# Bayes判别法

《美国数学建模竞赛》

完整课程请长按下方二维码

# 判别分析

判别分析方法最初应用于考古学，例如要根据挖掘出来的人头盖骨的各种指标来判别其性别年龄等. 近年来，在生物学分类，医疗诊断，地质找矿，石油钻探，天气预报等许多领域，判别分析方法已经成为一种有效的统计推断方法。

判别分析是一种在一些已知研究对象用某种方法已经分成若干类的情况下，确定新的样品的观测数据属于哪一类的统计分析方法。

肝病的判别

地震的判别

为了能识别待判断的对象$x = (x_1, x_2, \ldots, x_m)^T$是属于已知类$A_1, A_2, \ldots, A_r$中的哪一类？

事先必须要有一个一般规则，一旦知道了$x$的值，便能根据这个规则立即作出判断，称这样的一个规则为判别规则(用于衡量待判对象与各已知类别接近程度的方法准则)。

判别规则往往通过的某个函数来表达，我们把它称为判别函数，记作$W(i; x)$.

常用的方法有：距离判别法、Fisher判别法、贝叶斯判别法、逐步判别法。这里仅介绍后两种。

# Bayes判别法

Bayes判别法的基本思想：总是假设对所研究的对象已有一定的认识，计算新给样品属于各总体的条件概率 $P(G_i|x_0),\ (i$ 比较这个概率的大小，然后将新样品判归为来自概率最大的总体。

设有总体$G_i(i=1,2,\cdots,k)$，$G_i$具有概率密度函数$f_i(x)$。并且根据以往的统计分析，知道$G_i$出现的概率为$q_i$。即当样本$x_0$发生时，求他属于某类的概率。由贝叶斯公式计算后验概率，有：

$$P(G_i|x_0)=\frac{q_i f_i(x_0)}{\Sigma q_j f_j(x_0)}$$

判别规则 $P(G_h|X_0)=\max_{i<i<k} p(G_i|x_0)$

则$x_0$判给$G_h$。

**Bayes判别法的一般步骤：**

1.计算各类中变量的均值 $\bar{x}_j$ 及均值向量 $\bar{x}_h(h=1,2,..k)$ ，各变量的总均值 $\bar{x}_j(j=1,2..p)$ 及均值向量 $x$；

2.计算类内协方差矩阵S及其逆矩阵$S^{-1}$；

3.计算Bayes判别函数中，各个变量的系数及常数项并写出判别函数；

4.计算类内协方差矩阵W及总各协方差矩阵T作多个变量的全体判别效果的检验；

5.各个变量的判别能力的检验；

6.判别新样本应属于的类别。

例题：人文发展指数是联合国开发计划署于1990年5月发表的一份<<人类发展报告>>中公布的数据如下，试通过已知的样品建立判别函数,误判率是多少?并判断待判的归类.

| 类别 | 国家 | 寿命(X1) | 成人识字率%(X2) | 调整后GDP(X3) |
|---|---|---|---|---|
| 1 | 美国 | 76 | 99 | 5374 |
| 1 | 日本 | 79.5 | 99 | 5359 |
| 1 | 瑞士 | 78 | 99 | 5372 |
| 1 | 阿根廷 | 72.1 | 95.9 | 5242 |
| 1 | 阿联酋 | 73.8 | 77.7 | 5370 |
| 2 | 保加利亚 | 71.2 | 93 | 4250 |
| 2 | 古巴 | 75.3 | 94.9 | 3412 |
| 2 | 巴拉圭 | 70 | 91.2 | 3390 |
| 2 | 格鲁吉亚 | 72.8 | 99 | 2300 |
| 2 | 南非 | 62.9 | 80.6 | 3799 |
| 待判样品： | 中国 | 68.5 | 79.3 | 1950 |
| | 罗马丽亚 | 69.9 | 96.9 | 2840 |
| | 希腊 | 77.6 | 93.8 | 5233 |
| | 哥伦比亚 | 69.3 | 90.3 | 5159 |

- data ex;input g  x1-x3 @@;
- cards;
- 1 76  99    5374  1  79.5   99    5359  1  78  99    5372   1  72.1   95.9
  5242  1  73.8   77.7   53702    71.2    93   4250  2   75.3  94.9    3412 2
  70    91.2   3390 2    72.8   99    2300 2    62.9   80.6    3799
- ;
- data ex1; input x1-x3 @@;
- cards;
- 68.5 79.3   1950  69.9  96.9
  2840  77.6   93.8    5233
  69.3  90.3 5159
- ;
- proc discrim data=ex testdata=ex1
- anova manova simple list testout=ex2;
- class g; proc print data=ex2;run;

待判别类别的数据

将判别的结果输出到一个数据文件

Proc Discrim后的常用选择项有：

（1）Data=数据集名，指定输入数据集名，若缺省则指定最新建立的数据集。

（2）Testdata=数据集名，指定待作出判别的数据集名，其中的变量名须上Data数据集中的变量名一致。

（3）Testout=数据集名，指定输出数据集，输出Testdata数据集中所有观测值以及每个观测值的后验概率和判别后的类别。

（4）List，指定打印每个观测值的回代结果。

（5）Anova，指定输出各类均值检验的一元统计量。

（6）Manova，指定输出各类均值检验的多元统计量。

（7）Simple，指定打印总体和组内的简单统计量。

Linear Discriminant Function for g

| Variable | 1 | 2 |
|---|---|---|
| Constant | -323.21568 | -236.03823 |
| x1 | 5.79107 | 5.14034 |
| x2 | 0.26498 | 0.25167 |
| x3 | 0.03407 | 0.02533 |

**因此Bayes判别函数为**

**y1=-323.21568+5.79107x1+0.26498x2+0.03407x3**

**y2=-236.03823+5.14034x1+0.25167x2+0.02533x3**

Error Count Estimates for g

| | 1 | 2 | Total |
|---|---|---|---|
| Rate | 0.0000 | 0.0000 | 0.0000 |
| Priors | 0.5000 | 0.5000 | |

从上面运行结果得知，两类的误判率均为0

The SAS System

| 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 | 5159 | 0.98524 | 0.01476 | 1 |

因而得知中国与罗马尼亚归入第二类，希腊与哥伦比亚归入第一类。