第二章:

1、相似性度量计算

- SMC和Jaccard系数

$x = (1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)$

$y = (0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1)$

SMC = (F11 + F00) / (F01 + F10 + F11 + F00)

= (0+7) / (2+1+0+7) = 0.7

$J = (F_{11}) / (F_{01} + F_{10} + F_{11}) = 0 / (2 + 1 + 0) = 0$

2、相异性度量计算

•Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**
d_2 = **1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$||d_1||$ = $(3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5}$ = $(42)^{0.5}$ = 6.481

$||d_2||$ = $(1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5}$ = $(6)^{0.5}$ = 2.245

cos(d_1, d_2) = .3150

3、相关系数计算

- 对于**非线性的相关性**难以建模
- X = ($\boxed{-3, -2, -1,}$ 0, $\boxed{1, 2, 3}$)
- Y = ($\boxed{9, 4, 1,}$ 0, $\boxed{1, 4, 9}$)

$Y = X^2$

负相关　　正相关

- Mean(X) = 0, Mean(Y) = 4

- 皮尔森相关系数 Correlation

= (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5)

= **-15 + 0 + 3 +** 0 **- 3 + 0 + 15**

= 0 (即相关度为0)

第三章：线性回归

1、LDA 计算例题

由两类二维数据计算线性判别分析(LDA)投影向量

第一类采样数据ω_1： $\mathbf{X_1}=(x_1,x_2)=\{(4,2),(2,4),(2,3),(3,6),(4,4)\}$ (红色点)

第二类采样数据ω_2： $\mathbf{X_2}=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$ (蓝色点)

两个类的均值为：

$$\mu_1=\frac{1}{N_1}\sum_{x\in\omega_1}x=\frac{1}{5}\left[\begin{pmatrix}4\\2\end{pmatrix}+\begin{pmatrix}2\\4\end{pmatrix}+\begin{pmatrix}2\\3\end{pmatrix}+\begin{pmatrix}3\\6\end{pmatrix}+\begin{pmatrix}4\\4\end{pmatrix}\right]=\begin{pmatrix}3\\3.8\end{pmatrix}$$

$$\mu_2=\frac{1}{N_2}\sum_{x\in\omega_2}x=\frac{1}{5}\left[\begin{pmatrix}9\\10\end{pmatrix}+\begin{pmatrix}6\\8\end{pmatrix}+\begin{pmatrix}9\\5\end{pmatrix}+\begin{pmatrix}8\\7\end{pmatrix}+\begin{pmatrix}10\\8\end{pmatrix}\right]=\begin{pmatrix}8.4\\7.6\end{pmatrix}$$

第一类样本的**类内散度矩阵**为：

$$S_1=\sum_{x\in\omega_1}(x-\mu_1)(x-\mu_1)^T=\left[\begin{pmatrix}4\\2\end{pmatrix}-\begin{pmatrix}3\\3.8\end{pmatrix}\right]^2+\left[\begin{pmatrix}2\\4\end{pmatrix}-\begin{pmatrix}3\\3.8\end{pmatrix}\right]^2$$

$$+\left[\begin{pmatrix}2\\3\end{pmatrix}-\begin{pmatrix}3\\3.8\end{pmatrix}\right]^2+\left[\begin{pmatrix}3\\6\end{pmatrix}-\begin{pmatrix}3\\3.8\end{pmatrix}\right]^2+\left[\begin{pmatrix}4\\4\end{pmatrix}-\begin{pmatrix}3\\3.8\end{pmatrix}\right]^2$$

$$=\begin{pmatrix}1&-0.25\\-0.25&2.2\end{pmatrix}$$

第二类样本的**类内散度矩阵**为：

$$S_2=\sum_{x\in\omega_2}(x-\mu_2)(x-\mu_2)^T=\left[\begin{pmatrix}9\\10\end{pmatrix}-\begin{pmatrix}8.4\\7.6\end{pmatrix}\right]^2+\left[\begin{pmatrix}6\\8\end{pmatrix}-\begin{pmatrix}8.4\\7.6\end{pmatrix}\right]^2$$

