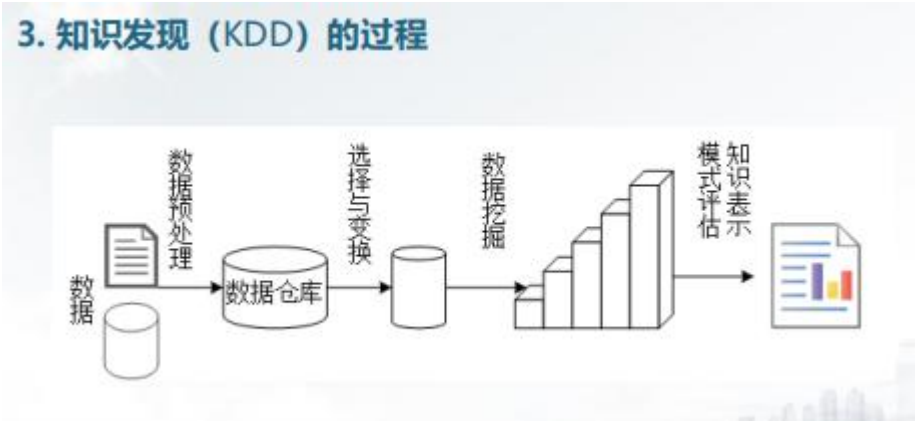
1. 第一章

1.1

数据分析是指采用适当的统计分析方法对收集到的数据进行分析、概括和总结，对数据进行恰当地描述，提取出有用的信息的过程。

数据挖掘(Data Mining，DM)是指从海量的数据中通过相关的算法来发现隐藏在数据中的规律和知识的过程。

通常将数据挖掘视为数据中“知识发现”的同义词，也可以认为数据挖掘是知识发现中的一个步骤。



内容	数据分析	数据挖掘
处理的数据量	不一定很大	海量
目标	比较明确	不明确的
侧重点	展现数据之间的关系	对未知的情况进行预测和估计

1.2

数据库系统（DataBase System，DBS）由一组内部相关的数据（称作数据库）和用于管理这些数据的程序组成，通过软件程序对数据进行高效的存储和管理。

数据仓库（Data Warehouse，DW）是一个**面向主题的、集成的、时变的、非易失的**数据集合，支持管理者决策过程。

事务数据库的每个记录代表一个事务

数据矩阵中的数据对象的所有属性都是具有相同性质的数值型数据。

1.3

频繁模式：数据中频繁出现的模式

频繁项集：频繁在事务数据集中一起出现的商品集合

分类与标签预测是找出描述和区分数据类或概念的模型或函数，以便能够

使用模型预测类标号未知的对象的类标号

分类预测类别（离散的、无序的）标号，**回归**建立连续值函数模型，也就是用来预测缺失的或难以获得的数值数据值。

聚类就是把一些对象划分为多个组或者“聚簇”，从而使得同组内对象间比较相似而不同组对象间差异较大。

离群点是指全局或局部范围内偏离一般水平的观测对象。

1.4

2. 第二章

2.1

数据集由数据对象组成。一个**数据对象**代表一个实体。

数据对象又称为样本、实例、数据点、对象或元组。

数据对象用属性描述。数据表的行对应数据对象；列对应属性

属性(特征，变量，维)是一个数据字段，表示数据对象的一个特征。

标称属性(nominal)

二元属性(binary)

序数属性(ordinal)

数值属性(numeric)

区间标度属性(interval-scaled)

比率标度属性(ratio-scaled)

标称属性(nominal attribute)：类别，状态或事物的名字

每个值代表某种类别、编码或状态，这些值不必具有有意义的序，可以看做是枚举的

二元属性(binary attribute)：布尔属性，是一种标称属性，只有两个状态：0 或 1。

对称的(symmetric)：两种状态具有同等价值，且具有相同的权重。例如：性别

非对称的(asymmetric)：其状态的结果不是同样重要。例如：体检结果（阴性和阳性），惯例：重要的结果用 1 编码（如，HIV 阳性）。

序数属性(ordinal attribute)，其可能的值之间具有有意义的序或者秩评定(ranking)，但是相继值之间的差是未知的

数值属性(numeric attribute）:定量度量，用整数或实数值表示

区间标度(interval-scaled)属性：使用相等的单位尺度度量。值有序，可以评

估值之间的差，不能评估倍数。没有绝对的零点。
例如：日期，摄氏温度，华氏温度

比率标度(ratio-scaled)属性：具有固定零点的数值属性。值有序，可以评估值之间 的差，也可以说一个值是另一个的倍数。
例如：开式温标(K)，重量，高度，速度

离散属性(discrete Attribute)：具有有限或者无限可数个值。有时，表示为整型量。
例如：邮编、职业或文库中的字集
二进制属性是离散属性的一个特例

连续属性(Continuous Attribute)：属性值为实数，一般用浮点变量表示。例如，温度，高度或重量，实际上，真实值只能使用一个有限的数字来测量和表示。

2.2

计算：

1. 中心趋势度量

– 均值（Mean） 掌握5、6、7、8、9、10、11页的计算

- 令x_1，x_2，…，x_N为某数值属性X的N个观测值，该值集合的均值如式(2-1)所示。

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} \qquad (2\text{-}1)$$

例：有学生考试成绩的值：60，45，33，77，80，100，100，90，70，65。

$$\bar{x} = \frac{60 + 45 + 33 + 77 + 80 + 100 + 100 + 90 + 70 + 65}{10} = \frac{720}{10} = 72$$

– 截尾均值

1. 中心趋势度量

– 加权算数平均数（Weighted Mean）

- 对于i=1，…，N，每个值x_i都有一个权重w_i。

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N} \qquad (2\text{-}2)$$

例：某同学的某一科的考试成绩：平时测验 80，期中 90，期末 95。科目成绩的计算方式是：平时测验占20%，期中成绩占30%，期末成绩占50%。这里，每个成绩所占的比重为权重。那么，

$$\bar{x} = \frac{80 \times 20\% + 90 \times 30\% + 95 \times 50\%}{20\% + 30\% + 50\%} = 90.5$$

1. 中心趋势度量

– 中位数(Median)：正中间的值

- 如果值有奇数个，取中间值，否则取中间两个数的平均值
- 有序数据值的中间值
- 如果观察值有偶数个，通常取最中间的两个数值的平均数作为中位数。

例：数据按递增排序为：33，45，60，65，70，77，80，90，100，100。有10个观测值，因此中位数不唯一。中间两个值为70和77，则中位数为

$$\frac{70+77}{2}=73.5$$

1. 中心趋势度量

– 分组数据中位数(Grouped Median)

- 根据N/2确定中位数所在的组

$$M_e = L + \frac{\frac{N}{2} - S_{m-1}}{f_m} \times d \qquad (2\text{-}3)$$

M_e：中位数，L：中位数所在组的下限，S_{m-1}：中位数所在组以下各组的累计频数，f_m：中位数所在组的频数，d：中位数所在组的组距.

1. 中心趋势度量

– 分组数据中位数

例：表2-1为某公司员工薪酬的分组数据，计算数据的近似分组数据中位数。

①判断中位数区间：
N=110+180+320+460+850+250+130+70+20+10=2400；
N/2=1200;
因为：110+180+320+460=1070<1200<1070+850=1920;
所以：1900~1999为对应区间。

②这里有：L=1900，N=2400，S_{m-1}=1070，f_m=850，d=100，由式(2-3)得：

$$M_e = 1900 + \frac{\frac{2400}{2} - 1070}{850} \times 100 \approx 1915.29$$

因此，近似分组数据中位数为1915.29。

表2-1员工薪酬分组数据

Salary	Frequency
1500~1599	110
1600~1699	180
1700~1799	320
1800~1899	460
1900~1999	850
2000~2099	250
2100~2199	130
2200~2299	70
2300~2399	20
2400~2499	10

众数(Mode)：数据中出现最频繁的值

中列数(Midrange)：数据集中最大值和最小值的算术平均值

极差（又称全距，Range）：是集合中最大值不最小值之间的差距， 即最大值减最小值后所得数据。

要求会计算分位数，四分位数：

- 分位数（Quantile）：取自数据分布的每隔一定间隔上的点，把数据划分成基本上大小相等的连贯集合。

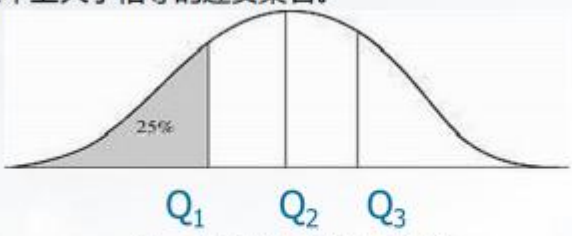


图2-1 某变量x的数据统计描述显示

- 给定数据分布的第k个q-分位数的值为x，使得小于x的数据值最多为k/q，而大于x的数据值最多为(q-k)/q，其中k是整数，使得0<k<q。这里有q-1个q-分位数。

