第6章 判别分析及R使用

武慧
wuh@hit.edu.cn

经济管理学院
哈尔滨工业大学（威海）

1

本章内容

1. 判别分析概述
2. 线性判别分析（LDA）及R实现
3. 二次判别分析（QDA）及R实现

1. 判别分析概述

3

判别分析的定义

- 判别分析（Discriminant Analysis, DA）：一种监督学习算法，通过已知类别标签的数据，建立分类模型，预测新样本的类别。
- 核心任务：寻找特征的“最优组合”，构造判别函数，将不同类别的样本尽可能分开。

如果给你身高和体重数据，如何区分男女生？

判别分析就是数学化的‘画分界线’工具。

4

判别分析应用场景

- 医学：根据肿瘤大小、细胞形态判别良性/恶性。
- 金融：基于收入、信用评分判别贷款违约风险。
- 营销：通过消费行为划分客户群体（如高/低价值用户）。

能否想到其他场景？比如图像识别中判别猫狗？

如果数据没有标签，能否用判别分析？

判别分析的类型

- 线性判别分析（LDA）：假设数据在每个类别中服从高斯分布，并且假设各类别具有相同的协方差矩阵，适用于线性可分的数据，计算较简单。
- 二次判别分析（QDA）：适用于类别间协方差矩阵不相同的情况，更加灵活但容易过拟合。需要估计每个类别的协方差矩阵，因此比LDA更复杂，计算开销较大。

6

单选题 1分

判别分析的主要目的是：

A 预测连续变量的值

B 对样本进行分类

C 计算不同变量之间的相关性

D 评估变量之间的回归关系

7

单选题 1分

判别分析与回归分析的主要区别在于:

A 判别分析的输出是类别标签，回归分析的输出是连续值

B 判别分析用于回归问题，回归分析用于分类问题

C 判别分析只能处理线性关系，回归分析只能处理非线性关系

D 判别分析无需训练集数据，回归分析需要

8

单选题 1分

在判别分析中，哪种方法假设各类别的特征变量遵循正态分布并且具有相同的协方差矩阵?

A 线性判别分析 (LDA)

B 二次判别分析

C K-最近邻 (KNN)

D 支持向量机 (SVM)

单选题 1分

下列关于判别分析的说法，哪一个是正确的?

A 判别分析在处理大规模数据集时表现比回归分析更差

B 判别分析常用于多类别问题，回归分析只适用于二分类问题

C 判别分析仅限于线性分类，无法处理非线性问题

D 判别分析常用于人脸识别、信用评分等分类问题

10

单选题 1分

判别分析中的“判别函数”是用来：

A 预测未来的数据趋势

B 对新样本进行分类，判断其所属类别

C 找出数据中的主成分

D 计算样本间的相关性

2. 线性判别分析（LDA）及R实现

12

贝叶斯决策规则

- LDA是一种基于贝叶斯决策规则的分类方法。
- 贝叶斯决策规则的核心思想是：

➢对于每个类别，计算给定特征向量的后验概率，然后选择后验概率最大的类别。

➢后验概率可以通过贝叶斯定理计算：

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)}$$

- $P(C_k|X)$：给定特征X的条件下类别C_k的概率（后验概率）。
- $P(X|C_k)$：类别C_k下特征X的似然。
- $P(C_k)$：类别C_k的先验概率。
- $P(X)$：特征X的边际概率。

13

LDA的基本原理

- 基本假设：

➢各类别的数据服从高斯分布。

➢不同类别的协方差矩阵相同。

- LDA的判别函数是线性函数。

$$p(\mathbf{x}|k) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k)\right)$$

$$\Downarrow$$

$$p(k|\mathbf{x}) \propto p(\mathbf{x}|k)p(k)$$

$$\Downarrow$$

$$\log p(k|\mathbf{x}) \propto -\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k) + \log p(k)$$

$$\Downarrow$$

$$\log p(k|\mathbf{x}) \propto -\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x} + \mu_k^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log p(k)$$

$$\Downarrow$$

$$\delta_k(\mathbf{x}) = \mu_k^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log p(k)$$

