Analisis Diskriminan

gdito

Note: output dari R pada dokumen ini diawali dengan tanda ##

Package

Pada Praktikum kali ini package yang dibutuhkan adalah

- rattle
- MASS (sudah otomatis ada di R)
- caret
- heplotss
- MVN

Silahkan install jika belum ada

```
install.packages("rattle")
install.packages("caret")
install.packages("MVN")
install.packages("heplots")
```

```
library(MASS)
library(MVN)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## sROC 0.1-2 loaded
```

```
library(heplots)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

Tahap analisis diskriminan

1. Membagi data menjadi data training dan testing

- pembagian data dilakukan dengan pengambilan acak

- data training biasanya berisi 70% atau 80% jumlah amatan dari data asal. Misalnya data asal memliki 100 amatan, maka data trainingnya bisa memiliki 70 amatan atau 30 amtan.
- data training digunakan untuk pemodelan
- data testing digunakan untuk menguji kemampuan klasifikasi model untuk data baru.

2. Dengan menggunakan data training lakukan langkah-langkah berikut:

a. uji normal ganda
b. Uji asumsi kesamaan ragam

jika uji ini menyimpulkan bahwa matriks ragam-peragam sama maka gunakan Linear Discriminant (LDA), jika kesimpulanya sebaliknya maka gunakan Quadratic Discriminant Analysis (QDA).

Note : Menurut Mattjik dan Sumertajaya pada buku sidik peubah ganda, "umumnya sangat sulit sekali untuk dapat memenuhi persyaratan (a) dan (b), yang dalam praktek sering kali tidak diuji; hal mana akan membuat akurasi dari analisis dengan fungsi diskriminan akan berkurang. Namun demikian, fungsi diskriminan selalu menghasilkan estimasi yang kokoh (robust estimates) terutama yang berkaitan dengan prediksi pengelompokan".

c. Estimasi Koefisien analisis diskriminan
d. Evaluasi kemampuan klasifikasi analisis diskriminan

3. Evaluasi kemampuan klasifikasi menggunakan data testing

Data Wine

The wine dataset contains the results of a chemical analysis of wines grown in a specific area of Italy. Three types of wine are represented in the 178 samples, with the results of 13 chemical analyses recorded for each sample. The Type variable has been transformed into a categoric variable.

Menyiapkan data di R

```
data(wine, package='rattle')
head(wine)
```

```
##   Type Alcohol Malic  Ash Alcalinity Magnesium Phenols Flavanoids
## 1    1   14.23  1.71 2.43       15.6       127    2.80       3.06
## 2    1   13.20  1.78 2.14       11.2       100    2.65       2.76
## 3    1   13.16  2.36 2.67       18.6       101    2.80       3.24
## 4    1   14.37  1.95 2.50       16.8       113    3.85       3.49
## 5    1   13.24  2.59 2.87       21.0       118    2.80       2.69
## 6    1   14.20  1.76 2.45       15.2       112    3.27       3.39
##   Nonflavanoids Proanthocyanins Color  Hue Dilution Proline
## 1          0.28            2.29  5.64 1.04     3.92    1065
## 2          0.26            1.28  4.38 1.05     3.40    1050
## 3          0.30            2.81  5.68 1.03     3.17    1185
## 4          0.24            2.18  7.80 0.86     3.45    1480
## 5          0.39            1.82  4.32 1.04     2.93     735
## 6          0.34            1.97  6.75 1.05     2.85    1450
```

analisis data di R

1. Membagi data menjadi data training dan testing

pembagian data dapat dilakukan dengan menggunakan fungsi `createDataPartition` dari pacakge `caret`. Sintaks `caret::createDataPartition` berarti kita memanggil fungsi `createDataPartition` dari `caret` tanpa perlu memanggil package `caret` menggunakan `library(caret)` . Argumen `y` merupakan peubah respon/ gerombol, `p` merupakan proporsi data training (dalam hal ini 0.7 atau 70%), `list=FALSE` berati hasil output dari `createDataPartition` disimpan dalam bentuk vektor (defaultnya `list=TRUE`).

