Chapter 2
Supervised learning

Xavier Gendre

M2 SE

Lubischew data

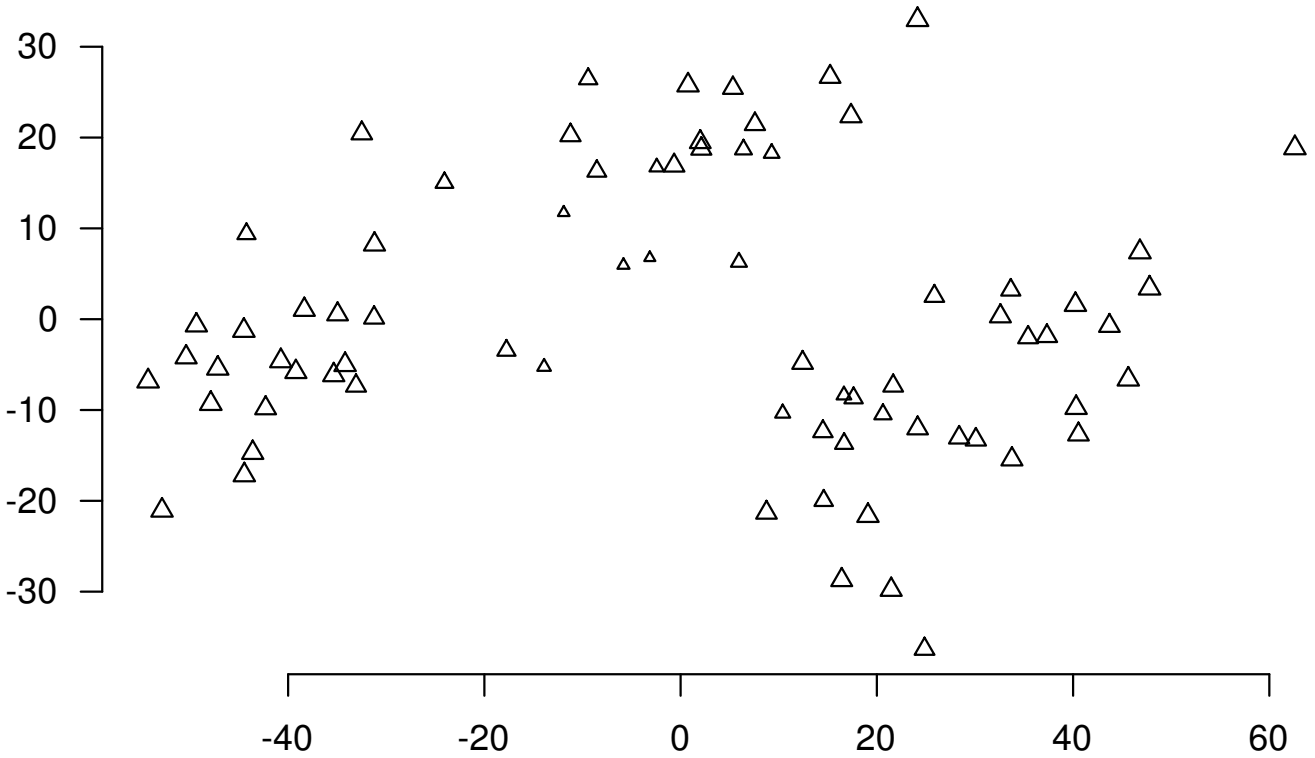
```
dim(X)

## [1] 74  6

Species

##  [1] A A A A A A A A A A A A A A A A A A A A A A B B B B B B B B B B B B B B
## [36] B B B B B B B B C C C C C C C C C C C C C C C C C C C C C C C C C C C C
## [71] C C C C
## Levels: A B C
```

Principal components analysis:

```
Xbar <- scale(X, scale = F)
Sigma <- t(Xbar) %*% Xbar/nrow(Xbar)
dg <- eigen(Sigma)
C <- Xbar %*% dg$vectors
cos2 <- rowSums(C[, 1:2]^2)/rowSums(C^2)
```

