80206085244011（40/2）

Pattern Recognition & Machine Learning

pre-processing the data

Yong Wang

University of Chinese Academy of Sciences

2018.03.25

数据预处理

跛足而不迷路能赶过虽健步如飞但误入歧途的人。

—— Francis Bacon

英国近代唯物主义哲学家、思想家和科学家 (1561 – 1626)

决定结果的正是我们在处理表面上微不足道、枯燥乏味、而且不胜麻烦的细枝末节时所采用的谨慎小心的态度。

—— Theobald Smith

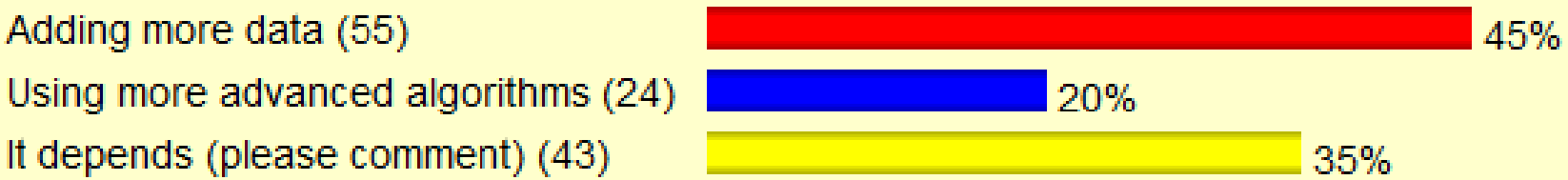
美国流行病学家 和病理学家(1859 – 1934)

网络调查1

KDnuggets : **Polls** : **More Data or Better Algorithm? (Apr 2008)**

Poll

What will usually give better improvement in data mining results: [122 votes total]



Comments

Ed Freeman, More Data or Better Algorithms?
What really helps:
1) Having a good problem to work on
2) Having a good approach to that problem
3) Having decent data. Quality is much more important than quantity
4) Handling the data well -- good transforms, good missing value handling, making sure that all approaches make sense for the problem.

数据分类

结构化数据

存储在数据库里，可以用二维表结构来逻辑表达实现的数据叫结构化数据。数据结构字段含义确定清晰。

非结构化数据

不方便用数据库二维逻辑表来表现的数据即称为非结构化数据。数据杂乱无章,很难按照一个概念去进行抽取,例如所有格式的办公文档、图片、音\视频等。

半结构化数据

半结构化数据模型是一种基于图的自描述的对象实例模型。具有一定结构,但语义不够确定，字段可根据需要扩充，例如XML、HTML网页。

典型示例 1

Attributes

Feature

Field

nominal

numeric

结构化数据（Table）

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

Record

Sample

典型示例 2

半结构化数据

```
- <rixml:XMLData xmlns:command="http://xmlns.xfy.com/command" xmlns:function="http://xmlns.xfy.com/function"
    xmlns:instruction="http://xmlns.xfy.com/instruction" xmlns:form="http://xmlns.xfy.com/form" xmlns:ctrl="http://xmlns.xfy.com/controls"
    xmlns="http://www.w3.org/1999/xhtml" xmlns:xhtml="http://www.w3.org/1999/xhtml" xmlns:tab="http://xmlns.xfy.com/tab"
    xmlns:frames="http://xmlns.xfy.com/frames" xmlns:event="http://xmlns.xfy.com/event" xmlns:rixml="http://www.rixml.org/2005/3/RIXML">
  - <xbrli:xbrl xmlns:xbrli="http://www.xbrl.org/2003/instance" xmlns:xbrll_2="http://www.xbrl.org/2003/linkbase"
      xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:UTX="http://investors.utc.com/sec/" xmlns:usfr-
      seccert="http://www.xbrl.org/us/fr/rpt/seccert/2005-02-28" xmlns:usfr-mda="http://www.xbrl.org/us/fr/rpt/mda/2005-02-28" xmlns:usfr-
      ar="http://www.xbrl.org/us/fr/rpt/ar/2005-02-28" xmlns:usfr-pte="http://www.xbrl.org/us/fr/common/pte/2005-02-28" xmlns:usfr-
      ptr="http://www.xbrl.org/us/fr/common/ptr/2005-02-28" xmlns:iso4217="http://www.xbrl.org/2003/iso4217" xmlns:us-gaap-
      ci="http://www.xbrl.org/us/fr/gaap/ci/2005-02-28">
      <xbrll_2:schemaRef xlink:href="utx-20061231.xsd" xlink:arcrole="http://www.w3.org/1999/xlink/properties/linkbase" xlink:type="simple" />
    - <xbrli:context id="QTD_Mar2006">
      - <xbrli:entity>
          <xbrli:identifier scheme="http://www.sec.gov">0000101829</xbrli:identifier>
        </xbrli:entity>
      - <xbrli:period>
          <xbrli:startDate>2006-01-01</xbrli:startDate>
          <xbrli:endDate>2006-03-31</xbrli:endDate>
        </xbrli:period>
      </xbrli:context>
    - <xbrli:context id="As_of_Dec31_2005">
      - <xbrli:entity>
          <xbrli:identifier scheme="http://www.se
        </xbrli:entity>
      - <xbrli:period>
          <xbrli:instant>2005-12-31</xbrli:instant>
        </xbrli:period>
      </xbrli:context>
    - <xbrli:context id="As_of_Sep30_2006">
      - <xbrli:entity>
          <xbrli:identifier scheme="http://www.se
        </xbrli:entity>
      - <xbrli:period>
          <xbrli:instant>2006-09-30</xbrli:instant>
        </xbrli:period>
      </xbrli:context>
    - <xbrli:context id="QTD_Sep2006">
      - <xbrli:entity>
          <xbrli:identifier scheme="http://www.se
        </xbrli:entity>
      - <xbrli:period>
```

公司信息

证券代码： 002002 查询

证券代码:002002 证券简称: 江苏琼花 起始年度： 2005 报告类型： 年度报告

公司主要指标

项目 / 年度	2008年度	2007年度	2006年度	2005年度	2004年度
投资与收益					
每股收益(元)	-	-0.207600	0.111100	0.219000	0.218300
每股净资产(元)	-	2.1558	3.3380	3.4270	3.5500
净资产收益率(%)	-	-9.630	3.330	6.390	6.150
扣除非经常性损益后的净利润(元)	-	-24,826,735	8,985,207	19,934,139	20,936,769
偿债能力					
流动比率(倍)	-	0.818	1.146	1.658	2.673
速动比率(倍)	-	0.571	0.867	1.443	2.401
应收帐款周转率(次)	-	9.037	8.155	8.121	29.087
资产负债比率(%)	-	50.172	40.607	38.998	29.049
盈利能力					

典型示例 3

非结构化数据

Offset	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
00000000	47	49	46	38	39	61	A3	00	CA	00	87	E2	00	07	07	0C	GIF89a£.Ê.‡â....
00000016	9F	83	45	8A	94	9A	83	2F	13	BB	89	48	3C	3F	4A	C7	ŸƒEŠ”šƒ/.»‰H<?JÇ
00000032	8A	A4	82	41	5D	46	1D	0C	CB	C3	AC	B9	8F	78	BF	B0	Š¤‚A]F..ËÃ¬¹x¿°
00000048	79	CE	CC	CF	85	59	1F	4C	3D	2C	C3	6B	57	80	82	7B	yÎÌÏ…Y.L=,ÃkW€‚{
00000064	AE	5A	35	94	67	7B	E2	90	72	5D	5A	57	E5	CF	A9	29	®Z5”g{â‐r]ZWåϩ)
00000080	26	28	9A	86	75	83	68	58	A5	AE	A7	83	59	3E	B8	94	&(š†uƒhX¥®§ƒY>¸”
00000096	85	A5	6A	5A	AF	A0	79	46	35	30	68	41	25	D8	E2	DE	…¥jZ¯ yF50hA%Øâ Þ
00000112	E2	B8	78	0E	10	28	8C	6E	3C	C5	AC	9F	E5	D1	C0	62	â¸x..(Œn<Å¬Ÿåѽb
00000128	66	6F	97	8B	89	E7	AC	9E	C2	A3	51	83	47	34	BB	A2	fo—‹‰ç¬žÂ£Qƒ G4»¢
00000144	77	66	4C	3D	25	10	0C	81	6B	6A	5E	3F	4B	A5	94	9C	wfL=%..kj^?K¥”œ
00000160	64	12	03	C8	C6	C1	EF	E8	DC	D8	88	60	BF	B1	8F	AF	d..ÈÆÁïèÜØˆ`¿±¯
00000176	BD	BF	C7	AB	B8	9E	5C	3C	71	5F	53	27	28	32	BB	A2	½¿Ç«¸ž\<q_S'(2»¢
00000192	8E	E8	DD	C4	C2	88	5D	AA	89	74	B1	72	3E	6A	56	43	Žè݄Âˆ]ª‰t±r>jVC

典型示例 4

一份商品说明信息

- 结构化数据
  - 将它存储到数据库的各个字段（名称、条形码等）中
- 半结构化数据
  - 用XML文档来存储
- 非结构化数据
  - 将它存储成一个普通的文本文件或者二进制文件

非结构化、半结构化数据和结构化数据的区别是模式（schema）对数据的约束程度不同。结构化是“强约束”的，半结构化是“弱约束”的，非结构化是“无约束”的。

典型示例 5

从非结构化到半结构化，从半结构化到结构化，从结构化到关联数据体系，从关联数据体系到数据挖掘，从数据挖掘到故事化呈现，从故事化呈现到决策导向。

数据的基本统计描述

For instance is no proof

把握数据的全貌是成功的数据预处理的关键

基本统计描述可以识别数据性质，凸显哪些数据值应该视为噪声或离群点

中心趋势度量	度量数据分布的中部或中心位置
数据散布度量	度量数值数据散布或发散的情况
图形显示	使用可视化技术审视数据

数据的基本统计描述

- 中心趋势度量
  - 均值（mean）
    - 算术平均 $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$
    - 加权平均 $\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$
    - 截尾平均　均值对极端值（例如，离群点）很敏感
  - 中位数（median）

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

倾斜（非对称）数据的数据中心的更好度量是中位数

当观测的数量很大时，中位数的计算开销很大

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

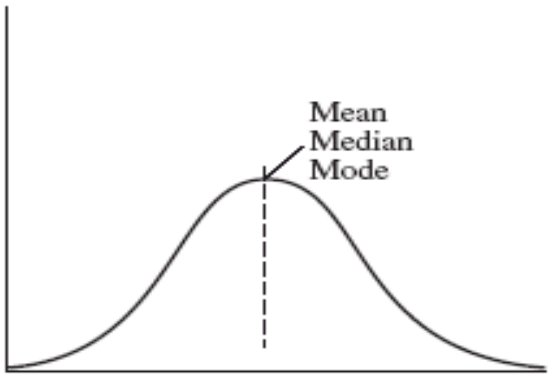
数据的基本统计描述

- 中心趋势度量
  - 众数（mode）
    - 数据集（定性或定量）的众数是集合中出现最频繁（最高频率）的值
    - 单峰（unimodal）、双峰（bimodal）、三峰（trimodal）。。。
    - 对于适度倾斜（非对称）的单峰数值数据存在经验近似等式

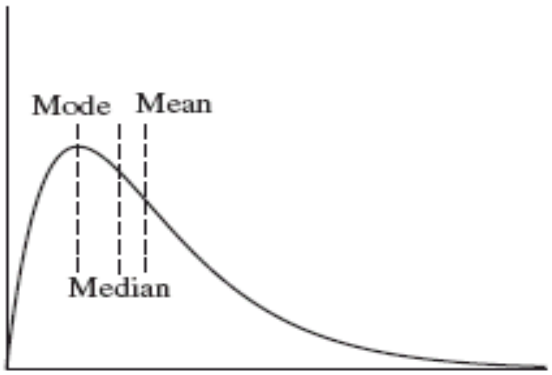
$$mean - mode \approx 3 \times (mean - median)$$

  - 中列数（midrange）

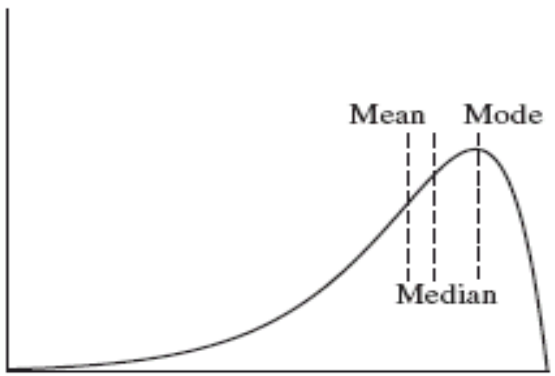
$$midrange = \frac{max(\ ) + min(\ )}{2}$$



(a) symmetric data



(b) positively skewed data

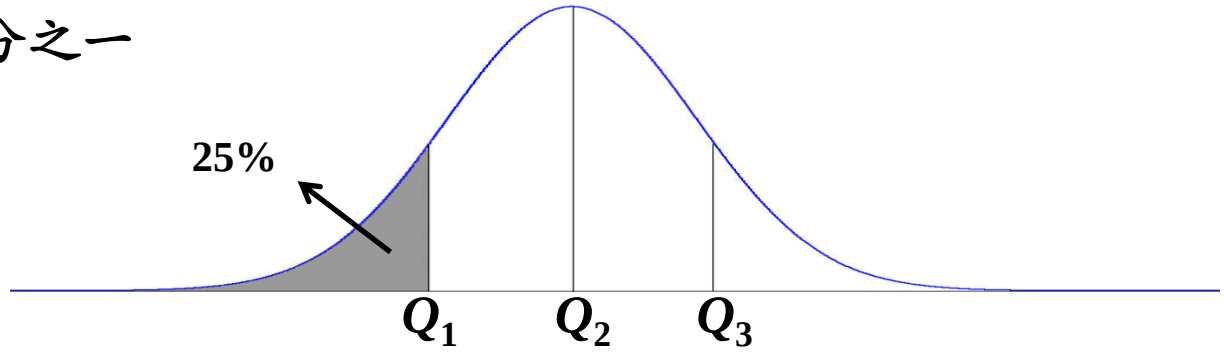


(c) negatively skewed data

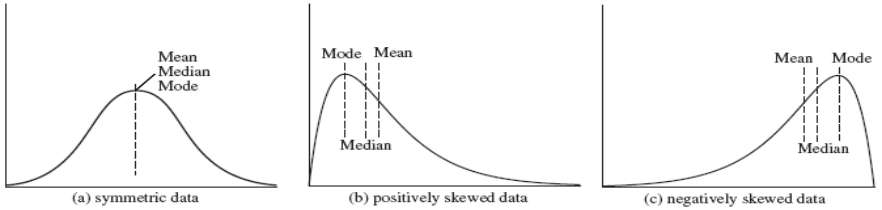
数据的基本统计描述

- 数据散布度量
  - 极差（range）
    - 数值属性上的观测值的最大值（max()）与最小值（min()）之差
  - 分位数（quantile）
    - 按照分布每隔一定间隔把数据划分成基本上大小相等的连贯集合
  - 四分位数（quartile） 百分位数（percentile）
    - 把数据分布划分为 4 个相等的部分，使得每部分表示数据分布的四分之一
  - 四分位数极差 （*IQR*）

$$IQR = Q_3 - Q_1$$



数据的基本统计描述



- 数据散布度量
  - 五数概括（five-number summary）
    - 描述倾斜分布，单个分布数值度量（例如，*IQR*）不是非常有用的。倾斜分布两边的分布是不等的。
    - 五数概括即用五个数来概括数据：最小值、第1四分位数(Q_1)、中位数(Q_2)、第3四分位数(Q_3)、最大值。依次写为*Minimum*，Q_1，*Median*，Q_3，*Maximum*。

在五数中相邻的每两个数之间大约有1／4或25％的数据项

例题

对12个月薪数据的样本，按照递增顺序排列如下：

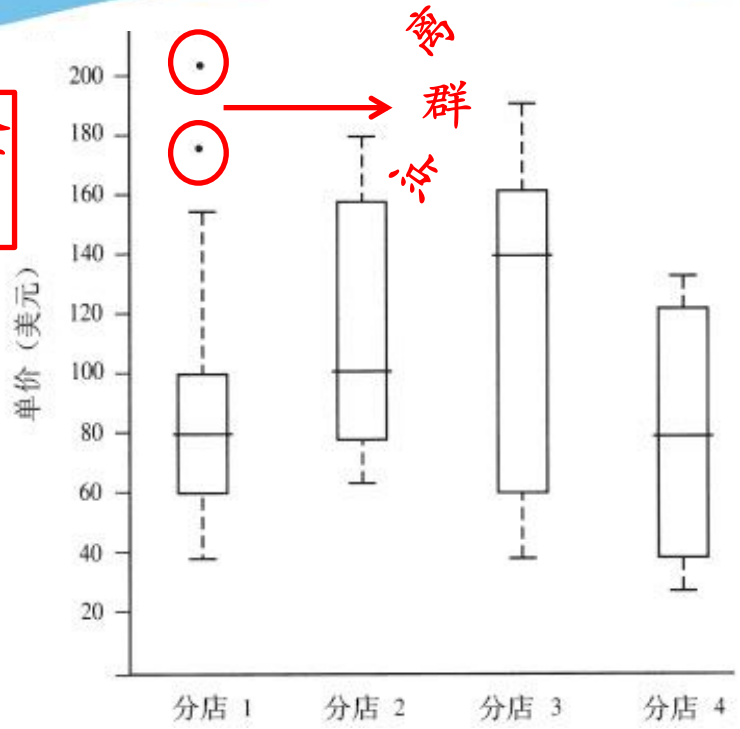
2210 2255 2350 | 2380 2380 2390 | 2420 2440 2450 | 2550 2630 2825

