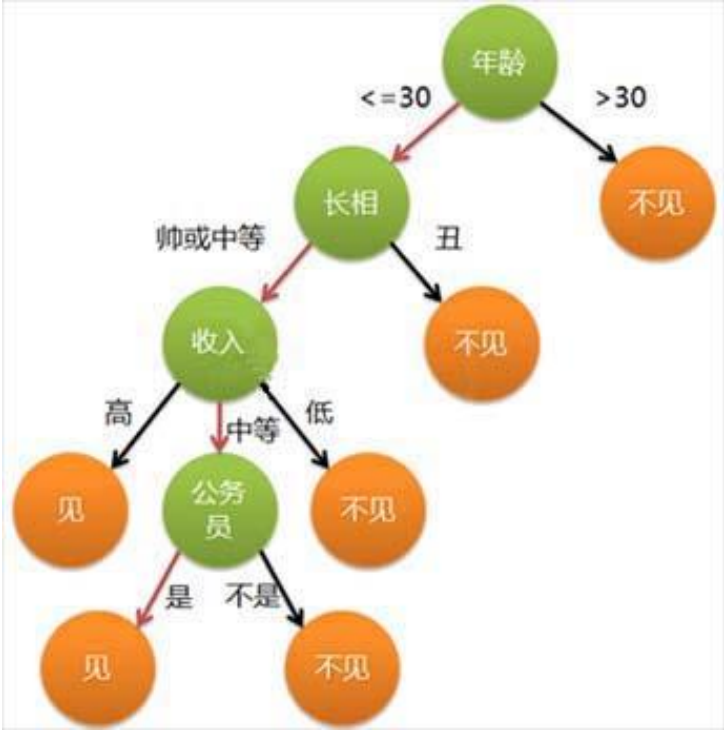


数据分析与R语言 第7周

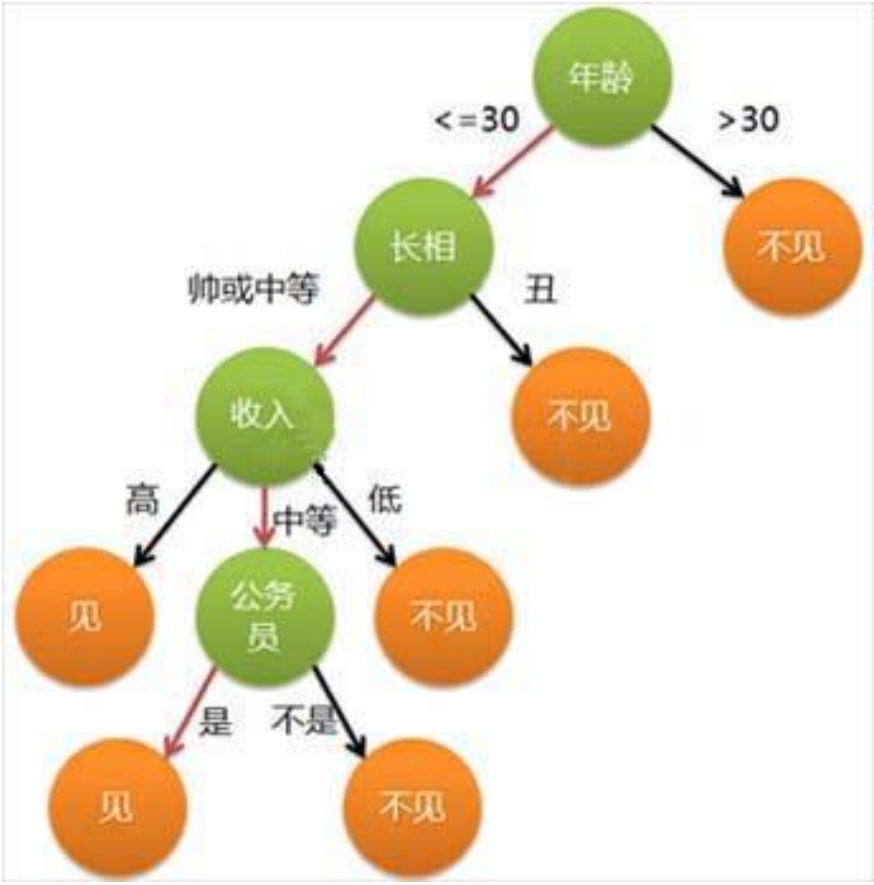
2012.6.28

分类：分类的意义

- 传统意义下的分类：生物物种
- 预测：天气预报
- 决策：yes or no
- 分类的传统模型
- 分类（判别分析）与聚类有什么差别？

常见分类模型与算法

- 线性判别法
- 距离判别法
- 贝叶斯分类器
- 决策树
- 支持向量机(SVM)
- 神经网络



线性判别法（Fisher）

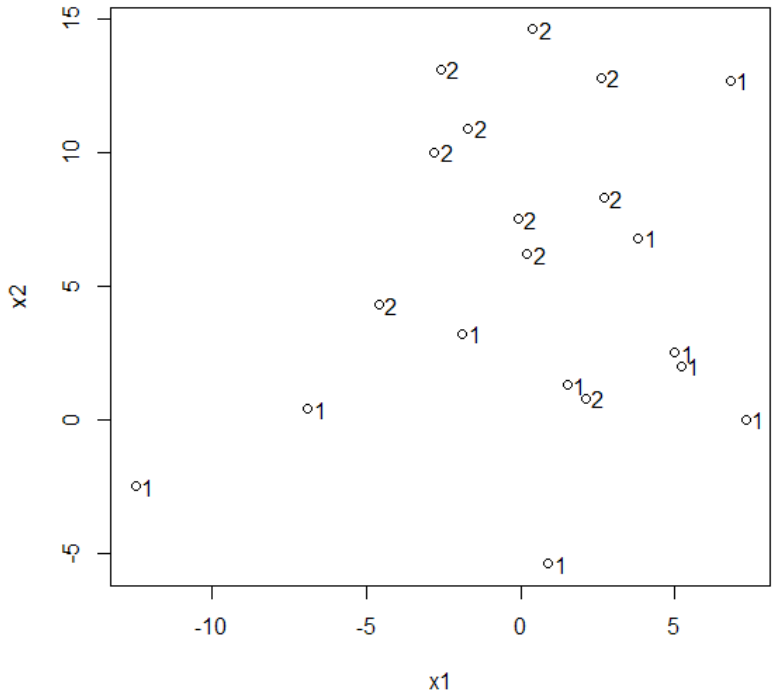
- 例子：天气预报数据

```
G=c(1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2)
x1=c(-1.9,-6.9,5.2,5.0,7.3,6.8,0.9,-12.5,1.5,3.8,0.2,-0.1,0.4,2.7,2.1,-4.6,-1.7,-2.6,2.6,-
    2.8)
x2=c(3.2,0.4,2.0,2.5,0.0,12.7,-5.4,-
    2.5,1.3,6.8,6.2,7.5,14.6,8.3,0.8,4.3,10.9,13.1,12.8,10.0)

a=data.frame(G,x1,x2)
plot(x1,x2)
text(x1,x2,G,adj=-0.5)
```

线性判别法的原理

- 用一条直线来划分学习集（这条直线一定存在吗？）
- 然后根据待测点在直线的哪一边决定它的分类



MASS包与线性判别函数lda( )

```
library(MASS)
ld=lda(G~x1+x2)
ld
```

```
> ld
Call:
lda(G ~ x1 + x2)