- 四分位数（Quantile）：把数据分布划分成4个相等的部分，使得每部分表示数据分布的四分之一。这3个数据点称为四分位数。

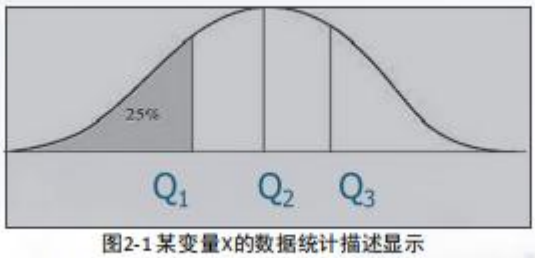


图2-1 某变量x的数据统计描述显示

- Q1：“下四分位数”；Q2：“中位数”；Q3：“上四分位数”。

- 四分位数极差（InterQuartile Range，IQR）：Q1和Q3之间的距离。

$$IQR = Q_3 - Q_1 \qquad (2\text{-}4)$$

确定四分位数的位置：

Q1的位置= (n+1)/4=(n+1) × 0.25

Q2的位置=2*(n+1)/4= (n+1) × 0.5

Q3的位置= 3*(n+1)/4=(n+1) × 0.75

n表示项数

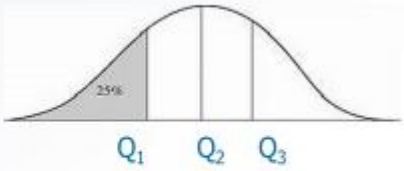


图2-1 某变量x的数据统计描述显示

- 四分位数极差（InterQuartile Range，IQR）：Q1和Q3之间的距离。

$$IQR = Q_3 - Q_1 \qquad (2\text{-}4)$$

例：由8人组成的旅游小团队年龄分别为：17，19，22，24，25，28，34，37，求其年龄的四分位差。

①计算Q1与Q3的位置：

Q1的位置=(n+1)/4=(8+1)/4=2.25；Q3的位置=3*(n+1)/4=3*(8+1)/4=6.75

②确定Q1与Q3的数值：

Q1=19+（22-19）*0.25=19.75；Q3=28+（34-28）*0.75=32.5

③计算四分位差：

IQR=Q3-Q1=32.5-19.75=12.75

- 四分位数极差（InterQuartile Range，IQR）：Q1和Q3之间的距离。

$$IQR = Q_3 - Q_1 \qquad (2\text{-}4)$$

另一种确定四分位数的位置：

Q1的位置=1+(n-1) × 0.25

Q2的位置=1+(n-1) × 0.5

Q3的位置=1+(n-1) × 0.75

n表示项数

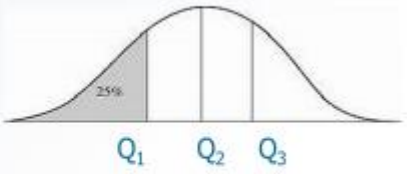
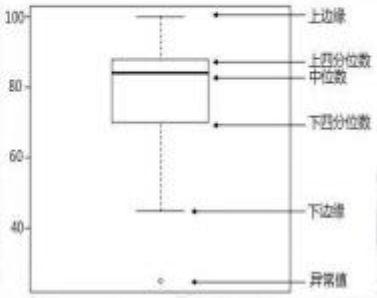


图2-1 某变量x的数据统计描述显示

3. 数据的图形显示

定义，图内容

- 盒图（又称箱线图，Box-plot），是一种用来描述数据分布的统计图形，可以表现观测数据的中位数、四分位数和极值等描述性统计量。
  - 用盒子表示数据
  - 盒子的端点在四分位数上，使得盒子长度为四分位数极差IQR
  - 中位数用盒内线标记
  - 盒子外线延伸到最小和最大的观测值
  - 离群点：绘制在离群阈值范围外的点



2.3

- 相似性(Similarity)
  - 两个对象相似程度的数量表示
  - 数值越高表明相似性越大
  - 通常取值范围为[0,1]
- 相异性(Dissimilarity)(例如距离)
  - 两个对象不相似程度的数量表示
  - 数值越低表明相似性越大
  - 相异性的最小值通常为0
  - 相异性的最大值（上限）是不同的
- 邻近性(Proximity):相似性和相异性都称为邻近性

– 相异性

$$d(i,j)=\frac{p-m}{p}=1-\frac{m}{p} \qquad (2\text{-}8)$$

- p是对象的属性总数，m是匹配的属性数目（即对象i和j状态相同的属性数）

– 相似性

$$sim(i,j)=1-d(i,j)=\frac{m}{p} \qquad (2\text{-}9)$$

– 相异性

- 对称的二进制属性

$$d(i,j)=\frac{p+n}{m+n+p+q}=\frac{p+n}{sum} \qquad (2\text{-}10)$$

- 非对称的二进制属性

$$d(i,j)=\frac{p+n}{m+n+p} \qquad (2\text{-}11)$$

i \ j	1	0	合计
1	*m*	*n*	*m+n*
0	*p*	*q*	*p+q*
合计	*m+p*	*n+q*	*sum*

– 相似性

$$sim(i,j)=1-d(i,j) \qquad (2\text{-}12)$$

例：计算二进制属性的相异性

表2-9 居民家庭情况调查表

姓名	婚否	买房否	买车否
张明	Y	N	N
李思	N	Y	Y
王刚	Y	Y	N

$$d(\text{张明},\text{李思})=\frac{2+1}{0+1+2+0}=1$$

$$d(\text{张明},\text{王刚})=\frac{1+0}{1+0+1+1}=0.33$$

$$d(\text{李思},\text{王刚})=\frac{1+1}{1+1+1+0}=0.67$$

- 欧几里得距离（Euclidean Distance）：又称直线距离。

  - 令$\boldsymbol{i}=(x_{i1},\ x_{i2},\ \cdots,\ x_{ip})$和$\boldsymbol{j}=(x_{j1},\ x_{j2},\ \cdots,\ x_{jp})$是两个被p个数值属性描述的对象。对象i和j之间的欧几里得距离为：

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+\cdots+(x_{ip}-x_{jp})^2} \qquad (2\text{-}13)$$

- 曼哈顿距离（Manhattan Distance）：又称城市块距离。

  - 令$\boldsymbol{i}=(x_{i1},\ x_{i2},\ \cdots,\ x_{ip})$和$\boldsymbol{j}=(x_{j1},\ x_{j2},\ \cdots,\ x_{jp})$是两个被p个数值属性描述的对象。对象i和j之间的曼哈顿距离为：

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\cdots+|x_{ip}-x_{jp}| \qquad (2\text{-}14)$$

- 欧几里得距离和曼哈顿距离都满足如下数学性质：

  ①非负性：$d(i,j)\geq 0$：距离是一个非负的数值。

  ②同一性：$d(i,i)=0$：对象到自身的距离为0。

  ③三角不等式：$d(i,j)\leq d(i,k)+d(k,j)$：从对象i到对象j的直接距离不会大于途经任何其他对象k的距离。

- 闵可夫斯基距离（Minkowski Distance）

  - 令$\boldsymbol{i}=(x_{i1},\ x_{i2},\ \cdots,\ x_{ip})$和$\boldsymbol{j}=(x_{j1},\ x_{j2},\ \cdots,\ x_{jp})$是两个被p个数值属性描述的对象。对象i和j之间的闵可夫斯基距离为：

$$d(i,j)=\sqrt[h]{|x_{i1}-x_{j1}|^h+|x_{i2}-x_{j2}|^h+\cdots+|x_{ip}-x_{jp}|^h} \qquad (2\text{-}15)$$

- 切比雪夫距离（Chebyshev Distance）：又称上确界距离，定义两个对象之间的上确界距离为其各坐标数值差的最大值。

$$d(i,j)=\lim_{h\to\infty}\left(\sum_{f=1}^{p}|x_{if}-x_{jf}|^h\right)^{\frac{1}{h}}=\max_{f\to p}|x_{if}-x_{jf}| \qquad (2\text{-}16)$$

4. 数值属性的相异性

例：数值属性的相异性计算

给定两个对象分别用元组(2，8，7，4)和(1，5，3，0)描述，计算这两个对象之间的欧几里得距离、曼哈顿距离、闵可夫斯基距离（h=4），以及切比雪夫距离。

- 欧几里得距离为：$d(i,j)=\sqrt{(2-1)^2+(8-5)^2+(7-3)^2+(4-0)^2}=\sqrt{42}=6.48$
- 曼哈顿距离为：$d(i,j)=|2-1|+|8-5|+|7-3|+|4-0|=1+3+4+4=12$
- 闵可夫斯基距离为：$d(i,j)=\sqrt[4]{|2-1|^4+|8-5|^4+|7-3|^4+|4-0|^4}=\sqrt[4]{594}\approx 4.94$
- 切比雪夫距离为：$d(i,j)=max\{|2-1|，|8-5|，|7-3|，|4-0|\}=max\{1，3，4，4\}=4$

- 余弦相似性（又称余弦相似度，Cosine Similarity）：是基于向量的，它利用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。
  - 令向量$\boldsymbol{x}=(x_1，x_2，\cdots，x_p)$，向量$\boldsymbol{y}=(y_1，y_2，\cdots，y_p)$，两个向量的余弦相似性定义为：

$$sim(\boldsymbol{x},\boldsymbol{y})=\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}=\frac{x_1y_1+x_2y_2+\cdots+x_py_p}{\sqrt{x_1^2+x_2^2+\cdots+x_p^2}\sqrt{y_1^2+y_2^2+\cdots+y_p^2}}=\frac{\sum_{i=1}^{p}x_iy_i}{\sqrt{\sum_{i=1}^{p}x_i^2}\sqrt{\sum_{i=1}^{p}y_i^2}}\qquad(2\text{-}19)$$

  - 其中，$\|\boldsymbol{x}\|$是向量$\boldsymbol{x}=(x_1，x_2，\cdots，x_p)$的欧几里得范数，定义为$\sqrt{x_1^2+x_2^2+\cdots+x_p^2}$。同理，$\|\boldsymbol{y}\|$是向量$\boldsymbol{y}=(y_1，y_2，\cdots，y_p)$的欧几里得范数，定义为$\sqrt{y_1^2+y_2^2+\cdots+y_p^2}$。

例：余弦相似度的计算

给定两个向量$\boldsymbol{x}=(1，2，5，4)$和$\boldsymbol{y}=(2，3，5，1)$，计算两个向量的余弦相似度。

余弦相似度为：$sim(\boldsymbol{x},\boldsymbol{y})=\frac{1*2+2*3+5*5+4*1}{\sqrt{1+4+25+16}*\sqrt{4+9+25+1}}=\frac{37}{\sqrt{46}*\sqrt{39}}\approx 0.87$

