聚类分析

距离定义个体间距离

例子： 计算向量(0,0)、(1,0)、(0,2)两两间的~~标准化欧氏~~前5个距离(假设两个分量的标准差分别为0.5和1)。

(1) 欧氏距离

$$d=\sqrt{\sum(x_i-y_i)^2}$$

$\sqrt{(1-0)^2+(0-0)^2}=1$，$\sqrt{(0-0)^2+(2-0)^2}=2$

$\sqrt{(1-0)^2+(0-2)^2}=\sqrt{5}$

(2) 曼哈顿距离

$$d=\sum|x_i-y_i|$$

$|0-1|+|0-0|=1$；$|0-0|+|0-2|=2$；$|1-0|+|0-2|=3$

(3) 切比雪夫距离

$$d=\max|x_i-y_i|$$

1，2，2

(4) 明氏距离

$$d=\sqrt[p]{|x_i-y_i|^p}\begin{cases}p=1\rightarrow 曼哈顿\\ p=2\rightarrow 欧式\\ p=+\infty\rightarrow 切比雪夫\end{cases}$$

(5) 标准化欧氏距离

$$d=\sqrt{\sum\left(\frac{x_{1k}-x_{2k}}{s_k}\right)^2}$$

$\sqrt{\left(\frac{0-1}{0.5}\right)^2+\left(\frac{0-0}{1}\right)^2}=2$

$\sqrt{\left(\frac{0-0}{0.5}\right)^2+\left(\frac{0-2}{1}\right)^2}=2$

$\sqrt{\left(\frac{1-0}{0.5}\right)^2+\left(\frac{0-2}{1}\right)^2}=2\sqrt{2}$

(6) 马氏距离

$$d(x_i,x_j)=\sqrt{(x_i-x_j)^T S^{-1}(x_i-x_j)}$$

已知二维正态总体G的分布为：$G\sim N(\mu,\Sigma)$，其中

$$\boldsymbol{\mu}=\begin{pmatrix}0\\0\end{pmatrix},\boldsymbol{\Sigma}=\begin{pmatrix}1&-0.9\\-0.9&1\end{pmatrix}$$

分别求点$A=(1,1)^T$和点$B=(1,-1)^T$到均值μ的欧氏距离和马氏距离。

解： 点A到μ的欧氏距离$=\sqrt{1^2+1^2}=\sqrt{2}$，点B到μ的欧氏距离$=\sqrt{1^2+1^2}=\sqrt{2}$。

$$\boldsymbol{\Sigma}^{-1}=\frac{1}{0.19}\begin{pmatrix}1&0.9\\0.9&1\end{pmatrix}$$

点A到μ的马氏距离$=\sqrt{\frac{1}{0.19}\begin{pmatrix}1&1\end{pmatrix}\begin{pmatrix}1&0.9\\0.9&1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}}=\cancel{\sqrt{10}}$ $\sqrt{20}$

点B到μ的马氏距离$=\sqrt{\frac{1}{0.19}\begin{pmatrix}1&-1\end{pmatrix}\begin{pmatrix}1&0.9\\0.9&1\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}}=\sqrt{1.05}$

二值变量个体间距离

(1) 简单匹配系数

	个体Y_0	个体Y_1
个体X_0	a个	b个
个体X_1	c个	d个

$$S(x,y)=\frac{b+c}{a+b+c+d}$$

(2) Jaccard系数

$$S(x,y)=\frac{b+c}{b+c+d}$$

基于相似系数的相似性度量

(1) 余弦相似度 $\cos\theta=\frac{X^TY}{|X||Y|}$

如 X=(3,2,0,5,0,0,0,2,0,0)^T 和 Y=(1,0,0,0,0,0,0,1,0,2)^T，它们的余弦相似度为 $\cos\theta=\frac{3+2}{\sqrt{3^2+2^2+5^2+2^2}*\sqrt{1^2+1^2+2^2}}=0.31$。

(2) 汉明距离：两个等长字符串对应位置不同字符的数量

(3) Jaccard相似系数　　Jaccard距离

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}\qquad J_\delta(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|}$$

如集合X={1,2,3,4};Y={3,4,5,6};

那么J(X,Y)= |{3,4}| / |{1,2,3,4,5,6}| =1/3;

例子：有两个物品A, B，调查7位用户是否购买了这两样物品，得以下向量：A=(0,0,1,1,1,0,1)，B=(1,0,1,0,1,0,0)

$|A\cap B|=2$

向量A：(0 0 1 1 1 0 1)

向量B：(1 0 1 0 1 0 0)

$|A\cup B|=5$

向量A：(0 0 1 1 1 0 1)

向量B：(1 0 1 0 1 0 0)

注意，因为忽略0-0匹配。所以$|A\cup B|\neq 7$。

因此，AB的杰卡德距离为$1-\frac{2}{5}=0.6$。

(4) Pearson相关系数

$$\rho_{XY}=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}=\frac{E(XY)-E(X)E(Y)}{\sigma_X\sigma_Y}$$

$$r=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2}\sqrt{\sum(Y_i-\bar{Y})^2}}$$

例1：计算压力x和压缩量y之间的相关系数r。

表 2-3　绝缘材料的压缩量和压力表

压力 $x(10\ lb/in^2)$	压缩量 $y(0.1\ in)$
1	1
2	1
3	2
4	2
5	4

$$\bar{X}=\frac{1+2+3+4+5}{5}=3,\quad \bar{Y}=\frac{1+1+2+2+4}{5}=2$$

$$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\sum_{i=1}^{5}(X_i-3)(Y_i-2)=7;$$

$$\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\sqrt{10},\quad \sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}=\sqrt{6};$$

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}=\frac{7}{\sqrt{10}\sqrt{6}}=0.904$$