$$+\left[\begin{pmatrix}9\\5\end{pmatrix}-\begin{pmatrix}8.4\\7.6\end{pmatrix}\right]^2+\left[\begin{pmatrix}8\\7\end{pmatrix}-\begin{pmatrix}8.4\\7.6\end{pmatrix}\right]^2+\left[\begin{pmatrix}10\\8\end{pmatrix}-\begin{pmatrix}8.4\\7.6\end{pmatrix}\right]^2$$

$$=\begin{pmatrix}2.3&-0.05\\-0.05&3.3\end{pmatrix}$$

$$S_w = S_1 + S_2 = \begin{pmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{pmatrix} + \begin{pmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{pmatrix}$$

$$= \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}$$

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$

$$= \left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right] \left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]^T$$

$$= \begin{pmatrix} -5.4 \\ -3.8 \end{pmatrix} \begin{pmatrix} -5.4 & -3.8 \end{pmatrix}$$

$$= \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix}$$

```
% between-cla
SB = (Mu1-Mu2
```

$$w^* = S_W^{-1}(\mu_1 - \mu_2) = \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}^{-1} \left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]$$

$$= \begin{pmatrix} 0.3045 & 0.0166 \\ 0.0166 & 0.1827 \end{pmatrix} \begin{pmatrix} -5.4 \\ -3.8 \end{pmatrix}$$

$$= \begin{pmatrix} 0.9088 \\ 0.4173 \end{pmatrix}$$

2、向量和矩阵求导

□ 几种特殊类型函数的梯度

1: $f(x) = b^T x + c \Longrightarrow \nabla f(x) = b$

2: $f(x) = x^T x \Longrightarrow \nabla f(x) = 2x$

3: $f(x) = x^T A x \Longrightarrow \nabla f(x) = (A + A^T)x$

$\nabla f(x) = (A + A^T)x = 2Ax$ （若A对称）

4: $f(x) = x^T A x + b^T x + c \Longrightarrow \nabla f(x) = (A + A^T)x + b$

问题: $\mathbf{x}=(x_i)_{n\times 1}\in\mathbb{R}^{n\times 1}$, $\mathbf{A}=(a_{ij})_{n\times n}\in\mathbb{R}^{n\times n}$, $f(\mathbf{x})=\mathbf{x}^T\mathbf{A}\mathbf{x}$ 为一个标量函数, 求 $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$.

解: 由标量值函数对向量的导数定义可知, $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}=\left[\frac{\partial f(\mathbf{x})}{\partial x_1},\ldots,\frac{\partial f(\mathbf{x})}{\partial x_k},\ldots,\frac{\partial f(\mathbf{x})}{\partial x_n}\right]^T$,

$$f(\mathbf{x})=\mathbf{x}^T\mathbf{A}\mathbf{x}=[x_1,\ldots,x_k,\ldots,x_n]^T\begin{bmatrix}a_{11} & a_{12} & \cdots a_{1k} & a_{1n}\\ \vdots & \vdots & \ddots\vdots & \vdots\\ a_{k1} & a_{k2} & \cdots a_{kk} & a_{kn}\\ \vdots & \vdots & \ddots\vdots & \vdots\\ a_{n1} & a_{n2} & \cdots a_{nk} & a_{nn}\end{bmatrix}\begin{bmatrix}x_1\\ \vdots\\ x_k\\ \vdots\\ x_n\end{bmatrix}$$

$$=[a_{11}x_1+\ldots+a_{k1}x_k+\ldots+a_{n1}x_n,\ldots,a_{1n}x_1+\ldots+a_{kn}x_k+\ldots+a_{nn}x_n]\begin{bmatrix}x_1\\ \vdots\\ x_k\\ \vdots\\ x_n\end{bmatrix}$$

