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 : 16 février 2021

Exercice 1 : On a effectué une enquête sur la relation des consommateurs vis-à-vis des magasins Champion. Un questionnaire a ainsi été administré à un échantillon représentatif de 60 clients. Le questionnaire qui a été administré dans le cadre de cette enquête est présenté à l'Annexe 1. Nous avons appelé `donnees_champion` le data frame contenant les réponses à ce questionnaire. Dans la suite, à la question numéro i on associe la variable statistique notée Q_i.

Partie I

On a effectué une classification hiérarchique des 60 clients sur les 8 items $(Q_{1.a}, ..., Q_{1.h})$ de la première question. La hiérarchie issue de cette classification est présentée à l'Annexe 2.

1) Déterminer, en justifiant votre réponse, la meilleure partition à retenir de cette hiérarchie.

2) Nous nous intéressons à la partition en 4 classes issue de cette hiérarchie. En utilisant la fonction `catdes` nous avons obtenu la description donnée ci-dessous des 4 classes obtenues. Commenter ces résultats.

```
Description of each cluster by quantitative variables
======================================================
$`1`
      v.test Mean in category Overall mean sd in category Overall sd      p.value
Q1c -3.195348         2.434783     2.783333      0.5768043  0.6605974 1.396622e-03
Q1d -5.256365         2.000000     2.650000      0.4170288  0.7488881 1.469306e-07

$`2`
      v.test Mean in category Overall mean sd in category Overall sd      p.value
Q1d  2.723147         3.055556     2.650000      0.6211300  0.7488881 0.0064663311
Q1b  2.589630         3.055556     2.700000      0.5241101  0.6904105 0.0096079183
Q1e -2.183443         2.166667     2.516667      0.7637626  0.8060535 0.0290032251
Q1h -3.880884         1.611111     2.183333      0.4874980  0.7414325 0.0001040774

$`3`
     v.test Mean in category Overall mean sd in category Overall sd      p.value
Q1e 4.642291         3.857143     2.516667      0.3499271  0.8060535 3.445676e-06
Q1c 2.122919         3.285714     2.783333      0.4517540  0.6605974 3.376064e-02
```

```
$`4`
       v.test Mean in category Overall mean sd in category Overall sd      p.value
Q1h  5.093603         3.166667     2.183333      0.3726780  0.7414325 3.513213e-07
Q1e -2.461742         2.000000     2.516667      0.4082483  0.8060535 1.382640e-02
```

3) Si on voulait effectuer une classification en utilisant l'ensemble des variables décrivant les individus, quelle serait la démarche à suivre.

Partie II

On voudrait expliquer la variable Q_4, appelée dans la suite `satisfaction`, par les variables `sexe`, `csp`,$Q_{1.a}$, ..., $Q_{1.h}$ à l'aide d'un arbre de décision. Les résultats obtenus sont présentés ci-dessous :

```
> arbre.full <- rpart(satisfaction~ ., data = donnees_champion, method = "class",
minsplit=10, cp =0.056)
> print(arbre.full)
n= 60

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

 1) root 60 27 0 (0.5500000 0.4500000)
   2) Q1a< 3.5 56 23 0 (0.5892857 0.4107143)
     4) Q1e< 1.5 5  0 0 (1.0000000 0.0000000) *
     5) Q1e>=1.5 51 23 0 (0.5490196 0.4509804)
      10) Q1d>=3.5 7  1 0 (0.8571429 0.1428571) *
      11) Q1d< 3.5 44 22 0 (0.5000000 0.5000000)
        22) Q1d< 2.5 24  9 0 (0.6250000 0.3750000) *
        23) Q1d>=2.5 20  7 1 (0.3500000 0.6500000) *
   3) Q1a>=3.5 4  0 1 (0.0000000 1.0000000) *
```

4) Commenter la commande

```
arbre.full <- rpart(satisfaction~ ., data = donnees_champion, method = "class",
minsplit=10, cp =0.056)}
```

5) Déterminer les règles issues de cet arbre.

6) Commenter ces règles.

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 deux modalités notées A et B, sur un échantillon I de dix individus notés $P_1, \ldots, P_{10}$. Chaque individu est muni du poids 1/10.

Par la suite, on applique une analyse discriminante à ces données afin d'expliquer Y en fonction de X^1 et X^2.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
X^1	0	0	1	1	2	2	1	0	2	1
X^2	0	0	1	1	2	2	0	1	1	2
Y	A	A	B	B	A	A	A	B	A	B

1) On a effectué l'AFD linéaire du tableau de données. Combien d'axes factoriels discriminants (non triviaux) existe-t-il ?

2) Quelle est la commande de R qui permet d'appliquer une AFD linéaire aux données. On précisera les arguments nécessaires pour cette fonction.

3) Les probabilités *a posteriori* et les scores des individus données par l'AFD linéaire du tableau de données sont donnés ci-dessous :

```
$posterior
           A          B
1  0.65555784 0.34444216
2  0.65555784 0.34444216
3  0.65555784 0.34444216
4  0.65555784 0.34444216
5  0.65555784 0.34444216
6  0.65555784 0.34444216
7  0.97070777 0.02929223
8  0.09853739 0.90146261
9  0.97070777 0.02929223
10 0.09853739 0.90146261

$x
              LD1
1  -6.661338e-16
2  -6.661338e-16
3   0.000000e+00
4   0.000000e+00
5   6.661338e-16
6   6.661338e-16
7  -1.851640e+00
8   1.851640e+00
9  -1.851640e+00
10  1.851640e+00
```

Construire la matrice de confusion puis calculer les taux d'erreurs dans chaque classe ainsi que le taux d'erreur global.

4) Indiquer la commande de R qui permet d'appliquer une AFD quadratique aux données.

Annexe 1 : Questionnaire

1. Veuillez cocher la case qui correspond le plus à votre jugement :

	1	2	3	4	5
1.a La modernité de l'equipement et le mobilier du magasin					
1.b L'attractivité et le design du magasin					
1.c La propreté des différents services offerts dans le magasin					
1.d La disponibilité des marchandises à temps pour la clientèle					
1.e La disponibilité du personnel à répondre aux questions					
1.f Votre degré de confiance à l'égard du personnel					
1.g La variété des marchandises					
1.h La qualité du service après vente					

où (1) = mauvais(e), (2) = moyen(ne) , (3) = normal(e), (4) = acceptable et (5) = excellent(e)

2. Catégorie socioprofessionnelle :
   Retraité ... Cadre ... Ouvrier ... Profession libérale ...

3. Sexe : Homme ... Femme ...

4. Etes-vous satisfait de Champion ? Oui ... Non ...

Annexe 2