

15 Klasyfikacja

15.1 Przykład

Przykład. Zbiór danych `iris` zawiera informacje na temat czterech cech trzech gatunków irysa.

```
head(iris)
```

```
## Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1           5.1           3.5           1.4           0.2 setosa
## 2           4.9           3.0           1.4           0.2 setosa
## 3           4.7           3.2           1.3           0.2 setosa
## 4           4.6           3.1           1.5           0.2 setosa
## 5           5.0           3.6           1.4           0.2 setosa
## 6           5.4           3.9           1.7           0.4 setosa
```

```
dim(iris)
```

```
## [1] 150  5
```

```
table(iris$Species)
```

```
##
## setosa versicolor virginica
##      50         50         50
```

Na przykładzie tego zbioru danych przedstawimy liniową analizę dyskryminacyjną (LDA).

- model liniowej analizy dyskryminacyjnej w R

```

library(MASS)

(model_lda <- lda(Species ~ ., data = iris))

## Call:
## lda(Species ~ ., data = iris)
##
## Prior probabilities of groups:
##      setosa versicolor  virginica
## 0.3333333 0.3333333 0.3333333
##
## Group means:
##           Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa           5.006         3.428         1.462         0.246
## versicolor       5.936         2.770         4.260         1.326
## virginica        6.588         2.974         5.552         2.026
##
## Coefficients of linear discriminants:
##           LD1          LD2
## Sepal.Length 0.8293776 0.02410215
## Sepal.Width  1.5344731 2.16452123
## Petal.Length -2.2012117 -0.93192121
## Petal.Width  -2.8104603 2.83918785
##
## Proportion of trace:
##      LD1      LD2
## 0.9912 0.0088

# Lub
# model_lda <- lda(iris[, 1:4], grouping = iris$Species)

```

- tablica kontyngencji

```
head(predict(model_lda)$posterior)
```

```
##      setosa      versicolor      virginica
## 1      1 3.896358e-22 2.611168e-42
## 2      1 7.217970e-18 5.042143e-37
## 3      1 1.463849e-19 4.675932e-39
## 4      1 1.268536e-16 3.566610e-35
## 5      1 1.637387e-22 1.082605e-42
## 6      1 3.883282e-21 4.566540e-40
```

```
head(predict(model_lda)$class)
```

```
## [1] setosa setosa setosa setosa setosa setosa
## Levels: setosa versicolor virginica
```

```
(conf_matrix <- table(predict(model_lda)$class, iris$Species))
```

```
##
##              setosa versicolor virginica
## setosa          50           0           0
## versicolor       0           48           1
## virginica        0           2          49
```

- błąd klasyfikacji metodą ponownego podstawiania

```
(1 - sum(diag(conf_matrix)) / nrow(iris))
```

```
## [1] 0.02
```

- błąd klasyfikacji metodą sprawdzania krzyżowego z $v = 1$ (1-CV, LOO, ang. *leave one out*)

```

pred_loo <- numeric(nrow(iris))
for (i in 1:nrow(iris)) {
  model_lda_i <- lda(Species ~ ., data = iris[-i, ])
  pred_loo[i] <- predict(model_lda_i, iris[i, ])$class
}
table(iris$Species, pred_loo)

```

```

##           pred_loo
##           1  2  3
## setosa      50  0  0
## versicolor  0 48  2
## virginica   0  1 49

```

```

(1 - sum(diag(table(iris$Species, pred_loo)))) / nrow(iris))

```

```

## [1] 0.02

```

- predykcja

```

new_data <- data.frame(Sepal.Length = 5.1,
                      Sepal.Width = 3.5,
                      Petal.Length = 1.3,
                      Petal.Width = 0.3)
predict(model_lda, new_data)

```

```
## $class
## [1] setosa
## Levels: setosa versicolor virginica
##
## $posterior
##      setosa      versicolor      virginica
## 1      1 4.850575e-22 6.605032e-42
##
## $x
##      LD1      LD2
## 1 8.000875 0.6775315
```

15.2 Zadania

Zadanie 1. Kontynuujemy przykład dotyczący zbioru danych `iris`.

