数据预处理

目录
CONTENTS

Chapter 7.1

数据存在的问题

7.1 数据存在的问题

- 数据预处理是数据挖掘中的重要一环，而且必不可少。要更有效地挖掘出知识，就必须为其提供干净，准确，简洁的数据。
- 现实世界中数据常常是不完整，不一致的脏数据，无法直接进行数据挖掘，或挖掘结果差强人意。

1. 原始数据存在的问题

– **数据的不一致：各系统间的数据存在较大的不一致性**

如属性重量的单位：A数据库重量单位kg、B数据库重量单位g

– **噪声数据：数据中存在着错误或异常（偏离期望值）**，如：血压和身高为0就是明显的错误。

收集数据的时候难以得到精确的数据，主要原因：

- 收集数据的设备可能出现故障；
- 数据输入时可能出现错误；
- 数据传输过程中可能出现错误；
- 存储介质有可能出现损坏等。

1. 原始数据存在的问题

- **缺失值：由于实际系统设计时存在的缺陷以及使用过程中的一些人为因素，数据记录可能会出现数据值的丢失或不确定。**
- 原因可能有：
  - 有些属性的内容有时没有（家庭收入，参与销售事务数据中的顾客信息）；
  - 有些数据当时被认为是不必要的；
  - 由于误解或检测设备失灵导致相关数据没有记录下来；
  - 与其它记录内容不一致而被删除；
  - 忽略了历史数据或对数据的修改。

2. 数据质量要求

- **准确性：**数据记录的信息是否存在异常或错误。
- **完整性：**数据信息是否存在缺失。
- **一致性：**指数据是否遵循了统一的规范，数据集合是否保持了统一的格式。
- **时效性：**某些数据是否能及时更新。
- **可信性：**用户信赖的数据的数量。
- **可解释性：**指数据自身是否易于人们理解。

3. 数据预处理的主要任务

- **数据清理（清洗）：** 去掉数据中的噪声，纠正不一致。
- **数据集成：** 将多个数据源合并成一致的数据存储，构成一个完整的数据集，如数据仓库。
- **数据归约（消减）：** 通过聚集、删除冗余属性或聚类等方法来压缩数据。
- **数据变换（转换）：** 将一种格式的数据转换为另一格式的数据(如规范化)。

Chapter 7.2

数据清理

数据清理就是对数据进行重新审查和校验的过程。其目的在于**纠正存在的错误，并提供数据一致性。**

- 缺失值的处理；
- 噪声数据；
- 不一致数据。

1. 空缺值的处理

- **引起空缺值的原因**
  - 设备异常
  - 与其他已有数据不一致而被删除
  - 因为误解而没有被输入的数据
  - 在输入时，有些数据因为得不到重视而没有被输入
  - 对数据的改变没有进行日志记录
- **空缺值要经过推断而补上**

如何处理空缺值?

1）忽略元组：

- 若一条记录中有属性值被遗漏了，则将该记录排除在数据挖掘之外。
- 但是，当某类属性的空缺值权重很大时，直接忽略元组会使挖掘性能变得非常差。

2）忽略属性列：

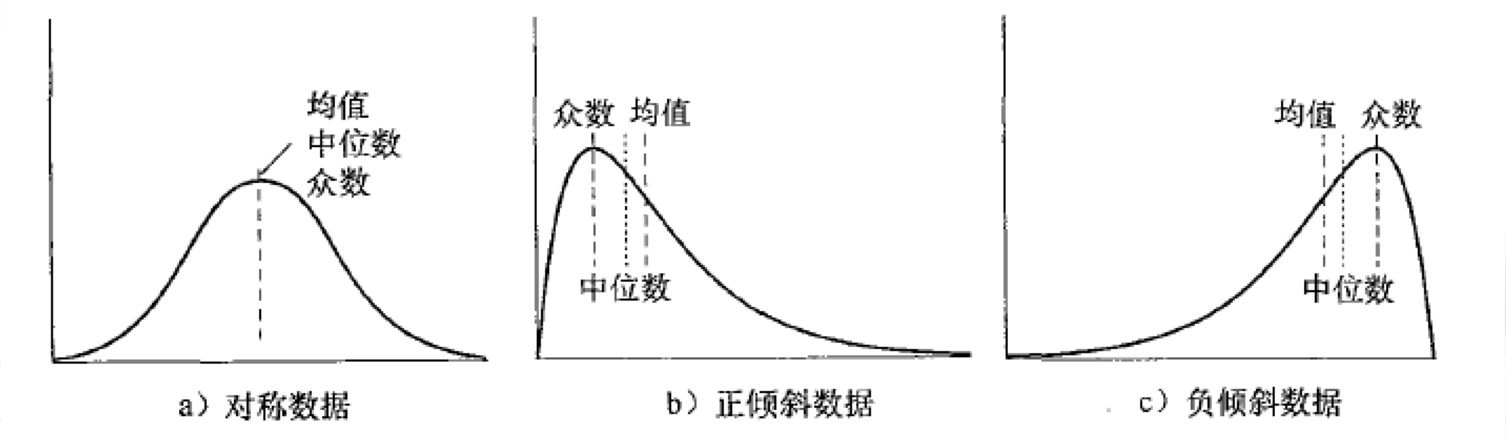
- 若某个属性的缺失值太多，则在整个数据集中可以忽略该属性。

3）人工填写空缺值：

- 工作量大，可行性低。

4）使用属性的中心度量值填充空缺值：

- 如果数据的分布是**正常**的，就可以使用**均值**来填充缺失值。
- 如果数据的分布是**倾斜**的，可以使用**中位数**来填充缺失值。



5）使用一个全局变量填充空缺值：

- 对一个所有属性的所有缺失值都使用一个固定的值来填补（如“Not sure”或∞）。

6）使用可能的特征值来替换空缺值（最常用）：

- 生成一个预测模型，来预测每个丢失值。
- 如可以利用回归、贝叶斯计算公式或判定树归纳确定，推断出该条记录特定属性最大可能的取值。

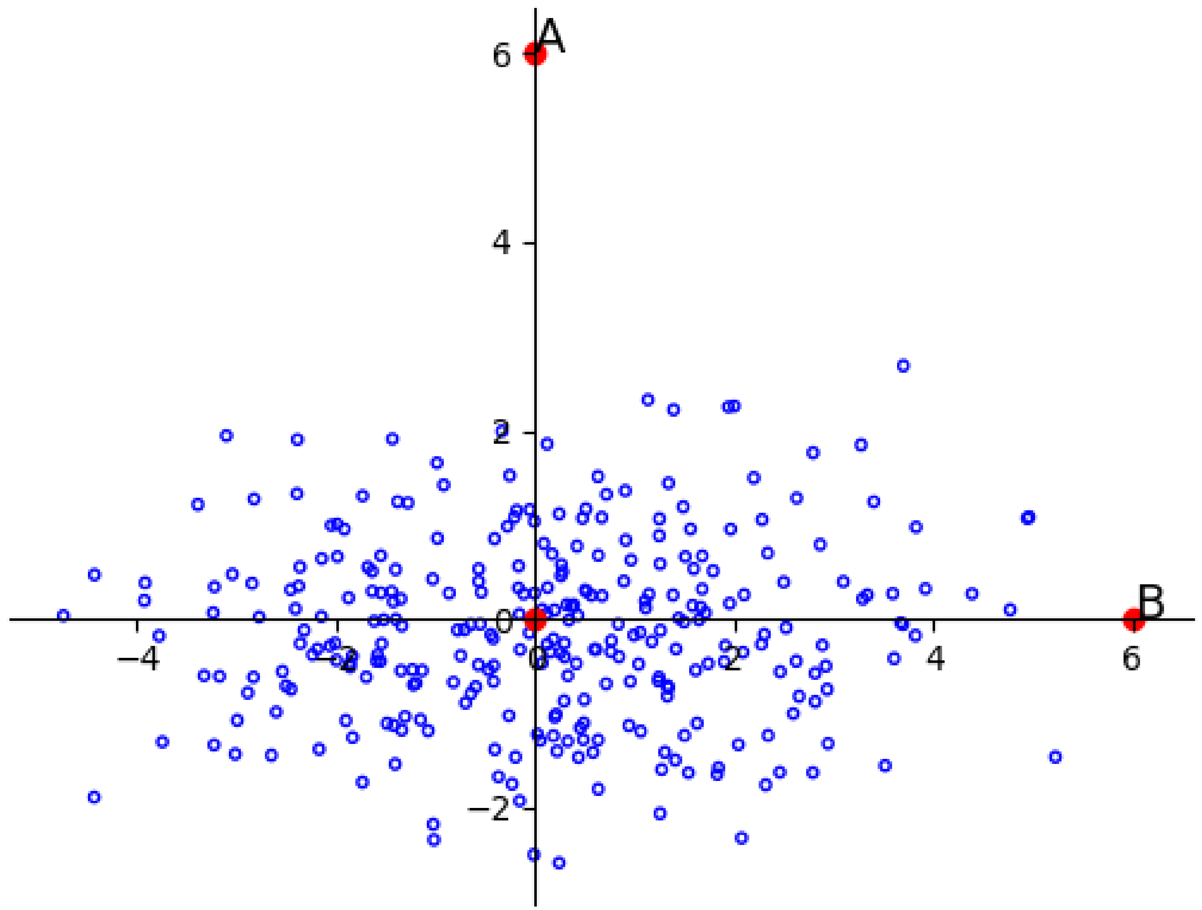
2. 噪声的处理

– **噪声(noise)：被测量的变量产生的随机错误或误差。**
  - 数据收集工具的问题
  - 数据输入错误
  - 数据传输错误
  - 技术限制
  - 命名规则的不一致