```
set.seed(123)
index_train <- caret::createDataPartition(y = wine$Type,p = 0.7,list = FALSE)
wine_train <- wine[index_train,]
wine_test <- wine[-index_train,]
```

index_train berisi vektor dari urutan amatan $(1, 2, ..., dst)$ yang telah dilakukan pengambilan contoh secara acak.

2. Dengan menggunakan data training lakukan langkah-langkah berikut:

a. Uji Normal ganda

H_0 : data menyebar normal ganda H_1 : data tidak menyebar normal ganda

Untuk menguji kenormalan ganda di R, bisa menggunakan fungsi `mvn` dari package `MVN`. fungsi mvn memiliki beberapa uji normal ganda yang bisa dilakukan. Pemilihan uji normal ganda bisa dilakukan melauli argumen mvnTest, seperti uji Mardia (`mvnTest="mardia"`), uji Henze-Zirkler(`mvnTest="hz"`), uji Royston (`mvnTest="royston"`),uji Doornik-Hansen (`mvnTest="dh"`) dan uji energy `mvnTest="energy"`. Argumen `subset` diisi dengan kolom data yang menyatakan gerombol.

```
uji_normalGanda <- mvn(data = wine_train,subset="Type",mvnTest = "hz")
uji_normalGanda$multivariateNormality
```

```
## $`1`
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.9917837 0.1761976 YES
##
## $`2`
##            Test       HZ      p value MVN
## 1 Henze-Zirkler 1.008078 0.0003652993  NO
##
## $`3`
##            Test        HZ   p value MVN
## 1 Henze-Zirkler 0.9909895 0.1661439 YES
```

Karena nilai dari p-value dari `$1dan `$`3` adalah 0.2604207 dan 0.3244687, yang mana lebih besar dari nilai α 0.05 maka dapat disimpulkan bahwa tidak cukup bukti untuk menolak H_0. Artinya untuk peubah-puebah penjelas pada wine tipe 1 dan tipe 3 berdistribusi normal ganda. Sementera itu, p-value dari `$2`“ sangat kecil yaitu 3.005905e-05 yang mana lebih kecil dari nilai α 0.05. Artinya peubah-puebah penjelas pada wine tipe 2 tidak berdistribusi normal ganda.

Note: jika salah satu uji normal ganda menyatakan tolak H_0 maka perlu dicoba uji kenormalan yang lain, karena berpotensi hasil dari uji normal ganda lainnya menghasilkan kesimpulan yang berbeda.

** Walaupun ada satu gerombol yang tidak memenuhi asumsi normal ganda kita akan tetap lanjutkan menggunakan analisis diskriminan karena berdasarkan Matjik dan Sumertajaya fungsi diskriminan masih dapat menghasilkan kemampuan klasifikasi yang baik.

b. Uji asumsi kesamaan ragam

Hipotesis asumsi kesamaan ragam H_0 : ragam antar populasi sama H_1 : ragam antar populasi tidak sama

Uji kesamaan ragam bisa dilakukan dengan menggunkan fungsi `boxM` yang berasal dari package `heplots`. Fungsi ini hanya membutuhkan 2 argumen, yaitu data dalam bentuk `data.frame` atau `matrix` tanpa kolom gerombol (dalam hal ini kolom **Type**) dan vektor gerombol yang diperoleh dari data (dalam hal ini kolom **Type**).

```
boxM(wine_train[,-1],wine_train$Type)
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  wine_train[, -1]
## Chi-Sq (approx.) = 502.42, df = 182, p-value < 2.2e-16
```

Karena nilai dari p-value dari uji Box's M adalah kurang dari 2.2e-16, yang mana lebih kecil dari nilai α 0.05 maka dapat disimpulkan bahwa cukup bukti untuk menolak H_0. Artinya untuk peubah-puebah penjelas pada wine memiliki ragam yang tidak sama. Hal ini berarti model yang lebih cocok digunakan adalah QDA.

Untuk ilustrasi kita akan menggunakan LDA juga karena LDA juga berpotensi mengungguli QDA dalam hal kemampuan klasifikasi berdasarkan argumen dari Matjik dan Sumertajaya.

c. Estimasi Koefisien analisis diskriminan

Estimasi LDA dan QDA dapat dilakukan dengan menggunakan fungsi `lda` dan `qda` dari package `MASS`. Argumen minimum yang dibutuhkan oleh kedua fungsi tersebut adalah `formula` dan `data`. Argumen `formula` berisi tentang rumus model yang digunakan tanpa coefisien, `Type~.` berarti kolom Type menjadi peubah respon dan `.` menandakan memakai semua kolom kecuali kolom Type sebagai peubah penjelas. Argumen data berisi tentang data yang kita gunakan.