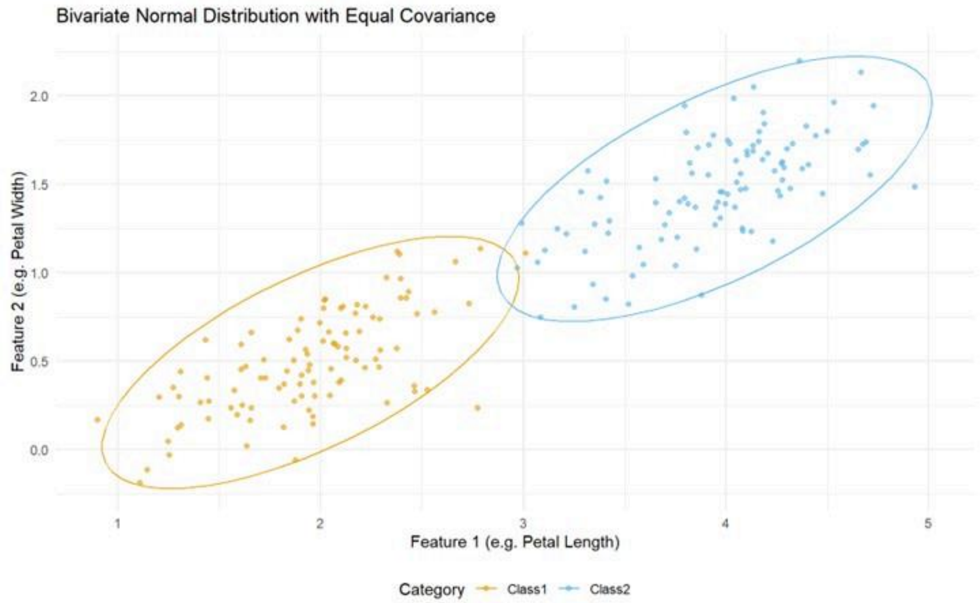
$$\Downarrow$$

$$\hat{k} = \arg\max_k \delta_k(\mathbf{x})$$

14

核心假设

- **多元正态分布**：每个类别k的特征向量服从$N(\mu_k, \Sigma)$。
- **同方差性**：所有类别共享相同的协方差矩阵Σ。

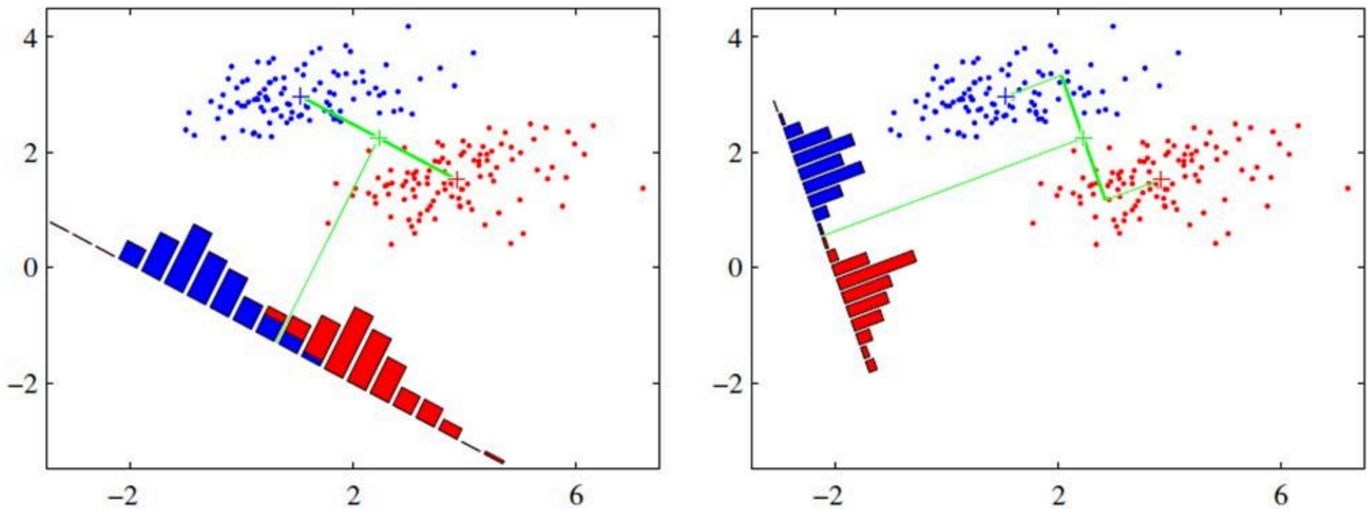


推导如何寻找最优投影方向，最大化类别可分性？

15

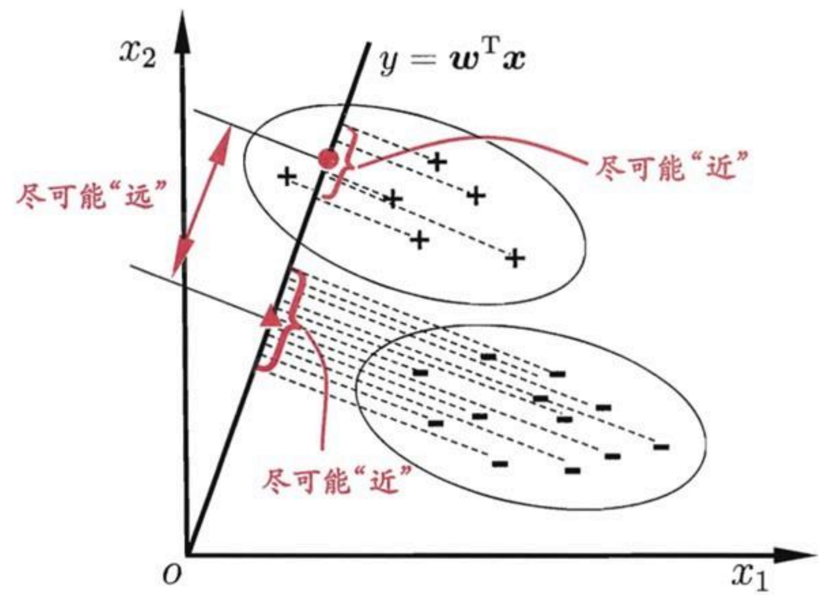
LDA的基本原理

两种投影方式，哪一种能更好的满足我们的标准呢？



16

LDA的基本原理



LDA 的二维示意图. "+"、"–" 分别代表正例和反例, 椭圆表示数据簇的外轮廓, 虚线表示投影, 红色实心圆和实心三角形分别表示两类样本投影后的中心点.

17

类间方差与类内方差

- LDA的核心目标是目标是找到一个线性变换，使得在这个变换下，类间方差与类内方差的比值最大化。
- 类内散度矩阵反映了各类别内部的样本分布情况：

$$S_W = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ik} - \mu_k)(x_{ik} - \mu_k)^T$$

- K 是类别数，n_k 是类别 k 的样本数量，x_{ik} 是类别 k 的样本，μ_k 是类别 k 的均值向量。

- 类间散度矩阵反映了不同类别之间均值的分布情况：

$$S_B = \sum_{k=1}^{K} n_k(\mu_k - \mu)(\mu_k - \mu)^T$$

- μ 是所有类别的全局均值，μ_k 是类别 k 的均值。

18

最优化目标与分类决策

- LDA的目标是选择投影方向w，使得类间散度与类内散度的比值最大化：

$$J(w)=\frac{w^T S_B w}{w^T S_W w}$$

最后，通过求解广义瑞利商问题，得到最佳投影方向。

- 分类决策：LDA的分类决策基于投影后的数据点的位置。将测试样本投影到相同的方向后，计算测试样本与每个类别均值的距离，并将测试样本分配给距离最小的类别。

19

数据准备

数据准备：使用iris数据集，该数据集包含150个样本，分为3个不同的花卉类别，每个类别有4个特征（如花瓣和萼片的长度和宽度）。

```
> data("iris")
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species  
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa    :50  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199                  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800                  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500                  
```

20

数据准备

```
# 随机划分训练集/测试集（70%训练）
library(caret)
set.seed(123)
train_index <- createDataPartition(iris$Species, p = 0.7, list = FALSE)
iris_train <- iris[train_index, ]
iris_test <- iris[-train_index, ]
```

建模与预测

```
library(MASS)
#建模
model_lda <- lda(Species ~ ., data = iris_train)

#预测新样本
pred_lda <- predict(model_lda, newdata = iris_test)
pred_lda$class   # 预测类别
pred_lda$posterior   # 后验概率
```

```
> pred_lda$class  # 预测类别
 [1] setosa     setosa     setosa     setosa     setosa     setosa
setosa     setosa     setosa     setosa
[11] setosa     setosa     setosa     setosa     setosa     versicolor
versicolor versicolor versicolor versicolor
[21] versicolor versicolor versicolor versicolor versicolor versicolor
versicolor versicolor versicolor versicolor
[31] virginica  virginica  virginica  virginica  virginica  virginica
virginica  virginica  virginica  virginica
[41] versicolor virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

23

