# 《多元统计分析》课后作业

姓名：李倩倩

学号：2024017349

班级：统计 24-1 班

中国石油大学（北京）克拉玛依校区文理学院数学与统计系

Thursday 28th May, 2026

# 作业要求

1. 可以和其他同学讨论作业当中的问题，但应当自己独立完成作业
2. 计算、证明等要有过程，要有主要步骤的说明
3. 请将计算、绘图所用的 R 代码以及生成的结果和图像一并添加在作业文件当中
4. 请使用 LaTeX 编辑并生成 PDF 格式的文件，第 X 周作业文件命名方式：学号-姓名-X.pdf
5. 评分标准：每一问得分 $\in \{2, 1, 0\}$
   - 2: 按时完成并上交作业，且答案基本正确
   - 1: 按时完成并上交作业，且答案部分正确
   - 0: 答案完全错误，或者迟交作业 (规定时间 72 小时之后)
6. 请将完成的 PDF 格式的作业文件发送至邮箱：xiaolei@cup.edu.cn
7. 每位同学可以有一次迟交作业的机会，但不得晚于规定时间三日之后
8. 第 11 周作业截止时间：2026 年 5 月 29 日 24:00

# 目录

# Chapter 1

# 第 11 周作业

第 11 周作业截止时间： 2026 年 5 月 29 日 24:00

第 11 周作业完成时间： Thursday 28th May, 2026 12:19

1. [2 分] 假设 $x \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$，并且

$$\begin{aligned} \Pi_1: &\quad X \sim b(10, 0.2) \quad \text{先验概率 } \pi_1 = 0.5; \\ \Pi_2: &\quad X \sim b(10, 0.3) \quad \text{先验概率 } \pi_2 = 0.3; \\ \Pi_3: &\quad X \sim b(10, 0.5) \quad \text{先验概率 } \pi_3 = 0.2. \end{aligned}$$

利用 Bayes 决策法则，确定集合 $R_1$, $R_2$ 及 $R_3$.

**【解】** Bayes 决策法则为：对每一个观测值 $x$，比较

$$\pi_i P_i(X = x), \qquad i = 1, 2, 3,$$

其中

$$P_i(X = x) = \binom{10}{x} p_i^x (1 - p_i)^{10-x}.$$

将 $x$ 判入使 $\pi_i P_i(X = x)$ 最大的总体.

计算所用 R 代码如下：

```
x <- 0:10
prior <- c(Pi1 = 0.5, Pi2 = 0.3, Pi3 = 0.2)
p <- c(Pi1 = 0.2, Pi2 = 0.3, Pi3 = 0.5)

score <- sapply(seq_along(p), function(i) {
  prior[i] * dbinom(x, size = 10, prob = p[i])
})
```

```
colnames(score) <- names(prior)

decision <- max.col(score, ties.method = "first")
bayes_table <- data.frame(x = x, round(score, 8),
                          Decision = paste0("R", decision))
bayes_table
split(x, decision)
```

得到如下比较表:

表 1.1: Bayes 判别法则的逐点比较.

| $x$ | $\pi_1 P_1(X=x)$ | $\pi_2 P_2(X=x)$ | $\pi_3 P_3(X=x)$ | 判入集合 |
|---|---|---|---|---|
| 0 | 0.05368709 | 0.00847426 | 0.00019531 | $R_1$ |
| 1 | 0.13421773 | 0.03631825 | 0.00195313 | $R_1$ |
| 2 | 0.15099494 | 0.07004233 | 0.00878906 | $R_1$ |
| 3 | 0.10066330 | 0.08004838 | 0.02343750 | $R_1$ |
| 4 | 0.04404019 | 0.06003628 | 0.04101562 | $R_2$ |
| 5 | 0.01321206 | 0.03087580 | 0.04921875 | $R_3$ |
| 6 | 0.00275251 | 0.01102707 | 0.04101562 | $R_3$ |
| 7 | 0.00039322 | 0.00270051 | 0.02343750 | $R_3$ |
| 8 | 0.00003686 | 0.00043401 | 0.00878906 | $R_3$ |
| 9 | 0.00000205 | 0.00004133 | 0.00195313 | $R_3$ |
| 10 | 0.00000005 | 0.00000177 | 0.00019531 | $R_3$ |

