7.3 判别分析

2014年6月25日

判别分析问题的一般形式

已知研究对象类别的若干样本数据。要求根据这些样本观测数据，建立一种判别方法，对一个未知类别的样品能够判定它属于哪一类，这种判定的方法就称为判别分析（**Discriminant Analysis**）

例 1 岩石分类

从某矿床取得14块已知是铀矿石的样品和14块已知是围岩的样品，分别测定其中7种成分的含量，取得了一批观测数据：

已知类别	样品编号	Pb	Zn	Mo	Cu	$CaO+MgO$	Al_2O_3	SiO_2
铀矿石	1	0.0049	0.488	0.22	0.0098	4.07	13.97	61.62
	2	0.0030	0.114	0.07	0.0077	1.51	11.47	69.69
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	14	0.0029	0.232	0.03	0.0090	1.83	13.35	66.96
围岩	15	0.0023	0.0134	0.14	0.0065	1.39	11.88	73.58
	16	0.0019	0.0099	0.10	0.0082	1.53	13.89	65.93
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	28	0.0014	0.0146	0.03	0.0073	1.87	12.79	65.85

要求建立一种判别方法，当我们从这个矿床取得一个新的岩石样品时，可以通过测定这个样品中7种成分的含量，判定它是铀矿石还是围岩

例 疾病诊断

对求诊者作心电图检测，测得5个心电图指标:X_1, X_2, X_3, X_4, X_5，要求诊断这个求诊者是属于正常，还是患主动脉硬化，还是患冠心病。已经对11个正常人，7个主动脉硬化病人，5个冠心病人测得数据如下：

已知类别	已知病例个数	编号	X_1	X_2	X_3	X_4	X_5
正常人	11	1	8.11	261.01	13.23	5.46	7.36
		2	9.36	185.39	9.02	5.66	5.99
		⋮	⋮	⋮	⋮	⋮	⋮
		11	8.06	231.03	14.41	5.72	6.15
主动脉硬化病人	7	12	6.80	308.90	15.11	5.52	8.49
		13	8.68	258.69	14.02	4.79	7.16
		⋮	⋮	⋮	⋮	⋮	⋮
		18	9.89	409.42	19.47	5.19	10.47
冠心病人	5	19	5.22	330.34	18.19	4.96	9.61
		20	4.71	331.47	21.26	4.30	13.72
		⋮	⋮	⋮	⋮	⋮	⋮
		23	8.27	189.56	12.74	5.46	6.94

要求建立一种诊断方法，可以通过对一个新来的求诊者的心电图检测，诊断出这个求诊者是属于正常，还是患主动脉硬化，还是患冠心病

判别分析问题的一般形式：

设有个已知的类别:$G_1,G_2,\cdots;G_p$,对各个类别分别取样,共得到n个样品，已知其中有n_1个属于G_1，n_2个属于G_2，…，n_p个属于G_p。对每一个样品进行观测检验，得到m个变量$X_1,X_2,\cdots;X_m$的观测值x_{ij}，$i=1,2,\cdots;n,j=1,2,\cdots;m$:

已知类别	样品个数	样品编号	变量 X_1	变量 X_2	…	变量 X_m
G_1	n_1	1	x_{11}	x_{12}	…	x_{1m}
		2	x_{21}	x_{22}	…	x_{2m}
		⋮	⋮	⋮		⋮
⋮	⋮	⋮	⋮	⋮		⋮
G_p	n_p	⋮	⋮	⋮		⋮
		$n-1$	$x_{n-1,1}$	$x_{n-1,2}$	…	$x_{n-1,m}$
		n	x_{n1}	x_{n2}	…	x_{nm}

要求建立一种判别方法，可以对一个未知类别的样品,通过对其m 个指标变量的观测结果，判定它属于哪一类

常用判别分析的方法

1) 距离判别

设有一个要判别类型的样品， $x=\begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$ 是对这个样品的m个指标变量