```
> pred_lda$posterior  # 后验概率
          setosa   versicolor    virginica
1   1.000000e+00 5.180307e-23 7.105444e-45
2   1.000000e+00 6.112832e-19 7.524191e-40
6   1.000000e+00 1.222997e-21 4.908094e-42
16  1.000000e+00 5.033088e-28 2.899621e-50
18  1.000000e+00 6.497181e-22 4.190740e-43
20  1.000000e+00 5.979798e-23 3.701092e-44
22  1.000000e+00 3.519153e-21 1.376569e-41
23  1.000000e+00 8.279076e-26 1.719910e-48
34  1.000000e+00 3.843511e-29 1.589131e-52
35  1.000000e+00 1.238706e-18 2.642630e-39
38  1.000000e+00 4.566464e-24 2.028376e-46
39  1.000000e+00 3.940714e-18 1.247896e-38
44  1.000000e+00 2.639596e-16 1.374334e-34
46  1.000000e+00 1.746224e-17 1.446013e-37
47  1.000000e+00 4.533411e-23 1.389874e-44
51  4.034130e-19 9.999432e-01 5.677908e-05
53  3.319791e-23 9.972221e-01 2.777924e-03
54  1.399896e-22 9.999001e-01 9.989128e-05
64  2.359039e-24 9.965518e-01 3.448213e-03
72  4.446081e-17 9.999973e-01 2.657600e-06
74  7.934156e-23 9.997898e-01 2.102467e-04
78  6.555931e-28 7.170027e-01 2.829973e-01
81  9.342777e-18 9.999994e-01 6.195122e-07
85  2.450876e-25 9.733615e-01 2.663852e-02
87  5.795655e-22 9.989503e-01 1.049657e-03
90  3.082332e-21 9.999447e-01 5.530622e-05
91  2.218765e-23 9.997424e-01 2.575835e-04
94  3.166380e-14 1.000000e+00 1.547191e-08
99  1.088353e-10 1.000000e+00 2.227743e-09
100 1.737428e-19 9.999741e-01 2.593075e-05
101 1.238967e-54 1.305533e-09 1.000000e+00
106 2.363996e-51 1.731195e-07 9.999998e-01
109 6.799656e-44 1.417482e-04 9.998583e-01
111 1.892038e-33 9.146281e-03 9.908537e-01
```

24

混淆矩阵与准确率

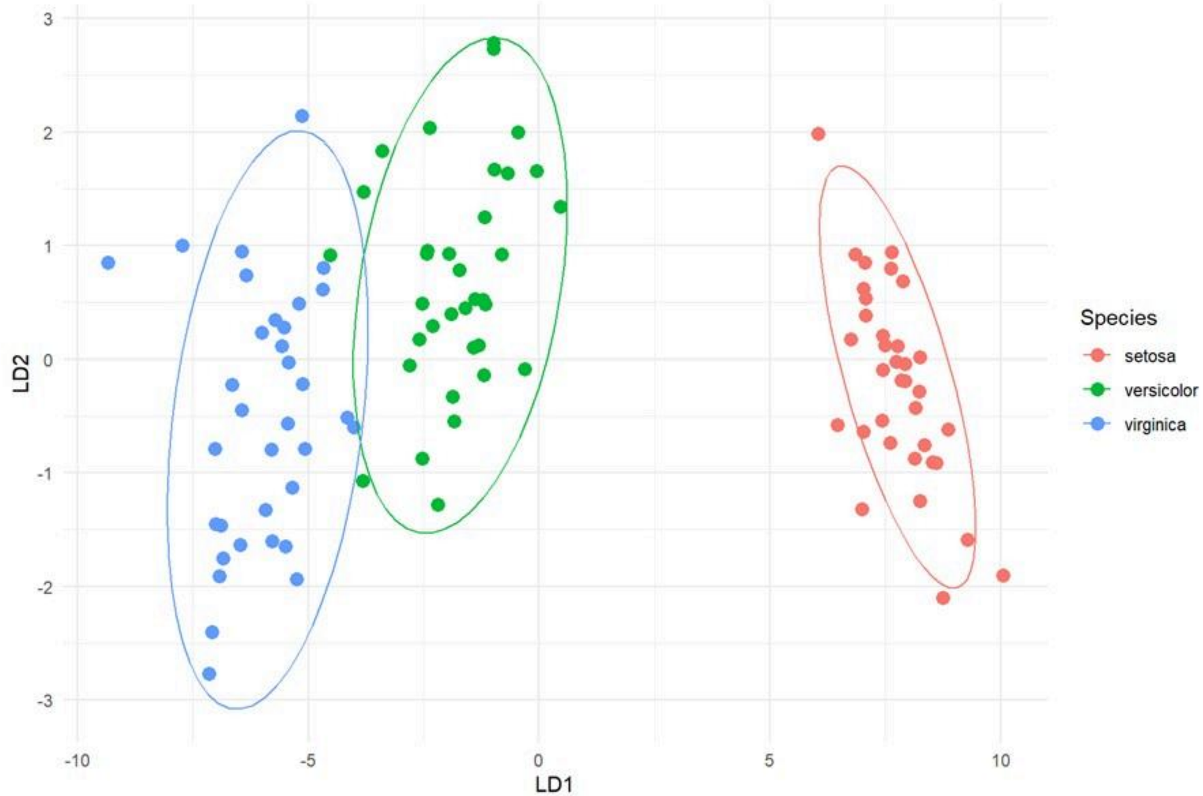
```r
# 计算混淆矩阵
confusion_matrix <- table(Predicted = pred_lda$class, Actual = iris_test$Species)