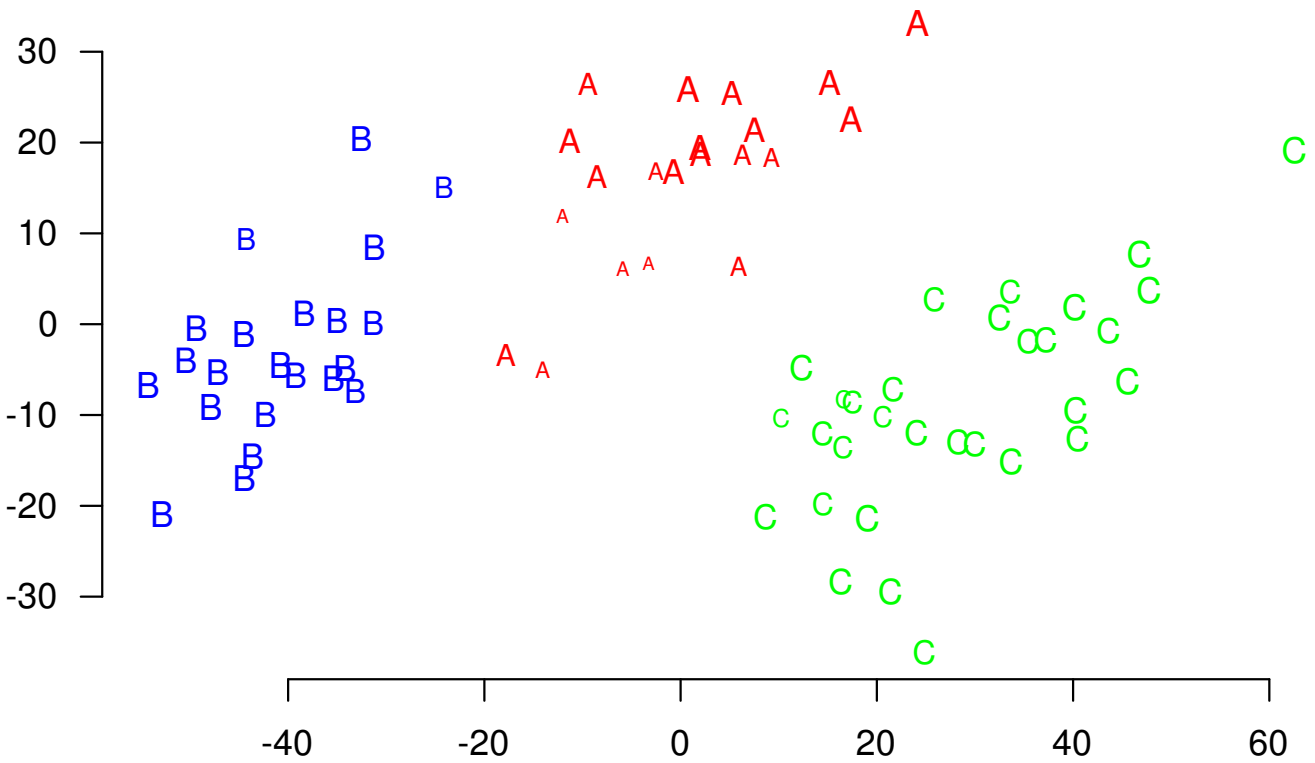
Lubischew PCA

```
plot(C[, 1:2], pch = 2, cex = cos2, xlab = "", ylab = "")
```



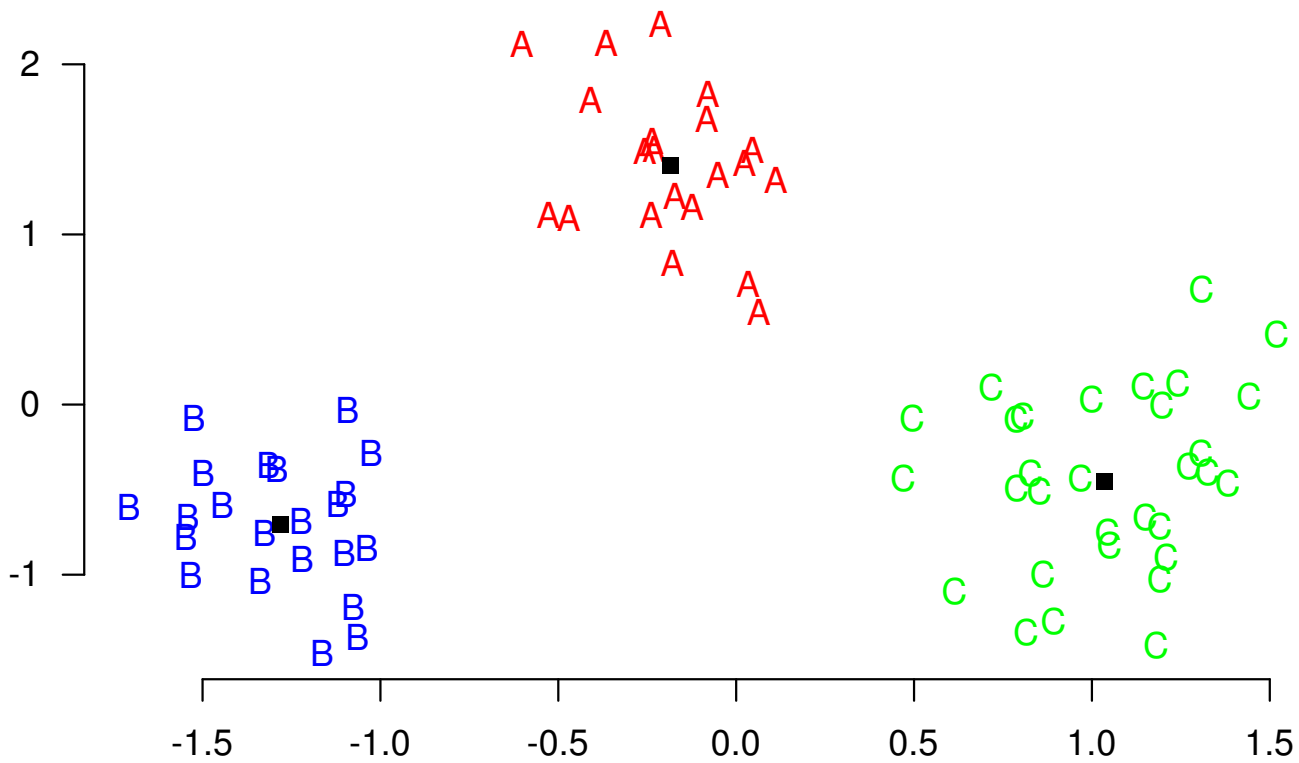
Lubischew PCA

```
text(C[, 1:2], labels = Species, cex = cos2, col = specol)
```