因此 Bayes 决策区域为

$$R_1 = \{0,1,2,3\}, \qquad R_2 = \{4\}, \qquad R_3 = \{5,6,7,8,9,10\}.$$

2. 数据 breast cancer wisconsin.csv 来自 University of Wisconsin Hospitals (July, 1992). 共有 699 个观测数据, 每一列对应下述 11 个变量之一:

   - 样本编号.
   - 肿块厚度 (取值 $1 \sim 10$).
   - 细胞大小的均匀性 (取值 $1 \sim 10$).
   - 细胞形状的均匀性 (取值 $1 \sim 10$).
   - 边缘附着力 (取值 $1 \sim 10$).
   - 单个上皮细胞大小 (取值 $1 \sim 10$).
   - 裸核 (取值 $1 \sim 10$).
   - 乏味染色体 (取值 $1 \sim 10$).

- 正常核 (取值 $1 \sim 10$).
- 线粒体 (取值 $1 \sim 10$).
- 分类: (2 代表良性，4 代表恶性).

分析该数据集并回答以下问题:

(a) [2 分] 有若干个观测数据含有一个标记为 “?” 的缺失值 (即，无可用数值)，找到它们，并将相应的观测值从数据集当中移除. 还剩下多少观测值?

**【解】** 读入数据时将 “?” 记为缺失值，并用 `complete.cases()` 找出含缺失值的观测.

```
bc_file <- "breast cancer wisconsin.csv"
var_names <- c("ID", "Clump", "Size", "Shape", "Adhesion",
               "SECSize", "Nuclei", "Chromatin", "Nucleoli",
               "Mitoses", "Class")

bc_raw <- read.csv(bc_file, header = FALSE, col.names = var_names,
                   na.strings = "?", stringsAsFactors = FALSE)

missing_rows <- which(!complete.cases(bc_raw))
missing_info <- bc_raw[missing_rows, c("ID", "Nuclei", "Class")]
missing_rows
missing_info

bc_clean <- bc_raw[complete.cases(bc_raw), ]
nrow(bc_raw)
nrow(bc_clean)
```

含缺失值的观测均在变量 `Nuclei` 上缺失，位置如下:

表 1.2: 含缺失值的观测.

| 原始行号 | 样本编号 | 缺失变量 | 分类 |
|---|---|---|---|
| 24 | 1057013 | Nuclei | 4 |
| 41 | 1096800 | Nuclei | 2 |
| 140 | 1183246 | Nuclei | 2 |
| 146 | 1184840 | Nuclei | 2 |
| 159 | 1193683 | Nuclei | 2 |
| 165 | 1197510 | Nuclei | 2 |
| 236 | 1241232 | Nuclei | 2 |
| 250 | 169356 | Nuclei | 2 |
| 276 | 432809 | Nuclei | 2 |
| 293 | 563649 | Nuclei | 4 |
| 295 | 606140 | Nuclei | 2 |
| 298 | 61634 | Nuclei | 2 |
| 316 | 704168 | Nuclei | 2 |
| 322 | 733639 | Nuclei | 2 |
| 412 | 1238464 | Nuclei | 2 |
| 618 | 1057067 | Nuclei | 2 |

原始数据共有 699 个观测，其中含缺失值的观测有 16 个. 删除这些观测后，还剩

$$699 - 16 = 683$$

个观测值.

(b) 最后一列给出了诊断结论：良性 (表示为 2) 与恶性 (表示为 4). 利用 MASS 包的函数 lda() 对 2a 中得到的数据 (除去样本编号) 作线性判别分析.

i. [2 分] 对这两类而言，你的先验概率是多少.

【解】 采用 lda() 的默认设置，即使用样本中各类别的比例作为先验概率. 清洗后的 683 个观测中，良性样本有 444 个，恶性样本有 239 个.

```
library(MASS)

bc_lda <- bc_clean[, -1]
bc_lda$Class <- factor(bc_lda$Class, levels = c(2, 4))
lda_fit <- lda(Class ~ ., data = bc_lda)

table(bc_lda$Class)
round(lda_fit$prior, 6)
```

输出为:

```
2 4
```

```
444 239

        2         4
0.650073 0.349927
```

因此两类的先验概率为

$$\hat{\pi}_2 = \frac{444}{683} = 0.650073, \qquad \hat{\pi}_4 = \frac{239}{683} = 0.349927.$$

ii. [2 分] 写出线性判别函数的表达式.