Min $Q_1 = 2365$ $Q_2 = 2405$ $Q_3 = 2500$ *Max*

数据的基本统计描述

- 数据散布度量

针对大数据绘制盒图依然是个挑战



  - 盒图（boxplot）

John W. Tukey
（1915-2000）

    - 以可视化的方式展现五数概括
    - 盒的端点在四分位数上使得盒的长度是四分位数极差*IQR*
    - 中位数用盒内的线标记
    - 盒外的两条线（称作胡须）延伸到最小和最大观测值
    - 当数据值超过四分位数达到1.5×*IQR*时，胡须出现在四分位数的1.5×*IQR*之内的最极端观测值处终止。剩下的情况个别地绘出。

  - 方差和标准差（variance，standard deviation）

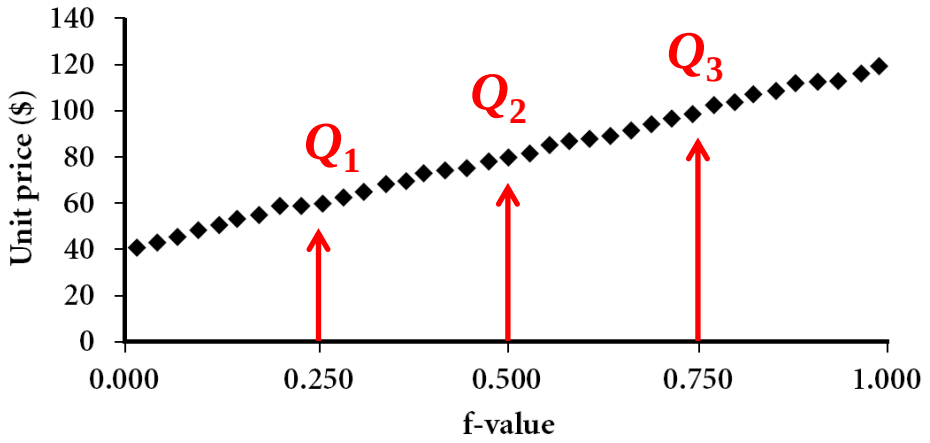
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2\right] = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2$$

数据的基本统计描述

- 数据基本统计描述的图形显示
  - 分位数图（quantile plot）
    - 每个观测值 x_i（序列或数值属性）与一个百分数 f_i 配对，指出大约 $f_i \times 100\%$ 的数据小于值 x_i

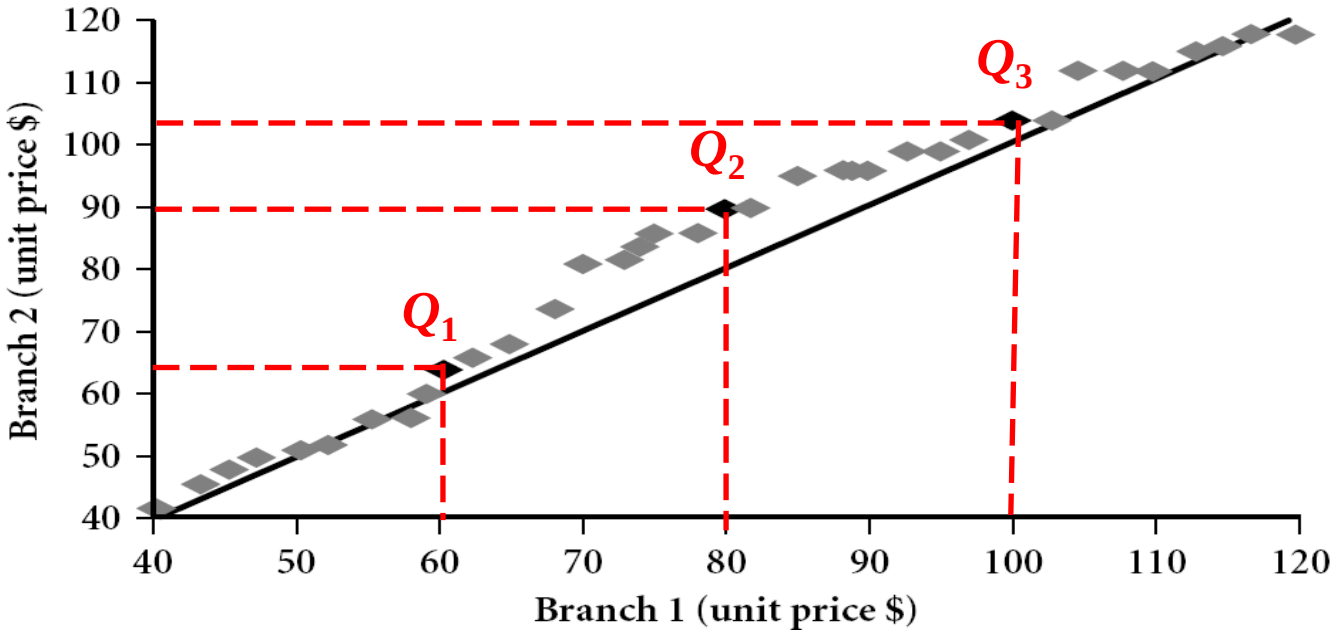
单价（美元）	商品销售量
40	275
43	300
47	250
…	…
74	360
75	515
78	540
…	…
115	320
117	270
120	350

$$f_i = \frac{i - 0.5}{N}$$



数据的基本统计描述

- 数据基本统计描述的图形显示
  - 分位数-分位数图（quantile-quantile plot，Q-Q图）
    - 直观验证某两组数据（样本个数相等）是否来自同一（族）分布，或观察对比从一个分布到另一个分布是否有漂移

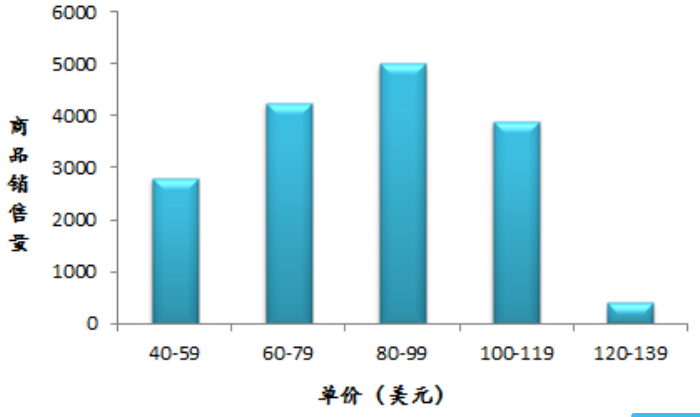


数据的基本统计描述

- 数据基本统计描述的图形显示
  - 直方图 (histogram) 或频率直方图 (frequency histogram)

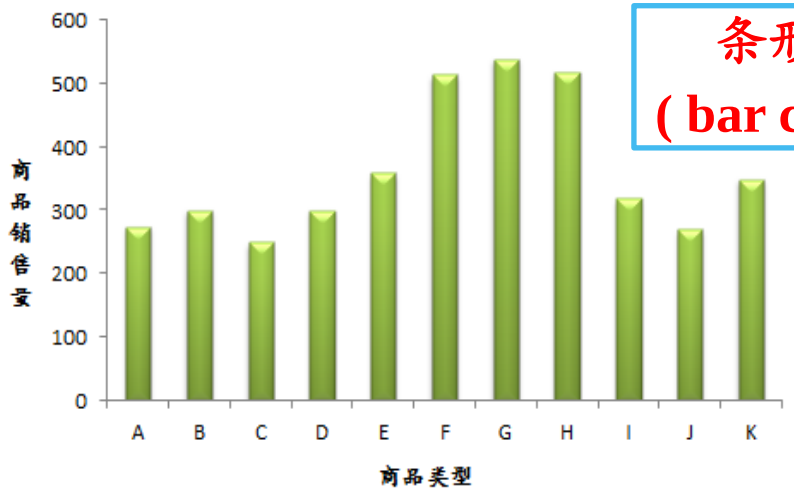
数值属性

单价（美元）	商品销售量
40	275
43	300
47	250
…	…
74	360
75	515
78	540
…	…
115	320
117	270
120	350



标称属性

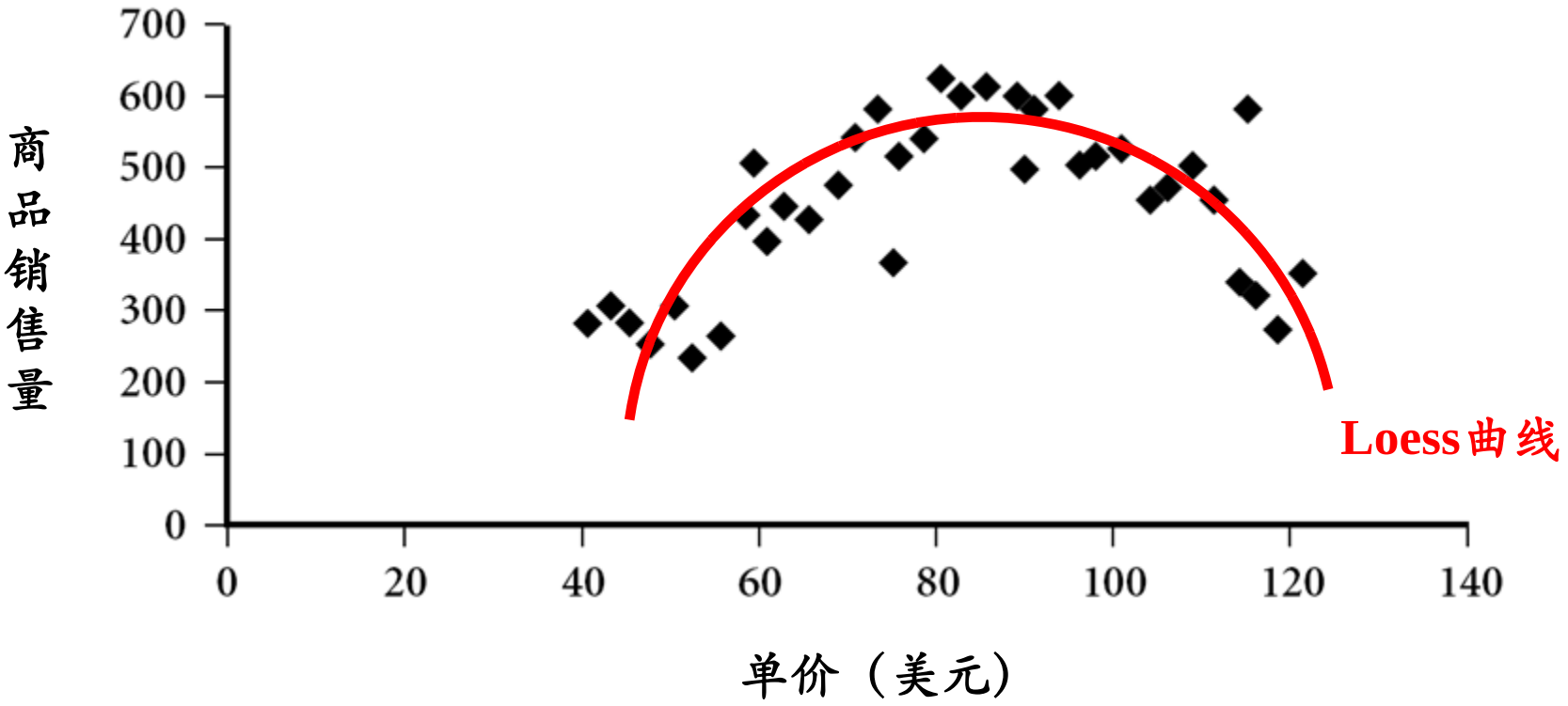
商品类型	商品销售量
A	275
B	300
C	250
D	300
E	360
F	515
G	540
H	520
I	320
J	270
K	350



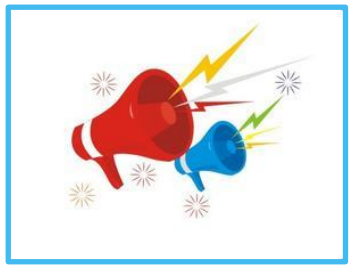
条形图 (bar chart)

数据的基本统计描述

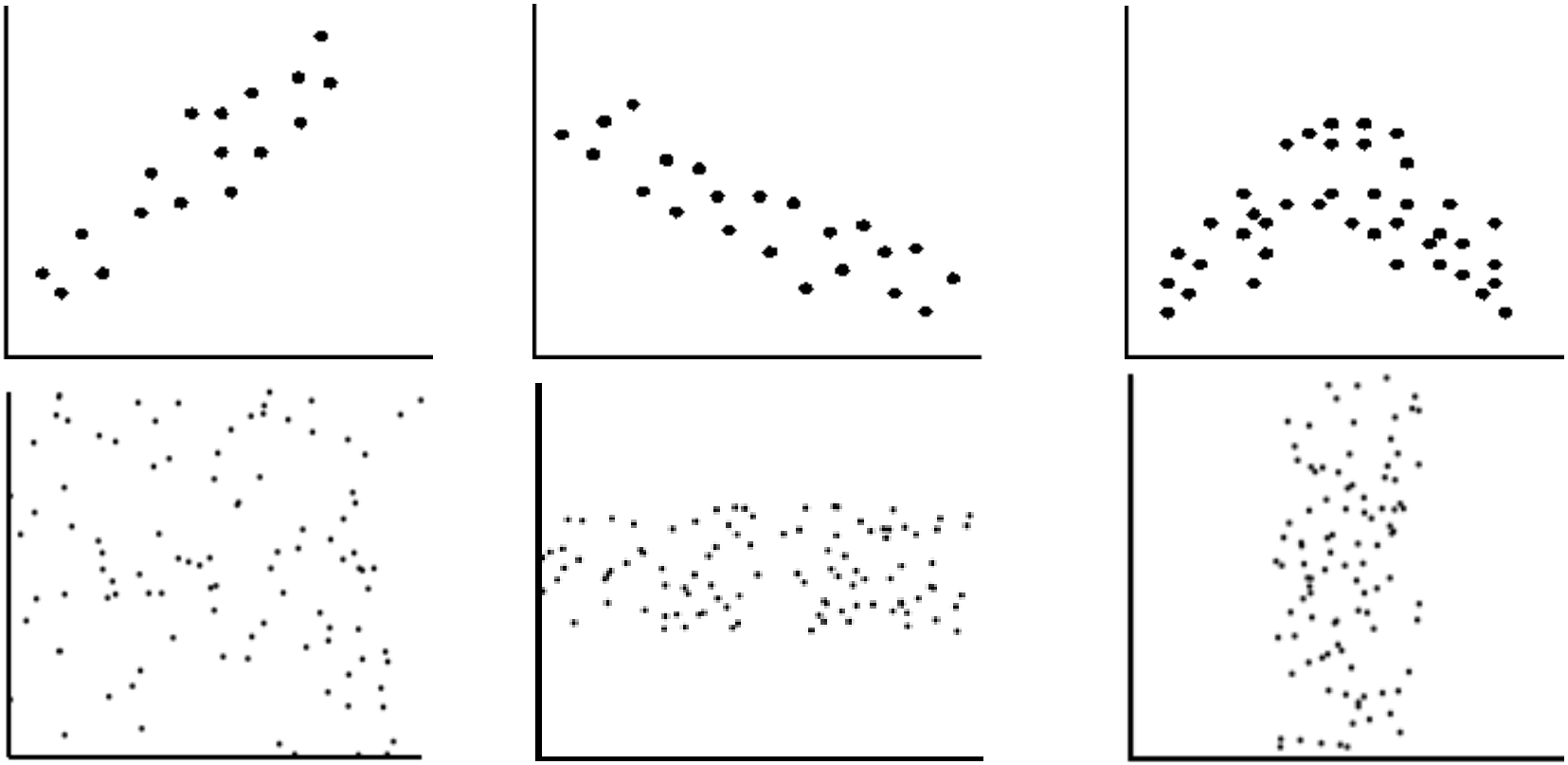
- 数据基本统计描述的图形显示
  - 散点图（scatter plot）
    - 直观确定两个数值变量之间是否存在联系、模式或趋势



数据的基本统计描述



- 数据基本统计描述的图形显示
  - 散点图（scatter plot）
    - 直观确定两个数值变量之间是否存在联系、模式或趋势



数据预处理原因 1

- **Data in the real world is dirty**
  - **incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data**
    - **e.g., occupation= “”**
  - **noisy: containing errors or outliers**
    - **e.g., Salary= “-10”**
  - **inconsistent: containing discrepancies in codes or names**
    - **e.g., age= “42” birthday= “03/07/1997”**
    - **e.g., was rating “1, 2, 3”, now rating “A, B, C”**
    - **e.g., discrepancy between duplicate records**

      name= “haha” income= “42$”

      name= “haha” income= “24$”

