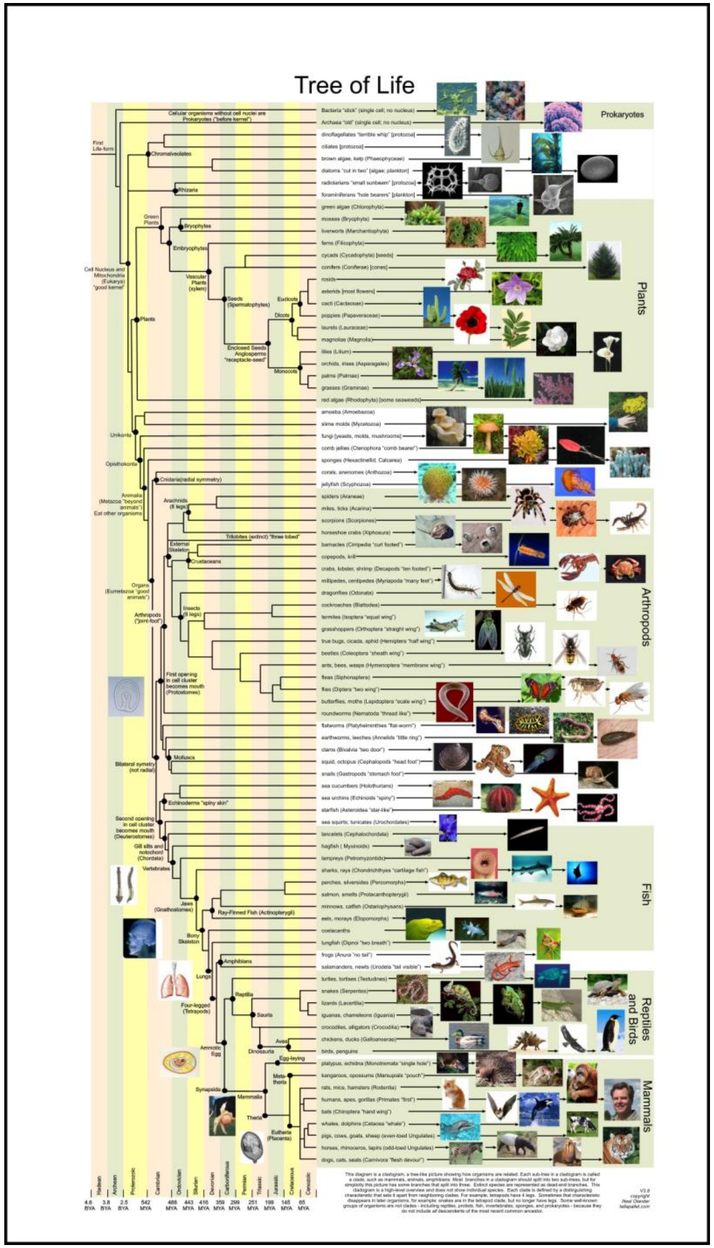
DATAGURU 炼数成金

数据分析与R语言 第9周



2012.7.20

支持向量机 SVM

- 问题的提出：最优分离平面（决策边界）

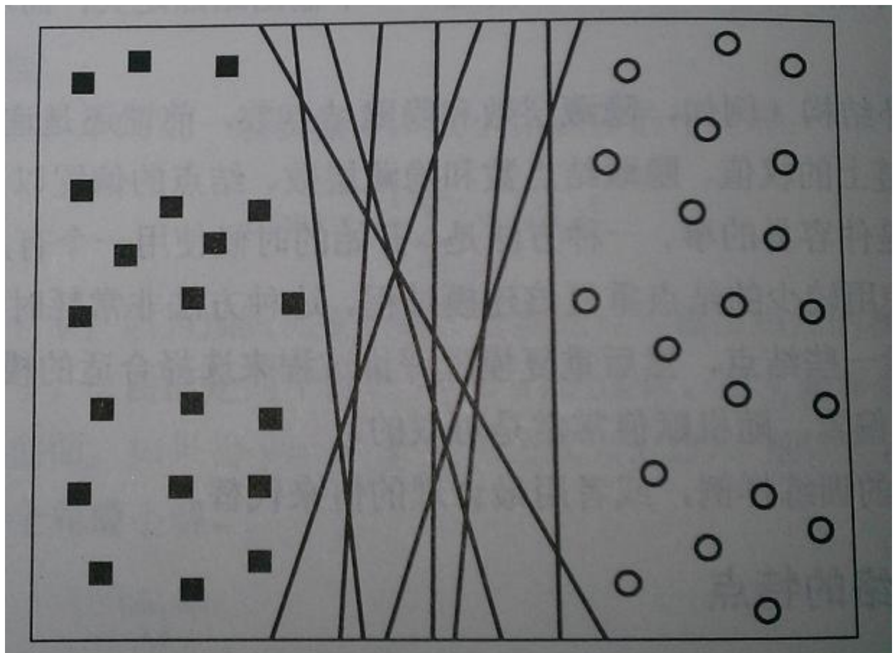


