Polytech' lille
Département gis.
M. Marbac

Classification non supervisée : les centres mobiles

Travaux pratiques (statistiques exploratoires).

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

1 Premiers pas

Construire une partition en deux classes des individus décrits par le tableau suivant à l'aide des centres mobiles. On choisira comme premiers centres mobiles les individus 1 et 3. Pour chaque itération de l'algorithme, vous indiquerez les centres et la partition obtenues. Lorsque l'algorithme aura convergé, vous représenterez le nuage de points et la partition estimée.

Individu	X	Y
1	-2	1.5
2	3	3
3	-2.2	1
4	0	-1
5	3.2	1.4

Table 1 – Données exercice 1.

2 Plusieurs initialisations nécessaires

On souhaite classifier l'ensemble des individus ci dessus (décrits par une seule variable) en utilisant l'algorithme des *k-means*. On cherche à construire une partition en trois classes c_1, c_2 c_3, en prenant

Individus	A	B	C	D	E	F	G
Valeurs	6	1	12	7	10	3	2

Table 2 – Données exercice 2

comme noyau initial respectivement la valeur des individus B, F et G.

1. Exécuter l'algorithme des *k-means* jusqu'à stabilisation des classes en indiquant à chaque itération la valeur des noyaux et la partition estimée.
2. Que pensez-vous de la classification finale obtenue ? Pouvez-vous imaginer et décrire une meilleure stratégie pour sélectionner les noyaux initiaux.

3 Simulation avec R

Il s'agit de simuler quatre échantillons gaussiens de même effectif dans le plan. Chaque échantillon gaussien suit une loi de matrice de variance identité. Les centres sont localisés sur les sommets d'un carré de côté δ.

1. Réaliser la fonction `simul` qui génère un échantillon provenant de ce mélange.
2. Identifier le fonctionnement de la fonction `kmeans`.

3. Réaliser la fonction `voir.partition` qui affiche le nuage de points et les centres de gravité en utilisant un symbole différent pour chacune des classes. Vérifier alors que l'on retrouve bien les quatre classes initiales lorsque l'on recherche une partition avec quatre classes (on prendra une valeur de δ de l'ordre de 5).
4. Réaliser la fonction `voir.W` qui lance la fonction `multi.kmeans` (à programmer) pour plusieurs nombres de classes puis trace la valeur de l'inertie de W en fonction de chacun des nombres de classes.
5. Pour différentes valeurs de δ, utiliser la méthode du coude pour proposer un nombre de classes plausible. Commenter.

4 Les Iris de Fisher (R puis SAS)

Les iris de Fisher sont un jeu de données constitué de 150 individus (des fleurs) décrits par quatre variables :
- largeur de pétale ;
- longueur de pétale ;
- largeur de sépale ;
- longueur de sépale.

Les 150 fleurs regroupent trois espèces d'iris différentes (Verginica, Setosa, Versicolor). Chaque espèce est représentée par 50 individus.

1. Dans R, ce jeu de données est disponible en faisant `data(iris)`. Exporter le de `R` vers SAS. Maintenant, seul le logiciel SAS sera utilisé.
2. Visualiser les données dans le premier plan factoriel de l'ACP en utilisant un symbole différent pour chaque espèce. Que constatez-vous ?
3. Utiliser la méthode du coude pour retenir un nombre plausible de classes.
4. Visualiser de nouveau les données dans le premier plan factoriel de l'ACP en faisant cette fois apparaître les classes estimées. Comparer alors le graphique de la question 2.

5 Classification de données qualitatives avec R

Effectuer puis visualiser une classification des données qualitatives `races de chiens` étudiées précédemment en ACM. Quelques directives :
- Utiliser l'ACM pour réaliser et visualiser la classification.
- Nombre de classes à déterminer.
- Interprétation des classes obtenues.