#计算准确率
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
```

```
> confusion_matrix
            Actual
Predicted    setosa versicolor virginica
  setosa         15          0         0
  versicolor      0         15         1
  virginica       0          0        14
> accuracy
[1] 0.9777778
```

可视化投影结果

```
# 提取LDA投影后的坐标（LD1和LD2）
iris_plot <- data.frame(
  LD1 = predict(model_lda)$x[, 1],
  LD2 = predict(model_lda)$x[, 2],
  Species = iris_train$Species
)
ggplot(iris_plot, aes(LD1, LD2, color = Species)) +
  geom_point(size = 3) +
  stat_ellipse(level = 0.95) +
  theme_minimal()
```



LD1为第一判别方向，解释了大部分类别差异。

28

单选题 1分

LDA的核心假设是?

A 特征服从均匀分布

B 各类别协方差矩阵相同

C 类别先验概率相等

D 数据线性可分

单选题 1分

在LDA中，贝叶斯决策规则的目标是：

A 最小化分类错误率

B 计算样本的均值

C 最大化后验概率

D 优化回归方程的参数

30

单选题 1分

LDA的优化目标是：

A 类间方差与类内方差的比值

B 最大化预测的准确性

C 最大化类内协方差

D 最小化均方误差

31

单选题 1分

以下关于LDA的说法，哪一个是正确的?

A LDA可以应用于非线性分类问题

B LDA假设特征之间存在明显的相关性

C LDA可以同时处理多类分类问题

D LDA不需要训练数据集

32

思考

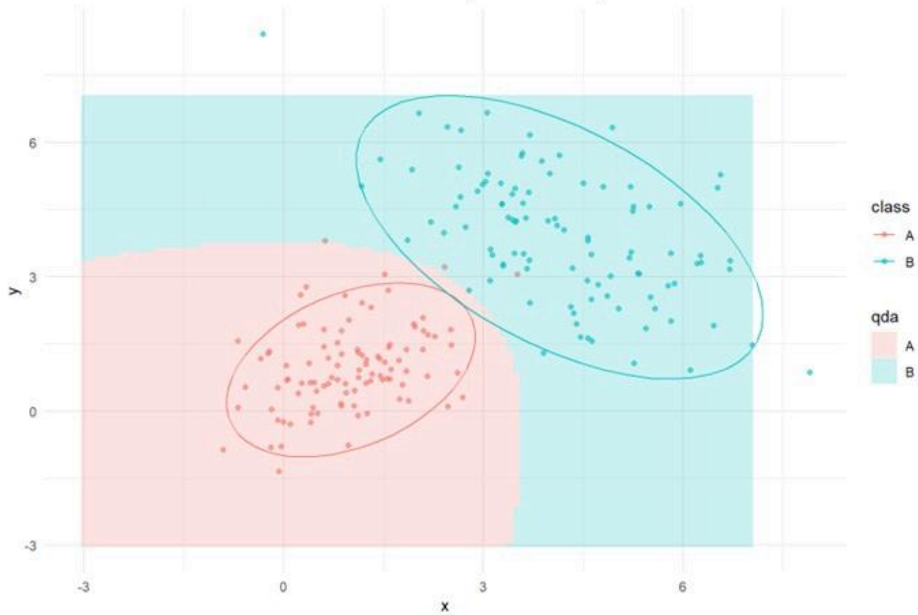
1. 如果数据不满足LDA的假设（如协方差矩阵不相同），如何调整或使用其他方法？

答：使用二次判别分析（QDA）；使用其他分类算法（SVM、KNN等）。

3. 二次判别分析（QDA）及R实现

核心假设

- 多元正态分布：每个类别k的特征向量服从$N(\mu_k, \Sigma_k)$。
- 异方差性：各类别协方差矩阵不同。



35

贝叶斯定理与后验概率

- 使用贝叶斯定理来计算给定x类别C_k的后验概率：

$$P(C_k|x) = \frac{P(x|C_k)P(C_k)}{P(x)}$$

- $P(x|C_k)$ 是类别 C_k 下的似然函数，表示数据点 x 在该类别下的概率密度。
- $P(C_k)$ 是类别 C_k 的先验概率。
- $P(x)$ 是总概率（是一个常数，不依赖于类别），在后续推导中可以忽略。

- 通过最大化后验概率选择类别：

$$\hat{k} = \arg\max_k P(C_k|x)$$

- QDA假设：

$$P(x|C_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right)$$

36

判别函数的推导

- 为了选择最有可能的类别，我们计算每个类别C_k的判别函数$g_k(\boldsymbol{x})$，并选择使判别函数最大的类别。

$$\ln P(C_k|x) = \ln P(x|C_k) + \ln P(C_k) - \ln P(x)$$

$$\Downarrow$$

$$g_k(x) = \ln P(x|C_k) + \ln P(C_k) = \ln\left(\frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right)\right) + \ln P(C_k)$$

$$\Downarrow$$

$$g_k(x) = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \ln P(C_k) \propto -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \ln P(C_k)$$

$$\Downarrow$$

判别函数为二次型

$$g_k(x) = -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k + \ln P(C_k) \Rightarrow \hat{k} = \arg\max_k g_k(x)$$

37

数据准备