图 5-21　一个线性可分数据集上的可能决策边界

优化目标

- 决策边界边缘距离最远

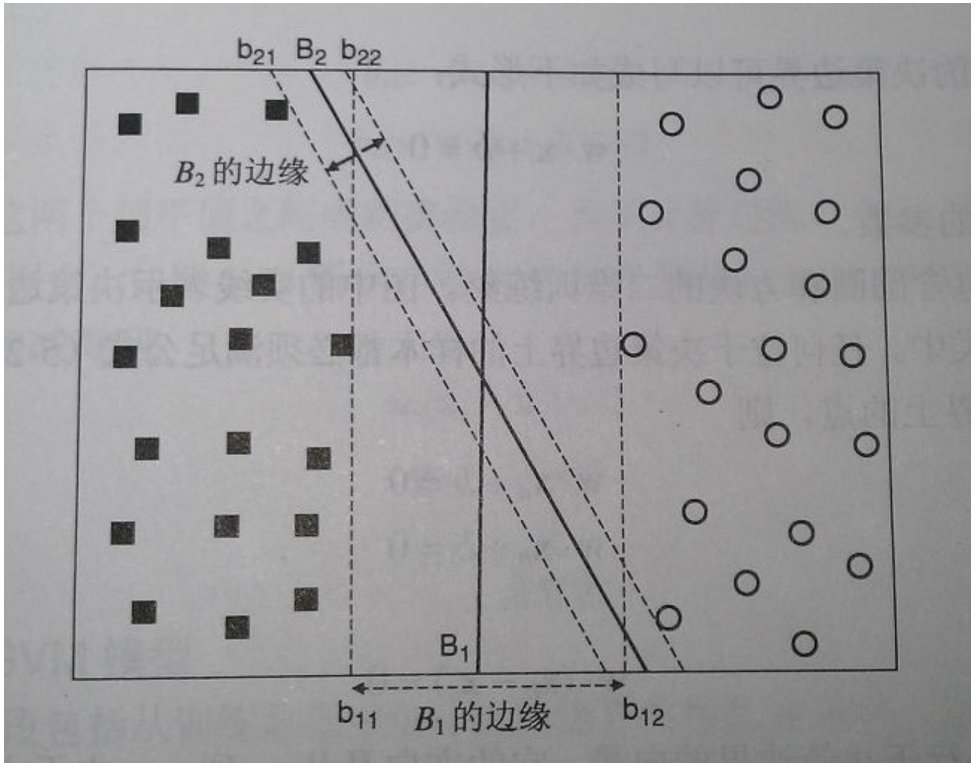
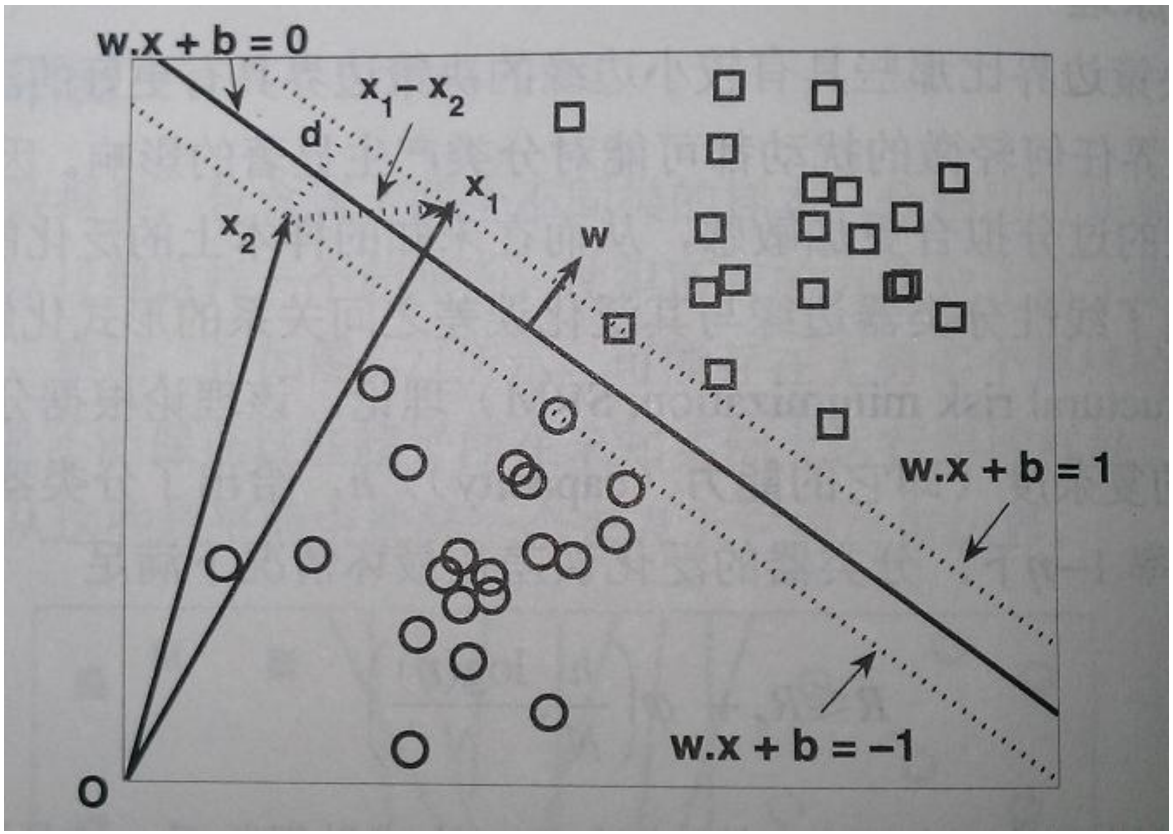


