阮敬 博士

首都经济贸易大学研究生院 副院长
首都经济贸易大学统计学院 教授

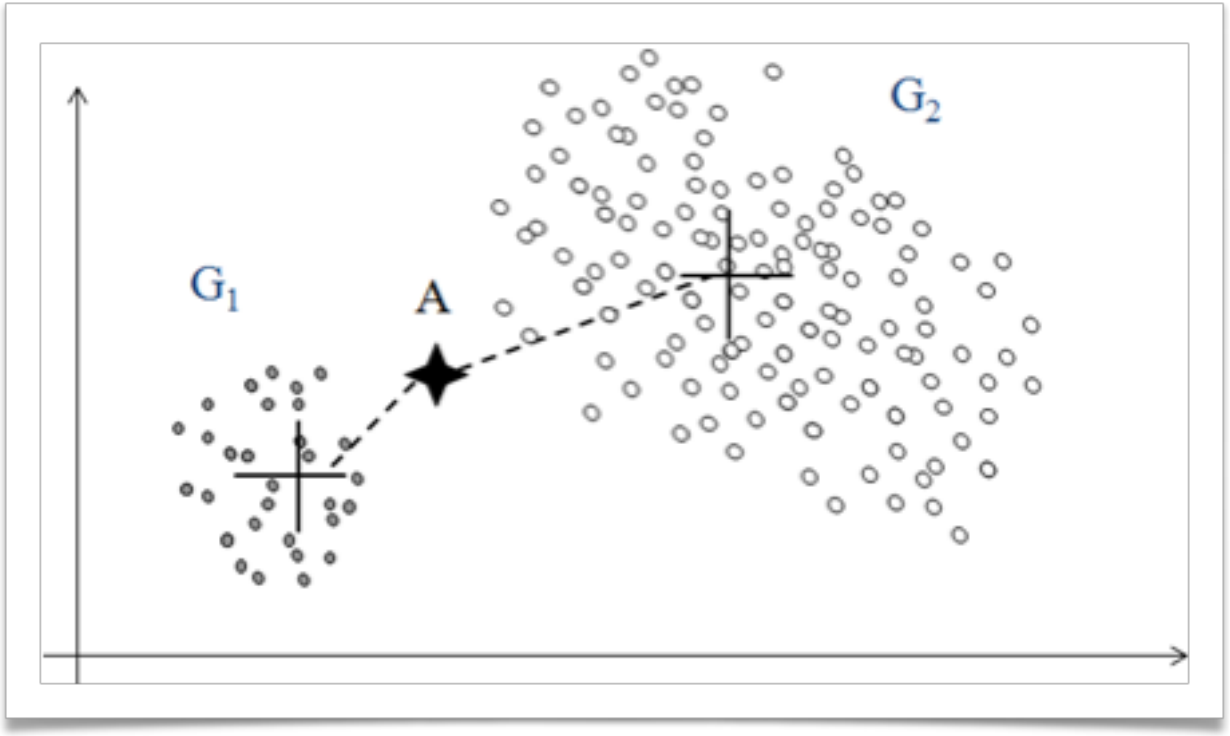
© ruanjing@msn.com

CH17 判别分析

- 现实生活中，人们不光要对现有事物分门别类，有些时候还需在已知分类的基础上对类型未确定的新样本依据特定特征进行了归类。即在给定现有分类的条件下，要求把新收集的样本，依据既定的特征，归入现有的某一个类别当中，因而有了本章所要介绍的判别分析。

判别分析的基本思想

- 如有G1和G2两个类别，对于新加入的样本A，考虑把A归入对应类别中去（如图17-1所示）。由于归类的过程涉及到对新样本与现有样本特征的判定与识别问题，才能对号入座，因此该过程也可称之为判别分析（Discriminant Analysis）。在数据挖掘分析方法中也可在高维数据下利用判别的思想进行分类分析，并称之为分类分析（Classification）。



判别分析的基本思想

- 判别分析和第16章介绍过的聚类分析有什么不同呢?