```r
# 生成异方差数据
library(mvtnorm)
set.seed(123)
mu1 <- c(1, 1); mu2 <- c(4, 4)
sigma1 <- matrix(c(1, 0.5, 0.5, 1), 2); sigma2 <- matrix(c(2, -1,
-1, 2), 2)
class1 <- rmvnorm(100, mu1, sigma1)
class2 <- rmvnorm(100, mu2, sigma2)
data_qda <- data.frame(
  x = c(class1[,1], class2[,1]),
  y = c(class1[,2], class2[,2]),
  class = factor(rep(c("A", "B"), each = 100)))
```

训练与预测

```
#训练QDA模型
library(MASS)
model_qda <- qda(class ~ x + y, data = data_qda)

# 预测与评估
pred_qda <- predict(model_qda, newdata = data_qda)
confusion_matrix_qda <- table(Predicted = pred_qda$class, Actual
= data_qda$class)
accuracy_qda <- sum(diag(confusion_matrix_qda)) / nrow(data_qda)
```

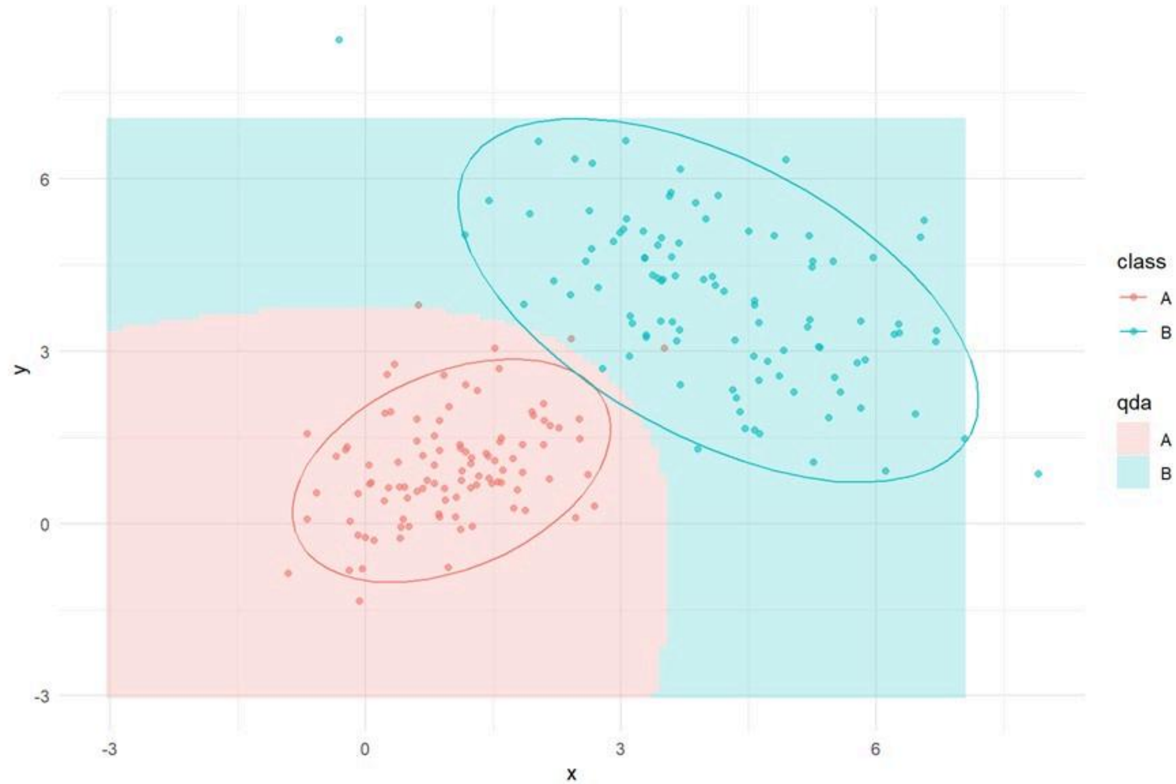
39

```
> confusion_matrix_qda
         Actual
Predicted   A   B
        A  97   0
        B   3 100
> accuracy_qda
[1] 0.985
```

绘制决策边界

```
# 生成网格点
grid <- expand.grid(x = seq(-3, 7, 0.1), y = seq(-3, 7, 0.1))
grid$qda <- predict(model_qda, grid)$class

# 绘制QDA边界
ggplot(data_qda, aes(x, y)) +
