判别分析

判别分析（discriminant analysis）是一种分类技术。它通过一个已知类别的“训练样本”来建立判别准则，并通过预测变量来为未知类别的数据进行分类。

判别分析的方法大体上有三类，即**Fisher判别（线性判别）**、**Bayes判别**和**距离判别**。**Fisher判别思想**是投影降维，使多维问题简化为一维问题来处理。选择一个适当的投影轴，使所有的样品点都投影到这个轴上得到一个投影值。对这个投影轴的方向的要求是：使每一组内的投影值所形成的组内离差尽可能小，而不同组间的投影值所形成的类间离差尽可能大。**Bayes判别思想**是根据先验概率求出后验概率，并依据后验概率分布作出统计推断。**距离判别思想**是根据已知分类的数据计算各类别的重心，对未知分类的数据，计算它与各类重心的距离，与某个重心距离最近则归于该类。

接下来将通过例题展示不同的判别方法。

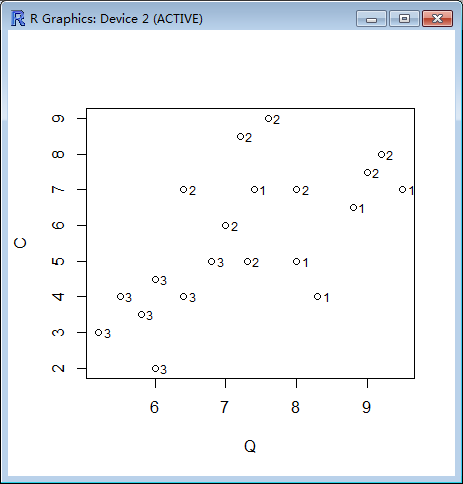
例1：在某市场抽取20种牌子的电视机中，5种畅销，8种平销，另外7种滞销。按电视质量评分、功能评分和销售价格三项指标衡量，销售状态：1为畅销，2为平销，3为滞销。

（数据集：d6.3）

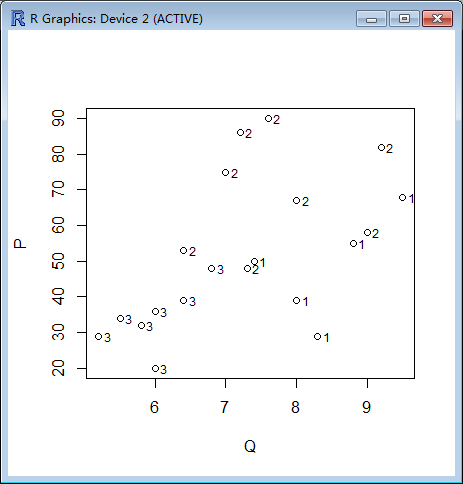
> X=read.table("clipboard",header=T) #读取数据存入X中

> plot(X$Q, X$C); #做横坐标为Q，纵坐标为C的散点图

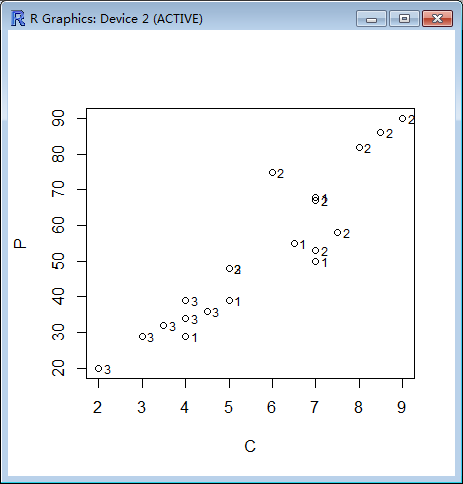
> text(X$Q, X$C, X$G,adj=-0.8,cex=0.75) #在上一句的散点图中为每个点加文本；Q,C,G表示依据Q和C加上G的文本名字；adj为调整文字与点距离的选项，+为向左，-为向右；cex为调整文字的大小；