3. 第三章

3.1

数据存在的问题？**数据不一致，噪声数据，缺失值。**

数据质量的要求？**准确性，完整性，一致性，时效性，可信性，可解释性。**

数据处理的主要任务：

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)。

3.2

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。

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载

如何处理空缺值？忽略元组，忽略属性列，人工填写空缺值，使用属性的中心度量值填写空缺值，使用一个全局变量填充空缺值，**使用可能的特征值来替换空缺值（常用）**

如何检测噪声数据？基于统计的技术，基于距离的技术。

3.3

检测冗余的方法：相关性分析

数值属性：蚕蛹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

1）相关系数：

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

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。

例：数值属性的相关性分析。

表3.3 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3.4 体重和血压的均值和标准差值

	均值	标准差
体重	63.83	10.14
血压	118.33	24.74

$r_{X,Y}=\sum_{i=1}^{m}(x_i-\bar{X})(y_i-\bar{Y})/m\sigma_X\sigma_Y=-0.1$

3.4

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。

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，

取样方法：

Ø 不放回简单随机取样 （Simple Random Sampling Without Replacement, SRSWOR）

Ø 放回简单随机取样（Simple Random Sampling With Replacement, SRSWR）

Ø 聚类取样（Clustered Sampling）

Ø 分层取样（Stra;fied Sampling）

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 1/）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-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。

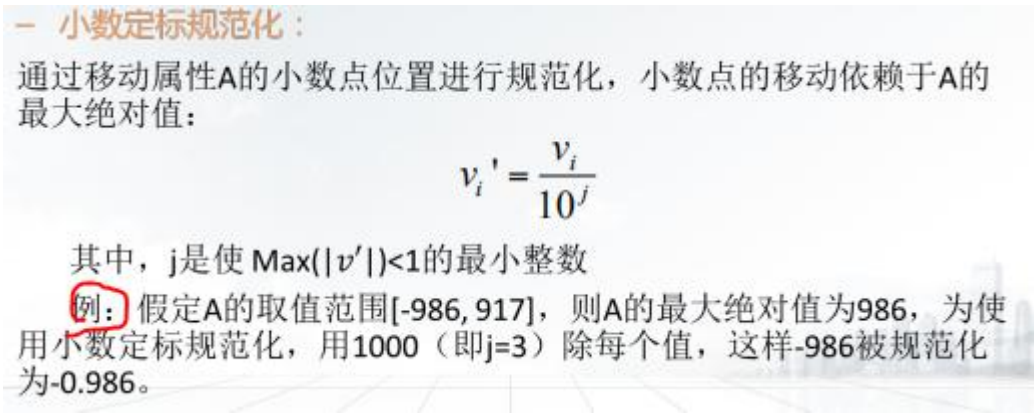
3.5

数据变换：将数据转换成适合数据挖掘的形式

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。

数据变换常用方法：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）。



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]中的v'

$$v_i' = \frac{v_i - \min A}{\max A - \min A}(b-a) + a$$

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）：
  - 将属性A的值根据其平均值和标准差进行规范化：
  - 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。

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

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

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。

例3.6 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

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

4. 第四章

4.1

数据仓库是一个面向主题的、集成的、时变的并且非易失的，用于支持管理者决策过程的数据集合

数据仓库的特征：

- 面向主题的
- 集成的
- 时变的
- 非易失的

联机分析处理（Online Analytical Processing, OLAP）是数据仓库系统前端分析服务的分析工具，能快速汇总大量数据并进行高效查询分析，为分析人员提供决策支持。

4.2

数据仓库的设计：

概念模型设计、 逻辑模型设计、 物理模型设计

4.4

OLAP 特点 ：

快速性 、可分析性 、多维性

表4-11 OLAP与OLTP的对比

比较项	OLAP	OLTP
特性	信息处理	操作处理
用户	面向决策人员	面向操作人员
功能	支持管理需要	支持日常操作
面向	面向数据分析	面向应用
驱动	分析驱动	事务驱动
数据量	一次处理的数据量大	一次处理的数据量小
访问	不可更新，但周期性刷新	可更新
数据	历史数据	当前值数据
汇总	综合性和提炼性数据	细节性数据
视图	导出数据	原始数据

5. 第五章

5.1

回归分析的分类：

按涉及变量个数划分： 一元回归分析 、多元回归分析

按自变量和因变量之间关系划分： 线性回归分析 、非线性回归分析

回归分析的步骤 ：

确定变量

建立预测模型

进行相关分析

计算预测误差

确定预测值

5.2

一元线性回归模型只包含一个解释变量（自变量）和一个被解释变量（因变量），是最简单的线性回归模型。

一元线性回归模型为： $Y=a+bX+\varepsilon$

其中，X 为自变量，Y 为因变量；a 为截距，是一常量；b 为回归系数，表示自变量对因变量的影响程度；ε 为随机误差项。

2. 回归方程

$$\hat{Y}=\hat{a}+\hat{b}X$$

- a 和 b 是回归方程的回归系数，a 是回归直线在y轴上的截距，b 是直线的斜率。
- 对于每一个 x_i，由回归方程可确定一个回归值 $\hat{y}_i=\hat{a}+\hat{b}x_i$。

例： 某种商品与家庭平均消费量的关系

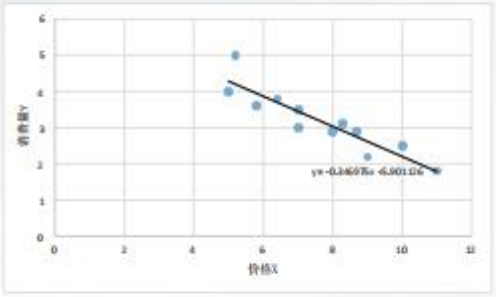
以某家庭为调查单位，某种商品在某年各月的家庭平均月消费量 Y (kg)与其价格 X (元/kg) 间的调查数据如表5-1。

表5-1 商品价格与消费量的关系

价格X	5.0	5.2	5.8	6.4	7.0	7.0	8.0	8.3	8.7	9.0	10.0	11
消费量Y	4.0	5.0	3.6	3.8	3.0	3.5	2.9	3.1	2.9	2.2	2.5	2.6

例： 某种商品与家庭平均消费量的关系（续）

在坐标轴上做出价格与消费量的相关关系。



由上图知，该商品在某家庭月平均消费量 Y 与价格 X 间基本呈线性关系，这些点与直线间的偏差是由其它一些无法控制的因素和观察误差引起的，根据 Y 与 X 之间的线性关系及表5-1中数据，可以求得两者之间的回归方程。

（1）求解一元线性回归方程：

① 求解 $\bar{x}$，$\bar{y}$，$\overline{xy}$，$\overline{x^2}$

$$\bar{x}=\frac{1}{12}(5.0+5.2+5.8+6.4+7.0+7.0+8.0+8.3+8.7+9.0+10.0+11)=7.616667$$

$$\bar{y}=\frac{1}{12}(4.0+5.0+3.6+3.8+3.0+3.5+2.9+3.1+2.9+2.2+2.5+2.6)=3.258333$$

$$\overline{xy}=\frac{1}{12}(5.0*4.0+5.2*5.0+5.8*3.6+6.4*3.8+7.0*3.0+7.0*3.5+8.0*2.9+8.3*3.1+8.7*2.9+9.0*2.2+10.0*2.5+11*2.6)=23.688333$$

$$\overline{x^2}=\frac{1}{12}(5.0^2+5.2^2+5.8^2+6.4^2+7.0^2+7.0^2+8.0^2+8.3^2+8.7^2+9.0^2+10.0^2+11^2)=61.268333$$

求解过程：

②根据 $\bar{x}$，$\bar{y}$，$\overline{xy}$，$\overline{x^2}$求解 $\hat{b}$

$$\hat{b}=\frac{\bar{x}\cdot\bar{y}-\overline{xy}}{\bar{x}^2-\overline{x^2}}=\frac{7.616667*3.258333-23.688333}{7.616667^2-61.268333}=-0.346975$$

③根据 $\hat{b}$，$\bar{x}$，$\bar{y}$求解 $\hat{a}$

$$\hat{a}=\bar{y}-\hat{b}\bar{x}=3.258333-(-0.346975)*7.616667=5.901126$$

故求得的线性回归方程为：$\hat{y}=-0.346975x+5.901126$

6. 第六章

6.1

项集 ： 包含 0 个或者多个项的集合

支持度

绝对支持度： 项集的出现频度，即包含项集的事务数。

相对支持度： 项集出现的百分比

频繁项集： 事务中同时包含集合 A 和集合 B 的事务数与包含集合 A 的事务数的百分比

要求会计算支持度和置信度：

- 关联规则

设$I=\{i1,i2,i3,\ldots,in\}$是事务数据中所有项的集合，$T=\{t1,t2,t3,\ldots,tn\}$是所有事务的集合，其中每个事务ti都有一个独一无二的标识符TID。

关联规则是形如$A\Rightarrow B$的蕴含式，其中A称为规则前件，B称为规则后件，并且A，B满足：A，B是I的真子集，并且A和B的交集为空集。

关联规则的支持度是指事务中同时包含集合A和集合B的百分比。

$$support(A\Rightarrow B)=P(A\cup B)$$

表6-1 某商店的事务数据

TID	Items
1	牛奶，面包，麦片
2	牛奶，面包，麦片，鸡蛋
3	牛奶，面包，黄油，麦片
4	糖，鸡蛋
5	黄油，麦片
6	糖，鸡蛋

- 置信度

置信度是指事务中同时包含集合X与Y的事务数与包含集合X的事务数的百分比。。

$$confidence(A\Rightarrow B)=P(B|A)$$

同时满足最小支持度和最小置信度阈值要求的所有关联规则被称为**强关联规则**。

例6.6 强关联规则

假设最小置信度阈值为30%，最小置信度阈值为70%，而关联规则：购买面包⇒购买牛奶[支持度=50%，置信度=100%]的支持度和置信度都满足条件，则该规则为强关联规则。