Prior probabilities of groups:
  1   2
0.5 0.5

Group means:
     x1   x2
1  0.92 2.10
2 -0.38 8.85

Coefficients of linear discriminants:
           LD1
x1 -0.1035305
x2  0.2247957
```

分类判断

```
z=predict(ld)

newG=z$class

 newG

 [1] 1 1 1 1 1 2 1 1 1 1 2 2 2 2 1 2 2 2 2 2

Levels: 1 2

cbind=(G,z$x,newG)

y=cbind(G,z$x,newG)

y
```

	G	LD1	newG
1	1	-0.28674901	1
2	1	-0.39852439	1
3	1	-1.29157053	1
4	1	-1.15846657	1
5	1	-1.95857603	1
6	1	0.94809469	2
7	1	-2.50987753	1
8	1	-0.47066104	1
9	1	-1.06586461	1
10	1	-0.06760842	1
11	2	0.17022402	2
12	2	0.49351760	2
13	2	2.03780185	2
14	2	0.38346871	2
15	2	-1.24038077	1
16	2	0.24005867	2
17	2	1.42347182	2
18	2	2.01119984	2
19	2	1.40540244	2
20	2	1.33503926	2

距离判别法

- 原理：计算待测点与各类的**距离**，取最短者为其所属分类
- 马氏距离（薛毅书p445，为什么不用欧氏距离？），计算函数mahalanobis( )

