Analyse de données
M1 Statistique et économétrie - 2012
V. Monbet
Classification

Partie 1 : Étude des caractéristiques d'un ensemble d'hôtels

Le fichier `hotels.CSV` contient des caractéristiques liées au confort, à la qualité et à la situation d'un ensemble d'hôtels. Importer les données sous `R`.

Quelles sont les différentes variables ? Quelle est leur nature ? Qui sont les individus sur qui on va faire porter la classification ? Obtenir les statistiques descriptives, les covariances et les corrélations entre les variables quantitatives du jeu de données.

Classification hiérarchique ascendante

Faire la classification hiérarchique ascendante des observations en utilisant les distances eulidienne et Manhattan et les liaisons simple, complète et de Ward. Dans chacun des cas, représenter le dendogramme, proposer un nombre de classes, lister les hôtels (et les pays) de chaque classe, donner la moyenne de chaque classe et interpréter les résultats.

Par exemple, si `hotels` est la table de données :

```
Pays = hotels[,1]
hotelsnum = hotels[,-1]
d.h = dist(hotelsnum,"euclidean")
cah.h = hclust(d.h,"ward")
str(cah.h)
plot(cah.h,hang =-1)
dev.new()
plot(cah.h$height)
nbcl = 3
```

```
split(rownames(hotelsnum),cutree(cah.h,k=nbcl))
(s.pays = split(Pays,cutree(cah.h,k=nbcl))
pca.h = PCA(hotelsnum)
plot(pca.h$ind$coord[,1:2], pch=21, bg=c("red", "green", "blue")[cutree(cah.h,k=nbc
```

Obtient-on les mêmes résultats si on réduit les données ?

Quelle est la meilleure classification ? Pourquoi ?

K-moyennes

1. Obtenir la classification des hôtels en trois groupes à l'aide de la méthode des K-moyennes qui portera sur toutes les variables numériques du tableau. Représenter graphiquement les trois groupes sur les premiers plans factoriels de l'ACP. Qu'observe t'on ? Lister les hôtels (et les pays) de chaque classe, donner les centres de chaque classe et interpréter les résultats.

   Pour obtenir la classification, on fait par exemple

   ```
   nbcl = 3
   centers.init = hotelsnum[sample(1:length(hotelsnum[,1]),nbcl),] ;
   K = kmeans(hotelsnum, centers.init)
   ```

   Comment est initialisé l'algorithme ? Obtient-on toujours le même résultat quand on change l'initialisation ?
   Proposer une autre méthode d'initialisation.
   Obtient-on le même résultat si on travaille avec des données réduites ? centrées-réduites ?

2. Obtenir la classification des hôtels en trois groupes à l'aide de la méthode des K-moyennes qui portera sur les coordonnées des hôtels dans le premier plan factoriel. Représenter graphiquement les trois groupes sur les premiers plans factoriels de l'ACP. Qu'observe t'on ? Lister les hôtels (et les pays) de chaque classe, donner les centres de chaque classe et interpréter les résultats.

3. Quelles sont les différences de classement entre les classements des questions 1 et 2 précédentes ? Le premier plan factoriel traduit-il fidèlement l'ensemble des données ? (voir valeurs propres).

4. Sélection de modèle.
   Il est possible d'associer un critère BIC au modèle obtenu quand on fait de la classification par la méthode des moyennes mobiles. On peut en effet supposer qu'on obtient un modèle de mélange gaussien : dans chaque classe le vecteur des variables X suit une loi de Gauss de moyenne μ_k et de variance Σ_k. Ainsi la densité de X est donnée par

$$f_X(x) = \sum_{k=1}^{K} \pi_k \phi_k(x; \mu_k, \Sigma_k)$$

   avec π_k la probabilité a priori de la classe k.
   Estimer le critère BIC pour des modèles à $K = 2$ jusqu'à $K = 7$ classes. Quel est le meilleur modèle selon ce critère ?
   Estimer aussi pour chacun des modèles le rapport de l'inertie inter-classe sur l'inertie totale. Quel est le meilleur modèle selon ce critère ?
   Ces résultats sont-ils cohérents ?

5. On décide de vérifier si l'attribution des étoiles est conforme aux critères de constitution des groupes par la méthode des K-moyennes. Puisqu'il existe 6 catégories d'étoiles de 0 à 5, classer les hôtels en 6 groupes à l'aide de la méthode des K-moyennes portant cette fois ci sur toutes les variables à l'exclusion des variables prix et étoiles. Comparer la répartition des hotels dans les groupes et leur nombre d'étoiles. On peut utiliser la fonction `table`. Attention, les groupes ne sont pas nécéssairement numérotés par ordre croissant d'étoiles.