$X_1, X_2, \cdots, X_m$ 测得的观测值， 定义一种从样品 x 到第 k 个类 G_k

的距离 $d(x, G_k)$ $k=1,2,\cdots,p$

判别方法:

样品 x 到哪个类的距离最近 ,就判定 x 属于哪个类

注： 样品到类的距离是指样品与类的"中心"之间的距离. 距离判别中的距离函数的定义,根据具体需要可以是欧式距离,闵氏距离,马氏距离，也可以是其他的距离

样品到类的闵氏(MINKOWSKI)距离

$d(x, G_k)$为样品 x 与 G_k的中心 $\overline{x}_k$（即 G_k类的样品的均值）分量差绝对值的 q次方之和 再开 q次方

特别地，$q=2$时 $d(x, G_k)$为欧氏距离：

$$d(x, G_k) = \sqrt{(x-\overline{x}_k)^T (x-\overline{x}_k)}$$

样品到类的马氏(Mahalanobis)距离

$$d(x, G_k) = \sqrt{(x - \bar{x}_k)^T S_k^{-1} (x - \bar{x}_k)}$$

S_k 是已知属于 G_k 的样品的样本协方差矩阵

$$S = \frac{1}{n-1} \boldsymbol{A} = [s_{ij}]_{m \times m}$$

其中: $s_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

2) Fisher（费希尔）判别

*Fisher*判别的基本思想是：在空间作一条方向为a的直线，把待判样品x和各类的样本均值$\overline{x}_1, \overline{x}_1, \cdots, \overline{x}_p$都投影到这条直线上，得到投影$y$ 和$\overline{y}_1, \overline{y}_1, \cdots, \overline{y}_p$ 。看投影之间的距离，y 到哪一个$\overline{y}_k$ 的距离最近，就将样品判别为哪一类。

设$X=\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix}$是观测值数据矩阵,$H=I_n-\frac{1}{n}\mathbf{1}\mathbf{1}^T$,$H_k=I_{n_k}-\frac{1}{n_k}\mathbf{1}\mathbf{1}^T$,

$k=1,2,\cdots,p$。$C=\begin{bmatrix} H_1 & & \\ & \ddots & \\ & & H_p \end{bmatrix}$是对角块为$H_1,H_2,\cdots,H_p$的矩阵。

$a=\begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix}$是矩阵$(X^THX)^{-1}X^T(H-C)X$的最大特征值$\lambda_1$对应的特征向量。

可以证明，按上述方法求出的投影方向a，

从某种意义上说，是能够最好地将各类别区分开来的方向。

设要判别类型的样品观测值为$x=\begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$，计算下列判别函数值

$$y=a^T x=a_1x_1+a_2x_2+\cdots+a_mx_m。$$

对类G_k的样本均值$\overline{x}_k=\begin{bmatrix} \overline{x}_{1k} \\ \vdots \\ \overline{x}_{mk} \end{bmatrix}$，也计算判别函数值

$$\overline{y}_k=a^T\overline{x}_k=a_1\overline{x}_{1k}+a_2\overline{x}_{2k}+\cdots+a_m\overline{x}_{mk},\ \ k=1,2,\cdots,p。$$

比较距离$|y-\overline{y}_1|,|y-\overline{y}_2|,\cdots,|y-\overline{y}_p|$的大小，到哪一类的距离最近，就将这个样品判别为哪一类。

3) 回归判别

基本思想是根据样本数据对每个类建立回归方程,把待判样品的观测值代入各回归方程求回归函数值,然后,根据这些回归函数值来判定样品的归类

把类别已知的样本观测值作为自变量$X_1, X_2, \cdots, X_m$ 的观测值。对每一类G_k ，人为给定一个因变量Y_k ，设它的观测值为

$$y_{ik} = \begin{cases} 1 & \text{第} i \text{个样品属于} G_k \\ 0 & \text{第} i \text{个样品不属于} G_k \end{cases}, \quad i = 1, 2, \cdots, n, \quad k = 1, 2, \cdots, p。$$