关联规则挖掘的任务：

①根据最小支持度阈值，找出数据集中所有的频繁项集；

②挖掘出频繁项集中满足最小支持度和最小置信度阈值要求的规则，得到强关联规则；

③对产生的强关联规则进行剪枝，找出有用的关联规则。

先验性质 ：如果一个项集是频繁的，那么它的所有非空子集也是频繁的。

6.2

关联规则挖掘的步骤 ：

1.找出所有频繁项集，即大于或等于最小支持度阈值的项集

2.由频繁项集产生强关联规则，这些规则必须大于或等于最小支持度阈值和最小置信度阈值。

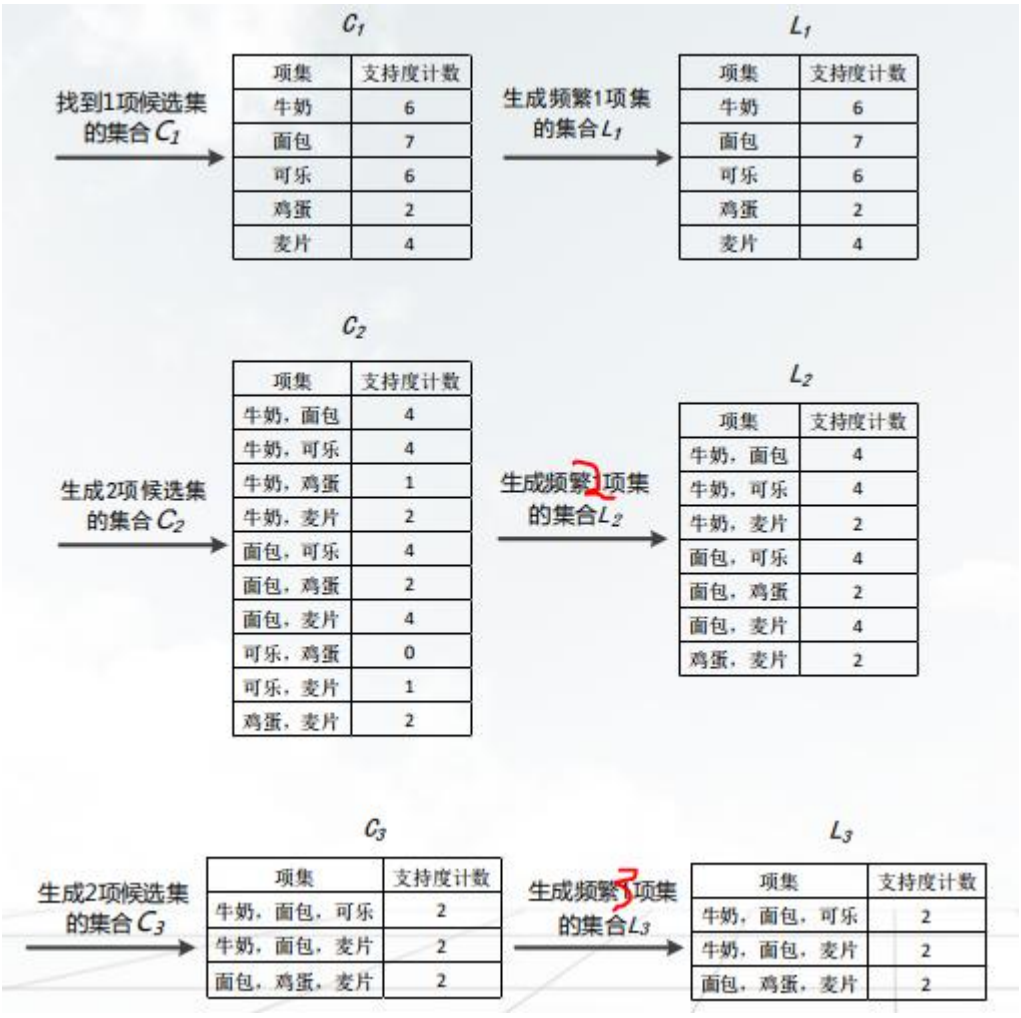
TID	Items
1	面包、可乐、麦片
2	牛奶、可乐
3	牛奶、面包、麦片
4	牛奶、可乐
5	面包、鸡蛋、麦片
6	牛奶、面包、可乐
7	牛奶、面包、鸡蛋、麦片
8	牛奶、面包、可乐
9	面包、可乐

例6.7 Apriori算法

假设使用表中的事务数据，该数据库具有9个事务，设最小支持度为2，试使用Apriori算法挖掘表6-3的事务数据中的频繁项集。

答：

L={{牛奶}：6，{面包}：7，{可乐}：6，
{鸡蛋}：2，{麦片}：4，{牛奶，面包}：4
{牛奶，可乐}：4，{牛奶，麦片}：2，
{面包，可乐}：4，{面包，鸡蛋}：2，
{面包，麦片}：4，{鸡蛋，麦片}：2，
{牛奶，面包，可乐}：2，
{牛奶，面包，麦片}：2，
{面包，鸡蛋，麦片}：2}



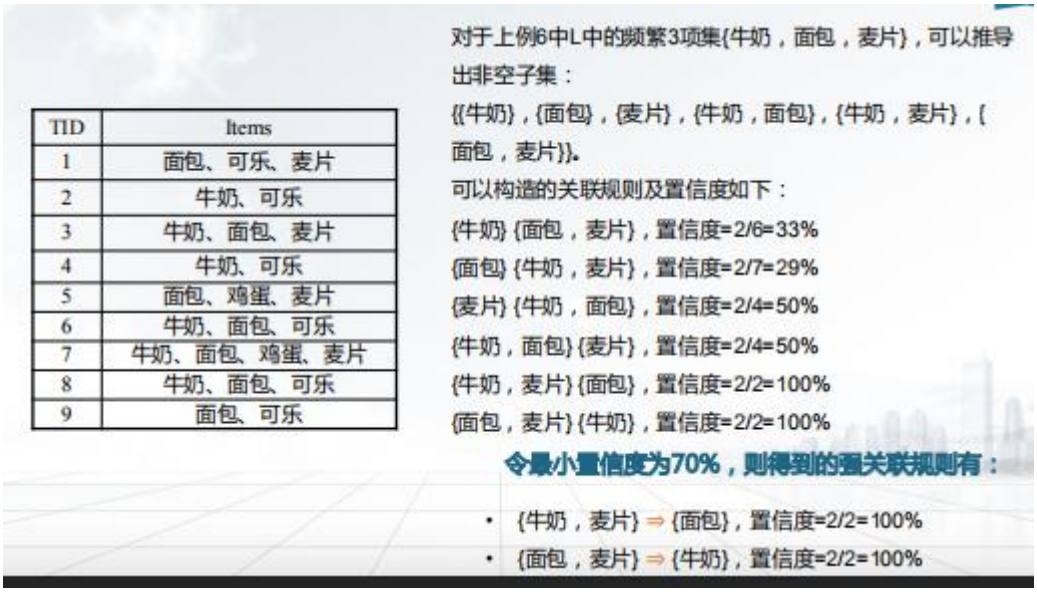
关联规则的生成过程包括两个步骤：

①对于L中的每个频繁项集X，生成X所有的非空真子集Y；

②对于X中的每一个非空真子集Y，构造关联规则$Y \Rightarrow (X-Y)$。

构造出关联规则后，计算每一个关联规则的置信度，如果大于最小置信度阈值，则该规则为强关联规则。

如何得出强相关联规则：



TID	Items
1	面包、可乐、麦片
2	牛奶、可乐
3	牛奶、面包、麦片
4	牛奶、可乐
5	面包、鸡蛋、麦片
6	牛奶、面包、可乐
7	牛奶、面包、鸡蛋、麦片
8	牛奶、面包、可乐
9	面包、可乐

对于上例6中L中的频繁3项集{牛奶，面包，麦片}，可以推导出非空子集：

{{牛奶}，{面包}，{麦片}，{牛奶，面包}，{牛奶，麦片}，{面包，麦片}}。

可以构造的关联规则及置信度如下：

{牛奶}⇒{面包，麦片}，置信度=2/6=33%

{面包}⇒{牛奶，麦片}，置信度=2/7=29%

{麦片}⇒{牛奶，面包}，置信度=2/4=50%

{牛奶，面包}⇒{麦片}，置信度=2/4=50%

{牛奶，麦片}⇒{面包}，置信度=2/2=100%

{面包，麦片}⇒{牛奶}，置信度=2/2=100%

令最小置信度为70%，则得到的强关联规则有：

- {牛奶，麦片}⇒{面包}，置信度=2/2=100%
- {面包，麦片}⇒{牛奶}，置信度=2/2=100%

一个频繁项集 X 能够生成 2|X|-2 个（即除去空集及自身之外的子集）候选关联规则

频繁项集的性质：

①如果 X 是频繁项集，则它的任何非空子集 X ‘也是频繁项集。即频繁项集的子集也是频繁项集。

②如果 X 是非频繁项集，则它的所有真超级都是非频繁项集。即非频繁项集的超集也是非频繁项集

6.3

Apriori 算法的优缺点：

优点 ：算法原理简单，易于理解。

缺点：

- 需要多次扫描数据集 ，如果频繁项集最多包含 10 个项，需要扫描事务数据集 10 次，这需要很大的 I/ O 负载
- 产生大量频繁项集 ，如数据集有 100 项，可能产生的候选项个数为 $1.27*10^{30}$

6.5

会计算：

表6-5 1000个人的手机偏爱

	买苹果手机	不买苹果手机	行和
买小米手机	400	350	750
不买小米手机	200	50	250
列和	600	400	1000

3.皮尔森相关系数

皮尔森相关系数能够反映两个变量的相似程度，皮尔森相关系数值越大表明两个变量的相关性越强。对于二元变量，皮尔森相关系数定义如式(6-4)所示。

$$\rho(A,B)=\frac{P(A\cup B)P(\overline{A}\cup\overline{B})-P(\overline{A}\cup B)P(A\cup\overline{B})}{\sqrt{P(A)P(\overline{A})P(B)P(\overline{B})}}$$

{苹果手机}和{小米手机}的皮尔森相关系数为

$$\rho(\{苹果\ 手机\},\{小米\ 手机\})=(0.4*0.05-0.35*0.2)/\sqrt{0.6*0.4*0.75*0.25}=-0.2357$$

