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 : 3 janvier 2019

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 deux modalités 0 et 1, sur un échantillon I de huit individus notés $i_1, \ldots, i_8$. 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
x^1	-2	2	-1	-1	-1	0	-1	2
x^2	0	-2	1	-1	-2	0	2	2
y	0	1	0	0	1	0	1	1

Par la suite, on applique différentes méthodes de classification supervisée à ces données afin d'expliquer y en fonction de x^1 et x^2. Pour cela, on utilise les commandes du logiciel R.

1- Calculer le centre de gravité g.

2- Calculer le tableau obtenu après centrage des variables x^1 et x^2 puis calculer les centres de gravité des classes associées à chaque modalité après centrage.

A- Analyse factorielle discriminante

On effectue l'AFD linéaire du tableau de données avec la commande `lda`.
3- Sachant que le facteur discriminant a pour coordonnées :

```
X1 -0.7385489
X2 0.0000000
```

Compléter la liste suivante qui indique les scores de chaque individu (un score manquant étant signalé par un " ?") :

```
i1 ?
i2 ?
i3 -0.5539117
i4 -0.5539117
i5 -0.5539117
i6  ?
i7 -0.5539117
i8 ?
```

4- Indiquer les scores des 2 centres de gravité.
5- En tenant compte des résultats précédents et sachant les probabilités *a posteriori* suivantes, déterminer de deux manières différentes la classe d'affectation de chacun des individus.

```
           0         1
i1 0.8071845 0.1928155
i2 0.1369438 0.8630562
i3 0.6487699 0.3512301
i4 0.6487699 0.3512301
i5 0.6487699 0.3512301
i6 0.4490412 0.5509588
i7 0.6487699 0.3512301
i8 0.1369438 0.8630562
```

6- Construire la matrice de confusion puis déterterminer le taux d'individus bien classés.

B- Régression logistique

On considère les résultats de la classification supervisée réalisée à l'aide d'une régression logistique. Plus précisément, on considère deux modèles de régression logistique : le premier, noté "modèle 1", expliquant la variable y par les variables x^1 et x^2, et le second, noté "modèle 2", expliquant y uniquement par la variable x^1. L'ajustement de ces deux modèles a conduit aux résultats suivants :

modèle 1 :

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 4.193e-01  9.940e-01   0.422    0.673
X1          1.051e+00  8.391e-01   1.252    0.211
X2          1.410e-16  5.958e-01   0.000    1.000

> anova(modele1,test = "Chisq")

     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    7    11.0904
X1    1   2.6532         6     8.4371   0.1033
X2    1   0.0000         5     8.4371   1.0000
```

modèle 2 :

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.4193     0.9940   0.422    0.673
X1            1.0507     0.8391   1.252    0.211

> anova(modele2,test = "Chisq")

     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    7    11.0904
X1    1   2.6532         6     8.4371   0.1033
```

7- Expliquer pourquoi le modèle 2 est préférable au modèle 1.
8- En utilisant le modèle 2, compléter les valeurs manquantes de la liste suivante :

```
> predict(modele2,type="response")
i1        i2        i3        i4        i5        i6        i7        i8
 ?         ? 0.3471799 0.3471799 0.3471799 0.6033092 0.3471799         ?
```

9- En déduire le taux global de bien classés obtenus à l'aide du Modèle 2.

C- Arbre de décision

On considère les résultats de la classification supervisée réalisée à l'aide de l'arbre de décision. L'arbre obtenu est présenté ci-dessous :

```
> modele_Arbre
n= 8

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

1) root 8 4 0 (0.5000000 0.5000000)
  2) X1< 1 6 2 0 (0.6666667 0.3333333)
    4) X2>=-1.5 5 1 0 (0.8000000 0.2000000)
      8) X2< 1.5 4 0 0 (1.0000000 0.0000000) *
      9) X2>=1.5 1 0 1 (0.0000000 1.0000000) *
    5) X2< -1.5 1 0 1 (0.0000000 1.0000000) *
  3) X1>=1 2 0 1 (0.0000000 1.0000000) *
```

10- Rappeler les 2 principaux paramètres de réglage de la taille d'un arbre.
11- Déterminer la classe prédite de chacun des huit individus et en déduire le taux global de bien classés selon cet arbre
12- Expliquer pourquoi les taux globaux de bien classés calculés précédemment (cf. questions 6, 9 et 11) ne sont pas suffisants pour déterminer la meilleure procédure de classification supervisée parmi celles qui ont été envisagées dans ce problème.