从这些数据出发，通过回归分析，对每一类G_k 建立一个线性回归方程：

$$\hat{Y}_k = \hat{\beta}_{0k} + \hat{\beta}_{1k} X_1 + \hat{\beta}_{2k} X_2 + \cdots + \hat{\beta}_{mk} X_m, k = 1, 2, \cdots, p。$$

将待判别的样品的观测值$x_1, x_2, \cdots, x_m$ 代入各个回归方程，求出因变量的估计值$\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_p$，看哪一个$\hat{y}_k$最接近1 ，就把这个样品判别为哪一类。

4) Bayes（贝叶斯）判别

若$B_1, B_2, \cdots, B_p$是一组互不相容的事件，有$P(B_k) > 0$，$k = 1, 2, \cdots, p$，事件$A \subset \sum_{k=1}^{p} B_k$，即$A$的发生总是与$B_1, B_2, \cdots, B_p$之一同时发生，则在事件$A$已经发生的条件下事件$B_k$发生的条件概率，可以由下式求出：

$$P(B_k \mid A) = \frac{P(B_k)P(A \mid B_k)}{\sum_{j=1}^{p} P(B_j)P(A \mid B_j)}, \quad k = 1, 2, \cdots, p。$$

Bayes公式常常被用来判断一个事件是什么原因引起的。设 $B_1, B_2, \cdots, B_p$ 是可能引起事件A发生的几个原因，$P(B_k)$是在事件A 发生之前，B_k就知道的 的概率，这个概率反映了在各种原因中所占的百分比大小，称为“先验概率”。$P(B_k|A)$是在事件 A发生之后，估计事件A 可能是由原因 B_k引起的概率，称为“后验概率”。

比较各个后验概率的大小，某个后验概率最大，说明 A最有可能是由这个原因引起的，某些后验概率比较小，说明A 不太可能是由这些原因引起的。这样，就可以比较有理由地对事件发生的原因作出判断。

设 $x_1, x_2, \cdots, x_m$ 是对样品的 m 个变量 $X_1, X_2, \cdots, X_m$ 测得的观测值

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

，。样品取值为 x ，相当于Bayes公式中的事件 A。

G_k B_k

样品属于类别 ，相当于Bayes公式中的事件 。

设从自然界任取一个样品，这个样品恰好属于类别 G_k 的概率为 π_k ，$k = 1, 2, \cdots, p$。称 π_k 为“先验概率”，先验概率 相当于Bayes公式中的概率 $P(B_k)$

对每一个类别G_k来说，变量$X_1, X_2, \cdots, X_m$ 的值，是一个服从多元分布的总体ξ_k ，设它的概率密度为$\phi_k(x)$。$\phi_k(x)$反映了当样品属于类别G_k 时，样品取值为x的概率大小，在某种意义上说，$\phi_k(x)$相当于*Bayes* 公式中的条件概率$P(A|B_k)$ 。

当一个样品取值为x 时，这个样品属于类别G_k 的概率相当于*Bayes*公式中的“后验概率”$P(B_k \mid A)$。

通常可以G_k认为ξ_k的总体服从的分布是一个m 元正态分布，即有

$$\xi_k \sim N_m(\mu_k, \Sigma_k),$$

概率密度为

$$\phi_k(x) = (2\pi)^{-\frac{m}{2}} (\det \Sigma_k)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\right],$$

其中，μ_k是ξ_k的数学期望向量，Σ_k 是ξ_k 的协方差矩阵，。

类似于*Bayes*公式中的后验概率

$$P(B_k \mid A) = \frac{P(B_k)P(A \mid B_k)}{\sum_{j=1}^{p} P(B_j)P(A \mid B_j)}, \quad k = 1, 2, \cdots, p,$$