Lubischew MDA

```
text(C[, 1:2], labels = Species, col = specol)
```



```
dg$values[1]
```

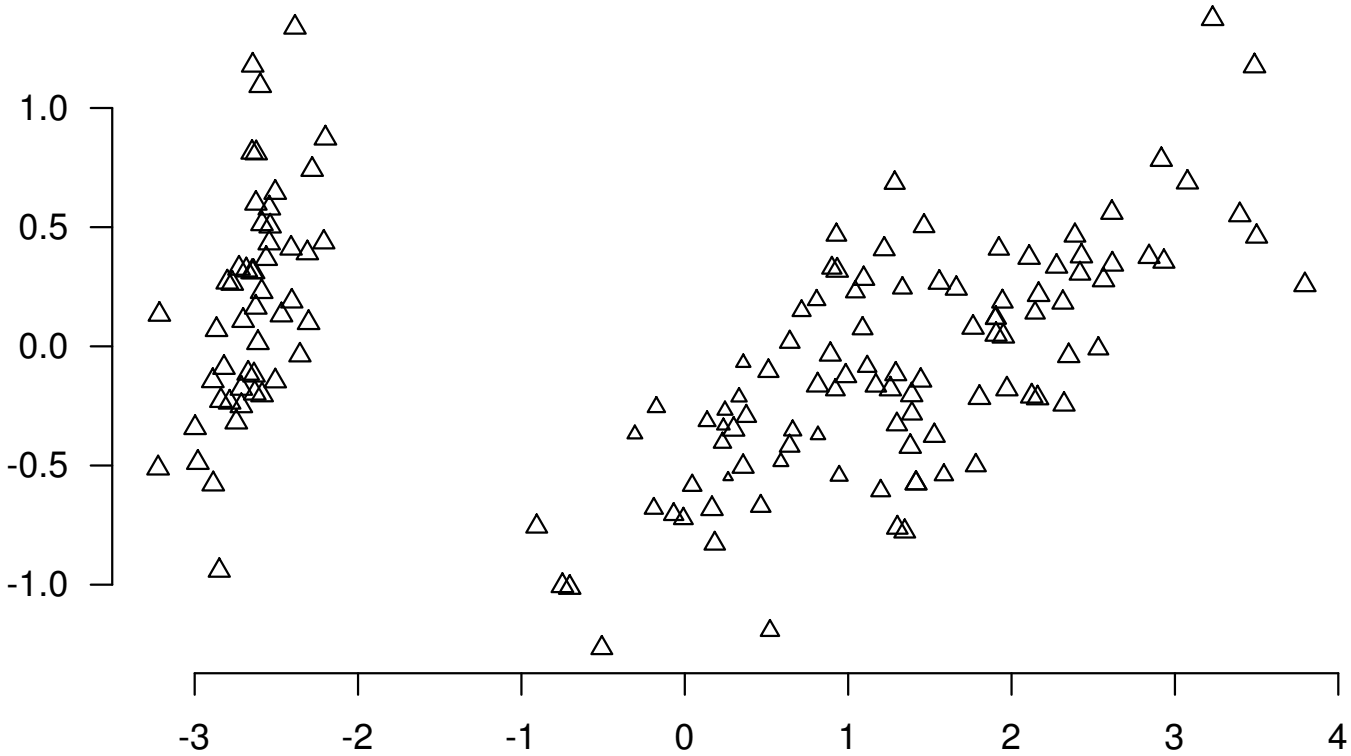
```
## [1] 0.9468
```

```
cumsum(dg$values)/sum(dg$values)
```

```
## [1] 0.5435 1.0000 1.0000 1.0000 1.0000 1.0000
```

