SY09 Printemps 2019
TP 5
Classification automatique

1 Classification hiérarchique

1. Effectuer la classification hiérarchique ascendante des données de `Mutations` (avec les différents critères d'agrégation disponibles). Commenter et comparer les résultats obtenus, en vous appuyant sur la représentation obtenue par AFTD. On pourra utiliser la fonction `hclust`.
2. Effectuer la classification hiérarchique ascendante des données `Iris`, après calcul des distances associées (on utilisera la fonction `dist` pour ce faire). Commenter les résultats obtenus, en vous appuyant sur votre connaissance de ce jeu de données.
3. Effectuer la classification hiérarchique descendante des données `Iris`, au moyen de la fonction `diana` (bibliothèque `cluster`). Comparer aux résultats obtenus au moyen de la CAH.

Remarque importante : dans les anciennes versions de R, il faut élever les distances au carré avant d'effectuer une CAH via la fonction `hclust` *avec le critère de Ward (lorsque celui-ci a un sens : tableau de distances euclidiennes). Dans les versions les plus récentes, il existe deux critères :* `ward.D` *et* `ward.D2` *; on choisira le second (*`ward.D2`*) qui implémente le critère de Ward.*

2 Méthode des centres mobiles (K-means)

Le but de cet exercice est de tester les performances de l'algorithme des centres mobiles sur trois jeux de données réelles : `Iris`, `Crabs` et `Mutations`. Sauf précision, on traitera les *données complètes* ; on pourra représenter cette classification obtenue sur les données complètes en utilisant une représentation des données dans le premier plan factoriel.

2.1 Données Iris

1. Tenter une partition en $K \in \{2, 3, 4\}$ classes avec la fonction `kmeans` ; visualiser et commenter.
2. On cherche à présent à étudier la stabilité du résultat de la partition. Effectuer plusieurs classifications des données en $K = 3$ classes. Observer les résultats, en termes de partition et d'inertie intra-classes. Ces résultats sont-ils toujours les mêmes ? Commenter et interpréter.
3. On cherche à déterminer le nombre de classes optimal.
   (a) Effectuer $N = 100$ classifications en prenant $K = 2$ classes ; puis à nouveau $N = 100$ classifications en $K = 3$ classes, $K = 4$ classes, ... jusqu'à $K = 10$ classes. On pourra faire deux boucles imbriquées pour cela.
   (b) Pour chaque valeur de K, calculer l'inertie intra-classe minimale (sur les 100 répétitions) $\widehat{I}_K = \min_{i=1,\ldots,100} I_{K\,i}$. Représenter la variation d'inertie minimale en fonction de K. On inclura à ce graphique l'inertie totale (assimilable à l'inertie intra-classe pour $K = 1$).

   Proposer un nombre de classes à partir de ces informations, en utilisant la méthode du coude.
4. Comparer les résultats de la partition obtenue par les centres mobiles avec la partition réelle des iris en trois groupes : quels individus sont placés dans le mauvais cluster ? Pourquoi ?

2.2 Données Crabs

1. Effectuer plusieurs classifications en $K = 2$ classes des données `Crabs` *pré-traitées de manière à supprimer l'effet taille*. Les résultats obtenus sont-ils toujours les mêmes d'une classification à l'autre ? À quoi sont dues les différences observées ?
2. Effectuer plusieurs classifications des données en $K = 3$ classes. Ici encore, qu'observe-t-on ?
3. Effectuer une classification en $K = 4$ classes des données. Comparer à la partition réelle suivant l'espèce et le sexe. Que peut-on conclure ?

2.3 Données Mutations

On calculera tout d'abord une représentation des données `mutations` (tableau individus-variables) dans un espace de dimension $d = 5$. On utilisera par la suite la fonction `kmeans` sur ces données.

1. Effectuer plusieurs classifications de cette représentation en $K = 3$ classes au moyen de l'algorithme des centres mobiles. On pourra représenter les résultats obtenus dans le premier plan factoriel de l'AFTD.
2. Étudier la stabilité du résultat de la partition. Commenter et interpréter.