可以证明

$$P(G_k \mid x) = \frac{\pi_k \phi_k(x)}{\sum_{j=1}^{p} \pi_j \phi_j(x)}, \quad k = 1, 2, \cdots, p$$

$P(G_k \mid x)$就是当一个样品取值为x时，这个样品属于类别G_k 的后验概率

作判别分析，只要比较这种后验概率的大小就可以了，哪一个后验概率最大，说明样品最有可能属于这一类别，就将样品判别为这一类。

这种做法，还可以从减少“错判损失”的角度来说明。

当一个样品实际上是属于G_i 类，如果将它错误地判别分类到另一类G_k ，就会有一个损失，可以定义一个“ 错判损失”函数：

$$L(i,k)=\begin{cases}0 & \text{当}i=k\text{时}\\ 1 & \text{当}i\neq k\text{时}\end{cases}$$

一个分类未知、取值为x 的样品，它的? 平均错判损失”就是$L(i,k)$的条件数学期望：

$$E(L(i,k)\big|x)=\sum_{i=1}^{p}L(i,k)P(G_i\big|x)=\sum_{i\neq k}P(G_i\big|x)=\sum_{i\neq k}\frac{\pi_i\phi_i(x)}{\sum_{j=1}^{p}\pi_j\phi_j(x)}$$

$$=\frac{\sum_{i=1}^{p}\pi_i\phi_i(x)-\pi_k\phi_k(x)}{\sum_{j=1}^{p}\pi_j\phi_j(x)}=1-\frac{\pi_k\phi_k(x)}{\sum_{j=1}^{p}\pi_j\phi_j(x)}$$

要使得平均错判损失最小，也就是要使得$\dfrac{\pi_k\phi_k(x)}{\sum_{j=1}^{p}\pi_j\phi_j(x)}$最大。

Bayes（贝叶斯）判别的基本思想：

对各个类别$G_1, G_2, \cdots, G_p$，分别写出判别函数

$$P(G_k \mid x) = \frac{\pi_k \phi_k(x)}{\sum_{j=1}^{p} \pi_j \phi_j(x)}，\ k = 1, 2, \cdots, p。$$

将要判别类型的样品观测值$x_1, x_2, \cdots, x_m$代入

$P(G_1 \mid x), P(G_2 \mid x), \cdots, P(G_p \mid x)$，算出后验概率，

比较后验概率的大小，哪一个后验概率最大，就判别样品为哪一类。

先验概率π_k，即从自然界任取一个样品，这个样品恰好属于类别G_k的概率，可以有下列几种取法：

⑴ 认为各个类别的先验概率相等，即取$\pi_k = \dfrac{1}{p}$ ，$k = 1, 2, \cdots, p$ 。

（2）认为先验概率π_k与已知属于这一类的样品个数n_k成正比，即取

$$\pi_k = \frac{n_k}{n}, \quad k = 1, 2, \cdots, p$$

（3）人为地给定先验概率π_k的值。

概率密度的计算

G_k类的总体ξ_k的概率密度为

$$\phi_k(x)=(2\pi)^{-\frac{m}{2}}(\det\Sigma_k)^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right]$$

其中，μ_k是ξ_k的数学期望向量，可以用已知属于G_k 类的样品的样本均值向量

$\overline{x}_k=\begin{bmatrix}\overline{x}_{1k}\\ \vdots \\ \overline{x}_{mk}\end{bmatrix}$作为它的估计值。

Σ_k是ξ_k的协方差矩阵，可以用样本协方差矩阵作为它的估计值。

⑴ 各类的协方差矩阵都相同，即有$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_p$。

可以将各个类别的样本观测数据合并在一起，计算一个总的样本协方差矩阵S，用它作为$\Sigma_1, \Sigma_2, \cdots, \Sigma_p$的估计值。