>plot(X$Q, X$P);text(X$Q, X$P, X$G,adj=-0.8,cex=0.75) #同上



> plot(X$C, X$P);text(X$C, X$P, X$G,adj=-0.8,cex=0.75) #同上



# 1 线性判别（等方差）

R中线性判别和贝叶斯判别的函数为lda()。贝叶斯判别的先验概率设为相等即为线性判别。

> library(MASS)

>attach(X)

>(ld=lda(G~Q+C+P,prior=c(1,1,1)/3)) #线性判别G为类别变量，按照Q、C、P三个自变量来分类，prior为设定先验概率，本例为三类，故先验概率相等，为1/3,1/3,1/3. 先验概率相等即为线性判别

Call:

lda(G ~ Q + C + P, prior = c(1, 1, 1)/3)

Prior probabilities of groups: #先验概率值，表示每类在原样本所占的比例

1 2 3

0.3333333 0.3333333 0.3333333

Group means: #每类的均值

Q C P

1 8.400000 5.900000 48.200

2 7.712500 7.250000 69.875

3 5.957143 3.714286 34.000

Coefficients of linear discriminants: #线性判别函数

LD1 LD2

Q -0.92307369 0.76708185

C -0.65222524 0.11482179

P 0.02743244 -0.08484154

Proportion of trace: #两个判别函数的迹（判别函数的判别能力）

LD1 LD2

0.7259 0.2741

> Z=predict(ld) #判别结果ld的预测类别结果存入Z中

> newG=Z$class #将Z中的class变量存为newG

> cbind(X$G,Z$x,newG) #将G（原来的类别）、z中的x（即两个判别函数）、newG（用线性判别后的结果）合并出结果

G LD1 LD2 newG

1 1 -0.40839476 2.37788417 1

2 1 -2.40289378 0.33402788 1

3 1 -0.50937350 1.41416605 1

4 1 -0.95822294 0.25030363 1

5 1 -1.78725129 0.84259965 1

6 2 -2.54179395 0.85631321 1

7 2 0.74904277 -2.29238928 2

8 2 -2.39414277 -0.96905637 2

9 2 -1.04571568 -0.73175336 2

10 2 -1.34999059 -2.76029783 2

11 2 -0.76437825 -2.78517532 2

12 2 0.04714807 -0.77130282 2

13 2 0.38367004 0.11363494 3

14 3 2.77223326 1.14752615 3

15 3 1.61976965 0.07201330 3

16 3 0.84520688 -0.26990599 3

17 3 3.10535893 -0.11489136 3

18 3 2.30769941 0.14824404 3

19 3 2.31337377 -0.19415269 3

20 3 1.58058919 0.07711606 3

>tab=table(X$G,newG) #做二联列表

> tab

newG

说明：原属于第1类的样品中，后判入第1类为5个，后判入第2类为0个，后判入第3类为0个；原属于第2类的样品中，后判入第1类为1个，后判入第2类为6个，后判入第3类为1个；原属于第3类的样品中，后判入第1类为0个，后判入第2类为0个，后判入第3类为7个；判定准确的有5+6+7=18个

G 1 2 3

1 5 0 0

2 1 6 1

3 0 0 7

[1] 0.9

> sum(diag(prop.table(tab))) #对角线上为判断正确的值，prop.table为求比例，diag为求对角线元素，sum为对角线元素加和

只有两个样品判错，判别符合率90%。判别效果还可以。

> plot(Z$x)

> text(Z$x[,1],Z$x[,2],X$G,col=X$G,adj=-0.8,cex=0.75)



# 2 二次判别（异方差）

> qd=qda(G~Q+C+P,data=X, prior = c(1, 1, 1)/3) #同上

> qd

Call:

qda(G ~ Q + C + P, data = X, prior = c(1, 1, 1)/3)

Prior probabilities of groups:

1 2 3

0.3333333 0.3333333 0.3333333

Group means:

Q C P

1 8.400000 5.900000 48.200

2 7.712500 7.250000 69.875

3 5.957143 3.714286 34.000

G newG

> Z=predict(qd)

> newG=Z$class

> cbind(G,Z$x,newG)