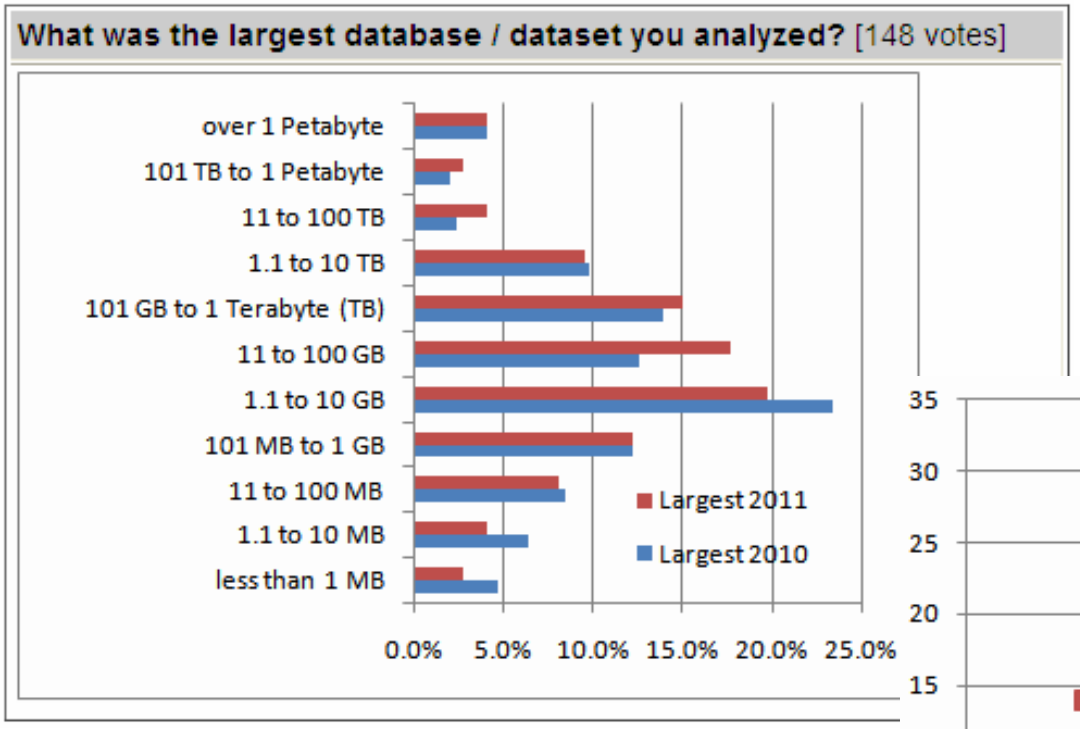
数据预处理原因 2

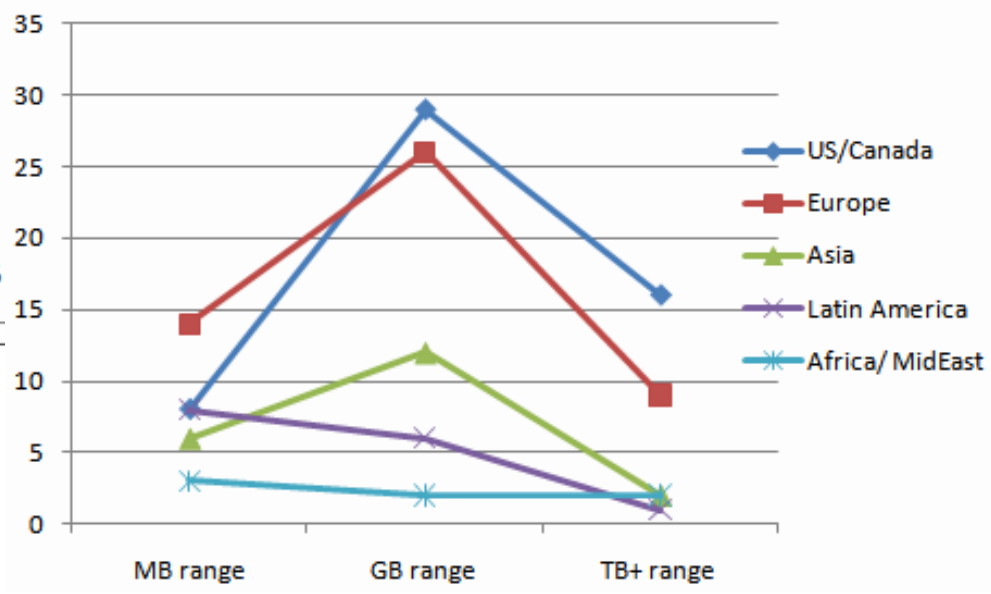
- 学习算法/参数选择
  - 合适的选择是由数据决定的（没有免费的午餐）
  - 要避免“过拟合”和“欠拟合”
    - 测试数据使用与训练数据集分离的大数据集（数据源充足）
    - 交叉验证（数据源不足）

- 机器学习从来就不是提取一个数据集，将学习算法应用于数据上那么简单的事情。每个问题都不相同。必须对数据进行思考，琢磨它的意义，然后从不同的角度来检验，具有创造性地找到一个合适的观点

数据预处理原因 3

Largest dataset analysed / data mined (May 2011)





数据缺失值的处理

- 忽略元组（记录、行、对象，tuple）

  会有信息损失

- 人工填写缺失值

  费时，工作量巨大

- 使用一个全局常量填充缺失值

  会产生类似“Unknown”的新属性概念，方法简单，但不可靠

- 使用属性的中心度量（如均值或中位数）填充缺失值

  对称数据使用均值，倾斜数据使用中位数，全体数据参与计算

- 使用与给定元组属同一类的所有样本的属性均值或中位数填充缺失值

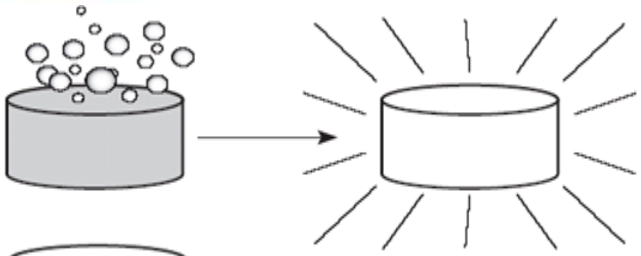
  对称数据使用均值，倾斜数据使用中位数，同类数据参与计算

- 使用最可能的值填充缺失值（回归方法、贝叶斯方法、决策树方法等）

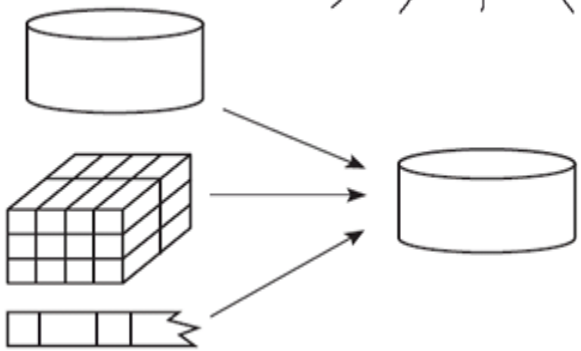
  使用已有数据的大部分信息来预测缺失值，方法最流行，但费时间

数据预处理基本功能

- 数据清理



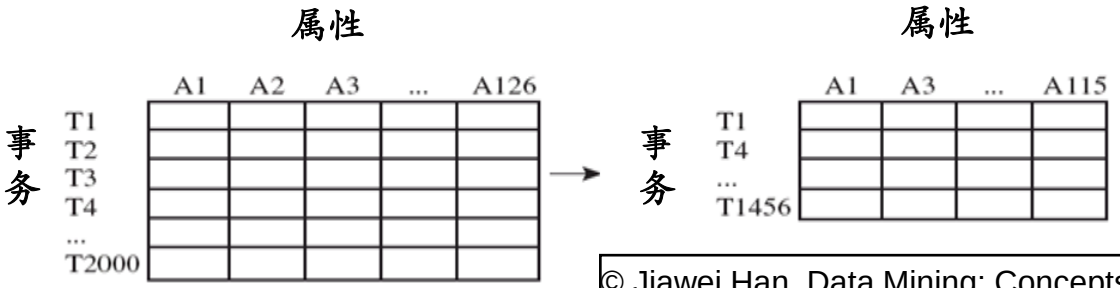
- 数据集成



- 数据变换

$-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

- 数据归约



数据预处理基本功能

- **数据清理**

  去除源数据集中的噪声数据；处理遗漏数据； 清洗脏数据

- **数据集成**

  将多文件或多数据库运行环境中的异构数据进行合并处理，其中主要涉及到实体识别问题、冗余问题、数据值冲突检测与处理问题

- **数据变换**

  找到数据的特征表示，用维变换或转换方法减少有效变量的数目或找到数据的不变式，包括规格化、归约、切换、旋转、投影等操作

- **数据归约**

  将数据库中的海量数据进行归约。归约后的数据仍接近保持原数据的完整性，但数据量相对小很多。数据归约的主要策略有数据立方体聚集、维归约、数据压缩、数值压缩、离散化、概念分层

对输入数据进行修改

- 操纵数据的策略
  - 将输入数据设计成一种能适合所选学习方案的形式
  - 将输出模型设计得更为有效
    - 样本重采样技术与元学习算法

属性选择

- 添加属性可能产生的问题
- 属性子集选择
- 独立于方案的选择

自动数据清理

- 改进决策树
- 稳健回归
- 离群点侦测

数值属性离散化

- 全局离散与局部离散
- 无监督离散与有监督离散
- 基于误差和基于熵的离散
- 离散属性转换成数值属性

属性转换

- 主成分分析
- 随机投影
- 从文本到属性向量
- 时间序列

对输入数据进行修改

- 操纵数据的策略
  - 将输入数据设计成一种能适合所选学习方案的形式
  - 将输出模型设计得更为有效
    - 样本重采样技术与元学习算法

属性选择

- 添加属性可能产生的问题
- 属性子集选择
- 独立于方案的选择

自动数据清理

- 改进决策树
- 稳健回归
- 离群点侦测

数值属性离散化

- 全局离散与局部离散
- 无监督离散与有监督离散
- 基于误差和基于熵的离散
- 离散属性转换成数值属性

属性转换

- 主成分分析
- 随机投影
- 从文本到属性向量
- 时间序列

添加属性可能产生的问题

- 添加随机属性（干扰）会导致决策树分类性能降低

  树总是会到达某个深度，那里只存在少量的数据可用于属性选择，数据越少，无关属性影响越大。

- 分治决策树学习器和割治规则学习器都存在这个问题

- 基于实例的学习器非常容易受到无关属性的影响

  在局部近邻范围内工作，每个决策只考虑少数几个实例便做出

- 朴素贝叶斯方法不受随机属性影响

  它假设所有属性都是相互独立，互不干扰的

- 添加相关属性也会导致决策树分类性能降低

  假设新属性与预测的类值是相同的，则新属性在决策树的上层便被（自然）选中用以分裂，以致其它属性选择只能基于稀疏的数据。

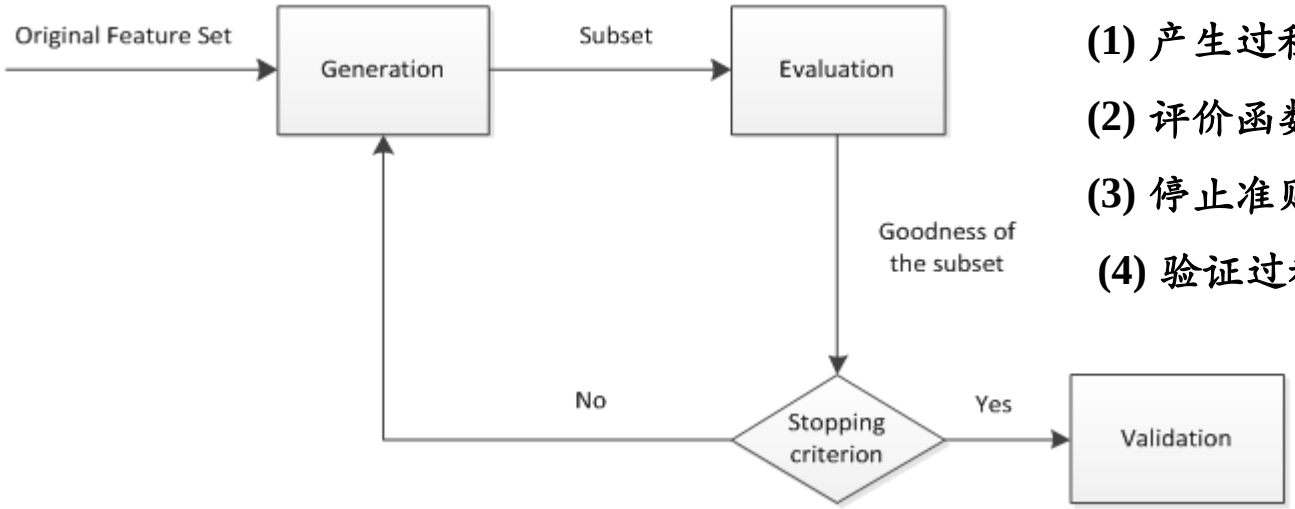
属性子集选择

- 数据集可能包含数以百计的属性，其中大部分属性可能与学习任务不相关。

  属性选择（特征选择）的目的是找到满足特定标准的最小的属性子集。

  - 首先，使用某种搜索方法找到一组属性子集；
  - 然后，测试这组属性是否满足特定标准；
  - 未满足则重新搜索，直到到达终止条件为止。

    终止条件一般是迭代次数、子集评估的阈值等。



(1) 产生过程(Generation Procedure)

(2) 评价函数(Evaluation Function)

(3) 停止准则(Stopping Criterion)

(4) 验证过程(Validation Procedure)

属性子集选择

- 属性子集选择方法

  从属性评价方法的角度，属性选择的方法可分为嵌入方法（Embedded）、过滤方法（Filter）和包装方法（Wrapper）。后两种也被称为独立于方案的选择。

  - **Embedded（嵌入）**

    将属性选择作为数据挖掘算法的一部分，特别是在算法运行期间，算法本身决定使用哪些属性和忽略哪些属性，例如构造决策树分类器的算法

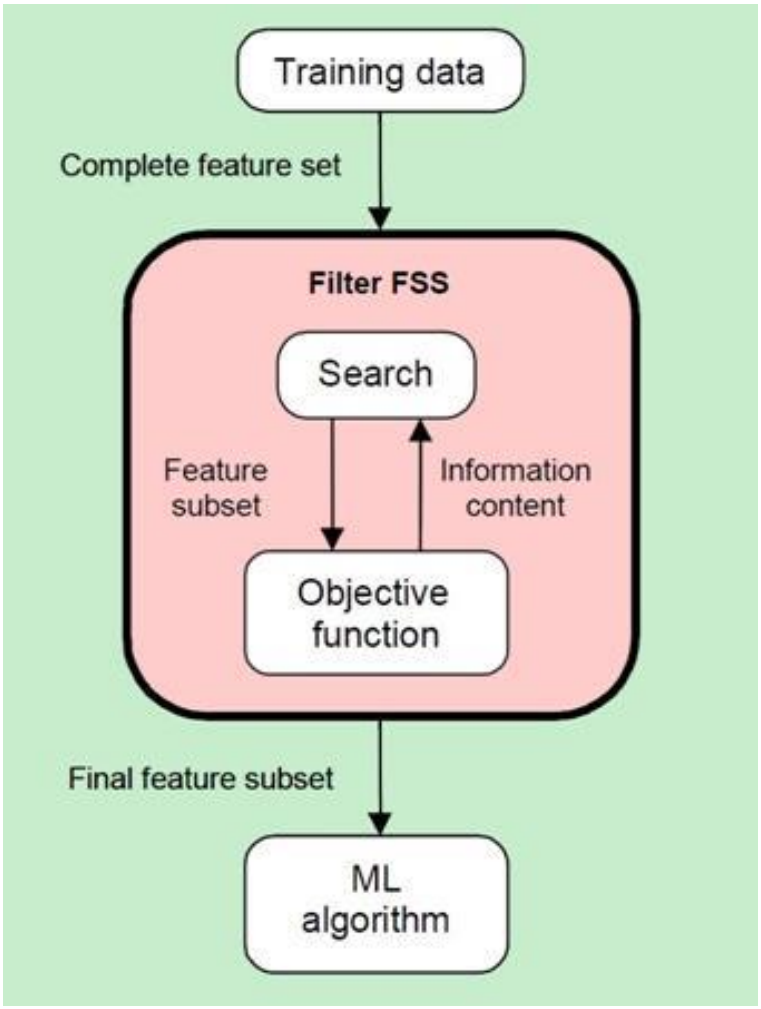
  - **Filter（过滤）**

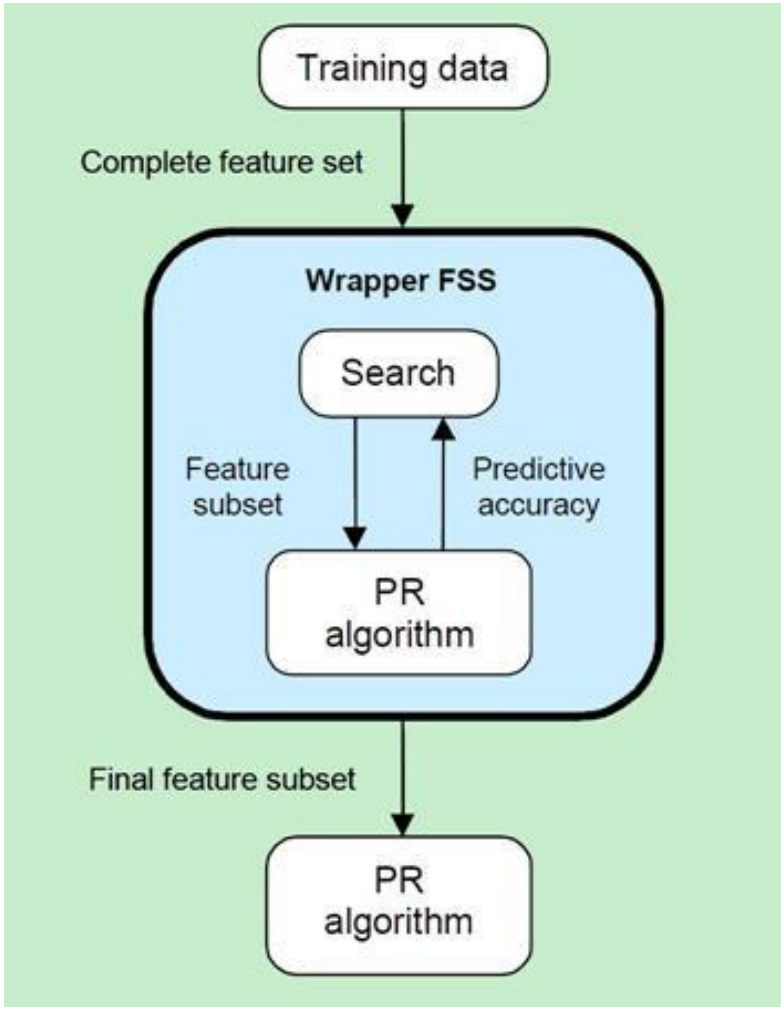
    构造一个独立于分类器的评价指标来计算属性的质量，即根据数据的普遍特性做出一个独立评估