⑵ 各类的协方差矩阵各不相同。

必须将各个类别的样本观测数据分开，分别计算样本协方差矩阵。

设用G_k类样品算出的样本协方差矩阵为S_k，$k = 1, 2, \cdots, p$。

分别用$S_1, S_2, \cdots, S_p$作为$\Sigma_1, \Sigma_2, \cdots, \Sigma_p$的估计值。

⑶ 采用组合方法，即通过假设检验，检验各个类别的协方差矩阵是否相等，如果检验认为相等，就将这几个类别的样本观测数据合并在一起，计算一个总的样本协方差矩阵；如果检验认为不相等，就将这几个类别的样本观测数据分开，分别计算样本协方差矩阵。

衡量判别分析效果好坏的标准

衡量判别分析效果的好坏，可以将已知属于各个类别的样品数据分别代入判别函数，判别它们属于哪一类。判别的结果可以写成一个表格，称为**判别矩阵**：

		判别认为的类别			
		G_1	G_2	$\cdots$	G_p
已知所属的类别	G_1	n_{11}	n_{12}	$\cdots$	n_{1p}
	G_2	n_{21}	n_{22}	$\cdots$	n_{2p}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	G_p	n_{p1}	n_{p2}	$\cdots$	n_{pp}

如果判别分析的结果完全正确，判别矩阵中非对角线上的数字应该全部等于 0 ，非 0 数字应该全部集中在判别矩阵的对角线上

		判别认为的类别			
		G_1	G_2	$\cdots$	G_p
已知所属的类别	G_1	n_{11}	n_{12}	$\cdots$	n_{1p}
	G_2	n_{21}	n_{22}	$\cdots$	n_{2p}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	G_p	n_{p1}	n_{p2}	$\cdots$	n_{pp}

判别矩阵中，非对角线上的等于 0 的数字越多，非 0 数字越是集中在对角线上, 说明误判的情况越少，判别的效果越好。
反之，非对角线上非 0 的数字越多，说明误判越多，判别的效果越不好

例2（国际数学建模竞赛1989年A题）蠓的分类

蠓是一种昆虫，分为很多类型，其中有一种名为Af，是能传播花粉的 益虫，另一种名为Apf，是会传播疾病的害虫，这两种类型的蠓在形态上十分相似，很难区分。现在有15只类型已知的蠓的标本，其中6只是 Af 蠓，9只是 Apf 蠓，测得它们的触角长度和翅膀长度数据如下：

已知类型	标本编号	X_1 触角长度 (mm)	X_2 翅膀长度 (mm)
Af	1	1.14	1.78
	2	1.18	1.96
	3	1.20	1.86
	4	1.26	2.00
	5	1.28	2.00
	6	1.30	1.96

已知类型	标本编号	X_1 触角长度 (mm)	X_2 翅膀长度 (mm)
Apf	7	1.24	1.72
	8	1.36	1.74
	9	1.38	1.64
	10	1.38	1.82
	11	1.38	1.90
	12	1.40	1.70
	13	1.48	1.82
	14	1.54	1.82
	15	1.56	2.08

另有3只类型未知的蠓的标本，触角长度和翅膀长度分别为（1.24，1.80），（1.28，1.84），（1.40，2.04）。要求建立一种判别方法

能够根据这两种蠓的触角长度和翅膀长度，判别它属于哪一种

这两类蠓的样本均值、样本协方差矩阵分别为

$$\overline{x}_1=\begin{bmatrix}1.22667\\1.92667\end{bmatrix},S_1=\begin{bmatrix}0.00394667 & 0.00434667\\0.00434667 & 0.00778667\end{bmatrix}$$

$$\overline{x}_2=\begin{bmatrix}1.41333\\1.80444\end{bmatrix},S_2=\begin{bmatrix}0.00980000 & 0.00808333\\0.00808333 & 0.01687778\end{bmatrix}$$