1）基于统计的技术-使用距离度量值（如马氏距离）来实现。

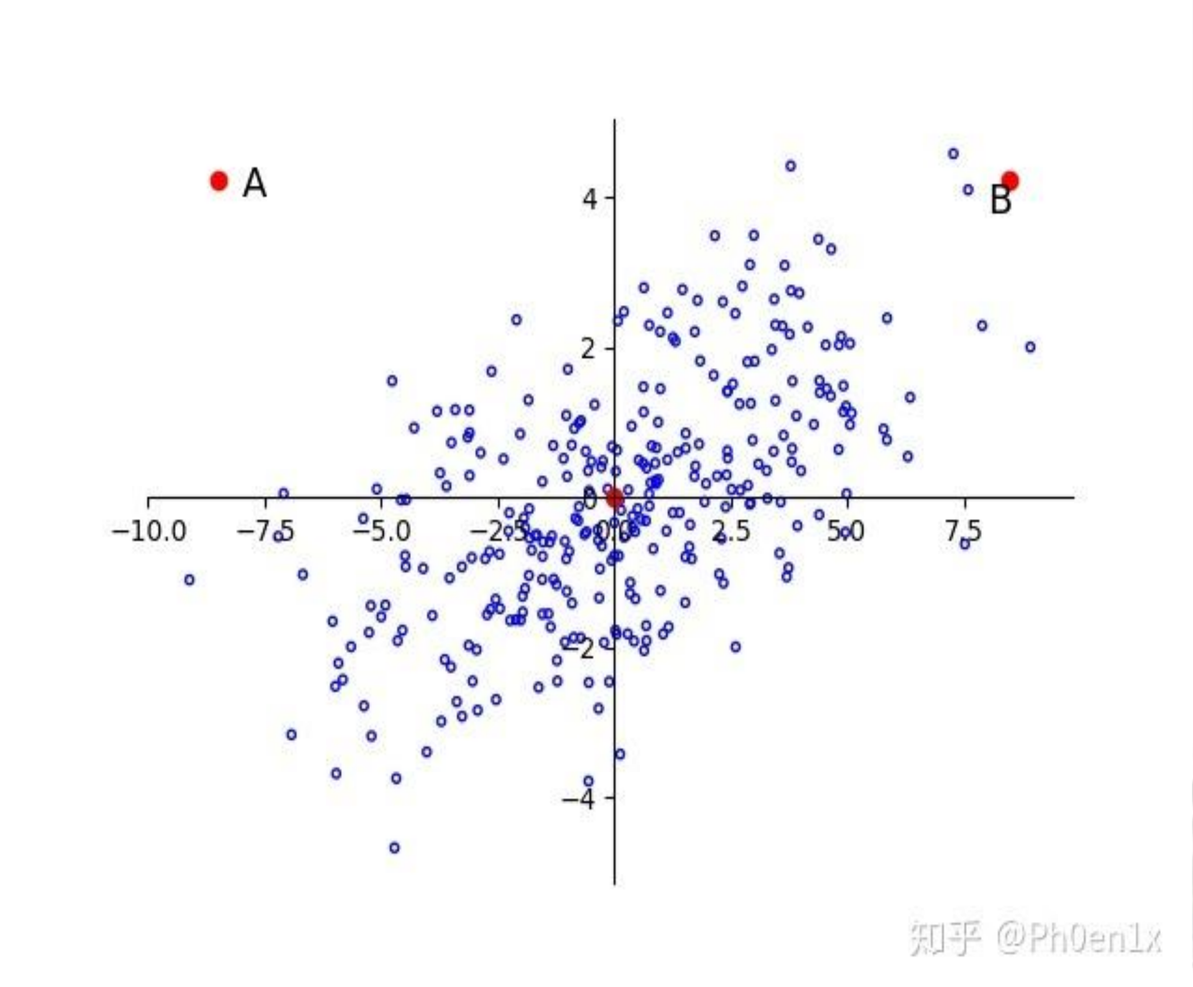
欧式距离问题

- **(1)** 在一个方差较小的维度下很小的差别就有可能成为离群点。就像右图一样，A与B相对于原点的欧式距离是相同的。
- 但是由于样本总体沿着横轴分布，所以B点更有可能是这个样本中的点，而A则更有可能是离群点。



(2)还有一个问题：如果维度间不独立同分布，样本点与欧氏距离近的样本点同类的概率更大吗?

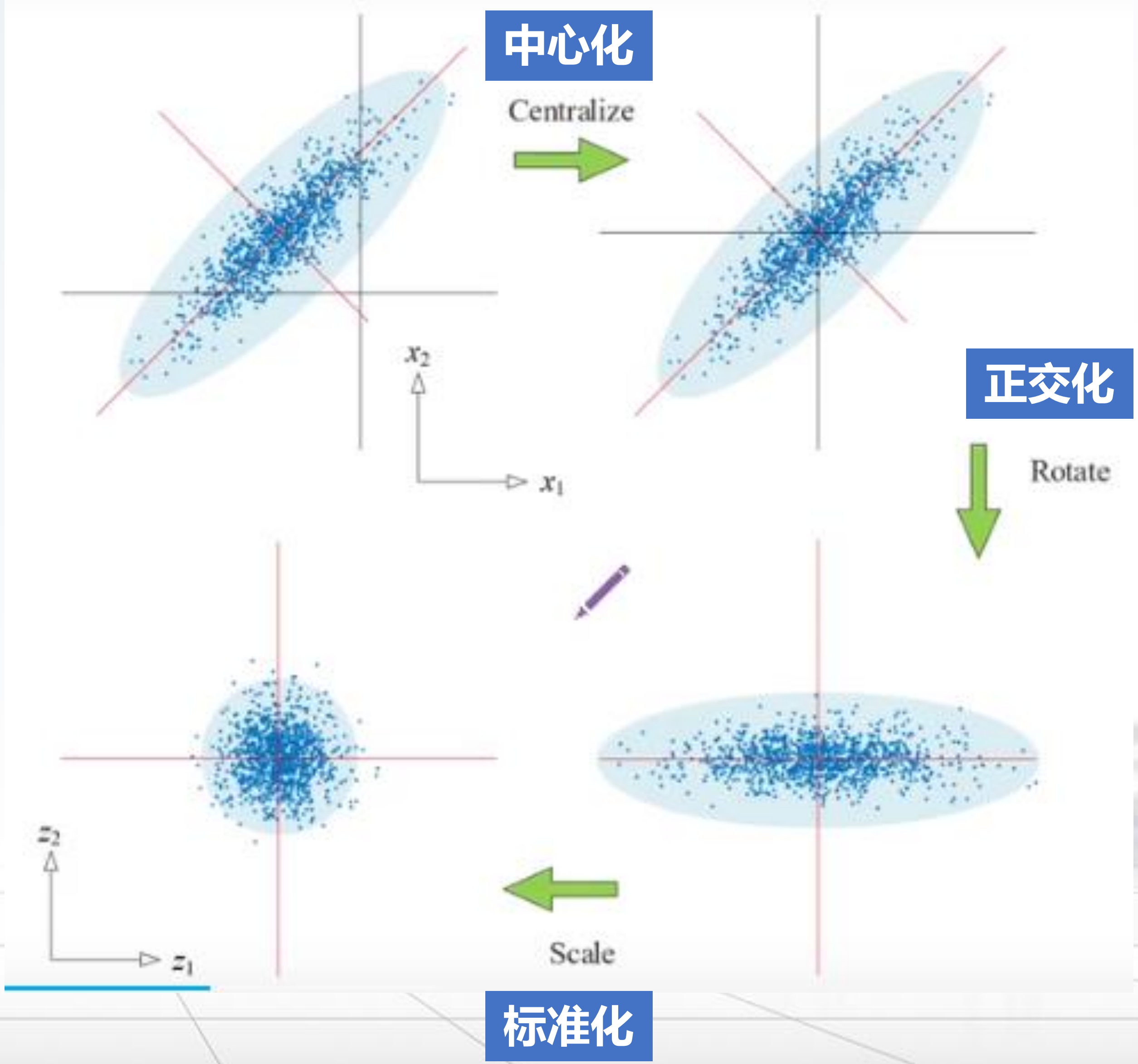
- 可以看到样本基本服从$f(x) = x$的线性分布，A与B相对于原点的距离依旧相等，显然A更像是一个离群点。
- 即使数据已经经过了标准化，也不会改变AB与原点间距离大小的相互关系。所以要本质上解决这个问题，就要针对主成分分析中的主成分来进行标准化。



