Final Project Math 7450

Wenjuan Bian

2023-12-09

1. C-reactive protein, ESR (erythrocyte sedimentation rate), and BMI are common factors in the diagnosis of rheumatoid arthritis. Assume that the factors follow normal models with

$$\mu_{\text{disease}} = (2.1, 35, 30), \quad \sigma_{\text{disease}} = (0.3, 5, 2.3)$$

for C-reactive protein, ESR, and BMI respectively. Also assume that correspondingly

$$\mu_{\text{healthy}} = (0.7, 15, 17), \quad \sigma_{\text{healthy}} = (0.3, 5, 2.3)$$

and the covariance matrix is

$$\Sigma = \begin{pmatrix} 0.09 & 0.3 & 0 \\ 0.3 & 25 & 0 \\ 0 & 0 & 5.29 \end{pmatrix}$$

for both diseased and healthy patients. If the disease rate of rheumatoid arthritis is 30%, do the following questions.

1.1 For a 0-1 loss function, derive a MRE for the diagnosis of rheumatoid arthritis on the basis of C-reactive protein, ESR, and BMI.

The 0-1 loss function implies an equal cost of misclassification, whether it's misclassifying a diseased person as healthy or vice versa.

Given:
X - Feature vector
μ_k - Mean vector of class k
Σ - Common covariance matrix for both classes
π_k - Prior probability of class k
The discriminant function for class k is given by:

$$\delta_k(X) = \mu_k^T \Sigma^{-1} X - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k)$$

Where:
μ_k^T - Transpose of the Mean vector
Σ^{-1} - Inverse of the covariance matrix
The decision rule:
Classify X into class k if $\delta_k(X)$ is the largest among all classes.

1.2 If a new patient features (0.9, 23, 18) for C-reactive protein, ESR, and BMI, what is your minimum risk estimator for the disease status of the patient?

```
# Covariance matrix
Sigma <- matrix(c(0.09, 0.3, 0,
                  0.3, 25, 0,
                  0, 0, 5.29),
                nrow = 3, ncol = 3, byrow = TRUE)

Sigma.inv <- solve(Sigma)

# Mean
mu0 <- c(0.7, 15, 17)
mu1 <- c(2.1, 35, 30)

# Data
X <- c(0.9, 23, 18)


delta0 <- t(mu0) %*% Sigma.inv %*% X - 1/2 * t(mu0) %*% Sigma.inv %*% mu0 + log(0.7)
delta1 <- t(mu1) %*% Sigma.inv %*% X - 1/2 * t(mu1) %*% Sigma.inv %*% mu1 + log(0.3)


delta0
```

```
##          [,1]
## [1,] 41.66337
```

```
delta1
```

```
##          [,1]
## [1,] 19.30926
```

Therefore, the new patient is classified as healthy

2. Refer to Question 1. If the covariance matrices for the disease and healthy populations are different:

$$\Sigma_{\text{disease}} = \begin{pmatrix} 0.09 & 0.3 & 0 \\ 0.3 & 25 & 0 \\ 0 & 0 & 5.29 \end{pmatrix}, \quad \Sigma_{\text{healthy}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

2.1 For a 0-1 loss function, derive a MRE for the diagnosis of rheumatoid arthritis on the basis of C-reactive protein, ESR, and BMI.

$$\delta_0(X) = -\frac{1}{2} X^T \Sigma_{\text{healthy}}^{-1} X + X^T \Sigma_{\text{healthy}}^{-1} \mu_0 - \frac{1}{2} \mu_0^T \Sigma_{\text{healthy}}^{-1} \mu_0 + \log(0.7)$$

and:

$$\delta_1(X) = -\frac{1}{2} X^T \Sigma_{\text{disease}}^{-1} X + X^T \Sigma_{\text{disease}}^{-1} \mu_1 - \frac{1}{2} \mu_1^T \Sigma_{\text{disease}}^{-1} \mu_1 + \log(0.3)$$

Decision Making: If $\delta_0(X) > \delta_1(X)$, the patient is classified as 'Healthy'. This implies that, based on the

patient's features, the likelihood of them being healthy is higher than the likelihood of them being diseased. If $\delta_1(X) > \delta_0(X)$, the patient is classified as 'Diseased'. In this case, the patient's features are more consistent with those typically observed in the diseased population.