设各类蠓的先验概率与已知属于各类的样品个数成正比，取

$$\pi_1=\frac{n_1}{n}=\frac{6}{15},\pi_2=\frac{n_2}{n}=\frac{9}{15}$$

采用组合方法, 对假设H_0：$\Sigma_1 = \Sigma_2$作χ^2 检验,
结论是可以认为两类总体的协方差矩阵是相同的，
所以可以将两类样品的数据合并在一起，计算出一个总的样本协方差矩阵

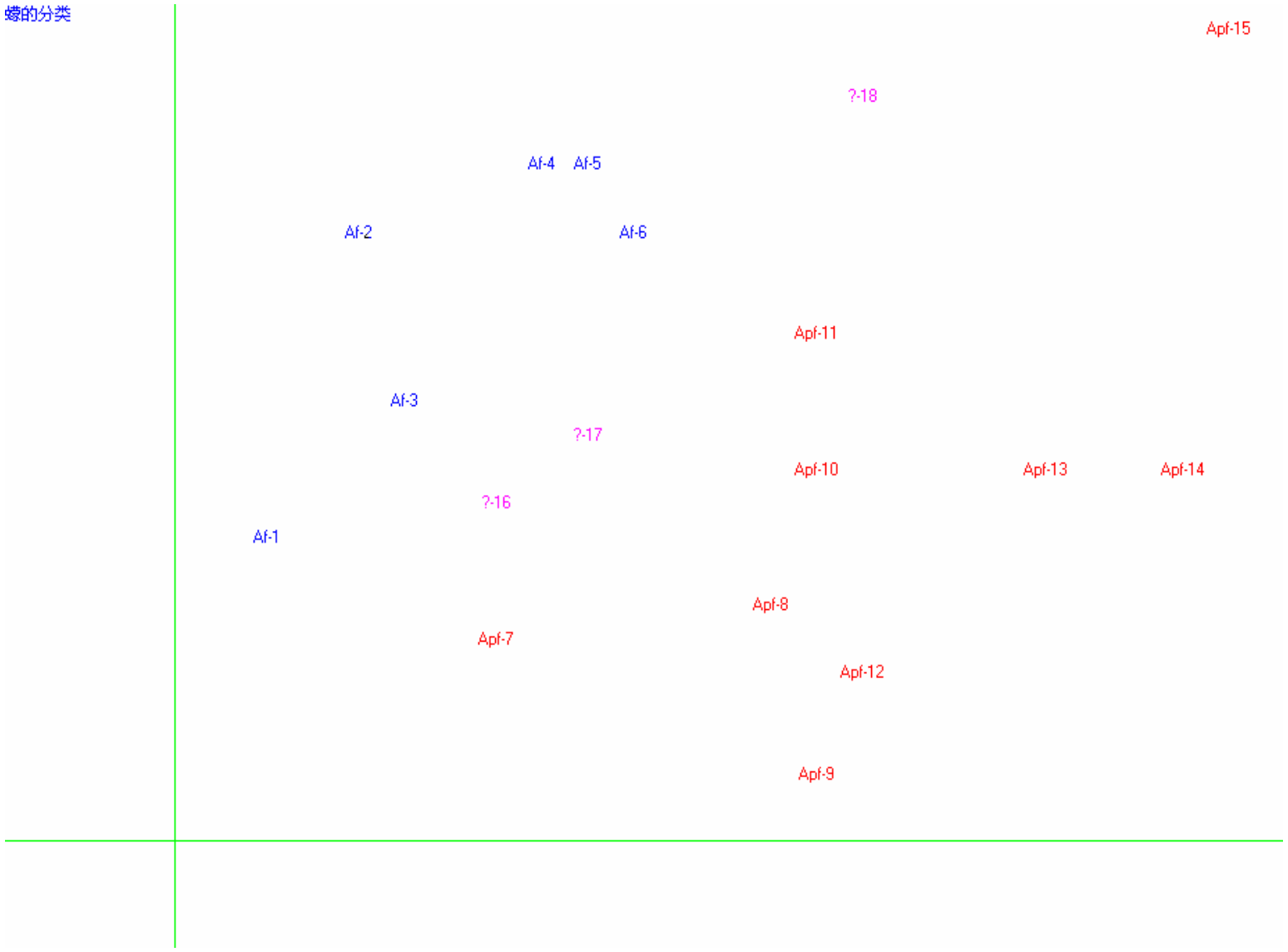
求出Bayes判别函数后，对原来15只类型已知蠓的标本进行判别，得到判别矩阵为：