图 5-22 决策边界的边缘

数学模型



$$b_{i1}:\quad \mathbf{w}\cdot\mathbf{x}+b=1$$

$$b_{i2}:\quad \mathbf{w}\cdot\mathbf{x}+b=-1$$

$$\mathbf{w}.(\mathbf{x}_1-\mathbf{x}_2)=2$$

$$\|\mathbf{w}\|\times d=2$$

$$\therefore d=\frac{2}{\|\mathbf{w}\|}$$

问题转化为凸优化

$$\begin{array}{ll} \mathbf{w}\cdot\mathbf{x_i}+b\geqslant 1 & \text{如果 } y_i=1 \\ \mathbf{w}\cdot\mathbf{x_i}+b\leqslant -1 & \text{如果 } y_i=-1 \end{array}$$

$$d=\frac{2}{\|\mathbf{w}\|}$$

$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2}$$

受限于 $y_i(\mathbf{w}\cdot\mathbf{x_i}+b)\geqslant 1, \quad i=1,2,\cdots,N$

拉格朗日乘子法——未知数太多

$$L_P = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N}\lambda_i\left(y_i(\mathbf{w}\cdot\mathbf{x_i}+b)-1\right)$$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{N}\lambda_i y_i \mathbf{x}_i$$

$$\frac{\partial L_p}{\partial_b} = 0 \Rightarrow \sum_{i=1}^{N}\lambda_i y_i = 0$$

$$y_i(\mathbf{w}\cdot\mathbf{x_i}+b)\geq 1$$

KKT变换和对偶公式

$$\lambda_i \geqslant 0$$
$$\lambda_i [y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1] = 0$$

