

## Διακρίνουσα ανάλυση

Λουκία Μελικοτσίδου

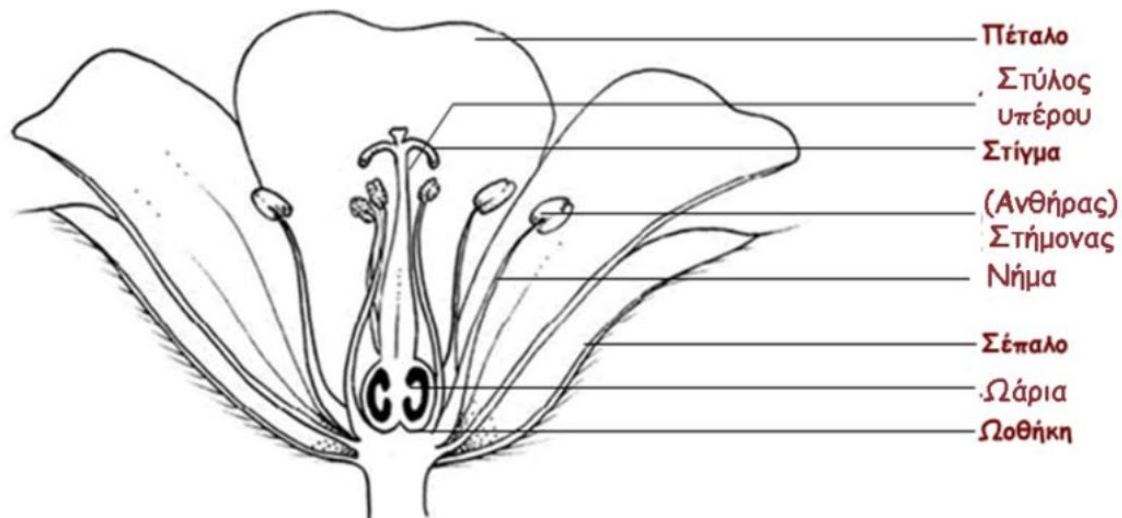
### Εισαγωγή

Η γραμμική διακρίνουσα ανάλυση, αρχικά προτάθηκε από τον Fisher το 1396 για να κατατάξει άτομα σε μια από 2 ξεκάθαρα ορισμένες ομάδες. Αργότερα επεκτάθηκε σε περισσότερες από δύο ομάδες (πληθυσμούς). Βασικά, χρησιμοποιεί έναν γραμμικό συνδυασμό από κάποιες μεταβλητές ώστε να επιτευχθεί ο καλύτερος δυνατός διαχωρισμός των ομάδων.

### Iris dataset

Σαν παράδειγμα, θα χρησιμοποιήσουμε τα δεδομένα που χρησιμοποίησε ο Fisher για 50 λουλούδια από 3 διαφορετικά είδη. Τα είδη είναι *Iris setosa*, *versicolor*, και *virginica*. Για κάθε είδος έχουμε μετρήσεις από τις 4 ακόλουθες μεταβλητές

- Sepal.Length (μήκος σεφάλου)
- Sepal.Width (πλάτος σεφάλου)
- Petal.Length (μήκος πετάλου)
- Petal.Width (πλάτος πετάλου)



Φορτώνουμε πρώτα τις βιβλιοθήκες που θα χρειαστούν

```
# Load libraries  
library(MASS)
```