- 二者主要不同点在于：在聚类分析中一般人们事先并不知道或一定要明确应该分成几类，类别的样本组成完全根据数据特征来确定；而在判别分析则要求至少有一个已经明确知道类别的“训练样本”，利用这个样本数据的特征，就可以建立判别准则，并通过预测变量来为未知类别的观测值进行判别。

判别分析的基本思想

- 人们通常把判别分析中已经明确知道类别的样本称之为“训练样本”，判别分析的整个过程就是通过归纳和提炼训练样本的特征来进行的。如某企业对其生产某种产品的消费者购买意愿进行调查，经过调查研究，有101个被调查到的消费者被划分为“潜在顾客”，另外有32个被调查到的消费者被划分为“非潜在顾客”。研究者希望从这些被调查到的消费者特征出发，从中找出一个分类标准，对那些还没有进行归类的消费者进行定位。而研究者所依据的这些被调查到的133个消费者数据就是一个“训练样本”。而那些没有进行归类的消费者数据或新样本数据可以看作是“测试样本”或“待判样本”。判别分析就是根据从训练样本中所归纳或总结出来的判别规则，对测试样本进行归类。

判别分析的步骤和过程

- 判别分析的基本思路就是根据从不同总体（设有*G1*，*G2*，…，*Gi*个总体）中随机抽取出来的训练样本，在分析训练样本特征的基础之上，然后建立一定的判别法则，根据新的样本特征和判别法则去判别新样本应该来自于哪一个总体。

- 在判别分析过程中，建立判别法则是尤为重要的步骤，也是判别分析的核心所在。根据不同的方法，可以建立不同的判别法则。如果已知或假定总体服从一定的分布（如多元正态分布），则可以使用参数判别规则；反之则可以采用非参数判别规则。

判别分析的步骤和过程

- SAS系统中可以用上述两种判别规则进行判别分析。
- 参数判别的基本思路具体如下：先根据协方差矩阵计算新样本点到各类中心的距离，并且依据广义距离的大小，把新样本点归入距离最近的一类；或先计算新样本点属于各类的后验概率，然后把新样本归入后验概率最大的一类。
- 而非参数方法以后验概率为依据进行判别，与参数判别规则不同的是其使用核估计或最近邻估计概率密度，这两种估计也需要定义距离。而后验概率通常也可以用距离来表示。
- 与聚类分析一样，判别规则中的距离同样可以选取不同定义的距离，如欧氏距离、马氏距离或相似系数等。判别规则所依据的最简单原则就是新样本点离哪一个类别的距离最近，那么它就属于哪一类。
- 除了上述主要两种判别规则和方法之外，SAS系统中还可以使用典型判别法、逐步判别法等多种方法进行判别分析。

距离判别

- 故名思意，距离判别的基本思想是：待判样本和哪个总体距离最近，就判它属于哪个总体。由于所有的类别已知，所以可求得每个类的中心。这样只要定义了如何计算距离，就可得到任何给定的点到类型中心的距离。这种根据距离远近判别的方法，原理简单，直观易懂。因此，距离判别也称为直观判别法。

- 通常情况下，距离判别过程一般采用马氏距离。马氏距离是样本点x到类中心μ_i的一种相对距离。该距离由印度数学家Mahalanobis于1936年依据协方差矩阵V提出，其计算公式为：

$$d_i^2 = [x - \mu_i]^T V_i^{-1} [x - \mu_i]$$

- 马氏距离不受总体空间大小的影响，也不受计量单位的影响，它反映了被判定样本按平均水平计算，到中心的相对距离（该距离以方差为单位），实质上是标准化变量的欧氏距离。

- 在距离判别中，把用来比较各样本点到各类中心距离的数学函数称为判别函数。通常情况下，用线性判别函数进行判别分析非常直观，使用起来最方便，在实际中的应用广泛。

BAYES 判别

- 距离判别虽然简单直观，很实用，但是在该种方法中，没有考虑到每个分类的观察值不同时，每类出现的机会是不同，也没有考虑误判之后所造成损失的差异。Bayes判别可以克服上述缺点，其判别效果更加理想，应用也更广泛。