马氏距离的几何意义-修正欧式距离：

将变量按照主成分进行**旋转**，让维度间**相互独立**，然后进行**标准化**，让维度同分布。

由主成分分析可知，由于主成分就是特征向量方向，每个方向的方差就是对应的特征值，所以只需要按照特征向量的方向旋转(独立)，然后缩放特征值倍即可(同分布)。



https://zhuanlan.zhihu.com/p/46626607

- 给定p维数据集中的n个观察值x_i（其中n>>p），用$\bar{x}_n$表示样本平均向量，V_n表示样本协方差矩阵，其中：

$$V_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)^T$$

- 每个多元数据点i(i=1,2,......,n)的马氏距离用M_i表示，为：

$$M_i = \left[\sum_{i=1}^{n}(x_i - \bar{x}_n)^T V_n^{-1}(x_i - \bar{x}_n)\right]^{\frac{1}{2}}$$

正交化:通过样本协方差矩阵,解决特征相关性问题
标准化:解决特征尺度不一致问题

马氏距离比欧氏距离好在哪里：

(1)欧式距离近就一定相似?

先举个比较常用的例子，身高和体重，这两个变量拥有不同的单位标准，也就是有不同的scale。比如身高用毫米计算，而体重用千克计算，显然差10mm的身高与差10kg的体重是完全不同的。但在普通的欧氏距离中，这将会算作相同的差距。

(2)归一化后欧氏距离近就一定相似?

当然我们可以先做归一化来消除这种维度间scale不同的问题，但是样本分布也会影响分类。举个一维的例子，现在有两个类别，统一单位，第一个类别均值为0，方差为0.1，第二个类别均值为5，方差为5。那么一个值为2的点属于第一类的概率大还是第二类的概率大？距离上说应该是第一类，但是直觉上显然是第二类，因为第一类不太可能到达2这个位置。

2）基于距离的技术

- 计算n维数据集中所有样本间的测量距离。
- 如果样本S中至少有一部分数量为p的样本到s_i的距离比d大，那么样本s_i就是数据集S中的一个噪声数据。

例：基于距离的噪声检测方法

给定一组三维样本S， $S=\{S_1,S_2,S_3,S_4,S_5,S_6\}=\{(1,2,0),(3,1,4),(2,1,5),(0,1,6),(2,4,3),(4,4,2)\}$

求在距离阈值d大于等于4，非邻点样本的阈值p大于等于3时的噪声数据。

- 首先，求数据集的欧几里得距离 $d=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2}$

表7.1 数据集S的距离表

	S_1	S_2	S_3	S_4	S_5	S_6
S_1		4.583	5.196	6.164	3.742	4.123
S_2			1.414	3.606	3.317	3.742
S_3				2.236	3.606	4.690
S_4					4.690	6.403
S_5						2.236

- 然后根据阈值距离d=4，计算出每个样本p的值，即距离大于等于d的样本数量

表7.2 S中每个样本的距离大于d的p点个数

	S_1	S_2	S_3	S_4	S_5	S_6
S_1		4.583	5.196	6.164	3.742	4.123
S_2			1.414	3.606	3.317	3.742
S_3				2.236	3.606	4.690
S_4					4.690	6.403
S_5						2.236

样本	p
S_1	4
S_2	1
S_3	2
S_4	3
S_5	1
S_6	3

3. 不一致数据的处理

- 数据的不一致性，就是指各类数据的矛盾性、不相容性。
- 数据库系统都会有一些相应的措施来解决并保护数据库的一致性。
  - 首先，我们需要确定数据库处于不一致状态的根本原因。可能的原因包括网络故障、硬件故障、软件错误、并发操作等。通过仔细分析日志和错误报告，可以帮助我们找到问题的源头。
  - 采取措施来恢复数据库到一致状态。以下是一些常见的方法：
  - 数据库备份和恢复。
  - 事务回滚：回滚到之前的一个稳定状态。
  - 数据修复工具。

Chapter 7.3

数据集成

- **数据集成**
  - 把不同来源、格式、特点和性质的数据合理地集中并**合并起来**。
  - 这些数据源可以是关系型数据库、数据立方体或一般文件。
- **它需要统一原始数据中的所有矛盾之处，如字段的：**
  - 同名异义；
  - 异名同义；
  - 单位不统一；
  - 字长不一致等。
- **集成过程中需要注意的问题：**
  - 集成的过程中涉及的实体识别问题；
  - 冗余问题。

1. 集成的过程中涉及的实体识别：

- **整合不同数据源中的元数据；**
- **进行实体识别：匹配来自不同数据源的现实世界的实体；**
  - 如：如何确定一个数据库中的brand和另一个数据库中的product是同一实体。
  - 通常，数据库的**数据字典**和数据仓库的**元数据**，可帮助避免模式集成中的错误。

2. 冗余问题

- 同一属性在不同的数据库或同一数据库的不同数据表中会有不同的字段名；
- 一个属性可以由另外的属性导出，如：一个顾客数据表中的平均月收入属性，可以根据月收入属性计算出来。

3. 检测冗余的方法

- **相关性分析**
  - 数值属性：采用相关系数和协方差进行相关性分析
  - 标称属性：采用χ^2（卡方）检验进行相关性分析

检测冗余的方法：

– **数值属性：**采用相关系数和协方差进行相关性分析

1）相关系数：

$$r_{X,Y}=\frac{\sum_{i=1}^{m}(x_i-\bar{X})(y_i-\bar{Y})}{m\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{m}(x_iy_i)-m\bar{X}\bar{Y}}{m\sigma_X\sigma_Y}$$

式中的m代表的是元组的个数， x_i是元组i在属性X上的值，y_i是元组i在属性Y上的值，$\bar{X}$表示X的均值，$\bar{Y}$表示Y的均值，σ_x表示X的标准差，σ_Y表示Y的标准差，$\sum_{i=1}^{m}(x_i,y_i)$表示每个元组中X的值乘Y的值。且$r_{X,Y}$的取值范围为$-1\le r_{X,Y}\le 1$。

- 如果$r_{X,Y}>0$，则X和Y是正相关的。
- 如果$r_{X,Y}=0$，则X和Y是独立的且互不相关。
- 如果$r_{X,Y}<0$，则X和Y是负相关的。

相关系数实例

例：数值属性的相关性分析。

表7.3 体重与血压表

	1	2	3	4	5	6	7	8	9	10	11	12
体重	68	48	56	60	83	56	62	59	77	58	75	64
血压	95	98	87	96	110	155	135	128	113	168	120	115

表7.4 体重和血压的均值和标准差值

	均值	标准差
体重	67. 83	10. 14
血压	118. 33	24. 74

$$r_{X,Y}=\frac{\sum_{i=1}^{m}(x_i-\bar{X})(y_i-\bar{Y})}{m\sigma_X\sigma_Y}=-0.112$$

– **数值属性：** 采用协方差进行相关性分析

2）协方差： 设有两个属性X和Y，以及有m次观测值的集合

$$\{(x_1,y_1),(x_2,y_2),\dots\dots,(x_m,y_m)\}$$

$$Cov(X,Y)=E\big((X-\bar{X})(Y-\bar{Y})\big)=\frac{\sum_{i=1}^{m}(x_i-\bar{X})(y_i-\bar{Y})}{m}=E(XY)-E(X)E(Y)$$

其中：

$$E(X)=\bar{X}=\frac{\sum_{i=1}^{m}x_i}{m} \qquad E(Y)=\bar{Y}=\frac{\sum_{i=1}^{m}y_i}{m}$$

$$r_{X,Y}=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

– **数值属性：**采用协方差进行相关性分析

2）协方差：

$$Cov(X,Y) = E\left((X-\bar{X})(Y-\bar{Y})\right) = \frac{\sum_{i=1}^{m}(x_i-\bar{X})(y_i-\bar{Y})}{m} = E(XY)-E(X)E(Y)$$