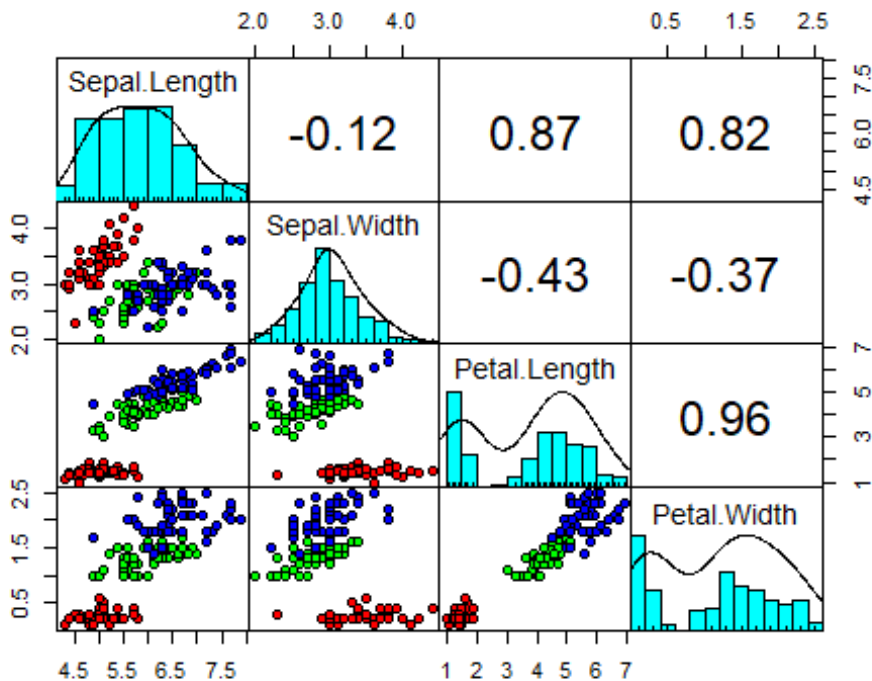
```
library(psych)
library(klaR)
```

Εισάγουμε τα δεδομένα `iris` ως `data.frame` στην R.

```
# Load data
data(iris)
```

Η συνάρτηση `pairs.panels` της βιβλιοθήκης `psych` αποτελεί μια γενίκευση της εντολής `pairs` της R που μας βοηθά να δημιουργήσουμε έναν πίνακα από διαγράμματα σημείων (`scatterplot`) μεταξύ των μεταβλητών. Δηλώνουμε τις ποσοτικές μεταβλητές που θέλουμε να εξετάσουμε `iris[1:4]`, η επιλογή `gap = 0` δηλώνει ότι δεν θέλουμε κενό μεταξύ των `scatterplots`, οι επιλογές `smooth = F` και `ellipses = F` δηλώνουν ότι δεν θέλουμε να εμφανισούν καμπύλες που να εκτιμούν την σχέση μεταξύ των δύο μεταβλητών και ελλείψεις που να δηλώνουν τη συσχέτιση μεταξύ των δύο μεταβλητών, ενώ η επιλογή `bg = c("red", "green", "blue")[iris$Species]` δηλώνει ότι θέλουμε διαφορετικό χρώμα για κάθε είδος λουλουδιών.

```
# Matrix plot
pairs.panels(iris[1:4],
             gap = 0, smooth = F, ellipses = F,
             bg = c("red", "green", "blue")[iris$Species],
             pch = 21)
```