$$=a_{11}x_1^2+a_{21}x_2x_1+\ldots+a_{n1}x_nx_1+a_{12}x_1x_2+a_{22}x_2^2+\ldots+a_{n2}x_nx_2+\ldots$$
$$=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_ix_j$$
$$=\sum_{i=1,i=j}^{n}a_{ii}x_i^2+\sum_{i=1,i\neq j}^{n}\sum_{j=1}^{n}a_{ij}x_ix_j=\sum_{i=1,i=j}^{n}a_{ii}x_i^2+\sum_{i=1,i\neq j}^{n}x_i\left(\sum_{j=1}^{n}a_{ij}x_j\right)$$

取 $\mathbf{x}$ 的第 k 个分量 x_k 为代表, 求 $f(\mathbf{x})$ 关于 x_k 的偏导数.

$$\frac{\partial f(\mathbf{x})}{\partial x_k}=2a_{kk}x_k+\sum_{i=1,i\neq j}^{n}x_i\left(\frac{\partial}{\partial x_k}\sum_{j=1}^{n}a_{ij}x_j\right)+\sum_{j=1,j\neq i}^{n}x_j\left(\frac{\partial}{\partial x_k}\sum_{i=1}^{n}a_{ij}x_i\right)$$
$$=2a_{kk}x_k+\sum_{i=1,i\neq k}^{n}x_ia_{ik}+\sum_{j=1,j\neq k}^{n}a_{kj}x_j$$
$$=(\mathbf{A}^T\mathbf{x})_k+(\mathbf{A}\mathbf{x})_k$$
$$=\left[(\mathbf{A}^T+\mathbf{A})\mathbf{x}\right]_k$$

第四章：决策树（无计算，这里列出）

第五章：神经网络

一、样本迭代次数和 epoch 计算

- 假设训练集有 2560000个样本。现在选择 Batch size = 256 对模型进行训练。
- 总的迭代（iteration）次数： 2560000/256 = 10000
- 每个 Epoch 要训练的样本数量： 2560000
- 需要10000次iteration完成一个epoch
- 不同epoch的训练，其实用的是同一个训练集的数据。第 1个 epoch和第个10 epoch虽然用的都是训练集的2560000 个样本，但是对模型的权重更新值却是完全不同的。因为不同epoch的模型处于代价函数空间上的不同位置，模型的训练epoch越靠后，越接近谷底，其代价越小

二、计算（特征图，参数，时间复杂度，空间复杂度）

卷积后输出特征图大小计算

输入图像大小：$\mathrm{H}_{in} \times W_{in} \times n_c$

每个滤波器(卷积核)大小：$k \times k \times n_c$

滤波器(卷积核)个数：K

加边填充 padding: P

卷积核滑动步幅(stride): S

输出特征图像大小：

$$\boxed{\underbrace{\left(\frac{H_{in}-k+2P}{S}+1\right)}_{\text{高}} \times \underbrace{\left(\frac{W_{in}-k+2P}{S}+1\right)}_{\text{宽}} \times \underbrace{K}_{\text{通道数}}}$$

计算卷积层输出特征图大小，当除不尽时，一般向**下**取整。

池化后输出特征图的大小计算

- 输入图像大小：$\mathrm{H}_{in} \times W_{in} \times n_c$
- 每个滤波器(卷积核)大小：$k \times k \times n_c$
- 滤波器(卷积核)个数：K
- 卷积滑动步幅(stride): S
- 加边填充 padding: P

□ **输出图像大小：**

$$\boxed{\left(\frac{H_{in}-k+2P}{S}+1\right) \times \left(\frac{W_{in}-k+2P}{S}+1\right) \times K}$$