  - **Wrapper（包装）**

    将学习算法的结果作为评价准则的一部分，直接使用分类器的分类性能估计属性的预测能力， 学习方法被包裹在选择过程中

属性子集选择





Ricardo Gutierrez-Osuna，Sequential Feature Selection

http://research.cs.tamu.edu/prism/lectures/pr/pr_l11.pdf

独立于方案的属性选择（filter）

- 构造一个独立于分类器的评价指标来计算属性的质量

  即，根据数据的普遍特性做出一个独立评估

- 根据语义上的差异属性选择有两类准则
  - 基于可分性的准则

    一般从考察类与类之间的距离或者分类的边界入手，如果各类之间的距离很大或者类与类之间的边界很小，那么在此特征空间中进行分类效果会比较好。

    - Fisher判别
    - 邻域覆盖

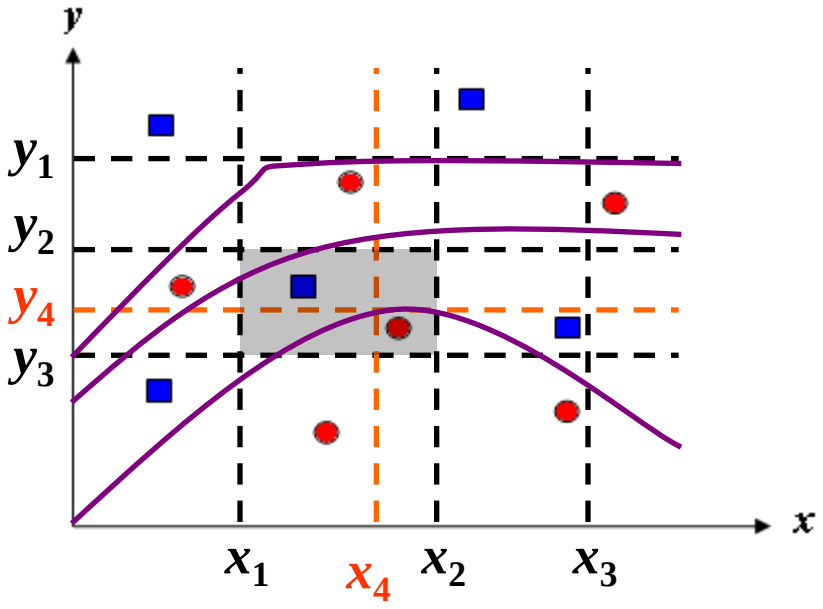
  - 基于相关性的准则（Correlation-based Feature Selection）

    从相关性语义的角度，希望描述分类问题的特征与决策变量的相关性较大，相关性越大，那么属性能够提供给分类的信息也就越丰富。

    - 互信息准则
    - 粗糙集中属性依赖度指标

独立于方案的属性选择：示例 1

- 使用足够的属性来分割实例空间，使得所有训练实例分割开来



计算耗时

过度拟合

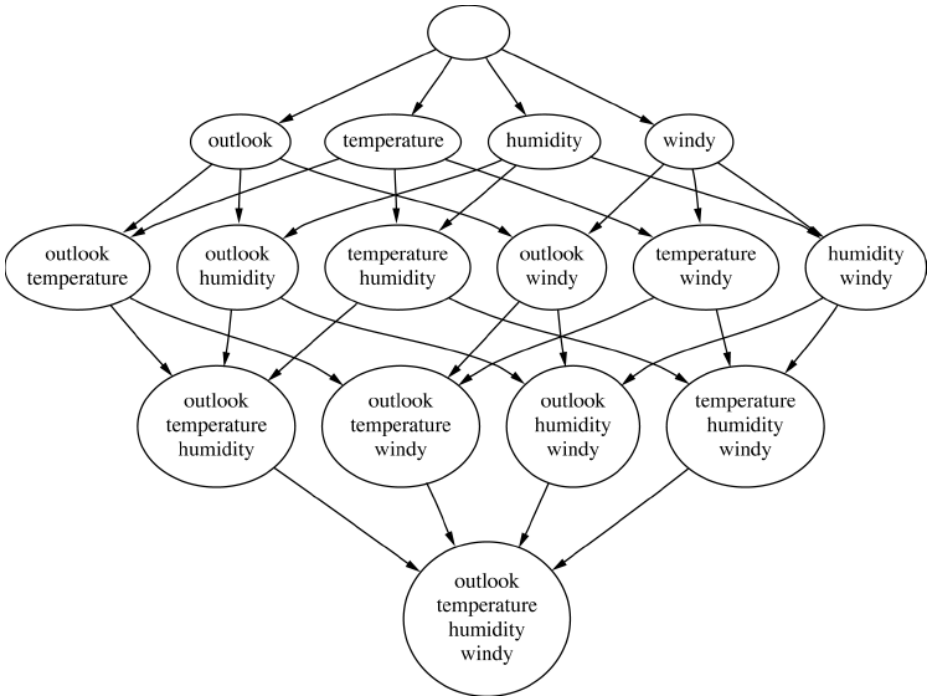
直觉告诉我们应该选择能区分所有实例的最小的属性集

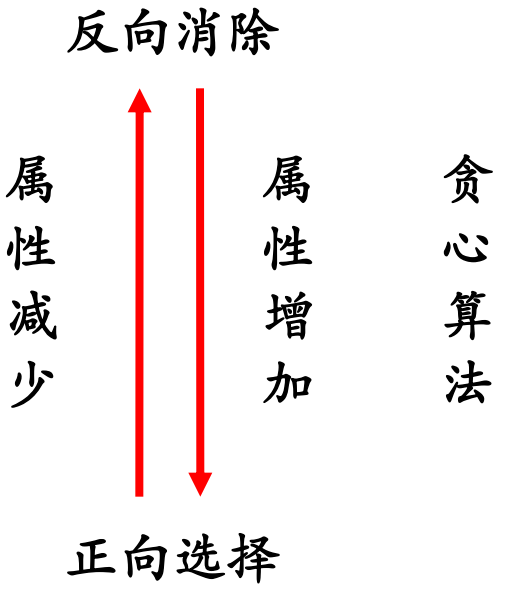
如何找出原属性的一个“好的”子集？

独立于方案的属性选择：示例 1（续）

- 搜索属性空间

  对于 n 个属性，有 2^n 个可能的子集





- 双向搜索
- 最佳优先搜索
- 粗糙集搜索
- 遗传算法搜索

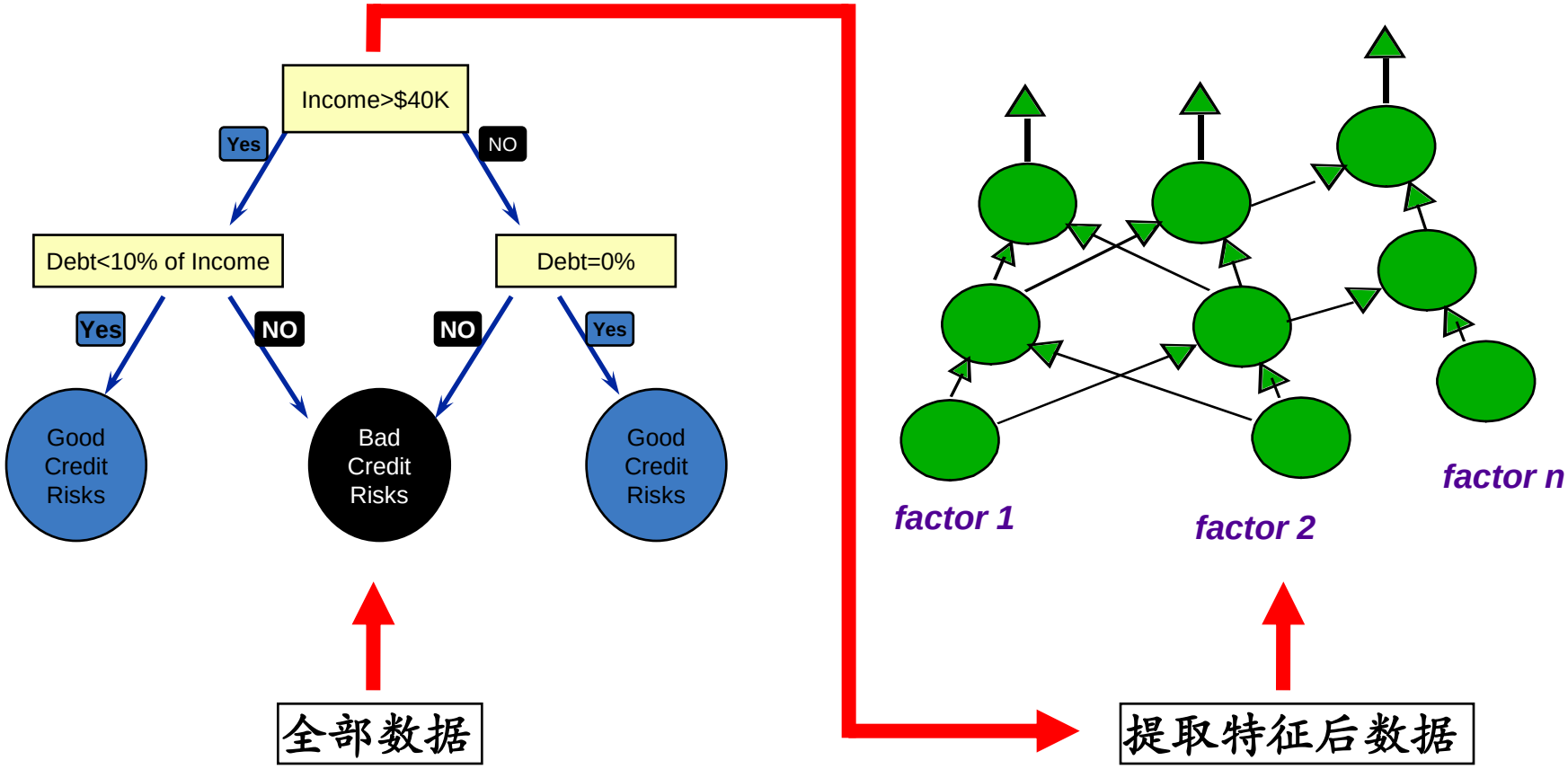
可能的属性子集数目随属性数量的增加呈指数增长

因此，穷举搜索不切实际，它只适合最简单的问题

独立于方案的属性选择：示例 2

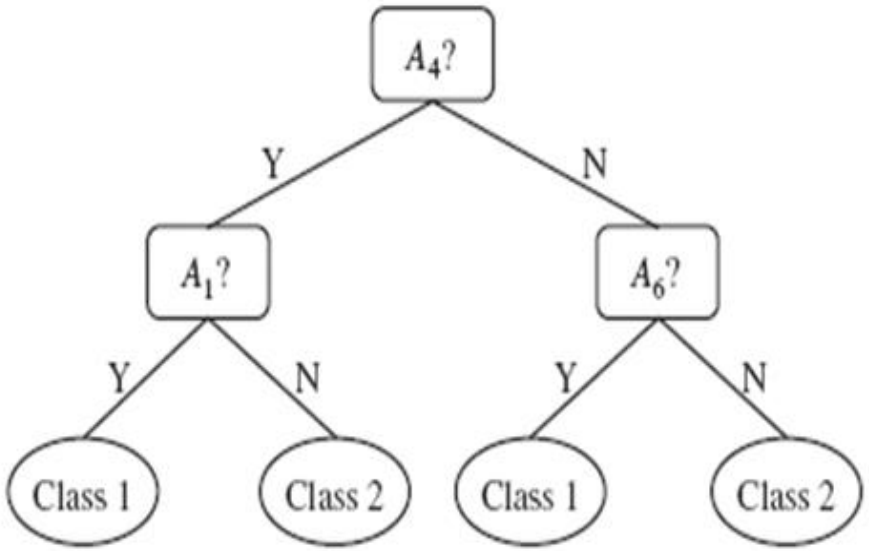
- 直接应用机器学习算法进行属性选择（不同学习方案组合）

  机器学习算法只是用来进行特征选择，而不参与后续的机器学习算法



独立于方案的属性选择：示例 2（续）

- 属性子集选择的贪心（启发式）方法

正向选择	反向消除	决策树归纳
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: {} => $\{A_1\}$ => $\{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_3, A_4, A_5, A_6\}$ => $\{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

独立于方案的属性选择：示例 3

- 基于相关性的准则总是希望:
  属性各自与类属性有较大关联但几乎没有内部关联

- 属性A与属性B之间的对称不定性（symmetric uncertainty）或互信息（mutual information）

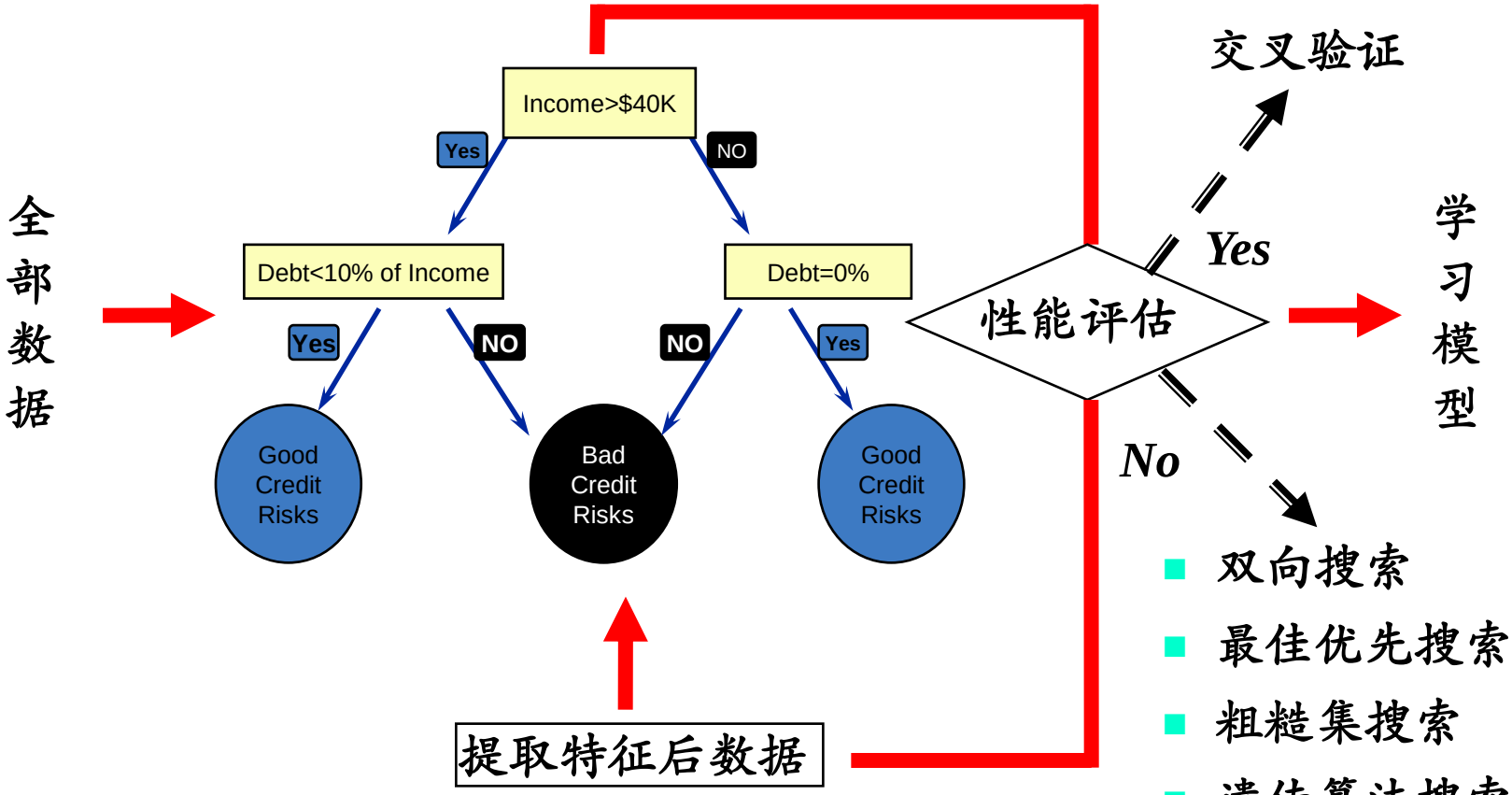
$$U(A,B)=2\frac{H(A)+H(B)-H(A,B)}{H(A)+H(B)}\in[0,1]$$