1. Wyznacz błąd klasyfikacji liniowej analizy dyskryminacyjnej metodą sprawdzania krzyżowego z $v = 10$ (10-CV).

```
## [1] 0.02
```

2. Błąd klasyfikacji można oszacować również następującą metodą bootstrapową.

- Przyjmijmy, że zbiór danych ma n obserwacji.
- Krok 1. Losujemy ze zwracaniem n obserwacji ze zbioru danych tworzących próbę bootstrapową.
- Krok 2. Konstruujemy klasyfikator na bazie próby bootstrapowej.
- Krok 3. Liczymy błąd klasyfikatora wyznaczonego w kroku 2 dla obserwacji, które nie znalazły się w próbie bootstrapowej.
- Krok 4. Powtarzamy kroki 1-3 n_boot razy, otrzymując błędy b_1, \dots, b_{n_boot} .
- Krok 5. Obliczamy błąd klasyfikacji metodą bootstrapową według wzoru

$$\frac{1}{n_boot} \sum_{i=1}^{n_boot} b_i.$$

Wyznacz błąd klasyfikacji liniowej analizy dyskryminacyjnej metodą bootstrapową. Przyjmij $n_boot = 100$.

```
## [1] 0.0259815
```

Zadanie 2. W pliku `wina.txt` zawarto informację o trzynastu cechach różnych gatunków win. Co więcej obserwacje podzielone są na trzy grupy.

```
##      V1   V2   V3   V4   V5   V6   V7   V8   V9  V10  V11  V12  V13 V14
## 1 14.23 1.71 2.43 15.6 127 2.80 3.06 0.28 2.29 5.64 1.04 3.92 1065  1
## 2 13.20 1.78 2.14 11.2 100 2.65 2.76 0.26 1.28 4.38 1.05 3.40 1050  1
## 3 13.16 2.36 2.67 18.6 101 2.80 3.24 0.30 2.81 5.68 1.03 3.17 1185  1
## 4 14.37 1.95 2.50 16.8 113 3.85 3.49 0.24 2.18 7.80 0.86 3.45 1480  1
## 5 13.24 2.59 2.87 21.0 118 2.80 2.69 0.39 1.82 4.32 1.04 2.93  735  1
## 6 14.20 1.76 2.45 15.2 112 3.27 3.39 0.34 1.97 6.75 1.05 2.85 1450  1
```

```
## ...
```

1. Jaki jest wymiar tych danych? Jakie są etykiety klas i ich liczebności?

```
## [1] 178  14
```

```
##
```

```
##  1  2  3
```

```
## 59 71 48
```

2. Wykonaj liniową analizę dyskryminacyjną bazując na trzech pierwszych zmiennych w tym zbiorze danych.

```
##      1      2      3
```

```
## 0.3314607 0.3988764 0.2696629
```

```
##      V1      V2      V3
```

```
## 1 13.74475 2.010678 2.455593
```

```
## 2 12.27873 1.932676 2.244789
```

```
## 3 13.15375 3.333750 2.437083
```

```
##          LD1          LD2
## V1 -1.8725417 -0.2943580
## V2 -0.0862327  1.0473192
## V3 -1.4493443  0.1419408
```

3. Wyznacz oceny prawdopodobieństw a posteriori i przewidywaną przynależność do klas obserwacji oraz tablicę kontyngencji otrzymanego klasyfikatora.

```
##          1          2          3
## 1 0.9705550 0.0006735689 0.02877140
## 2 0.3933512 0.3924750849 0.21417373
## 3 0.5316537 0.0682685490 0.40007778
## 4 0.9723331 0.0002235964 0.02744332
## 5 0.5798070 0.0197639349 0.40042907
## 6 0.9668517 0.0007345077 0.03241381
```

```
## [1] 1 1 1 1 1 1
## Levels: 1 2 3
```

```
##
##      1  2  3
## 1 51  5  7
## 2  4 62  8
## 3  4  4 33
```

4. Wyznacz błąd klasyfikacji metodą ponownego podstawiania.

```
## [1] 0.1797753
```

5. Wyznacz błąd klasyfikacji metodą sprawdzania krzyżowego z $v = 1$.

```
##      pred_loo
##      1  2  3
##    1 49  5  5
##    2  5 61  5
##    3 10  8 30
```

```
## [1] 0.2134831
```

6. Wyznacz błąd klasyfikacji metodą sprawdzania krzyżowego z $v = 10$.

```
## [1] 0.2078652
```

7. Wyznacz błąd klasyfikacji metodą bootstrapową. Przyjmij $n_{boot} = 100$.

```
## [1] 0.2111531
```

8. Do których klas i z jakimi prawdopodobieństwami a posteriori należy zaklasyfikować poniższe nowe obserwacje?

V1	V2	V3
13.64	3.10	2.56
13.94	1.73	2.27
13.08	3.90	2.36
12.29	3.17	2.21


```

## $class
## [1] 1 1 3 2
## Levels: 1 2 3
##
## $posterior
##           1           2           3
## 1 0.531302523 0.007133455 0.46156402
## 2 0.924346812 0.007006399 0.06864679
## 3 0.061216479 0.054434582 0.88434894
## 4 0.005015639 0.810915785 0.18406858
##
## $x
##           LD1           LD2
## 1 -1.5435449  0.6390430
## 2 -1.5668588 -0.9252545
## 3 -0.2740389  1.6133507
## 4  1.4856206  1.0600594

```