```
# LDA
wine_lda <- lda(Type~.,data = wine_train)
coef(wine_lda)
```

```
##                         LD1          LD2
## Alcohol         -0.483076356  0.854816559
## Malic            0.224630564  0.354730236
## Ash             -0.672781084  2.639186194
## Alcalinity       0.152999471 -0.168693427
## Magnesium       -0.001907630 -0.004035584
## Phenols          0.705495150  0.523268740
## Flavanoids      -1.735333782 -0.868717551
## Nonflavanoids   -0.760204649 -1.589205566
## Proanthocyanins  0.092670993 -0.365830979
## Color            0.410568456  0.277565083
## Hue             -0.089314921 -1.096511470
## Dilution        -0.882807209  0.087609503
## Proline         -0.002645145  0.002808837
```

Karena terdapat tiga gerombol pada data wine maka fungsi diskriminan yang terbentuk sebanyak dua. Secara umum fungsi diskriminan yang terbentuk dari g gerombol adalah $g-1$ gerombol.

Fungsi diskriminan pertama dapat ditulis

$$D_1 = -0.664359840 * Alcohol + 0.114575321 * Malic - 0.949984960 * Ash +$$

$$0.174421943 * Alcalinity - 0.005920473 * Magnesium + 0.340435417 * Phenols -$$

$$1.324937540 * Flavanoids - 1.205951463 * Nonflavanoids + 0.134118792 * Proanthocyanins +$$

$$0.347847475 * Color - 0.403983767 * Hue - 1.116033719 * Dilution - 0.003384957 * Proline$$

Fungsi diskriminan kedua dapat ditulis

$$D_2 = 1.07856939515304 * Alcohol + 0.272532494819719 * Malic + 2.74929435419577 * Ash$$

$$-0.109652670989459 * Alcalinity + 0.00536435585977734 * Magnesium -$$

$$0.0403549073529205 * Phenols - 0.812515733054733 * Flavanoids -$$

$$2.31726701046428 * Nonflavanoids - 0.22853350687638 * Proanthocyanins +$$

$$0.281988543029379 * Color - 2.57998408336665 * Hue + 0.1981034 * Dilution + 0.002545 * Proline$$

