逐步判别法

《美国数学建模竞赛》

完整课程请长按下方二维码

逐步判别法

在判别问题中，当判别变量个数较多时，如果不加选择地一概采用来建立判别函数，不仅计算量大，还由于变量之间的相关性，可能使求解逆矩阵的计算精度下降，建立的判别函数不稳定。因此适当地筛选变量的问题就成为一个很重要的事情。凡具有筛选变量能力的判别分析方法就统称为逐步判别法。

逐步判别法其基本思路类似于逐步回归分析，按照变量是否重要逐步引入变量，每引入一个“最重要”的变量进入判别式，同时要考虑较早引入的变量是否由于其后的新变量的引入使之丧失了重要性变得不再显著了（例如其作用被后引入地某几个变量的组合所代替），应及时从判别式中把它剔除，直到判别式中没有不重要的变量需要剔除，剩下来的变量也没有重要的变量可引入判别式时，逐步筛选结束。也就是说每步引入或剔除变量，都作相应的统计检验，使最后的判别函数仅保留“重要”的变量。

逐步判别法的步骤：

1.计算各总体中各变量的均值和总均值以及似然统计量，规定引入变量和剔除变量的临界值$F_{进}$、$F_{出}$。

2.逐步计算，计算全部变量的判别能力，在已入选变量中考虑剔除可能存在的最不显著变量。在未选入变量中选出最大判别能力的变量，对变量作F检验通过检验则接受，否则剔除变量。直到能剔除又不能增加新变量，逐步计算结束。

3.建立判别式，使用第2步中选入的变量，用Bayes判别法建立判别式。

4.对待判样本进行判别分类。

```
data ex;
input g  x1-x3 @@;
cards;
/*数据省略*/
;
data ex1;
input x1-x3 @@;
cards;
/*数据省略*/
;
proc stepdisc data=ex method=stepwise sle=0.3 sls=0.3;
/*选择用逐步判别法，选择后验概率大于0.3，不注明时系统默认为0.15*/
class g; var x1-x3;
run;
proc discrim  data=ex testdata=ex1  /*待判别集合*/
anova manova simple list testout=ex2;
class g; var x1 x3;  /*选用x1和x3作为判别指标*/
proc print data=ex2;
run;
```

1.逐步回归选取变量：

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	x3		0.8133	34.85	0.0004	0.18667969	0.0004	0.81332031	0.0004
2	2	x1		0.2701	2.59	0.1516	0.13626590	0.0009	0.86373410	0.0009

结果显示：通过逐步判别分析选入两个变量X_1、X_3作为判别分析的指标，接下来再以变量X_1,X_3为参考变量作Bayes判别。

2. 得到判别函数：

Linear Discriminant Function for g

Variable	1	2
Constant	-322.01658	-234.95662
x1	6.13904	5.47083
x3	0.03335	0.02465

$$y_1 = -322.01658 + 6.13904x_1 + 0x_2 + 0.03335x_3$$

$$y_2 = -234.95662 + 5.47083x_1 + 0x_2 + 0.02465x_3$$

3. 误判概率：

Error Count Estimates for g

	1	2	Total
Rate	0.0000	0.0000	0.0000
Priors	0.5000	0.5000	

两类的误判率均为0，说明判别能力很强，于是可以利用已经得到的判别函数去判别新样本。

4. 待判样本分类结果：

Obs	x1	x2	x3	_1	_2	_INTO_
1	68.5	79.3	1950	0.00000	1.00000	2
2	69.9	96.9	2840	0.00000	1.00000	2
3	77.6	93.8	5233	0.99997	0.00003	1
4	69.3	90.3	5158	0.98420	0.01580	1

结果表明，中国与罗马尼亚归入第二类，希腊与哥伦比亚归入第一类。

由以上两个例子可知，逐步判别法所得到的结果可看出来，尽管这里没有利用变量X_2(成人识字率)，但是最终的判别结果与利用全部变量所得得判别结果完全一致，这说明了三个变量在判别式中所起到的作用不同。由此可见，在解决现实问题中应结合两种方法使得更加科学的使用已知数据得到更加合理的结论。