计算池化层输出特征图大小，当除不尽时，通常向**上**取整。

- 1. Input

  输入图像统一归一化为32*32。
- 2. C1卷积层

  经过(5*5*1)*6卷积核，stride=1，pad=0，生成feature map为28*28*6。
- 3. S2池化层

  经过(2*2)池化核，平均池化，stride=2，pad=0，生成feature map为14*14*6。
- 4. C3卷积层

  经过(5*5*6)*16卷积核，stride=1，pad=0，生成feature map为10*10*16。
- 5. S4池化层

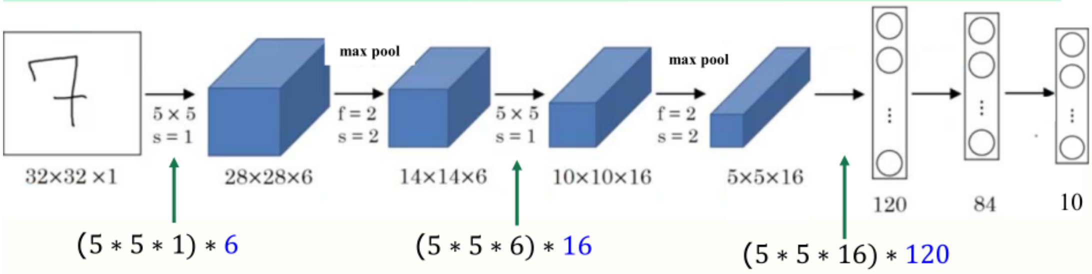
  经过(2*2)池化核，平均池化， stride=2，pad=0，生成feature map为5*5*16。
- 6. C5卷积层

  经过(5*5*16)*120卷积核，stride=1，pad=0，生成feature map为1*1*120。
- 7. F6全连接层

  输入为1*1*120，输出为1*1*84，总参数量为120*84。
- 8. F7 全连接层 (输出层)

  输入为1*1*84，输出为1*1*10，总参数量为84*10。10就是分类的类别数。输出层激活函数是 softmax。
- 中间隐层激活函数是ReLU

LeNet-5 要估计的参数量（空间复杂度）



待估计的权重参数量：

$5*5*1*6+5*5*6*16+5*5*16*120+120*84+84*10$

待估计全部参数量 (权重+偏置)：

$5*5*1*6$(卷积核)$+6$(偏置)$+5*5*6*16$(卷积核)$+16$(偏置)$+$
$5*5*16*120$(卷积核)$+120$(偏置)
$+120*84$(全连接权重)$+84$(偏置)$+84*10$(全连接权重)$+10$(偏置)

注意：CNN中权重和偏置共享

CNN的浮点计算量 (时间复杂度)

- 衡量卷积计算量的指标是FLOPs（Floating Point Operations，浮点运算次数）
- 一次乘法或一次加法表示一个浮点运算次数
- CNN 中单个卷积层的乘法和加法浮点运算次数：

$$[(k \times k \times n_c) + (k \times k \times n_c - 1) + 1] \times \mathrm{H}_{out} \times W_{out} \times K$$

- 卷积核每滑动一次的乘法浮点计算量： $k \times k \times n_c$
- 卷积核每滑动一次的加法浮点计算量： $k \times k \times n_c - 1$
- 输出单个特征图的卷积**乘法**浮点计算量：$(k \times k \times n_c) \times \mathrm{H}_{out} \times W_{out}$
- 输出单个特征图的卷积**加法**浮点计算量：$(k \times k \times n_c - 1) \times \mathrm{H}_{out} \times W_{out}$
- 输出 K 个特征图的卷积乘法浮点计算量：$(k \times k \times n_c) \times \mathrm{H}_{out} \times W_{out} \times K$
- 输出 K 个特征图的卷积加法浮点计算量：$(k \times k \times n_c - 1) \times \mathrm{H}_{out} \times W_{out} \times K$
- 输出单个特征图的偏置浮点加法计算量：$\mathrm{H}_{out} \times W_{out}$
- 输出 K 个特征图的偏置浮点加法计算量：$\mathrm{H}_{out} \times W_{out} \times K$

CNN 的浮点计算量 (时间复杂度)

