Université Paris X - Nanterre
Master ISIFAR, M1
Mathilde Mougeot - Hélène Boistard

Analyse des données
Novembre 2007

TD 7 - Classification hiérarchique

But du TD : connaître des méthodes de classification hiérarchique sous **R** pour des données quantitatives.

1 Exemple d'étude simple : données Animals de la librairie MASS

- Récupérer les données Animals de la librairie MASS et les renommer A. Il est préférable de travailler sur les données centrées réduites, dans cet exemple. Rappeler pourquoi. Nommer As les données centrées réduites obtenues avec la commande scale. Nommer Ad le tableau des distances euclidiennes des données centrées réduites, obtenu avec la fonction dist. Exécuter une classification à l'aide de la fonction hclust, avec la méthode "ward". Appeler le résultat de cette classification : Ac. Pour visualiser le dendogramme : appliquer plclust sur Ac.
- Exécuter la commande cutree(Ac,5). A quoi cela correspond-il ?
- On veut maintenant choisir de manière adaptative le nombre de classes retenues. On va employer 2 méthodes basées sur la comparaison variance inter/variance intra. Voici quelques rappels :

$$Vinter = \sum_{c=1}^{k} \frac{n_c}{n} \frac{1}{n_c} \sum_{j=1}^{p} \left(\bar{X}_c^j - \bar{X}^j\right)^2,$$

$$Vintra = \sum_{c=1}^{k} \frac{n_c}{n} \sum_{j=1}^{p} \sum_{i=1}^{n_c} \left(X_{ic}^j - \bar{X}_c^j\right)^2,$$

avec les notations du cours.
Le code suivant permet de calculer la variance inter-classe et la variance intra-classe de la première variable lorsque le nombre de classes choisi est 3. C'est-à-dire :

$$\sum_{c=1}^{3} \frac{n_c}{n} \frac{1}{n_c} \left(\bar{X}_c^1 - \bar{X}^3\right)^2 \text{ et } Vintra = \sum_{c=1}^{3} \frac{n_c}{n} \sum_{i=1}^{n_c} \left(X_{ic}^1 - \bar{X}_c^1\right)^2.$$

Effectuer ces commandes. Puis, modifier le code pour calculer la variance inter-classe et la variance intra-classe sur l'ensemble des variables et pour un nombre de classes variant de 1 à $n = 28$. Pour cela, rajouter convenablement deux boucles for imbriquées.

```
n=dim(As)[1]
#Pour la variable j=1 et le nombre de classes k=3:
#On considère la colonne As[,j]
j=1; k=3;
classe<-cutree(Ac,1:n)
classek<-classe[,k]
data<-As[,j]
#variance inter classe :
moycl<-tapply(data, classek, mean)
longcl<-tapply(data,classek,length)
Vinter<-sum(longcl*(moycl-mean(data))^2/length(data))
#variance sur chaque classe :
variance<-function(vect)
{
l<-length(vect)
(l-1)/l*var(vect)
}
x<-tapply(data,classek, variance)
x[is.na(x)]<-0
#Variance intra-classe :
Vintra<-sum(longcl*x/length(data))
#verif eq analyse de la variance :
VT<-(n-1)/n*var(data)
Vinter+Vintra
```

- Tracer sur un graphique l'évolution de la variance inter-classe et de la variance intra-classe en fonction du nombre de classes. Utiliser la commande matplot(1 :n, cbind(Vinter, Vintra),pch="o"). Déterminer le nombre de classes obtenu lorsqu'on choisit l'équilibre variance inter/variance intra. Avec la fonction cutree, déterminer les classes. Retrouver les classes sur le dendogramme.
- Déterminer le nombre de classes obtenu lorsqu'on choisit la règle du rapport variance inter-classe sur variance totale supérieur à 95%. Déterminer les classes. Les retrouver sur le dendogramme.
- Tracer pour chacune des deux méthodes un graphique représentant les données Animals, avec une couleur pour chaque groupe obtenu. Mettre ces deux graphiques sur une même fenêtre. Commenter les différentes classes obtenues. On constate qu'on a isolé 3 espèces d'animaux complètement atypiques. Que se passe-t-il si on retire ces animaux et que l'on refait une classification ?

2 Classification sur les données USArrests

- Récupérer les données USArrests, les décrire. On va classifier les variables 1, 2 et 4. On centre et réduit d'abord les données. Justifier : que se passe-t-il si on ne les réduit pas ?
- Appliquer la procédure décrite au paragraphe 1 pour faire une classification hiérarchique des données. Représenter les données en les projetant sur les axes correpondant à deux variables, pour tous les choix de couples de deux variables. Utiliser différentes couleurs pour les différents groupes de la classification.
- Effectuer une analyse en composantes principales des données centrées réduites (toujours en se restreignant aux variables 1, 2 et 4). Faire un résumé de l'ACP pour ces données (axes, valeurs propres, etc.) Représenter les groupes obtenus lors de la classification précédente sur un graphique où les données sont projetées sur les première et deuxième composantes principales, puis sur les première et troisième composantes.

3 Classification sur les données iris

Récupérer les données iris. On veut classifier les 4 variables quantitatives. Suivre le même schéma que dans la question précédente.