Partie 2 : Sélection de variables par classification

On dispose d'une base de données collectant certains facteurs qui sont susceptibles d'influencer la maladie cardiaque d'hommes du Western Cape, Afrique de Sud (voir `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`, South African Heart Disease).

Les variables renseignées sont les suivantes :

- `sbp` : pression sanguine systolique *(systolic blood pressure)*
- `tobacco` : la quantité de tabac consommé cumulée *(cumulative tobacco (kg))*
- `ldl` : taux de cholestérol dans le sang *(low density lipoprotein cholesterol)*
- `adiposity` : adiposité
- `famhist` : antécédents familiaux : `Present` s'il y a eu des antécédants familiaux *(family history of heart disease (Present, Absent))*
- `typea` : comportement de type A *(type-A behavior)*
- `obesity` : obésité
- `alcohol` : consommation courante d'alcool *(current alcohol consumption)*
- `age` : age au moment de l'attaque cardiaque *(age at onset)*
- `chd` : variable réponse codée 1 si la maladie du coeur est présente, 0 sinon *(response, coronary heart disease)*

Vous trouverez ces données sur la page web du cours.

Matrice de similarité

On calcule tout d'abord une matrice de similarité que l'on pourra utiliser pour faire de la classification et/ou du positionnement multidimensionnel. Ici on choisit d'utiliser le T de Tschuprow qui permet de mesurer la similarité entre deux variables qualitatives. Considérons deux variables qualitatives, X à r modalités et Y à s modalités. Le T de Tschuprow défini par

$$T = \sqrt{\frac{\chi^2_{\text{observé}}}{n\sqrt{\nu}}}$$

où $\nu = (r-1)(s-1)$ et

$$\chi^2_{\text{observé}} = n\left[\sum_{l=1}^{r}\sum_{h=1}^{s}\frac{n_{lh}^2}{n_{l+}n_{+h}} - 1\right]$$

Ici n_{lh} représente le nombre d'individus tels que $x = l$ et $y = h$, n_{l+} est le nombre d'individus tels que $x = l$ et n_{+h} le nombre d'individus tels que $y = h$. On observe

que T est une fonction croissante du $\chi_{\text{observé}}$. Et le premier terme du $\chi_{\text{observé}}$ donne une approximation empirique du rapport

$$\frac{P(X = x, Y = y))}{P(X = x)P(Y = y)}$$

On rappelle que si X et Y sont indépendantes (et donc distantes), ce rapport est égal à 1 et, en conséquence, le T de Tschuprow proche de 0.

L'éxecution de ce programme peut être fastidieuse car elle requiert de nombreux lancement de la `proc freq` (pour discrétiser les variables continues notamment). Il est conseillé de réfléchir en amont à la sélection de variables et de réduire un peu le jeu de données avant de lancer `%dtprow` quand le jeu de données est important (ce qui n'est pas le cas ici).

```
%let listev = sbp tobacco ldl adiposity famhist  typea
    obesity alcohol age chd ;
%dtprow(heart_disease,&listev,outprox = sasuser.dtschvp) ;
```

Classification

Une classification hiérarchique est ensuite opérée sur le tableau `dtschvp` obtenu par `dtprow`. L'idée est alors de sélectionner un sous-ensemble de variables en ne retenant, par exemple qu'une variable dans certains groupes ou encore de réaliser une analyse factorielle (afcm) par groupe pour aider au choix.

```
proc cluster data = sasuser.dtschvp method=ward outtree=tree  ;
var &listev ;
copy varname;

proc tree data=tree graphics hor out=chclasse ncluster = 5 ;
id varname ;
run;
```

Positionnement multidimensionnel

Un positionnement multidimensionnel des variables tenant compte de leurs distances respectives apporte une vision complémentaire :

```
proc mds data=sasuser.dtschvp shape=square out=result;
var &listev ;
object varname ;
```

```
run ;
proc sort data=result out = result1 ;
by varname ;
proc sort data =chclasse out= result2 ;
by varname ;
run ;
data resul;
merge result1 result2;
by varname ;
run ;
%couleur(resul,cluster) ;
%gafcx(ident=varname,nc=6,col=coul) ;
```

Choix de variables

Aidez vous des résultats précédents pour proposer une sélection de 2 à 5 variables discriminantes pour prédire la présence/absence de maladie. Ajuster des modèles discriminant à l'aide de la `discrim` pour les variables que vous avez retenues. Valider et comparer les modèles obtenus.