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 : 31 janvier 2019

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. Un extrait du questionnaire qui a été administré dans le cadre de cette enquête est présenté à l'Annexe 1.

N.B. : Dans la suite, à la question numéro i on associe la variable statistique notée Q_i.

A- Analyse en Composantes Principales

On a effectué une Analyse en Composantes Principales (ACP) sur les 9 items $(Q_{1,j}, j \in \{1, ..., 9\})$ de la première question. Les résultats de cette ACP sont présentés à l'Annexe 2.

1- Déterminer le nombre d'axes à retenir.

Dans la suite, on suppose que l'on retient les 4 premières composantes principales.

2- Justifier l'intérêt de faire une rotation à l'issu de cette ACP.

3- Donner une interprétation des axes retenus.

Dans la suite, les 4 premières composantes principales de l'ACP seront appellées `dim1`, `dim2`, `dim3`, `dim4`.

B- Arbre de décision

Dans cette partie, on voudrait expliquer la variable Q_7, appelée dans la suite `satisfaction`, par les variables `revenu`, `sexe`, `csp`, `dim1`, `dim2`, `dim3` et `dim4` à l'aide d'un arbre de décision. Les résultats obtenus sont présentés ci-dessous :

```
> arbre.full <- rpart(satisfaction~ revenu+sexe+csp+dim1+dim2+dim3+dim4,
 data = donnees_champion, method = "class")
> print(arbre.full)
n= 60
node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 60 27 0 (0.55000000 0.45000000)
    2) revenu>=725 23  2 0 (0.91304348 0.08695652) *
```

```
   3) revenu< 725 37 12 1 (0.32432432 0.67567568)
     6) dim3< -1.06706 5  1 0 (0.80000000 0.20000000)
      12) dim4< 0.6927249 4  0 0 (1.00000000 0.00000000) *
      13) dim4>=0.6927249 1  0 1 (0.00000000 1.00000000) *
     7) dim3>=-1.06706 32  8 1 (0.25000000 0.75000000)
      14) dim1< -1.795057 2  0 0 (1.00000000 0.00000000) *
      15) dim1>=-1.795057 30  6 1 (0.20000000 0.80000000)
        30) dim4< -0.8938268 3  1 0 (0.66666667 0.33333333) *
        31) dim4>=-0.8938268 27  4 1 (0.14814815 0.85185185)
          62) dim2>=1.359888 1  0 0 (1.00000000 0.00000000) *
          63) dim2< 1.359888 26  3 1 (0.11538462 0.88461538)
           126) dim4>=2.104005 1  0 0 (1.00000000 0.00000000) *
           127) dim4< 2.104005 25  2 1 (0.08000000 0.92000000)
             254) dim2< -0.4813271 7  2 1 (0.28571429 0.71428571)
               508) dim2>=-0.9902607 3  1 0 (0.66666667 0.33333333) *
               509) dim2< -0.9902607 4  0 1 (0.00000000 1.00000000) *
             255) dim2>=-0.4813271 18  0 1 (0.00000000 1.00000000) *
> printcp(arbre.full)

Classification tree:
rpart(formula = satisfaction ~ ., data = donnees_champion, method = "class")

Variables actually used in tree construction:
[1] dim1   dim2   dim3   dim4   revenu

Root node error: 27/60 = 0.45

n= 60

        CP nsplit rel error  xerror    xstd
1 0.481481      0   1.00000 1.00000 0.14272
2 0.111111      1   0.51852 0.59259 0.12687
3 0.074074      2   0.40741 0.74074 0.13524
4 0.037037      3   0.33333 0.55556 0.12423
5 0.018519      7   0.18519 0.62963 0.12928
6 0.010000      9   0.14815 0.66667 0.13147

> pred <- predict(arbre.full, newdata = donnees_champion, type = "class")
> mc <- table(donnees_champion$satisfaction,pred)
> print(mc)
   pred
     0  1
  0 33  0
  1  4 23
```

4- Rappeler le principe qui permet d'obtenir l'arbre optimal.