2.2 If a new patient features (0.9, 23, 18) for C-reactive protein, ESR, and BMI, what is your minimum risk estimator for the disease status of the patient according to the criterion in 2.1)?

```
Sigmah <- matrix(c(1, 0, 0,
                   0, 1, 0,
                   0, 0, 1),
                 nrow = 3, ncol = 3, byrow = TRUE)
Sigmah.inv <- solve(Sigmah)

delta0 <- -1/2*t(X)%*%Sigmah.inv%*%X + t(mu0) %*% Sigmah.inv %*% X - 1/2 * t(mu0) %*% solve(Sigm
ah.inv) %*% mu0 + log(0.7)
delta1 <- -1/2*t(X)%*%Sigma.inv%*%X + t(mu1) %*% Sigma.inv %*% X - 1/2 * t(mu1) %*% Sigma.inv %
*% mu1 + log(0.3)

delta.t <- c(delta0, delta1)
delta.t
```

```
## [1] -32.87667 -24.14789
```

3. Refer to Question 1. If the loss function is changed to the following:

$\text{diagnosis}\backslash\text{true}$	disease	healthy
disease	0	2
healthy	30	0

3.1 Derive a MRE for the diagnosis of rheumatoid arthritis on the basis of C-reactive protein, ESR, and BMI.

The discriminant function for class k is given by:

$$\delta_k(X) = \mu_k^T \Sigma^{-1} X - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \log(\pi_k) + log(loss_k)$$

3.2 If a new patient features (0.9, 23, 18) for C-reactive protein, ESR, and BMI, what is your minimum risk estimator for the disease status of the patient according to the criterion in 3.1)? \end{enumerate}

```
delta0.3 <- t(mu0) %*% solve(Sigma.inv) %*% X - 1/2 * t(mu0) %*% solve(Sigma.inv) %*% mu0 + log
(0.7) +log(2)
delta1.3 <- t(mu1) %*% solve(Sigma.inv) %*% X - 1/2 * t(mu1) %*% solve(Sigma.inv) %*% mu1 + log
(0.3) +log(30)
delta.t3 <- c(delta0.3, delta1.3)
delta.t3
```

```
## [1] 6672.936 5292.659
```

4. Refer to Question 1. Use the covariance matrices in Question 2 and the loss function in Question 3 for the following:

Derive a MRE for the diagnosis of rheumatoid arthritis on the basis of C-reactive protein, ESR, and BMI.

If a new patient features (0.9, 23, 18) for C-reactive protein, ESR, and BMI, what is your minimum risk estimator for the disease status of the patient according to the criterion in 4.1)?

The functions are:

$$\delta_0(X) = -\frac{1}{2}X^T\Sigma_{\text{healthy}}^{-1}X + X^T\Sigma_{\text{healthy}}^{-1}\mu_0 - \frac{1}{2}\mu_0^T\Sigma_{\text{healthy}}^{-1}\mu_0 + \log(0.7) + log(2)$$

and:

$$\delta_1(X) = -\frac{1}{2}X^T\Sigma_{\text{disease}}^{-1}X + X^T\Sigma_{\text{disease}}^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma_{\text{disease}}^{-1}\mu_1 + \log(0.3) + log(30)$$

```
delta0.4 <- t(mu0) %*% solve(Sigmah.inv) %*% X - 1/2 * t(mu0) %*% solve(Sigmah.inv) %*% mu0 + lo
g(0.7) +log(2)
delta1.4 <- t(mu1) %*% solve(Sigma.inv) %*% X - 1/2 * t(mu1) %*% solve(Sigma.inv) %*% mu1 + log
(0.3) + log(30)
delta.t4 <- c(delta0.4, delta1.4)
delta.t4
```

```
## [1]  394.7215 5292.6589
```

5. Use the R-code and variable description in class to describe the dataset Smarket.

Summary of the Dataset (summary(Smarket)):

The dataset is from 2001 to 2005 (Year variable).

Lag1 to Lag5 represent previous days' returns, showing a wide range of values from approximately -4.92 to 5.73.

