Scoring Appliqué à la Détection du Risque – Chapitre 2 – Exercices

Exercice 1.

Le fichier `client`[d] donne pour cent clients bancaires : la liquidité (`cash`), le flux (`flow`), l'épargne (`saving`), le niveau de consommation (`consume`) et la classe de risque (`risk`).

A quelle classe de risque l'analyse discriminante probabiliste basée sur un modèle Gaussien affecte-t-elle un nouveau client avec les caractéristiques suivantes : liquidité $= 6,2$; flux $= 2,9$; épargne $= 4,9$; consommation $= 1,7$?

Représentation des données

```
client <- read.table(file='http://alexandrelourme.free.fr/M2IREF/SCORING/client.csv',
sep=',',dec='.',header=TRUE)
names(client)

## [1] "cash"    "flow"    "saving"  "consume" "risk"

desc=as.data.frame(client[,1:4])
risk=as.factor(client[,5]+1)
# plot(desc,col=risk) # enlever # pour représenter les données
```

Apprentissage du classifieur Gaussien

```
library(Rmixmod)
gausslearn <- mixmodLearn(desc, knownLabels=risk)
```

Affectation du nouveau client

```
new=data.frame(cash=6.2, flow=2.9, saving=4.9, consume=1.7) # nouveau client
prediction <- mixmodPredict ( data = new , classificationRule = gausslearn["bestResult"])
print(prediction[6]) # probabilités associées aux classes

##           [,1]      [,2]
## [1,] 0.4381953 0.5618047

print(prediction[5]) # classe de risque estimée

## [1] 2
```

Exercice 2. On considère les données `bank{gclus}` de R pour lesquels deux modèles sont proposés.

Modèle G : dans chaque classe (genuine/counterfeit) de billets, les descripteurs continus (Length, Left, Right, Bottom, Top, Diagonal) forment un vecteur gaussien.

Modèle T : dans chaque classe (genuine/counterfeit) de billets, les descripteurs continus (Length, Left, Right, Bottom, Top, Diagonal) forment un vecteur distribué selon une loi de Student multidimensionnelle.

1. Quel modèle parmi G et T le critère *BIC* sélectionne-t-il ?

Représentation des données

```
library(gclus)
data(bank)
desc=as.data.frame(bank[,2:7])
class=as.factor(bank[,1]+1)
# plot(desc,col=class) # enlever # pour représenter les données
```

Apprentissage d'un classifieur à classes Gaussiennes

[d] disponible sous : `http://alexandrelourme.free.fr/M2IREF/SCORING/client.csv`

```
library(Rmixmod)
gausslearn <- mixmodLearn(desc, knownLabels=class,criterion=c('BIC'))
print(-gausslearn[8][3]/2) # valeur de BIC du modèle à classes Gaussiennes

## [1] -891.2999
```

Apprentissage d'un classifieur à classes Student

```
library(teigen)
studentlearn <- teigen(desc, known=class)

print(studentlearn[[5]]/2) # valeur de BIC du modèle à classes Student

## [1] -1386.187
```

2. A quelle classe l'analyse discriminante basée sur le modèle T affecte-t-elle un billet présentant les caractéristiques suivantes : `Length:214.9`, `Left:130.1`, `Right:129.8`, `Bottom:9.4`, `Top:10.6`, `Diagonal:140.3` ?

```
new=data.frame(Length=214.9,Left=130.1,Right=129.8,Bottom=9.4,Top=10.6,Diagonal=140.3)
predict(studentlearn,new)

## $fuzzy
##           [,1]       [,2]
## [1,] 0.9606948 0.03930517
##
## $classification
## [1] 1
```

Exercice 3.

Dans le fichier `loan` disponible sous : `http://alexandrelourme.free.fr/M2IREF/SCORING/loan.csv`, soixante individus sont décrits par sept variables dont voici l'interprétation.

variable	*interprétation*
salaire	salaire mensuel moyen au cours de l'année 2014
solde	solde au 15 du mois moyenné sur l'année 2014
age	âge atteint par le sujet au cours de l'année
nbenfant	nombre d'enfants au 31 Dec. 2014
vehicule	type de véhicule dont il dispose (1: aucun, 2: voiture, 3: deux roues)
reside	lieu de résidence (1: ville, 2: campagne)
classe	risque de défaut pour le prêt *trankilis* (1: élevé, 2: faible)

On cherche à estimer la solvabilité des dix derniers sujets en apprenant un modèle utile à la classification sur les cinquante premiers individus.

