Université de Carthage
Ecole Supérieure de la Statistique et de l'Analyse de l'Information

Examen de Data Mining

3 ème année du cycle de formation d'ingénieurs

Durée de l'épreuve : 1 heure 30 - Documents non autorisés
Nombre de pages : 3 - Date de l'épreuve : 1er février 2018

On considère la base de données Breast Cancer (fichier `breast-cancer.csv`). Chacun des 699 individus de cette base est décrit par trois variables explicatives quantitatives (ucellsize, normnucl et mitoses) ainsi que par la variable à expliquer (class) prenant les valeurs 1 ou 0 selon que l'individu soit atteint ou pas du cancer du sein. Ci-dessous les statistiques descriptives des données :

```
> breast_cancer<-read.table("breast-cancer.csv",header=TRUE, sep=";")
> summary(breast_cancer)
   ucellsize         normnucl         mitoses       class
 Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   0:458
 1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.000   1:241
 Median : 1.000   Median : 1.000   Median : 1.000
 Mean   : 3.134   Mean   : 2.867   Mean   : 1.589
 3rd Qu.: 5.000   3rd Qu.: 4.000   3rd Qu.: 1.000
 Max.   :10.000   Max.   :10.000   Max.   :10.000
```

Afin d'expliquer Y on réalise des classifications supervisées à l'aide du logiciel R.

Nous avons expliqué la variable `class` en effectuant une régression logistique. Les résultats obtenus sont présentés ci-dessous :

```
> modele_glm

Call:  glm(formula = class ~ ucellsize + normnucl + mitoses, family = "binomial")

Coefficients:
(Intercept)    ucellsize     normnucl      mitoses
    -5.8574       1.1790       0.3579       0.5422

Degrees of Freedom: 698 Total (i.e. Null);  695 Residual
Null Deviance:      900.5
Residual Deviance: 237.6        AIC: 245.6

> anova(modele_glm,test = "Chisq")
Analysis of Deviance Table
```

```
Model: binomial, link: logit

Response: class

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        698     900.53
ucellsize  1   624.97       697     275.55 < 2.2e-16 ***
normnucl   1    26.42       696     249.14 2.751e-07 ***
mitoses    1    11.49       695     237.64 0.0006984 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> mc_glm<-table(modele_glm$fitted.values>0.5,class==1)
```

1. Commenter les résultats de la commande : `anova(modele_glm,test = "Chisq")`.

2. Commenter la commande : `table(modele_glm$fitted.values>0.5,class==1)`.

Nous avons aussi expliqué la variable `class` en utilisant la méthode des arbres de décision. Les résultats obtenus sont présentés ci-dessous :

```
> library(rpart)
> modele_arbre<- rpart(class ~ ., data = breast_cancer,method = "class")
> print(modele_arbre)
n= 699

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 699 241 0 (0.65522175 0.34477825)
   2) ucellsize< 2.5 429  12 ? (0.97202797 0.02797203)
     4) normnucl< 3.5 420   5 ? (0.98809524 0.01190476) *
     5) normnucl>=3.5 9   2 ? (0.22222222 0.77777778) *
   3) ucellsize>=2.5 270  41 ? (0.15185185 0.84814815)
     6) ucellsize< 3.5 52  25 ? (0.51923077 0.48076923)
      12) normnucl< 2.5 27   7 ? (0.74074074 0.25925926) *
      13) normnucl>=2.5 25   7 ? (0.28000000 0.72000000) *
     7) ucellsize>=3.5 218  14 ? (0.06422018 0.93577982) *

> printcp(modele_arbre)

Classification tree:
rpart(formula = class ~ ., data = breast_cancer, method = "class")
Root node error: 241/699 = 0.34478

n= 699
```

```
        CP nsplit rel error  xerror     xstd
1 0.780083      0   1.00000 1.00000 0.052142
2 0.026971      1   0.21992 0.25726 0.031190
3 0.020747      3   0.16598 0.21992 0.029040
4 0.010000      4   0.14523 0.19917 0.027743
```

3. Représenter l'arbre obtenu après avoir remplacé chaque " ?" par la valeur adéquate.

4. Determiner les règles issues de cet arbre.

5. Commenter les résultats de la commande `printcp(modele_arbre)`.

Nous avons enfin expliqué la variable `class` en utilisant la méthode des Random Forest. Les résultats obtenus sont présentés ci-dessous :

```
> library(randomForest)
> modele_RF <- randomForest(class~.,data=breast_cancer, ntree=2500)
> imp <- importance(modele_RF)
> order(imp,decreasing=TRUE)
[1] 1 2 3
```

6. Donner l'algorithme de la la méthode des Random Forest.

7. A partir des résultats des 3 méthodes, déterminer, en justifiant votre réponse, les variables les plus explicatives de la variable `class`.