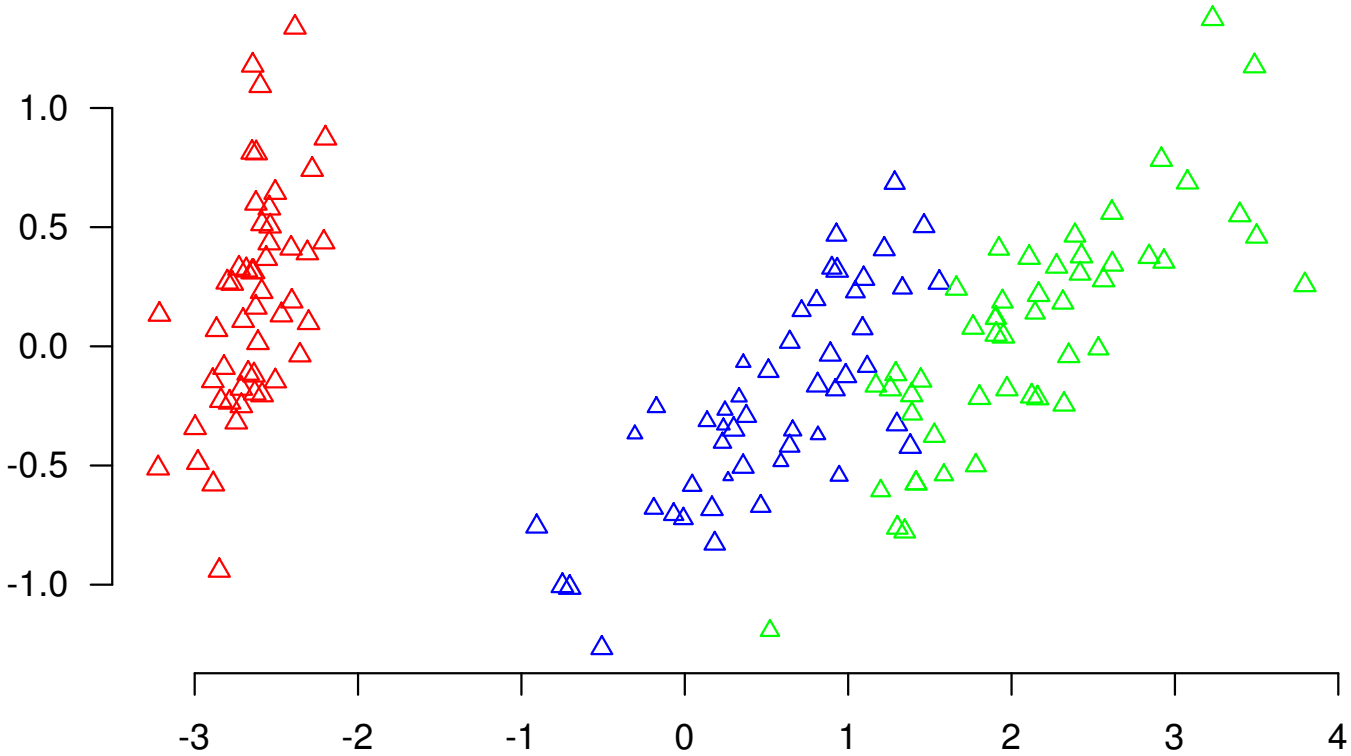
Fisher Iris PCA

```
plot(C[, 1:2], pch = 2, cex = cos2, xlab = "", ylab = "")
```



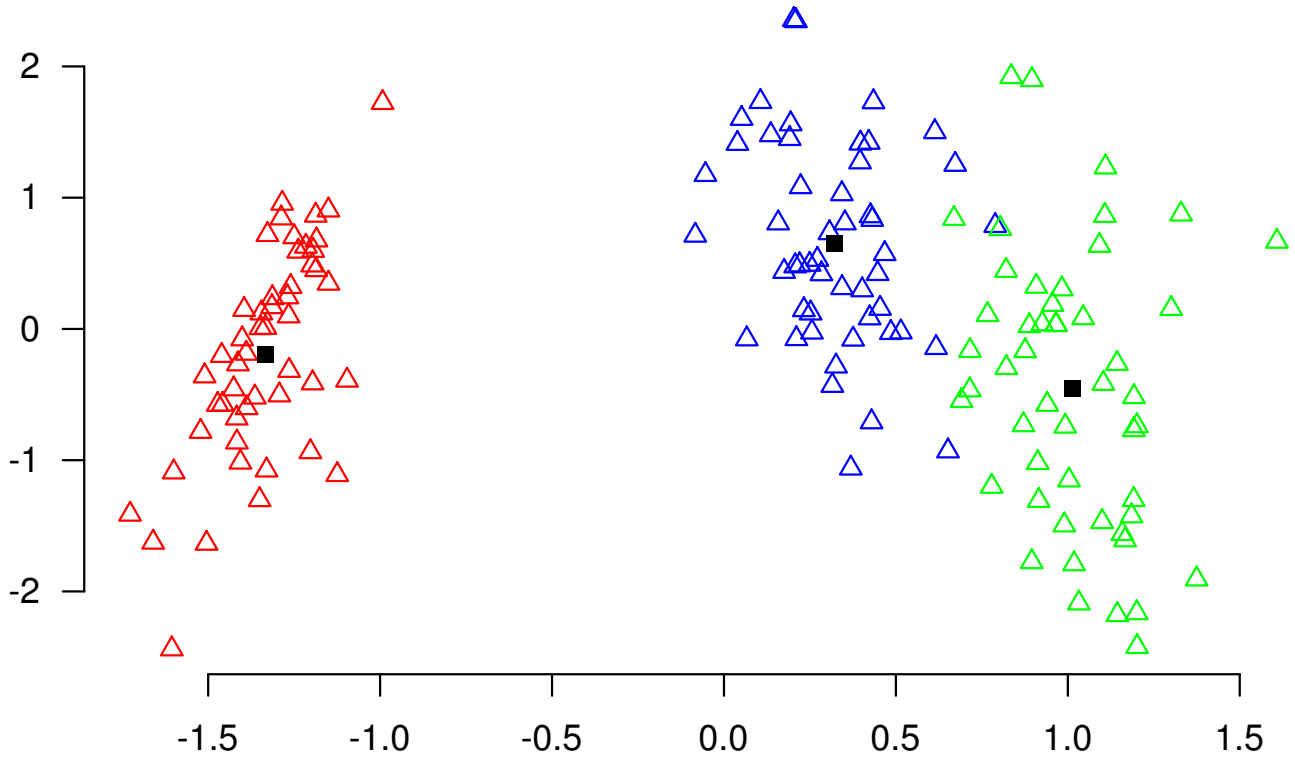
Fisher Iris PCA

```
plot(C[, 1:2], pch = 2, cex = cos2, xlab = "", ylab = "", col = specol)
```