说明两者一定程度负相关。

1.提升度

- 令A和B表示不同的项集，P(*)表示项集*在总体数据集中的出现概率。根据统计学定义，如果项集A和项集B的$P(A\cup B)=P(A)P(B)$，那么项集A和项集B是相互独立的，否则两者是相互依赖的。
- 项集A和项集B的提升度定义如式(6-3)所示。

$$\mathrm{lift}(A,B)=\frac{P(A\cup B)}{P(A)P(B)}$$

如果A和B的提升度的值等于1，说明A和B相互独立；若是A和B的提升度的值大于1，说明A和B正相关；如果A和B的提升度的值小于1，说明A和B负相关。

Lift({苹果手机}，{小米手机})=0.4/(0.75*0.6)=0.89

7. 第七章

7.1

什么是分类？

- 分类就是根据以往的数据和结果对另一部分数据进行结果的预测。
- 模型的学习在被告知每个训练样本属于哪个类的“指导”下进行新数据使用训练数据集中得到的规则进行分类

分类的基本过程：

Ø 学习阶段：建立一个分类模型，描述预定数据类或概念集。
评估模型的预测准确率
如果准确率可以接受，那么使用该模型来分类标签为未知的样本。

Ø 分类阶段：即使用分类模型，对将来的或未知的对象进行分类。

分类与预测

- 不同点

– 分类是预测类对象的分类标号（或离散值），根据训练数据集和类标号属性，构建模型来分类现有数据，并用来分类新数据。

– 预测是建立连续函数值模型评估无标号样本类，或评估给定样本可能具有的属性值或值区间，即用来估计连续值或量化属性值，比如预测空缺值。

- 相同点

– 分类和预测的共同点是两者都需要构建模型，都用模型来估计未知值。预测中主要的估计方法是回归分析。

计算：

– 1、信息熵

信息熵用来衡量事件的不确定性的大小，计算公式如下：

$$Infor(x)=-p(x)\times log_2 p(x)$$

信息熵具有可加性，即多个期望信息，计算公式如下：

$$Infor(X)=-\sum_{i=1}^{m} p(x_i)\times log_2 p(x_i)$$

– 2、信息增益

信息增益表示某一特征的信息对类标签的不确定性减少的程度。

$$gDA=Infor(D)-Infor(D|A)$$

其中$Infor(D|A)$是在特征A给定条件下对数据集合D进行划分所需要的期望信息，它的值越小表示分区的纯度越高，计算公式如式(7-4)所示。

$$Infor(D|A)=\sum_{j=1}^{n}|D_j|/|D| \times Info(D_j) \qquad (7\text{-}4)$$

其中n是数据分区数，$|D_j|$表示第j个数据分区的长度，$|D_j|/|D|$表示第j个数据分区的权重。

- 例7.1 信息增益的计算
- 表7-1是带有标记类的训练集D，训练集的列是一些特征，表中最后一列的类标号为是否提供贷款，有两个不同的取值，计算按照每个特征进行划分的信息增益。

表7-1 贷款申请的训练集

ID	学历	婚否	是否有车	收入水平	类别
1	专科	否	否	中	否
2	专科	否	否	高	否
3	专科	是	否	高	是
4	专科	是	是	中	是
5	专科	否	否	中	否
6	本科	否	否	中	否
7	本科	否	否	高	否
8	本科	是	是	高	是
9	本科	否	是	很高	是
10	本科	否	是	很高	是
11	研究生	否	是	很高	是
12	研究生	否	是	高	是
13	研究生	是	否	高	是
14	研究生	是	否	很高	是
15	研究生	否	否	中	否

①根据公式计算信息熵 $Infor(D)$。

$$Infor(D)=-9/15\times log_2 9/15-6/15\times log_2 6/15=0.971$$

②计算按照每个特征进行划分的期望信息，A代表特征“学历”，B代表特征“婚否”，C代表特征“是否有车”，E代表特征“收入水平”。

$$Infor(D|A)=5/15\times(-2/5\ log_2 2/5-3/5\ log_2 3/5)+5/15\times(-3/5\ log_2 3/5-2/5\ log_2 2/5)$$
$$+5/15\times(-4/5\ log_2 4/5-1/5\ log_2 1/5)=0.888$$

$$Infor(D|B)=Infor(D|B)=10/15\times(-6/10\ log_2 6/10-4/10\ log_2 4/10)+5/15\times(-5/5\ log_2 5/5)=0.647$$

$$Infor(D|C)=9/15\times(-6/9\ log_2 6/9-3/9\ log_2 3/9)+6/15\times(-6/6\ log_2 6/6)=0.951$$

$$Infor(D|E)=5/15\times(-4/5\ log_2 4/5-1/5\ log_2 1/5)+6/15\times(-2/6\ log_2 2/6-4/6\ log_2 4/6)+4/15\times(-4/4\ log_2 4/4)=0.608$$

③计算信息增益 $gDA=Infor(D)-Infor(D|A)=0.083$

$$gDB=Infor(D)-Infor(D|B)=0.324$$
$$gDC=Infor(D)-Infor(D|C)=0.019$$
$$gDE=Infor(D)-Infor(D|E)=0.363$$

– 3、信息增益率

信息增益率是指按照某一特征进行划分的信息增益与训练集关于这个特征的信息熵的比值。：

$$gain\_ratio(D,A)=gain(D,A)/SplitInfo_A(D)$$

其中：

$$SplitInfo_A(D)=-\sum_{j=1}^{m}\frac{|D_j|}{|D|}\times\log_2\left(\frac{|D_j|}{|D|}\right)$$

例7.2 信息增益率的计算

基于例7.1的数据，计算按照每个特征进行划分的信息增益率。

解：①根据例7.1计算出的按照每个特征划分的信息增益，*A*代表特征“学历”，*B*代表特征“婚否”，*C*代表特征“是否有车”，*E*代表特征“收入水平”，计算$SplitInfo_A(D)$。

$$SplitInfo_A(D)=-5/15\times\log_2 5/15-5/15\times\log_2 5/15-5/15\times\log_2 5/15=1.585$$

$$SplitInfo_B(D)=-10/15\times\log_2 10/15-5/15\times\log_2 5/15=0.918$$

$$SplitInfo_C(D)=-9/15\times\log_2 9/15-6/15\times\log_2 6/15=0.971$$

$$SplitInfo_E(D)=-5/15\times\log_2 5/15-6/15\times\log_2 6/15-4/15\times\log_2 4/15=1.566$$

②按照公式(7-5)计算信息增益率。

$$gain\_ratio(D,A)=0.083/1.585=0.052$$

$$gain\_ratio(D,B)=0.324/0.918=0.331$$

$$gain\_ratio(D,C)=0.420/0.971=0.433$$

$$gain\_ratio(D,E)=0.363/1.566=0.232$$

– 4、基尼系数

基尼指数是度量数据分区或者训练数据的不纯度。

$$Gini(D)=1-\sum_{i=1}^{m}p_i^2$$

其中p_i是数据集*D*中任何一个记录属于C_i类的概率，可通过$|C_i,D|/|D|$ 进行计算，$|C_i,D|$是*D*中属于C_i类的集合的记录个数，$|D|$是所有记录的个数。如果所有的记录都属于同一个类，则$p_i=1$，*m*是分区数量。基尼指数考虑的是二元化，即将某一特征中的数值分为两个子集，然后进行划分。如果按照特征*A*作为数据的二元划分准则将*D*分成D_1 和D_2 ，则*D*的基尼指数为：

$$Gini_A(D)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2)$$

对于属性*A*的二元划分导致的不纯度降低为

$$\Delta Gini(A)=Gini(D)-Gini_A(D)\qquad(7\text{-}9)$$

例7.3 计算属性的不纯度降低值

根据表7-1中的数据计算“学历”属性的基尼指数。

解 ①使用基尼指数计算公式(7-7)计算*D*的不纯度：

$$Gini(D)=1-(9/15)^2-(6/15)^2=0.48$$

②计算属性“学历”的基尼指数。此特征有三个取值：“专科”、“本科”、“硕士”。所以划分值有三个，即三种划分集合，分别为：

以“专科”划分：{专科}、{本科、研究生}。

以“本科”划分：{本科}、{专科、研究生}。

以“研究生”划分：{研究生}、{专科、本科}。

考虑集合{研究生}、{本科，专科}，*D*被划分成两个部分，基于这样的划分计算基尼指数为：

$$Gini_{\{本科，专科\},\{研究生\}}(D)=10/15\,Gini(D_1)+5/15\,Gini(D_2)$$

$$=10/15\times(1-(1/2)^2-(1/2)^2)+5/15\times(1-(1/5)^2-(4/5)^2)=0.44$$

例7.3 计算属性的不纯度降低值

根据表7-1中的数据计算“学历”属性的基尼指数。

解

类似地可以求出属性“学历”其余子集的基尼指数：

以“专科”划分的基尼指数为：

$$Gini_{\{本科，研究生\},\{专科\}}(D)=10/15\ Gini(D1)+5/15\ Gini(D2)$$
$$=10/15\times(1-(3/10)^2-(7/10)^2)+5/15\times(1-(2/5)^2-(3/5)^2)=0.44$$

以“本科”划分的基尼指数为：

$$Gini_{\{专科，研究生\},\{本科\}}(D)=10/15\ Gini(D1)+5/15\ Gini(D2)$$
$$=10/15\times(1-(2/5)^2-(3/5)^2)+5/15\times(1-(2/5)^2-(3/5)^2)=0.48$$

