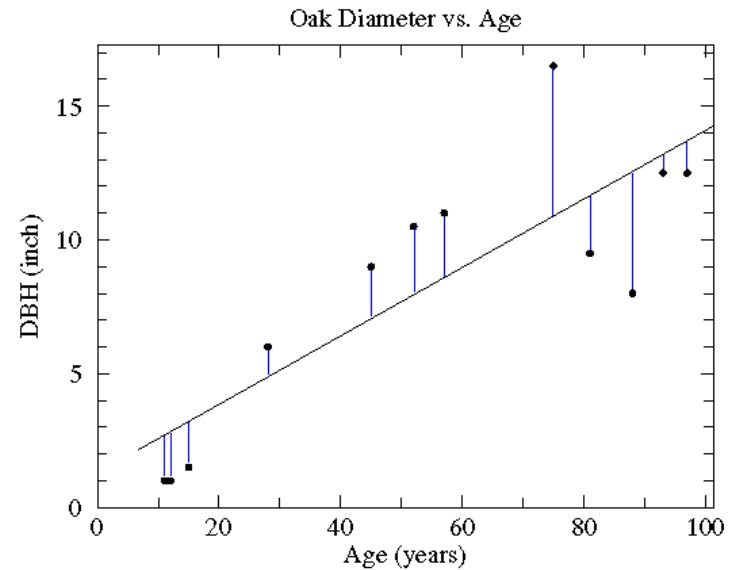
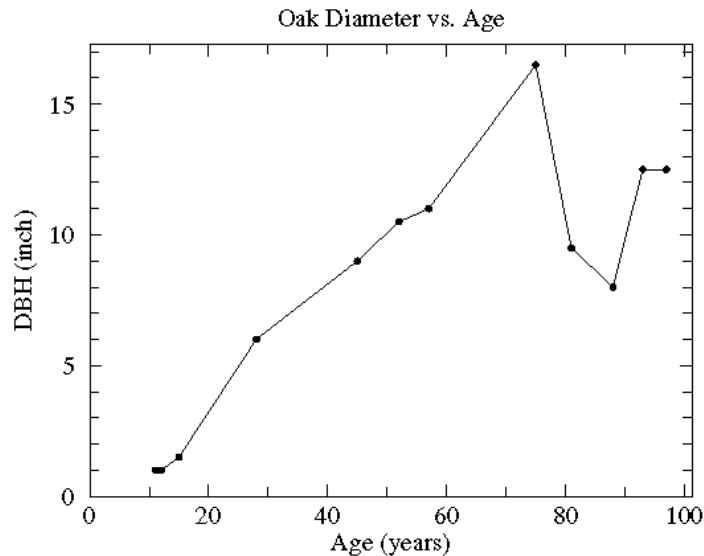
- **Basic idea**

- **Least square principle or 最小二乘原理**
- **Based the correlation between genes**

Least square regression

Y: --diameter at breast height(*DBH*) \leftrightarrow 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



Least square regression(cont.)

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

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

- Least squares estimation

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

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

is the empirical covariance between x and y ,

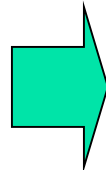
$$s_{xx} = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

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

A simple example

■ Absolute Pearson Correlation Coefficient (APCC)

	sample2	...	sample5
Gene	1.32	0.62	6.1
G_1	2.23	?	0.89
G_2	-2.15	2.21	0.23
G_3	2.01	-3.10	2.1
G_4	1.14	-0.88	4.31
...			
G_n



	sample2	...	sample5
Gene	1.32	0.62	6.1
G_1	2.23	?	0.89
G_2	-2.15	2.21	0.23
G_3	2.01	-3.10	2.1
G_4	1.14	-0.88	4.31
...			
G_n

APCC	\hat{Y}
0.0647	
0.6102	
0.9614	
	...

A simple example Gene_2

	sample2	sample2	...	sample5 0
Gene	?	1.32	0.62	6.1
G_1	3.12	2.23	?	0.89
G_2	-1.65	-2.15	2.21	0.23
G_3	0.68	2.01	-3.10	2.1
G_4	1.23	1.14	-0.88	4.31
...				
G_n	

APCC	$\hat{Y}=bX+a$	Prediction
0.0647	?	?
0.6102	?	?
0.9614	?	?
	...	

A simple example Gene_2_regression

	sample2	sample2	...	sample5 0
Gene	?	1.32	0.62	6.1
G_1	3.12	2.23	?	0.89
G_2	-1.65	-2.15	2.21	0.23
G_3	0.68	2.01	-3.10	2.1
G_4	1.23	1.14	-0.88	4.31
...				
G_n	

APCC	$\hat{Y}=bX+a$	Prediction
0.0647	$Y=-0.0883+2.6885$?
0.6102	?	?
0.9614	?	?
	...	

A simple example Gene_2-prediction

	sample2	sample2	...	sample5 0
Gene	?	1.32	0.62	6.1
G_1	3.12	2.23	?	0.89
G_2	-1.65	-2.15	2.21	0.23
G_3	0.68	2.01	-3.10	2.1
G_4	1.23	1.14	-0.88	4.31
...				
G_n	

APCC	$\hat{Y}=bX+a$	Prediction
0.0647	$Y=-0.0883+2.6885$	$(-1.65) \rightarrow 2.8343$
0.6102	?	?
0.9614	?	?
	...	

A simple example Gene_3-regression

	sample2	sample2	...	sample5 0
Gene	?	1.32	0.62	6.1
G_1	3.12	2.23	?	0.89
G_2	-1.65	-2.15	2.21	0.23
G_3	0.68	2.01	-3.10	2.1
G_4	1.23	1.14	-0.88	4.31
...				
G_n	

APCC	$\hat{Y}=bX+a$	Prediction
0.0647	$Y=-0.0883+2.6885$	$(-1.65) \rightarrow 2.8343$
0.6102	$Y=0.6114x+2.4742$?
0.9614	?	?
	...	

A simple example Gene_3-prediction

	sample2	sample2	...	sample5 0
Gene	?	1.32	0.62	6.1
G_1	3.12	2.23	?	0.89
G_2	-1.65	-2.15	2.21	0.23
G_3	0.68	2.01	-3.10	2.1
G_4	1.23	1.14	-0.88	4.31
...				
G_n	

APCC	$\hat{Y}=bX+a$	Prediction
0.0647	$Y=-0.0883+2.6885$	$(-1.65) \rightarrow -0.798$
0.6102	$Y=0.6114x+2.4742$	$0.68 \rightarrow 2.8899$
0.9614	?	?
	...	

A simple example Gene_4-regression

	sample2	sample2	...	sample5 0
Gene	?	1.32	0.62	6.1
G_1	3.12	2.23	?	0.89
G_2	-1.65	-2.15	2.21	0.23
G_3	0.68	2.01	-3.10	2.1
G_4	1.23	1.14	-0.88	4.31
...				
G_n	

APCC	$\hat{Y}=bX+a$	Prediction
0.0647	$Y=-0.0883+2.6885$	$(-1.65) \rightarrow -0.798$
0.6102	$Y=0.6114x+2.4742$	$0.68 \rightarrow 2.8899$
0.9614	$Y=1.096x+1.0104$?
	...	