- 把对每个样本可能属于某个总体（类别）的可能估计值称之为“先验概率”（Prior Probability），把其记为$P(G_i)$。先验概率的值可以从经验中得出，也可使用每组样本占全部样本的百分比来估计。

- 每个样本可以根据判别函数计算出得分，在属于类别G_i条件下判别得分S的条件概率为$P(S/G_i)$；把样本根据判别函数得分而判为某个类别G_i的概率称之为“后验概率”（Post Probability），则根据贝叶斯公式可以计算出后验概率为：

$$P(G_i/S)=\frac{P(S/G_i)P(G_i)}{\sum P(S/G_i)P(G_i)}$$

BAYES 判别

- 则可以依据每个样本被判入某个类别的后验概率进行归类。
- 因而Bayes判别的基本思路是：对每个样本，首先计算出判别函数得分，然后根据先验概率$P(G_i)$和判别得分S的条件概率$P(S/G_i)$ ，计算出该样本被判为每一类的后验概率$P(G_i/S)$ ，被判入哪类的后验概率最大，则把该样本判为哪一类。

BAYES 判别

- DISCRIM过程可以依据上述方法进行距离判别、贝叶斯判别以及最邻近法和核密度法等参数和非参数判别分析，其主要语法为：

  PROCDISCRIM<选项> ;
  CLASS变量;
  BY变量列表;
  FREQ变量;
  ID变量;
  PRIORS概率列表;
  TESTCLASS变量;
  TESTFREQ变量;
  TESTID变量;
  VAR变量列表;
  WEIGHT变量;

BAYES 判别

- CLASS语句的主要作用是指定分类的标注标量，即存储了类别信息的变量；PRIORS语句主要用于指定类别之间先验概率的关系，也可指定每个类别的先验概率，默认情况下是所有类别先验概率相等；TESTCLASS、TESTFREQ、TESTID语句主要是用DISCRIM过程选项中的关键字TESTDATA所指定数据集中的变量来对观测值进行判别；VAR语句用于指定进行判别分析的变量；BY、FREQ、ID、WEIGHT语句与前面章节介绍的功能相同。

BAYES 判别

- DISCRIM的过程选项有四十多项，本书结合本章内容介绍如下常用选项：
  - DATA=：指定用于判别分析的数据集或训练样本数据集；
  - TESTDATA=：指定存储有测试样本的数据集；
  - CANONICAL：进行典型判别分析；
  - DISTANCE：显示马氏距离平方、F统计量及其他相关概率的分析结果；
  - METHOD=：指定分类方法。NORMAL关键字表示依据参数判别方法使用线性判别函数判别（为默认选项）；NPAR关键字表示使用非参数方法判别；
  - POOL=：该选项有YES、NO和TEST等3个关键字。POOL=YES表示使用混合协方差矩阵（即所有类别合并计算的协方差）计算平方距离，并计算线性判别函数；POOL=NO表示使用各类间协方差矩阵计算距离，并计算二次判别函数；POOL=TEST表示使用Bartlett's类内协方差矩阵同质性的极大似然比检验修正；系统默认选择YES关键字；

BAYES 判别

- LIST：显示对每个训练样本进行判别分析的后验概率、类别等分类结果；
- TESTLIST：显示对每个测试样本进行判别分析的后验概率、类别等分类结果；
- OUT=：指定存储分类结果的数据集；
- OUTSTAT=：指定存储协方差矩阵等统计量的数据集；
- CROSSVALIDATE：指定进行交叉核实验证；
- R=：在采用核估计判别方法时指定核估计半径；
- K=：在使用近邻估计判别方法时指定近邻的个数；
- TESTOUT=：指定存储测试样本分类结果的数据集。