```
# QDA
wine_qda <- qda(Type~.,data = wine_train)
coef(wine_qda)
```

```
## NULL
```

Berbeda dengan LDA, QDA tidak memiliki koefisien yang bisa ditampilkan.

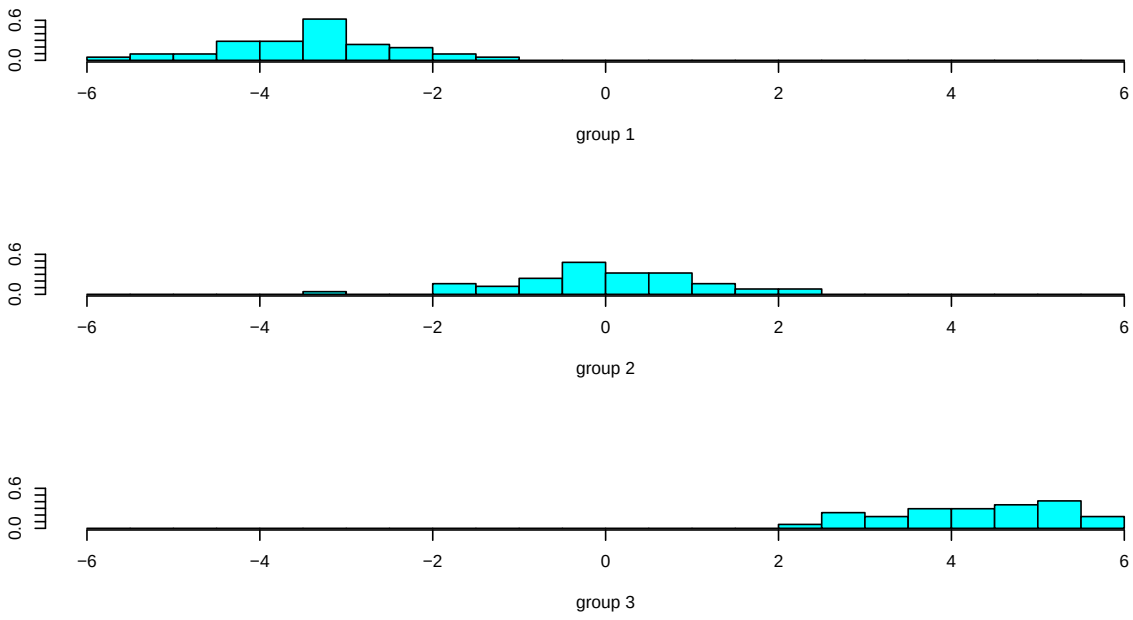
d. Evaluasi kemampuan klasifikasi analisis diskriminan

Sebelum kita mengevaluasi kedua model, maka kita akan mengekstrak prediksi gerombol yang dihasilkan oleh kedua model.

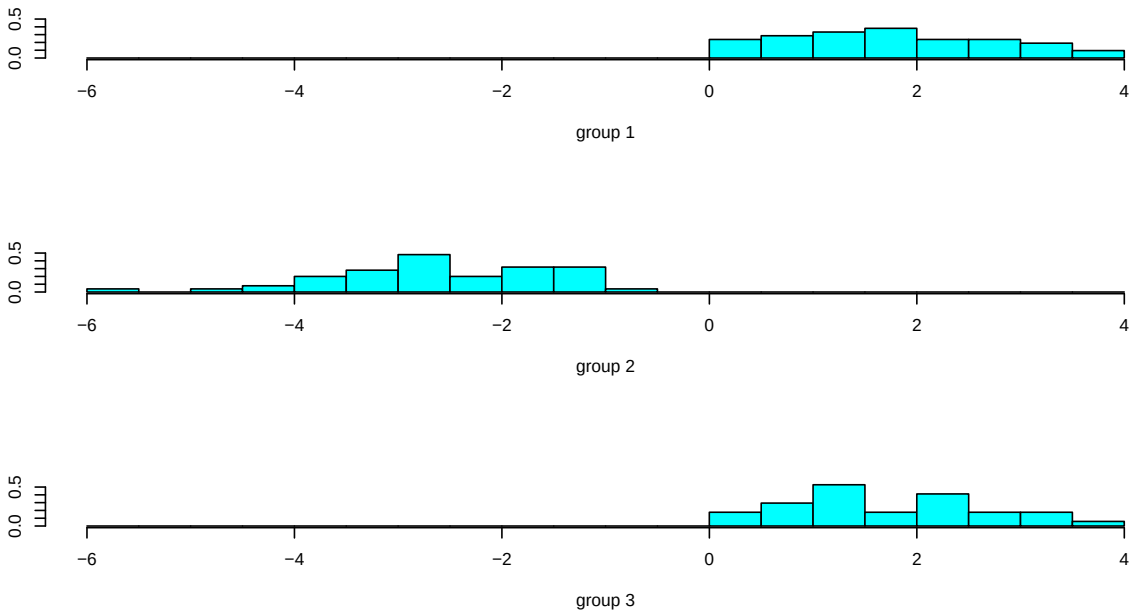
```
predict_lda <- predict(wine_lda)

predict_qda <- predict(wine_qda)
```

- Menggunakan histogram (khusus lda)

```
predict_lda <- predict(wine_lda)

predict_qda <- predict(wine_qda)

# first discriminat function
ldahist(predict_lda$x[,1],g = wine_train$Type)
```




```
# second discriminat function
ldahist(predict_lda$x[,2],g = wine_train$Type)
```



Histogram first discriminant function, digunakan untuk melihat kemampuan fungsi diskriminan yang pertama untuk membedakan ketiga gerombol. Karena histogram pertama dan kedua hanya beririsan di dipinggirnya saja, maka dapat dikatakan bahwa fungsi diskriminan pertama cukup baik untuk membedakan gerombol 1 dan gerombol 2. Sedangkan histogram pertama dan ketiga tidak beririsan sama sekali, yang berarti fungsi diskriminan pertama memiliki kemampuan membedakan gerombol 1 dan gerombol 3 dengan

sangat baik. Disisi lain histogram 2 dan histogram 3 relatif besar irisanya sehingga bisa dikatakan fungsi diskriminan pertama tidak terlalu baik dalam membedakan kelompok 2 dan kelompok 3.

Histogram second discriminant function memiliki interpretasi yang mirip seperti yang diatas. Berdasrkan histogram ini, fungsi discriminan tidak mampu membedakan gerombol 1 dan gerombol 3.

- Tabel Klasfikasi dan tingkat kesalahan klasifikasi

Tabel klasifikasi bisa dimunculkan dengan menggunakan fungsi `table`, yang argumen pertamanya merupakan gerombol asli dan argumen keduanya merupakan gerombol hasil prediksi.