- 单个卷积层的乘法和加法的浮点计算量：

$$FLOPs = 2 \times k \times k \times n_c \times \mathrm{H}_{out} \times W_{out} \times K$$

- 上式是乘法和加法运算的总和，将一次乘运算或加运算都视作一次浮点运算
- 在计算机视觉论文中，**常常将一个'乘-加'组合视为一次浮点运算，**英文表述为'Multi-Add'，运算量正好是上面的算法减半，此时的运算量为：

$$FLOPs = k \times k \times n_c \times \mathrm{H}_{out} \times W_{out} \times K$$

全连接层的浮点计算量 (FLOPs) 和参数量 (parameters)

示例

$$a_1 = W_1 1 * x_1 + W_1 2 * x_2 + W_1 3 * x_3 + b_1$$

$$a_2 = W_2 1 * x_1 + W_2 2 * x_2 + W_2 3 * x_3 + b_2$$

$$a_3 = W_3 1 * x_1 + W_3 2 * x_2 + W_3 3 * x_3 + b_3$$

$$\boldsymbol{a}^l = \sigma\left(\boldsymbol{W}^l \boldsymbol{a}^{l-1} + \boldsymbol{b}^l\right)$$

- 单个全连接层的乘法和加法浮点计算量（权重+偏置）：

$$FLOPs = [N_{in} + (N_{in} - 1) + 1] \times N_{out} = 2 \times N_{in} \times N_{out}$$

- 其中 N_{in} 表示输入层神经元个数， N_{out} 表示输出层神经元个数。上述式子中第一个 N_{in} 表示乘法运算量， N_{in} -1 表示加法运算量，+1表示 N_{out}个偏置项计算量，$\times N_{out}$表示计算 N_{out}个神经元的值。

- 单层全连接层的网络模型参数量（权重+偏置）：

$$parameters = (N_{in} + 1) \times N_{out}$$

全连接层的浮点计算量（FLOPs）

- **如果将一个'乘-加'组合视为一次浮点运算，** 则此时单个全连接层的浮点运算量为：

$$FLOPs = N_{in} \times N_{out}$$

- 其中 N_{in} 表示输入层神经元个数，N_{out} 表示输出层神经元个数。

第七章：关联分析

频繁项集 $L_1 = \{1, 2, 3, 4, 5\}$ $L_2 = \{\{1, 2\}, \{2, 3\}\}$

C_3 候选项集产生方法1： 频繁1-项集与频繁2-项集进行连接

$$\{X \cup p \mid X \in L_k, p \in L_1, p \notin X\}$$

候选项集 $C_3 = \{\{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}, \{2, 3, 4\}, \{2, 3, 5\}\}$ 不紧凑

非频繁 {1,3}不在L_2中

非频繁 {1,5}不在L_2中

C_3 候选项集产生方法2：频繁2-项集与其自身进行连接

条件：X 和 Y 只有一位不同

$$\{X \cup Y \mid X, Y \in L_k, |X \cap Y| = k-1\}$$

候选项集 $C_3 = \{\{1, 2, 3\}\}$ 不紧凑

非频繁 (因{1,3}非频繁)

Apriori 算法例子,挖掘频繁项集, 要求最小支持度=50%(即支持度计数

数据库 D

TID	项集
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

扫描 D

C_1

项集	支持度计数
{1}	2
{2}	3
{3}	3
~~{4}~~	~~1~~
{5}	3

L_1

项集	支持度计数
{1}	2
{2}	3
{3}	3
{5}	3

C_2

项集
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

扫描 D

C_2

项集	支持度计数
~~{1 2}~~	~~1~~
{1 3}	2
~~{1 5}~~	~~1~~
{2 3}	2
{2 5}	3
{3 5}	2

L_2

项集	支持度计数
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

项集
{2 3 5}

扫描 D

L_3

项集	支持度计数
{2 3 5}	2

注意{1,2,3}, {1,2,5}, {1,3,5} 不在C_3中

Apriori 算法例子, 由频繁项集构造强规则

数据库 D

TID	项集
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

要求规则最小支持度=50%(即支持度计数≥2), 置信度≥70%

频繁项集	支持度计数
{1}	2
{2}	3
{3}	3
{5}	3
{1,3}	2
{2,3}	2
{2,5}	3
{3,5}	2
{2,3,5}	2

{1,3}产生规则: **1→3(sup=2/4=50%, conf=2/2=1**)
3→1(sup=2/4=50%, conf=2/3≈66.7%)