BAYES 判别

- 在例16-3中，本书已经利用聚类分析方法对各种游戏鼠标依据多个指标进行了分类。现假定这13个鼠标的样本来自于已有类别的总体（即已知具体鼠标类别的训练样本，本例用A、B、C表示类别）。现又有两款鼠标的评测数据如图17-2所示，试利用判别分析的方法把两款鼠标归入对应的类别。

Dataset of SASUSER.MOUSE_DISCRIM

Obs	Brand 品牌型号	Touch 外观及手感	Chips 芯片及微动	Driver 功能及驱动	Compatibility 兼容性	Game 游戏性	Type 类别
1	Brand1	7.5	17.5	7	8	8	A
2	Brand2	7.5	19.5	7	7	9	B
3	Brand3	8.5	18	8.5	8	9.5	B
4	Brand4	9	18.5	8.5	8	9.5	B
5	Brand5	7	14	6.5	7	7.5	C
6	Brand6	7	16	6.5	7.5	8	C
7	Brand7	7.5	17	8	7.5	8	A
8	Brand8	8	17.5	8.5	7.5	8.5	A
9	Brand9	7	16.5	6	8	7	C
10	Brand10	7.5	17	7.5	8.5	8	A
11	Brand11	8	16	6.5	7	7	C
12	Brand12	7	15.5	6	8	7	C
13	Brand13	7.5	17	8	7	7	A
14	Brand14	7	16.5	6	7.5	7	
15	Brand15	7.5	18	7.5	7.5	8.5	

BAYES 判别

- 在图17-2中，利用变量TYPE对各个鼠标所归属的类别进行标记。而最后两支鼠标在变量TYPE的值为缺失，是待进行判别的对象。

- 在SAS系统中，对于判别分析的原始数据可有两种数据预处理方式。

- 第一种方式是把已经分好类别的训练样本和未分类的样本放在同一个数据集之中，并且用一个分类变量来标注各个样本所属的类别，如本例中的变量TYPE便是标记有类别的分类变量；而未进行归类或待判样本在该分类变量的数值用缺失值表示。图17-2所示的数据就是采用该种方式进行数据预处理。

BAYES 判别

- 第二种方式是把训练样本数据和测试样本数据分别存储为两个数据集，这两个数据集当中作为判别依据的变量的变量名不一定相同。如本例把存储训练样本的数据集命名为Mouse_D1；把测试样本数据存储为Mouse_D2，如图17-3和图17-4所示。

Dataset of SASUSER.MOUSE_D1

Obs	Brand 品牌型号	Touch 外观及手感	Chips 芯片及微动	Driver 功能及驱动	Compatibility 兼容性	Game 游戏性	Type 类别
1	Brand1	7.5	17.5	7	8	8	A
2	Brand2	7.5	19.5	7	7	9	B
3	Brand3	8.5	10	8.5	8	9.5	D
4	Brand4	9	18.5	8.5	8	9.5	D
5	Brand5	7	14	6.5	7	7.5	C
6	Brand6	7	16	6.5	7.5	8	C
7	Brand7	7.5	17	8	7.5	8	A
8	Brand8	8	17.5	8.5	7.5	8.5	A
9	Brand9	7	16.5	6	8	7	C
10	Brand10	7.5	17	7.5	8.5	8	A
11	Brand11	8	16	6.5	7	7	C
12	Brand12	7	15.5	6	8	7	C
13	Brand13	7.5	17	8	7	7	A

BAYES 判别

- 在DISCRIM过程中，当CLASS语句指定的分类变量有缺失值时，该缺失值对应的样本会被系统自动排除，不参加判别规则的制定。如果样本中只有CLASS语句指定的分类变量存在缺失值，其他变量没有缺失值时，则该样本会自动依据判别规则进行归类并在归类结果中显示出来。这也是第一种数据预处理方法的理论依据。

Dataset of SASUSER.MOUSE_D2

Obs	Brand 品牌型号	Touch 外观及手感	Chips 芯片及微动	Driver 功能及驱动	Compatibility 兼容性	Game 游戏性	Type 类别
1	Brand14	/	16.5	6	7.5	/	
2	Brand15	7.5	18	7.5	7.5	8.5	