【解】 设

$$x_1 = \text{Clump}, \quad x_2 = \text{Size}, \quad x_3 = \text{Shape},$$
$$x_4 = \text{Adhesion}, \quad x_5 = \text{SECSize}, \quad x_6 = \text{Nuclei},$$
$$x_7 = \text{Chromatin}, \quad x_8 = \text{Nucleoli}, \quad x_9 = \text{Mitoses}.$$

用 `lda()` 得到的第一线性判别变量为:

```
round(lda_fit$means, 4)
```

```
ld_score <- predict(lda_fit)$x[, 1]
scaling1 <- lda_fit$scaling[, 1]
if (mean(ld_score[bc_lda$Class == "4"]) <
    mean(ld_score[bc_lda$Class == "2"])) {
  scaling1 <- -scaling1
}
round(cbind(LD1 = scaling1), 6)
```

将符号方向统一为恶性样本的 $LD_1$ 均值大于良性样本后，判别系数为:

| | LD1 |
|---|---|
| Clump | 0.182656 |
| Size | 0.125819 |
| Shape | 0.090079 |
| Adhesion | 0.047478 |
| SECSize | 0.058029 |
| Nuclei | 0.261409 |
| Chromatin | 0.110445 |
| Nucleoli | 0.106722 |
| Mitoses | 0.005638 |

因此可以写成

$$\begin{aligned} LD_1 = {} & 0.182656x_1 + 0.125819x_2 + 0.090079x_3 + 0.047478x_4 \\ & + 0.058029x_5 + 0.261409x_6 + 0.110445x_7 + 0.106722x_8 + 0.005638x_9. \end{aligned}$$

这里符号方向取为恶性样本的 $LD_1$ 平均值大于良性样本；若软件输出整体变号，则判别结果不变.

从两类线性判别函数之差也可直接写出判别规则：

```
X <- as.matrix(bc_lda[, var_names[2:10]])
grp <- bc_lda$Class
mu <- rowsum(X, grp) / as.vector(table(grp))

S <- Reduce("+", lapply(levels(grp), function(g) {
  Z <- sweep(X[grp == g, , drop = FALSE], 2, mu[g, ], "-")
  t(Z) %*% Z
})) / (nrow(X) - nlevels(grp))

Sinv <- solve(S)
w <- Sinv %*% (mu["4", ] - mu["2", ])
b <- -0.5 * (mu["4", ] %*% Sinv %*% mu["4", ] -
               mu["2", ] %*% Sinv %*% mu["2", ]) +
     log(lda_fit$prior["4"] / lda_fit$prior["2"])

round(c(Intercept = as.numeric(b),
        setNames(as.numeric(w), colnames(X))), 6)
```

$$\begin{aligned} D(x) &= \delta_4(x) - \delta_2(x) \\ &= -20.878096 + 0.887204x_1 + 0.611135x_2 + 0.437534x_3 + 0.230613x_4 \\ &\quad + 0.281861x_5 + 1.269724x_6 + 0.536456x_7 + 0.518376x_8 + 0.027385x_9. \end{aligned}$$

当 $D(x) > 0$ 时判为恶性 (4)，否则判为良性 (2).

iii. [2 分] 利用 `plot()` 函数，使用参数 `dimen = 1` 以及 `type = "both"`，对你的判别结果进行可视化.

**【解】** 绘图所用 R 代码如下：

```
ld_score <- predict(lda_fit)$x[, 1]
lda_plot <- lda_fit
if (mean(ld_score[bc_lda$Class == "4"]) <
    mean(ld_score[bc_lda$Class == "2"])) {
```

```
  lda_plot$scaling[, 1] <- -lda_plot$scaling[, 1]
}

plot(lda_plot, dimen = 1, type = "both")

pdf("figures/p11_lda_plot.pdf", width = 7, height = 5)
plot(lda_plot, dimen = 1, type = "both")
dev.off()
```

由图可以看到，良性样本主要集中在较小的 $LD_1$ 区域，恶性样本主要集中在较大的 $LD_1$ 区域，两类样本之间有较明显的分离，但在中间区域仍存在少量重叠.