```
# Tabel kasifikasi

# LDA
table(wine_train$Type,predict_lda$class)
```

```
##
##      1  2  3
##   1 42  0  0
##   2  0 50  0
##   3  0  0 34
```

```
# QDA
table(wine_train$Type,predict_lda$class)
```

```
##
##      1  2  3
##   1 42  0  0
##   2  0 50  0
##   3  0  0 34
```

Angka 1 2 dan 3 pada baris pertama melambangkan gerombol asli dan kolom pertama yang berisi 1,2 dan 3 melambangkan gerombol hasil perdiksi. Contoh membaca tabel ini adalah sebagai berikut: misalnya saja banyaknya gerombol 1 yang terprediksi sebagai gerombol 1 juga adalah 42, banyaknya gerombol 2 yang terprediksi gerombol 3 adalah 0.

Berdasarkan kedua tabel klasifikasi ini bisa dilihat bahwa lda dan qda tidak memiliki kesalahan dalam memprediksi ketiga gerombol tersebut. Hal ini ditunjukkan dengan hanya diagonal tabel saja yang berisi nilai.

Tingkat kesalahan klasifikasi dihitung dengan menjumlahkan berapa banyak kesalahan prediksi gerombol yang dilakukan oleh model dibagi dengan banyknya amatan data.

```
# tingkat kesalahan klasifikasi

#LDA
sum(wine_train$Type!=predict_lda$class)/length(predict_lda$class)
```

```
## [1] 0
```

```
#QDA
sum(wine_train$Type!=predict_qda$class)/length(predict_qda$class)
```

```
## [1] 0
```

Hasil tingkat kesalahan klasifikasi 0, berarti model lda mampu memprediksi gerombol untuk semua amatan dengan benar.

Berdasarkan langkah 2 ini, dapat disimpulkan bahwa kemampuan klasifikasi lda dan qda sama walaupun menurut uji kesamaan ragam model yang lebih cocok adalah model qda.

3. Evaluasi kemampuan klasifikasi menggunakan data testing

Pada tahap terakhir ini akan dilakukan evaluasi kemampuan klasifikasi jika seandainya terdapat data baru yang tidak terlibat dalam proses pemodelan.

Sebelum kita mengevaluasi kedua model, maka kita akan mengekstrak prediksi gerombol yang dihasilkan oleh kedua model pada data baru ini. Argumen `newdata` disi dengan data baru.

```
predict_lda_test <- predict(wine_lda,newdata = wine_test)
predict_qda_test <- predict(wine_qda,newdata = wine_test)
```

```
# Tabel kasifikasi

#LDA
table(wine_test$Type,predict_lda_test$class)
```

```
##
##      1  2  3
##   1 17  0  0
##   2  0 21  0
##   3  0  1 13
```

```
#QDA
table(wine_test$Type,predict_qda_test$class)
```

```
##
##      1  2  3
##   1 17  0  0
##   2  0 21  0
##   3  0  0 14
```

Berdasarkan tabel klasifikasi diatas, terlihat bahwa model QDA memiliki kesalahan prediksi gerombol. Kesahalah prediksi ini terjadi pada gerombol 3, dimana prediksi gerombolnya dua.

```
# tingkat kesalahan klasifikasi

#LDA
sum(wine_test$Type!=predict_lda_test$class)/length(predict_lda_test$class)
```

```
## [1] 0.01923077
```

```
#QDA
sum(wine_test$Type!=predict_qda_test$class)/length(predict_qda_test$class)
```

```
## [1] 0
```

Berdasarkan tingkat kesalahan klasifikasi model LDA memiliki nilai yang lebih kecil sehingga dapat dikatakan model lda lebih baik daripada model QDA.

Note Dalam praktiknya data baru yang dimaksud belum memiliki gerombol asli, sehingga tidak memungkinkan untuk dilakukan evaluasi.