UFR IMA
Informatique & Mathématiques Appliquées

UNIVERSITE JOSEPH FOURIER GRENOBLE 1
SCIENCES, TECHNOLOGIE, MEDECINE
Sciences, Technologie, Médecine

Master 2 P Génie Informatique
Systèmes d'Information Décisionnels

TP5: Classification non-supervisée/supervisée

Partie 1 : Analyse non supervisée

Considérons la base SPAM7 (http://www.ics.uci.edu/~Here) donnant la description de 4601 emails dont les 1813 premiers identifiés de type spam. Chaque email est décrit par les variables suivantes :

Crl.tot : longueur totale des mots en capital
dollar : nombre d'occurrence du symbole '$'
bang : nombre d'occurrence du symbole '!'
money : nombre d'occurrence du mot `money'
n000 : nombre d'occurrence de la chaine `000'
make : nombre d'occurrence du mot `make'
yesno : type de l'email : "n" pour non spam, "y" pour spam.

Nous listons un petit extrait de la base « spam7» portant sur les 5 premiers et 5 derniers emails :

crl.tot dollar bang money n000 make yesno

1 278 0.000 0.778 0.00 0.00 0.00 y
2 1028 0.180 0.372 0.43 0.43 0.21 y
3 2259 0.184 0.276 0.06 1.16 0.06 y
4 191 0.000 0.137 0.00 0.00 0.00 y
5 191 0.000 0.135 0.00 0.00 0.00 y
……

4597 88 0.000 0.000 0.00 0.00 0.31 n
4598 14 0.000 0.353 0.00 0.00 0.00 n
4599 118 0.000 0.000 0.00 0.00 0.30 n
4600 78 0.000 0.000 0.00 0.00 0.96 n
4601 40 0.000 0.125 0.00 0.00 0.00 n

a) Interprétez les résultats

```
> summary(spam7)
> spam.sample <- spam7[sample(seq(1,4601),500, replace=FALSE), ]
> par(mfrow=c(2,3))
> boxplot(split(spam.sample$crl.tot,spam.sample$yesno), main="crl.tot")
> boxplot(split(spam.sample$dollar,spam.sample$yesno), main="dollar")
> boxplot(split(spam.sample$bang,spam.sample$yesno), main="bang")
> boxplot(split(spam.sample$money,spam.sample$yesno), main="money")
> boxplot(split(spam.sample$n000,spam.sample$yesno), main="n000")
> boxplot(split(spam.sample$make,spam.sample$yesno), main="make")
```

b) Procédez au partitionnement des données spam7 par Kmeans, Hclust et pam.
c) Analysez, interprétez et comparez les résultats.
d) Visualisez l'évolution de la fonction critère et justifiez le nombre de classes retenu.

Partie 2 : Analyse supervisée

Notre objectif est de pouvoir prédire la nature spam ou pas d'un email sur la base des nombres d'occurrence des symboles « *dollar », « bang », « money », « n000 », et « make »*. Pour cela, un arbre de classification est construit.

```
> S<-spam7
> sub <- c(sample(1:max(which(S$yesno=="y")), round(0.8*max(which(S$yesno=="y")),digits=0)),
   sample((max(which(S$yesno=="y"))+1):4601,round(0.8*length(which(S$yesno=="n")),digits=0)))
> fit <- rpart(S$yesno~ ., data=S, subset=sub)
>plot(fit)
>text(fit)
>table(predict(fit, S[-sub,], type="class"), S[-sub, "yesno"])
```

a) Expliquez l'effet des six commandes ci-dessus et interprétez les résultats obtenus.
c) Extrayez l'ensemble des règles permettant d'identifier la nature d'un email.
d) Évaluez le taux de mauvais classement par validation-croisée.
Pour cela, partitionnez aléatoirement la base de données « spam7 » en k=10 sous-ensembles (K-fold stratifiées). Les 10 sous-ensembles sont combinés afin de constituer 10 échantillons d'apprentissage. Par exemple, un échantillon d'apprentissage est composé de 9 sous-ensembles, le dernier étant réservé pour l'évaluation du taux d'erreurs.