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 : 4 - Date de l'épreuve : 6 janvier 2018

Exercice 1 : Le traitement du cancer de la prostate change si le cancer a atteint ou non les noeuds lympatiques entourant la prostate. Pour éviter une investigation lourde un certain nombre de variables sont considérées comme explicatives de la variable Y : $Y = 0$ si le cancer n'a pas atteint le réseau lympatique et $Y = 1$ sinon. Le but de cette étude est donc d'expliquer et de prédire Y par les variables suivantes :

- `age` : âge du patient au moment du diagnostic ;
- `acide` : le niveau d'acide phosphate sérique ;
- `rayonx` : le résultat d'une analyse par rayon X, 0= négatif et 1= positif ;
- `taille` : la taille de la tumeur, 0= petite et 1= grande ;
- `grade` : l'état de la tumeur déterminé par biopsie, 0= moyen et 1= grave ;
- `log.acid` : le logarithme népérien du niveau d'acidité ;

On dispose de la base de données `cancer_prostate` (fichier cancerprostate.txt) constituée de 53 individus. Chacun des 53 individus est décrit par les 6 variables prédictives présentées ci-dessus ainsi que par sa valeur sur la variable Y. Ci-dessous les statistiques descriptives des données :

```
> cancer_prostate<-read.table("cancerprostate.txt",sep=";",header=T)
> summary(cancer_prostate)
      age            acide          rayonx  taille  grade   Y       log.acid
 Min.   :45.00   Min.   :0.4000   0:38    0:26    0:33    0:33    Min.   :-0.9163
 1st Qu.:56.00   1st Qu.:0.5000   1:15    1:27    1:20    1:20    1st Qu.:-0.6931
 Median :60.00   Median :0.6500                                   Median :-0.4308
 Mean   :59.38   Mean   :0.6942                                   Mean   :-0.4189
 3rd Qu.:65.00   3rd Qu.:0.7800                                   3rd Qu.:-0.2485
 Max.   :68.00   Max.   :1.8700                                   Max.   : 0.6259
```

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

1) Si l'on devait utiliser l'analyse discriminante pour expliquer la variable Y, indiquer la démarche à suivre.

Dans la suite, la classification est réalisée à l'aide d'une régression logistique.

Dans une première étape nous avons obtenu les résultats suivants :

```
> modele1 = glm(Y ~., family=binomial,cancer_prostate)
> summary(modele1)

Call:
glm(formula = Y ~ ., family = binomial, data = cancer_prostate)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0960  -0.6102  -0.2863   0.4834   2.2000

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.08672    7.83450   1.287   0.1979
age         -0.04289    0.06166  -0.696   0.4867
acide       -8.48006    7.63305  -1.111   0.2666
rayonx       2.06673    0.85469   2.418   0.0156 *
taille       1.38415    0.79546   1.740   0.0819 .
grade        0.85376    0.81247   1.051   0.2933
log.acid     9.60912    6.21652   1.546   0.1222
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 44.768  on 46  degrees of freedom
AIC: 58.768

Number of Fisher Scoring iterations: 5

> mc1 <- table(modele1$fitted.values>0.5,Y==1)
> mc1
       FALSE TRUE
  FALSE    28    7
  TRUE      5   13
```