当Cov(X,Y)>0时，表明X与Y正相关；

当Cov(X,Y)<0时，表明X与Y负相关；

当Cov(X,Y)=0时，表明X与Y不相关。

- **数值属性：**采用协方差进行相关性分析

假设属性X和Y是相互独立的，有

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

协方差的公式是

$$Cov(X,Y) = E(X \cdot Y) - \bar{X}\bar{Y} = E(X) \cdot E(Y) - \bar{X}\bar{Y} = 0。$$

但是，它的逆命题是不成立的。

– 协方差实例

例：数值属性的协方差计算。

求上例中血压是否会随着体重一起变化。

$$E(X)=\frac{68+48+56+60+83+56+62+59+77+58+75+64}{12}=63.83$$

$$E(Y)=\frac{95+98+87+96+110+155+135+128+113+168+120+115}{12}=118.33$$

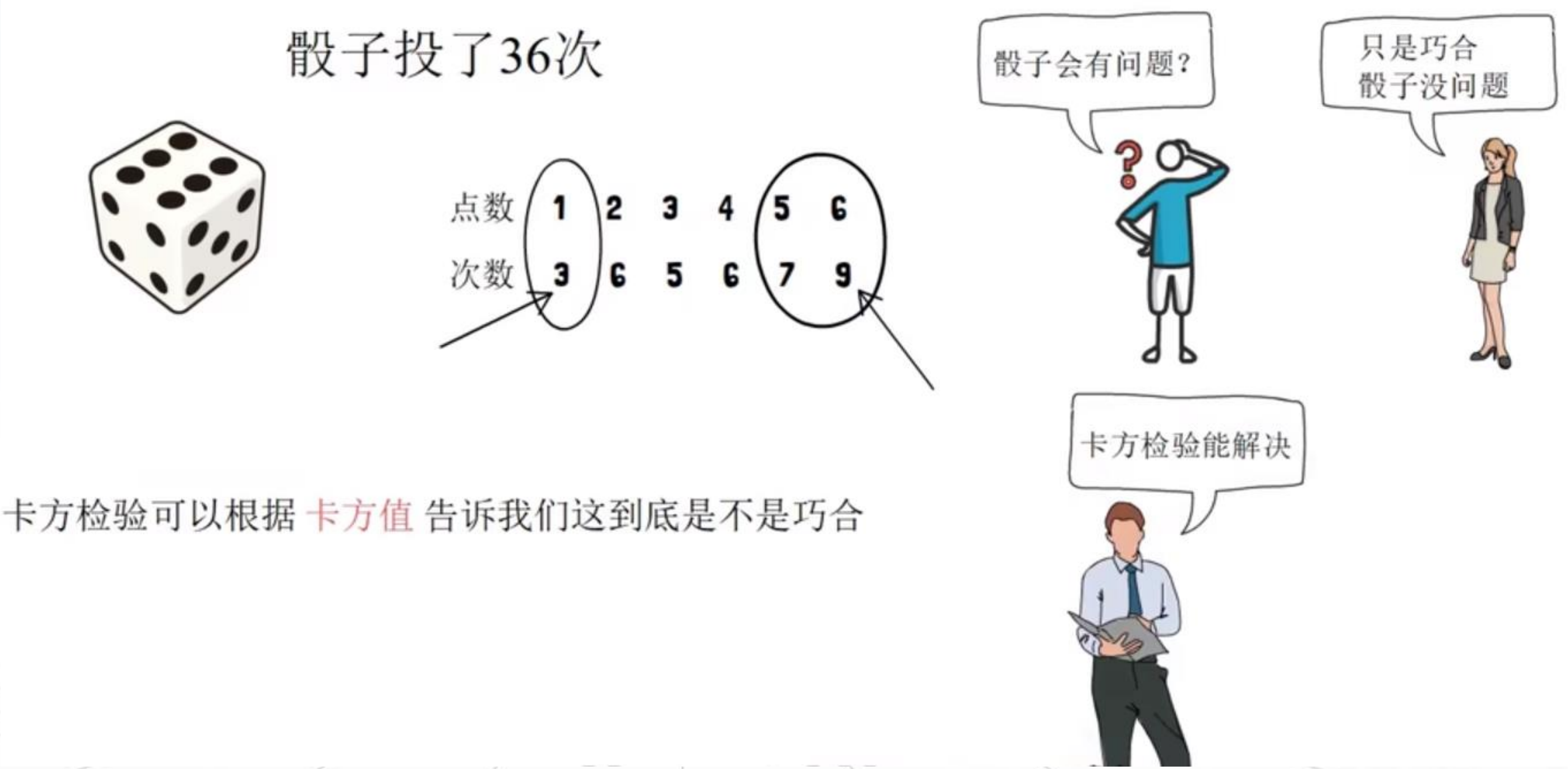
$$Cov(X,Y)=r_{X,Y}\cdot\sigma_X\cdot\sigma_Y=-0.112\times10.14\times24.74=-28.10$$

负相关

– **标称属性：**使用卡方检验进行相关性分析

卡方检验主要用于检测观察到的类别变量的分布是否与期望的不同。

举例:



– 标称属性：使用卡方检验进行相关性分析

卡方检验主要用于检测观察到的类别变量的分布是否与期望的不同。

第一步：提出假设

零假设：期望值 和 观测值之间没有显著差异

证明假设成立可能性特别低

就能够说明这个假设是不合理的

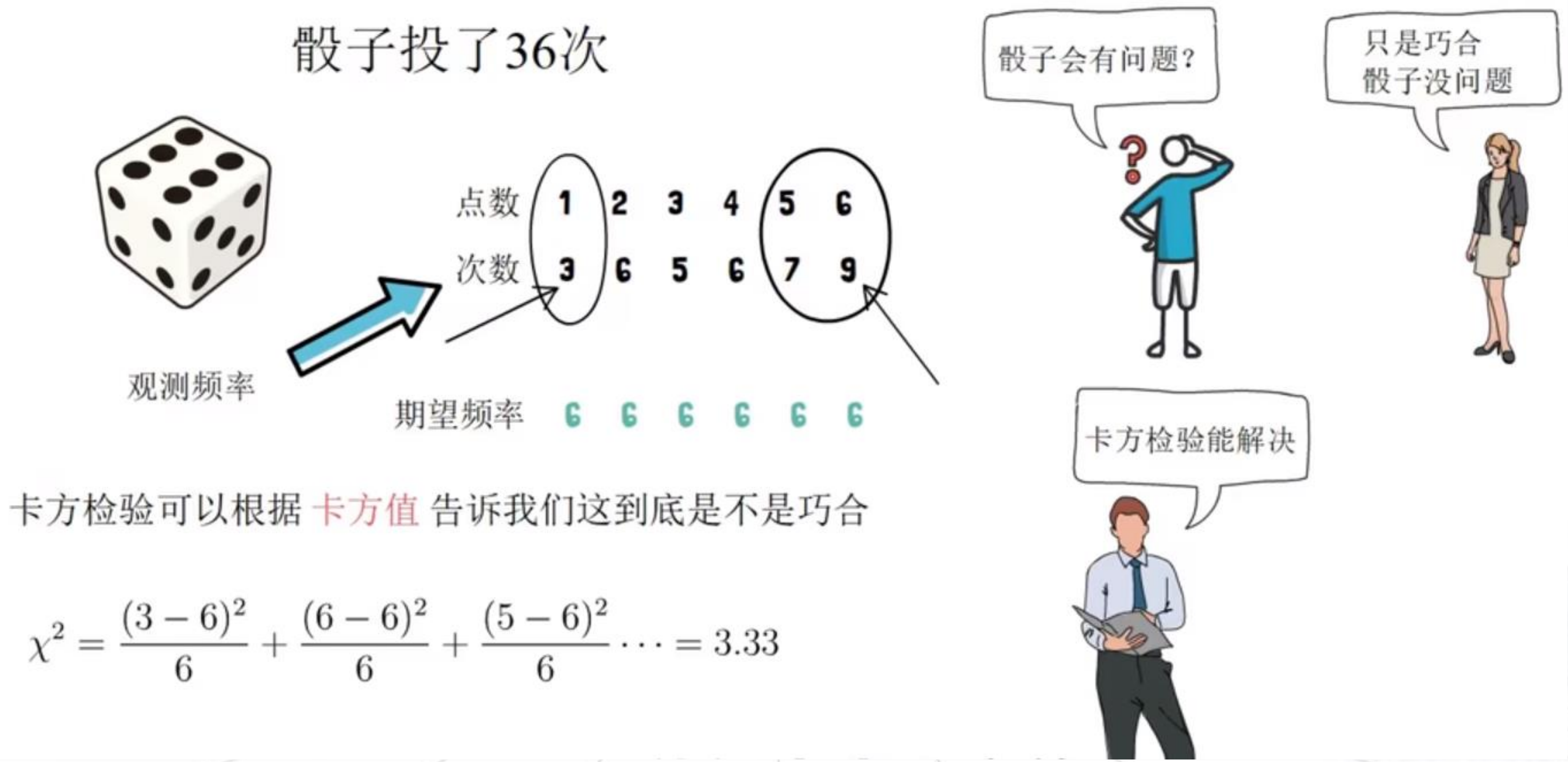
因此拒绝这个假设

α (显著性水平) = 0.05 ⇒ p值小于0.05时，则拒绝我们的假设

认为期望值和观测值之间是存在显著差异的。

– **标称属性：**使用卡方检验进行相关性分析

第二步,计算卡方值



卡方值多大合适-临界值-查表。

– 标称属性：使用卡方检验进行相关性分析

卡方临界值 (χ2 critical value)