类间距离聚类

例子：二维空间中有6个点，分别是x_1、x_2、x_3、x_4、x_5、x_6，数据如表所示。用最近(短)距离法对这6个点进行层次聚类。

最近/最远/中间

数据集

	x	y
x_1	1	1
x_2	2	1
x_3	1	3
x_4	4	1
x_5	4	4
x_6	5	4

Euclidean Distance

(1) 最短距离

数据集

	x	y
x_1	1	1
x_2	2	1
x_3	1	3
x_4	4	1
x_5	4	4
x_6	5	4

①计算初始化距离

$D^2(0)$	C_1	C_2	C_3	C_4	C_5	C_6
C_1	0					
C_2	1	0				
C_3	4	5	0			
C_4	9	4	13	0		
C_5	18	13	10	9	0	
C_6	25	18	17	10	1	0

C_1, C_2合并成C_7

C_5, C_6合并成C_8

②合并、更新

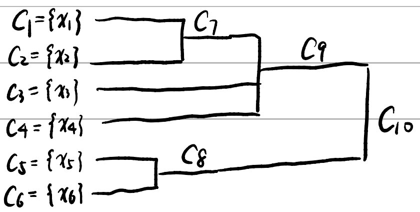
$D^2(1)$	$C_7=\{x_1,x_2\}$	C_3	C_4	$C_8=\{x_5,x_6\}$
$C_7=\{x_1,x_2\}$	0			
C_3	4	0		
C_4	4	13	0	
$C_8=\{x_5,x_6\}$	13	10	9	0

C_7和C_3, C_4合并成C_9

$D^2(2)$	$C_9=\{x_1,x_2,x_3,x_4\}$	$C_8=\{x_5,x_6\}$
$C_9=\{x_1,x_2,x_3,x_4\}$	0	
$C_8=\{x_5,x_6\}$	9	0

③最后二合一

把C_8, C_9合成一类C_{10}



(2)最大距离

①计算初始化距离

$D^2(0)$	C_1	C_2	C_3	C_4	C_5	C_6
C_1	0					
C_2	1	0				
C_3	4	5	0			
C_4	9	4	13	0		
C_5	18	13	10	9	0	
C_6	25	18	17	10	1	0

C_1, C_2合并成C_7

C_5, C_6合并成C_8

②合并、更新

$D^2(1)$	$C_7=\{x_1,x_2\}$	C_3	C_4	$C_8=\{x_5,x_6\}$
$C_7=\{C_1,C_2\}$	0			
C_3	5	0		
C_4	9	13	0	
$C_8=\{x_5,x_6\}$	25	17	10	0

C_1, C_3合并成C_9

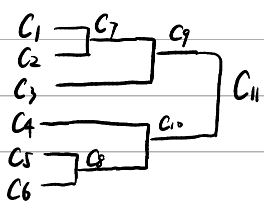
$D^2(2)$	$C_9=\{x_1,x_2,x_3\}$	C_4	C_8
$C_9=\{x_1,x_2,x_3\}$	0		
C_4	13	0	
$C_8=\{x_5,x_6\}$	25	10	0

C_4, C_8合并成C_{10}

$D^2(3)$	$C_9=\{x_1,x_2,x_3\}$	$C_{10}=\{x_4,x_5,x_6\}$
$C_9=\{x_1,x_2,x_3\}$	0	
$C_{10}=\{x_4,x_5,x_6\}$	25	0

③最后,二合一

C_9, C_{10}合并成C_{11}



(3)中间距离

①计算初始化距离

$D^2(0)$	C_1	C_2	C_3	C_4	C_5	C_6
C_1	0					
C_2	1	0				
C_3	4	5	0			
C_4	9	4	13	0		
C_5	18	13	10	9	0	
C_6	25	18	17	10	1	0

C_1,C_2合并成C_7

C_5,C_6合并成C_8

② 合并,更新

$D^2(1)$	$C_7=\{x_1,x_2\}$	C_3	C_4	$C_8=\{x_5,x_6\}$
$C_7=\{x_1,x_2\}$	0			
C_3	4.25	0		
C_4	6.25	13	0	
$C_8=\{x_5,x_6\}$	18	13.25	9.25	0

C_7和C_3合并成C_9

$$D_{37}^2=\frac{1}{2}\left(D_{13}^2+D_{23}^2\right)-\frac{1}{4}D_{12}^2=\frac{1}{2}(4+5)-\frac{1}{4}\times1=4.25$$

$$D_{47}^2=\frac{1}{2}\left(D_{14}^2+D_{24}^2\right)-\frac{1}{4}D_{12}^2=\frac{1}{2}(9+4)-\frac{1}{4}\times1=6.25$$

$$D_{38}^2=\frac{1}{2}\left(D_{35}^2+D_{36}^2\right)-\frac{1}{4}D_{56}^2=\frac{1}{2}(10+17)-\frac{1}{4}\times1=13.25$$

$$D_{48}^2=\frac{1}{2}\left(D_{45}^2+D_{46}^2\right)-\frac{1}{4}D_{56}^2=\frac{1}{2}(9+10)-\frac{1}{4}\times1=9.25$$

$$D_{78}^2=\frac{1}{4}\left(D_{15}^2+D_{25}^2+D_{16}^2+D_{26}^2-D_{12}^2-D_{56}^2\right)=\frac{1}{4}(18+25+13+18-2)=18$$

$D^2(2)$	$C_9=\{x_1,x_2,x_3\}$	$C_4=\{x_4\}$	$C_8=\{x_5,x_6\}$
$C_9=\{x_1,x_2,x_3\}$	0		
$C_4=\{x_4\}$	8.5625	0	
$C_8=\{x_5,x_6\}$	14.5625	9.25	0