BAYES 判别

- 本例如果使用第一种数据处理方式的程序如下：

```
proc discrim data=sasuser.mouse_discrim
        List    /*显示对每个训练样本进行判别分析的后验概率、类别等分类结果*/
        out=mouse_discrim_out    /*把判别结果存储在指定数据集当中*/
        Distance    /*显示用于判别的距离*/
        pool=yes;    /*使用混合协方差矩阵计算线性判别函数*/
    class type;    /*指定分类变量*/
    var touch chips driver compatibility game;    /*指定用于判别所依据的变量*/
    id brand;    /*指定标注样本名称的变量*/
run;
```

BAYES 判别

- 程序运行之后可得到非常多的输出结果。首先是判别分析的基本情况，因为本例没有指定样本归类的先验概率，因此系统自动假定归入每一个类别的先验概率相等（本例每类的先验概率为0.3333），如图17-5所示。

Class Level Information

Type	Variable Name	Frequency	Weight	Proportion	Prior Probability
A	A	5	5.0000	0.384615	0.333333
B	B	3	3.0000	0.230769	0.333333
C	C	5	5.0000	0.384615	0.333333

BAYES 判别

- 接下来系统给出了每个类别两两配对的马氏距离平方，以及依据马氏距离测度的类别之间差异性是否显著的*F*统计量值及其对应的*P*值，如图17-6所示。

The DISCRIM Procedure

Squared Distance to Type

From Type	A	B	C
A	0	20.82822	29.42673
B	20.82822	0	59.65702
C	29.42673	59.65702	0

F Statistics, NDF=5, DDF=6 for Squared Distance to Type

From Type	A	B	C
A	0	4.68635	8.82802
B	4.68635	0	13.42283
C	8.82802	13.42283	0

Prob > Mahalanobis Distance for Squared Distance to Type

From Type	A	B	C
A	1.0000	0.0434	0.0098
B	0.0434	1.0000	0.0033
C	0.0098	0.0033	1.0000

BAYES 判别

- 图17-6中的Squared Distance to Type表格列示了本例中鼠标A、B、C 3个类别之间的距离平方，同时也给出了各类距离差异检验的*F*统计量值，Prob>Mahalanobis Distance for Squared Distance to Type表格给出了对应*F*统计量的*P*值，各组差异在显著性水平α=0.05的条件下可以通过显著性检验，即可以认为这3个类别之间是有显著差异的，在此基础上进行归类才有意义。

- DISCRIM过程的输出结果还会给出用于距离判别的判别函数，如图17-7所示。

Linear Discriminant Function for Type

Variable	Label	A	B	C
Constant		-788.16350	-903.03606	-626.22517
Touch	外观及手感	-7.81528	-1.50078	1.39421
Chips	芯片及微动	51.34464	54.06714	44.29113
Driver	功能及驱动	54.93436	50.29628	40.25220
Compatibility	兼容性	24.40398	21.03482	21.75651
Game	游戏性	17.24008	26.32372	18.45311

BAYES 判别

- 根据图17-7所示的结果，可以建立如下线性判别函数：

$$S_A = -788.16350 - 7.81528 \times Touch + 51.34464 \times Chips + 54.93436 \times Driver + 24.40398 \times Compatibility + 17.24008 \times Game$$

$$S_B = 903.03606 \quad 1.50078 \times Touch \mid 54.06714 \times Chips \mid 50.29628 \times Driver + 21.03482 \times Compatibility + 26.32372 \times Game$$

$$S_C = 626.22517 \mid 1.39421 \times Touch \mid 44.29113 \times Chips \mid 40.25220 \times Driver + 21.75651 \times Compatibility + 18.45311 \times Game$$

- 根据判别函数及各样本指标值，可计算各样本在各类的判别函数得分，并依据判别函数得分把测试样本归入其对应的类别中去。即测试样本在哪个判别函数得分最高，就把该样本归入哪一类中去。