$\alpha = 0.05$

$$df = k - 1$$

组数

硬币有六个面，因此自由度为

6 − 1 = 5

临界值 = 11.070

然而，我们的卡方值为 3.33 < 11.07

接受零假设 ⟶ 观测值和期望值之间没有显著差异

骰子没有问题

显著性水平

自由度

Degrees of freedom (df)	.99	.975	.95	.9	.1	.05	.025	.01
				Significance level (α)				
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	[illegible]	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	[illegible]	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980

– **标称属性：**使用卡方检验进行相关性分析

列联表

Y \ X	x_1	x_2	…	x_i	…	x_n	sum
y_1	O_{11}	O_{12}	…	O_{1i}	…	O_{1n}	$O_{1.}$
y_2	O_{21}	O_{22}	…	O_{2i}	…	O_{2n}	$O_{2.}$
…	…	…	…	…	…	…	…
y_j	O_{j1}	O_{j2}	…	O_{ji}	…	O_{jn}	$O_{j.}$
…	…	…	…	…	…	…	…
y_r	O_{r1}	O_{r2}	…	O_{ri}	…	O_{rn}	$O_{r.}$
sum	$O_{.1}$	$O_{.2}$	…	$O_{.i}$	…	$O_{.n}$	m

$$O_{i.} = count(Y = y_i)；\ O_{.j} = count(X = x_j)$$

o_{ij}是联合事件(x_i, y_j)的观测频度(实际计数),m是总的频度。

– **标称属性：**使用卡方检验进行相关性分析

独立性检验的步骤如下：

（1）统计假设：H_0：属性X和属性Y之间是独立的

（2）期望频数的计算，计算公式如式所示。

$$e_{ij}=\frac{(O_{i.}\times O_{.j})}{m}=\frac{count(X=x_i)\times count(Y=y_j)}{m}$$

（3）自由度的确定

$$df=(r-1)\times(n-1)$$

r和n是检验条件的分类数，即行数和列数

– **标称属性：**使用卡方检验进行相关性分析

（4）***Pearson*** χ^2**统计量**的计算：

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

$$e_{ij} = \frac{count(X = x_i) \times count(Y = y_j)}{m}$$

（5）统计推断

- $\chi^2 >$ 临界值（具有自由度df和显著水平α）：拒绝假设H_0
- $\chi^2 <$ 临界值（具有自由度df和显著水平α）：接受假设H_0

– **卡方检验实例：**二分类情况

例：对从事两种工种的某一年龄段男性患某种疾病的情况进行调查，如下表。分析某一年龄段男性患某种疾病与从事工种是否相关。

表 四方格列联表

患病情况 / 从事工种	患病	不患病	合计
工种1	386	895	1281
工种2	65	322	387
合计	451	1217	1668

– 卡方检验实例：二分类情况

（1）统计假设：

H0：某一年龄段男性患某种疾病与从事工种不相关

（2）期望频数的计算。

表 四方格列联表（期望频数）

患病情况 / 从业情况	患病	不患病	合计
工种1	386（**346. 36**）	895（934. 64）	1281
工种2	65（104. 64）	322（282. 36）	387
合计	451	1217	1668

期望频数： e_{11} =（451×1281）÷1668=346. 36

– 卡方检验实例：二分类情况

（3）自由度的确定：df=（2-1）×（2-1）=1

（4）卡方统计量的计算

表 四方格列联表（期望频数）

从业情况 \ 患病情况	患病	不患病	合计
工业	386（346.36）	895（934.64）	1281
农业	65（104.64）	322（282.36）	387
合计	451	1217	1668

$$\chi^2=\frac{(386-346.36)^2}{346.36}+\frac{(895-934.64)^2}{934.64}+\frac{(65-104.64)^2}{104.64}+\frac{(322-282.36)^2}{282.36}=26.80$$

– 卡方检验实例：二分类情况

（5）统计判断

表 卡方检验**临界值表**（部分）

显著水平α / 自由度	0.99	0.98	0.95	0.90	0.50	0.10	0.05	0.02	0.01	0.005
1	0.000	0.001	0.004	0.016	0.045	2.71	**3.84**	5.41	6.64	10.83
2	0.020	0.040	0.103	0.211	1.36	4.61	5.99	7.82	9.21	17.82
3	0.115	0.185	0.352	0.584	2.366	6.25	7.82	9.84	11.34	16.27

卡方值大于临界值,拒绝假设

Chapter 7.4

数据归约

- 对大规模数据库内容进行复杂的数据分析常需要**消耗大量的时间**，使得这样的分析变得不现实和不可行；
- **数据归约（data reduction）**：数据消减或约简，是在不影响最终挖掘结果的前提下，缩小所挖掘数据的规模；
- 数据归约技术可以用来得到数据集的归约表示，它小得多，但仍接近保持原数据的**完整性**；
- 对归约后的数据集进行挖掘可提高**挖掘的效率**，并产生相同（或几乎相同）的结果。

– **数据归约的标准：**

- 用于数据归约的时间不应当超过或“抵消”在归约后的数据集上挖掘节省的时间。
- 归约得到的数据比原数据小得多，但可以产生相同或几乎相同的分析结果。

– 数量归约：直方图

- 直方图（Histogram）是一种常见的数据归约的形式。属性X的直方图将X的数据分布划分为不相交的**子集或桶**。通常情况下，子集或桶表示给定属性的一个**连续区间**。单值桶表示每个桶只代表单个属性值/频率对（单值桶对于存放那些高频率的离群点，非常有效。）

- 划分桶和属性值的方法有两种：

①**等宽**：在等宽直方图中，每个桶的宽度区间是一致的。

②**等频（或等深）**：在等频直方图中，每个桶的频率粗略地计为常数，即每个桶大致包含相同个数的邻近数据样本。

例：用直方图表示数据

已知某人在不同时刻下所量血压值为：95，98，87，96，110，155，135，128，113，168，120，115，110，155，135，128，113，158，87，96，110，98，87，94，80，93，89，95，99，101，111，123，128，113，158，128，113，168，87，96，110。