[1,] 1 1

[2,] 1 1

[3,] 1 1

[4,] 1 1

[5,] 1 1

[6,] 2 2

[7,] 2 2

[8,] 2 2

[9,] 2 2

[10,] 2 2

[11,] 2 2

[12,] 2 2

[13,] 2 3

[14,] 3 3

[15,] 3 3

[16,] 3 3

[17,] 3 3

[18,] 3 3

[19,] 3 3

[20,] 3 3

> (tab=table(G,newG))

newG

G 1 2 3

1 5 0 0

2 0 7 1

3 0 0 7

> sum(diag(prop.table(tab)))

[1] 0.95

# 3 贝叶斯判别

取q1=5/20，q2=8/20，q3=7/20

> (ld2=lda(G~Q+C+P,prior=c(5,8,7)/20)) #同上

Call:

lda(G ~ Q + C + P, prior = c(5, 8, 7)/20)

Prior probabilities of groups:

1 2 3

0.25 0.40 0.35

Group means:

Q C P

1 8.400000 5.900000 48.200

2 7.712500 7.250000 69.875

3 5.957143 3.714286 34.000

Coefficients of linear discriminants:

LD1 LD2

Q -0.81173396 0.88406311

C -0.63090549 0.20134565

P 0.01579385 -0.08775636

Proportion of trace:

LD1 LD2

0.7403 0.2597

> Z2=predict(ld2)

> cbind(G,Z2$x,Z2$class)

G LD1 LD2 newG

1 1 -0.1409984 2.582951755 1

2 1 -2.3918356 0.825366275 1

3 1 -0.3704452 1.641514840 1

4 1 -0.9714835 0.548448277 1

5 1 -1.7134891 1.246681993 1

6 2 -2.4593598 1.361571174 1

7 2 0.3789617 -2.200431689 2

8 2 -2.5581070 -0.467096091 2

9 2 -1.1900285 -0.412972027 2

10 2 -1.7638874 -2.382302324 2

11 2 -1.1869165 -2.485574940 2

12 2 -0.1123680 -0.598883922 2

13 2 0.3399132 0.232863397 3

14 3 2.8456561 0.936722573 3

15 3 1.5592346 0.025668216 3

16 3 0.7457802 -0.209168159 3

17 3 3.0062824 -0.358989534 3

18 3 2.2511708 0.008852067 3

19 3 2.2108260 -0.331206768 3

20 3 1.5210939 0.035984885 3

> table(G,Z2$class)

newG

G 1 2 3

1 5 0 0

2 1 6 1

3 0 0 7

由判别符合率可知，用贝叶斯判别函数进行判别的效果还是不错的。

> Z2$post #后验概率

1 2 3

1 9.747144e-01 0.0088198914 1.646573e-02

2 7.069622e-01 0.2929273720 1.104490e-04

3 9.066215e-01 0.0667161901 2.666229e-02

4 5.420713e-01 0.4473098858 1.061882e-02

5 8.565024e-01 0.1428619923 6.355979e-04

6 8.893310e-01 0.1106146104 5.440087e-05

7 2.123192e-03 0.8789838980 1.188929e-01

8 1.188275e-01 0.8811091490 6.337076e-05

9 1.241051e-01 0.8713952505 4.499663e-03

10 1.781107e-03 0.9980655088 1.533841e-04

11 1.374025e-03 0.9977941856 8.317893e-04

12 7.358709e-02 0.8254938102 1.009191e-01

13 2.163910e-01 0.3857093332 3.978997e-01

14 5.425064e-04 0.0002260742 9.992314e-01

15 8.646125e-03 0.0260022554 9.653516e-01

16 5.604906e-02 0.2740442884 6.699067e-01

17 5.674877e-05 0.0004331326 9.995101e-01

18 9.943552e-04 0.0032114683 9.957942e-01

19 7.112081e-04 0.0049072777 9.943815e-01

20 9.840783e-03 0.0288659040 9.612933e-01

后验概率给出了样品落在各个类的大小.