选择基尼指数最小值0.44作为属性“学历”的基尼指数，

因此属性“学历”的不纯度降低值为：

$$\Delta Gini(A)=Gini(D)-Gini_A(D)=0.48-0.44=0.04$$

同样可以求出每个属性的基尼指数及不纯度降低值。

虽然模型在训练数据上有较好的效果，但是对未知的测试数据可能结果会不好，这种现象叫做**过拟合**

7.2

决策树构造过程：

①输入数据，主要包括训练集的特征和类标号。

②选取一个属性作为根节点的分裂属性进行分裂。

③对于分裂的每个分支，如果已经属于同一类就不再分了，如果不是同一类，依次选取不同的特

征作为分裂属性进行分裂，同时删除已经选过的分列属性。

④不断的重复③，直到到达叶子节点，也就是决策树的最后一层，这时这个节点下的数据都是一

类了。

⑤最后得到每个叶子节点对应的类标签以及到达这个叶子节点的路径。

计算：（温馨提示，我觉得这道题可以放弃了）

ID3

例7.4 使用ID3算法进行分类预测

表7-2和表7-3为训练数据和测试数据，其中“患病与否”是类标记，使用ID3算法构建决策树然后进行分类预测。

表7-2 某疾病患病情况的训练数据

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
1	23	无	无	较低	否
2	25	无	无	中	否
3	27	0-5年	无	中	否
4	30	0-5年	有	低	是
5	39	无	无	较低	否
6	41	无	无	低	否
7	43	无	无	高	否
8	45	5年以上	有	高	是
9	46	无	有	高	是
10	47	无	有	高	是
11	62	无	有	较高	是
12	63	无	有	高	是
13	66	5年以上	无	高	是
14	66	5年以上	无	较高	是
15	68	0-5年	无	中	否

表7-3 某疾病患病情况的测试数据

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
1	25	无	无	较低	?
2	42	无	无	高	?
3	67	5年以上	无	较高	?

ID3算法的构建方法和决策树的构建基本是一致的，不同的是分裂节点的特征选择的标准。该算法在分裂节点处将信息增益作为分裂准则进行特征选择，递归的构建决策树。

ID3算法的步骤如下：

① 输入数据，主要包括训练集的特征和类标号。

② 如果 所有实例都属于一个类别，则决策树是一个单结点树，否则执行③。

③ 计算训练数据中每个特征的信息增益。

④ 从根节点开始选择最大信息增益的特征进行分裂。依次类推，从上向下构建决策树，每次选择具有最大信息增益的特征进行分裂，选过的特征后面就不能继续进行选择使用了。

⑤ 不断的构建决策树，至没有特征可以选择或者分裂后的所有元组属于同一类别时候停止构建。

⑥ 决策树构建完成，进行预测。

解：①连续型数据的离散化。ID3算法不能直接处理连续型数据，只有通过离散化将连续型数据转化成离散型数据再进行处理。

此例采用等宽分箱法对连续型特征“年龄”离散化：

设定区域范围（设箱子数为3，箱子宽度为（68-23）/3=15），分箱结果为：

箱1：23 25 27 30

箱2：39 41 43 45 46 47

箱3：62 63 66 66 68

图7-4 对特征“年龄”分箱

离散后训练数据集如表7-4所示。

表7-4 某疾病患病情况的训练数据（离散化后）

ID	年龄	家族史	有无支气管炎史	体重指数	患病与否
1	青年	无	无	较低	否
2	青年	无	无	中	否
3	青年	0-5年	无	中	否
4	青年	0-5年	有	低	是
5	中年	无	无	较低	否
6	中年	无	无	低	否
7	中年	无	无	高	否
8	中年	5年以上	有	高	是
9	中年	无	有	高	是
10	中年	无	有	高	是
11	老年	无	有	较高	是
12	老年	无	有	高	是
13	老年	5年以上	无	高	是
14	老年	5年以上	无	较高	是
15	老年	0-5年	无	中	否

②根据训练数据构造ID3算法的决策树，其中Z代表训练集，A、B、C、D分别代表特征“年龄”、“吸烟史”、“有无家族病史”、“体重范围”，按照每个特征计算其分裂的信息增益。

$$Infor(Z)=-8/15\times\log\downarrow2\ 8/15\ -7/15\times\log\downarrow2\ 7/15\ =0.997$$

$$Infor(Z|A)=4/15\times(-3/4\times\log\downarrow2\ 3/4\ -1/4\times\log\downarrow2\ 1/4\ )+6/15\times(-3/6\times\log\downarrow2\ 3/6\ -3/6\times\log\downarrow2\ 3/6\ )+5/15\times(-4/5\times\log\downarrow2\ 4/5\ -1/5\times\log\downarrow2\ 1/5\ )=0.857$$

$$Infor(Z|B)=9/15\times(-5/9\times\log\downarrow2\ 5/9\ -4/9\times\log\downarrow2\ 4/9\ )+3/15\times(-2/3\times\log\downarrow2\ 2/3\ -1/3\times\log\downarrow2\ 1/3\ )+3/15\times(-3/3\times\log\downarrow2\ 3/3\ )=0.778$$

②根据训练数据构造ID3算法的决策树，其中Z代表训练集，A、B、C、D分别代表特征“年龄”、“吸烟史”、“有无家族病史”、“体重范围”，按照每个特征计算其分裂的信息增益。

$$Infor(Z|C)=9/15\times(-7/9\times\log\downarrow2\ 7/9\ -2/9\times\log\downarrow2\ 2/9\ )+6/15\times(-6/6\times\log\downarrow2\ 6/6\ )=0.459$$

$$Infor(Z|D)=2/15\times(-2/2\times\log\downarrow2\ 2/2\ )+2/15\times(-1/2\times\log\downarrow2\ 1/2\ -1/2\times\log\downarrow2\ 1/2\ )+3/15\times(-3/3\times\log\downarrow2\ 3/3\ )+6/15\times(-5/6\times\log\downarrow2\ 5/6\ -1/6\times\log\downarrow2\ 1/6\ )+2/15\times(-2/2\times\log\downarrow2\ 2/2\ )=0.393$$

②根据训练数据构造ID3算法的决策树，其中Z代表训练集，A、B、C、D分别代表特征“年龄”、“吸烟史”、“有无家族病史”、“体重范围”，按照每个特征计算其分裂的信息增益。

$$gZA=Infor(Z)-Infor(Z|A)=0.997-0.857=0.140$$

$$gZB=Infor(Z)-Infor(Z|B)=0.997-0.778=0.219$$

$$gZC=Infor(Z)-Infor(Z|C)=0.997-0.459=0.538$$

$$gZD=Infor(Z)-Infor(Z|D)=0.997-0.393=0.604$$

选择信息增益最大特征“体重范围”作为根结点的分裂属性，将训练集Z划分为5个子集$Z\downarrow1$、$Z\downarrow2$、$Z\downarrow3$、$Z\downarrow4$和$Z\downarrow5$，对应的“体重范围”取值分别为“低”、“较低”、“中”、“较高”、“高”。由于$Z\downarrow2$、$Z\downarrow3$和$Z\downarrow4$只有一类数据，所以它们各自成为一个叶结点，三个结点的类标签分别为“否”、“否”、“是”。

表 Z_1训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
4	青年	0-5年	有	低	是
6	中年	无	无	低	否

表 Z_2训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
1	青年	无	无	较低	否
5	中年	无	无	较低	否

表 Z_3训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
2	青年	无	无	中	否
3	青年	0-5年	无	中	否
15	老年	0-5年	无	中	否

表 Z_4训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
11	老年	无	有	较高	是
14	老年	5年以上	无	较高	是

表 Z_5训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
7	中年	无	无	高	否
8	中年	5年以上	有	高	是
9	中年	无	有	高	是
10	中年	无	有	高	是
12	老年	无	有	高	是
13	老年	5年以上	无	高	是

③对于Z_1继续进行分裂，选择剩余特征中信息增益最大的作为分裂属性。

$$Infor(Z_1)=-1/2\times\log_2 1/2-1/2\times\log_2 1/2=1$$

$$Infor(Z_1|A)=1/2\times(-1/1\times\log_2 1/1)+1/2\times(-1/1\times\log_2 1/1)=0$$

$$Infor(Z_1|B)=1/2\times(-1/1\times\log_2 1/1)+1/2\times(-1/1\times\log_2 1/1)=0$$

$$Infor(Z_1|C)=1/2\times(-1/1\times\log_2 1/1)+1/2\times(-1/1\times\log_2 1/1)=0$$

③对于Z_1继续进行分裂，选择剩余特征中信息增益最大的作为分裂属性。

$$gZ_1 A=Infor(Z_1)-Infor(Z_1|A)=1-0=1$$

$$gZ_1 B=Infor(Z_1)-Infor(Z_1|B)=1-0=1$$

$$gZ_1 C=Infor(Z_1)-Infor(Z_1|C)=1-0=1$$

选择信息增益最大特征作为Z_1结点的分裂属性，由于三个属性的信息增益相同，随机挑选一个作为分裂属性，此处选取“有无家族病史”作为分裂属性，它将数据集Z_1分成2个子集Z_{21}和Z_{22}，对应的“有无家族病史”的取值分别为“有”和“无”，在这2个子集中的数据都各自属于同一类，于是就不需要再继续分裂。

表 Z_{21}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
6	中年	无	无	低	否

表 Z_{22}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
4	青年	0-5年	有	低	是

④对于Z_5继续进行分裂，选择剩余特征中信息增益最大的作为分裂属性。

$$Infor(Z_5)=-1/6\times\log_2 1/6-5/6\times\log_2 5/6=0.650$$

$$Infor(Z_5|A)=4/6\times(-1/4\times\log_2 1/4-3/4\times\log_2 3/4)+2/6\times(-2/2\times\log_2 2/2)=0.541$$

$$Infor(Z_5|B)=4/6\times(-3/4\times\log_2 3/4-1/4\times\log_2 1/4)+2/6\times(-2/2\times\log_2 2/2)=0.541$$

$$Infor(Z_5|C)=4/6\times(-4/4\times\log_2 4/4)+2/6\times(-1/2\times\log_2 1/2-1/2\times\log_2 1/2)=0.333$$