使用等宽直方图表示数据，如图所示。由于需要继续压缩数据，所以一般都是使用桶来表示某个属性的一个连续值域。

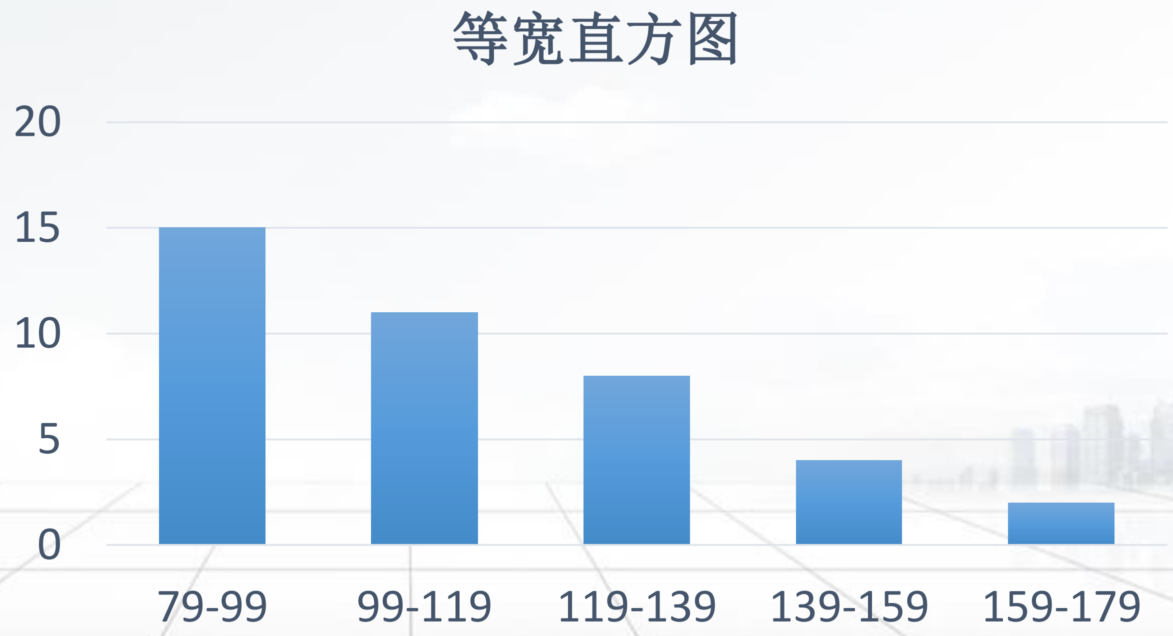


图7-2 等宽直方图

– **数量归约：数据立方体**

- 数据立方体是一类多维矩阵，可以使用户从多个角度探索和分析数据集，它的数据是已经处理过的，并且聚合成了立方形式。
- 数据立方体的基本概念。

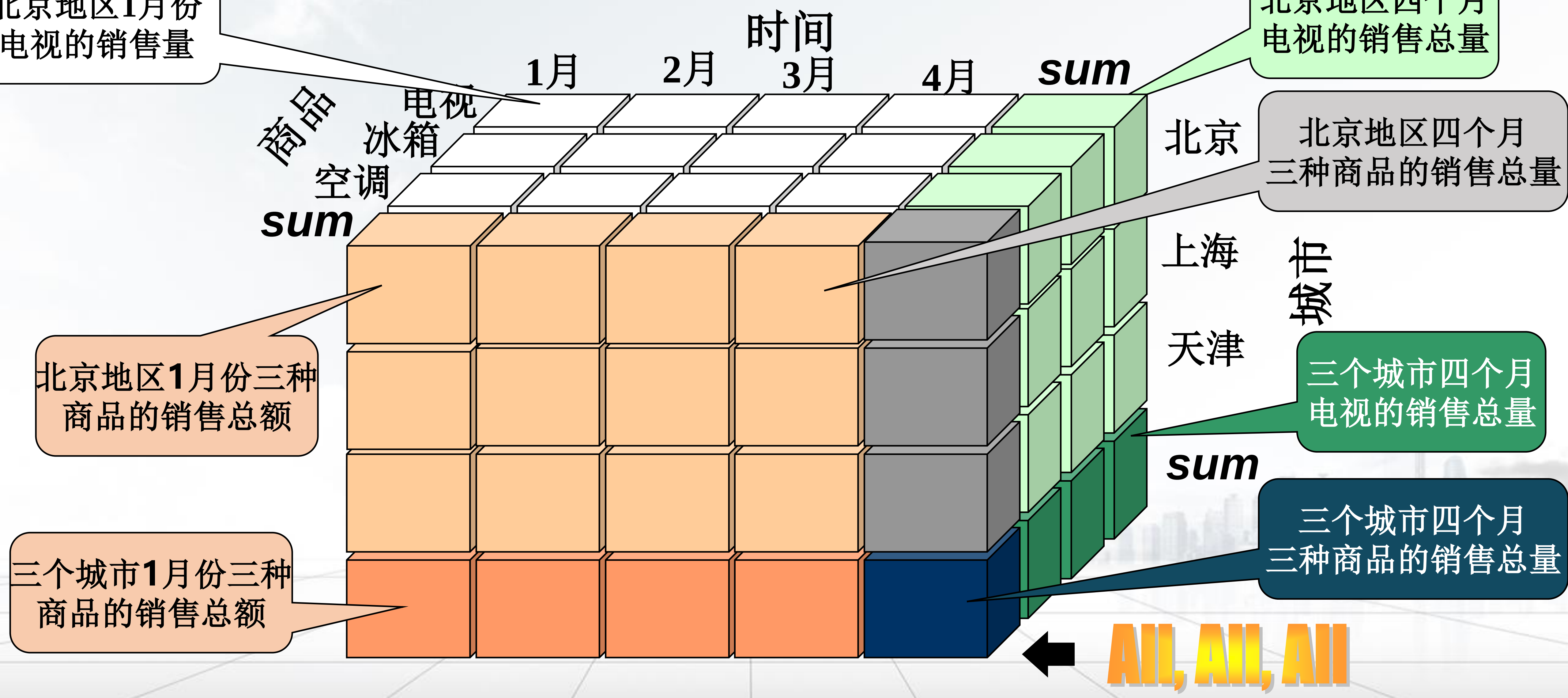
①**方体**：不同层创建的数据立方体。

②**基本方体**：最低抽象层创建的立方体。

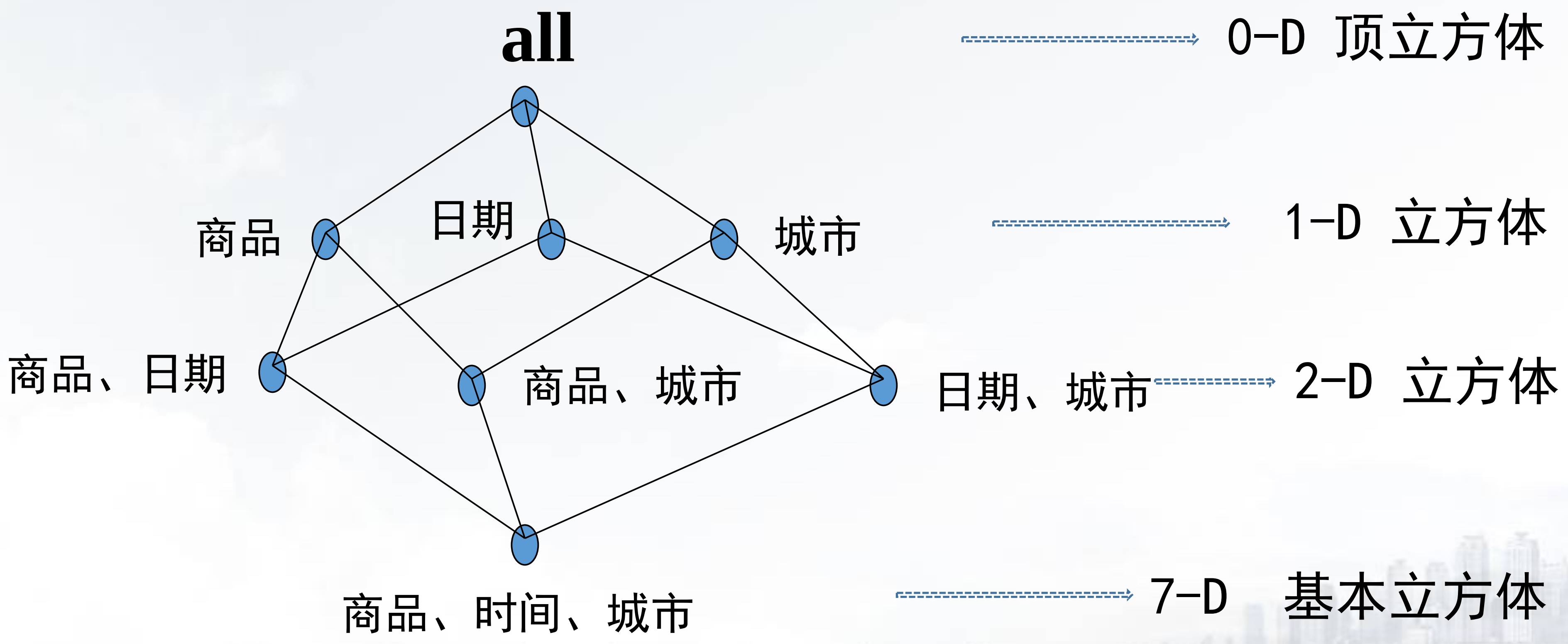
③**顶方体**：最高层抽象的立方体。

④**方体的格**：每一个数据立方体。

数量归约：数据立方体


北京地区1月份电视的销售量
时间
1月
2月
3月
4月
sum
商品
电视
冰箱
空调
sum
北京地区四个月电视的销售总量
北京
北京地区四个月三种商品的销售总量
上海
城市
天津
三个城市四个月电视的销售总量
sum
三个城市四个月三种商品的销售总量
北京地区1月份三种商品的销售总额
三个城市1月份三种商品的销售总额
All, All, All