1. Quelle est la variable à prédire ? Quels sont les descripteurs ? Quelle est la nature de chaque descripteur ?
2. Quel modèle permettrait de réaliser l'analyse discriminante de ces données ?
3. Estimez le paramètre de ce modèle en utilisant mixtcomp[a]. Déterminez la valeur de BIC du modèle.
4. Estimez la probabilité de défaut de chacun des dix derniers sujets. A qui accorderiez vous le prêt ?

Exercice 4.

On considère le jeu `Hdma` du package `Ecdat` de R ; ces données sont disponibles sous `http://alexandrelourme.free.fr/M2IREF/SCORING/hdma.rda`. Un échantillon d'apprentissage S est constitué des travailleurs indépendants décrits par : `dir`, `hir`, `lvr` et répartis en deux classes : prêt hypotécaire accepté/refusé. Les travailleurs non indépendants forment un échantillon de test S'.

[a] mixtcomp est une application permettant de réaliser l'analyse discriminante de données mixtes ; son utilisation est détaillée sous : `https://modal.lille.inria.fr/wikimodal/doku.php?id=mixtcomp` ; elle est utilisable en ligne sur le serveur de l'Inria de Lille en créant un compte à l'adresse : `https://modal-research.lille.inria.fr/BigStat/project/mixtcomp/`

```r
load('/home/alourme/Bureau/hdma.rda')
attach(Hdma)
```

1. Interprétez les variables `dir`, `hir`, `lvr`.

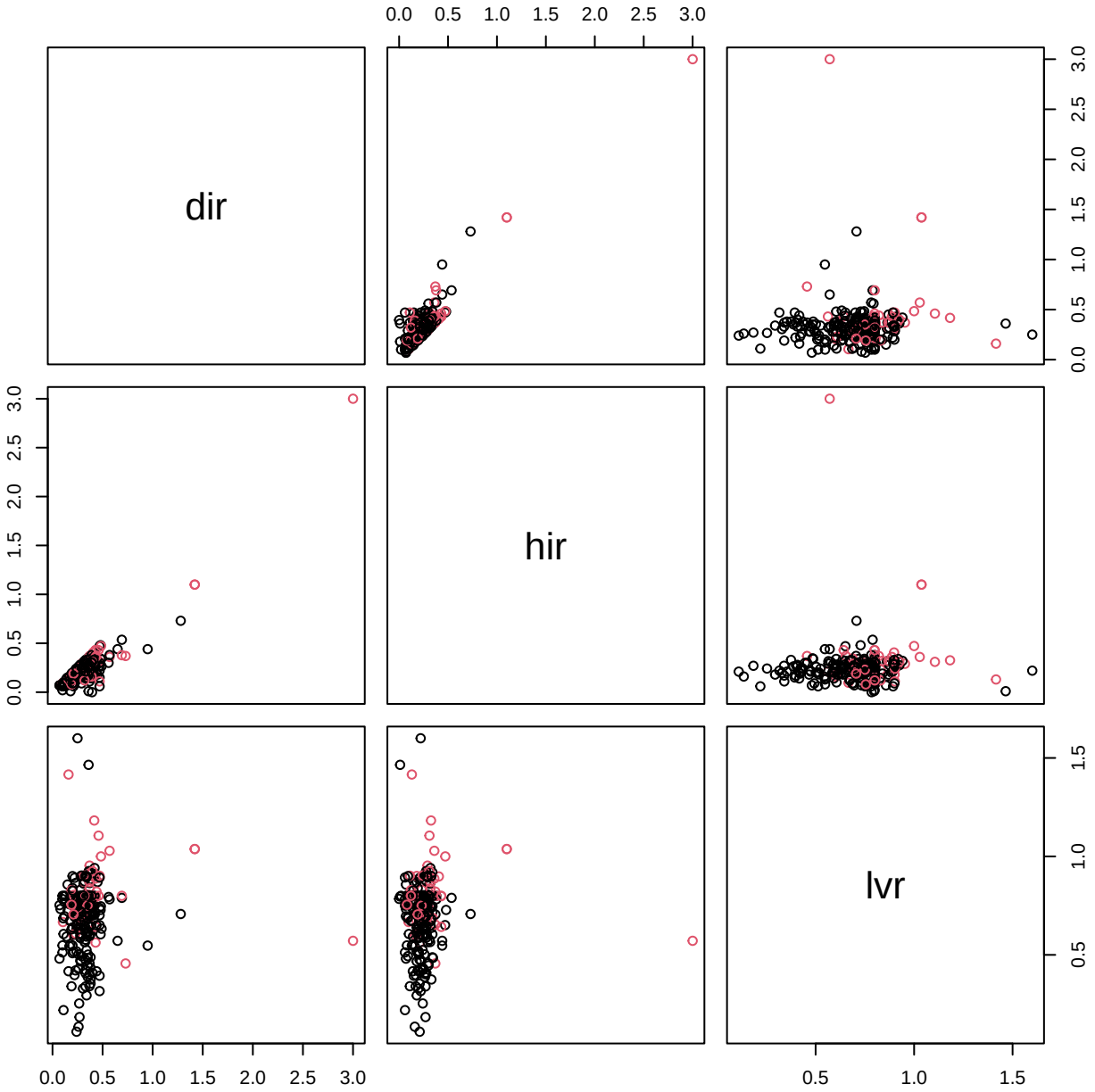
```r
 # dir : paiement de la dette/revenu total
 # hir : dépenses logement/revenu
 # lvr : taille du prêt/valeur imposable des biens
```