A simple example Gene_4-prediction

	sample2	sample2	...	sample5 0
Gene	?	1.32	0.62	6.1
G_1	3.12	2.23	?	0.89
G_2	-1.65	-2.15	2.21	0.23
G_3	0.68	2.01	-3.10	2.1
G_4	1.23	1.14	-0.88	4.31
...				
G_n	

APCC	$\hat{Y}=bX+a$	Prediction
0.0647	$Y=-0.0883+2.6885$	$(-1.65) \rightarrow -0.798$
0.6102	$Y=0.6114x+2.4742$	$0.68 \rightarrow 2.8899$
0.9614	$Y=1.096x+1.0104$	$1.23 \rightarrow 2.3585$
	...	

Weighted average of several single prediction

- Given the APCC between y and xi, the weighting (wi) to the prediction is

weight w_i assigned to the estimate \hat{y}_i is

$$w_i = \left(\frac{r_{yx_i}^2}{1 - r_{yx_i}^2 + \epsilon} \right)^2,$$

■ Where epsilon=0.000001

APCC(r_{yxi})	$\hat{Y}=bX+a$	Prediction	Weighting (Wi)
0.0647	$Y=-0.0883+2.6885$	$(-1.65) \rightarrow -0.798$?
0.6102	$Y=0.6114x+2.4742$	$0.68 \rightarrow 2.8899$?
0.9614	$Y=1.096x+1.0104$	$1.23 \rightarrow 2.3585$?
	...		



The final result of prediction

$$\hat{y} = \frac{\sum_{i=1}^k w_i \cdot \hat{y}_i}{w_1 + w_2 + \cdots + w_k}$$

■ where $w_i = \left(\frac{r_{yx_i}^2}{1 - r_{yx_i}^2 + \epsilon} \right)^2$

■ **Authors:** The empirical suitable value of k is 10

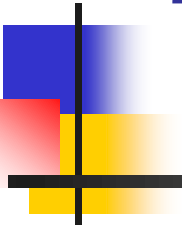


Standard for evaluating the performance

- **Root mean squared deviation (RMSD)**

$$RMSD = \sqrt{\text{mean}[(y_{\text{guess}} - y_{\text{answer}})^2]}$$

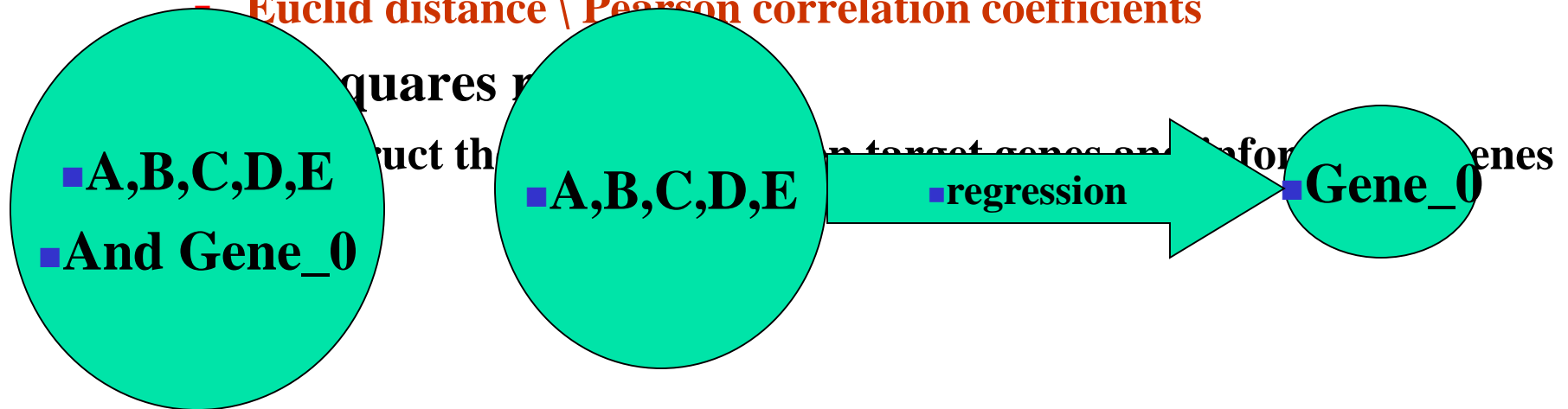
Missing value estimation for DNA microarray gene expression data: local least squares imputation ---LLSimpute



**Bioinformatics, 2005,
21(2): 187-198**

LSimpute: Rationale or Basic idea

- Using the local structure between genes or arrays
 - Local similarity structures \ correlations in genes
 - co-expression \ interaction \ similar expression pattern
 - Euclid distance \ Pearson correlation coefficients



LLSimpute: basic idea

- Given gene expression matrix (Table) with missing values
- Processing every gene (**record**) with missing values by turns

■ **One:** Select k genes(records) similar to the processing one, by some distance

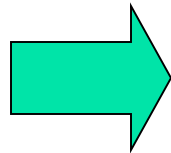
■ **Two:** Regress and estimate. Regardless of how the k genes are selected.

	sample1	sample2	...	sample100
Gene1	2.23	1.32	0.62	6.1
Gene 2	3.12	■2.50	3.12	0.89
Gene 3	-0.65	2.15	■1.09	0.23
Gene 4	■2.04	2.001	-3.10	■0.03
...	1.23	5.14	2.88	4.31
...				
Gene 10000

One simple example1

- **Sorting genes by the similarity to target (gene1)**
 - **Based on attributes (sample 2 ~sample50)**

	sample1	sample2	...	sample50
Gene1	?	1.32	0.62	6.1
Gene 2	3.12	2.23	3.12	0.89
Gene 3	-1.65	-2.15	2.21	0.23
Gene 4	0.68	2.01	-3.10	2.1
Gene 5	1.23	1.14	-0.88	4.31
...				
Gene 10000



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

$$= \begin{pmatrix} \alpha & w_1 & w_2 & w_3 & w_4 & w_5 \\ b_1 & A_{1,1} & A_{1,2} & A_{1,3} & A_{1,4} & A_{1,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b_k & A_{k,1} & A_{k,2} & A_{k,3} & A_{k,4} & A_{k,5} \end{pmatrix}$$

One simple example2

$$\begin{pmatrix} \mathbf{g}_1^T \\ \mathbf{g}_{s_1}^T \\ \vdots \\ \mathbf{g}_{s_k}^T \end{pmatrix} = \begin{pmatrix} \alpha & \mathbf{w}^T \\ \mathbf{b} & A \end{pmatrix}$$
$$= \begin{pmatrix} \alpha & w_1 & w_2 & w_3 & w_4 & w_5 \\ b_1 & A_{1,1} & A_{1,2} & A_{1,3} & A_{1,4} & A_{1,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b_k & A_{k,1} & A_{k,2} & A_{k,3} & A_{k,4} & A_{k,5} \end{pmatrix}$$