④对于Z_5继续进行分裂，选择剩余特征中信息增益最大的作为分裂属性。

$$gZ_5 A=Infor(Z_5)-Infor(Z_5|A)=0.650-0.541=0.109$$

$$gZ_5 B=Infor(Z_5)-Infor(Z_5|B)=0.650-0.541=0.109$$

$$gZ_5 C=Infor(Z_5)-Infor(Z_5|C)=0.650-0.333=0.317$$

选择信息增益最大特征“有无家族病史”作为Z_5结点的分裂属性，将数据集Z_5分成2个子集Z_{51}和Z_{52}，对应的“有无家族病史”的取值分别为“有”和“无”，“有”对应的Z_{51}中的数据都属于同一类，不需要再继续分裂，“无”对应的Z_{52}中的数据属于不同类，需要对其进行分裂。

表 Z_{S1}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
8	中年	5年以上	有	高	是
9	中年	无	有	高	是
10	中年	无	有	高	是
12	老年	无	有	高	是

表 Z_{S2}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
7	中年	无	无	高	否
13	老年	5年以上	无	高	是

⑤对于$ZI52$继续进行分裂，选择剩余特征中信息增益最大的作为分裂属性。

$Infor(ZI52)=-1/2\times\log\downarrow2\ 1/2\ -1/2\times\log\downarrow2\ 1/2\ =1$

$Infor(ZI52|A)=1/2\times(-1/1\times\log\downarrow2\ 1/1\ )+1/2\times(-1/1\times\log\downarrow2\ 1/1\ )=0$

$Infor(ZI52|B)==1/2\times(-1/1\times\log\downarrow2\ 1/1\ )+1/2\times(-1/1\times\log\downarrow2\ 1/1\ )=0$

$gZI52\ A=Infor(ZI51)-Infor(ZI51|A)=1-0=1$

$gZI52\ B=Infor(ZI51)-Infor(ZI51|B)=1-0=1$

由于两个属性的信息增益相同，随机挑选一个作为分裂属性，此处选取“吸烟史”作为$ZI52$结点的分裂属性，将数据集$ZI52$分成2个子集$ZI521$和$ZI522$，对应的“吸烟史”的取值分别为“无”和“5年以上”，在这两个子集中的数据都各自属于同一类，于是就不需要再继续分裂，决策树构造完毕。

表 Z_{S21}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
7	中年	无	无	高	否

表 Z_{S22}训练数据集

ID	年龄	吸烟史	有无家族病史	体重范围	患病与否
13	老年	5年以上	无	高	是

利用ID3算法构建的决策树如图7-5所示。

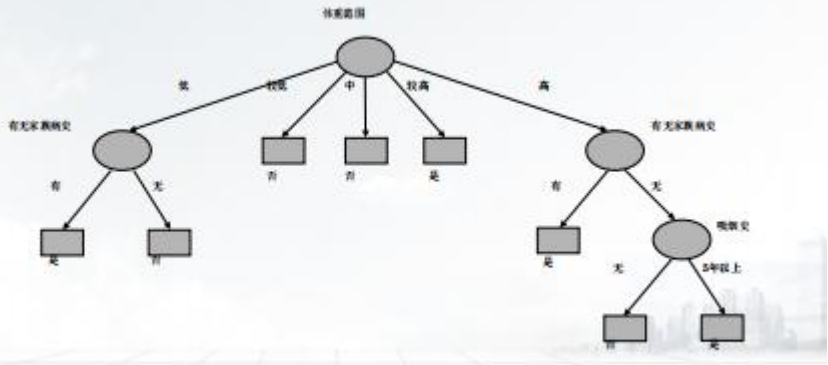


图7-5 ID3算法构造的决策树

根据构建的决策树，可以提取分类规则如下。

①IF “体重范围” = “低” AND “有无家族病史” = “有” THEN 患病

②IF “体重范围” = “低” AND “有无家族病史” = “无” THEN 没有患病

③IF “体重范围” = “较低” THEN 没有患病

④IF “体重范围” = “中” THEN 没有患病

⑤IF “体重范围” = “较高” THEN 患病

⑥IF “体重范围” = “高” AND “有无家族病史” = “有” THEN 患病

⑦IF “体重范围” = “高” AND “有无家族病史” = “无” AND “吸烟史” = “无” THEN 没有患病

⑧IF “体重范围” = “高” AND “有无家族病史” = “无” AND “吸烟史” = “5年以上” THEN 患病

7.3

计算：

例朴素贝叶斯分类算法实例：

训练数据如表7-7所示，其中$X\uparrow(1)$和$X\uparrow(2)$为特征，取值分别来自特征集合$A\downarrow1=\{1,2,3\}$，$A\downarrow2=\{S,P,Q\}$，C为类标记，$C=\{1,-1\}$，即有1和-1两类。根据训练数据学习一个朴素贝叶斯分类器并确定$X=(2,S)\uparrow T$的类标记。

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
	S	P	P	S	S	S	P	P	Q	Q	Q	P	P	Q	Q
C	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

- 计算先验概率：

$$P(C=1)=9/15\ ,\ P(C=-1)=6/15$$

- 计算条件概率：

$$P(X\uparrow(1)=1|C=1)=2/9\ ,\ P(X\uparrow(1)=2|C=1)=3/9\ ,\ P(X\uparrow(1)=3|C=1)=4/9$$
$$P(X\uparrow(2)=S|C=1)=1/9\ ,\ P(X\uparrow(2)=P|C=1)=4/9\ ,\ P(X\uparrow(2)=Q|C=1)=4/9$$
$$P(X\uparrow(1)=1|C=-1)=3/6\ ,\ P(X\uparrow(1)=2|C=-1)=2/6\ ,\ P(X\uparrow(1)=3|C=-1)=1/6$$
$$P(X\uparrow(2)=S|C=-1)=3/6\ ,\ P(X\uparrow(2)=P|C=-1)=2/6\ ,\ P(X\uparrow(2)=Q|C=-1)=1/6$$

- 对于给定的$x=(2,S)\uparrow T$，依照分类器模型计算：

$$P(C=1|X)=P(C=1)\cdot P(X\uparrow(1)=2|C=1)\cdot P(X\uparrow(2)=S|C=1)=9/15\cdot3/9\cdot1/9=1/45$$
$$P(C=-1|X)=P(C=-1)\cdot P(X\uparrow(1)=2|C=-1)\cdot P(X\uparrow(2)=S|C=-1)=6/15\cdot2/6\cdot3/6=1/15$$

$X=(2,S)\uparrow T$属于-1类别的概率最大，依照朴素贝叶斯中概率最大化准则，该分类器输出的类标记为-1。

7.5

神经网络是一组连接的输入/输出单 元，其中每个连接都与一个权重相关联。在学习阶段，通过调整这些权重，能够正确预测输入样本的类标号。

神经网络由三个要素组成：拓扑结构、连接方式和学习规则。

拓扑结构

- 拓扑结构可以分为单层、两层或者三层。其中单层神经网络只有一组输入单元和一个输出单元。两层神经网络由输入单元层和输出单元层组成。三层神经网络用于处理更复杂的非线性问题。在这种模型中，除了输入层和输出层外，还引入了中间层，也称为隐藏层，隐藏层可以有一层或多层。每层单元的输出作为下一层单元的输入，神经网络的拓扑结果如下图所示。

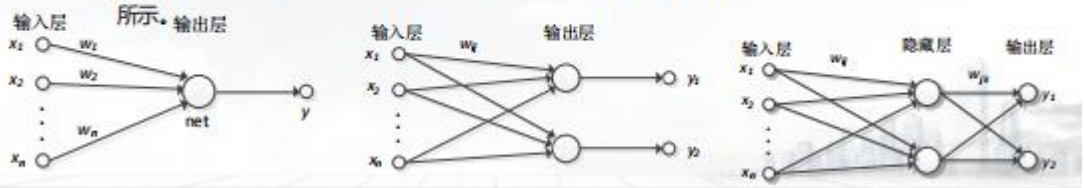


图7-57(a) 单层神经网络的拓扑结构　图7-57(b) 两层神经网络的拓扑结构　图7-57(c) 三层神经网络的拓扑结构

7.6

计算：

二分类问题，即分类目标只有两类，正类（positive）即感兴趣的主要类和负类（negtive）即其他类，正例即为正类的实例或元组，负例即为负类的实例或元组。

表7-13 二分类的混淆矩阵

实例分类 \ 预测类	正类	负类	合计
正类	TP	FN	P
负类	FP	TN	N
合计	P'	N'	P+N

①真正例(True Positives，*TP*):被正确地划分为正类的实例数，即实际为正例且被分类器划分为正例的实例数。
②假正例(False Positives，*FP*):被错误地划分为正类的实例数，即实际为负例但被分类器划分为正例的实例数。
③假负例(False Negatives，*FN*):被错误地划分为负类的实例数，即实际为正例但被分类器划分为负例的实例数。
④真负例(True Negatives，*TN*):被正确地划分为负类的实例数，即实际为负例且被分类器划分为负例的实例数。

1. 分类模型的评价指标

①准确率（accuracy）

又称为分类器的总体识别率，准确率表示分类器对各类元组的正确识别情况，它定义为被正确分类的元组数占预测总元组数的百分比，计算公式如（7-59）所示。

$$accuracy=TP+TN/P+N \qquad (7\text{-}59)$$

②错误率（error rate）

又称为误分辨率，错误率表示分类器对各类元组的错误识别情况，是1-accuracy ，具体计算公式如（7-60）所示。

$$error\ rate=FP+FN/P+N \qquad (7\text{-}60)$$

③特效性（specificity）

又称为真负例识别率(True Negative Rate，TNR)，特效性表示分类器对负元组的正确识别情况，它定义为正确识别的负元组数量占实际为负元组总数的百分比，计算公式如（7-61）所示。

$$specificity=TN/N \quad (7\text{-}61)$$

④灵敏度（sensitivity）

灵敏度也被称为真正例识别率(True Positive Rate，TPR)，即正确识别的正元组的百分比，衡量了分类器对正类的识别能力，具体计算公式如（7-63）所示。

$$sensitivity=TP/P \quad (7\text{-}63)$$

⑤精度（precision）：精度可以看做精确性的度量，即正确识别的正元组数量占预测为正元组总数的百分比

⑥召回率（recall）：召回率用来评价模型的灵敏度和识别率，是完全性的度量，即正元组被标记为正类的百分比，即为灵敏度（或真正例率）

⑦综合评价指标（F 度量）：将精度和召回率组合到一个度量中，即为 F 度量（又称为 F1 分数或 F 分数）和 Fβ 度量的方法，

计算:

例7.10 分类模型评价指标的计算

混淆矩阵如表7-14所示，计算准确率、错误率、特效性等评价指标。

表7-14混淆矩阵实例

实例分类 \ 预测类	正类	负类	合计
正类	150	40	190
负类	60	250	310
合计	210	290	500

解：

①计算准确率。

$$accuracy=150+250/500\times100\%=80\%$$

②计算错误率。

$$errorrate=1-accuracy=20\%$$

③计算特效性。

$$specificity=250/310\times100\%\approx80.65\%$$

④计算灵敏度。

$$sensitivity=150/190\times100\%=78.94\%$$

⑤计算精度。

$$precision=150/150+60=150/210\times100\%=71.42\%$$

交叉验证的方法：

①留出法(holdout cross validation)

②k 折交叉验证(k-fold Cross Validation)

③留一法(Leave-One-Out Cross Validation)

8. 第八章

8.5

聚类分析(cluster analysis)简称聚类(clustering)，是一个把数据对象(或观测)划分成子集的过程。

每个子集是一个**簇(cluster)**，使得簇中的对象彼此相似，但与其他簇中的对象不相似。

数据挖掘对聚类的典型要求如下：

- 处理不同属性类型的能力
- 可伸缩性
- 对于确定输入参数的领域知识的要求
- 发现任意形状的簇

聚类过程遵循的基本步骤：

- 特征选择
- 近邻测度
- 准则定义
- 算法调用
- 结果验证
- 结果判定

基本聚类方法：基于划分的方法、基于层次的方法、基于密度的方法、基于网格的方法。

k-均值算法的基本过程：

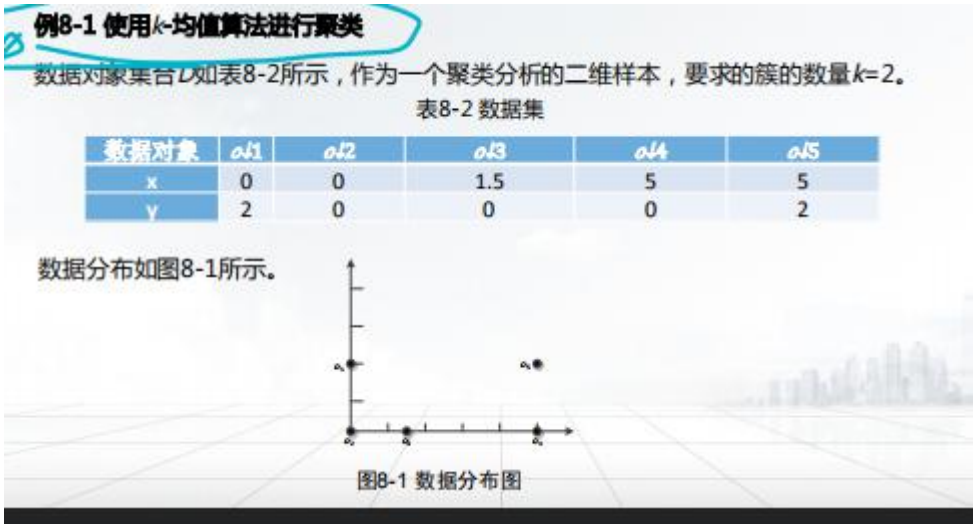
①首先输入 k 的值，即具有 n 个对象的数据集/span>={ 1， 2,…， }经过聚类将得到 k 个分类或分组。

②从数据集 D 中随机选择 k 个对象作为簇质心，每个簇质心代表一个簇，得到的簇质心集合为={ 1， 2,…， }。

③对 D 中每一个对象 ，计算 与 (/span>=1,2,…,/span>)的距离，得到一组距离值，选择最小距离值对应的簇质心 c↓，则将对象 划分到以 c↓为质心的簇中。

④根据每个簇所包含的对象集合，重新计算簇中所有对象的平均值得到一个新的簇质心，返回步骤③，直到簇质心不再变化。

计算：（看不懂，看不懂）



解：

①任意选择$o_1(0,2)$，$o_2(0,0)$为簇C_1和C_2初始的簇质心，即$c_1=o_1=(0,2)$，$c_2=o_2=(0,0)$。

②对剩余的每个对象，根据其与各个簇质心的距离，将它赋给最近的簇。

o_3：$d(c_1,o_3)=\sqrt{(0-1.5)^2+(2-0)^2}=2.5$；　$d(c_2,o_3)=\sqrt{(0-1.5)^2+(0-0)^2}=1.5$

显然，$d(c_1,o_3)>d(c_2,o_3)$,故将o_3分配给C_2。

o_4：$d(c_1,o_4)=\sqrt{(0-5)^2+(2-0)^2}=\sqrt{29}$；　$d(c_2,o_4)=\sqrt{(0-5)^2+(0-0)^2}=5$

显然，$d(c_1,o_4)>d(c_2,o_4)$，故将o_4分配给C_2。

o_5：$d(c_1,o_5)=\sqrt{(0-5)^2+(2-2)^2}=5$；　$d(c_2,o_5)=\sqrt{(0-5)^2+(2-2)^2}=\sqrt{29}$

显然，$d(c_1,o_5)<d(c_2,o_5)$,故将o_5分配给C_1。

更新，得到两个簇：$C_1=\{o_1,o_5\}$ $C_2=\{o_2,o_3,o_4\}$

计算每个簇的平方误差准则:

$E_1=[(0-0)^2+(2-2)^2]+[(5-0)^2+(2-2)^2]=25$

$E_2=[(0-0)^2+(0-0)^2]+[(1.5-0)^2+(0-0)^2]+[(5-0)^2+(0-0)^2]=27.25$

$E=E_1+E_2=25+27.25=52.25$

形成的两个簇如图8-2所示。

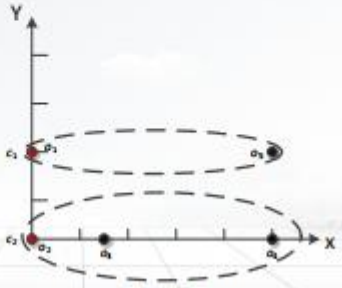


图8-2 初始簇划分

③计算新簇质心

$c_1=[0+5/2,2+2/2]=(2.5,2)$　$c_2=[0+1.5+5/3,0+0+0/3]=(2.17,0)$

④重复②，对每个对象，根据其与各个簇新质心的距离，将它赋给最近的簇。

o_1：$d(c_1,o_1)=\sqrt{(2.5-0)^2+(2-2)^2}=2.5$　$d(c_2,o_1)=\sqrt{(2.17-0)^2+(0-2)^2}=2.95$

显然，$d(c_1,o_1)<d(c_2,o_1)$,故将o_1分配给C_1。

o_2：$d(c_1,o_2)=\sqrt{(2.5-0)^2+(2-0)^2}=3.20$　$d(c_2,o_2)=\sqrt{(2.17-0)^2+(0-0)^2}=2.17$

显然，$d(c_1,o_2)>d(c_2,o_2)$,故将o_2分配给C_2。

o_3：$d(c_1,o_3)=\sqrt{(2.5-1.5)^2+(2-0)^2}=2.24$　$d(c_2,o_3)=\sqrt{(2.17-1.5)^2+(0-0)^2}=0.67$

显然，$d(c_1,o_3)>d(c_2,o_3)$,故将o_3分配给C_2。

④重复②，对每个对象，根据其与各个簇新质心的距离，将它赋给最近的簇。

$o4$ ：$d(c1,o4)=\sqrt{(2.5-5)^2+(2-0)^2}=3.20$　$d(c2,o4)=\sqrt{(2.17-5)^2+(0-0)^2}=2.83$

显然，$d(c1,o4)>d(c2,o4)$，故将$o4$ 分配给$C2$。

$o5$ ：$d(c1,o5)=\sqrt{(2.5-5)^2+(2-2)^2}=2.5$　$d(c2,o5)=\sqrt{(2.17-5)^2+(0-2)^2}=3.47$

显然，$d(c1,o5)<d(c2,o5)$，故将$o5$ 分配给$C1$。

更新，得到两个新簇：$C1=\{o1,o5\}$，$C2=\{o2,o3,o4\}$

计算每个簇的平方误差准则：

$E1=[(0-2.5)^2+(2-2)^2]+[(5-2.5)^2+(2-2)^2]=12.5$

$E2=[(0-2.17)^2+(0-0)^2]+[(1.5-2.17)^2+(0-0)^2]+[(5-2.17)^2+(0-0)^2]=13.17$

$E=E1+E2=12.5+13.17=25.67$

形成的两个簇如图8-3所示。

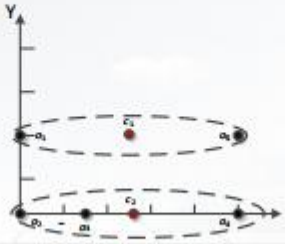


图8-3 更新簇划分

K-均值算法的优点：

① 擅长处理球状分布的数据，当结果聚类是密集的，而且类和类之间的区别比较明显时，k-均值聚类算法的效果比较好。

② 对于处理大数据集，是相对可伸缩的和高效的，它的复杂度是 O(nkt),n 是对象的个数，k 是簇的数目，t 是迭代的次数。

③ 相比其他的聚类算法，k-均值聚类算法比较简单、容易掌握。

K-均值算法的缺点：

① 初始质心的选择与算法的运行效率密切相关

② 要求用户事先给定簇数 k

③ 对噪声和离群点敏感，少量的这类数据对结果产生很大影响