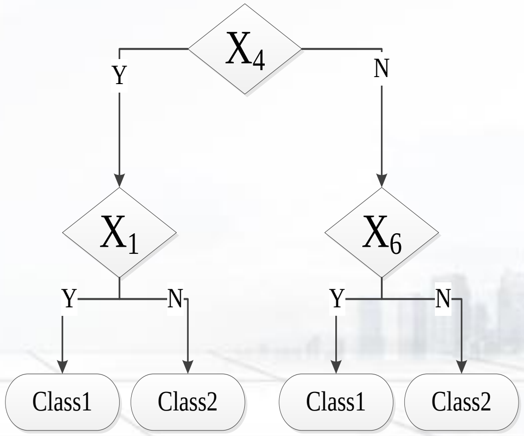
–数量归约：数据立方体——方体的格



– 数据归约—属性子集选择：检测并删除不相关、弱相关或冗余的属性。

属性子集选择的基本启发式方法包括逐步向前选择、逐步向后删除、逐步向前选择和逐步向后删除的组合以及决策树归纳，表7.7给出了属性子集选择方法。

表7.7 属性子集选择方法

向前选择	向后删除	决策树归纳
初始属性集： $\{X_1,X_2,X_3,X_4,X_5,X_6\}$	初始属性集： $\{X_1,X_2,X_3,X_4,X_5,X_6\}$	初始属性集： $\{X_1,X_2,X_3,X_4,X_5,X_6\}$
初始化归约集： $\{\}$ $\Rightarrow\{X_1\}$ $\Rightarrow\{X_1,X_4\}$ $\Rightarrow\{X_1,X_4,X_6\}$	$\Rightarrow\{X_1,X_2,X_3,X_4,X_5,X_6\}$ $\Rightarrow\{X_1,X_3,X_4,X_5,X_6\}$ $\Rightarrow\{X_1,X_4,X_5,X_6\}$ $\Rightarrow\{X_1,X_4,X_6\}$	
归约后的属性集： $\{X_1,X_4,X_6\}$	归约后的属性集： $\{X_1,X_4,X_6\}$	归约后的属性集： $\{X_1,X_4,X_6\}$

– 数据归约—取样（抽样）

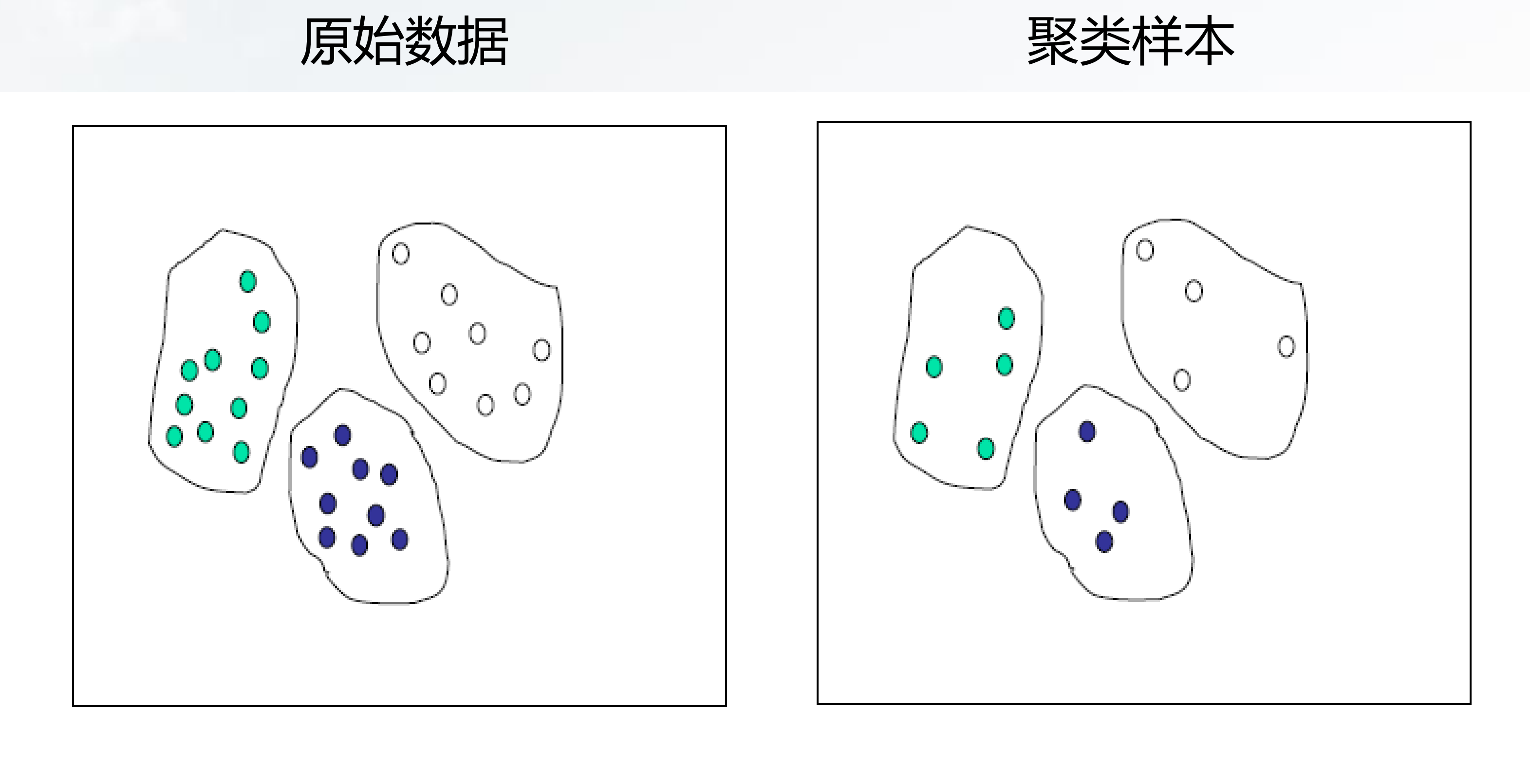
- 允许用数据的较小随机样本（子集）表示大的数据集。
- 取样方法：
  - **不放回简单随机取样** （Simple Random Sampling Without Replacement, SRSWOR）
  - **放回简单随机取样**（Simple Random Sampling With Replacement, SRSWR）
  - **聚类取样**（Clustered Sampling）
  - **分层取样**（Stratified Sampling）

– 数据归约—无放回简单随机取样& 放回简单随机取样

- 假定大型数据集D包含N个元组
- 无放回的简单随机抽样方法，从N个元组中随机（每一数据行被选中的概率为$\frac{1}{N}$）抽取出n个元组，以构成抽样数据子集。
- 有放回的简单随机抽样方法，与无放回简单随机抽样方法类似，也是从N个元组中每次抽取一个元组，但是抽中的元组接着放回原来的数据集D中，以构成抽样数据子集。这种方法可能会产生相同的元组。

– 数量规约-聚类采样：

首先将大数据集D划分为M个**互不相交的聚类**，然后再从M个类中的数据对象分别进行随机抽取，可最终获得聚类采样的数据子集。



– 数量规约-分层取样：

- 首先将大数据集D划分为**互不相交的层**，然后对每一层简单随机选样得到D的分层选样。
- 如，根据顾客的年龄组进行分层，然后再在每个年龄组中进行随机选样，从而确保了最终获得分层采样数据子集中的年龄分布具有代表性。

Chapter 7.5

数据变换与数据离散化

数据变换： 将数据转换成适合数据挖掘的形式

- **平滑：** 去掉数据中的噪声，将连续的数据离散化
  - 分箱
  - 回归
  - 聚类
- **聚集：** 对数据进行汇总和聚集
  - avg(), count(), sum(), min(), max(),…
  - 如，每天销售额（数据）可以进行聚集操作以获得每月或每年的总额
  - 可用来构造数据立方体

数据变换：将数据转换成适合数据挖掘的形式

- **数据泛化：**使用概念分层，用更抽象（更高层次）的概念来取代低层次或数据层的数据对象。
  - 如：街道属性，可以泛化到更高层次的概念，如城市、国家；
  - 同样，对于数值型的属性，如年龄属性，可以映射到更高层次的概念，如年轻、中年和老年。

- **规范化：**把属性数据按比例缩放，使之落入一个特定的小区间
- **属性构造：**通过已知的属性构建出新的属性，然后放入属性集中，有助于挖掘过程。
- **离散化：**数值属性的原始值用区间标签或概念标签替换。

1、数据变换：数据泛化——概念分层

- 概念分层定义了一组由低层概念到高层概念集的**映射**。允许在各种**抽象级**别上处理数据，从而在多个**抽象层**上发现知识。
- 用较高层概念替换低层次的概念，以此来**减少取值个数**。
- 概念分层结构可以用**树**来表示，树的每个节点代表一个概念。

