多元统计 9

罗震林

1 Question 8.2

```r
library(MASS)
library(ggplot2)
library(ggpubr)
wine <- read.table("wine.train.txt")  #最后一列表示酒的种类

lda_wine <- lda(V14~., data = wine)
# 预测结果
result <- predict(lda_wine, wine)
## 作图
pre_wine <- data.frame(result$x, type = result$class)

p1 <- ggplot(pre_wine, aes(x = LD1,y = LD2,colour = type)) +
  geom_point() # 预测

p2 <- ggplot(pre_wine, aes(x = LD1,y = LD2,colour = factor(wine$V14))) +
  geom_point() # 实际

ggarrange(p1,p2,nrow = 1, ncol = 2,
          common.legend = T, legend = "bottom")
# 结果分析
table(wine$V14, result$class)
mean(result$class!=wine$V14)  # 误差率
```

表 1: LDA 分类结果

实际 \ 预测	1	2	3
1	38	0	0
2	0	45	1
3	0	0	34

从表1中可以看到，通过 LDA 判别得到的分类结果只有 1 个实际为种类 2 的观测值被错判成了种类 3，图形化结果见图1，总的分类错误率计算得到为 0.008474576。综上，说明 LDA 可以取到较好的分类结果，使用 LDA 方法是可行的。图1

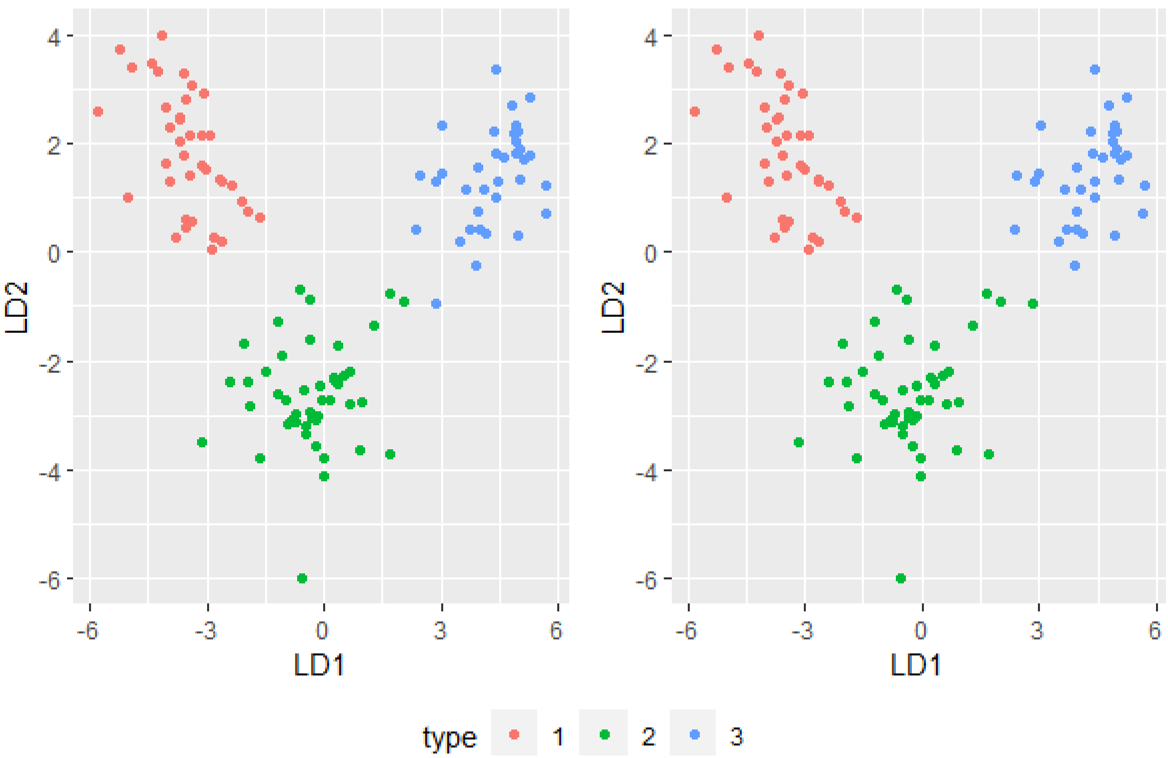

图 1: LDA 预测结果（左）和实际结果（右）

2 Question 8.3

将 $\mu_1, \mu_2, \Sigma_{XX}$ 代入统计量中，替换 $E(\mathbf{X}_1), E(\mathbf{X}_2)$ 和 $\mathrm{var}(\mathbf{X}_i), i = 1, 2$，且由 $\mathbf{X_1}$ 和 $\mathbf{X_2}$ 相互独立，则关于 $\mathbf{a}$ 的函数可以转换为

$$\begin{aligned} f(\mathbf{a}) &= \frac{\mathbf{a^T}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\mathbf{T}\mathbf{a}}{\mathbf{a}^T\mathrm{var}(\mathbf{X_1} - \mathbf{X_2})\mathbf{a}} \\ &= \frac{\mathbf{a}^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\mathbf{a}}{2\mathbf{a}^T\Sigma_{XX}\mathbf{a}} \end{aligned}$$