定义 8.1 设 x, y 是服从均值为 μ, 协方差阵为 Σ 的总体 X 中抽取的样本，则总体 X 内两点 x 与 y 的 *Mahalanobis* 距离（简称马氏距离）定义为

$$d(x,y) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)}. \tag{8.1}$$

定义样本 x 与总体 X 的 *Mahalanobis* 距离为

$$d(x,X) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}. \tag{8.2}$$

算法

- 情形一（薛毅书p445）

首先考虑两个总体 X_1 和 X_2 的协方差相同的情况，即

$$\mu_1 \neq \mu_2, \quad \Sigma_1 = \Sigma_2 = \Sigma.$$

要判断 x 是属于哪一个总体，需要计算 x 到总体 X_1 和 X_2 的 Mahalanobis 距离的平方 $d^2(x, X_1)$ 和 $d^2(x, X_2)$，然后进行比较，若 $d^2(x, X_1) \leq d^2(x, X_2)$，则判定 x 属于 X_1；否则判定 x 来自 X_2. 由此得到如下判别准则：

$$R_1 = \{x \mid d^2(x, X_1) \leq d^2(x, X_2)\}, \quad R_2 = \{x \mid d^2(x, X_1) > d^2(x, X_2)\}. \tag{8.3}$$

令

$$w(x) = (x - \overline{\mu})^T \Sigma^{-1} (\mu_1 - \mu_2), \tag{8.5}$$

称 $w(x)$ 为两总体距离的判别函数，因此判别准则 (8.3) 变为

$$R_1 = \{x \mid w(x) \geq 0\}, \quad R_2 = \{x \mid w(x) < 0\}. \tag{8.6}$$

算法

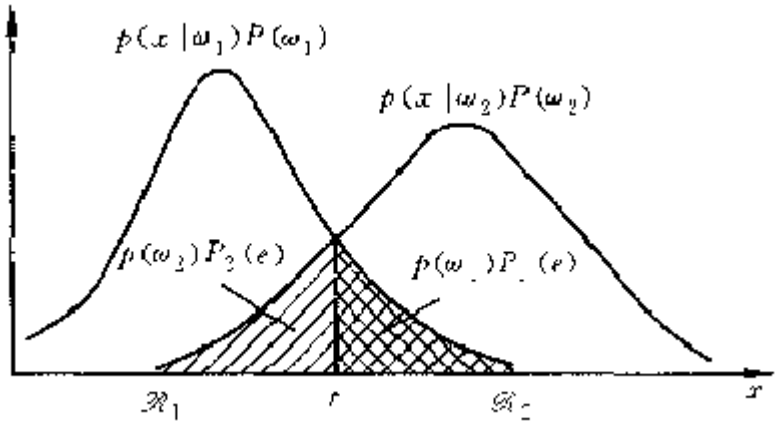
- 情形二（薛毅书p447）
- 例子（薛毅书p449）

对于样本 x, 在协方差阵不同的情况下，判别函数为

$$w(x) = (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) - (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1). \qquad (8.12)$$

贝叶斯分类器

- 原理（薛毅书p455)



$$R_1 = \left\{ x \mid \frac{f_1(x)}{f_2(x)} \geq \frac{L(1|2)}{L(2|1)} \cdot \frac{p_2}{p_1} \right\}, \quad R_2 = \left\{ x \mid \frac{f_1(x)}{f_2(x)} < \frac{L(1|2)}{L(2|1)} \cdot \frac{p_2}{p_1} \right\}.$$

结果

- 对于总体协方差矩阵相同的情形

$$R_1 = \left\{ x \;\middle|\; W(x) \geq \beta \right\}, \quad R_2 = \left\{ x \;\middle|\; W(x) < \beta \right\}, \tag{8.26}$$

其中

$$\begin{aligned} W(x) &= \frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) \\ &= \left[x - \frac{1}{2}(\mu_1+\mu_2) \right]^T \Sigma^{-1}(\mu_1-\mu_2), \qquad (8.27) \\ \beta &= \ln \frac{L(1|2)\cdot p_2}{L(2|1)\cdot p_1}. \qquad (8.28) \end{aligned}$$

结果

- 对于总体协方差矩阵不同的情形

$$R_1 = \left\{ x \;\middle|\; W(x) \geq \beta \right\}, \quad R_2 = \left\{ x \;\middle|\; W(x) < \beta \right\}, \tag{8.29}$$