Volume varies significantly, indicating different levels of trading activity across the days.Today gives the current day's return, similar in distribution to the lag variables. Direction is a factor indicating the market movement, either 'Down' or 'Up'. Analysis of the Dataset (lda.fit and related output):

Linear Discriminant Analysis (LDA) is applied to predict Direction based on Lag1 and Lag2. The LDA model is trained on data before 2004 and tested on 2004 data.

The output includes prior probabilities of each group ('Down' and 'Up'), group means for Lag1 and Lag2, and coefficients of linear discriminants.

The model's predictions for 2004 are evaluated through a confusion matrix.

This analysis provides a comprehensive view of the dataset, including its structure, the nature of its variables, and an initial approach to predicting market direction using statistical techniques.

6. Use the data before year 2004 as training data to build a linear discrimination analysis that predicts the market trend (up or down) in 2004 and 2005. The new LDA uses information in the past three lag times in the stock market data. Compare your LDA results with the two-lag time prediction.

```
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.3.2
```

```
attach(Smarket)
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:ISLR2':
##
##     Boston
```

```
train <- Smarket[Year < 2004,]

lda.fit2 <- lda(Direction ~ Lag1 + Lag2, data = train)

lda.fit3 <- lda(Direction ~ Lag1 + Lag2 + Lag3, data = train)


Smarket.2004_2005 <- Smarket[Year >= 2004,]
Direction.2004_2005 <- Direction[Year >= 2004]


lda.pred2 <- predict(lda.fit2, Smarket.2004_2005)
table2 <- table(lda.pred2$class, Direction.2004_2005)


lda.pred3 <- predict(lda.fit3, Smarket.2004_2005)
table3 <- table(lda.pred3$class, Direction.2004_2005)

# Print tables for comparison
print("Confusion Matrix for Two-Lag Model:")
```

```
## [1] "Confusion Matrix for Two-Lag Model:"
```

```
print(table2)
```

```
##       Direction.2004_2005
##        Down  Up
##   Down  179 224
##   Up     44  57
```

```
print("Confusion Matrix for Three-Lag Model:")
```

```
## [1] "Confusion Matrix for Three-Lag Model:"
```

```
print(table3)
```

```
##       Direction.2004_2005
##        Down  Up
##   Down  181 223
##   Up     42  58
```

The three-lag model shows a marginal improvement over the two-lag model in all considered metrics. However, the improvements are slight, and both models exhibit low precision, suggesting that they might not be highly effective in accurately predicting market trends based on the lag variables alone. The small differences in performance also suggest that the addition of the third lag time (Lag3) provides only a minimal increase in predictive ability over using just two lag times.

7. Refer to Question 6. Assume that the data were in the bear market in which the market has 20% up stocks and 80% down stocks. Theoretically, how to adjust the classification criterion in Question 6? Prove your answer.

The adjustment involves setting the prior probabilities in the LDA model to reflect the actual market condition:

Prior probability for 'Up' class: 0.2 (20%)

Prior probability for 'Down' class: 0.8 (80%)

```
lda.fit4 <- lda(Direction ~ Lag1 + Lag2, data = train, priors = c(0.2, 0.8))

lda.fit5 <- lda(Direction ~ Lag1 + Lag2 + Lag3, data = train, priors = c(0.2, 0.8))


lda.pred4 <- predict(lda.fit4, Smarket.2004_2005)
table4 <- table(lda.pred4$class, Direction.2004_2005)

lda.pred5 <- predict(lda.fit5, Smarket.2004_2005)
table5 <- table(lda.pred5$class, Direction.2004_2005)

print("Confusion Matrix for Two-Lag Model:")
```

```
## [1] "Confusion Matrix for Two-Lag Model:"
```

```
print(table4)
```

```
##       Direction.2004_2005
##        Down  Up
##   Down  179 224
##   Up     44  57
```

```
print("Confusion Matrix for Three-Lag Model:")
```

```
## [1] "Confusion Matrix for Three-Lag Model:"
```

```
print(table5)
```

```
##       Direction.2004_2005
##        Down  Up
##   Down  181 223
##   Up     42  58
```