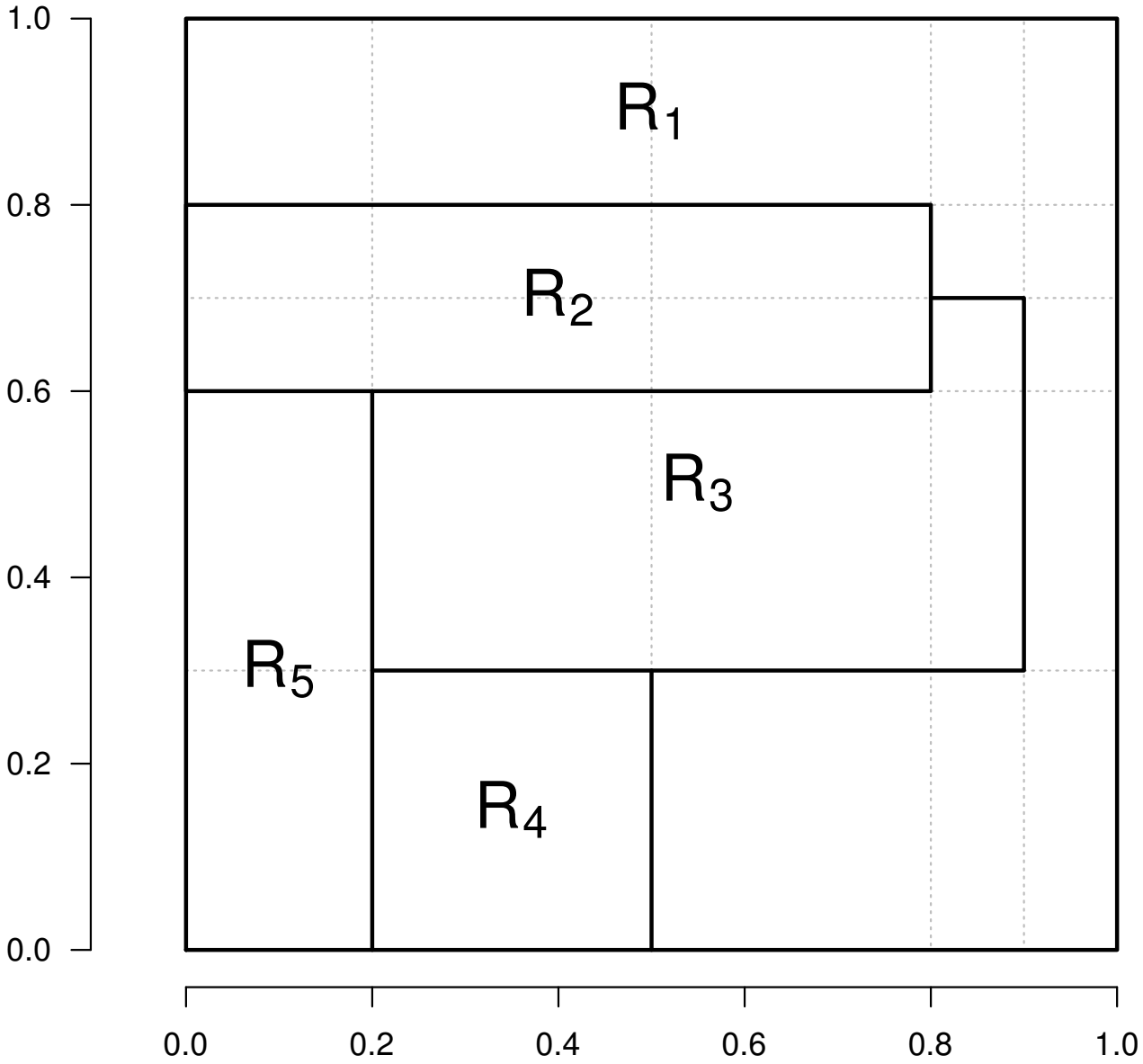
Fisher Iris MDA

```
plot(C[, 1:2], pch = 2, col = specol, xlab = "", ylab = "")
```