`c("red", "green", "blue")[iris$Species]` είναι ένα character vector με διαφορετικό χρώμα για κάθε είδος

```
cbind(iris$Species,c("red", "green", "blue")[iris$Species])[1:60,]
```

```
##      [,1] [,2]
## [1,] "1"  "red"
## [2,] "1"  "red"
## [3,] "1"  "red"
## [4,] "1"  "red"
## [5,] "1"  "red"
## [6,] "1"  "red"
## [7,] "1"  "red"
## [8,] "1"  "red"
## [9,] "1"  "red"
## [10,] "1" "red"
## [11,] "1" "red"
## [12,] "1" "red"
## [13,] "1" "red"
## [14,] "1" "red"
## [15,] "1" "red"
## [16,] "1" "red"
## [17,] "1" "red"
## [18,] "1" "red"
## [19,] "1" "red"
## [20,] "1" "red"
## [21,] "1" "red"
## [22,] "1" "red"
## [23,] "1" "red"
## [24,] "1" "red"
## [25,] "1" "red"
## [26,] "1" "red"
## [27,] "1" "red"
## [28,] "1" "red"
## [29,] "1" "red"
## [30,] "1" "red"
## [31,] "1" "red"
## [32,] "1" "red"
## [33,] "1" "red"
## [34,] "1" "red"
## [35,] "1" "red"
## [36,] "1" "red"
## [37,] "1" "red"
## [38,] "1" "red"
## [39,] "1" "red"
## [40,] "1" "red"
## [41,] "1" "red"
## [42,] "1" "red"
## [43,] "1" "red"
## [44,] "1" "red"
## [45,] "1" "red"
## [46,] "1" "red"
## [47,] "1" "red"
```

```
## [48,] "1" "red"
## [49,] "1" "red"
## [50,] "1" "red"
## [51,] "2" "green"
## [52,] "2" "green"
## [53,] "2" "green"
## [54,] "2" "green"
## [55,] "2" "green"
## [56,] "2" "green"
## [57,] "2" "green"
## [58,] "2" "green"
## [59,] "2" "green"
## [60,] "2" "green"
```

Εδώ εστιάζουμε κυρίως στο ότι στις περισσότερες περιπτώσεις οι μεταβλητές φαίνεται να μπορούν να διαχωρίσουν τις 3 ομάδες.

## Εκπαιδευτικό (training) και επικυρωτικό (validation) δείγμα

Χωρίζουμε το συνολικό δείγμα σε εκπαιδευτικό και επικυρωτικό. Το εκπαιδευτικό δείγμα χρησιμοποιείται για να εκπαιδευτεί ο αλγοριθμός (να δημιουργηθούν δηλαδή οι κανόνες ταξινόμησης) και το επικυρωτικό για να αξιολογήσουμε την απόδοση του μοντέλου όσον αφορά την ταξινόμηση νέων παρατηρήσεων (όπου το είδος του λουλουδιού είναι **άγνωστο**) Αυτό συμβαίνει διότι για να εξετάσουμε την απόδοση του μοντέλου δεν μπορούμε να χρησιμοποιήσουμε μόνο ένα εκπαιδευτικό δείγμα - έχει αποδειχθεί ότι μια τέτοια αξιολόγηση θα οδηγούσε σε πιο αισιόδοξα αποτελέσματα επειδή θα χρησιμοποιούσαμε τα ίδια δεδομένα για εκπαίδευση και επικύρωση!

```
set.seed(123)
ind <- sample(2, nrow(iris),
              replace = TRUE,
              prob = c(0.6, 0.4))
training <- iris[ind==1,]
testing <- iris[ind==2,]
```

Η εντολή `sample` κάνει δειγματοληψία με/χωρίς επανάθεση. Στην συγκεκριμένη περίπτωση ουσιαστικά ζητάμε να δειγματολοπήσουμε μεταξύ των αριθμών 1 (εκπαίδευση) και 2 (επικύρωση) με πιθανότητες 60% και 40%, αντίστοιχα. Το αποτέλεσμα θέλουμε να είναι `vector` με μήκος όσες οι γραμμές του `data.frame` των συνολικών δεδομένων (`nrow(iris)`). Το `data.frame training` είναι το εκπαιδευτικό δείγμα ενώ το `testing` το επικυρωτικό.