不妨令 $2\mathbf{a}^T\Sigma_{XX}\mathbf{a} = 1$（因为分母为任意值时，都可以令 $\mathbf{a}$ 乘以一个实数，使分母为 1，且不会改变 $f(\mathbf{a})$ 的大小，所以不妨对分母进行限制），则问题可以转化为

$$\begin{aligned} \max \quad & \mathbf{a}^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\mathbf{a} \\ \text{s.t.} \quad & 2\mathbf{a}^T\Sigma_{XX}\mathbf{a} = 1 \end{aligned}$$

运用拉格朗日乘子法，有

$$L(\mathbf{a}) = \mathbf{a}^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\mathbf{a} - \lambda(2\mathbf{a}^T\Sigma_{XX}\mathbf{a} - 1)$$

对 $\mathbf{a}$ 求导，并令其值为 0，得

$$\frac{\partial f}{\partial \mathbf{a}} = 2(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\mathbf{a} - 4\lambda\Sigma_{XX}\mathbf{a} = 0$$

进而有

$$\frac{1}{2}\Sigma_{XX}^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T\mathbf{a} = \lambda\mathbf{a}$$

因为 $(\mu_1-\mu_2)(\mu_1-\mu_2)^T\mathbf{a}$ 的方向始终为 $\mu_1-\mu_2$, 故可以用 $k(\mu_1-\mu_2)$ 代替（k 为任意实数），所以有

$$\mathbf{a}\propto\Sigma_{XX}^{-1}(\mu_1-\mu_2)$$

3 Question 8.6

```
library(MASS)
library(dplyr)
library(iris)
library(ggplot2)
data(iris)

# 数据变换
iris_adjust <- iris %>% mutate(Sepal.Shape = log(Sepal.Length/Sepal.Width),
                Petal.Shape = log(Petal.Length/Petal.Width))

# 转换后数据
boxplot(iris_adjust$Sepal.Shape, iris_adjust$Petal.Shape,
        names = c("Sepal.Shape","Petal.Shape"))
```

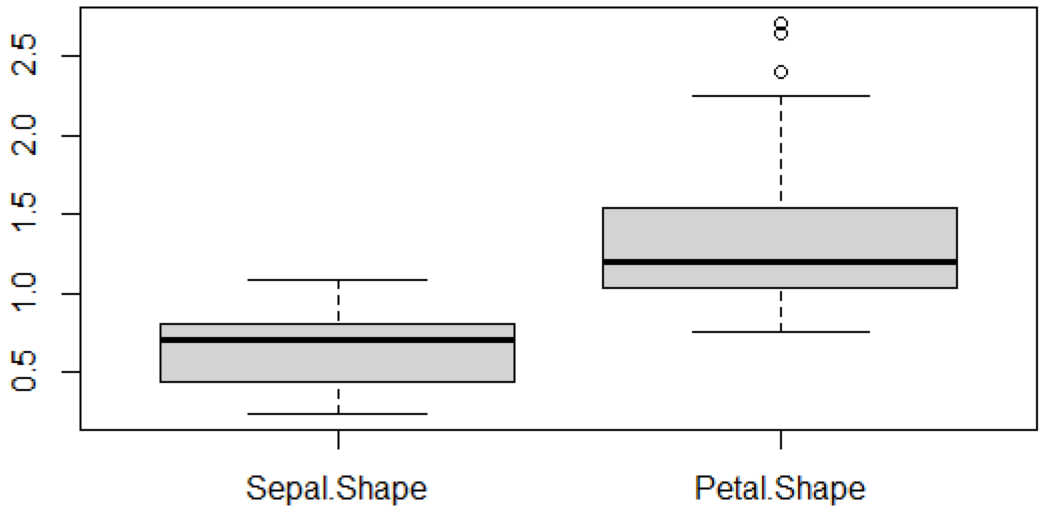

图 2: 转换后数据 X_5（左）和 X_6（右）

```
# LDA
lda_iris <- lda(Species~Sepal.Shape+Petal.Shape, data = iris_adjust)
result <- predict(lda_iris, iris_adjust[,c(6,7)])
table(iris_adjust$Species, result$class)
mean(iris_adjust$Species!=result$class) # 误判率

# QDA
qda_iris <- qda(Species~Sepal.Shape+Petal.Shape, data = iris_adjust)
qda_result <- predict(qda_iris, iris_adjust[,c(6,7)])
table(iris_adjust$Species, qda_result$class)
mean(iris_adjust$Species!=qda_result$class) # 误判率
```

表 2: 分析结果对比

预测 \ 实际	setosa	versicolor	virginica
setosa	49	1	0
versicolor	0	41	9
virginica	0	16	34

(a) LDA 判别结果

预测 \ 实际	setosa	versicolor	virginica
setosa	49	1	0
versicolor	0	42	8
virginica	0	11	39

(b) QDA 判别结果

由计算，LDA 的误判率为 0.1733333，QDA 的误判率为 0.1333333. 因此对于该数据集，QDA 的判别效果会好于 LDA，但仍有较大的误判率，需要进一步改善方法