例7.5 根据每个属性的不同值的个数产生概念分层。

服装类的级别可以分为男装和女装，然后接下去可以分为上装和下装。

服装的概念分层可以自动产生，如图7.3所示。

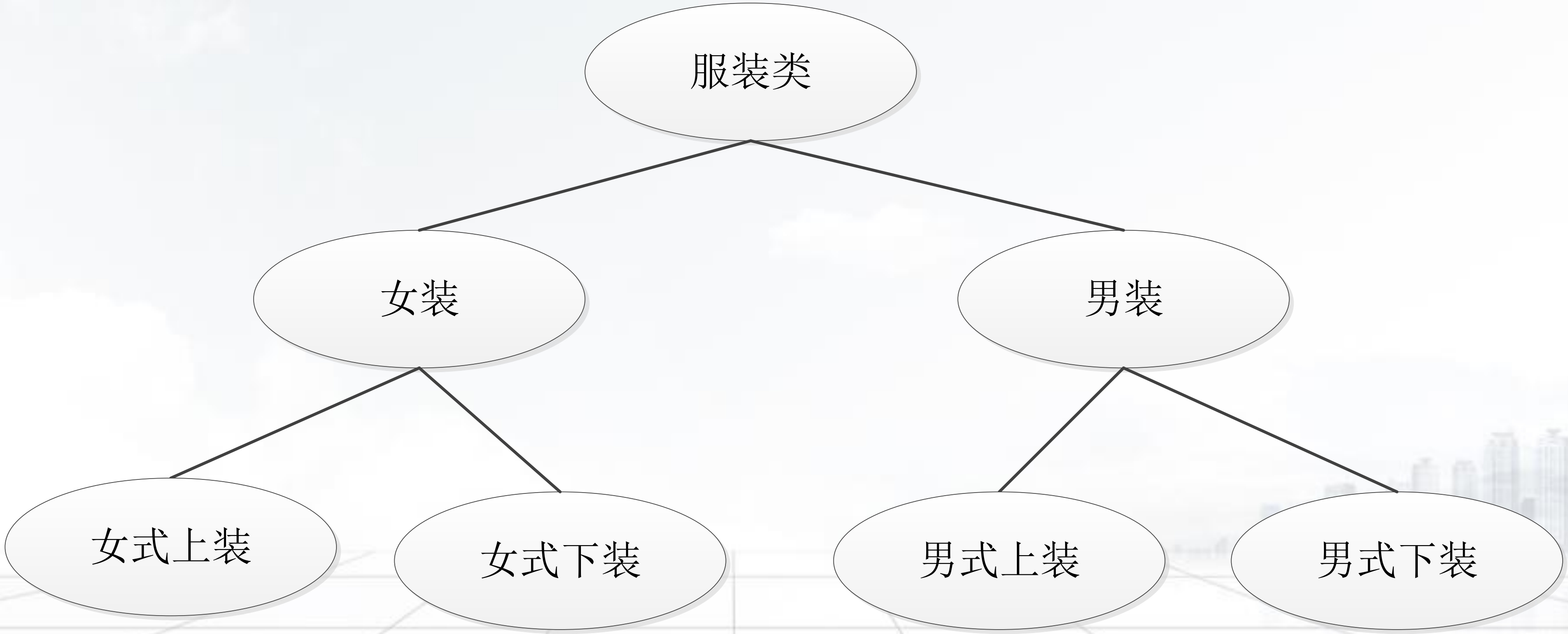


图7.3 服装的概念分层

2、数据变换：规范化

将数据按比例进行缩放，使之落入一个特定的区域，以消除数值型属性因大小不一而造成的挖掘结果的偏差。

如将工资收入属性值映射到[-1.0, 1.0]的范围内

– **规范化的目的：**

- 将一个属性取值范围影射到一个特定范围之内，以消除数值性属性因大小不一而造成挖掘结果的偏差。

2、数据变换：规范化

– **常用的方法：**

- 小数定标规范化；
- 最小-最大规范化；
- 零-均值规范化（z-score规范化）。

小数定标规范化：

通过移属性值的小数点的位置进行规范化，通俗地说就是**将属性值除以10的*j*次幂**，使其值落在-1到1的范围内。属性A的值v_i被规范化为v_i'

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

例如：假设属性A的取值范围为[-986,917]，用1000（即j=3）去除每个值，这样-986被规范化为0.986.

– 最小—最大规范化：

- 假定minA和maxA分别为属性A的最小和最大值，则将A的值映射到区间[a, b]中：$v_i' = \frac{v_i - \min A}{\max A - \min A}(b-a)+a$
- 其中：v_i表示对象i的原属性值，v_i'表示规范化的属性值， a为规范化后的最小值，b为规范化后的最大值。

例：假定某公司员工的最大年龄为52岁，最小年龄为21岁，请将年龄映射到区间[0.0,1.0]的范围内：

根据最小-最大值规范化，44岁将变换为：$\frac{44-21}{52-21}(1.0-0)+0 \approx 0.742$

- **z-score规范化（零均值规范化）：**
  - 常用于属性最大值与最小值未知，或使用最小最大规范化方法会出现异常数据的情况。
  - 将属性A的值根据其平均值和标准差进行规范化：$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$

其中v_i表示对象 的原属性值， v_i'表示规范化的属性值，$\bar{A}$表示属性A的平均值，σ_A 表示属性A的标准差。

例：某公司员工年龄的平均值和标准差分别为25岁和11岁。请根据z-score规范化，将44岁这个数据规范化。(44-25）/11≈1.727

3、数据变换：属性构造

- 利用已有属性集构造出新的属性，并加入到现有属性集中以帮助挖掘更深层次的模式知识，提高挖掘结果的准确性；
- 如，根据宽、高属性，可以构造一个新属性：面积。

4、数据变换：离散化

- 连续变量的离散化，就是将具体性的问题抽象为概括性的问题，即是将它取值的连续区间划分为小的区间，再将每个小区间重新定义为一个唯一的取值。
- 数据离散化的基本方法主要有**分箱法**和**直方图分析法**。

对连续变量进行离散化处理，一般经过以下步骤：

①对此变量进行排序。

②选择某个点作为候选断点，根据给定的要求，判断此断点是否满足要求。

③若候选断点满足离散化的要求，则对数据集进行分裂或合并，再选择下一个候选断点。

④重复步骤②和③，如果满足停止准则，则不再进行离散化过程，从而得到最终的离散结果。

数据离散化—分箱

- 首先**排序**数据，并将它们分到等深（等宽）的箱中；
- 然后可以按箱的**平均值**、或中值或者边界值等进行平滑。
  - 按箱的**平均值**平滑：箱中每一个值被箱中的平均值替换；
  - 按箱的**中值**平滑：箱中的每一个值被箱中的中值替换；
  - 按箱的**边界**平滑：箱中的最大和最小值被视为箱边界，箱中的每一个值被最近的边界值替换。

① 等深分箱:

·按记录数进行分箱，每箱具有相同的记录数，每箱的记录数称为箱的权重，也称箱子的深度。

例7.6 分箱法。
某公司存储员工信息的数据库里表示收入的字段“income”排序后的值（人民币元）：900，1000，1300，1600，1600，1900，2000，2400，2600，2900，3000，3600，4000，4600，4900，5000，请按照等深分箱法分箱。

设定权重（箱子深度）为4，分箱后
箱1：900，1000，1300，1600
箱2：1600，1900，2000，2400
箱3：2600，2900，3000，3600
箱4：4000，4600，4900，5000

用平均值平滑结果为：
箱1：1200，1200，1200，1200
箱2：1975，1975，1975，1975
箱3：3025，3025，3025，3025
箱4：4625，4625，4625，4625

②等宽分箱 (binning):

- 在整个属性值的区间上平均分布，即每个箱的区间范围设定为一个常量，称为箱子的宽度。

上例中设定区间范围（箱子宽度）为1000元人民币，按等宽分箱法分箱后

箱1：900，1000，1300，1600，1600，1900

箱2：2000，2400，2600，2900，3000

箱3：3600，4000，4600

箱4：4900，5000

用**平均值平滑**结果为：

箱1：1383，1383，1383，1383，1383，1383

箱2：2580，2580，2580，2580，2580

箱3：4067，4067，4067

箱4：4950，4950

数据离散化—直方图分析法:

- 直方图也可以用于数据离散化。它能够递归的用于每一部分，可以自动产生多级概念分层，直到满足用户需求的层次水平后结束。
- 例如，图7-8是某数据集的分布直方图，被划分成了范围相等的区间（79~99，99~119，……，159~179）。这就产生了多级概念分层。