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j}$$

问题的解决和神经网络化

- 对偶公式是二次规划问题，有现成的数值方法可以求解
- 大部分的拉格朗日乘子为0，不为0的对应于“支持向量”（恰好在边界上的样本点）
- 只要支持向量不变，修改其他样本点的值，不影响结果，当支持变量发生改变时，结果一般就会变化
- 求解出拉格朗日乘子后，可以推出**w**和**b**，判别函数可以写成以下神经网络的样式

$$f(\mathbf{z}) = sign(\mathbf{w}\cdot\mathbf{z} + b) = sign(\sum_{i=1}^{N}\lambda_i y_i \mathbf{x}_i \cdot \mathbf{z} + b)$$

支持向量机神经网络

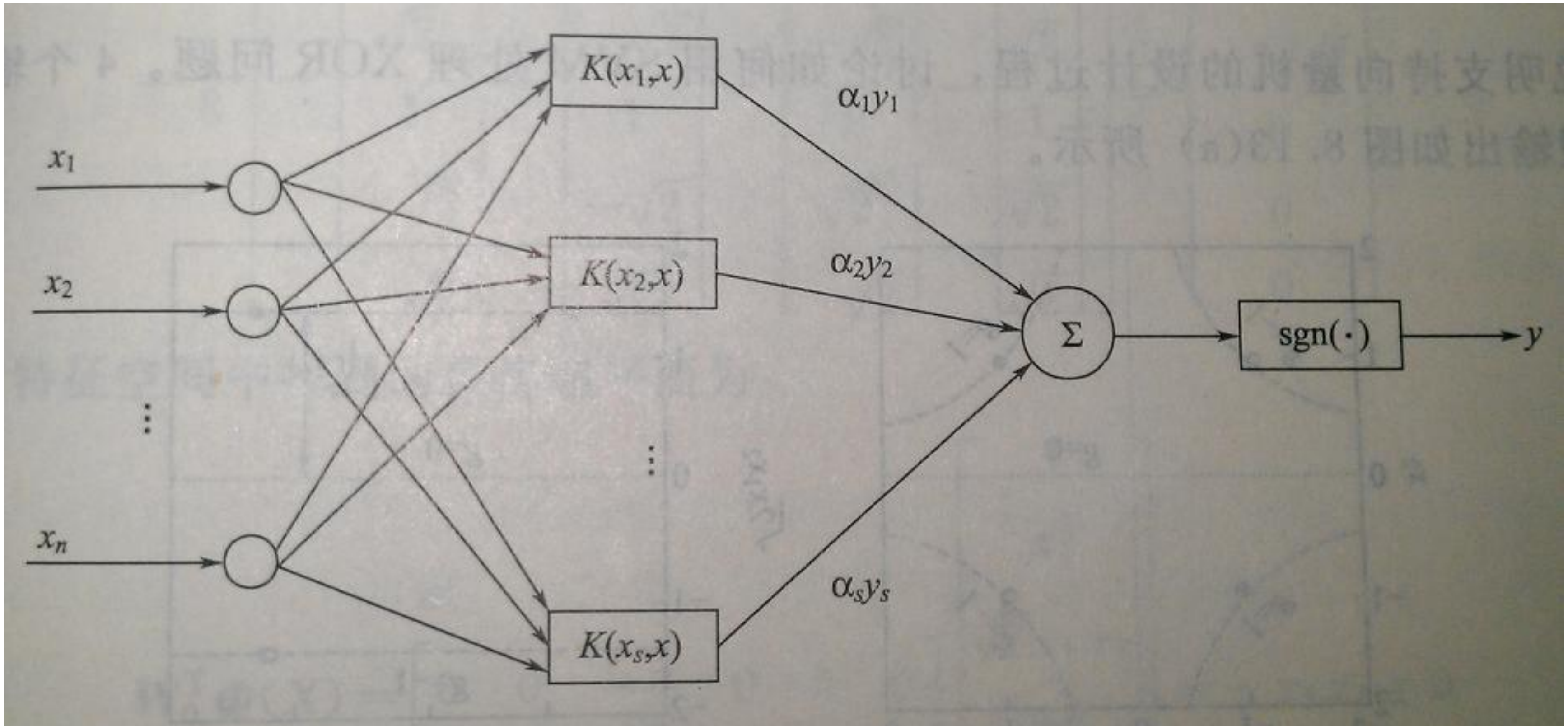
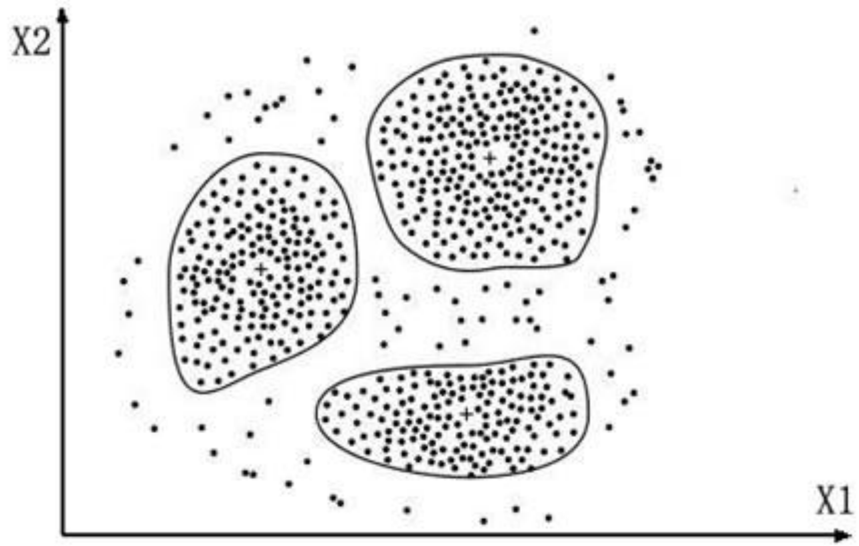