- 基于相关性的属性选择决定一个属性集的优良是采用以下公式来选择最小的子集

$$\sum_j U(A_j,C)\Big/\sqrt{\sum_i\sum_j U(A_i,A_j)}$$

独立于方案的属性选择（wrapper）

- 将学习算法的结果作为评价准则的一部分，直接使用分类器的分类性能估计属性的预测能力， 学习方法被包裹在选择过程中
  - 计算量大，效率低，小规模数据容易出现过拟合



对输入数据进行修改

- 操纵数据的策略
  - 将输入数据设计成一种能适合所选学习方案的形式
  - 将输出模型设计得更为有效
    - 样本重采样技术与元学习算法

属性选择

- 添加属性可能产生的问题
- 属性子集选择
- 独立于方案的选择

自动数据清理

- 改进决策树
- 稳健回归
- 离群点侦测

数值属性离散化

- 全局离散与局部离散
- 无监督离散与有监督离散
- 基于误差和基于熵的离散
- 离散属性转换成数值属性

属性转换

- 主成分分析
- 随机投影
- 从文本到属性向量
- 时间序列

数值属性离散化的原因

- 一些分类和聚类算法只能处理名词属性而不能处理数值属性或，即使能够处理数值属性其处理方法也并不完全令人满意
  - 朴素贝叶斯

    一般假设数值属性呈正态分布
- 一些分类和聚类算法虽然能够处理数值属性，但处理速度相当慢
  - 决策树和决策规则学习器

    当出现数值属性时，需要重复对属性值进行排序

典型示例

@ data

sunny,85,85,false,no

sunny,80,90,true,no

overcast,83,86,false,yes

rainy,70,96,false,yes

rainy,68,80,false,yes

rainy,65,70,true,yes

overcast,64,65,true,yes

sunny,72,95,false,no

@ data

sunny,hot,high,false,no

sunny,hot,high,true,no

overcast,hot,high,false,yes

rainy,mid,high,false,yes

rainy,cool,normal,false,yes

rainy,cool,normal,true,yes

overcast,cool,normal,true,yes

sunny,mid,high,false,no

数值属性离散化的策略 1

- 全局离散化
  - 单规则（1R）学习方案

根据属性值将实例进行排序，在类出现变化时设定属性值值域

Temperature	64	65	68	69	70	71	72	72	75	75	80	81	83	85
Play	yes	no	yes	yes	yes	no	no	yes	yes	yes	no	yes	yes	no

过度拟合

	yes	no	yes	yes	yes	no	no	yes	yes	yes	no	yes	yes	no

最小数量样本是3

	yes	no	yes	yes	yes	no	no	yes	yes	yes	no	yes	yes	no

合并拥有相同的多数类的相邻段

	yes	no	yes	yes	yes	no	no	yes	yes	yes	no	yes	yes	no

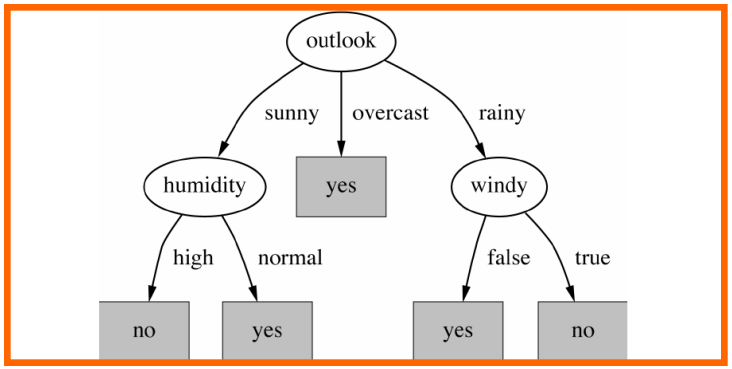
Temperature <= 77.5 Play = Yes

> 77.5 Play = No

数值属性离散化的策略 2

- 局部离散化
  - C4.5决策树学习方案

    在树的每个节点处，当决定是否值得分支时对属性进行检查，并且只在这个节点处决定连续属性的最佳分裂点



- 全局离散要优于局部离散

  因为随着树的深度增加，决定局部离散的数据越来越少，可靠性越来越差
- 离散后的属性很难保留原有属性的有序性

  数值属性是有序的，把它当作名词属性处理便丢弃了它潜在的排序信息

数值属性离散化的方法：无监督离散（装箱）

- 在训练集实例类未知的情况下，对每个属性量化，即为所谓的无监督离散（unsupervised discretization）

只有在处理类未知或类不存在的聚类问题时，才有可能碰到

  - 等值区间装箱（equal-interval binning）

    将值域分割成预先设定的相等区间：一个固定的独立于数据的尺度

Temperature	64	65	68	69	70	71	72	72	75	75	80	81	83	85
Play	yes	no	yes	yes	yes	no	no	yes	yes	yes	no	yes	yes	no

    实例分布不均匀，使用的等级过于粗糙而破坏了在学习阶段中可能有用的差别，或者选择了将不同类的实例不必要地混在一起的分割边界

  - 等频区间装箱（equal-frequency binning）
    直方图均衡化（histogram equalization）

    每个区间内的训练实例数量相等

Temperature	64	65	68	69	70	71	72	72	75	75	80	81	83	85
Play	yes	no	yes	yes	yes	no	no	yes	yes	yes	no	yes	yes	no

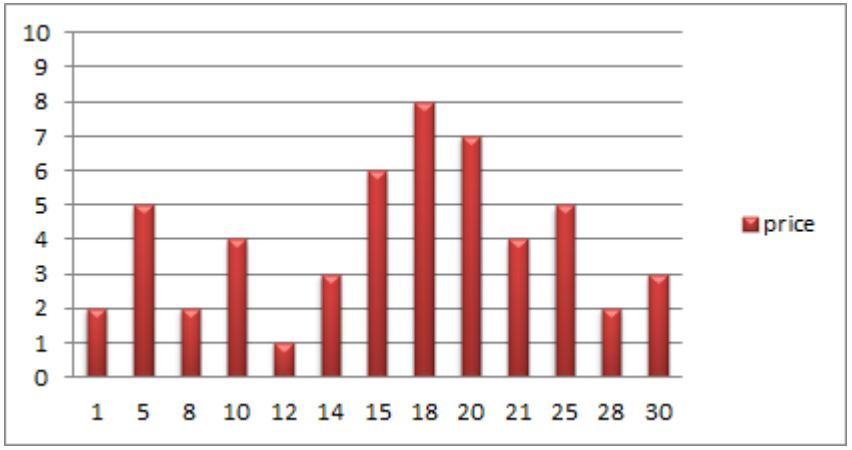
数值属性离散化的方法：无监督离散（直方图）

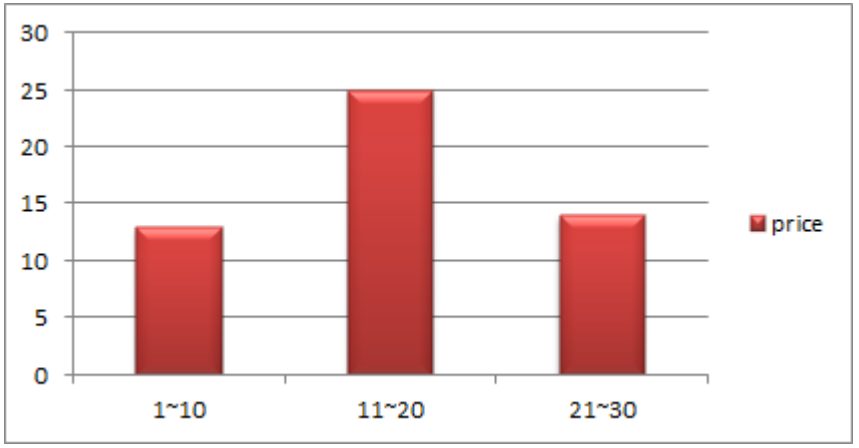
- 直方图分析（discretization by histogram analysis）

直方图使用各种划分规则把属性A的值划分成不相交的区间（桶或箱）

DATA（price）

1，1，5，5，5，5，5，8，8，10，10，10，10，12，14，14，14，
15，15，15，15，15，15，18，18，18，18，18，18，18，18，20，
20，20，20，20，20，20，21，21，21，21，25，25，25，25，25，
28，28，30，30，30





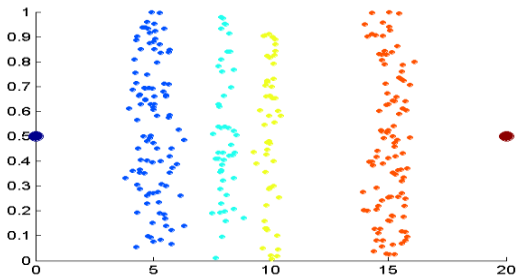
与装箱方法有何不同？

数值属性离散化的方法：无监督离散（聚类）

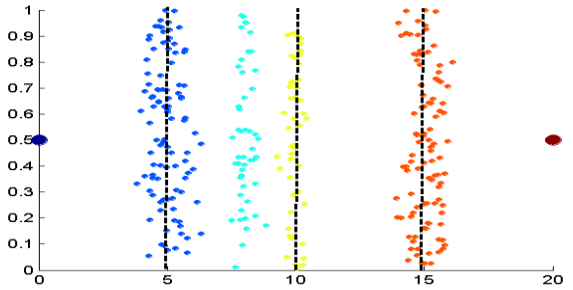
- 聚类（discretization by cluster）

  通过聚类算法将数值属性 A 的值划分成簇或组以实现离散化

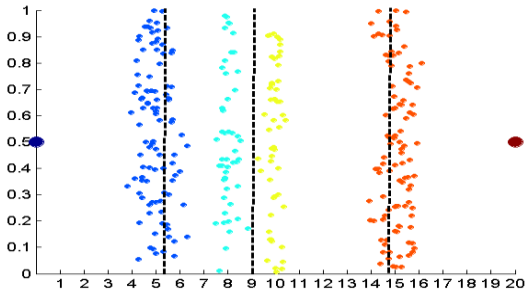
  聚类考虑 A 的分布以及数据点的临近性，因此可以产生高质量的离散化结果



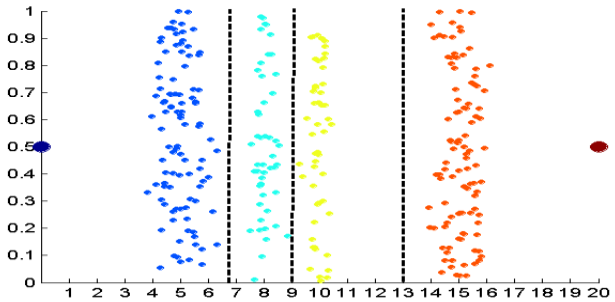
Data



Equal interval width (binning)



Equal frequency (binning)



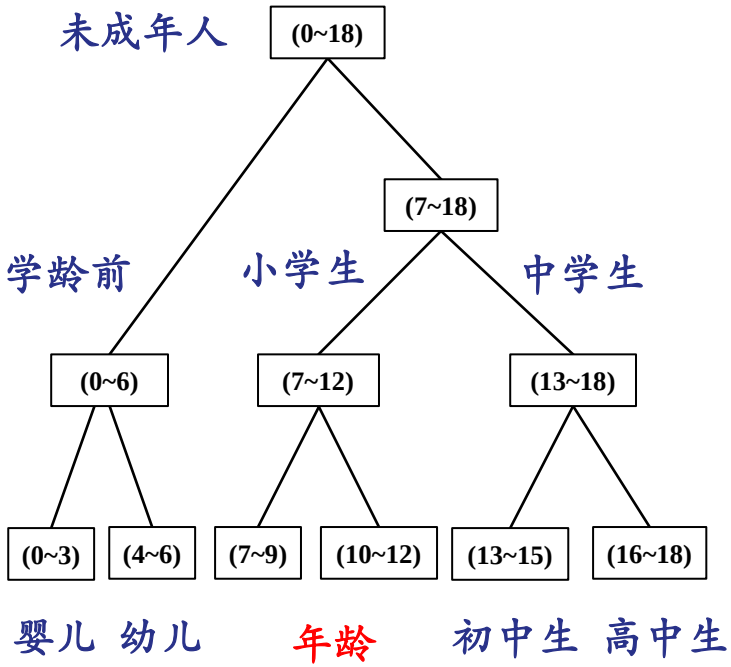
K-means clustering leads to better results

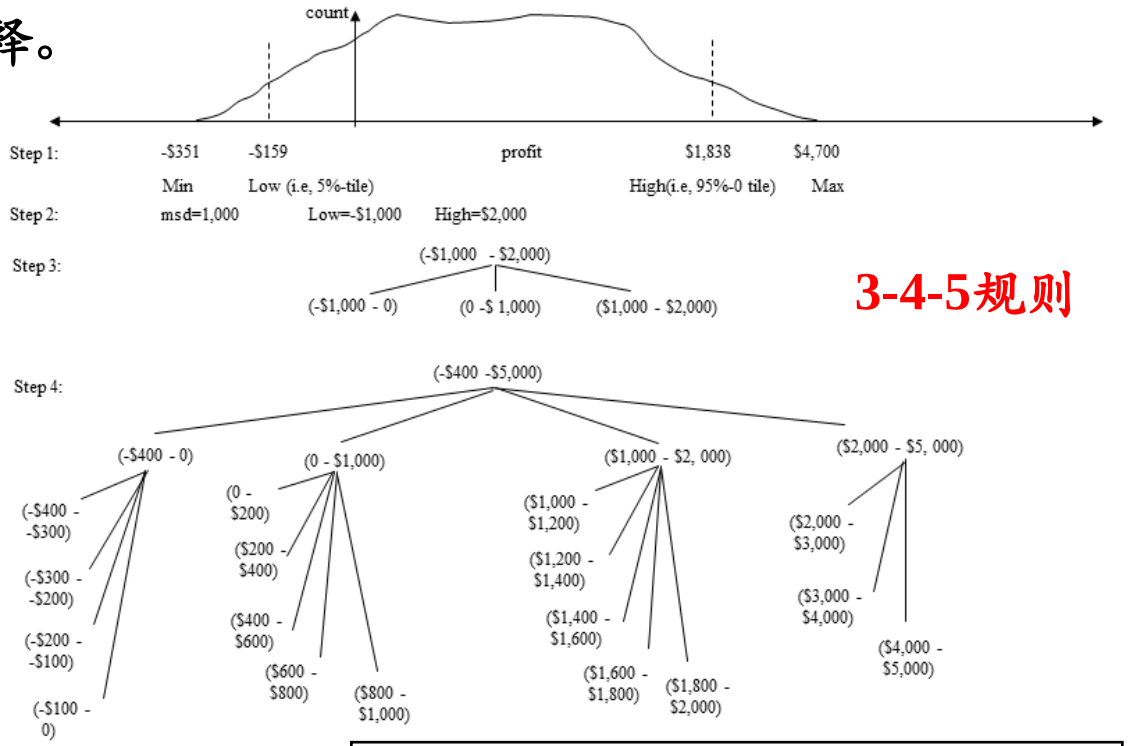
数值属性离散化的方法：无监督离散（概念分层）

- 数值属性的概念分层（concept hierarchy generation by numeric data）

  对一个属性递归地进行离散化，产生属性值的分层或多分辨率划分，称做概念分层。

  通过收集较高层的概念并用它们替换较低层的概念泛化数据产生更有意义、更容易理解的解释。





3-4-5规则

数值属性离散化的方法：有监督离散

- 在数值属性离散时需要考虑类属性，即为有监督离散（supervised discretization）
  - 基于熵和最小描述长度（Minimum Description Length，MDL）的离散
    - 根据信息量（熵）对属性离散，，第一次分裂一旦决定，分裂过程可以在上部值域或下部值域重复、递归进行
    - 应用MDL原则停止熵分裂

