Examen
Systèmes d'Information Décisionnels
RICM5 - 2014-2015

Ahlame Douzal
Durée : 2 heures, documents autorisés

21-janvier-2014

1 Quelques questions de cours ! ! (4 pts)

1. Précisez le rôle de la puissance de Minkowski ?
2. Indiquez les principales différences entre les coefficients de corrélation de Kendall et Spearman ?
3. Comment distinguer une variable binaires non symétrique ? Illustrez votre explication par un exemple.
4. Expliquez les principales stratégies adoptées pour mesurer la proximité entre deux variables binaires non symétriques ?
5. En classification non supervisée par partitionnement, comment procède-t-on pour déterminer le bon nombre de classes ?
6. En classification non supervisée hiérarchique, comment procède-t-on pour déterminer la bonne partition ?
7. Comment évaluer la distance entre deux individus décrits par des variables hétérogènes ? Donnez un exemple illustrant votre explication.
8. Comment se fait la prédiction de la classe d'un individu dans un modèle de classification par arbre ?

2 Analyse des admissions à Berkeley (4 pts)

L'objectif de cette étude est l'analyse de la dépendance entre les résultats des admissions à l'université de Californie à Berkeley et le sexe des candidats. Nous considérons les données recensées dans la base *UCBAdmissions*[1].

1. Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975) Sex bias in graduate admissions : Data from Berkeley. Science, 187, 398–403.

UCBAdmissions donne la description de 4526 candidatures décrites par 3 variables : "Admit" indiquant l'acceptation ou le rejet de la candidature (variable binaire : Admitted vs. Rejected), la variable "Gender" donnant le sexe du postulant (binaire : Male vs. Female) et "Dept" précisant le département visé par le candidat (variable nominale : A, B, C, D, E, F).

La répartition des candidatures selon le sexe des candidats et le taux d'acceptation correspondant est indiquée dans le tableau (Table 1) :

Gender	Nb. candidatures	Nb. retenues	% Acceptation
Male	2691	1198	44.5%
Female	1835	557	30.4%

TABLE 1 –

Le taux d'acceptation des candidatures peut corroborer l'hypothèse d'une dépendance forte entre les variables "Gender" et "Admit". Cependant, cette hypothèse peut être remise en cause dans le cas d'une dépendance forte entre les départements visés et les taux de rejet. En effet, un taux de rejet élevé chez les candidates peut s'expliquer par le faite qu'elles visent essentiellement des départements à fort taux de rejet. Pour valider ces hypothèses, observons la distribution jointe des variables "Gender" et "Admit" par département de la table 2.

Dept.		Admitted	Rejected
All	Male	1198	1493
	Female	557	1278
A	Male	512	313
	Female	89	19
B	Male	353	207
	Female	17	8
C	Male	120	205
	Female	202	391
D	Male	138	279
	Female	131	244
E	Male	53	138
	Female	94	299
F	Male	22	351
	Female	24	317

TABLE 2 –

Questions

1. Sur la base de la distribution globale donnée ligne 1 de la table 2, acceptez-vous l'hypothèse de dépendance entre les taux d'acceptation et le sexe des

candidats à un risque de 5 % ? (la valeur du χ_2 à 1 degré de liberté et à un risque 5% est de 3.48).

2. Acceptez-vous cette même hypothèse à un risque de 5 % pour les candidatures visant le département F (à fort taux de rejet) ?

2.1 Classification supervisée (12 pts)

Les données de cette étude [2] porte sur la description pour 53 patients atteint d'un cancer de la prostate d'un ensemble de paramètres permettant d'orienter le type de traitement à mettre en place, en particulier : on y indique l'age du patient, le niveau d'acide phosphatase sérique (acide), le résultat d'une analyse par rayon X, 0=négatif, 1 positive (rayonx), la taille de la tumeur, 0=petite, 1=grande(taille), l'état de la tumeur, 0=moyen, 1 grave (grade), le log du niveau d'acidité (log.acid).

Nous procédons dans un premier temps à une analyse descriptive des données :

```
Data<- read.table("cancerprostate.txt", header=TRUE, sep=';')
for(i in 3:6)
Data[,i]<-factor(Data[,i])
> summary(Data)
> summary(Data)
      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
```

1) Interprétez les résultats : 3rd Qu. :0.7800 pour la variable acide et 1 :27 pour taille.

Nous réalisons ensuite une classification par arbre pour la prédiction de la diffusion des cellules cancéreuses dans la région lymphatique :

```
> library(rpart)
> prostate.arbre<-rpart(Y~., data=Data)
> prostate.arbre
n= 53
```

2. http ://www.agrocampus-ouest.fr/math/livreR/cancerprostate.txt

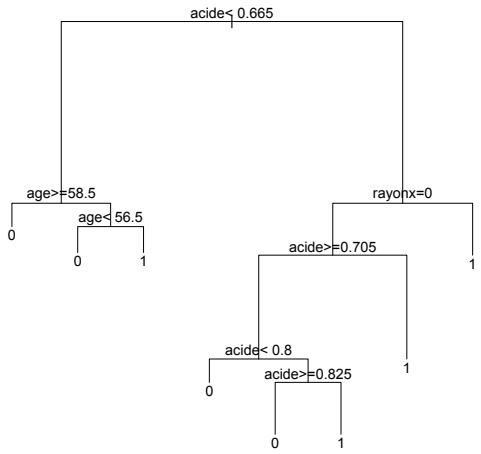
```
node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 53 20 0 (0.6226415 0.3773585)
  2) acide< 0.665 28  4 0 (0.8571429 0.1428571) *
  3) acide>=0.665 25  9 1 (0.3600000 0.6400000)
    6) rayonx=0 16  7 0 (0.5625000 0.4375000) *
    7) rayonx=1 9  0 1 (0.0000000 1.0000000) *
```

2) Interprétez chacun des noeuds de l'arbre
3) Estimer la probabilité d'une bonne prédiction par l'arbre induit

En indiquant le nombre minimum de données au sein de chaque noeud de l'arbre, un arbre plus détaillé est obtenu

```
> library(rpart)
> prostate.arbre2<-rpart(Y~., data=Data, minsplit=5)
> plot(prostate.arbre2)
>text(prostate.arbre2, pretty=0)
```



4) Extraire et interpréter toutes les règles de décisions permettant de prédire le niveau de diffusion des cellules cancéreuses.