{2,3}产生规则: 2→3(sup=2/4=50%, conf=2/3≈66.7%)
3→2(sup=2/4=50%, conf=2/3≈66.7%)

{2,5}产生规则: **2→5(sup=3/4=75%, conf=3/3=100%)**
5→2(sup=3/4=75%, conf=3/3=100%)

{3,5}产生规则: 3→5(sup=2/4=50%, conf=2/3≈66.7%)
5→3(sup=2/4=50%, conf=2/3≈66.7%)

{2,3,5}产生规则: 2→3∪5(sup=2/4=50%, conf=2/3≈66.7%)
3→5∪2(sup=2/4=50%, conf=2/3≈66.7%)
5→2∪3(sup=2/4=50%, conf=2/3≈66.7%)
2∪3→5(sup=2/4=50%, conf=2/2=100%)
2∪5→3(sup=2/4=50%, conf=2/3≈66.7%)
3∪5→2(sup=2/4=50%, conf=2/3=100%)

强关联规则:
1→3(50%, 100%)
2→5(75%, 100%)
5→2(75%, 100%)
2∪3→5(50%, 100%)
3∪5→2(50%, 100%)

数据库 D

TID	项集
10	A B C
20	A C
30	A D
40	B E F

(1) 根据给定的数据库D计算所有的频繁项集

(2) 根据频繁项集给出满足最小支持度和最小置信度的强关联规则

频繁项集最小支持度=50%(即支持度计数≥2)

关联规则支持度≥ 50%(即支持度计数≥2),
置信度≥70%

数据库 D

TID	项集
10	A B C
20	A C
30	A D
40	B E F

扫描 D

C_1

项集	支持度计数
{A}	3
{B}	2
{C}	2
~~{D}~~	~~1~~
~~{E}~~	~~1~~
~~{F}~~	~~1~~

L_1

项集	支持度计数
{A}	3
{B}	2
{C}	2

C_2

项集
{A B}
{A C}
{B C}

扫描 D

C_2

项集	支持度计数
~~{A B}~~	~~1~~
{A C}	2
~~{B C}~~	~~1~~

L_2

项集	支持度计数
{A C}	2

频繁项集	支持度
{A}	75%
{B}	50%
{C}	50%
{A, C}	50%

要求最小支持度=50%, 最小置信度=70%

对于规则 A→C:

支持度: Support(A∪C)=50%

置信度: Support(A∪C)/Support(A)=66.6%

对于规则 C→A:

支持度: Support(C∪A)=50%

置信度: Support(C∪A)/Support(C)=100%

- 设 *I* ={1, 2, 3, 4, 5, 6, 7, 8, 9}
- 序列<{3} {4,5} {8}>包含在<{6} {3,7} {9} {4,5,8} {3,8}> (或者说前者是后者的子序列)
  - {3} ⊆ {3,7}, {4,5} ⊆ {4,5,8}, {8} ⊆ {3,8}
  - 但是<{3},{8}>并不包含在<{3,8}>中, 反之也成立
  - 序列<{3} {4,5} {8}>的大小是3, 序列长度是4
- 如果一个序列只有一个项集，则括号可以省略
- 看 t 是否为 s 的子序列

s	t	Y/N
<{2, 4} {3, 6, 5} {8}>	<{2} {3, 6} {8}>	Yes
<{2, 4} {3, 6, 5} {8}>	<{2} {8}>	Yes
<{1, 2} {3, 4}>	<{1} {2}>	No
<{2, 4} {2, 4} {2, 5}>	<{2} {4}>	Yes