共有11处可能的分裂点

Temperature	64	65	68	69	70	71	72	75	80	81	83	85
Play	yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no

分裂后的两个值域分别包含有4个yes、2个no和5个yes、3个no

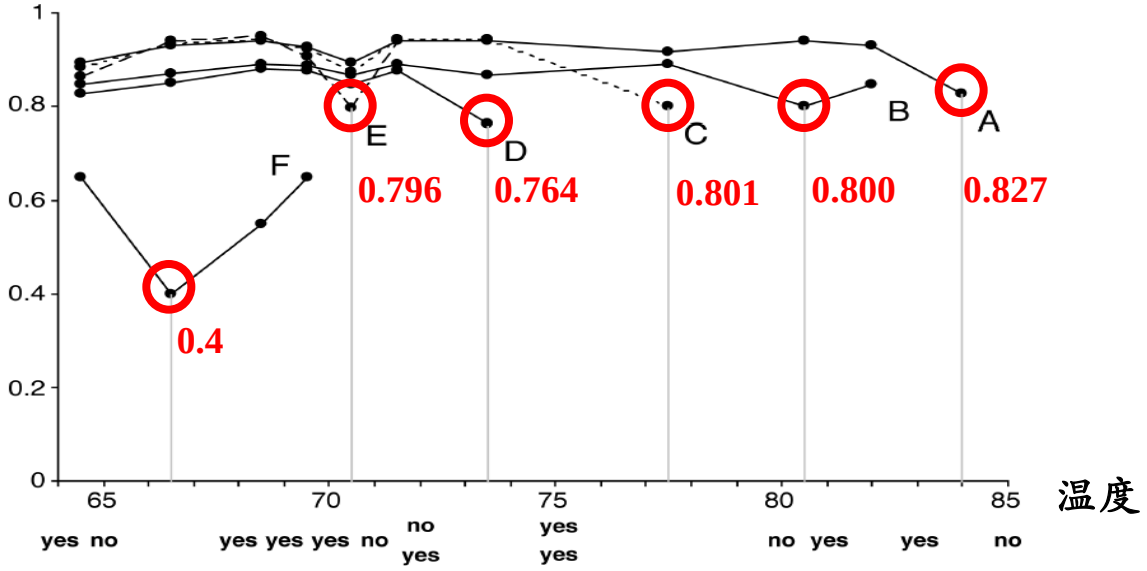
$$\text{info}([4,2],[5,3]) = (6/14)\times \text{info}([4,2]) + (8/14)\times \text{info}([5,3]) = 0.939\text{位}$$

数值属性离散化的方法：有监督离散（续1）

- 在信息量小的点进行离散化

Temperature	64	65	68	69	70	71	72	75	80	81	83	85
Play	yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no

信息量



64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no
		F			E		D	C	B		A
		66.5			70.5		73.5	77.5	80.5		84

- 一般来说，上部和下部区间都将被进一步分裂
- 从理论上看，最小的信息量绝对不会出现在两个同属一类的实例之间 ？
- 采用基于熵并应用MDL停止标准是有指导离散的最好的通用技术之一
- 其它方法

数值属性离散化的方法：有监督离散（续2）

- 基于误差和基于熵的离散

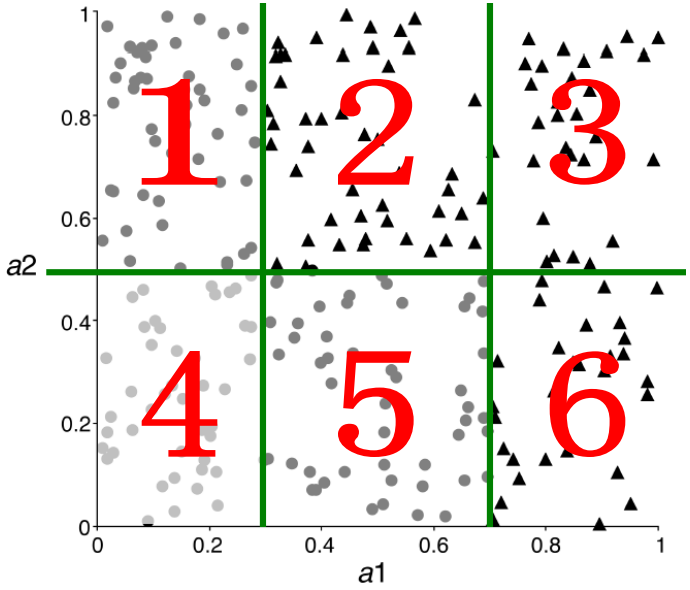
64	65	68	69	70	71	72	75	80	81	83	85
yes	no	yes	yes	yes	no	no yes	yes yes	no	yes	yes	no
		F			E		D	C	B		A
		66.5			70.5		73.5	77.5	80.5		84

基于相关性的离散

χ^2 离散

（*ChiMerge*方法）

- 缺点：基于误差的离散不让相邻区间有相同的类标签
- 原因：合并这样两个区间不影响误差累计，却能释放一个区间用于离散别处以降低误差累计
- 为什么要生成两个相同类表的相邻区间呢？
- 基于熵的离散方法对于类变化敏感而基于误差的方法不敏感



离散属性转换为数值属性

- 原因
  - 有些学习算法要求输入的变量为连续型变量

    基于实例的最近邻方法、支持向量机方法等
  - 牵扯到回归的数值预测技术也只处理数值属性
- 方法
  - 多变量法编码（对距离不敏感）

    将一个离散变量变换为多个数值变量

    （适用于变量值之间没有明显关系的变量）
  - 二值属性编码

    一个排序的名词性属性可以用一个整数代替（对距离敏感）

    为含k个值的名词性属性建立$k-1$个合成二值属性（对距离不敏感）

    （适用于变量值之间可以排序的变量）

多变量编码变换

变量：地区	华北	华东	华中	西南	西北
地区=华北	1	0	0	0	0
地区=华东	0	1	0	0	0
地区=华南	0	0	1	0	0
地区=西南	0	0	0	1	0
地区=西北	0	0	0	0	1

若原变量中取值和新变量名一致，则为1，否则为0

变量：地区	北方	中原	南方
地区=华北	1	0	0
地区=华东	0	0	1
地区=华南	0	1	0
地区=西南	0	1	1
地区=西北	1	1	0

对输入数据进行修改

- 操纵数据的策略
  - 将输入数据设计成一种能适合所选学习方案的形式
  - 将输出模型设计得更为有效
    - 样本重采样技术与元学习算法

属性选择

- 添加属性可能产生的问题
- 属性子集选择
- 独立于方案的选择

自动数据清理

- 改进决策树
- 稳健回归
- 离群点侦测

数值属性离散化

- 全局离散与局部离散
- 无监督离散与有监督离散
- 基于误差和基于熵的离散
- 离散属性转换成数值属性

属性转换

- 主成分分析
- 随机投影
- 从文本到属性向量
- 时间序列

属性转换

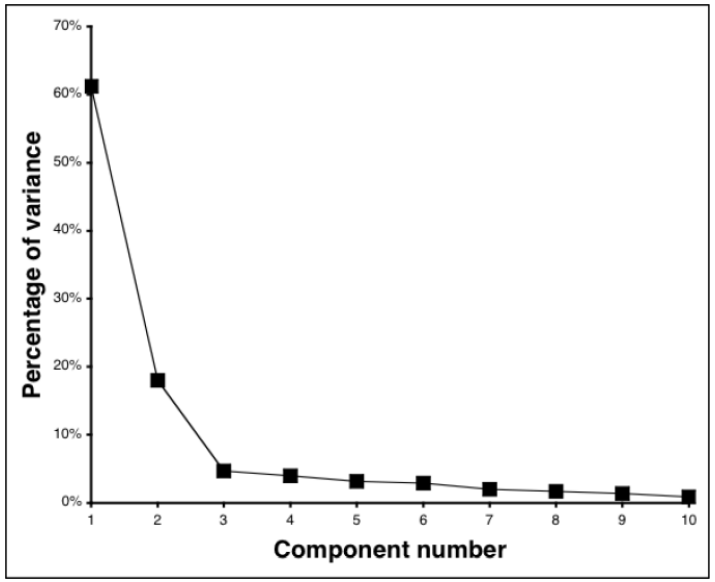
- 从数据集中原有变量的角度对数据进行分析叫特征提取 (Feature Selection)
- 将数据集中原有变量进行变换产生新的特征叫特征抽取 (Feature Extraction)
  - 将数学函数应用于现有的属性上，定义出一个新的属性
  - 在数据集上应用聚类程序，定义出一个新的属性
  - 在数据上添加一些干扰属性也会帮助检测一个算法的鲁棒性
    - 随机添加一些新的属性
    - 随机删除一些属性
    - 改变属性性质或打乱属性顺序使得数据混乱
  - 针对特殊输入的属性转换
    - 稀疏数据输入
    - 文本输入
    - 时间序列输入

属性转换：主成分分析

- 无论采用何种坐标，数据点在每个方向上都存在一个方差，预示了在这个方向上平均值周围的延伸度。奇怪的是如果将各个方向上的方差相加，然后将数据点转换到一个不同的坐标系统中并做同样的操作，两种情况下的方差总和是相同的。只要坐标系是正交的，这种关系始终是成立的。
- 主成分分析（Principle Components Analysis，PCA）的思想是使用一个特殊的、由数据点决定的坐标系，将第一个坐标轴设在数据点方差最大的方向上，第二个坐标轴与第一个坐标轴正交，且选择沿轴向的方差达到最大值的方向作为第二个轴向。依次选择下去，使坐标轴轴向的方差在所剩方差中占的分额是最大的。
  - 计算数据点原始坐标的协方差矩阵（covariance matrix）
  - 对协方差矩阵进行对角化（diagonalize）
  - 根据对角化矩阵找到特征向量（eigenvector）：转换后的空间上的坐标轴
  - 根据特征值（eigenvalue）对特征向量排序：特征值提供了这个轴向上的方差

属性转换：主成分分析（续）

- 含有10个数值属性的样本集的主成分转换



Axis	Variance	Cumulative
1	61.2%	61.2%
2	18.0%	79.2%
3	4.7%	83.9%
4	4.0%	87.9%
5	3.2%	91.1%
6	2.9%	94.0%
7	2.0%	96.0%
8	1.7%	97.7%
9	1.4%	99.1%
10	0.9%	100.0%

- 新产生的数值属性与原有数据中的数值属性不同，是它们的线性组合
- 主成分之间互不相关，因此主成分分析常用于需要属性变量不相关假设的数据挖掘算法中，如回归分析、贝叶斯方法等
- 属性的衡量尺度会影响主成分分析的结果，通常惯例是先将所有属性进行标准化，使之平均值为0且方差单元化
- 主成分分析不适宜处理离散属性，且没有考虑类信息

属性转换：随机投影

- 主成分分析将数据线性转换到低维空间，时间代价昂贵
  - 要找出转换（一个由协方差矩阵的特征向量所组成的矩阵）花费的时间将是维数的立方，对于属性数目庞大的数据集不可行
- 随机投影（Random Projection）将数据投影到维数预先设定好的子空间
  - 理论表明随机投影能相当好地保持距离关系
  - 在最近邻方法中，使用多个随机矩阵来建立一个联合分类器能使结果更加稳定，而且更少依赖于随机投影的选择
  - 实验表明，随着维数的升高，随机投影效果与主成分分析效果的差距呈减小趋势
  - 随机投影计算成本要低得多

属性转换：规范化或标准化

- 属性的度量单位可能影响数据分析

  一般而言，用较小的单位表示属性将导致该属性具有较大值域，因此趋向于使这样的属性具有较大的影响或较高的“权重”。

  - 最小 - 最大规范化

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

把属性 A 的值 v_i 映射到区间 [new_min_A，new_max_A] 中的 v_i'

如果今后的输入实例落在 A 的原数据值域之外，

则该方法面临“越界”的错误

属性转换：规范化或标准化（续1）

- *z* 分数（*z-score*）规范化 （或零均值规范化）

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

当属性 *A* 的实际最小值和最大值未知，或离群点左右了最小-最大规范化时，该方法是有用的

标准差可以用 *A* 的均值绝对偏差（mean absolute deviation）S_A 替换

$$v_i' = \frac{v_i - \bar{A}}{S_A} \qquad S_A = \frac{1}{n}\left(|v_1 - \bar{A}| + |v_2 - \bar{A}| + \cdots + |v_n - \bar{A}|\right)$$

对于离群点，均值绝对偏差 S_A 比标准差更加鲁棒

？

属性转换：规范化或标准化（续2）

- 小数定标规范化

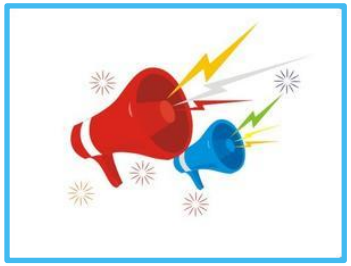
  通过移动属性 A 的值的小数点位置进行规范化。小数点的移动位置依赖于 A 的最大绝对值。

$$v_i' = \frac{v_i}{10^j}$$

j 是使得 $max(|v_i'|) < 1$ 的最小整数

规范化可能将原来的数据改变很多，特别是使用 z 分数规范化或小数定标规范化时尤其如此。因此有必要保留规范化参数（如均值和标准差），以便将来的数据可以用一致的方式规范化。

从文本到属性向量



- 字符串属性从本质上看是未指明属性值数目的名词性属性
- 字符串属性中的文本可以分解为一个个段落、句子或词组
- 文本所包含的单词是最有用的单元
- 转变为单词（tokenization）的问题：
  - 是否只保留字母而丢弃非字母字符？例如：数字、＋－号、幂次方等
  - 如何定义分割符？例如：空格、、换行符、标点符号等
  - 所有单词在被加入词汇表之前是否需要先转变为小写？
  - 预先设定的功能词或停止词（stopwords）是否可以忽略？例如：the
    停止词是取决于语言的。大写习惯、数字语法、标点符号习惯、字符集本身都是取决于语言的
  - 低频率词，譬如只用过一次的词（hapax legomena），是否被丢弃？
- 每个单词的属性值应该是什么？
  - 将文件j中的单词i所出现的频率f_{ij}按照不同的标准方式进行转换

这苹果不大好吃

$$\log\left(1+f_{ij}\right)$$

$$f_{ij}\log\frac{\text{文件数量}}{\text{含有词 } i \text{ 的文件数量}}$$

→ TF × IDF

时间序列

注意样本属性中是否含有时间属性，否则样本处理方法不同

- 在时间序列数据中，每个实例代表不同的时间间隙，属性给出了该时间间隙所对应的值，如气象预报或股市行情预测
  - 将当前实例的一个属性值用过去或将来的实例所对应的属性值来替换
  - 用当前实例与过去实例属性值的差值来替换当前的属性值（更常用）

    差值从本质上来说是以由时间间隙大小所决定的某个常量为量度的第一次求导，连续差值转换即为更高次的求导
- 在一些时间序列中，实例并不代表定期的样本，每个实例的时间由特定的时间戳（**timestamp**）属性给出，如股票市场中的临时公告
  - 不同的时间戳之间的差别在于实例的时间间隙大小不同，如果要取其它属性的连续差值，必须除以时间戳大小以使求导正常化
- 每个属性代表不同的时间，而非每个实例代表不同时间，如上学经历
  - 时间序列从一个属性到下一个属性，而非从一个实例到下一个实例。

    如果需要差值，必须取每个实例的一个属性和下一个属性之间的差值

一些算法对数据预处理的要求

算法名称	缺值处理	异常点处理	归一化	数值化	离散化	变换与合成	变量选择
决策树	不需	不需	不需	不需	不需	需	需
BP网络	需	需	需	需	不需	需	需
规则归纳	需	不需	不需	不需	需	需	不需
支持向量机	需	需	不需	需	不需	需	需
朴素贝叶斯	需	不需	不需	不需	不需	需	需
粗糙集方法	需	不需	不需	不需	需	需	不需
回归分析	需	需	不需	需	不需	需	需
k－最近邻	需	需	需	不需	不需	需	需
聚类	需	需	需	不需	不需	需	需
关联规则	需	不需	不需	不需	需	需	不需

一些算法适用的数据类型与学习问题

	Data Type				*Data Mining Problem*		
Data Mining Methodology	*Labeled Data*	*Unlabeled Data*	*Separate Data Records*	*Time Series Data*	*Predication and Classification*	*Discovery of Data Patterns, Associations, and Structure*	*Recognition of Data Similarities and Differences*
Decision trees	X		X		X	X	X
Association rules		X	X			X	X
Artificial neural networks	X	X	X	X	X		X
Statistical analysis of normal and abnormal data		X	X		X		X
Bayesian data analysis	X	X	X	X	X	X	X
Hidden Markov processes and sequential pattern mining	X	X		X	X		X
Prediction and classification models	X		X	X	X	X	X
Principal components analysis		X	X			X	X
Psychometric methods of latent variable modeling	X	X	X		X	X	X
Scalable clustering		X	X			X	X
Time series similarity and indexing	X	X		X	X		X
Nonlinear time series analysis	X	X		X	X	X	X