图 8.12 支持向量机神经网络

2012.7.20

聚类

聚类和分类判别有什么区别？



关键度量指标：距离

- 距离的定义
- 常用距离（薛毅书P469）

绝对值距离

欧氏距离

闵可夫斯基距离

切比雪夫距离

马氏距离

Lance和Williams距离

离散变量的距离计算

dist( )函数

x1=c(1,2,3,4,5)

x2=c(3,2,1,4,6)

x3=c(5,3,5,6,2)

x=data.frame(x1,x2,x3)

```
> dist(x,method="euclidean")
         1        2        3        4
2 2.449490
3 2.828427 2.449490
4 3.316625 4.123106 3.316625
5 5.830952 5.099020 6.164414 4.582576

> dist(x,method="minkowski")
         1        2        3        4
2 2.449490
3 2.828427 2.449490
4 3.316625 4.123106 3.316625
5 5.830952 5.099020 6.164414 4.582576
> dist(x,method="minkowski",p=5)
         1        2        3        4
2 2.024397
3 2.297397 2.024397
4 3.004922 3.143603 3.004922
5 4.323101 4.174686 5.085057 4.025455
```

dist( )函数

```
> y1=c("F","F","M","F","M")
> y2=c("A","B","B","C","A")
> y3=c(2,3,1,2,3)
> y=data.frame(y1,y2,y3)
> dist(y,method="binary")
  1 2 3 4
2 0
3 0 0
4 0 0 0
5 0 0 0 0
警告信息:
In dist(y, method = "binary") : 强制改变过程中产生了NA
> y1=c(1,0,1,1,0,0,1)
> y2=c(1,0,0,0,1,1,1)
> y3=c(1,1,1,0,0,1,1)
> y=data.frame(y1,y2,y3)
> dist(y,method="binary")
          1         2         3         4         5         6
2 0.6666667
3 0.3333333 0.5000000
4 0.6666667 1.0000000 0.5000000
5 0.6666667 1.0000000 1.0000000 1.0000000
6 0.3333333 0.5000000 0.6666667 1.0000000 0.5000000
7 0.0000000 0.6666667 0.3333333 0.6666667 0.6666667 0.3333333
```