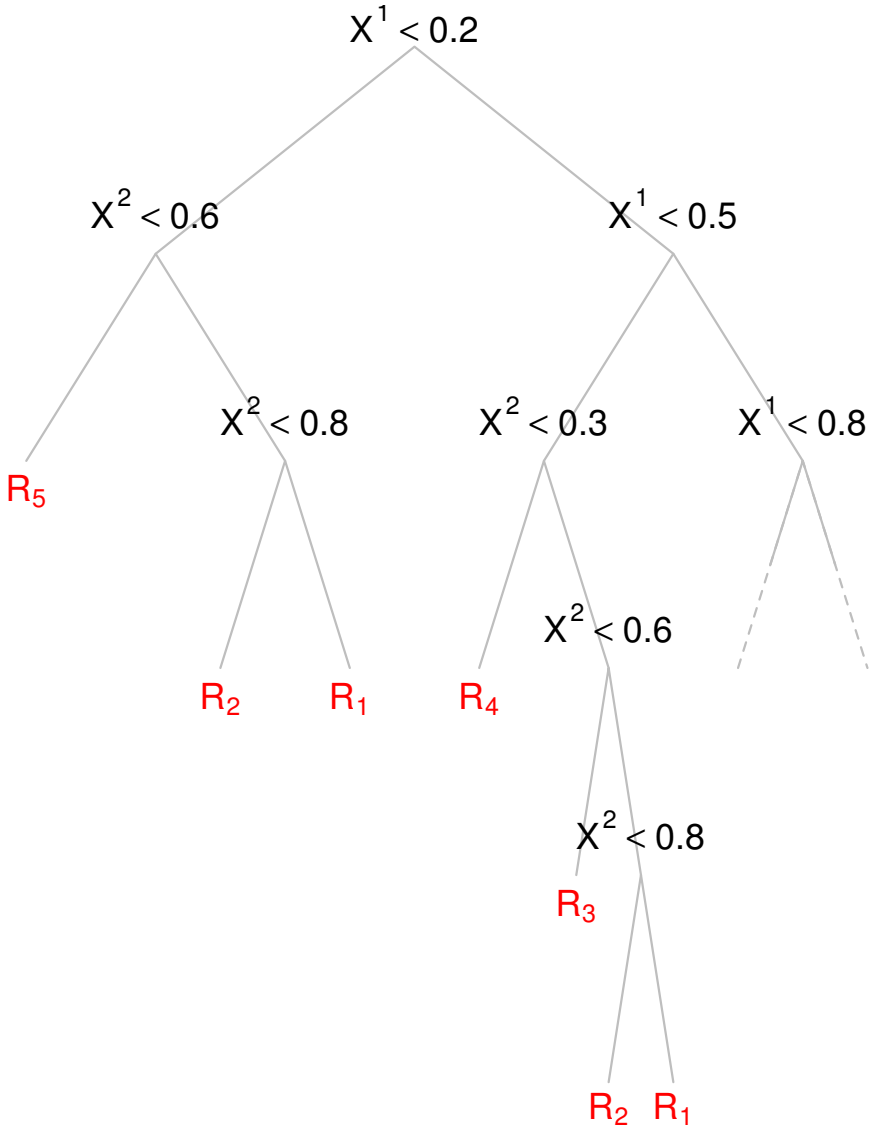


```
dg$values[1]

## [1] 0.9699

cumsum(dg$values)/sum(dg$values)

## [1] 0.8137 1.0000 1.0000 1.0000
```