Par la suite, on a procédé à l'élagage de `arbre.full`, le résultat est donné ci-dessous :

```
> arbre.full.prune<-prune(?)
> print(arbre.full.prune)
```

```
n= 60
node), split, n, loss, yval, (yprob)
      * denotes terminal node
 1) root 60 27 0 (0.55000000 0.45000000)
   2) revenu>=725 23  2 0 (0.91304348 0.08695652) *
   3) revenu< 725 37 12 1 (0.32432432 0.67567568)
     6) dim3< -1.06706 5  1 0 (0.80000000 0.20000000) *
     7) dim3>=-1.06706 32  8 1 (0.25000000 0.75000000)
      14) dim1< -1.795057 2  0 0 (1.00000000 0.00000000) *
      15) dim1>=-1.795057 30  6 1 (0.20000000 0.80000000) *

> pred.prune <- predict(arbre.full.prune, newdata = donnees_champion, type = "class")
> mc.prune <- table(donnees_champion$satisfaction,pred.prune)
> print(mc.prune)
   pred.prune
     0  1
  0 27  6
  1  3 24
```

5- Compléter la commande `prune` par les paramètres adéquats afin d'obtenir `arbre.full.prune`.

6- Déterminer les régles issues de `arbre.full.prune`.

7- Que peut-on conclure quant aux variables explicatives de la satisfaction d'un client ?

8- Comparer les taux d'erreur des deux arbres.

9- Quel arbre choisiriez-vous ? Justifier votre réponse.

Annexe 1 : Extrait du questionnaire

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

	1	2	3	4	5
1.1 La modernité de l'equipement et le mobilier du magasin					
1.2 L'attractivité et le design du magasin					
1.3 La propreté des différents services offerts dans le magasin					
1.4 La disponibilité des marchandises à temps pour la clientèle					
1.5 La disponibilité du personnel à répondre aux questions					
1.6 La sécurité des transactions dans le magasin					
1.7 Votre degré de confiance à l'égard du personnel					
1.8 La variété des marchandises					
1.9 La qualité du service après vente					

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

2. Le nombre de fois par semaine où vous fréquentez Champion ...
3. Le nombre de produits achetés auprès de Champion par semaine ....
4. Quel est votre revevu ?
5. Catégorie socioprofessionnelle :
   Retraité ... Cadre ... Ouvrier ... Profession libérale ...
6. Sexe : Homme ... Femme ...
7. Etes-vous satisfait de Champion ? Oui ... Non ...

Annexe 2 : Résultats de l'ACP

Les 6 premières valeurs propres

Composantes	Valeurs propres
1	2,02
2	1,41
3	1,27
4	1,09
5	0,99
6	0,74

Matrice des composantes

	Composante			
	1	2	3	4
1. La modernité de l'equipement et le mobilier du magasin	,845	-,139	-,142	,165
2. L'attractivité et le design du magasin	,648	-,353	,297	-,148
3. La propreté des différents services offerts dans le magasin	,706	-,021	-,289	,313
4. La disponibilité des marchandises à temps pour la clientèle	,300	,373	-,298	-,447
5. La disponibilité du personnel à répondre aux questions	,027	,314	,608	,347
6. La sécurité des transactions dans le magasin	-,040	-,022	,622	-,062
7. Votre degré de confiance à l'égard du personnel	,514	,297	,461	-,382
8. La variété des marchandises	,027	,730	-,142	-,330
9. La qualité du service après vente	,104	,641	-,032	,605

Matrice des composantes après rotation

	Composante			
	1	2	3	4
1. La modernité de l'equipement et le mobilier du magasin	,882	,052	-,018	,015
2. L'attractivité et le design du magasin	,619	-,056	,390	-,342
3. La propreté des différents services offerts dans le magasin	,775	,039	-,195	,201
4. La disponibilité des marchandises à temps pour la clientèle	,153	,687	-,122	-,094
5. La disponibilité du personnel à répondre aux questions	-,046	-,128	,577	,489
6. La sécurité des transactions dans le magasin	-,126	-,130	,599	-,038
7. Votre degré de confiance à l'égard du personnel	,287	,474	,631	-,072
8. La variété des marchandises	-,176	,757	,001	,241
9. La qualité du service après vente	,095	,107	-,027	,876