C_4 C_9合并成C_{10}

$$D_{49}^2=\frac{1}{2}\left(D_{47}^2+D_{34}^2\right)-\frac{1}{4}D_{37}^2=\frac{1}{2}\times(6.25+13)-\frac{1}{4}\times4.25=8.5625$$

$$D_{89}^2=\frac{1}{2}\left(D_{78}^2+D_{38}^2\right)-\frac{1}{4}D_{37}^2=\frac{1}{2}\times(18+13.25)-\frac{1}{4}\times4.25=14.5625$$

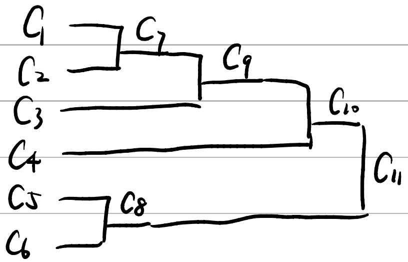
$D^2(3)$ $\quad C_{10}=\{x_1, x_2, x_3, x_4\}$ $\quad C_8=\{x_5, x_6\}$

	$C_{10}=\{\_\}$	$C_8=\{\_\}$
$C_{10}=\{\_\}$	0	
$C_8=\{\_\}$	9.765625	0

$$D^2_{810}=\frac{1}{2}(D^2_{48}+D^2_{89})-\frac{1}{4}D^2_{49}=\frac{1}{2}\times(9.25+14.5625)-\frac{1}{4}\times 8.5625$$
$$=9.765625$$

③最后二合一

把C_8, C_{10}合成一个类C_{11}



K-Means

假设给定如下要进行聚类的元组：{2,4,10,12,3,20,30,11,25},并假设k=2。

- 初始时用前两个数值作为类的均值；m_1=2和m_2=4。

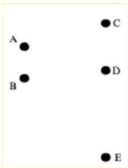
簇1均值	簇2均值	新簇1	新簇1均值	新簇2	新簇2均值
2	4	{2,3}	2.5	{4,10,11,12,20,25,30}	16
2.5	16	{2,3,4}	3	{10,11,12,20,25,30}	18
3	18	{2,3,4,10}	4.75	{11,12,20,25,30}	19.6
4.75	19.6	{2,3,4,10,11,12}	7	{20,25,30}	25
4.75	19.6	{2,3,4,10,11,12}	7	{20,25,30}	25

算法已收敛

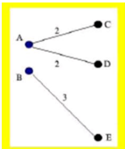
K-Medoids 算法

假设空间中的五个点{A、B、C、D、E}，如下图所示，各点之间的距离关系如下表所示。根据所给的数据对其运行k--medoids算法实现划分聚类(设k=2)。

样本点	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0

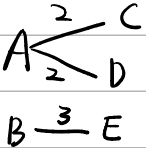


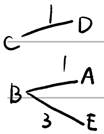
第一步 建立阶段：假如从5个对象中随机抽取的2个中心点为{A，B}，则样本被划分为{A、C、D}和{B、E}，如图所示。



第二步 交换阶段：假定中心点A、B分别被非中心点{C、D、E}替换，根据PAM算法需要计算下列代价TC_{AC}、TC_{AD}、TC_{AE}、TC_{BC}、TC_{BD}、TC_{BE}。

预设、

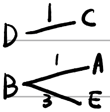




$TC_{AC} = C_{AAC} + C_{BAC} + C_{CAC} + C_{DAC} + C_{EAC}$

$= (1-0) + 0 + (0-2) + (1-2) + (3-3)$

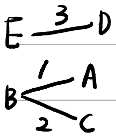
$= -2$



$TC_{AD} = C_{AAD} + C_{BAD} + C_{CAD} + C_{DAD} + C_{EAD}$

$= (1-0) + 0 + 0 + (0-2) + (2-3)$

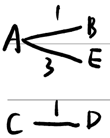
$= -2$



$TC_{AE} = C_{AAE} + C_{BAE} + C_{CAE} + C_{DAE} + C_{EAE}$

$= (1-0) + 0 + 0 + (3-2) + (0-3)$

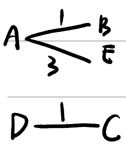
$= -1$



$TC_{BC} = C_{ABC} + C_{BBC} + C_{CBC} + C_{DBC} + C_{EBC}$

$= 0 + (1-0) + (0-2) + (1-2) + (3-3)$

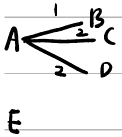
$= -2$



$TC_{BD} = C_{ABD} + C_{BBD} + C_{CBD} + C_{DBD} + C_{EBD}$

$= 0 + (1-0) + (1-2) + (0-2) + (3-0)$

$= -2$



$TC_{BE} = C_{ABE} + C_{BBE} + C_{CBE} + C_{DBE} + C_{EBE}$

$= 0 + (1-0) + (2-2) + (2-2) + (0-3)$

$= -2$

DIANA算法

例子：有如下表所示的数据集，使用DIANA算法对该数据集进行分裂层次聚类。

序号	属性 1	属性 2
1	1	1
2	1	2
3	2	1
4	2	2
5	3	4
6	3	5
7	4	4
8	4	5

$$
\begin{array}{c|cccccccc}
1 & 0 \\
2 & 1 & 0 \\
3 & 1 & 1.4 & 0 \\
4 & 1.4 & 1 & 1 & 0 \\
5 & 3.6 & 2.8 & 3.2 & 2.2 & 0 \\
6 & 4.5 & 3.6 & 4.1 & 3.2 & 1 & 0 \\
7 & 4.2 & 3.6 & 3.6 & 2.8 & 1 & 1.4 & 0 \\
8 & 5 & 4.2 & 4.5 & 3.6 & 1.4 & 1 & 1 & 0 \\
\hline
 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8
\end{array}
$$