- **Top k similar genes**
 - Top left corner
- **Least squares problem**
 - Top right corner

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

value α is estimated as a linear combination of the expression values of the top k genes

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

$$\mathbf{w} \simeq x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + \cdots + x_k \mathbf{a}_k,$$

where x_1, x_2, \dots, x_k are the coefficients of the linear combination that best fits the least squares formulation (2). According to this, the value α in \mathbf{g}_1 can be estimated by

$$\alpha = \mathbf{b}^T \mathbf{x} = b_1 x_1 + b_2 x_2 + \cdots + b_k x_k.$$



Least squares regression

- **Target gene: Y**
- **Information genes: X1,X2,...,Xk.**
- **"n" is number of samples, "k" :number of variances**

$$Y \quad X_1 \quad X_2 \quad \dots \quad X_k \quad (1)$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x1_1 & x2_1 & & xk_1 \\ x1_2 & x2_2 & & xk_2 \\ \vdots & \vdots & & \vdots \\ x1_n & x2_n & & xk_n \end{pmatrix} * \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (2)$$

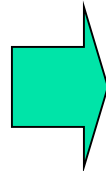
$$Y \approx a_1 \cdot X_1 + a_2 \cdot X_2 + \dots + a_k \cdot X_k \quad (3)$$

Estimating “Parameter k”

Heuristic algorithm for determining k

- Using previous formula to predict artificial missing values
- Determine k by Error (the actual value, the predicted value)

	sample2	...	sample5 0
Gene	1.32	0.62	6.1
G_1	2.23	3.12	0.89
G_2	-2.15	2.21	0.23
G_3	2.01	-3.10	2.1
G_4	1.14	-0.88	4.31
...			
G_n



	sample2	...	sample5 0
Gene	1.32	0.62	?
G_1	2.23	3.12	0.89
G_2	-2.15	2.21	0.23
G_3	2.01	-3.10	2.1
G_4	1.14	-0.88	4.31
...			
G_n

Error	Artificial missing
k=1	0.23
k=2	0.65
k=3	0.21
k=4	0.36
...	
...	
k=...	...

Global k

- In table below, there are just 3 genes having missing values

	■sample 1	■sample 2	■...	■sample 100
■Gene1	■2.23	■1.32	■0.62	■6.1
■Gene 2	■3.12	■2.50	■3.12	■0.89
■Gene 3	■-0.65	■2.15	■1.09	■0.23
■Gene 4	■1.09	■2.001	■-3.10	■0.03
■...	■1.23	■5.14	■2.88	■4.31
■...				
■Gene 10000	■...	■...	■...	■...

Error

k=1

k=2

k=3

k=4

...

...

k=...

Artificial missing on gene2	Artificial missing on gene3	Artificial missing on gene4	Sigma $\sum error$
0.23	0.62	1.2	2.05
0.65	0.15	0.23	0.93
0.23	0.84	0.31	1.38
...	



Standard for evaluating the performance

values. The performance of the missing value estimation is evaluated by normalized root mean squared error (NRMSE):

$$\text{NRMSE} = \sqrt{\text{mean}[(y_{\text{guess}} - y_{\text{ans}})^2]} / \text{std}[y_{\text{ans}}] \quad (12)$$

Research article

Open Access

Improving missing value imputation of microarray data by using spot quality weights

Peter Johansson* and Jari Häkkinen

Address: Computational Biology, Department of Theoretical Physics, Lund University, SE-223 62 Lund, Sweden

Email: Peter Johansson* - peter@thep.lu.se; Jari Häkkinen - jari@thep.lu.se

* Corresponding author

Published: 16 June 2006

Received: 14 March 2006

BMC Bioinformatics 2006, 7:306 doi:10.1186/1471-2105-7-306

Accepted: 16 June 2006

This article is available from: <http://www.biomedcentral.com/1471-2105/7/306>

© 2006 Johansson and Häkkinen; licensee BioMed Central Ltd.

■ BMC Bioinformatics 2006,7: 306



噪音数据Noisy Data

- **Noise:** 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, etc.
- **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**
 - 人工检查可疑值 (e.g., deal with possible outliers)
- **回归 Regression**
 - 回归函数拟合数据



分箱：简单的离散化方法

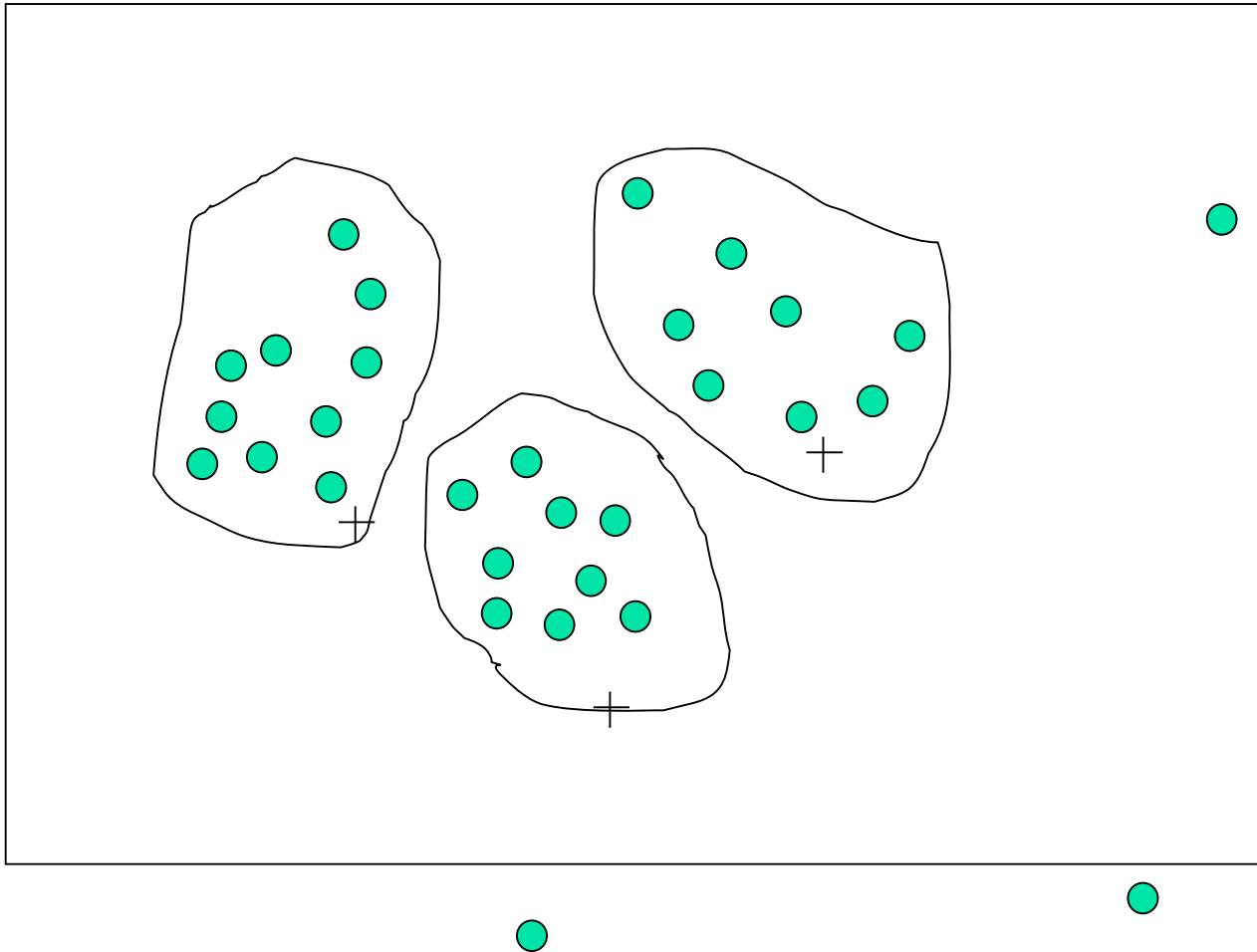
- **等宽度Equal-width (distance) 剖分:**
 - 分成大小相等的 n 个区间: 均匀网格 **uniform grid**
 - 若 A 和 B 是属性的最低和最高取值, 区间宽度为: $W = (B - A)/N$.
 - 孤立点可能占据重要影响 **may dominate presentation**
 - 倾斜的数据处理不好.
- **等频剖分 (frequency) /等深equi-depth :**
 - 分成 n 个区间, 每一个含近似相同数目的样本
 - **Good data scaling**
 - 类别属性可能会非常棘手.