BAYES 判别

- 如对于品牌为Brand14的待判样本，依据判别函数计算得分如下：

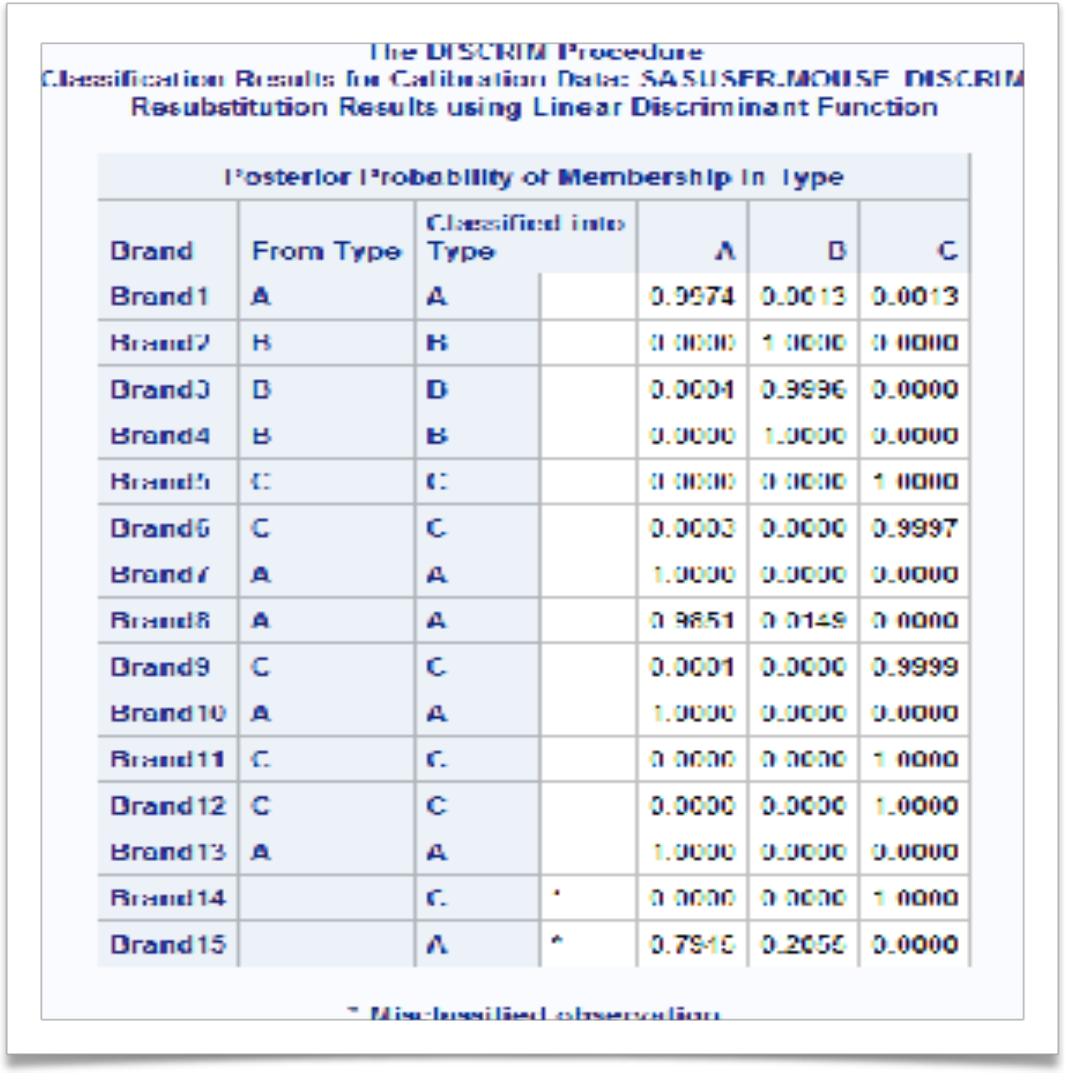
$$
\begin{aligned}
S_A &= -788.16350 - 7.81528\times 7 + 51.34464\times 16.5 + 54.93436\times 6 \\
&\quad + 21.40398\times 7.5 + 17.24008\times 7 \\
&= 637.6327 \\
S_B &= 903.03606 \quad 1.50078\times 7 \mid 54.06714\times 16.5 \mid 50.29628\times 6 \\
&\quad + 21.03482\times 7.5 + 26.32372\times 7 \\
&= 622.3721 \\
S_C &= -626.22517 + 1.39421\times 7 + 44.29113\times 16.5 + 40.25220\times 6 \\
&\quad + 21.75651\times 7.5 - 18.45311\times 7 \\
&= 648.1967
\end{aligned}
$$