3 Convergence des K-means

On cherche ici à montrer que l'algorithme des K-means peut être interprété comme une procédure de minimisation alternée qui répète deux étapes :
— le calcul d'un représentant pour chaque groupe,
— le calcul d'une affectation des points à chacun des groupes.

1. On définit une affectation comme un ensemble de variables indicatrices $z_{ik} \in \{0, 1\}$, pour tout $i = 1, \ldots, n$ et $k = 1, \ldots, K$, telles que $\sum_k z_{ik} = 1$ (une seule est non nulle). Exprimer le critère d'inertie optimisé par l'algorithme des K-means en fonction de ces variables d'affectation $\boldsymbol{z}_i$ et des représentants des groupes $\boldsymbol{\mu}_k$.

Le critère d'inertie est

$$I(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \frac{1}{n} \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \omega_k} d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k),$$

où la distance utilisée est la distance euclidienne :

$$d^2(\boldsymbol{x}_1, \boldsymbol{x}_2) = \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2 = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^T (\boldsymbol{x}_1 - \boldsymbol{x}_2).$$

2. On suppose disposer d'une affectation $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. Montrer que pour un groupe, le centre de gravité est le représentant qui minimise le critère d'inertie.

Pour choisir le représentant $\boldsymbol{\mu}_k$ d'un groupe, on peut calculer la dérivée partielle du critère d'inertie par rapport à $\boldsymbol{\mu}_k$:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} I(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = -\frac{2}{n} \sum_{i=1}^{n} z_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k).$$

L'annulation de ce vecteur de dérivées premières donne :

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} I(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = 0 \Leftrightarrow \sum_{i=1}^{n} z_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) = 0 \Leftrightarrow \boldsymbol{\mu}_k = \frac{\sum_{i=1}^{n} z_{ik} \boldsymbol{x}_i}{\sum_{i=1}^{n} z_{ik}}.$$

Remarquons que la matrice des dérivées secondes (matrice hessienne) est définie par blocs :

$$\frac{\partial^2 I(\boldsymbol{\mu}_k, \boldsymbol{z}_i)}{\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_k^T} = \frac{2}{n} \sum_{i=1}^{n} z_{ik} \mathrm{I}_\mathrm{p}, \text{ et } \frac{\partial^2 I(\boldsymbol{\mu}_k, \boldsymbol{z}_i)}{\partial \boldsymbol{\mu}_k \partial \boldsymbol{\mu}_\ell^T} = 0_\mathrm{p} \text{ pour tout } \ell \neq k,$$

avec $\mathrm{I_p}$ la matrices identité, et 0_p la matrice nulle, de dimensions $p \times p$; elle est donc diagonale et définie positive si les classes ne sont pas vides (auquel cas $\sum_i z_{ik} > 0$).

3. Étant donné des centres $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$, montrer que l'affectation de chaque point au centre le plus proche minimise le critère d'inertie.

Remarquons tout d'abord que le critère d'inertie est séparable — il s'écrit comme une somme de termes dont chacun ne concerne qu'un exemple $\boldsymbol{x}_i$:

$$I(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k).$$

Pour chaque $\boldsymbol{x}_i$, il est alors évident qu'il faut choisir l'affectation $\boldsymbol{z}_i^\star$ définie par

$$z_{ik}^\star = \begin{cases} 1 & \text{pour } k^\star = \arg \min\limits_{\ell=1,\ldots,K} d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_\ell), \\ 0 & \text{pour } \ell \neq k, \end{cases}$$

car $d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_\ell) \geqslant d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_{k^\star})$ pour tout $\ell \neq k^\star$, et donc pour tout $\boldsymbol{z}_i \neq \boldsymbol{z}_i^\star$,

$$\sum_{k=1}^{K} z_{ik} d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k) \geqslant \sum_{k=1}^{K} z_{ik}^\star d^2(\boldsymbol{x}_i, \boldsymbol{\mu}_k).$$