## Γραμμική διακρίνουσα ανάλυση

Θα χρησιμοποιήσουμε την εντολή `lda` της βιβλιοθήκης `MASS`. Δηλώνουμε το μοντέλο μέσω φόρμουλας (`formula`), όπως συνήθως στην R. Στο αριστερό μέρος της φόρμουλας δηλώνουμε την κατηγορική μεταβλητή (το είδος του λουλουδιού) ενώ στο δεξί βάζουμε τις μεταβλητές που θέλουμε να χρησιμοποιήσουμε για να

προβλέψουμε το είδος του λουλουδιού (χωρισμένες με +). Η κατηγορική μεταβλητή που δηλώνει την ομάδα (είδος λουλουδιού) ξεχωρίζει από τις υπόλοιπες μέσω του ~

```
# Linear LDA
linear <- lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length +
Petal.Width, data = training,
              prior = c(1/3,1/3,1/3))
linear

## Call:
## lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length +
Petal.Width,
##      data = training, prior = c(1/3, 1/3, 1/3))
##
## Prior probabilities of groups:
##      setosa versicolor virginica
## 0.3333333 0.3333333 0.3333333
##
## Group means:
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## setosa              4.946667      3.380000      1.443333      0.250000
## versicolor          5.943333      2.803333      4.240000      1.316667
## virginica           6.527586      2.920690      5.489655      2.048276
##
## Coefficients of linear discriminants:
##              LD1          LD2
## Sepal.Length  0.3628515  0.05249291
## Sepal.Width   2.2263073  1.47790102
## Petal.Length -1.7839369 -1.61086637
## Petal.Width  -3.9784152  4.10159748
##
## Proportion of trace:
##      LD1      LD2
## 0.9932 0.0068
```

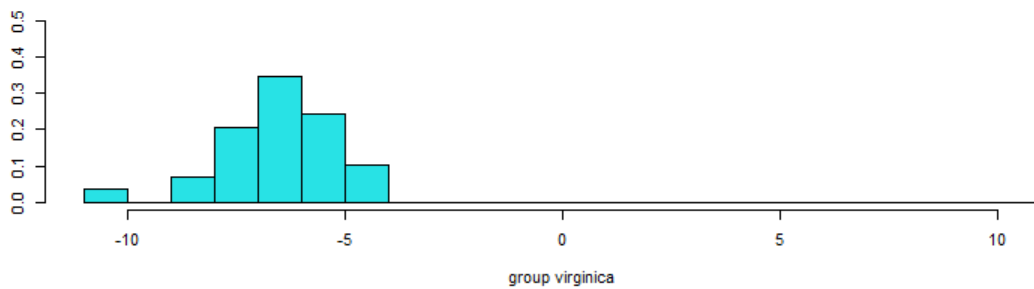
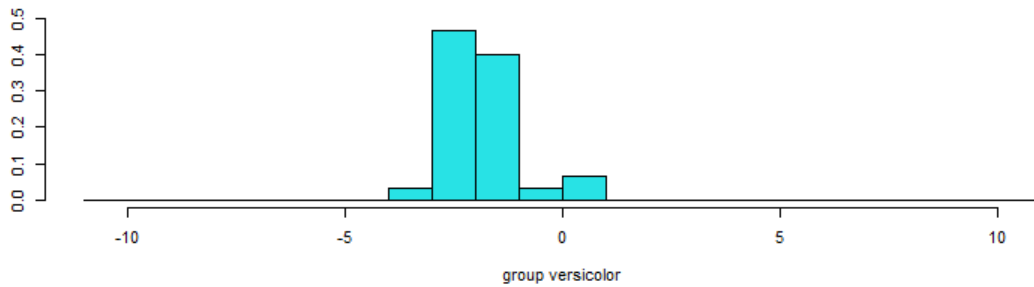
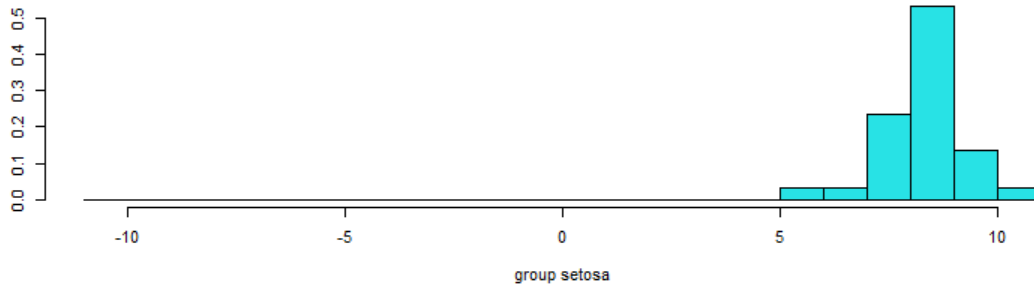
Το data.frame με τις μεταβλητές δηλώνεται με την επιλογή data = training, ενώ επίσης δηλώνουμε ότι θέλουμε ίσες εκ των προτέρων πιθανότητες για τις 3 ομάδες μέσω του prior = c(1/3,1/3,1/3). Οι μέσες τιμές των μεταβλητών δηλώνονται κάτω από το Group means:, όπου βλέπουμε αρκετή διαφοροποίηση των μέσων τιμών ανά ομάδα. Οι συντελεστές των διακρίνουσων συναρτήσεων φαίνονται κάτω από το Coefficients of linear discriminants:. Η πρώτη διακρίνουσα εξηγεί το 99.3% της διαφοροποίησης μεταξύ των ομάδων.