# 4 距离判别

>library(mvstats)

> D=discrim.dist(cbind(X$Q, X$C, X$P),as.factor(X$G)) #距离判别函数discrim.dist，as.factor表示以G作为分类变量

## 4.1距离判别（异方差）

G D.1 D.2 D.3 newG

1 1 2.523945 9.1098063 67.45440177 1

2 1 2.647988 4.0185551 44.62599853 1

3 1 2.351227 4.5484097 39.84664742 1

4 1 2.834838 3.0750203 45.15294171 1

5 1 1.642002 2.6003823 43.92182886 1

6 2 36.094079 3.3813060 71.64252058 2

7 2 524.483775 3.5490324 89.58755724 2

8 2 29.617174 2.7431000 45.06306177 2

9 2 55.716748 0.1739398 17.90918643 2

10 2 159.307720 2.2098606 54.82120347 2

11 2 220.973242 1.7891625 54.33005806 2

12 2 51.324207 4.1752738 20.49558450 2

13 2 108.302944 2.9783249 6.30248460 2

14 3 277.546516 16.4806425 4.64954551 3

15 3 225.150834 6.5934388 1.10207828 3

16 3 166.717076 3.1779814 3.39680714 2

17 3 420.877625 13.4992376 3.50621981 3

18 3 273.863395 9.7542839 0.09427008 3

19 3 272.231240 9.2729988 1.49170488 3

20 3 154.746222 6.7145131 3.75937430 3

> table(D$G,D$newG)

1 2 3

1 5 0 0

2 0 8 0

3 0 1 6

## 4.2 距离判别（等方差）

> D=discrim.dist(cbind(X$Q, X$C, X$P),as.factor(X$G),var.equal=T) #距离判别函数discrim.dist，as.factor表示以G作为分类变量，var.equal=T表示方差相等，默认为不相等即F；

G D.1 D.2 D.3 newG

1 1 1.4453619 7.537584 5.60401471 1

2 1 2.2664280 3.815572 8.03923348 1

3 1 0.2180557 3.847188 3.33168261 1

4 1 4.2187373 4.998226 6.77363946 1

5 1 0.1387577 2.772724 5.25312540 1

6 2 1.2052321 4.163269 7.96097682 1

7 2 11.5119044 5.038159 7.06315239 2

8 2 3.9971279 2.160277 8.04431766 2

9 2 1.9114280 0.175065 3.36826837 2

10 2 8.5080434 1.622153 7.96253787 2

11 2 8.8262908 1.651934 6.98824286 2

12 2 7.6853199 5.428475 6.73903800 2

13 2 1.7100195 1.621708 1.19880008 3

14 3 5.8449048 7.527269 1.54883666 3

15 3 3.4594845 2.790551 0.23425362 3

16 3 2.5944645 1.334076 0.58787430 3

17 3 7.0518669 5.329790 0.38554940 3

18 3 4.6223349 3.888282 0.01685556 3

19 3 5.6101515 3.985109 0.54399031 3

20 3 3.9324003 3.291692 0.79815067 3

> table(D$G,D$newG)

1 2 3

1 5 0 0

2 1 6 1

3 0 0 7

例2：对21个破产企业收集它们在破产前两年的财务数据，对25个财务良好的企业也收集同一时期的数据。数据涉及四个变量：CF\_TD（现金流量/总债务），NI\_TA （净收入/总资产）， CA\_CL（流动资产/流动债务），CA\_NS（流动资产/净销售额），一个分组变量：企业现状G（1：非破产企业，2：破产企业）

（数据集：case5）

> library(MASS) ##载入MASS函数包

> ld=lda(G~.,data=Case5,prior = c(1, 1)/2);ld #线性判别

> Case5=read.table("clipboard",header=T) ##读取数据

> head(Case5,30) ##显示数据的前30行

（1）线性判别

Call:

lda(G ~ ., data = Case5)

Prior probabilities of groups:

1 2

0.5 0.5

Group means:

CF\_TD NI\_TA CA\_CL CA\_NS

1 0.23520000 0.05560000 2.593600 0.426800

2 -0.06809524 -0.08142857 1.366667 0.437619

Coefficients of linear discriminants:

LD1

CF\_TD -0.6291667

NI\_TA -4.4458516

CA\_CL -0.8892843

CA\_NS 1.1844801

> Zld=predict(ld) ##判别

> data.frame(Case5$G,Zld$class,round(Zld$x,3))

Case5.G Zld.class LD1

1 1 1 -1.096

2 1 1 -0.055

3 1 1 -1.977

4 1 1 -0.708

5 1 1 -2.293

6 1 1 -2.313

7 1 1 -0.477

8 1 1 -0.241

9 1 1 -0.866

10 1 1 -0.159

11 1 1 -0.351

12 1 1 -0.185

13 1 2 1.188

14 1 1 -0.860

15 1 1 -0.437

16 1 1 -0.896

17 1 1 -0.391

18 1 1 -1.088

19 1 1 -0.268

20 1 1 -0.348

21 1 1 -2.891

22 1 1 -0.652

23 1 1 -1.738

24 1 1 -1.054

25 1 1 -3.644

26 2 2 2.914

27 2 2 1.821

28 2 2 0.693

29 2 2 0.707

30 2 2 1.113

31 2 2 1.344

32 2 2 0.682

33 2 2 0.804

34 2 2 0.429

35 2 2 1.195

36 2 2 2.642

37 2 2 0.242

38 2 2 0.155

39 2 2 2.166

40 2 1 -0.425

41 2 1 -0.797

42 2 2 0.805

43 2 2 0.711

44 2 2 0.828

45 2 1 -0.133

46 2 2 2.095

> addmargins(table(Case5$G,Zld$class))#在列联表上添加边缘列

1 2 Sum

1 24 1 25

2 3 18 21

Sum 27 19 46

准确率为(24+18)/46=91.3%

（2）二次判别

> qd=qda(G~.,data=Case5,prior=c(1,1)/2);qd #二次判别

Call:

qda(G ~ ., data = Case5, prior = c(1, 1)/2)

Prior probabilities of groups:

1 2

0.5 0.5

Group means:

CF\_TD NI\_TA CA\_CL CA\_NS

1 0.23520000 0.05560000 2.593600 0.426800

> Zqd=predict(qd)

> data.frame(Case5$G,Zqd$class,round(Zqd$post,3)\*100)

2 -0.06809524 -0.08142857 1.366667 0.437619

Case5.G Zqd.class X1 X2

1 1 1 100.0 0.0

2 1 1 61.9 38.1

3 1 1 100.0 0.0

4 1 1 97.5 2.5

5 1 1 100.0 0.0

6 1 1 100.0 0.0

7 1 1 93.4 6.6

8 1 1 82.3 17.7

9 1 1 96.9 3.1

10 1 1 51.0 49.0

11 1 1 83.3 16.7

12 1 1 98.5 1.5

13 1 2 26.1 73.9

14 1 1 98.6 1.4

15 1 1 85.0 15.0

16 1 1 97.5 2.5

17 1 1 95.4 4.6

18 1 1 100.0 0.0

19 1 1 100.0 0.0

20 1 1 100.0 0.0

21 1 1 100.0 0.0

22 1 1 72.5 27.5

23 1 1 100.0 0.0

24 1 1 99.6 0.4

25 1 1 100.0 0.0

26 2 2 0.0 100.0

27 2 2 0.0 100.0

28 2 2 43.2 56.8

29 2 2 0.2 99.8

30 2 2 0.0 100.0

31 2 2 6.9 93.1

32 2 2 7.8 92.2

33 2 2 1.5 98.5

34 2 2 17.9 82.1

35 2 2 0.0 100.0

36 2 2 0.0 100.0

37 2 2 32.6 67.4

38 2 2 36.4 63.6

39 2 2 0.0 100.0

40 2 1 72.2 27.8

41 2 1 97.4 2.6

42 2 2 0.3 99.7

43 2 2 1.7 98.3

44 2 2 14.4 85.6

45 2 2 1.3 98.7

46 2 2 0.0 100.0

> addmargins(table(Case5$G,Zqd$class))

1 2 Sum

1 24 1 25

2 2 19 21

Sum 26 20 46

准确率为(24+19)/46=93.5%

（3）贝叶斯判别

> ld2=lda(G~.,data=Case5);ld2 ##不设定先验概率，即默认为样本中的比例。

Call:

lda(G ~ ., data = Case5)