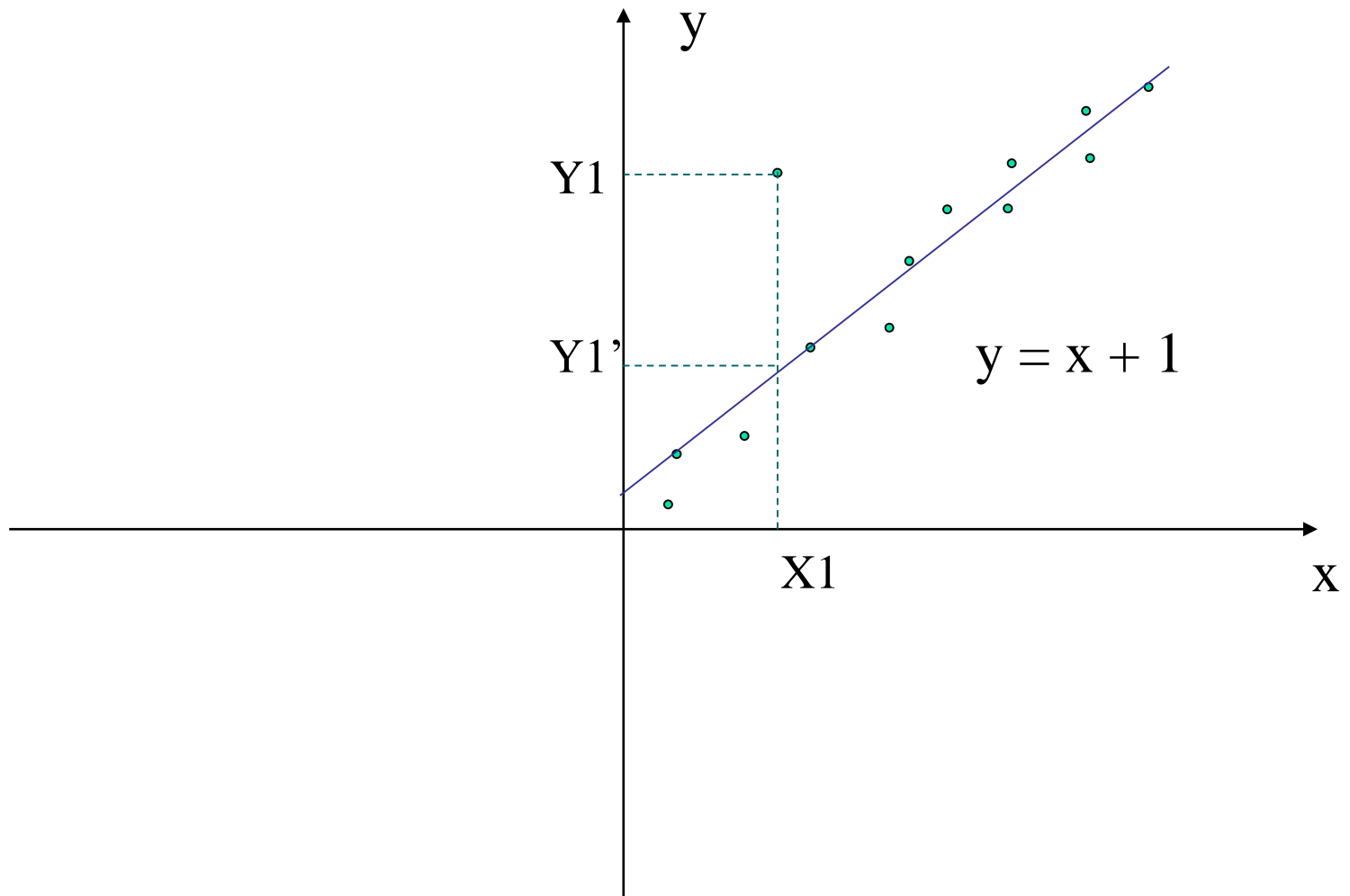
Binning Methods for Data Smoothing

- * Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- * Partition into (等频frequency / 等深equi-depth) bins:
 - **Bin 1:** 4, 8, 9, 15
 - **Bin 2:** 21, 21, 24, 25
 - **Bin 3:** 26, 28, 29, 34
- * Smoothing by bin means:
 - **Bin 1:** 9, 9, 9, 9
 - **Bin 2:** 23, 23, 23, 23
 - **Bin 3:** 29, 29, 29, 29
- * Smoothing by bin boundaries:
 - **Bin 1:** 4, 4, 4, 15
 - **Bin 2:** 21, 21, 25, 25
 - **Bin 3:** 26, 26, 26, 34

聚类分析



Regression





第2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、冗余数据、冲突数据值

■ 模式集成 **Schema integration**

- 例如., $A.cust-id \equiv B.cust-#$
- 实体识别问题 **Entity identification problem**:
 - 多个数据源的真实世界的实体的识别, e.g., Bill Clinton = William Clinton
- 集成不同来源的元数据

■ 冲突数据值的检测 and 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, e.g., 年收入
- 相关分析 *correlation analysis* / 协方差分析 *covariance analysis*
 - 可用于检测冗余数据
- 小心地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

相关分析 (数值数据)