- 假定没有时限约束，列举包含在下面的数据序列中的所有4-子序列

<{1,3}, {2}, {2,3}, {4}>

- 列举如下:
- < {1, 3} {2} {2}>, <{1, 3} {2} {3} >, < {1, 3} {2} {4}>,
- <{1, 3}{2, 3} >, < {1, 3} {3} {4}>, <{1} {2} {2, 3} >,
- < {1} {2} {2} {4}>, <{1} {2} {3} {4} >, < {1} {2, 3} {4}>,
- <{3} {2} {2, 3} >, < {3} {2} {2} {4}>, <{3} {2} {3} {4} >,
- < {3} {2, 3} {4}>,<{2} {2, 3} {4} >

- 假定没有时限约束，列举包含在下面的数据序列中的所有3个元素 (项集) 的子序列

<{1,3}, {2}, {2,3}, {4}>

- 列举如下:
- < {1, 3} {2} {2, 3}>， <{1, 3} {2} {4} >
- < {1, 3} {3} {4}>， <{1, 3} {2} {2} >
- < {1, 3} {2} {3}>， <{1, 3} {2, 3} {4} >
- < {1} {2} {2, 3}>， <{1} {2} {4} >
- < {1} {3} {4}>， <{1} {2} {2} >
- < {1} {2} {3}>， <{1} {2, 3} {4} >
- < {3} {2} {2, 3}>， <{3} {2} {4} >
- < {3} {3} {4}>， <{3} {2} {2} >
- < {3} {2} {3}>， <{3} {2, 3} {4} >

区分不同的客户

CID	时间戳	项
A	1	1, 2, 4
A	2	2, 3
A	3	5
B	1	1, 2
B	2	2, 3, 4
C	1	1, 2
C	2	2, 3, 4
C	3	2, 4, 5
D	1	2
D	2	3, 4
D	3	4, 5
E	1	1, 3
E	2	2, 4, 5

数据集 S 含5个数据序列，A、B、C、D、E各一个

支持度	
<{1, 2}>	60%
<{2, 3}>	60%
<{2, 4}>	80%
<{3} {5}>	80%
<{1} {2}>	80%
<{2} {2}>	60%
<{1} {2, 3}>	60%
<{2} {2, 3}>	60%
<{1, 2} {2, 3}>	60%

$$\mathrm{supp}(<\{1,2\}>) = \frac{\# <\{1,2\}>}{\#S} = \frac{3}{5} = 60\%$$

$$\mathrm{supp}(<\{1\},\{2\}>) = \frac{\# <\{1\},\{2\}>}{\#S} = \frac{4}{5} = 80\%$$

□考虑以下频繁3-序列：< {1, 2, 3} >, < {1, 2}{3} >, < {1}{2, 3} >, < {1, 2}{4} >, < {1, 3}{4} >, < {1, 2, 4} >, < {2, 3}{3} >, < {2, 3}{4} >, < {2}{3}{3} >, 和 < {2}{3}{4} >

- (1) 列举出候选生成步骤产生的所有候选4-序列

  所有候选4-序列列举如下：

  < {1, 2, 3} {3} >, < {1, 2, 3} {4} >, < {1, 2} {3} {3} >, < {1, 2} {3} {4} >, < {1} {2, 3} {3} >, < {1} {2, 3} {4} >

- (2) 列出候选剪枝步骤剪掉的所有候选4-序列(假定没有时限约束)。

  如果没有时间限制，则所有候选子序列都必须频繁。 因此，经过修剪的候选子序列为：

  < {1, 2, 3} {3} >, < {1, 2} {3} {3} >, < {1, 2} {3} {4} >, < {1} {2, 3} {3} >, < {1} {2, 3} {4} >

  剪枝后的候选序列为： < {1, 2, 3} {4} >