Prior probabilities of groups:

1 2

0.5434783 0.4565217

Group means:

CF\_TD NI\_TA CA\_CL CA\_NS

1 0.23520000 0.05560000 2.593600 0.426800

2 -0.06809524 -0.08142857 1.366667 0.437619

Coefficients of linear discriminants:

LD1

CF\_TD -0.6291667

NI\_TA -4.4458516

CA\_CL -0.8892843

CA\_NS 1.1844801

> Zld2=predict(ld2)

> data.frame(Case5$G,Zld2$class,round(Zld2$x,3))

Case5.G Zld2.class LD1

1 1 1 -1.013

2 1 1 0.028

3 1 1 -1.895

4 1 1 -0.625

5 1 1 -2.210

6 1 1 -2.230

7 1 1 -0.395

8 1 1 -0.158

9 1 1 -0.783

10 1 1 -0.076

11 1 1 -0.268

12 1 1 -0.102

13 1 2 1.271

14 1 1 -0.778

15 1 1 -0.354

16 1 1 -0.813

17 1 1 -0.308

18 1 1 -1.005

19 1 1 -0.185

20 1 1 -0.265

21 1 1 -2.808

22 1 1 -0.569

23 1 1 -1.655

24 1 1 -0.971

25 1 1 -3.562

26 2 2 2.997

27 2 2 1.904

28 2 2 0.776

29 2 2 0.790

30 2 2 1.196

31 2 2 1.426

32 2 2 0.764

33 2 2 0.887

34 2 2 0.512

35 2 2 1.278

36 2 2 2.725

37 2 2 0.325

38 2 2 0.238

39 2 2 2.249

40 2 1 -0.342

41 2 1 -0.715

42 2 2 0.888

43 2 2 0.793

44 2 2 0.911

45 2 1 -0.050

46 2 2 2.178

> addmargins(table(Case5$G,Zld2$class))

1 2 Sum

1 24 1 25

2 3 18 21

Sum 27 19 46

> Zld2$post##后验概率

1 2

1 0.905608977 0.0943910231

2 0.569237025 0.4307629752

3 0.980908153 0.0190918472

4 0.820808758 0.1791912420

5 0.989435734 0.0105642655

6 0.989835480 0.0101645200

7 0.747111991 0.2528880090

8 0.653080409 0.3469195908

9 0.860945375 0.1390546253

10 0.616985520 0.3830144802

11 0.698955706 0.3010442943

12 0.628563744 0.3714362558

13 0.110259488 0.8897405118

14 0.859637468 0.1403625315

15 0.732239184 0.2677608155

16 0.867687085 0.1323129147

17 0.714904470 0.2850955299

18 0.904318147 0.0956818533

19 0.664841806 0.3351581943

20 0.697624901 0.3023750989

21 0.996594251 0.0034057486

22 0.804583642 0.1954163583

23 0.970195809 0.0298041907

24 0.898470539 0.1015294615

25 0.999186327 0.0008136726

26 0.004617768 0.9953822320

27 0.035798353 0.9642016473

28 0.241368867 0.7586311334

29 0.236616984 0.7633830162

30 0.125030492 0.8749695080

31 0.084408130 0.9155918698

32 0.245407356 0.7545926440

33 0.204797404 0.7952025958

34 0.344725231 0.6552747688

35 0.109038796 0.8909612045

36 0.007717942 0.9922820585

37 0.428740788 0.5712592121

38 0.469849960 0.5301500400

39 0.018887025 0.9811129747

40 0.727657162 0.2723428384

41 0.844534191 0.1554658092

42 0.204583491 0.7954165094

43 0.235328496 0.7646715036

44 0.197478875 0.8025211246

45 0.605228988 0.3947710119

46 0.021561060 0.9784389399

此外还可以使用predict(model)$posterior提取后验概率。

>predict(ld2)$posterior

在使用lda和qda函数时注意：其假设是总体服从多元正态分布，若不满足的话则谨慎使用。