对输入数据进行修改

- 操纵数据的策略
  - 将输入数据设计成一种能适合所选学习方案的形式
  - 将输出模型设计得更为有效
    - 样本重采样技术与元学习算法

属性选择

- 添加属性可能产生的问题
- 属性子集选择
- 独立于方案的选择

自动数据清理

- 改进决策树
- 稳健回归
- 离群点侦测

数值属性离散化

- 全局离散与局部离散
- 无监督离散与有监督离散
- 基于误差和基于熵的离散
- 离散属性转换成数值属性

属性转换

- 主成分分析
- 随机投影
- 从文本到属性向量
- 时间序列

自动数据处理

数据质量低劣是令实际机器学习任务头痛的事情

- 人工艰辛地检查数据
- 应用机器学习技术自动过滤数据（作用有限）
  - 改进决策树
    - 去除错分样本再训练（在判定错分样本是错误样本下更有效）
    - 添加属性干扰（最好对类无干扰）

      只在训练数据中添加属性干扰或系统类干扰，去除非系统类干扰
  - 离群点检测
    - 统计回归中应用可视化技术
    - 多方法综合判断

      保守策略：所有方法判断都失败时，删除样本

      存在问题：可能会牺牲某个小类别的实例来提高剩余类别的正确性

由于缺乏咨询专家，没有办法知道某个实例真的是一个错误或者只是所应用的模型不适合它

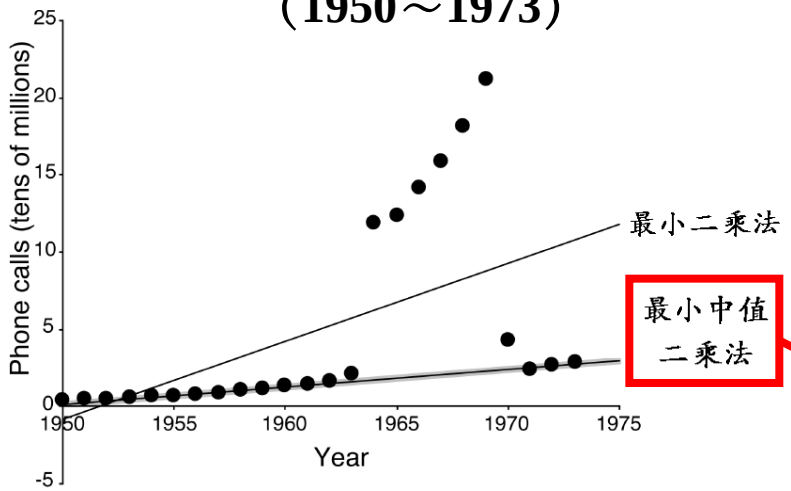
自动数据处理（续）

- 稳健回归（处理离群点的统计方法称为稳健型，robust）

  离群点大大影响了最小二乘回归（least-squares regression）

  二乘方的距离衡量加强了远离回归线的数据影响

比利时国际长途电话数量
（1950～1973）



- 采用绝对值距离衡量
- 自动识别离群点并将其删除
- 最小化距离回归线二乘中值（median，非平均值）

寻找一条覆盖半数观察点的最窄带，带的厚度是从垂直方向衡量

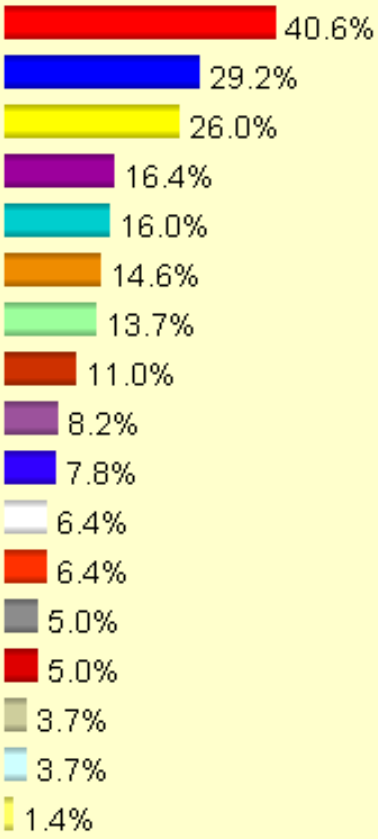
- 最小中值二乘回归线比最小二乘回归线更抗离群点干扰
- 稳健型通常是重新定义距离公式来实现对离群点的处理
- 计算成本高，对实际问题经常是不可行的 ？

网络调查 2

KDnuggets : **Polls** : Tools / Languages for Data Cleaning (Sep 2008)

Poll

What tools/languages you typically use for data manipulation/cleaning [219 voters total]



参考文献

- ***Exploratory Data Mining and Data Cleaning***. John Wiley & Sons, 2003

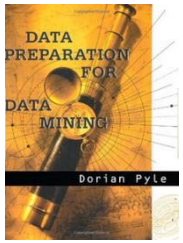
  作者: Tamraparni Dasu, Theodore Johnson

- ***Data Quality: The Accuracy Dimension***. Morgan Kaufmann, 2003

  作者: Jack E. Olson

- ***Data preparation for data mining***. San Francisco: Morgan Kaufmann, 1999

  作者: D. Pyle

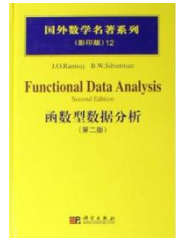


- **数据质量工程实践:获取高质量数据和可信信息的十大步骤**. 电子工业出版社, 2010

  作者: (美国) Danette McGilvray, 翻译: 刁兴春, 曹建军, 张健美 等

- **函数型数据分析 (第二版) (Functional Data Analysis)**. 科学出版社, 2006

  作者: Jim O. Ramsay, Bernard W. Silverman



参考文献

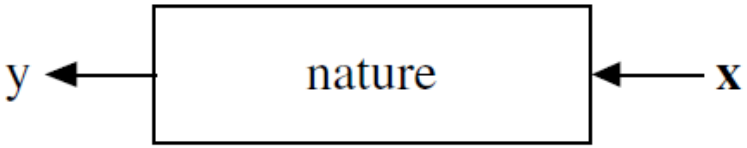
- H. Liu, F. Hussain, C. L. Tan, and M. Dash. ***Discretization: An enabling technique***. Data Mining and Knowledge Discovery, Vol. 6(4): 393-423, 2002 *http://www.comp.nus.edu.sg/~tancl/Papers/Journals/DMKD-Liu-2002.pdf*
- Kohavi, Ron, John H. George. ***Wrappers for Feature Subset Selection***. Artificial Intelligence, Vol.97: 273-324, 1997 *http://emf.mit.edu/afs/cs.pitt.edu/usr0/hwa/docs/toread/kohavi97.pdf*
- Surajit Chaudhuri, Umeshwar Dayal. ***An overview of data warehousing and OLAP technology***. ACM SIGMOD Record, Vol. 26(1): 65-74, 1997 *http://www.acm.org/sigmod/record/issues/9703/chaudhuri.ps*
- Elmagarmid, Ahmed; Ipeirotis, Panagiotis; Verykios, Vassilios. ***Duplicate Record Detection: A Survey***. IEEE Trans. on Knowledge and Data Engineering, Vol.19(1): 1-16, 2007 *http://homepages.inf.ed.ac.uk/wenfei/tdd/reading/tkde07.pdf*
- Xin Luna Dong and Wang-Chiew Tan. ***Special Issue on Towards Quality Data with Fusion and Cleaning***. Bulletin of the Tech. Committee on Data Eng., Vol.34(3), Sept. 2011 *http://sites.computer.org/debull/A11sept/issue1.htm*

数据建模的困境（1）

Leo Breiman 加州伯克利统计系教授，美国国家科学院院士，20世纪伟大的统计学家，囊括多项统计领域大奖。机器学习先驱者，分类回归树作者之一，**Bagging、Random Forests**方法发明者，对机器学习、模式识别领域有巨大贡献。于**2005**年逝世。

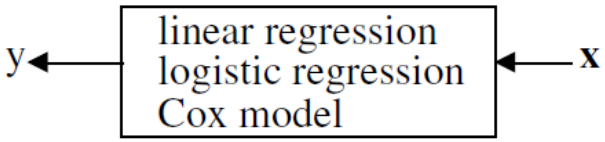
http://www.stat.berkeley.edu/users/breiman/

Breiman, L. (2001). "Statistical Modeling: the Two Cultures". Statistical Science 16 (3): 199–215.

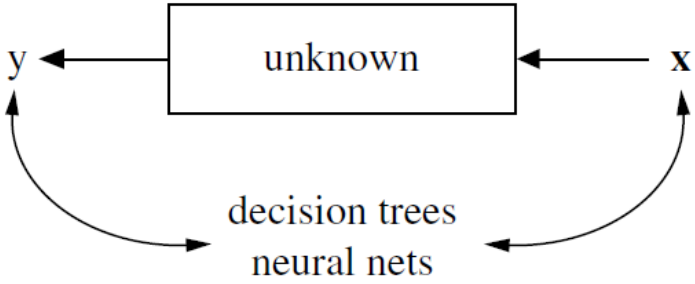


学习目标1：对于新输入的样本$\tilde{x}$，是否可以准确预测输出 $\tilde{y}$

学习目标2：是否可以依据已有数据得出“nature” 的一些信息



Data Modeling Culture

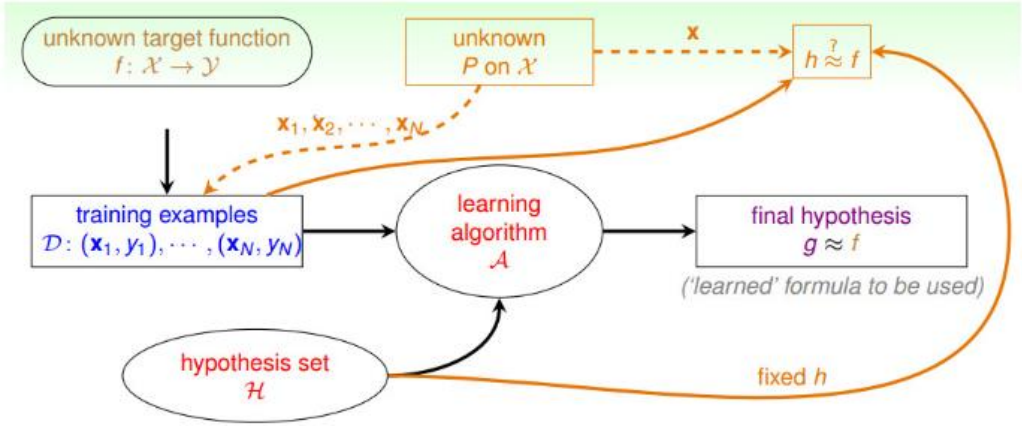


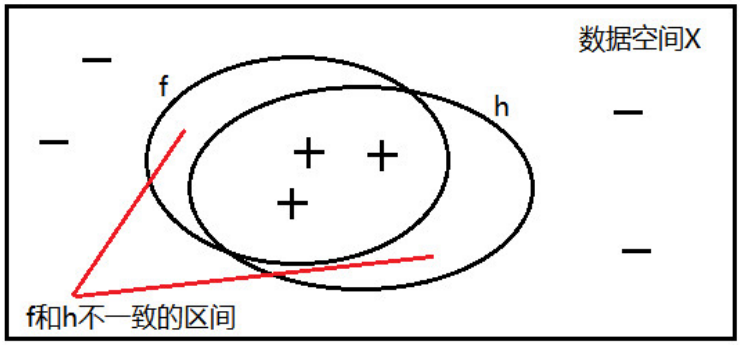
Algorithmic Modeling Culture

数据建模的困境（2）

- 可能近似正确（Probably Approximately Correct，PAC）学习模型

	x	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
$\mathcal{D}$	000	o	o	o	o	o	o	o	o	o	o
	001	×	×	×	×	×	×	×	×	×	×
	010	×	×	×	×	×	×	×	×	×	×
	011	o	o	o	o	o	o	o	o	o	o
	100	×	×	×	×	×	×	×	×	×	×
	101		?	o	o	o	o	×	×	×	×
	110		?	o	o	×	×	o	o	×	×
	111		?	o	×	o	×	o	×	o	×





机器学习的现实情况：

1、除非对每个可能的数据进行训练，否则总会存在多个假设使得真实错误率不为0，即学习器无法保证和目标函数完全一致

2、训练样本是随机选取的，训练样本总有一定的误导性

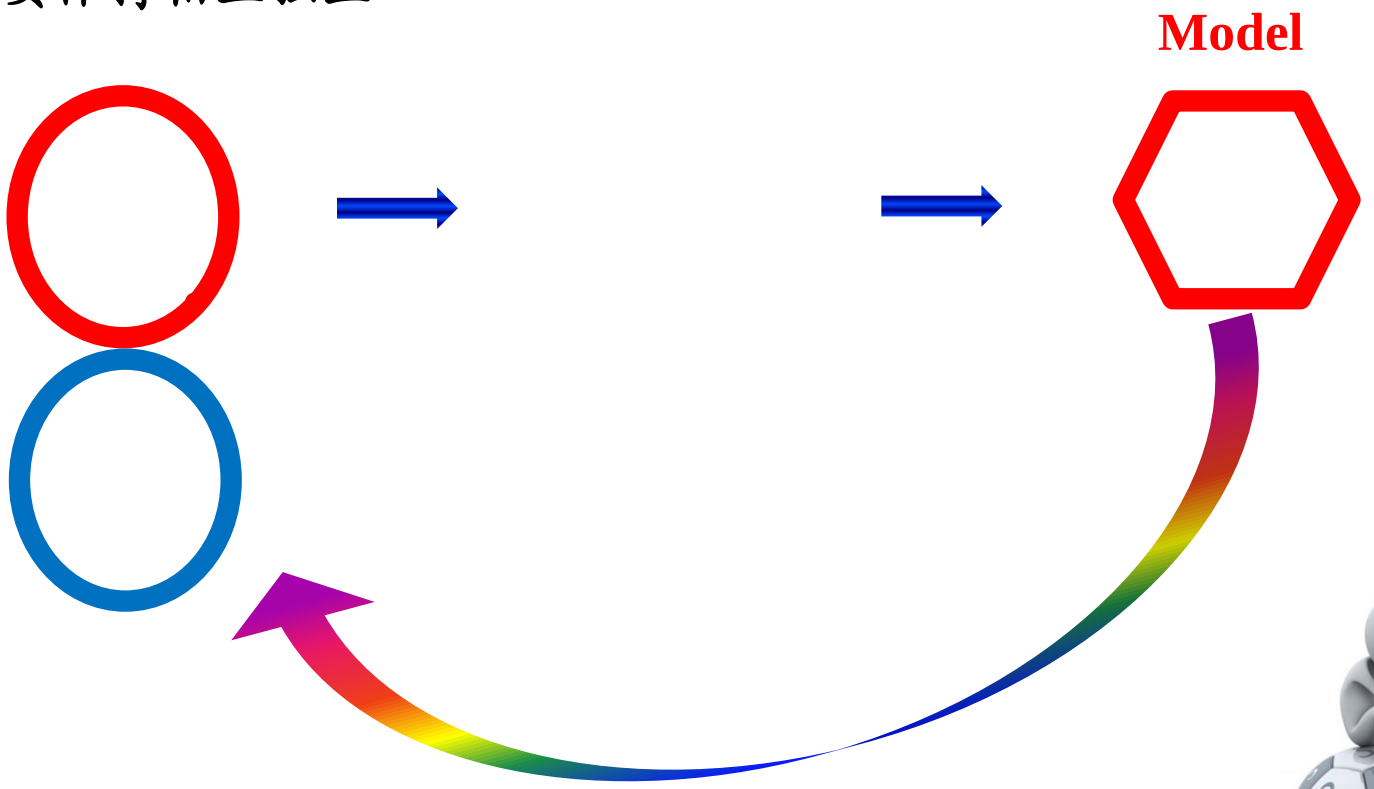
一个可PAC学习的学习器要满足两个条件：

- 学习器必须以任意高的概率输出一个错误率任意低的假设
- 学习过程的时间最多以多项式方式增长

经验风险最小化
(Empirical Risk Minimization, ERM)
结构风险最小化
(Structural Risk Minimization, SRM)