Binary parition



Binary tree



Spam data

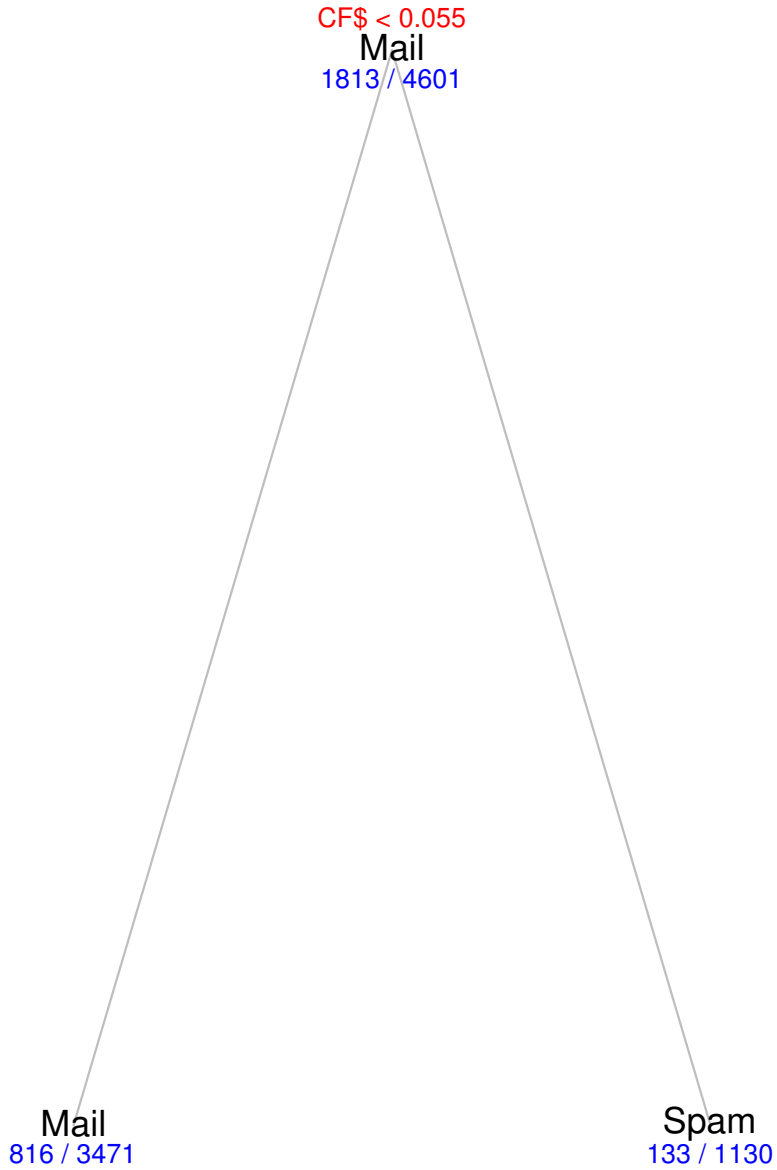
```
# Mail n°1

## [1] "spam"
```

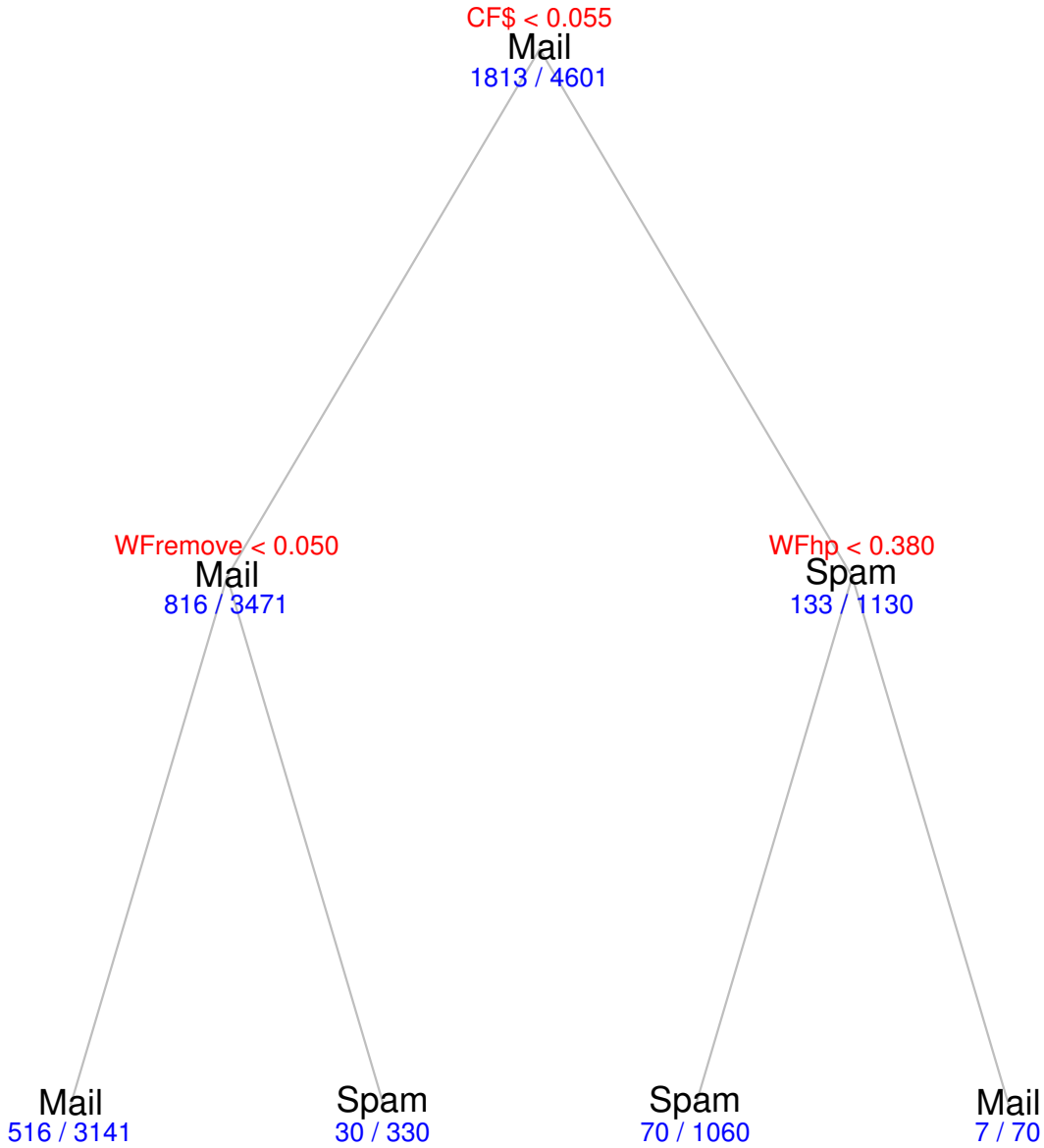
```
##     WFmake  WFaddress       WFall        WF3d      WFour      WFover
##       0.00       0.64        0.64        0.00       0.32        0.00
##   WFremove WFinternet    WForder     WFmail
##       0.00       0.00        0.00        0.00
## [ ... ]
##        WFedu      WFtable WFconference          CF;          CF(
##        0.000        0.000        0.000        0.000        0.000
##          CF[          CF!          CF$          CF#       CAPave
##        0.000        0.778        0.000        0.000        3.756
##       CAPlon       CAPtot
##       61.000      278.000
```