对于所给的数据进行DIANA算法，(设n=8,用户输入的终止条件为2个类)，初始类{1,2,3,4,5,6,7,8}。

①计算平均距离（①找直径最大的类）

序号1的平均距离（就是1距离其它各个点的距离长度之和除以7）
s1=(1+1+1.1414+3.6+4.47+4.24+5)/7=2.96;
序列2的平均距离 s2=(1+1.414+1+2.828+3.6+3.6+4.24)/7=2.526;
序列3的平均距离 s3=(1+1.414+1+3.16+4.12+3.6+4.27)/7=2.68;
序列4的平均距离 s4=(1.414+1+1+2.24+3.16+2.828+3.6)/7=2.18
序列5的平均距离 s5=2.18;
序列6的平均距离 s6=2.68;
序列7的平均距离 s7=2.526;
序列8的平均距离 s8=2.96;

②挑平均距离最大的放到splinter group，剩下的放old party

splinter group={1}　old party={2,3,4,5,6,7,8}

③找离splinter group最近的，若该距离≤该点到old Party中其他点的最小距离，则合并

一直找直到找不到

splinter group={1}

$$
\begin{array}{c|cccccccc}
1 & 0 \\
2 & 1 & 0 \\
 & 1 & 1.4 & 0 \\
 & 1.4 & 1 & 1 & 0 \\
 & 3.6 & 2.8 & 3.2 & 2.2 & 0 \\
 & 4.5 & 3.6 & 4.1 & 3.2 & 1 & 0 \\
 & 4.2 & 3.6 & 3.6 & 2.8 & 1 & 1.4 & 0 \\
 & 5 & 4.2 & 4.5 & 3.6 & 1.4 & 1 & 1 & 0
\end{array}
$$

2 → {1,2}

$$
\begin{array}{c|cccccccc}
1 & 0 \\
2 & 1 & 0 \\
3 & 1 & 1.4 & 0 \\
 & 1.4 & 1 & 1 & 0 \\
 & 3.6 & 2.8 & 3.2 & 2.2 & 0 \\
 & 4.5 & 3.6 & 4.1 & 3.2 & 1 & 0 \\
 & 4.2 & 3.6 & 3.6 & 2.8 & 1 & 1.4 & 0 \\
 & 5 & 4.2 & 4.5 & 3.6 & 1.4 & 1 & 1 & 0
\end{array}
$$

3 → {1,2,3}

$$
\begin{array}{c|cccccccc}
1 & 0 \\
2 & 1 & 0 \\
3 & 1 & 1.4 & 0 \\
4 & 1.4 & 1 & 1 & 0 \\
 & 3.6 & 2.8 & 3.2 & 2.2 & 0 \\
 & 4.5 & 3.6 & 4.1 & 3.2 & 1 & 0 \\
 & 4.2 & 3.6 & 3.6 & 2.8 & 1 & 1.4 & 0 \\
 & 5 & 4.2 & 4.5 & 3.6 & 1.4 & 1 & 1 & 0
\end{array}
$$

4 →

没了

④spliter group 为一类，old party 为一类
　　　　‖　　　　　　　　　　　　‖
{1,2,3,4}　　　　　　　　{5,6,7,8}

分类分析

协方差矩阵相等的距离判别法

例1： 记二维正态总体$N_2(\mu^{(i)},\Sigma)$为$G_i\,(i=1,2)$（两总体协差阵相同），已知来自$G_i\,(i=1,2)$的样本数据为

$$X^{(1)}=\begin{pmatrix}2&12\\4&10\\3&8\\3&10\end{pmatrix},\ X^{(2)}=\begin{pmatrix}5&7\\3&9\\4&5\end{pmatrix}\qquad \begin{pmatrix}k=2, & m=2\\ n_1=4, & n_2=3\end{pmatrix}$$

（1）试求两总体的样本离差阵S_1, S_2和合并样本协差阵S。

（2）今有样本$x_0=(2,8)'$，试问按马氏距离准则样本x_0应判归哪一类。

样本离差阵

$$S_1=\sum_{j=1}^{n_1}(x_j^{(1)}-\bar{x}^{(1)})'(x_j^{(1)}-\bar{x}^{(1)})$$

$$S_2=\sum_{j=1}^{n_2}(x_j^{(2)}-\bar{x}^{(2)})'(x_j^{(2)}-\bar{x}^{(2)})$$

样本合并组内离差阵 S_1+S_2

合并样本协差阵：$\frac{1}{n_1+n_2-2}(S_1+S_2)$

判别函数 $W(x)=(x-\bar{x})'S^{-1}(\bar{x}^{(1)}-\bar{x}^{(2)})$

(1) ①求每个类的样本的均值向量

$\bar{X}^{(1)}$: $\frac{2+4+3+3}{4}=3$ $\quad\frac{12+10+8+10}{4}=10$ $\quad \bar{X}^{(1)}=\begin{pmatrix}3\\10\end{pmatrix}$

$\bar{X}^{(2)}$: $\frac{5+3+4}{3}=4$ $\quad\frac{7+9+5}{3}=7$ $\quad \bar{X}^{(2)}=\begin{pmatrix}4\\7\end{pmatrix}$

②样本和均值作差

$$\tilde{X}^{(1)}=\begin{pmatrix}2-3&12-10\\4-3&10-10\\3-3&8-10\\3-3&10-10\end{pmatrix}=\begin{pmatrix}-1&2\\1&0\\0&-2\\0&0\end{pmatrix}$$

$$\tilde{X}^{(2)}=\begin{pmatrix}5-4&7-7\\3-4&9-7\\4-4&5-7\end{pmatrix}=\begin{pmatrix}1&0\\-1&2\\0&-2\end{pmatrix}$$

③自己乘自己求样本离差阵

$$S_1=(\tilde{X}^{(1)})'(\tilde{X}^{(1)})=\begin{pmatrix}-1&1&0&0\\2&0&-2&0\end{pmatrix}\begin{pmatrix}-1&2\\1&0\\0&-2\\0&0\end{pmatrix}=\begin{pmatrix}2&-2\\-2&8\end{pmatrix}$$

$$S_2=(\tilde{X}^{(2)})'(\tilde{X}^{(2)})=\begin{pmatrix}1&-1&0\\0&2&-2\end{pmatrix}\begin{pmatrix}1&0\\-1&2\\0&-2\end{pmatrix}=\begin{pmatrix}2&-2\\-2&8\end{pmatrix}$$

