Shenduo Zhang

Problem 1

从健康人群 G_1、硬化症患者 G_2 和冠心病患者 G_3 中分别随机选取 10,6,4 人考差了各自心电图的五个不同指标 $(X_1 \sim X_5)$. 假定各总体的协方差矩阵均相等. 由此训练样本建立距离判别准则, 并对其中的两个待判样品做判别.

(数据见 exercise5_2)

Solution **1.a** 首先对均值, 方差进行估计. 然后构造 $W_i(x), i = 1, 2, 3$ 函数, 再将待测样本带入函数, 取最大值

```
> #新样本判别
> distance(c(8.06,231.03,14.41,5.72,6.15))
[1]  TRUE FALSE FALSE
> distance(c(9.89,409.42,19.47,5.19,10.49))
[1] FALSE  TRUE FALSE
>
```

可以看到对于第一个待测样本, 最大的函数值由 $W_1(x)$ 给出, 故归位第一类. 第二个待测样本的最大函数值由 $W_2(x)$ 给出, 故其属于第二类.

回带验证的结果如下

```
> #回带验证
> apply(E1.g1, 1, distance)
         1     2     3     4     5     6     7     8     9    10
[1,]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[2,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[3,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> apply(E1.g2, 1, distance)
        11    12    13    14    15    16
[1,] FALSE FALSE FALSE FALSE FALSE FALSE
[2,]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[3,] FALSE FALSE FALSE FALSE FALSE FALSE
> apply(E1.g3, 1, distance)
        17    18    19    20
[1,] FALSE FALSE FALSE FALSE
[2,] FALSE FALSE FALSE FALSE
[3,]  TRUE  TRUE  TRUE  TRUE
```

回带验证给出的误判率的估计为 0.

□

Problem 2

某气象站预报某地区有无春旱的观测资料中, X_1 与 X_2 是与气象有关的两个综合预报因子. 数据包括发生春旱的六个年份的 X_1, X_2 的观测值和无春旱的 8 个年份的相应观测值. 假设两总体均服从正态分布且协方差矩阵 $\Sigma_1 \neq \Sigma_2$, 误判损失相同又先验概率按比例分配, 即

(数据见 exercise5_3)

$$p_1 = \frac{6}{14} = 0.4286, p_2 = \frac{8}{14} = 0.5714$$

进行两总体的 Bayes 预测.

***Solution* 2.a** 首先计算类内的均值以及每个类内的协方差矩阵的估计. 然后构造两个待测样本, 对其计算两正态总体的二次决策函数, 取广义距离函数最小的那个, 可以看到结果如下

```
> #新样本预测
> bayes(c(23.7,-3.1))
[1]  TRUE FALSE
> bayes(c(21,-0.1))
[1] FALSE  TRUE
>
```

第一个的最小广义距离函数由第一类给出, 第二个的最小广义距离由第二类给出. 故判定第一个样本属于第一类第二个样本属于第二类.

进行回带验证,

```
> apply(E2.g1,1,bayes)
      1     2     3     4     5     6
[1,]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[2,] FALSE FALSE FALSE FALSE FALSE FALSE
> apply(E2.g2,1,bayes)
      7     8     9    10    11    12    13    14
[1,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[2,]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

回带验证给出的预判率的预测为 0.

□

Problem 3

对于 IRIS 数据集进行 LDA 分类实验.

Solution **3.a** 首先将 IRIS 数据集随机分成百分之七十的训练集以及百分之三十的测试集. 使用 MASS 工具包里的 LDA 模型, 进行对其花萼花瓣长度宽度进行拟合.

```
> summary(fisher_model)
        Length Class  Mode
prior    3     -none- numeric
counts   3     -none- numeric
means   12     -none- numeric
scaling  8     -none- numeric
lev      3     -none- character
svd      2     -none- numeric
N        1     -none- numeric
call     3     -none- call
terms    3     terms  call
xlevels  0     -none- list
```

然后将拟合的模型在新的数据集进行预测, 结果如下

```
> fisher_model.predict
$class
 [1] setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     setosa     versicolor versicolor
[17] versicolor versicolor versicolor versicolor versicolor versicolor virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica  virginica
[33] virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica

$posterior
        setosa   versicolor    virginica
3   1.000000e+00 7.612214e-22 7.582317e-42
5   1.000000e+00 4.480793e-25 6.225106e-46
7   1.000000e+00 5.144210e-21 2.538166e-40
8   1.000000e+00 2.114020e-22 1.556710e-42
9   1.000000e+00 2.112499e-17 4.230470e-36
26  1.000000e+00 4.493812e-18 4.409046e-37
28  1.000000e+00 7.409626e-24 1.923879e-44
31  1.000000e+00 2.845615e-18 3.149912e-37
32  1.000000e+00 2.327615e-21 1.044361e-40
```

给出正误列联表如下

```
> table(test_data$Species,fisher_model.predict$class)

             setosa versicolor virginica
  setosa         14          0         0
  versicolor      0          9         0
  virginica       0          0        14
```

给出的误判率为

```
> sum(diag(table(test_data$Species,fisher_model.predict$class)))/sum(table(test_data$Species,fisher_model.predict$class))
[1] 1
```

□