- 因为$S_c>S_a>S_b$，故把该样本归入C类；同理对于品牌为Brand15的鼠标，计算判别函数得分分别为$(S_a=819.0037)>(S_b=817.6515)>(S_c=803.3885)$，故可把该样本归入A类。> =

BAYES 判别

- 此外，根据得分和先验概率，可以计算出每个样本归入每一类的后验概率，输出结果如图17-9所示。

- 图17-8所示结果详细列示出了每个样本的原始类别（From Type）、进行判别分析之后应该归入的类别（Classified into Type）以及每个样本归入每个类别的后验概率（本例训练样本有A、B、C 3大类，每个类别的标记对应的列即归入该类的后验概率）。



The DISCRIM Procedure
Classification Results for Calibration Data: SASUSER.MONUSE DISCRIM
Resubstitution Results using Linear Discriminant Function

		Posterior Probability of Membership in Type			
Brand	From Type	Classified into Type		A	B	C
Brand1	A	A		0.9974	0.0013	0.0013
Brand2	B	B		0.0000	1.0000	0.0000
Brand3	B	B		0.0004	0.9996	0.0000
Brand4	B	B		0.0000	1.0000	0.0000
Brand5	C	C		0.0000	0.0000	1.0000
Brand6	C	C		0.0003	0.0000	0.9997
Brand7	A	A		1.0000	0.0000	0.0000
Brand8	A	A		0.9851	0.0149	0.0000
Brand9	C	C		0.0001	0.0000	0.9999
Brand10	A	A		1.0000	0.0000	0.0000
Brand11	C	C		0.0000	0.0000	1.0000
Brand12	C	C		0.0000	0.0000	1.0000
Brand13	A	A		1.0000	0.0000	0.0000
Brand14		C	*	0.0000	0.0000	1.0000
Brand15		A	*	0.7946	0.2054	0.0000

* Misclassified observation

BAYES 判别

- DISCRIM过程对于事先没有进行分类的对象即测试样本用“*”标注（同时“*”也用于标注经过判别后归类与原有类别不一致的情况），即图17-8中最后两个样本Brand14和Brand15。依据判别函数得分以及后验概率的大小，Brand14归入C类的后验概率为1，故把其归入C类；而Brand15归入A类的后验概率最大，为0.7945，故把其归入A类。

- 对每个样本的判别结果及计算出来的后验概率，都会保存在DISCRIM语句选项OUT关键字指定的数据集中，读者可在临时数据库中查看MOUSE_DISCRIM_OUT数据集的结果，其内容与图17-8所列示的结果类似。

- 在对事先没有进行分类的样本进行判别的同时，DISCRIM过程还对原有训练样本中的每个样本进行判别，除了在图17-8中列示出进行判别之后的归类结果之外，系统还给出了判别分析的交叉核实过程及错判概率，如图17-9所示。

The DISCRIM Procedure
Classification Summary for Calibration Data: SASUSER.MOUSE_DISCRIM
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Type

From Type	A	B	C	Total
	1	0	1	2
	50.00	0.00	50.00	100.00
A	5	0	0	5
	100.00	0.00	0.00	100.00
B	0	3	0	3
	0.00	100.00	0.00	100.00
C	0	0	5	5
	0.00	0.00	100.00	100.00
Total	6	3	6	15
	40.00	20.00	40.00	100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Type