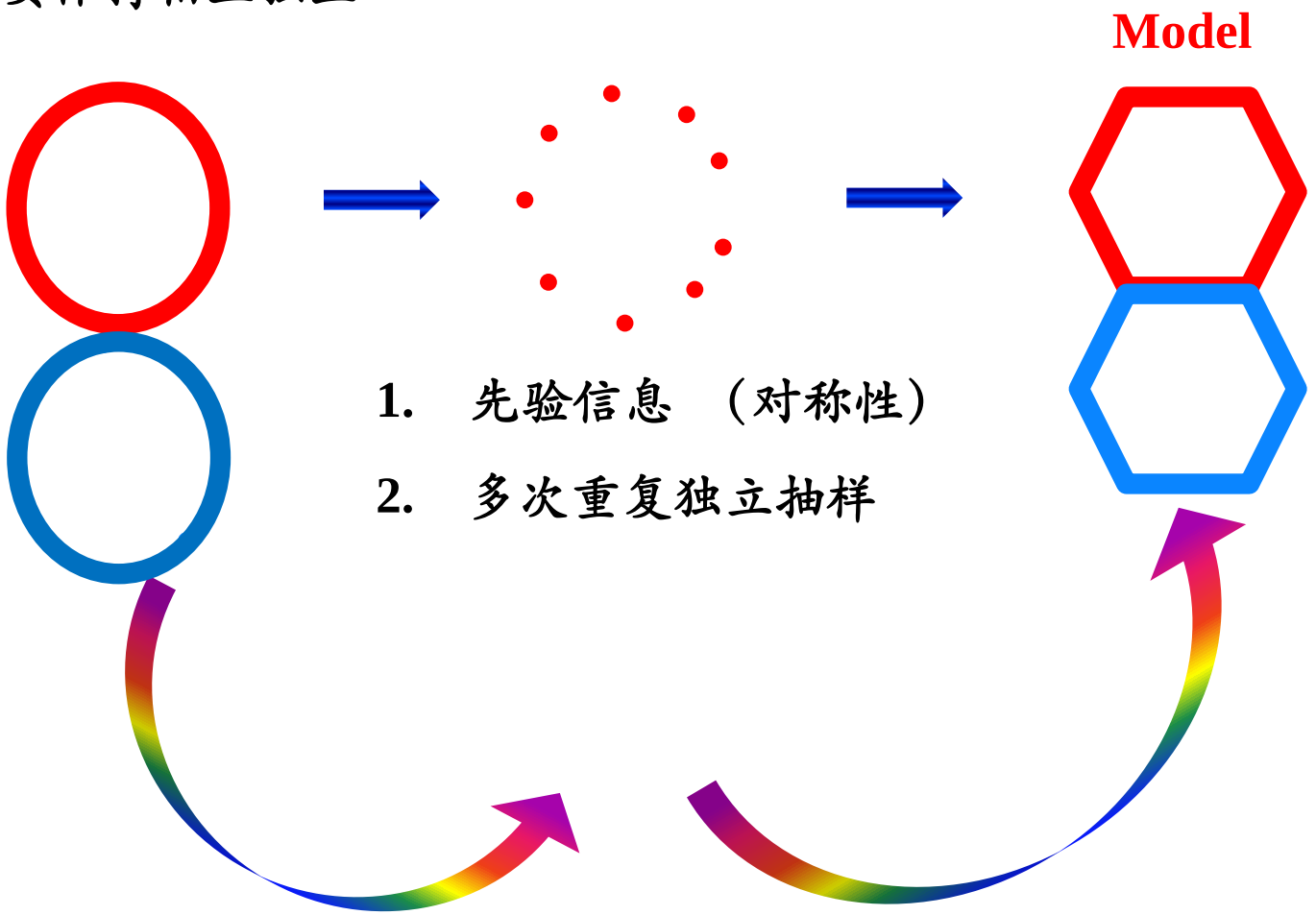
数据建模的困境（3）

基于有限数据建立的学习模型，源自数据但不能依赖数据，学习模型与学习数据要保持相互独立



数据建模的困境（4）

基于有限数据建立的学习模型，源自数据但不能依赖数据，学习模型与学习数据要保持相互独立



数据建模的困境（5）

- 统计数字会撒谎（真实数据后面的谎言，畅销美国50年的投资经典），中国城市出版社，2009

  作者: (美国) Darrell Huff　翻译: 廖颖林

  - 根据样本得到的结论不会比样本更精确
  - 在所有抽样研究中都有误差，忽略这些误差将导致一些愚蠢的举动
  - 为了确保结论有价值，根据抽样得出的结论一定要采用具有代表性的样本，这种样本才能排除各种误差
  - 采用严重有偏的样本几乎能够产生任何人需要的任何结果
  - 多少才算够呢？这又是个棘手的问题。它取决于其他的因素，即你采用抽样方式所研究的总体容量有多大、变化程度有多大。值得一提的是，这时样本的规模与看上去的并不一致

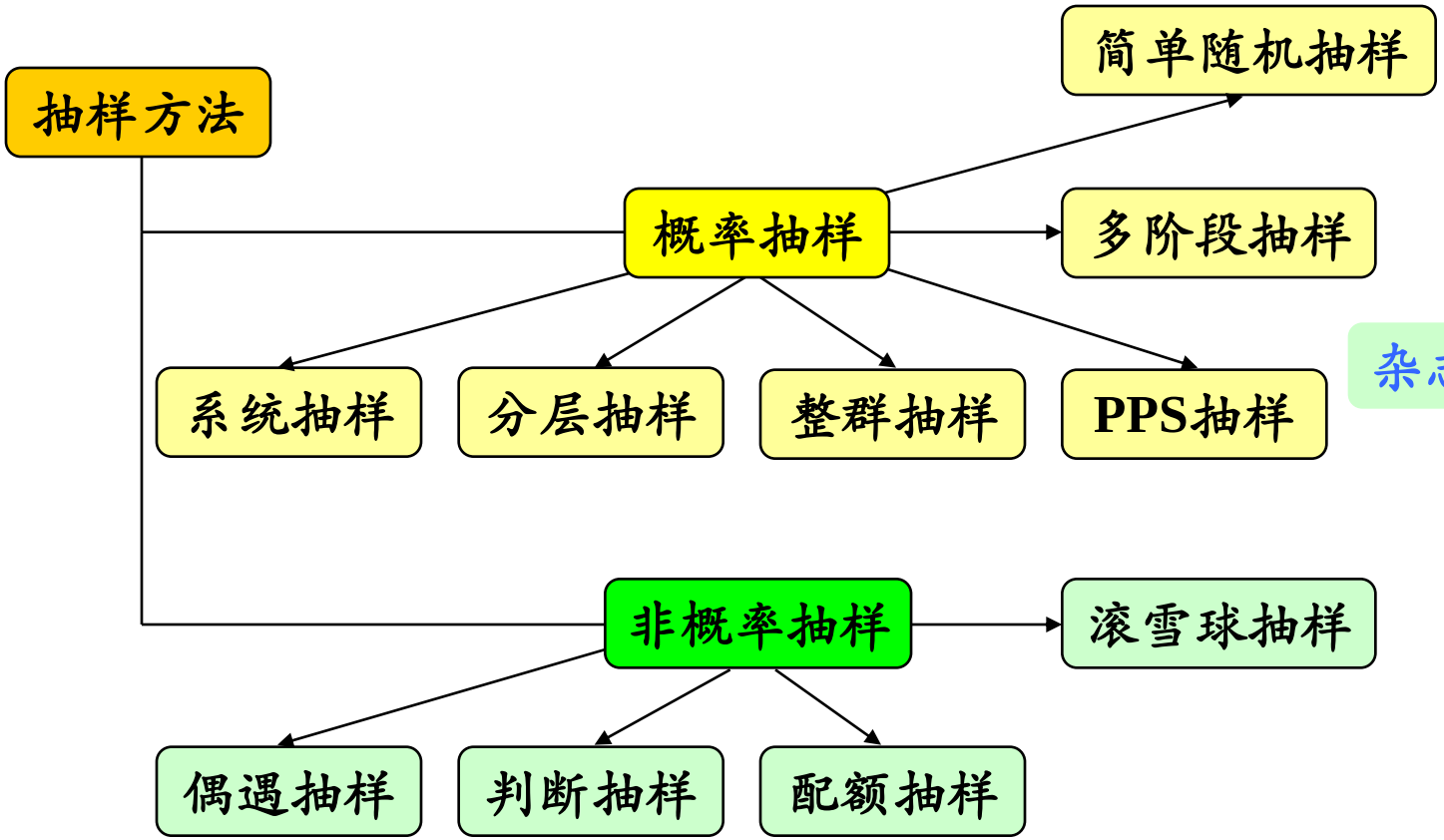
“据统计，麻省理工学院某系有50%的女生与男教师同居”，这一结果震动了校方，但一调查，确实如统计结果所说:
该系共有两名女生，其中一个与男教师堕入爱河……

数据抽样方法的选择（Sampling Techniques）

你不必为知道牛肉的滋味而吞掉整头牛

—— Paul A Samuelson

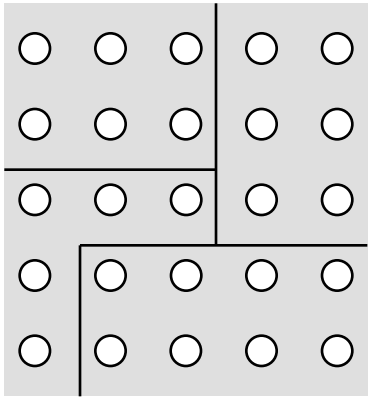
美国经济学家 (1915 －2009)



街头随访

邮寄式调查

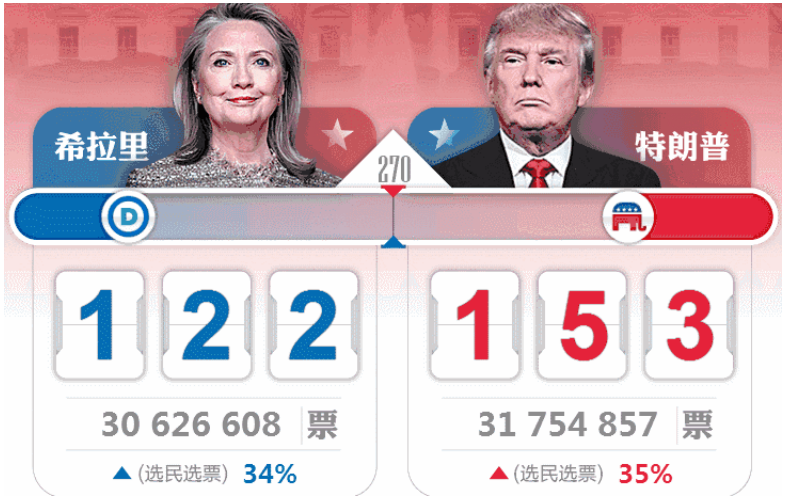
杂志上的问卷调查



数据抽样方法的选择（Sampling Techniques）

抽样方法		优点	缺点
概率抽样	简单随机抽样	易理解，结果容易推广到总体	抽样框不易构造，费用高，样本代表性有时不能保证
	系统抽样	很强的代表性，操作简单	代表性有时会很差
	分层抽样	精度高，可对子总体进行估计	分层困难，成本高
	整群抽样	费用省，易操作	精度差
	多阶段抽样	易操作，精度高于整群抽样	计算复杂，误差大
	PPS抽样	提高样本对总体的代表性	计算复杂
非概率抽样	偶遇抽样	省时省力，易操作	样本代表性差，易产生误差
	判断抽样	费用省，易操作	受主观影响
	定额抽样	某些特征上可以对样本进行控制	有选择偏差
	滚雪球抽样	成本低，易于对特殊群体做调查	调查可能不全面

数据抽样方法的选择：示例



Trump, Failure of Prediction, and Lessons for Data Scientists

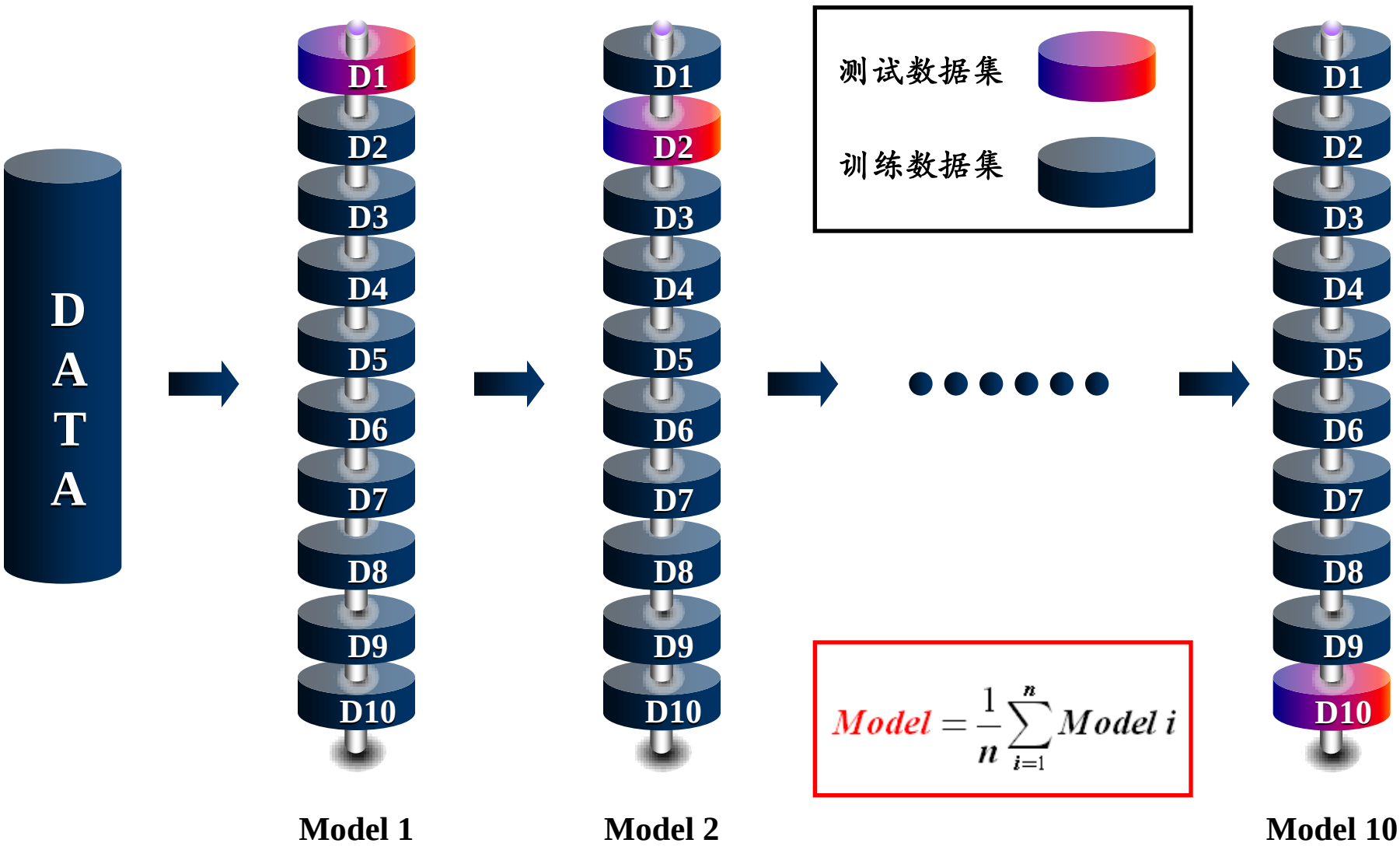
Gregory Piatetsky ***The shocking and unexpected win of Donald Trump of presidency of the United States has once again showed the limits of Data Science and prediction when dealing with human behavior.***

http://www.kdnuggets.com/2016/11/trump-shows-limits-prediction.html

- Correct prediction is based on statistics and statistics requires history of similar events and assumptions like independent variables to function correctly
- So a good lesson for Data Scientists is to question their assumptions and to be especially skeptical when predicting a rare event with limited history using human behavior.

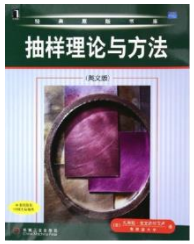
抽样样本的可靠性

交叉验证（Cross Validation）



参考文献

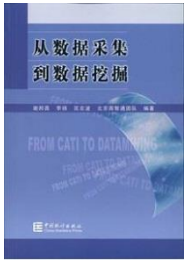
- **抽样理论与方法（英文影印版）**. 机械工业出版社, 2005
  作者: (美国) Zakkula Govindarajulu



- **抽样技术及其应用**. 清华大学出版社, 2005
  作者: 杜子芳

- **抽样技术（第三版）**. 中国人民大学出版社, 2012
  作者: 金勇进, 杜子芳, 蒋妍

- **从数据采集到数据挖掘**. 中国统计出版社, 2009
  作者: 谢邦昌, 李扬, 匡洪波, 北京商智通团队



- **数据学**. 复旦大学出版社, 2009
  作者: 朱扬勇, 熊赟

- B.H. Gu, F.F. Hu and H. Liu. ***Sampling and Its Application in Data Mining: A Survey***. Technical Report TRA6/00, National University of Singapore, Singapore, 2000
  https://dl.comp.nus.edu.sg/dspace/bitstream/1900.100/1408/1/report.pdf

总结

- 数据分类
- 数据预处理原因与操作数据策略
- 属性选择与属性转换
- 数值属性离散化与数据自动处理
- 数据建模的困境（模型评估）
- 数据抽样方法的选择

思维活动中最困难的是重新排版整理一组熟悉的资料，从不同的角度着眼看待它，并且摆脱当时流行的理论

—— Herbert Butterfield

英国历史学家和哲学家 (1900 – 1979)

To be continued ……

wangyong@ucas.ac.cn

http://people.ucas.ac.cn/~wangyong