④相加，乘系数，求合并样本协差阵

$$\underbrace{S}_{\text{合并样本协差阵}}=\frac{1}{n_1+n_2-2}\underbrace{(S_1+S_2)}_{\text{合并组内离差阵}}=\frac{1}{5}\begin{pmatrix}4&-4\\-4&16\end{pmatrix}=\frac{4}{5}\begin{pmatrix}1&-1\\-1&4\end{pmatrix}$$

(2) f_1: 飞利用马氏距离

先求S^{-1} $\quad |S|=\frac{16}{25}\begin{vmatrix}1&-1\\-1&4\end{vmatrix}=\frac{16}{25}\times3=\frac{48}{25}$

$$S^*=\frac{4}{5}\begin{pmatrix}4&1\\1&1\end{pmatrix}$$

$$S^{-1}=\frac{1}{|S|}S^*=\frac{25}{48}\times\frac{4}{5}\begin{pmatrix}4&1\\1&1\end{pmatrix}=\frac{5}{12}\begin{pmatrix}4&1\\1&1\end{pmatrix}$$

$$D_1^2(x_0)=(x_0-\overline{X}^{(1)})'S^{-1}(x_0-\overline{X}^{(1)})$$

$$=\begin{pmatrix}-1 & -2\end{pmatrix}\frac{5}{12}\begin{pmatrix}4 & 1\\ 1 & 1\end{pmatrix}\begin{pmatrix}-1\\ -2\end{pmatrix}=\frac{5}{12}\begin{pmatrix}-6 & -3\end{pmatrix}\begin{pmatrix}-1\\ -2\end{pmatrix}=5$$

$$D_2^2(x_0)=(x_0-\overline{X}^{(2)})'S^{-1}(x_0-\overline{X}^{(2)})$$

$$=\begin{pmatrix}-2 & 1\end{pmatrix}\frac{5}{12}\begin{pmatrix}4 & 1\\ 1 & 1\end{pmatrix}\begin{pmatrix}-2\\ 1\end{pmatrix}=\frac{5}{12}\begin{pmatrix}-7 & -1\end{pmatrix}\begin{pmatrix}-2\\ 1\end{pmatrix}=\frac{5\times 13}{12}=\frac{65}{12}$$

$D_1^2(x_0)<D_2^2(x_0)$ $\quad\therefore x_0\in G_1$

f_2 用判别函数

$$W(X)=(X-\overline{X})'S^{-1}(\overline{X}^{(1)}-\overline{X}^{(2)})$$

$$\overline{X}=\frac{1}{2}(\overline{X}^{(1)}+\overline{X}^{(2)})=\begin{pmatrix}3.5\\ 8.5\end{pmatrix}$$

$$W(x_0)=\begin{pmatrix}-1.5 & -0.5\end{pmatrix}\frac{5}{12}\begin{pmatrix}4 & 1\\ 1 & 1\end{pmatrix}\begin{pmatrix}-1\\ 3\end{pmatrix}$$

$$=\frac{5}{12}\begin{pmatrix}-6.5 & -2\end{pmatrix}\begin{pmatrix}-1\\ 3\end{pmatrix}$$

$$=\frac{5}{12}\times 0.5=\frac{5}{24}>0 \quad \therefore x_0\in G_1$$

协方差矩阵不相等的距离判别法

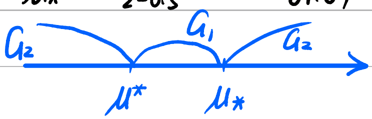
例子： 已知有两个类G_1和G_2，分别为设备A、B生产的产品。设备A生产的产品平均耐磨度$\mu_1=80$，精度$\sigma_1^2=0.25$；设备B的平均耐磨度$\mu_2=75$，精度$\sigma_2^2=4$。现有一耐磨度为78的产品x，试判断它为哪一台设备生产的。

f_1 ① 先算 $\mu^*=\dfrac{\mu_1\sigma_2+\mu_2\sigma_1}{\sigma_1+\sigma_2}$ $\quad \mu_*=\left|\dfrac{\mu_1\sigma_2-\mu_2\sigma_1}{\sigma_1-\sigma_2}\right|$

$$\mu^*=\frac{80\times 2+75\times 0.5}{0.5+2}=79$$

$$\mu_*=\frac{80\times 2-75\times 0.5}{2-0.5}=81.67$$

②



$x=78<\mu^*$. $\therefore x\in G_2$

f_2

首先计算马氏距离：

$$D^2(x,G_1)=\frac{(x-\mu_1)^2}{\sigma_1^2}=\frac{(78-80)^2}{0.25}=16=4^2$$

$$D^2(x,G_2)=\frac{(x-\mu_2)^2}{\sigma_2^2}=\frac{(78-75)^2}{4}=2.25=1.5^2$$

$D^2(x,G_2)<D^2(x,G_1)$

$x\in G_2$

ID3算法构造决策树

表 4-1 高尔夫活动决策表

编号	天气	温度	湿度	风速	活动
1	晴	炎热	高	弱	取消
2	晴	炎热	高	强	取消
3	阴	炎热	高	弱	进行
4	雨	适中	高	弱	进行
5	雨	寒冷	正常	弱	进行
6	雨	寒冷	正常	强	取消
7	阴	寒冷	正常	强	进行
8	晴	适中	高	弱	取消
9	晴	寒冷	正常	弱	进行
10	雨	适中	正常	弱	进行
11	晴	适中	正常	强	进行
12	阴	适中	高	强	进行
13	阴	炎热	正常	弱	进行
14	雨	适中	高	强	取消