2. Quels sont les paramètres du modèle de régression logistique appris sur S ?

```r
Hdma <- Hdma[complete.cases(Hdma),] # ne conserve que les lignes sans données manquantes
train <- Hdma[Hdma$self=='yes',c(1,2,3,13)] ; attach(train)
# ne conserve que les variables 1 (dir), 2 (hir), 3 (lvr), 13 (deny) et les indépendants

test <- Hdma[Hdma$self=='no',c(1,2,3,13)]
# ne conserve que les variables 1 (dir), 2 (hir), 3 (lvr) et les non indépendants

LR <- glm(deny~dir+hir+lvr,data=train,family=binomial(lin='logit'))

print(summary(LR))

##
## Call:
## glm(formula = deny ~ dir + hir + lvr, family = binomial(lin = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9523  -0.6028  -0.4723  -0.2944   2.3811
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.0294     0.9938  -6.067  1.3e-09 ***
## dir           0.9957     1.8960   0.525 0.599466
## hir           3.3383     2.5514   1.308 0.190736
## lvr           4.2442     1.1031   3.848 0.000119 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 249.07  on 276  degrees of freedom
## Residual deviance: 213.99  on 273  degrees of freedom
## AIC: 221.99
##
## Number of Fisher Scoring iterations: 5

# Pour connaitre la correspondance entre classes : yes/no et codes classes : 1/2
score <- predict(LR,new=train)
print(table(score>0,train[,4]))

##
##          no yes
##   FALSE 228  38
##   TRUE    3   8

# Ainsi, la classe 2 (score <0) contient les sujets dont le prêt a été accepté (deny=no)
```

3. Représentez les données de S dans chacun des plans canoniques.

```
plot(train[,1:3],col=as.factor(train[,4]))
```



4. Déterminez la probabilité de refus d'un prêt hypothécaire aux individus de S'.

```
score <- predict(LR,new=test)

proba_classe_1 <- exp(score)/(1+exp(score))

print(head(proba_classe_1))

##         1         2         3         4         5         6
## 0.1576053 0.2752407 0.2840069 0.2272085 0.1239057 0.0449557
```

5. Quelle est la classe estimée (refus/acceptation du prêt hypothécaire) pour les individus de S' ?

```r
decision <- NULL
decision[proba_classe_1 < 0.5]='no' # acceptation (deny=no)
decision[proba_classe_1 >= 0.5]='yes' # refus (deny=yes)
print(head(decision))

## [1] "no" "no" "no" "no" "no" "no"
```

6. Dressez le tableau de confusion entre classe réelle et classe estimée pour les individus de S'.

```r
cat('table de confusion classe réelle x classe estimée pour l\'échantillon de test','\n')

## table de confusion classe réelle x classe estimée pour l'échantillon de test

cat('[yes: refus/no:acceptation(non_refus)]','\n')

## [yes: refus/no:acceptation(non_refus)]

print(table(test[,4],decision))

##       decision
##          no  yes
##   no  1857    7
##   yes  227   12
```