	A	B	C	Total
Rate	0.0000	0.0000	0.0000	0.0000
Priors	0.3333	0.3333	0.3333	

BAYES 判别

- 图17-9中的Number of Observations and Percent Classified into Type表格即为交叉核实表（该表也可通过在DISCRIM的过程选项中使用关键字CROSSVALIDATE得到）。其中From Type列表示样本的原有类别，而列A、B、C表示把样本进行判定的类别，即进行判别分析之后的现有类别。

- 在图17-9所示结果中，原有类别为空白（即缺失值）的一共有2个样本，分别归入A类和C类各一个样本；原有类别为A的样本一共有5个，经过判别分析之后，归入A类5个样本；依次类推，原有类别为B的3个样本经过判别后全部归入B类；原有类别为C的5个样本也全部归入到C类中。总结归类的全过程，一共有6个样本归入了A类，占总样本的40%；3个样本归入B类，占总样本的20%，6个样本归入了C类，占40%。

BAYES 判别

- 对于上述判别分析的交叉核实过程，系统自动给出了每个类别的错判概率。简单错判概率由如下公式计算：

$$p = \frac{1}{n} \sum_{i=i}^{k} \sum_{\substack{j=1 \\ j \neq i}}^{k} m_{ij}$$

- 如果考虑先验概率p，则错判概率可由如下加权形式计算：

$$P = \sum_{i=1}^{k} q_i p_i$$

BAYES 判别

- 从图17-9的Error Count Estimates for Type表格可以看出，本例每一个类别的错判概率均为0，总错判概率为0，表明本例所考察的鼠标分类是正确的，且分类精度和可靠性非常高。

- 本例数据如果采用第二种方式进行判别，即把训练样本和测试样本分为两个数据集进行分析，利用DISCRIM过程的分析程序如下：

```
proc discrim data=sasuser.mouse_d1 testdata=sasuser.mouse_d2testlist testout=mouse_d2_out;
    class type;
    testclass type;
    testid brand;
    var touch chips driver compatibility game;
run;
```

BAYES 判别

- 在第二种数据预处理方式下，要注意在DISCRIM的过程选项中，用关键字DATA指定训练样本数据集，用关键字TESTDTAT指定待判样本数据集，关键字TESTLIST用于列示待判样本的判别结果，关键字TESTOUT用于指定存储待判样本输出结果的数据集。此外还要在该过程中增加指定测试样本类别分类变量的语句TESTCLASS，同时可用TESTID指定标注测试样本名称的变量。

- 运行上述程序之后，可以得到与第一种数据预处理方式相同的结果。

- 前面讨论的距离判别需要估计总体的参数，估计总体参数的前提是已知总体服从什么样的分布；而进行Bayes判别时，也应假定总体服从正态分布。一般情况下，各类总体分布是否就是正态分布往往是未知的，当总体分布未知时，可以使用非参数判别法。通常可用核方法和近邻方法进行非参数判别分析。

BAYES 判别

- 非参数判别仍然使用Bayes 后验概率作为判别的依据，设有*n* 个类别，由于各类总体分布未知，故每个类别具有的概率密度函数$f_n(x)$未知，于是可对$f_n(x)$利用核方法或近邻方法进行估计，将估计的先验概率和密度函数结果再代入判别规则中计算后验概率，然后再按照前面介绍过的方法进行归类。

- SAS 系统中的DISCRIM 过程也可进行非参数判别，通过DISCRIM 选项中的METHOD=NPAR 关键字来指定该过程进行非参数判别。选用非参数判别方法时，还应指定R=核估计半径来指定系统用核估计方法进行估计，或者指定K=近邻个数来指定系统使用近邻估计方法进行估计。

非参数判别

- DISCRIM过程还可以进行非参数判别分析，主要有两种非参数方法。
- 1、核方法判别
- 2、*k*最邻近法判别