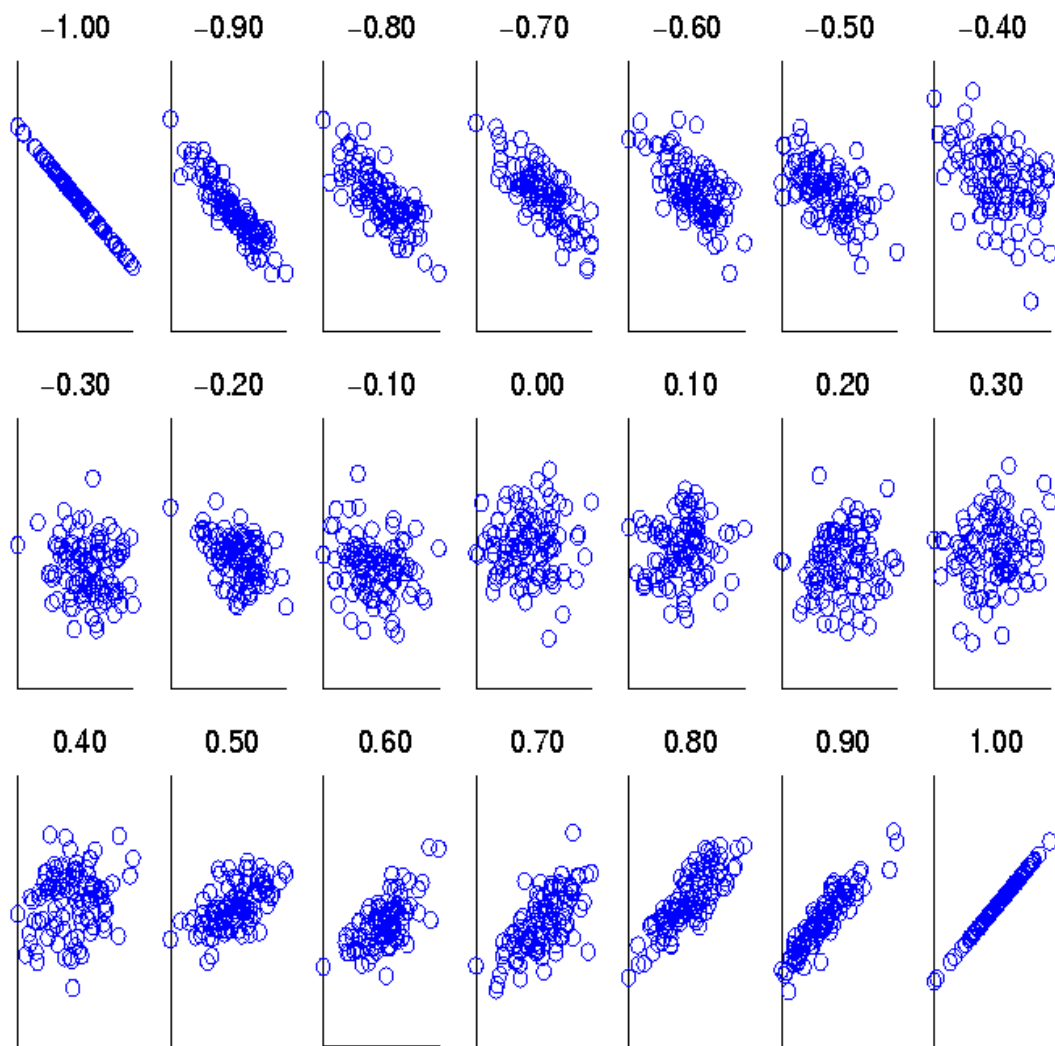
- Correlation coefficient (also called **Pearson's product moment coefficient**)
- 相关系数 (皮尔逊相关系数)

$$r_{A,B} = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum_{i=1}^n (a_i b_i) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

n元组个数, \bar{A} 和 \bar{B} 属性A和B上的平均值, σ_A and σ_B 分别为各自标准差, $\Sigma(a_i b_i)$ is the AB叉积 cross-product之和.

- If $r_{A,B} > 0$, A and B 正相关 (A's values increase as B's). 值越大相关程度越高.
- $r_{A,B} = 0$: 不相关; $r_{A,B} < 0$: 负相关

相关性的视觉评价



**Scatter plots
showing the
similarity from
-1 to 1.**



相关 (线形关系)

- 相关测量的是对象间的线性关系
- **To compute correlation, we standardize data objects, A and B, and then take their dot product**

$$a'_k = (a_k - \text{mean}(A)) / \text{std}(A)$$

$$b'_k = (b_k - \text{mean}(B)) / \text{std}(B)$$

$$\text{correlation}(A, B) = A' \bullet B'$$

协方差Covariance (Numeric Data)

- Covariance is similar to correlation

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

Correlation coefficient:

$$r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}$$

n元组个数, \bar{A} 和 \bar{B} 属性A和B上的平均值, σ_A and σ_B 分别为各自标准差.

- 正covariance: If $Cov_{A,B} > 0$, 则A 和B 同时倾向于大于期望值.
- 负covariance: If $Cov_{A,B} < 0$, 则如果 A > 其期望值, B is likely to be smaller than its expected value.
- Independence: $Cov_{A,B} = 0$ but the converse is not true:
 - Some pairs of random variables may have a covariance of 0 but are not independent. Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence



Co-Variance: An Example

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}$$

- It can be simplified in computation as

$$Cov(A, B) = E(A \cdot B) - \bar{A}\bar{B}$$

- 设两个股票 A 和 B 一周内值如下 (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- 问：如果股票是由同行业趋势的影响，它们的价格将一起上升或下降？
 - $E(A) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$
 - $E(B) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9.6$
 - $Cov(A, B) = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14) / 5 - 4 \times 9.6 = 4$
- Thus, A and B rise together since $Cov(A, B) > 0$.

相关分析 (名义数据Nominal Data)

■ X² (chi-square) test 开方检验

- Σ_{ij} --是 (a_i, b_j) 的观测频度 (实际计数)
- e_{ij} --是 (a_i, b_j) 的期望频度
- N --数据元组的个数

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(\sigma_{ij} - e_{ij})^2}{e_{ij}}$$

属性 A

		a_1	a_2	$i \rightarrow$	a_c
属性 B	b_1				
	b_2				
	$j \downarrow$				
	b_r				

$(A=a_i, B=b_j)$

$$e_{ij} = \frac{\text{count}(A = a_i) * \text{count}(B = b_j)}{N}$$

- X^2 值越大,相关的可能越大
- 对 X^2 值贡献最大的项, 其实际值与期望值相差最大的相
- 相关不意味着因果关系

Chi-Square 卡方值计算: 例子

	Play chess	Not play chess	Sum (row)
看小说	250(90)	200(360)	450
不看小说	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

$$e_{11} = \frac{\text{count}(\text{看小说}) * \text{count}(\text{下棋})}{N} = \frac{450 * 300}{1500} = 90$$

- χ^2 (chi-square) 计算(括号中的值为期望计值, 由两个类别的分布数据计算得到)

$$\chi^2 = \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} = 507.93$$

- 结果表明like_fiction 和play_chess 关联



数据变换Data Transformation

- 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Attribute/特征feature 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 min-max normalization