Είναι σημαντικό να εξετάσουμε τα ιστογράμματα των διακρίνουσων συναρτήσεων επειδή οι τιμές των διακρίνουσων συναρτήσεων χρησιμοποιούνται πρακτικά για την κατάταξη των παρατηρήσεων. Η εντολή predict χρησιμοποιείται γενικά για να κάνουμε προβλέψεις με βάση ένα μοντέλο lda που έχει ήδη τρέξει (για περισσότερες λεπτομέρειες ?predict.lda)

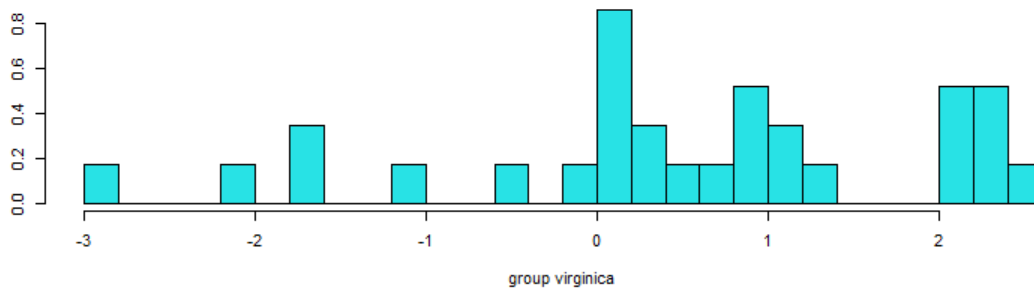
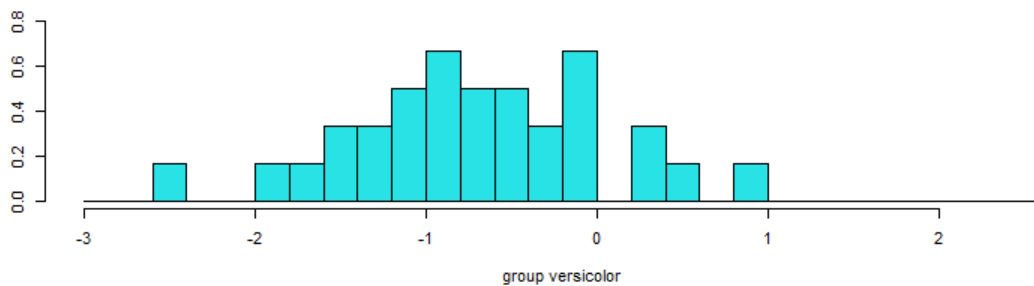
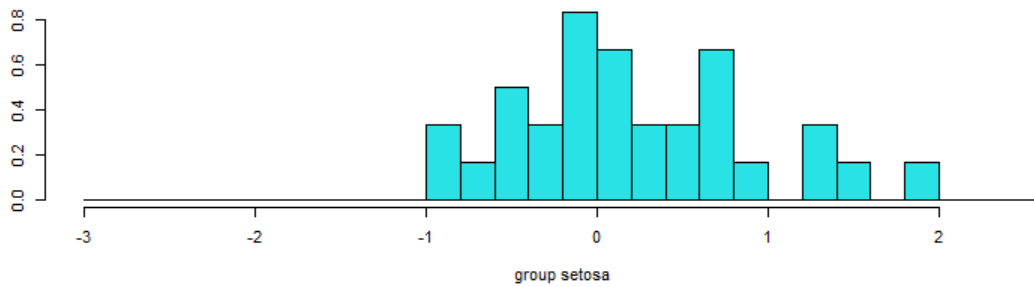
```
p <- predict(linear, data = training)
```