Par la suite, nous avons effectué une régression logistique pas à pas sur cette base pour obtenir ce que nous appelons modele2. Les résultats sont donnés ci-dessous :

```
> modele2<-step(modele1,dir = "backward")
Start:  AIC=58.77
Y ~ age + acide + rayonx + taille + grade + log.acid

           Df Deviance    AIC
- age       1   45.259 57.259
- grade     1   45.883 57.883
- acide     1   46.560 58.560
<none>          44.768 58.768
- taille    1   47.949 59.949
- log.acid  1   48.126 60.126
```

```
- rayonx    1   51.368 63.368

Step:  AIC=57.26
Y ~ acide + rayonx + taille + grade + log.acid

           Df Deviance    AIC
- grade     1   46.425 56.425
<none>          45.259 57.259
- acide     1   47.776 57.776
- taille    1   48.300 58.300
- log.acid  1   49.615 59.615
- rayonx    1   51.742 61.742

Step:  AIC=56.43
Y ~ acide + rayonx + taille + log.acid

           Df Deviance    AIC
<none>          46.425 56.425
- acide     1   48.986 56.986
- log.acid  1   50.660 58.660
- taille    1   51.246 59.246
- rayonx    1   53.707 61.707
> summary(modele2)

Call:
glm(formula = Y ~ acide + rayonx + taille + log.acid, family = binomial,
    data = cancer_prostate)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9452  -0.6464  -0.2999   0.4517   2.2676

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   9.0668     7.9335   1.143   0.2531
acide        -9.8617     7.8364  -1.258   0.2082
rayonx        2.0934     0.8273   2.530   0.0114 *
taille        1.5909     0.7574   2.100   0.0357 *
log.acid     10.4097     6.2649   1.662   0.0966 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.252  on 52  degrees of freedom
Residual deviance: 46.425  on 48  degrees of freedom
AIC: 56.425

Number of Fisher Scoring iterations: 5

> mc2<-table(modele2$fitted.values>0.5,Y==1)
> mc2
       FALSE TRUE
  FALSE    28    6
  TRUE      5   14
```

2) Rappeler l'expression de l'indice AIC puis expliquer le principe de la sélection pas à pas utilisée ci-dessus.

3) Commenter les résultats du modele2.

4) Comparer la qualité des deux modèles obtenus.

5) En utilisant le modele2, classer l'individu ayant les caractéristiques suivantes : (age=61 ; acide=0.70 ; rayonx= 1 ; taille= 0 ; grade = 1 ; log.acid= -0.51)

6) Si vous deviez déterminer le modèle le plus performant en utilisant la régression logistique comment procèderiez-vous ?

Exercice 2 : On considère le tableau de données ci-dessous contenant les valeurs observées de deux variables quantitatives x^1 et x^2, et d'une variable qualitative y possédant les trois modalités A, B et C, sur un échantillon I de dix individus notés $i_1, \ldots, i_{10}$. Par la suite, on cherche à expliquer y en fonction de x^1 et x^2.

	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}
x^1	-1	0	-2	1	2	0	0	2	2	0
x^2	0	1	2	0	-1	-2	0	0	2	2
y	A	A	A	B	B	B	C	C	C	C

On applique la commande : `mod2 <- rpart(Y~., don, minsplit=3)`
où don désigne le tableau des données enregistré sous R. Le résultat est un arbre de décision que l'on peut décrire de la façon suivante où chaque noeud est identifié par une valeur de m :

- La racine, noeud $m = 1$, est divisée en deux régions : $\{x^1 < -0.5\}$ $(m = 2)$ et $\{x^1 \geq -0.5\}$ $(m = 3)$.

- Le noeud $m = 3$ est divisé en deux régions : $\{x^2 < -0.5\}$ $(m = 4)$ et $\{x^2 \geq -0.5\}$ $(m = 5)$.

- Le noeud $m = 5$ est divisé en deux régions : $\{x^1 < 1.5\}$ $(m = 6)$ et $\{x^1 \geq 1.5\}$ $(m = 7)$.

- Le noeud $m = 6$ est divisé en deux régions : $\{x^1 \geq 0.5\}$ $(m = 8)$ et $\{x^1 < 0.5\}$ $(m = 9)$.

1) Dessiner l'arbre de décision ainsi défini. Pour chaque noeud non terminal, on indiquera la coupure utilisée pour diviser ce noeud. Pour chaque noeud terminal, on indiquera les individus appartenant à ce noeud et la classe d'affectation des individus qui appartiennent à ce noeud.

2) Sachant que la commande `rpart` utilise l'indice de Gini pour construire l'arbre, quelle est la qualité de la coupure (appelée aussi réduction de l'impureté) effectuée pour diviser le noeud $m = 1$?