  geom_point(aes(color = class), alpha = 0.6) +
  geom_tile(data = grid, aes(fill = qda), alpha = 0.2) +
  stat_ellipse(aes(color = class), level = 0.95) +
  theme_minimal()
```



6
3
0
-3
y
-3
0
3
6
x
class
A
B
qda
A
B

单选题 1分

二次判别分析（QDA）与线性判别分析（LDA）的主要区别是:

A QDA假设类别间的协方差矩阵相同，而LDA假设不同

B QDA假设每个类别的特征遵循独立分布，而LDA假设特征有相同的分布

C QDA允许每个类别有不同的协方差矩阵，而LDA假设所有类别有相同的协方差矩阵

D QDA能够处理大规模数据，而LDA无法

43

单选题 1分

在QDA中，判别函数的推导主要依据以下哪个原理?

A 最大似然估计

B 最小二乘法

C 最大后验概率原则

D 最优特征选择

单选题 1分

在QDA中，判别函数通常是特征的：

A 线性函数

B 指数函数

C 二次函数

D 对数函数

45

单选题 1分

在进行QDA分类时，如果假设每个类别的协方差矩阵相同，会导致：

A 使用QDA时会更慢

B QDA的性能提高

C 分类结果更复杂

D QDA退化为LDA

46

思考

1. QDA与LDA相比，何时更适合使用QDA？

答：类别间协方差矩阵不同；决策边界复杂；数据量较大时。

2. 如何处理QDA模型中可能存在的过拟合问题？

答：通过正则化协方差矩阵、增加训练数据量、降维、交叉验证等方法来降低模型复杂性并提高泛化能力。

本章小结

- 判别分析概述：定义、应用场景、类型。
- 线性判别分析（LDA）及R实现：贝叶斯决策规则、LDA基本原理（假设、目标、优化、分类决策）、R实现（MASS::lda）。
- 二次判别分析（QDA）及R实现：与LDA的区别（基本假设）、贝叶斯定理与后验概率、判别函数的推导、R实现（MASS::qda）。

习题

1. 在糖尿病数据集（如Pima.te）上复现LDA流程，并报告测试集准确率。

2. 在iris数据集上分别运行LDA和QDA，比较准确率。