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

- 新数据可能“越界”

■ z-score normalization

$$v' = \frac{v - \text{均值}_A}{\text{标准差}_A}$$

■ normalization by decimal scaling

- 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)

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



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**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：**
 - **维度规约**
 - 特征子集选择**Feature subset selection**,
 - 属性产生-主成份分析**Principal Components Analysis (PCA)**
 - **数据压缩 Data Compression**
 - 基于离散小波变换的数据压缩：图像压缩
 - **数量(数值)规约—用某种表示方式替换/估计原数据**
 - **Regression and Log-Linear Models**
 - **Histograms, clustering, sampling**
 - 数据立方体聚集：聚集数据立方体结构的数据
 - **离散化和产生概念分层**



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
- 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):

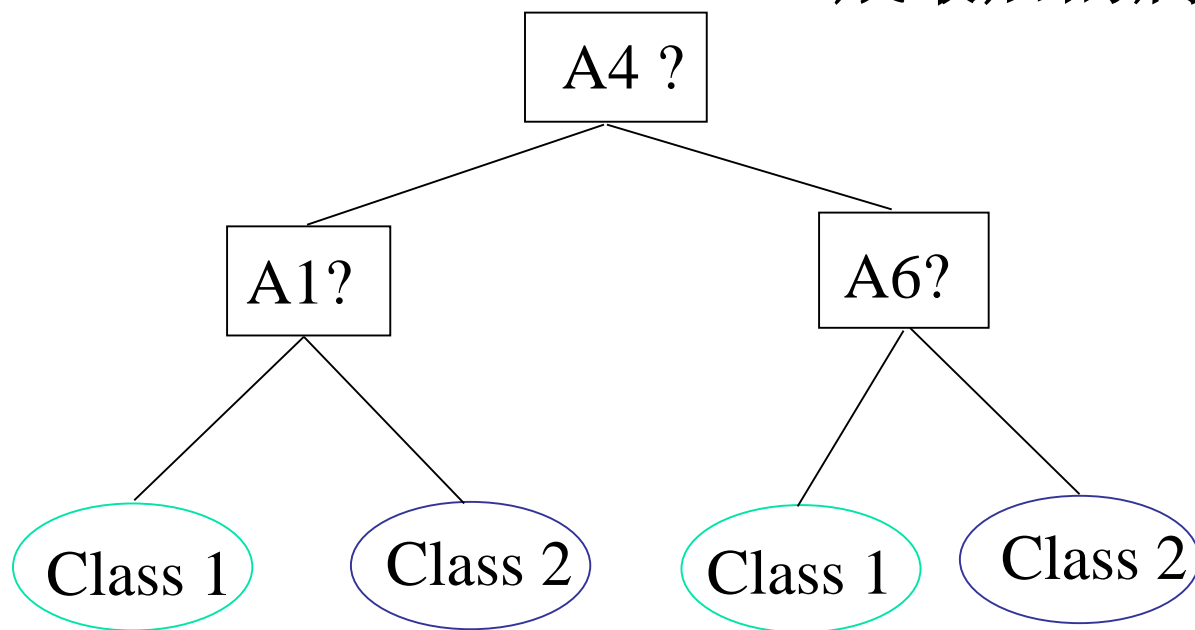
- 局部最优选择, 期望获得全局最优解
- 逐步向前选择
- 逐步向后删除 **step-wise backward elimination**
- 向前选择和向后删除结合
- 决策树归纳 **decision-tree induction**

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-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**
 - 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
 - **PCA**



维度规约-主成分分析

- **principal component analysis(PCA)**也称主分量分析,由霍特林(Hotelling)于1933年提出, 利用降维的思想, 在损失很少信息的前提下, 把多个指标转化为几个综合指标的多元统计方法
 - Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 417-441, and 498-520.
- 通常把转化生成的综合指标称为主成分, 每个成分都是原始变量(指标)的线性组合, 并且主成分之间互不相关, 使得主成分比原始变量在某些方面具有更优越的性能



数量(数值)规约

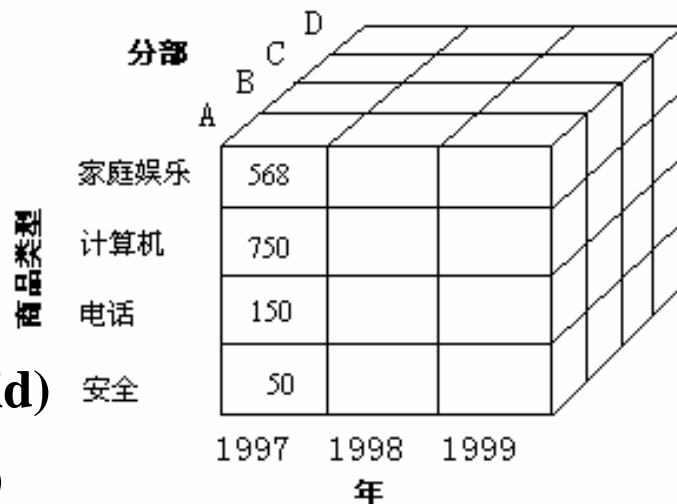
- 用替代的,较小的数据表示形式来替换原始数据
- 参数方法:使用模型估计数据,存放模型参数+离群点.
 - 线性回归: 数据拟合到一条直线上; 多元线性回归; 对数线性模型-近似离散的多维概率分布
- 非参数方法
 - 直方图; 聚类; 抽样; 数据立方体聚集

数据立方体聚集

数据立方体存储多维聚集信息

- 某抽象层上建的数据立方体称为方体(cuboid)
- 最底层建的方体称为基本方体(base cuboid)
- 最高层的立方体称为 顶点方体(apex cuboid)

每个更高层的抽象将减少数据的规模



年=1999			
年=1998			
年=1997			
季度	销售额		
Q1	\$224, 000		
Q2	\$408, 000		
Q3	\$350, 000		
Q4	\$586, 000		