2012.7.20

数据中心化与标准化变换

- 目的：使到各个变量平等地发挥作用
- scale( )函数
- 极差化。 sweep( )函数（薛毅书P473）

```
> x
  x1 x2 x3
1  1  3  5
2  2  2  3
3  3  1  5
4  4  4  6
5  5  6  2
> scale(x,center=TRUE,scale=TRUE)
             x1         x2         x3
[1,] -1.2649111 -0.1039750  0.4868645
[2,] -0.6324555 -0.6238503 -0.7302967
[3,]  0.0000000 -1.1437255  0.4868645
[4,]  0.6324555  0.4159002  1.0954451
[5,]  1.2649111  1.4556507 -1.3388774
attr(,"scaled:center")
 x1  x2  x3
3.0 3.2 4.2
attr(,"scaled:scale")
      x1       x2       x3
1.581139 1.923538 1.643168
```

2012.7.20

对变量进行分类的指标：相似系数

- 距离：对样本进行分类
- 相似系数：对变量进行分类
- 常用相似系数：夹角余弦，相关系数（薛毅书P475）

（凝聚的）层次聚类法

- 思想

1 开始时，每个样本各自作为一类

2 规定某种度量作为样本之间的距离及类与类之间的距离，并计算之

3 将距离最短的两个类合并为一个新类

4 重复2-3，即不断合并最近的两个类，每次减少一个类，直至所有样本被合并为一类

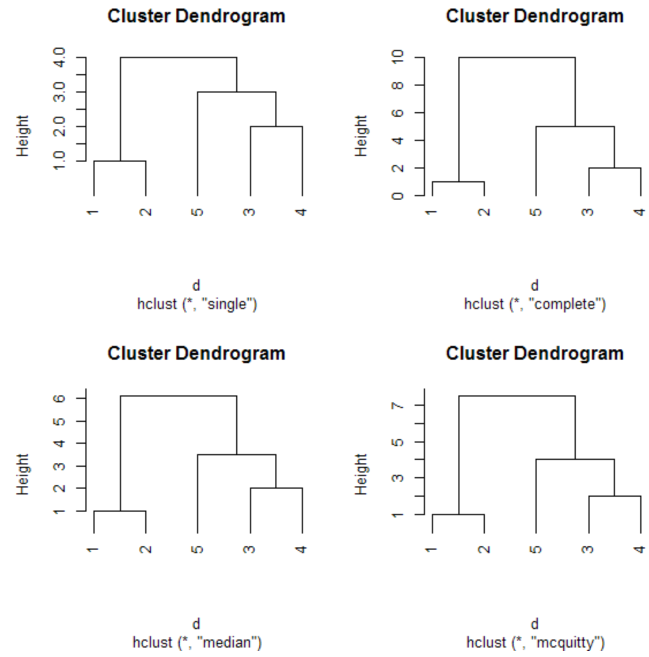
各种类与类之间距离计算的方法

- 薛毅书P476
- 最短距离法
- 最长距离法
- 中间距离法
- 类平均法
- 重心法
- 离差平方和法

hclust( )函数