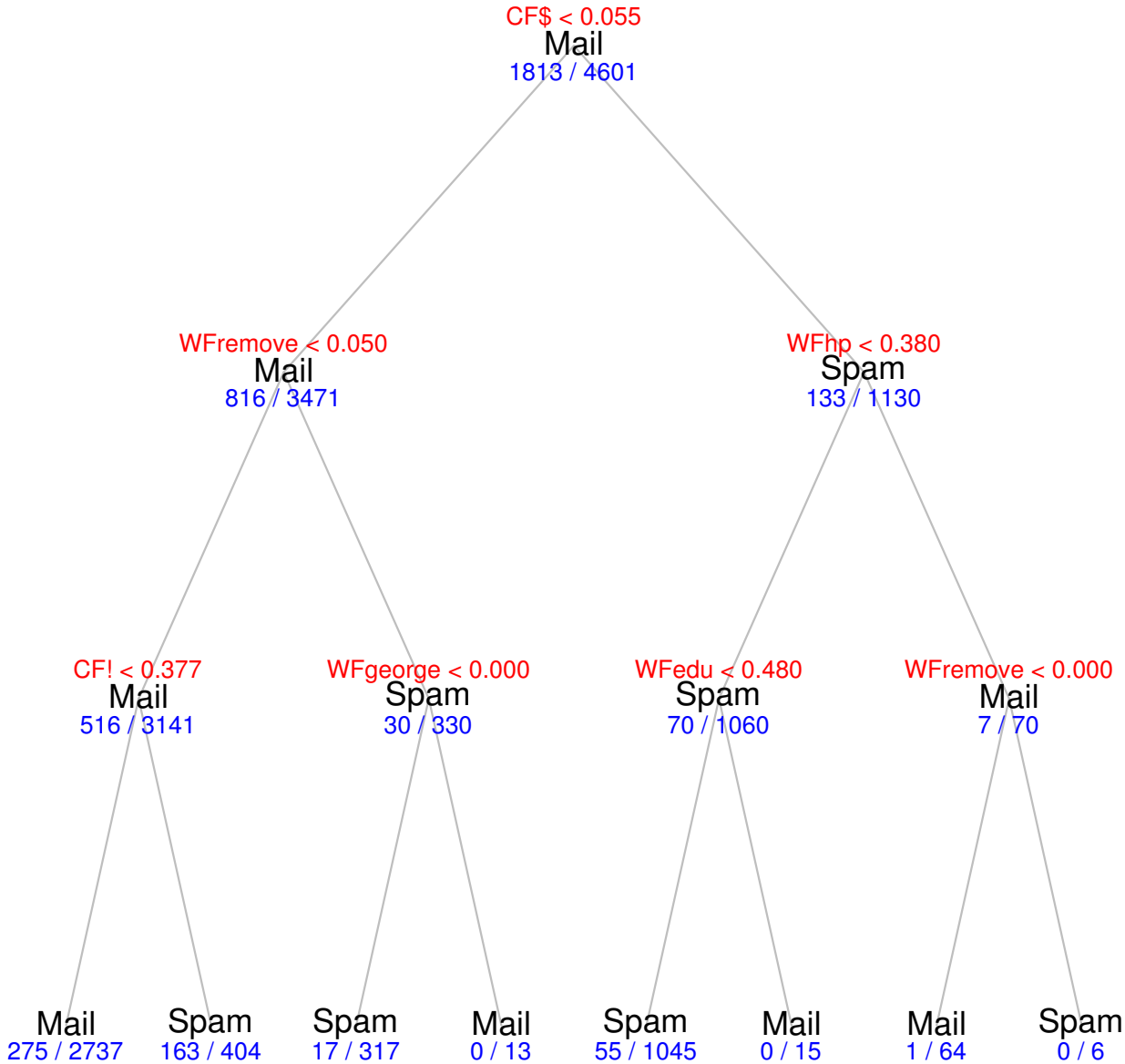
Spam data

```
# Mail n°1900

## [1] "mail"
```

```
##     WFmake  WFaddress       WFall        WF3d      WFour      WFover
##          0          0           0           0          0           0
##   WFremove WFinternet    WForder     WFmail
##          0          0           0           0
## [ ... ]
##        WFedu      WFtable WFconference          CF;          CF(
##         7.14         0.00         0.00         0.00         0.00
##          CF[          CF!          CF$          CF#       CAPave
##         0.00         0.00         0.00         0.00         5.50
##       CAPlon       CAPtot
##        10.00        11.00
```


CF$ < 0.055
Mail
1813 / 4601
Mail
816 / 3471
Spam
133 / 1130


CF$ < 0.055
Mail
1813 / 4601
WFremove < 0.050
Mail
816 / 3471
WFhp < 0.380
Spam
133 / 1130
Mail
516 / 3141
Spam
30 / 330
Spam
70 / 1060
Mail
7 / 70


CF$ < 0.055
Mail
1813 / 4601
WFremove < 0.050
Mail
816 / 3471
WFhp < 0.380
Spam
133 / 1130
CF! < 0.377
Mail
516 / 3141
WFgeorge < 0.000
Spam
30 / 330
WFedu < 0.480
Spam
70 / 1060
WFremove < 0.000
Mail
7 / 70
Mail
275 / 2737
Spam
163 / 404
Spam
17 / 317
Mail
0 / 13
Spam
55 / 1045
Mail
0 / 15
Mail
1 / 64
Spam
0 / 6

Bibliography

- *The Elements of Statistical Learning : Data Mining, Inference and Prediction*, J. Friedman, T. Hastie and R. Tibshirani (2009)
- *Classification and Regression Trees*, L. Breiman, J. Friedman, R. Olshen and C. Stone (1984)
- *Sélection de variables pour la discrimination en grande dimension et classification de données fonctionnelles (PhD. Thesis)*, C. Tuleau (2005)
- *Wiki Stat*, `http://wikistat.fr/`