Introduction au data mining
L3 MIS - STA 1616 - 2010
V. Monbet
Classification et choix de variables

L'objectif de ce TD est d'illustrer l'intérêt du positionnement multidimensionnel. On montrera aussi en quoi le positionnement multidimensionnel et la classification peuvent permettre d'aider au choix de variables. En effet, différentes techniques telles que la classification à partir d'un tableau de distances ou de similarités permettent d'aider au choix de variables quand on est confronté à une grande base de données comportant de nombreuses variables. Ces méthodes reposent sur le calcul d'indices de similarité entre les variables : le carré du coefficient de corrélation pour des variables quantitatives et l'indice de Tschuprow pour des variables qualitatives.

A RENDRE Compte rendu de TP portant sur la section 3. Ces comptes rendus sont à rendre en début de cours le lundi 15 mars 2010 dernier délai.

1 Positionnement multidimensionnel pour les individus

Dans un premier temps, nous allons illuster le positionnement multidimensionnel à partir des données de criminalité dans les départements français.

La lecture des données peut être menée de la façon suivante.

```
dep <- read.table("h:/DATAMINING/DONNEES/depart_names.dat",header=TRUE)
# dep <- read.table(file.choose(),header=TRUE)
str(dep)
dep$num <- factor(dep$num)
summary(dep)
```

1. Centrer et réduire les données correspondant à des variables numériques dans la table `dep`. Puis calculer les distances entre individus en utilisant la fonction `dist`.

2. Utiliser les commandes ci-dessous pour réaliser un positionnement multidimensionnel des individus. `D` représente la matrice des distances entre individus.

```
library(MASS)
require(graphics)
loc <- cmdscale(D)
x <- loc[,1]
y <- -loc[,2]
plot(x, y, type="n", xlab="", ylab="", main="MDS - Départements")
text(x, y, dep[,2], cex=0.8)
```

3. Réaliser une analyse en composantes principales et comparer ce qu'on obtient au résultat du positionnement multidimensionnel. Commenter.

4. Etudiez l'effet du choix d'autres distances telle que la distance de Manhattan par exemple.

2 Choix de variables pour des variables quantitatives

2.1 Positionnement multidimensionnel

1. Utiliser maintenant le positionnement multidimensionnel pour représenter les variables. On pourra s'aider des commandes ci-dessous. Interpréter les résultats.

```
vdat = data.frame(t(dep.red),row.names=c("txcr","etra","urbr","jeun","age","chom",
"agri","arti","cadr","empl","ouvr","prof","fisc","crim","fe90"))
Dv <- dist(scale(vdat), method = "euclidean")
loc <- cmdscale(Dv)
x <- loc[,1]
y <- -loc[,2]
X11()
plot(x, y, type="n", xlab="", ylab="", main="MDS - Départements")
text(x, y, c("txcr","etra","urbr","jeun","age","chom","agri",
"arti","cadr","empl","ouvr","prof","fisc","crim","fe90"), cex=0.8)
```

2. La distance euclidienne n'a pas réellement de sens pour les variables. Mener la même analyse mais en utilisant maintenant la distance définie en cours et basée sur un coefficient de corrélation. On pourra programmer soi même une nouvelle fonction pour le calcul de cette distance.

2.2 Classification de variables

On peut utiliser la matrice de distances en entrée d'une classification hiérarchique ascendante sur les variables. Sous R, on utilise par exemple la fonction `hclust` avec la méthode de Ward. Représenter l'arbre obtenu. Discuter les résultats.

3 Choix de variables pour des variables quantitatives et qualitatives

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 sur le forum le jeu de données sous forme de table SAS.

3.1 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. L'éxecution de ce programme peut être fastidieuse car elle requiert de nombreux lancement de la `proc freq`. 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).

```
libname TP4 'H:\DATAMINING\DONNEES' ;
%let listev = sbp tobacco ldl adiposity famhist  typea
    obesity alcohol age chd ;
%dtprow(TP4.heart_disease,&listev,outprox = sasuser.dtschvp) ;
```

3.2 Classification

Une classification hiérarchique est ensuite opérée sur le tableau. 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;
```

3.3 Positionnement multidimensionnel

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

```
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) ;
```

3.4 Choix de variables

Aidez vous des résultats précédent pour proposer une sélection de 2 à 5 variables discriminantes pour prédire la présence/absence de maladie. Comparez ce resultat aux modèles que nous avions obtenus lors du premier TP.