① 信息熵: $Ent(D)=-\sum_{i=1}^{m} p_i \log_2(p_i)$

② 信息增益: 划分前-划分后

$$Gain(D,a)=Ent(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}Ent(D^v)$$

(D^v表示在特征a^v上取值的所有样本)

③ 增益率: $Gain\_ratio(D,a)=\frac{Gain(D,a)}{IV(a)}$

$$IV(a)=-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\log_2\frac{|D^v|}{|D|}$$

①先算样本的熵

$$I(s_1,s_2)=I(9,5)=-\frac{9}{14}\log_2\frac{9}{14}-\frac{5}{14}\log_2\frac{5}{14}=0.940$$

②算每个属性的信息增益，取最大的对树分支，一直重复直到全叶

天气：晴天 $I_{晴}=I(s_{11},s_{21})=I(2,3)=-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}=0.971$

阴天: $I_{阴}=I(s_{12},s_{22})=I(4,0)=-\frac{4}{4}\log_2 1=0$

雨天 $I_{雨}=I(s_{13},s_{23})=I(3,2)=-\frac{3}{5}\log_2\frac{3}{5}-\frac{2}{5}\log_2\frac{2}{5}=0.971$

$Ent(天气)=\frac{5}{14}\times0.971+\frac{5}{14}\times0.971=0.694$

$Gain(天气)=0.940-0.694=0.246$

温度: 炎热: $I_{炎}=I(s_{11},s_{21})=I(2,2)=-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}=1$

适中: $I_{中}=I(s_{12},s_{22})=I(4,2)=-\frac{4}{6}\log_2\frac{4}{6}-\frac{2}{6}\log_2\frac{2}{6}=0.918$

寒冷: $I_{冷}=I(s_{13},s_{23})=I(3,1)=-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}=0.811$

$Ent(温度)=\frac{4}{14}\times1+\frac{6}{14}\times0.918+\frac{4}{14}\times0.811=0.911$

$Gain(温度)=0.940-0.911=0.029$

湿度: 高: $I_{高}=I(s_{11},s_{21})=I(3,4)=-\frac{3}{7}\log_2\frac{3}{7}-\frac{4}{7}\log_2\frac{4}{7}=0.985$

正常: $I_{正}=I(s_{12},s_{22})=I(6,1)=-\frac{6}{7}\log_2\frac{6}{7}-\frac{1}{7}\log_2\frac{1}{7}=0.592$

$Ent(湿度)=\frac{1}{2}\times 0.985+\frac{1}{2}\times 0.592=0.7885$

$Gain(湿度)=0.940-0.7885=0.1515$

风速：弱：$I(s_1,s_2)=I(6,2)=-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}=0.811$

强：$I(s_1,s_2)=I(3,3)=-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}=1$

$Ent(风速)=\frac{8}{14}\times 0.811+\frac{6}{14}\times 1=0.892$

$Gain(风速)=0.940-0.892=0.048$

显然，天气的信息增益最大，按天气划分

再对晴中的{1,2,8,9,11}划分

$I(s_1,s_2)=I(2,3)=-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}=0.971$

温度：炎热 $I(s_{11},s_{21})=I(0,2)=-\frac{2}{2}\log_2 1=0$

适中 $I(s_{12},s_{22})=I(1,1)=-\frac{1}{2}\log_2-\frac{1}{2}\log_2=1$

寒冷 $I(s_{13},s_{23})=I(1,0)=\log_2 1=0$

$Gain(温度)=0.971-\frac{2}{5}=0.571$

湿度：高 $I(s_{11},s_{21})=I(0,3)=-\frac{3}{3}\log_2 1=0$

正常 $I(s_{21},s_{22})=I(2,0)=0$

$Gain(湿度)=0.971$ 最大，不用再算风速

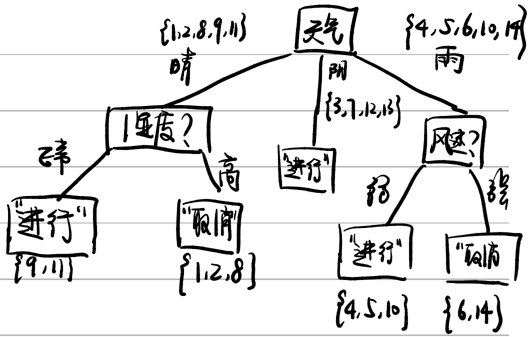
对雨中的{4,5,6,10,14}划分

$I(s_1,s_2)=I(3,2)=0.971$

风速：弱：$I(s_{11},s_{21})=I(3,0)=0$

强：$I(s_{12},s_{22})=I(0,2)=0$

$Gain(风速)=0.971$ 最大



KNN

问题：上面数据集中序号1-12为已知的电影分类，分为喜剧片、动作片、爱情片三个种类，使用的特征值分别为搞笑镜头、打斗镜头、拥抱镜头的数量。那么来了一部新电影《唐人街探案》，它属于上述3个电影分类中的哪个类型？

电影分类数据集（纯属虚构）：

序号	电影名称	搞笑镜头	拥抱镜头	打斗镜头	电影类型
1.	宝贝当家	45	2	9	喜剧片
2.	美人鱼	21	17	5	喜剧片
3.	澳门风云3	54	9	11	喜剧片
4.	功夫熊猫3	39	0	31	喜剧片
5.	谍影重重	5	2	57	动作片
6.	叶问3	3	2	65	动作片
7.	伦敦陷落	2	3	55	动作片
8.	我的特工爷爷	6	4	21	动作片
9.	奔爱	7	46	4	爱情片
10.	夜孔雀	9	39	8	爱情片
11.	代理情人	9	38	2	爱情片
12.	新步步惊心	8	34	17	爱情片
13.	唐人街探案	23	3	17	?