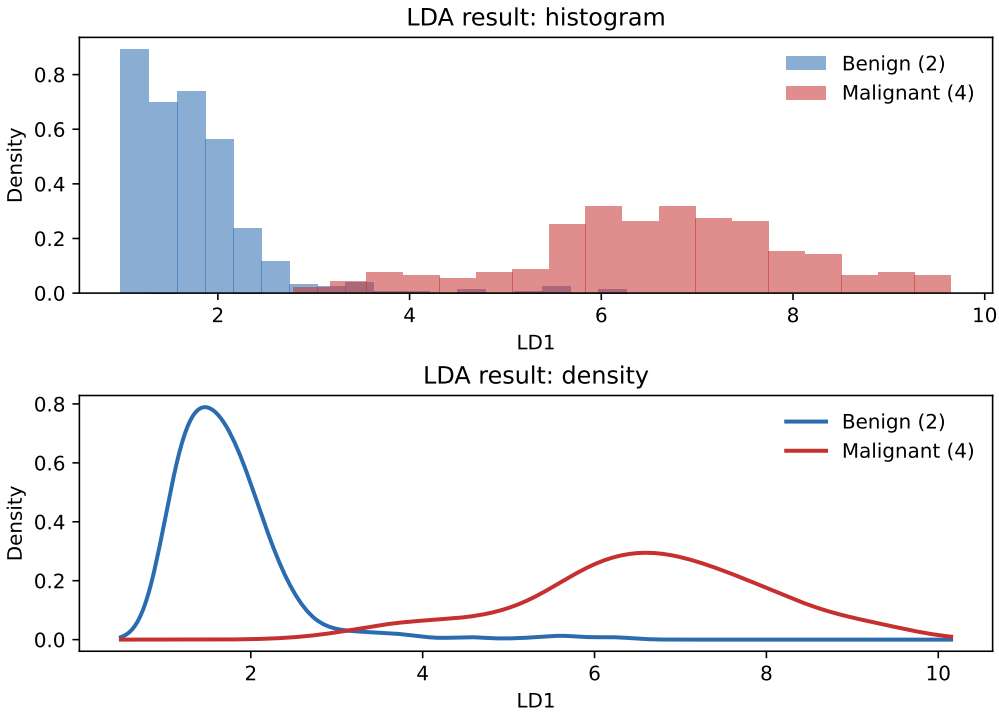


图 1.1: 乳腺癌数据的 LDA 一维判别结果可视化.

iv. [2 分] 应用你的判别规则于整个数据集，计算误判的数目，给出混淆矩阵和总误判率.

【解】 将训练得到的 LDA 判别规则应用于清洗后的全部 683 个观测:

```
train_pred <- predict(lda_fit, newdata = bc_lda)
conf_mat <- table(Actual = bc_lda$Class,
                  Predicted = train_pred$class)
conf_mat

mis_num <- sum(train_pred$class != bc_lda$Class)
mis_rate <- mean(train_pred$class != bc_lda$Class)
mis_num
```

mis_rate

混淆矩阵为：

表 1.3: LDA 在清洗后全数据集上的混淆矩阵.

| 真实类别 | 预测类别 | |
|---|---|---|
| | 2 | 4 |
| 2 | 436 | 8 |
| 4 | 19 | 220 |

误判数为

$$8 + 19 = 27,$$

总误判率为

$$\frac{27}{683} = 0.039531 \approx 3.953\%.$$

v. [2 分] 数据集 breast cancer new.csv 是 16 个新患者的检查结果，利用你的判别规则对其进行判别，有多少患者属于良性? 有多少患者属于恶性?

【解】 用训练好的 LDA 模型预测 16 个新患者：

```
new_file <- "breast cancer new.csv"
new_patients <- read.csv(new_file, stringsAsFactors = FALSE)

new_pred <- predict(lda_fit, newdata = new_patients)
data.frame(No = 1:nrow(new_patients),
           Predicted = new_pred$class)
table(new_pred$class)
```

逐个患者的预测类别为：

表 1.4: 16 个新患者的 LDA 判别结果.

| 编号 | 预测类别 | 编号 | 预测类别 | 编号 | 预测类别 | 编号 | 预测类别 |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 2 | 4 | 3 | 2 | 4 | 2 |
| 5 | 2 | 6 | 4 | 7 | 2 | 8 | 2 |
| 9 | 2 | 10 | 4 | 11 | 2 | 12 | 2 |
| 13 | 4 | 14 | 2 | 15 | 2 | 16 | 2 |

因此，在这 16 个新患者中，预测为良性 (2) 的有 12 人，预测为恶性 (4) 的有 4 人.