其中

$$W(x) = \frac{1}{2}(x-\mu_2)^T \Sigma_2^{-1}(x-\mu_2) - \frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1), \tag{8.30}$$

$$\beta = \ln \frac{L(1|2)\cdot p_2}{L(2|1)\cdot p_1} + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right). \tag{8.31}$$

程序与例子

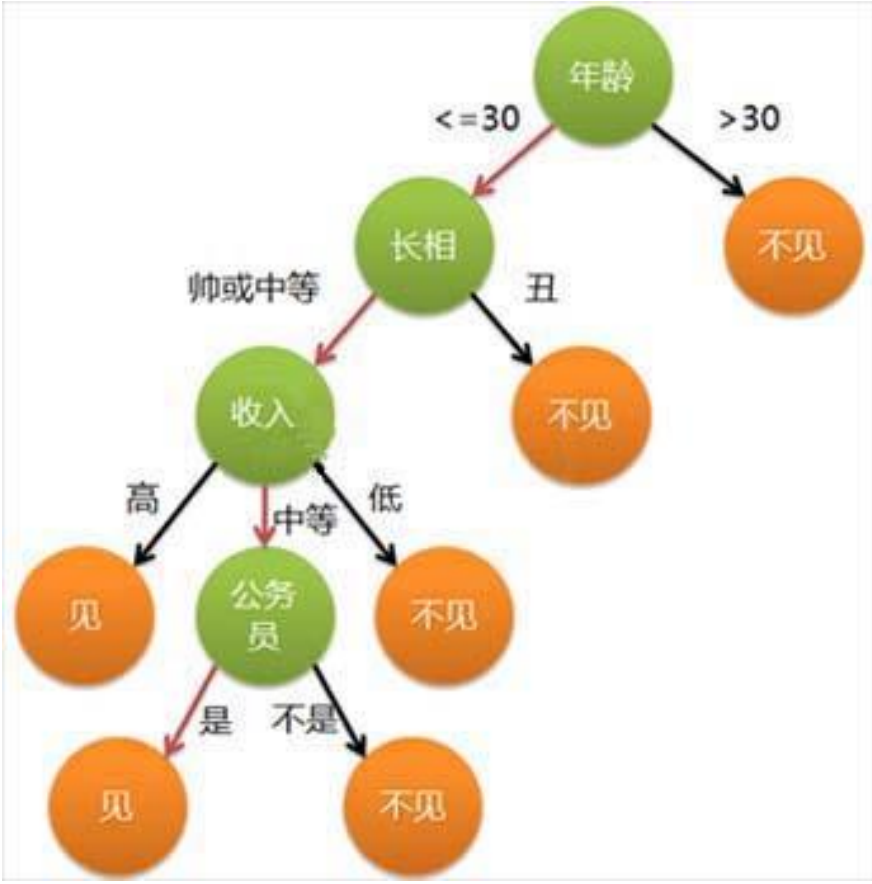
- 薛毅书P457
- 利用贝叶斯分类器判断垃圾邮件

多分类的情况

- 多分类下的距离判别法（薛毅书p452）
- 多分类下的贝叶斯（薛毅书p460）

决策树 decision tree

- 什么是决策树
- 输入：学习集
- 输出：分类规则（决策树）



例子

- 用SNS社区中不真实账号检测的例子说明如何使用ID3算法构造决策树。为了简单起见，我们假设训练集合包含10个元素。其中s、m和l分别表示小、中和大。

日志密度	好友密度	是否使用真实头像	账号是否真实
s	s	no	no
s	l	yes	yes
l	m	yes	yes
m	m	yes	yes
l	m	yes	yes
m	l	no	yes
m	s	no	no
l	m	no	yes
m	s	no	yes
s	s	yes	no