①算距离(升序排序)

②前K个里面哪个类多就选哪个

取K=5

```
[['我的特工爷爷', 17.49],   —— 动作
['美人鱼', 18.55],          —— 喜剧
['功夫熊猫3', 21.47],       —— 喜剧
['宝贝当家', 23.43],        —— 喜剧
['澳门风云3', 32.14],       —— 喜剧
['新步步惊心', 34.44],
['夜孔雀', 39.66],
['代理情人', 40.57],
['伦敦陷落', 43.42],
['谍影重重', 43.87],
['奔爱', 47.69],
['叶问3', 52.01]]
```

所以是喜剧

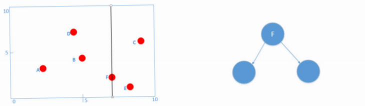
KD树(及其搜索算法)

例子： 给定一个二维空间的数据集：$T=\{(2,3), (5,4), (9,6), (4,.7), (8, 1), (7,2)\}$，构造一个平衡KD树。

根节点对应包含数据集 T 的矩形，首先选择 $x^{(1)}$ 维，按照 $x^{(1)}$维对数据集进行排序，得到：A(<u>2</u>, 3)、D(<u>4</u>, 7)、B(<u>5</u>, 4)、F(<u>7</u>, 2)、E(<u>8</u>, 1)、C(<u>9</u>, 6)。

对于 $x^{(1)}$维，其中位数为：7，选择F (7,2)作为根节点。

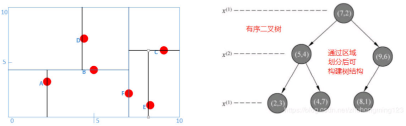
接着按照 $x^{(1)}$维进行划分，将小于7的划分到根节点的左子树中，大于7的划分到根节点的右子树中（该过程类似搜索二叉树）。对应的树结构如下：



此时对于左右子树，需要<u>选取相同的维进行划分</u>（此处选取 $x^{(2)}$维）。

✓ 左子树：对于节点 (2,3), (4,7), (5,4)，按照 $x^{(2)}$ 维进行排序 (2,<u>3</u>), (5,<u>4</u>), (4,<u>7</u>), 选择中位数 (5,4)作为子树根节点；同理，将小于4的放到 B(5,4)节点的左子树中，大于4的放到B (5,4)节点的右子树中。

✓ 对于根节点 F(7,2)的右子树，对节点 (8,<u>1</u>), (9,<u>6</u>)进行排序，并选择中位数6,由于 1<6,将E(8,1)作为C (9,6)的左孩子。到此，KD树构造完成。



Naive Bayes

例1：给出如表所示的训练样本，目的是判定一个人是否会购买电脑。这个人的属性为X=（年龄<=30，收入=中等，学生=是，信用率=一般）。

判定一个人是否会购买电脑的训练样本

编号	年龄	收入	学生	信用等级	类别：购买电脑
1	<=30	高	否	一般	不会购买
2	<=30	高	否	良好	不会购买
3	31…40	高	否	一般	会购买
4	>40	中等	否	一般	会购买
5	>40	低	是	一般	会购买
6	>40	低	是	良好	不会购买
7	31…40	低	是	良好	会购买
8	<=30	中等	否	一般	不会购买
9	<=30	低	是	一般	会购买
10	>40	中等	是	一般	会购买
11	<=30	中等	是	良好	会购买
12	31…40	中等	否	良好	会购买
13	31…40	高	是	一般	会购买
14	>40	中等	否	良好	不会购买

解：设类别C_1：购买电脑=“是”，类别C_2：购买电脑=“否”，所以可求：

$P(C_1)$=P(购买电脑=“是”)

=9/14=0.643

$P(C_2)$= P(购买电脑=“否”)

=5/14=0.357

计算每个类别的$P(X|C_i)$:

$P(X \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i)$

P(年龄=“<30”|购买电脑=“是”)=2/9=0.222

P (年龄=“<30”|购买电脑=“否”)=3/5=0.6

P (收入=“中等”|购买电脑=“是”)=4/9=0.444

P (收入=“中等”|购买电脑=“否”)=2/5=0.4

P (学生=“是”|购买电脑=“是”)

=6/9=0.667

P (学生=“是”|购买电脑=“否”)

=1/5=0.2

P (信用率=“一般”|购买电脑=“是”)

=6/9=0.667

P (信用率=“一般”|购买电脑=“否”)

=2/5=0.4

编号	年龄	收入	学生	信用等级	类别：购买电脑
1	<=30	高	否	一般	不会购买
2	<=30	高	否	良好	不会购买
3	31…40	高	否	一般	会购买
4	>40	中等	否	一般	会购买
5	>40	低	是	一般	会购买
6	>40	低	是	良好	不会购买
7	31…40	低	是	良好	会购买
8	<=30	中等	否	一般	不会购买
9	<=30	低	是	一般	会购买
10	>40	中等	是	一般	会购买
11	<=30	中等	是	良好	会购买
12	31…40	中等	否	良好	会购买
13	31…40	高	是	一般	会购买
14	>40	中等	否	良好	不会购买

从而得到：

$P(X \mid C_i) = \prod_{k=1}^{n} P(X_k \mid C_i)$

P(X|购买电脑=“是”)=0.222 × 0.444 × 0.667 × 0.667=0.044

P(X|购买电脑=“否”)=0.6 × 0.4 × 0.2 × 0.4=0.019

又由于：

$P(X \mid C_i)\, P(C_i)$

$P(C_1)$=P(购买电脑=“是”) =9/14=0.643

$P(C_2)$= P(购买电脑=“否”) =5/14=0.357

所以：

P(X|购买电脑=“是”) × P(购买电脑=“是”)=0.028

P(X|购买电脑=“否”) × P(购买电脑=“否”)=0.007

所以判定X处于类别C_1，此人会购买电脑。

关联分析

GBDT树

GBDT学习例子：训练集: (A, 14岁), (B, 16岁), (C, 24岁), (D, 26岁)。训练数据的均值为：20岁；决策树的个数为：2棵。每个样本的特征有两个：购买金额是否小于1K，经常去百度提问还是回答？