- 简单的例子（薛毅书P480）

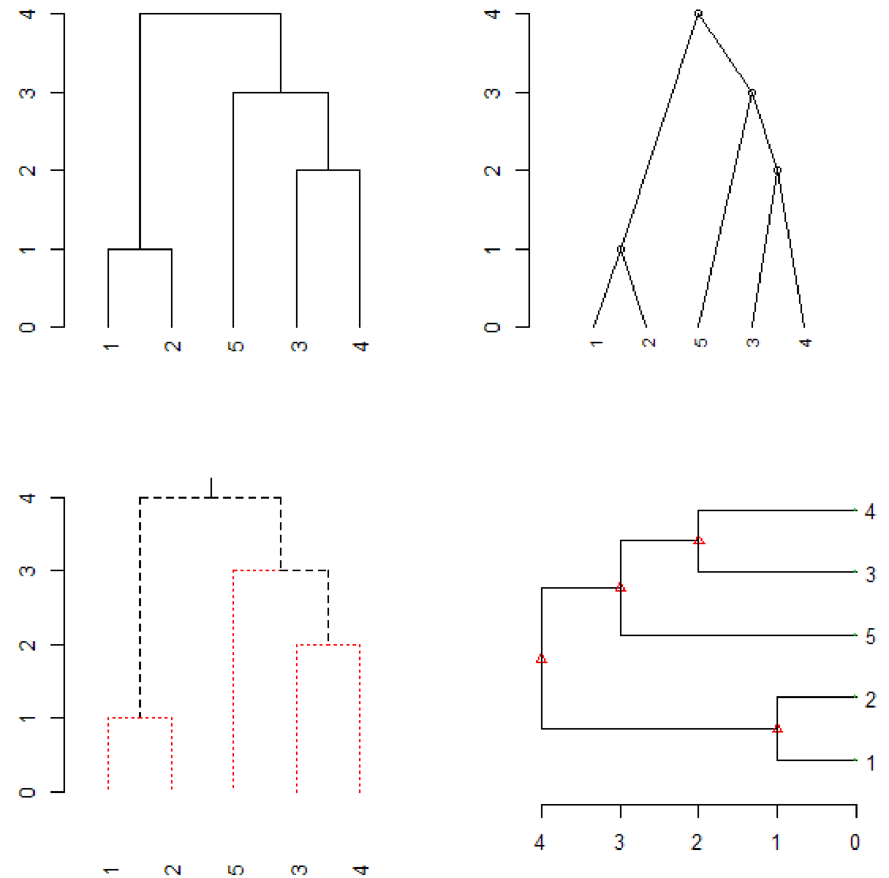
```
> x<-c(1,2,6,8,11); dim(x)<-c(5,1);
> x
     [,1]
[1,]    1
[2,]    2
[3,]    6
[4,]    8
[5,]   11
> d<-dist(x)
> d
   1  2  3  4
2  1
3  5  4
4  7  6  2
5 10  9  5  3
> hc1<-hclust(d, "single"); hc2<-hclust(d, "complete")
> hc3<-hclust(d, "median"); hc4<-hclust(d, "mcquitty")
> opar <- par(mfrow = c(2, 2))
> plot(hc1,hang=-1); plot(hc2,hang=-1)
> plot(hc3,hang=-1); plot(hc4,hang=-1)
> par(opar)
```



各种谱系图画法

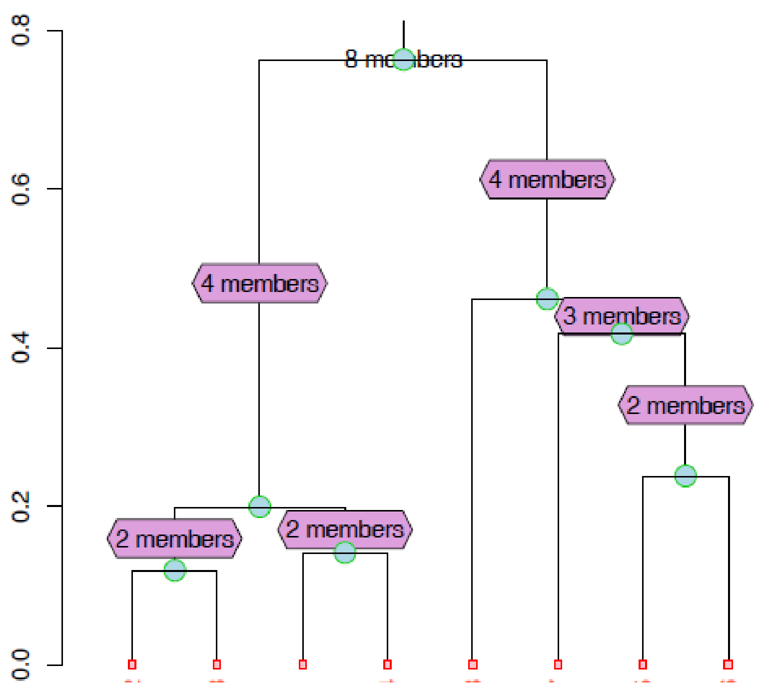
- as.dendrogram( )函数（薛毅书P482）

```
dend1<-as.dendrogram(hc1)
opar <- par(mfrow = c(2, 2),mar = c(4,3,1,2))
plot(dend1)
plot(dend1, nodePar=list(pch = c(1,NA),
     cex=0.8, lab.cex=0.8),
type = "t", center=TRUE)
plot(dend1, edgePar=list(col = 1:2, lty = 2:3),
dLeaf=1, edge.root = TRUE)
plot(dend1, nodePar=list(pch = 2:1,
     cex=.4*2:1, col=2:3),
horiz=TRUE)
par(opar)
```



