Fouille de données
Classication automatique : agrégation autour de k centres mobiles

L'objectif de ce TP est de se familiariser avec la pratique de la classification automatique fondée sur l'agrégation autour de centres mobiles sous R. Un compte-rendu de TP vous est demandé. Ce compte rendu doit contenir le code `R` utilisé et d'abondants commentaires.

1 Exemple introductif

1. Générez des données bivariées artificielles en tapant :
   ```
   n <- 200
   x1 <- rnorm(n/2)
   y1 <- rnorm(n/2)
   x2 <- rnorm(n/2, 4, 2)
   y2 <- rnorm(n/2, 4, 2)
   d <- data.frame(c(x1, x2), c(y1, y2))
   names(d) <- c("x", "y")
   ```
   Comme d'habitude, les individus sont en ligne et les variables en colonne. Expliquez comment sont générées ces données et représentez le nuage de points correspondant. Combien de groupes voyez-vous ?
2. Nous allons maintenant utiliser la fonction `kmeans` pour identifier automatiquement deux groupes "homogènes" dans les données. Idéalement, parmi toutes les partitions possibles en deux classes, on souhaite obtenir la partition qui a la plus petite inertie intraclasse (ou, de façon équivalente, la plus forte inertie interclasse). Le code `R` à utiliser est :
   ```
   km <- kmeans(d, 2)
   km
   ```
   La partie "Clustering vector" de la sortie donne la partition obtenue. Vérifiez que cela est équivalent au vecteur `km$cluster`.
3. Nous allons maintenant représenter le résultat de la classification dans le plan :
   ```
   plot(d, pch="")
   text(d, label=km$cluster)
   ```
   La classification obtenue automatiquement a-t-elle du sens d'après vous ? Essayez aussi `plot(d, col=km$cluster)`.
4. Les centres de gravité des classes sont donnés par `km$centers` et leurs tailles par `km$size`. Si on divise `km$withinss` par `km$size`, on obtient l'inertie de chaque classe. Laquelle des deux classes a la plus forte inertie ? Est-ce que cela est conforme à l'intuition ? Expliquez. Calculez l'inertie intraclasse qui est la moyenne des inerties des classes pondérées par la proportion d'individus dans les classes. Vérifiez que cela est égal à `km$tot.withinss/n`. L'inertie intraclasse est donc donnée par `km$tot.withinss/n`.
5. L'inertie interclasse est quant à elle donnée par `km$betweenss/n`. Calculez l'inertie totale.

6. C'est uniquement parce que l'on travaille sur des données bivariées que l'on a pu ici suggérer un nombre de classes pertinent. En général, on ne sait pas quelle valeur donner au deuxième argument de `kmeans`. Une approche possible consiste à essayer plusieurs valeurs consécutives pour le nombre de classes est à regarder l'évolution de l'inertie intraclasse en fonction du nombre de classes (ou, de façon équivalente, l'évolution de l'inertie interclasse en fonction du nombre de classes). Faites varier le nombre de classes et, lors de chaque exécution, calculez l'inertie intraclasse, l'inertie interclasse et l'inertie totale. Par exemple :

```
km1 <- kmeans(d, 1)
intra1 <- km1$tot.withinss/n
inter1 <-  km1$betweenss/n
tot1 <- intra1 + inter1

km2 <- ...
```

Représentez l'évolution des trois inerties en tapant :

```
plot(c(intra1,intra2,intra3,intra4,intra5),type="l")
points(c(inter1,inter2,inter3,inter4,inter5),type="l",lty=2)
points(c(tot1,tot2,tot3,tot4,tot5),type="l",lty=3)
```

Combien de classes suggérez-vous de prendre ? Comme en ACP, on recherchera des coudes dans ce graphique.

7. Pour terminer, représentez la partition en trois classes dans le plan. Que pourrait-on reprocher à cette partition ?

2 Les données USArrests

1. Chargez les données `USArrests`. Pour éviter que les différences de dispersion entre variables n'affectent trop la classification automatique, travaillez sur les données centrées-réduites.
2. Appliquez la fonction `kmeans` aux données pour 2 à 6 classes et comparez les partitions en terme d'inertie intraclasse. Combien de classes proposez-vous de garder ?
3. Caractérisez et interprétez chaque classe en examinant son centre de gravité et son inertie.
4. Effectuez une ACP des données et représentez les classes obtenues par `kmeans` dans les plans factoriels retenus afin d'inspecter visuellement la qualité de la classification.