→

年	销售额
1997	\$1, 568, 000
1998	\$2, 356, 000
1999	\$3, 594, 000

使用合适的抽象层上的数据

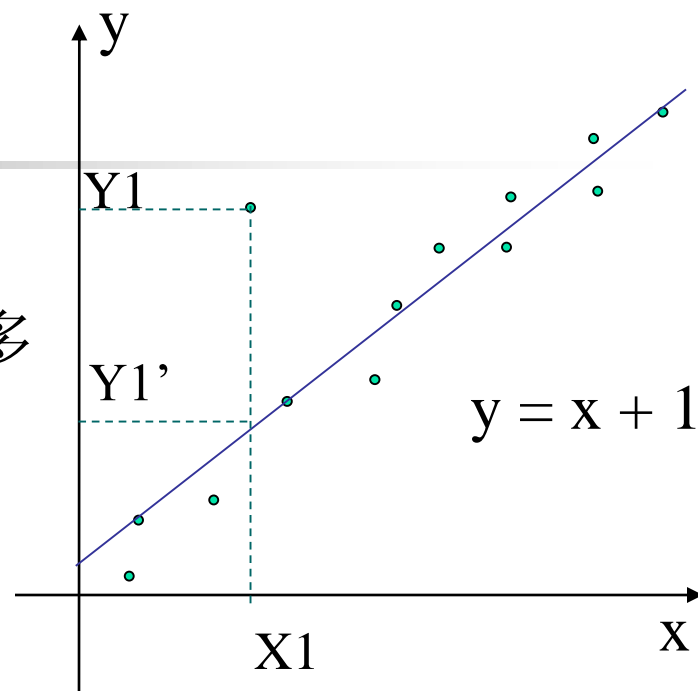
- 对数据立方体聚集得到与任务相关的最小立方体

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

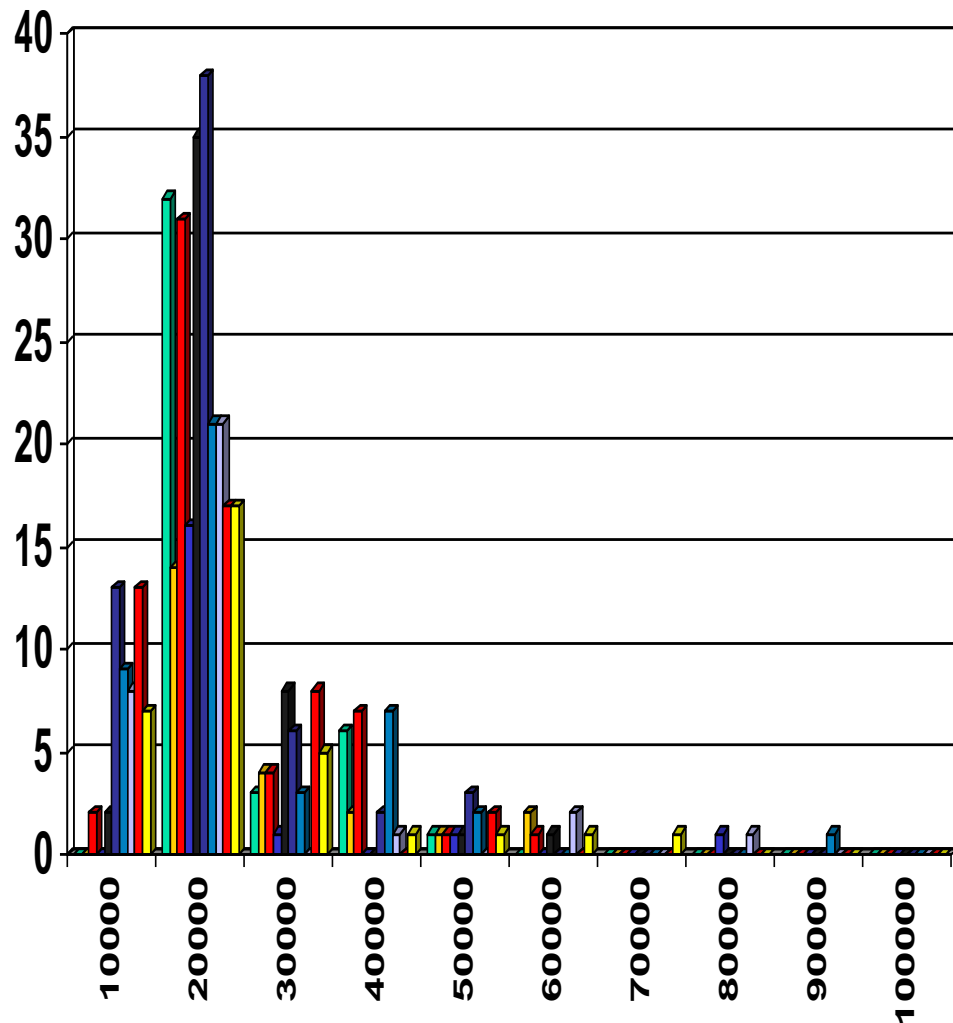
- 绝大多数情况“最好的拟合”是由最小二乘法(*least squares method*)实现, 其他的方法也有



- 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

直方图Histograms

- 使用分箱来近似数据分布.
- 对于近似稀疏和稠密数据, 以及高倾斜和均匀数据, 直方图非常有效.
- 多维直方图能有效地近似多达5个属性的数据.
 - 更高维的有效性待验证





聚类Clustering

- 将对象划分成集/簇, 用簇的表示替换实际数据
- 该技术的有效性依赖于数据的质量



抽样Sampling

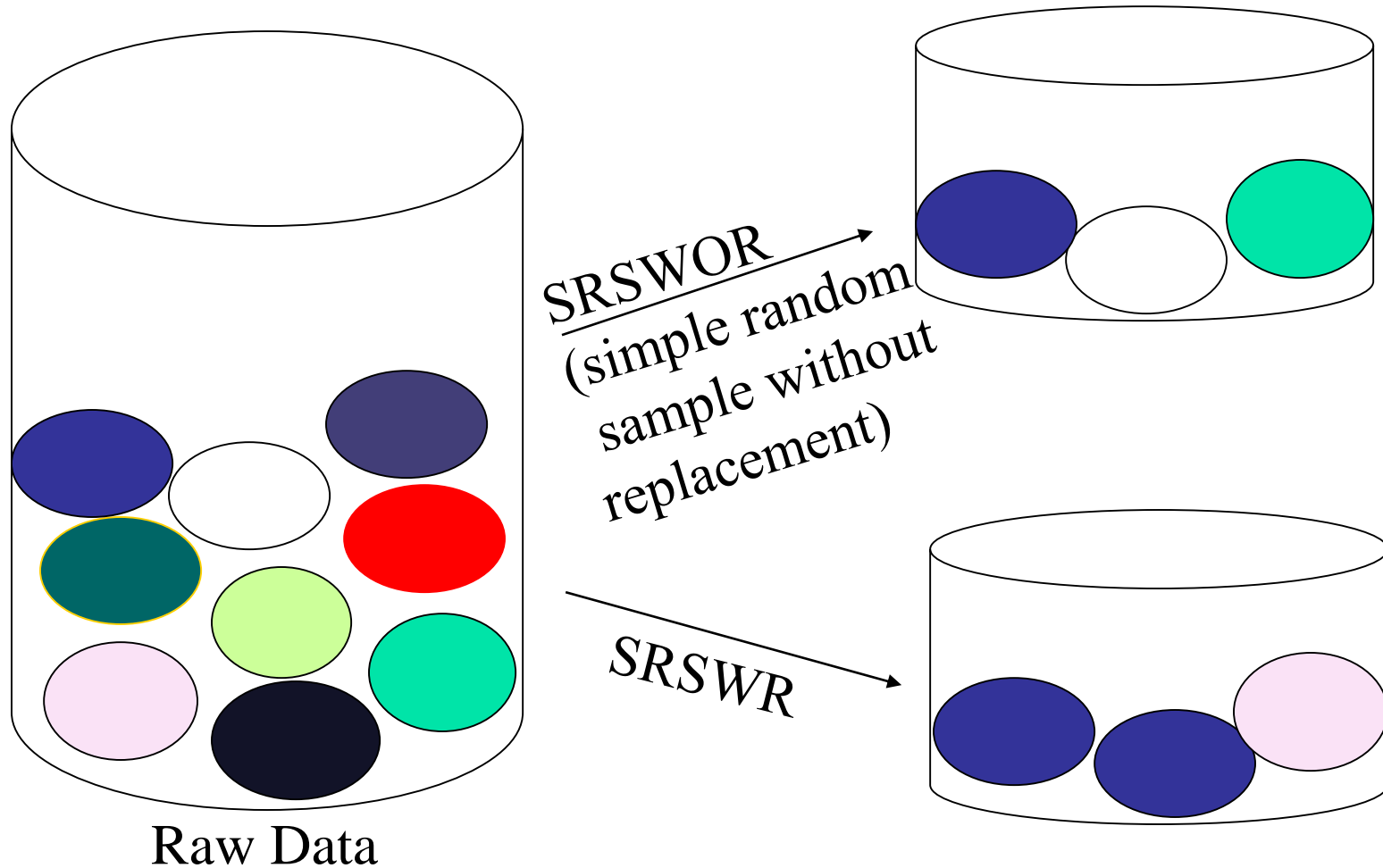
- 抽样: 获得一个小的样本集 s 来表示整个数据集 N
- 允许一个挖掘算法运行复杂度子线性于样本大小
- 关键原则: 选择一个有代表性的数据子集
 - 数据偏斜时简单随机抽样的性能很差
 - 发展适应抽样方法: 分层抽样
- **Note: Sampling may not reduce database I/Os (page at a time)**



