Polytech' lille
Département gis.
M. Marbac

Classification non supervisée : classification ascendante hiérarchique.

Travaux pratiques (statistiques exploratoires).

Les jeux de données étudiés sont disponibles sur `http://math.univ-lille1.fr/~marbaclo/`

1 Premiers pas

Voici un tableau de données de 5 individus décrits par 2 variables quantitatives :

	x	y
a	0	0
b	3	3
c	9	0
d	3	6
e	9	8

Table 1 – Données premiers pas.

1. Avec le critère de saut maximum sur la distance euclidienne, effectuer la classification ascendante hiérarchique et représenter le dendrogramme.
2. En extraire une classification à deux classes et donner sa décomposition en inertie (inertie intra-classe et inertie inter-classe).
3. Faites la même chose avec le critère de Ward.

2 Définition d'une distance

On a la base de patients suivante

	Température	Antécédent	Age
A	36	oui	25
B	38	non	30
C	41	oui	75
D	42	non	80
E	39	non	40

Table 2 – Base de patients.

Les trois variables sont définies dans les domaines de valeurs suivants :
- température : entier $[32, 42]$
- antécédent : booléen
- age : entier $[0, 99]$.

1. Définir une mesure de distance dans l'intervalle $[0..1]$ pour chaque variable.
2. Donner la matrice de ressemblance entre les 5 patients en utilisant une distance de Manhattan pour les comparer :

$$DM(x, y) = \sum_{j=1}^{d} Dist(x_i, y_i)$$

3. Organiser les patients à l'aide de l'algorithme de classification ascendante hiérarchique en utilisant le critère du saut minimal.

3 Simulation avec R

Reprendre la fonction `simul` du TP des centres mobiles qui simule quatre échantillons gaussiens de même effectif dans le plan, de matrice de variance identité et dont les centres sont localisés sur les sommets d'un carré de côté δ.

1. Appliquer l'algorithme de CAH avec la méthode de ward à ce tableau (fonction `agnes` du package `cluster`).
2. Visualiser la hiérarchie indicée obtenue sur un dendogramme.
3. Pour différents nombre de classes, visualiser sur un graphique les partitions obtenues.
4. Reproduire les étapes précédentes avec d'autres méthodes que celle de Ward.

4 Les états des USA avec R

De nombreuses données concernant les 50 états des USA sont disponibles dans `R` en faisant `data(state)`. Les variables `state.center, state.x77` et `sate.abb` correspondent respectivement aux centres géographiques des états (longitude et latitude), à diverse statistiques les concernant (population, espérance de vie, etc...). et aussi à l'abbréviation de leur nom.

1. Concaténer en `R` dans un unique data.frame les trois sources de données `state.center, state.x77` et `sate.abb`.
2. Visualiser la position géographique des 50 états américains.
3. Enchaîner une ACP puis une CAH sur les statistiques issues de `state.x77`. Visualiser alors la structure de classification obtenue sur un dendrogramme.
4. Visualiser la partition en cinq classes des états dans le premier plan factoriel de l'ACP.
5. Faire une manipulation permettant maintenant de visusaliser la partition en cinq classes des états en fonction de leur position géographique. (Piste : on peut exporte les données de format label/x/y/classe et l'importer ensuite dans SAS ou R). Que remarque-t-on ?

5 Utilisation de SAS

En 1994, les indicateurs démographiques données par le taux de mortalité infantile pour 1000 naissances vivantes et par le taux de mortalité néonatale précoce pour 1000 naissances vivantes étaient les suivants pour six pays (ou provinces) occidentaux :

Pays	Infantile	Néonatale
France	5.9	2.3
Allemagne	5.6	2.4
Royaume-Uni	5.9	3.4
Canada	6.3	3.5
Québec	5.6	3.1
États-Unis	8.0	4.2

TABLE 3 – Taux de mortalité infantile et néonatale.

1. Entrer ces données dans SAS.
2. Effectuer une CAH, voir le dendrogramme et discuter le nombre de classes.
3. Visualiser graphiquement dans le plan la partition obtenue.