		判别认为的类型	
		Af	Apf
已知所属类型	Af	6	0
	Apf	0	9

将3只类型未知的蠓的标本的数据代入判别函数，可得到判别结果如下：

标本编号	X_1触角长度 (mm)	X_2翅膀长度 (mm)	判别认为的类型	最大后验概率
16	1.24	1.80	Af	0.853021
17	1.28	1.84	Af	0.721393
18	1.40	2.04	Af	0.828459

判别结果认为这3只标本都属于Af类。


蟓的分类
Apf-15
?-18
Af-4
Af-5
Af-2
Af-6
Apf-11
Af-3
?-17
Apf-10
Apf-13
Apf-14
?-16
Af-1
Apf-8
Apf-7
Apf-12
Apf-9

例3 人文发展水平分类

1990年5月联合国开发计划署发布“人文发展指数”，认为人生有三大要素，衡量人生三大要素的指标为：“出生时的预期寿命”，“成人识字率”和“人均GDP”。“人文发展指数”就是由这3个指标综合而成。

按照人文发展指数的高低，从高发展水平国家和中等发展水平国家中，各选取5个作为样品。另选4个国家，作为待判样品，要求判别它们分别属于哪一类型。这些国家的3个指标的统计数据如下：

已知类型	国家名称	X_1 预期寿命	X_2 成人识字率	X_3 人均GDP
高人文发展水平国家	美国	76.0	99.0	5374
	日本	79.5	99.0	5359
	瑞士	78.0	99.0	5372
	阿根廷	72.1	95.9	5242
	阿联酋	73.8	77.7	5370
中等人文发展水平国家	保加利亚	71.2	93.0	4250
	古巴	73.8	94.9	3412
	巴拉圭	70.0	91.2	3390
	格鲁吉亚	72.8	99.0	2300
	南非	62.9	80.6	3799

待判别国家名称	X_1 预期寿命	X_2 成人识字率	X_3 人均GDP
中国	68.5	79.3	1950
罗马尼亚	69.9	96.9	2840
希腊	77.6	93.8	5233
哥伦比亚	69.3	90.3	5158

设两种类型国家的先验概率相等，取先验概率$\pi_1=\frac{1}{2}$ ，$\pi_2=\frac{1}{2}$ 。

认为两种类型国家的数据的协方差矩阵是相同的，将两类国家的数据合并在一起计算出一个总的样本协方差矩阵。

求出Bayes判别函数后对原来10个类型已知的国家进行判别：

		判别认为的类型	
		高人文发展水平国家	中等人文发展水平国家
已知所属类型	高人文发展水平国家	5	0
	中等人文发展水平国家	0	5

将待判定类型国家的统计数据带入判别函数，可得判别结果如下：

国家名称	X_1预期寿命	X_2成人识字率	X_3人均GDP	判别认为的类型	属于这一类的概率
中国	68.5	79.3	1950	中等人文发展水平国家	1.000000
罗马尼亚	69.9	96.9	2840	中等人文发展水平国家	1.000000
希腊	77.6	93.8	5233	高人文发展水平国家	0.999966
哥伦比亚	69.3	90.3	5158	高人文发展水平国家	0.985110

判别结果认为中国和罗马尼亚属于中等人文发展水平国家希腊和哥伦比亚则属于高人文发展水平国家