测试样本：预测一个购物金额为3K，经常去百度问淘宝相关问题的女生的年龄。

标签数据:

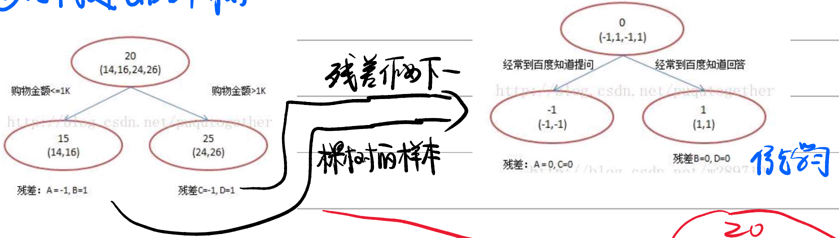
A: 14岁, 购物金额≤1K ; 经常到百度知道提问

B: 16岁, ≤1K ; 回答

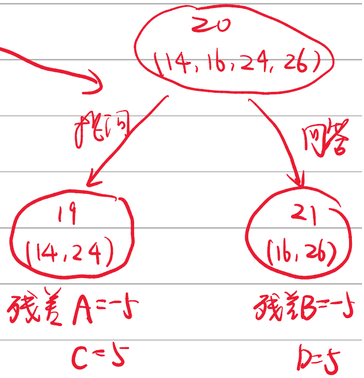
C: 24岁 >1K ; 提问

D: 26岁 >1K ; 回答

①先构建GBDT树

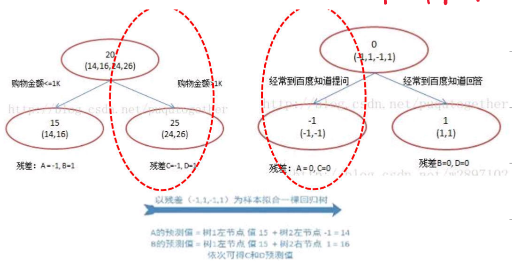


为什么不能用第二个属性



残差平方和大

②使用GBDT树进行测试



- 提取2个特征：购物金额3K，经常去百度上面提问问题。
- 第一棵树 →购物金额大于1K→右叶子，初步说明这个女生25岁。
  第二棵数-→经常去百度提问→左叶子，说明这个女生的残差为-1。
- 叠加前面每棵树得到的结果：25-1 = 24岁，最终预测结果为24岁。

关联规则分析

①支持度：相集的支持度是包含项集的事务数与总事务数的比值.
S　　规则 X⇒Y 的支持度是D中事务同时包含X,Y的百分比

②置信度：规则 X⇒Y 的置信度是D中已包含X的情况下,包含Y的百分比
C

③提升度：置信度/含有Y的比例
Lift

④兴趣因子：$I(X\Rightarrow Y)=\frac{S(A\Rightarrow B)}{S(A)\times S(B)}$
I　　= 二项集的提升度 ⇔ 兴趣因子

⑤频繁项集：支持度>最小支持度阈值

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example:

$$\{Milk, Diaper\} \Rightarrow Beer$$

$$s=\frac{\sigma(Milk, Diaper, Beer)}{|T|}=\frac{2}{5}=0.4$$

Example:

$$\{Milk, Diaper\} \Rightarrow Beer$$

$$c=\frac{\sigma(Milk, Diaper, Beer)}{\sigma(Milk, Diaper)}=\frac{2}{3}=0.67$$

关联规则分析

例子：一个Apriori的具体例子，该例基于下图的AllElectronics的事务数据库。数据库中有9个事务，即|D| = 9。Apriori假定事务中的项按字典次序存放。

TID	商品ID的列表
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

支持度阈值=2

置信度阈值=70%

求频繁项集

求最高频繁项集产生的关联规则

频繁1项集	候选2项集	2-项集	候选3项集	3-项集
{1} 6	{1 2} 4	{1 2} 4	{1 2 3} 2	{1 2 3} 2
{2} 7	{1 3} 4	{1 3} 4	{1 2 5} 2	{1 2 5} 2
{3} 6	~~{1 4} 1~~	{1 5} 2	~~{1 3 5} 1~~	
{4} 2	{1 5} 2	{2 3} 4	~~{2 3 4} 0~~	
{5} 2	{2 3} 4	{2 4} 2	~~{2 3 5} 1~~	
	{2 4} 2	{2 5} 2	~~{2 4 5} 0~~	
	{2 5} 2			
	~~{3 4} 0~~			
	~~{3 5} 1~~			
	~~{4 5} 0~~			

4-项集 {1 2 3 5} 不再是频繁项集，算法结束

求关联规则:

频繁3-项集 {1 2 3}　{1 2 5}

先求{1 2 3}

{1} {2} {3} {1 2} {1 3} {2 3}

{1}⇒{2 3} 2/6=33.3%

{2}⇒{1 3} 2/7

{3}⇒{1 2} 2/6

{1 2}⇒{3} 2/4=50%

{1 3}⇒{2} 2/4=50%

{2 3}⇒{1} 2/4=50%

没有满足置信度70%的

{1 2 5}

{1} {2} {5} {12} {15} {25}

{1}⇒{2 5} 2/6

{2}⇒{1 5} 2/7

{5}⇒{1 2} 2/2=100% ✓

{12}⇒{5} 2/4=50%

{15}⇒{2} 2/2=100% ✓

{25}⇒{1} 2/2=100% ✓

最终产生的规则 {5}⇒{1 2}, {1 5}⇒{2}, {2 5}⇒{1}

算置信度时分子就是该3-项集的支持度,分母是蕴涵式左侧集合的支持度