抽样类型 Types of Sampling

- 无放回抽样 Sampling without replacement
 - **Once an object is selected, it is removed from the population**
- 有放回抽样 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





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和概念分层

■ 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)
 - 自顶向下: 由一个/几个点开始递归划分整个属性区间

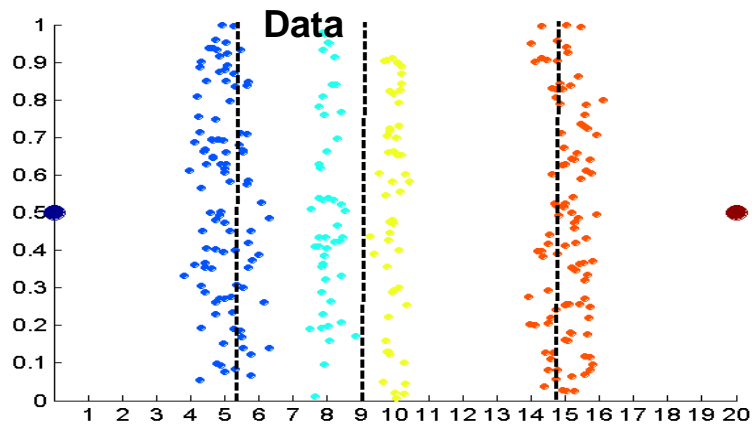
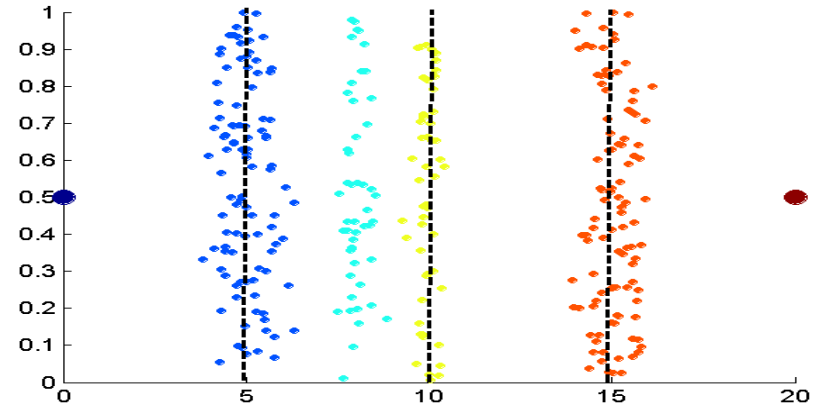
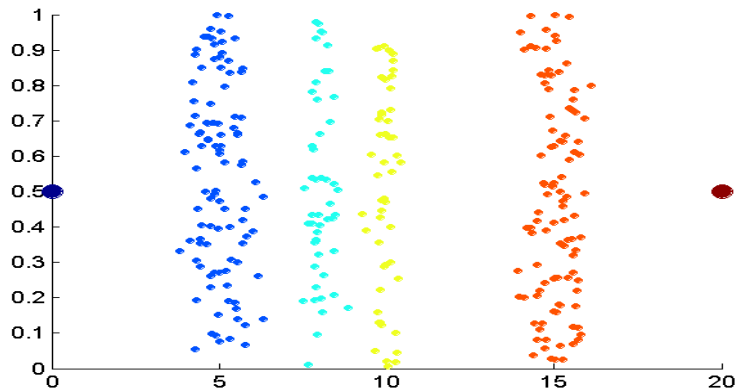
■ 递归离散化属性, 产生属性值分层/多分辨率划分: 概念分层



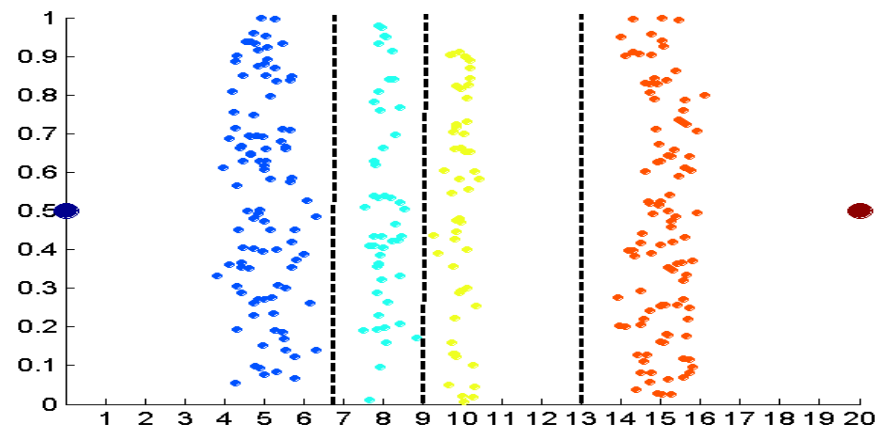
数值数据离散化

- 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. 选择的阈值 T 使其后划分得到的信息增益最大

■ 具有最小期望信息需求的点
选为 A 的分裂点

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

其中， S_1 和 S_2 分别对应于 S 中满足条件 $A < T$ 和 $A \geq T$ 的样本。

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

其中， p_i 是类 i 在 S_1 中的概率，

等于 S_1 中类 i 的样本数除以 S_1 中的样本总数。

3. 直到满足某个终止条件 $Ent(S) - I(S, T) > \delta$



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**