信息增益

- 设L、F、H和R表示日志密度、好友密度、是否使用真实头像和账号是否真实，下面计算各属性的信息增益。

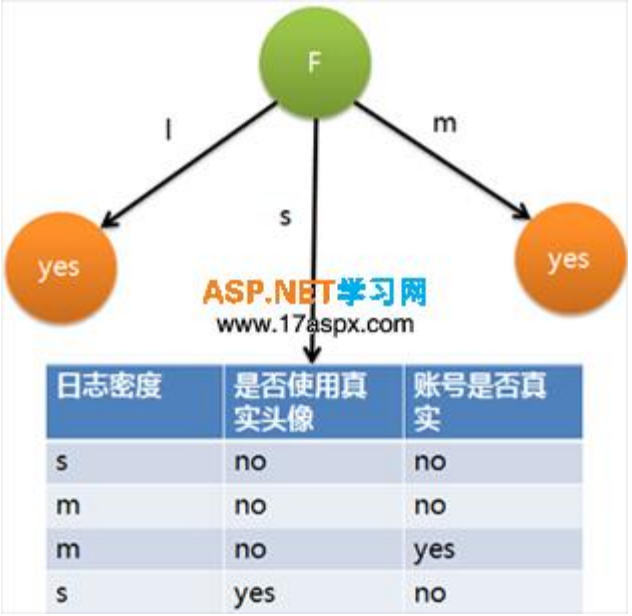
$$info(D) = -0.7log_20.7 - 0.3log_20.3 = 0.7 * 0.51 + 0.3 * 1.74 = 0.879$$

$$info_L(D) = 0.3 * (-\frac{0}{3}log_2\frac{0}{3} - \frac{3}{3}log_2\frac{3}{3}) + 0.4 * (-\frac{1}{4}log_2\frac{1}{4} - \frac{3}{4}log_2\frac{3}{4}) + 0.3 * (-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3}) = 0 + 0.326 + 0.277 = 0.603$$

$$gain(L) = 0.879 - 0.603 = 0.276$$

根据信息增益选择分裂属性

- 因此日志密度的信息增益是0.276。用同样方法得到H和F的信息增益分别为0.033和0.553。因为F具有最大的信息增益，所以第一次分裂选择F为分裂属性，分裂后的结果如下图表示：



日志密度	是否使用真实头像	账号是否真实
s	no	no
m	no	no
m	no	yes
s	yes	no

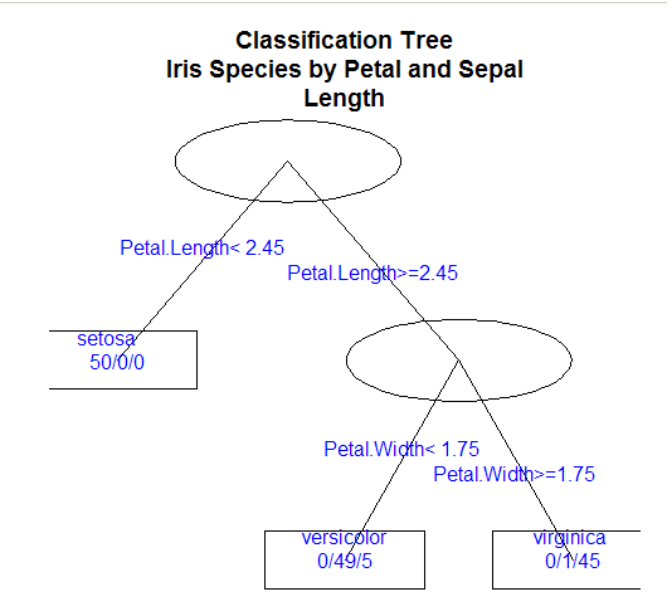
递归+分而治之

- 在上图的基础上，再递归使用这个方法计算子节点的分裂属性，最终就可以得到整个决策树。
- 这个方法称为ID3算法，还有其它的算法也可以产生决策树
- 对于特征属性为**连续值**，可以如此使用ID3算法：先将D中元素按照特征属性排序，则每两个相邻元素的中间点可以看做潜在分裂点，从第一个潜在分裂点开始，分裂D并计算两个集合的期望信息，具有最小期望信息的点称为这个属性的最佳分裂点，其信息期望作为此属性的信息期望。

R语言实现决策树：rpart扩展包

- 以鸢尾花数据集作为算例说明

```
iris.rp = rpart(Species~., data=iris,
    method="class")
plot(iris.rp, uniform=T, branch=0,
    margin=0.1, main= “ Classification
    Tree\nIris Species by Petal and Sepal
    Length")
text(iris.rp, use.n=T, fancy=T, col="blue")
```



Rule 1: if Petal.Length>=2.45&Petal.Width<1.75, then it is versicolor(0/49/5)
Rule2: if Petal.Length>=2.45&Petal.Width>=1.75, then it is virginica (0/1/45)
Rule 3: if Petal.Length<2.45, then it is setosa (50/0/0)

Thanks

FAQ时间