附录四　判别分析

在生产、科学研究和日常生活中，经常会遇到对某一研究对象属于哪种情况作出判断。例如要根据这两天天气情况判断明天是否会下雨；医生要根据病人的体温、白血球数目及其它症状判断此病人是否会患某种疾病等等。

从概率论的角度看，可把判别问题归结为如下模型。设共有 n 个总体：

$$\xi_1,\xi_2,\cdots,\xi_n$$

其中 ξ_i 是 m 维随机变量，其分布函数为

$$F_i(x_1,\cdots,x_m)，\ i=1,2,\cdots,n$$

而 $(x_1,\cdots,x_m)$ 是表征总体特性的 m 个随机变量的取值。在判别分析中称这 m 个变量为判别因子。现有一个新的样本点 $x=(x_1,\cdots,x_m)^T$，要判断此样本点属于哪一个总体。

Matlab 的统计工具箱提供了判别函数 classify。

函数的调用格式为：

[CLASS,ERR] = CLASSIFY(SAMPLE,TRAINING,GROUP, TYPE)

其中 SAMPLE 为未知待分类的样本矩阵，TRAINING 为已知分类的样本矩阵，它们有相同的列数 m，设待分类的样本点的个数，即 SAMPLE 的行数为 s，已知样本点的个数，即 TRAINING 的行数为 t，则 GROUP 为 t 维列向量，若 TRAINING 的第 i 行属于总体 ξ_i，则

GROUP 对应位置的元素可以记为 i，TYPE 为分类方法，缺省值为'linear'，即线性分类，TYPE 还可取值'quadratic'，'mahalanobis'（mahalanobis 距离）。返回值 CLASS 为 s 维列向量，给出了 SAMPLE 中样本的分类，ERR 给出了分类误判率的估计值。

例　已知 8 个乳房肿瘤病灶组织的样本，其中前 3 个为良性肿瘤，后 5 个为恶性肿瘤。数据为细胞核显微图像的 10 个量化特征：细胞核直径，质地，周长，面积，光滑度。根据已知样本对未知的三个样本进行分类。已知样本的数据为：

13. 54, 14. 36, 87. 46, 566. 3, 0. 09779
13. 08, 15. 71, 85. 63, 520, 0. 1075
9. 504, 12. 44, 60. 34, 273. 9, 0. 1024
17. 99, 10. 38, 122. 8, 1001, 0. 1184
20. 57, 17. 77, 132. 9, 1326, 0. 08474
19. 69, 21. 25, 130, 1203, 0. 1096
11. 42, 20. 38, 77. 58, 386. 1, 0. 1425
20. 29, 14. 34, 135. 1, 1297, 0. 1003

待分类的数据为：

16.6, 28.08, 108.3, 858.1, 0.08455

20.6, 29.33, 140.1, 1265, 0.1178

7.76, 24.54, 47.92, 181, 0.05263

解：编写程序如下：

```
a=[13.54,14.36,87.46,566.3,0.09779
13.08,15.71,85.63,520,0.1075
9.504,12.44,60.34,273.9,0.1024
17.99,10.38,122.8,1001,0.1184
20.57,17.77,132.9,1326,0.08474
19.69,21.25,130,1203,0.1096
11.42,20.38,77.58,386.1,0.1425
20.29,14.34,135.1,1297,0.1003]
x=[16.6,28.08,108.3,858.1,0.08455
20.6,29.33,140.1,1265,0.1178
7.76,24.54,47.92,181,0.05263]
g=[ones(3,1);2*ones(5,1)];
[class,err]=classify(x,a,g)
```