对变量进行聚类分析

- 例子（薛毅书P483）

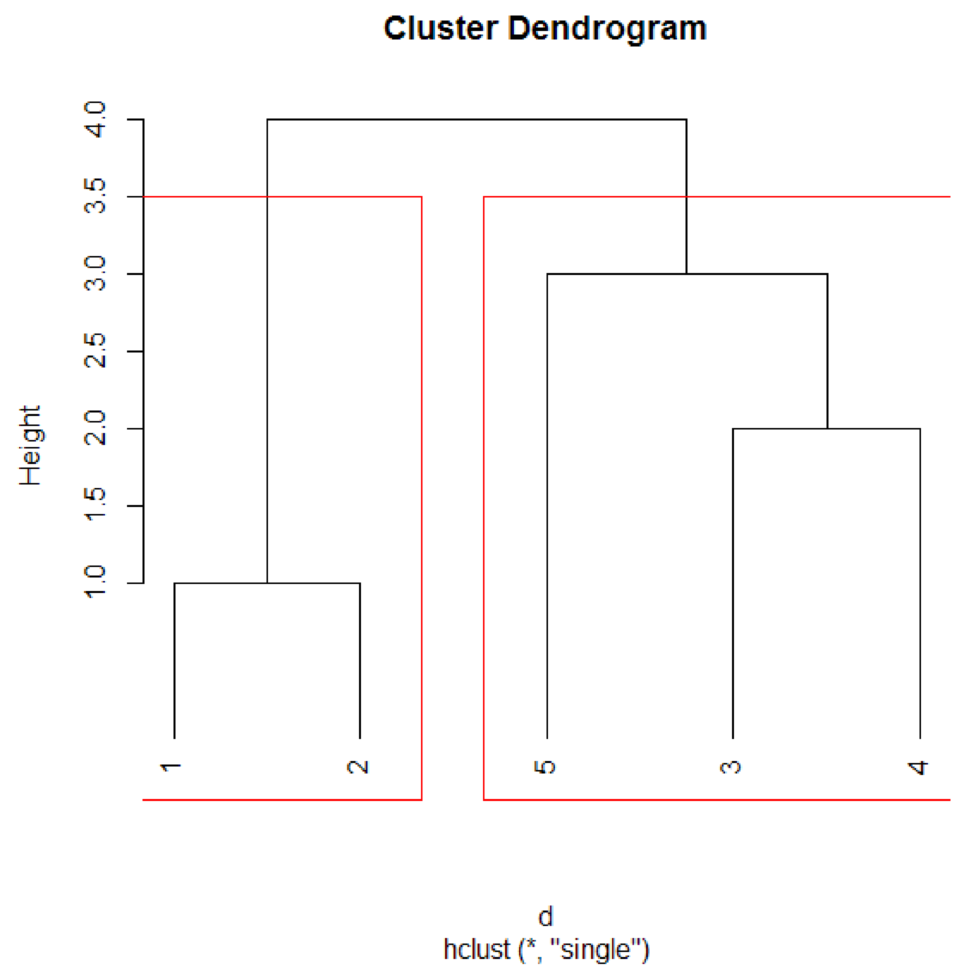


2012.7.20

分多少个类？

- rect.hclust( )函数

```
> plot(hc1,hang=-1)
> rect.hclust(hc1,k=2)
```



2012.7.20

综合实例

- 薛毅书P487

Thanks

FAQ时间