Εδώ αποθηκεύουμε τη λίστα με τις προβλέψεις στο αντικείμενο `p`. Η `p$x` έχει αποθηκεύσει τις τιμές των δύο διακρίνουσων συναρτήσεων.

```
ldahist(data = p$x[,1], g = training$Species)
```



```
ldahist(data = p$x[,2], g = training$Species)
```



Με τη σύνταξη `Species~.` δηλώνουμε ότι θέλουμε να χρησιμοποιήσουμε όλες τις ανεξάρτητες στο `data.frame data = training` (εκπαιδευτικό δείγμα). Η κατάταξη των ομάδων εμφανίζεται με διαφορετικά χρώματα - παρατηρήστε ότι οι κανόνες ταξινόμησης είναι γραμμικοί (χωρίζουν το χώρο με γραμμές). Οι παρατηρήσεις που έχουν ταξινομηθεί λανθασμένα εμφανίζονται με κόκκινο χρώμα. Στα περισσότερα γραφήματα ο αριθμός των λανθασμένα ταξινομημένων παρατηρήσεων είναι μικρός, εκτός από το ζευγάρι `Sepal.Length` και `Sepal.Width` όπου η πιθανότητα δυσταξινόμησης είναι μεγαλύτερη.

## Ακρίβεια στο εκπαιδευτικό δείγμα

Η εντολή `predict` μπορεί να προβλέψει την ομάδα του λουλουδιού με βάση το μοντέλο `linear` (αποθηκεύεται στο `$class`).

```
# Confusion matrix and accuracy - training data
p1 <- predict(linear, training)$class
tab <- table(Predicted = p1, Actual = training$Species)
tab

##               Actual
## Predicted   setosa versicolor virginica
##   setosa      30         0         0
##   versicolor  0         30         0
##   virginica   0         0        29

sum(diag(tab))/sum(tab)

## [1] 1
```

Εδώ έχουμε χρησιμοποιήσει το εκπαιδευτικό δείγμα και προβλέπουμε την ομάδα με βάση το μοντέλο. Στη συνέχεια δημιουργούμε έναν πίνακα συνάφειας με την πραγματική και προβλεπόμενη ομάδα λουλουδιού. Η ακρίβεια είναι 100%!

## Ακρίβεια στο επικυρωτικό δείγμα

```
# Confusion matrix and accuracy - testing data
p2 <- predict(linear, testing)$class
tab1 <- table(Predicted = p2, Actual = testing$Species)
tab1

##               Actual
## Predicted   setosa versicolor virginica
##   setosa      20         0         0
##   versicolor  0        19         1
##   virginica   0         1        20

sum(diag(tab1))/sum(tab1)

## [1] 0.9672131
```

Για ακριβέστερα συμπεράσματα, αξιολογούμε το μοντέλο στο επικυρωτικό δείγμα. Παρατηρείστε τώρα ότι η ακρίβεια μειώθηκε λίγο (96.7%).