Cours7 : Classification automatique de données par la méthode hierarchique ascendante.

La classification automatique (clustering) est une méthode mathématique d'analyse de données : pour faciliter l'étude d'une population d'effectif important (animaux, plantes, malades, gènes, etc...), on les regroupe en plusieurs classes de telle sorte que les individus d'une même classe soient le plus semblables possible et que les classes soient le plus distinctes possibles. Pour cela il y a diverses façons de procéder (qui peuvent conduire à des résultats différents...). Dans ce cours nous présentons deux algorithmes, un premier appelé *classification hierarchique ascendante* et un second que nous étudierons lors du prochain cours appelé *méthode des centres mobiles*.

1 Distances (ou dissimilarité) entre individus d'une même population

Pour regrouper les individus qui se ressemblent (et bien séparer ceux qui ne se ressemblent pas), il faut choisir un "critère de ressemblance". Pour cela on examine l'ensemble des informations dont on dispose concernant les individus (pression artérielle, température, taux de métabolisme, ... par exemple s'il s'agit de malades) notées $(x_i, y_i, \ldots)$ pour le ième individu, et on imagine que chaque individu est un point $M_i = (x_i, y_i, z_i, \ldots)$ de l'espace. S'il n'y a que deux variables relevées (x_i, y_i) on obtient ainsi un nuage Γ de points dans le plan, chaque point M_i ayant pour coordonnées (x_i, y_i). Ce nuage $\Gamma = \{M_i, i = 1, \ldots, n\}$ contient $n$points, si n est l'effectif total de la population. La *distance euclidienne* de deux individus M_i et M_j est par définition

$$d_2(M_i, M_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Elle est d'autant plus petite que les deux individus sont semblables (du point de vue des valeurs des deux critères retenus) et d'autant plus grande qu'ils sont différents.

On peut associer à chaque nuage d'individus une matrice $\mathbb{D} = (d_{ij})_{0 \leq i \leq n, 0 \leq j \leq n} = (d_2(M_i, M_j))$, dite *matrice des distances*. C'est une matrice à n lignes et n colonnes, à coefficients positifs, symétrique (puisque $d_2(M_i, M_j) = d_2(M_j, M_i)$) et nulle sur la diagonale (puisque $d_2(M_i, M_i) = 0$). Pour un nuage d'effectif n, il y a donc $\frac{n(n-1)}{2}$ distances à calculer.

A coté de la distance euclidienne, on peut définir d'autres distances (et donc d'autres matrices des distances). Par exemple

$$d_1(M_i, M_j) = |x_i - x_j| + |y_i - y_j|$$

$$d_\infty(M_i, M_j) = \text{Max}\,\{|x_i - x_j|, |y_i - y_j|\}$$

2 Ecarts entre classes (cas euclidien)

Supposons le nuage $\Gamma = \{M_i, i = 1, \ldots, n\}$ décomposé en plusieurs classes Γ_1, Γ_2, ... , Γ_k et notons G_1, G_2, ... , G_k les centres de gravité respectifs de chaque classes et notons p_1, p_2, ... , p_k les *poids* respectifs de chaque classe que l'on définit de la façon suivante : si l'on suppose que tous les individus ont le même poids égal à $\frac{1}{n}$, le poids p_l de la classe Γ_l est égal à l'effectif de Γ_l divisé par n. De cette façon la somme des poids de toutes les classes vaut 1. Rappelons que le centre de gravité G d'un nuage de points Γ est le *point moyen* du nuage, c'est-à-dire le point $G = (\overline{x}, \overline{y}, \ldots)$ de coordonnées $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, ...

Pour mesurer la proximité ou l'écart entre deux classes Γ_l et Γ_m, il existe plusieures façons de procéder. On calcule par exemple la quantité $\text{Min}\,\{d(M_i, M_j), M_i \in \Gamma_m, M_j \in \Gamma_l\}$ appelée *distance du plus proche voisin* ou encore $\text{Max}\,\{d(M_i, M_j), M_i \in \Gamma_m, M_j \in \Gamma_l\}$ ou simplement la distance des centres de gravité

$d_2(G_m, G_l)$ des classes. Mais la mesure que l'on utilise le plus souvent lors des classifications automatiques, appelée é*cart de Ward*, est définie par :

$$d(\Gamma_m, \Gamma_l) := \frac{p_m p_l}{p_m + p_l} d_2(G_m, G_l)^2$$

où p_l et p_m sont les poids des deux classes.

3 Inertie interclasse et inertie intraclasse

On appelle *inertie totale* d'un nuage $\Gamma = \{M_i, i = 1, \ldots, n\}$ la somme pondérée des carrés des distances de ses points au centre de gravité du nuage. Donc, si G désigne le centre de gravité de Γ, l'inertie totale de Γ est, si tous les points du nuage sont de même poids égal à $\frac{1}{n}$,

$$\mathcal{I}(\Gamma) = \frac{1}{n} \left(d_2(M_1, G)^2 + d_2(M_2, G)^2 + \ldots + d_2(M_n, G)^2 \right). \quad (1)$$

Notons que le centre de gravité est précisément le point G pour laquelle cette somme pondérée est minimal. L'inertie "mesure" la dispersion du nuage. Si le nuage Γ est composé de k classes Γ_1, Γ_2, ... , Γ_k (disjointes deux à deux), celles-ci seront d'autant plus homogènes que les inerties de chaque classe, $\mathcal{I}(\Gamma_1)$, $\mathcal{I}(\Gamma_2)$, ... , $\mathcal{I}(\Gamma_k)$, calculées par rapport à leurs centres de gravité G_1, G_2, ... , G_k respectifs, sont faibles. La somme de ces inerties est appelée *inertie intraclasse* :

$$\mathcal{I}_{intra} = \mathcal{I}(\Gamma_1) + \mathcal{I}(\Gamma_2) + \ldots + \mathcal{I}(\Gamma_k)$$

Les inerties des classes $\mathcal{I}(\Gamma_1)$, $\mathcal{I}(\Gamma_2)$, ... sont simplement calculées avec la formule (1) ci-dessus où l'on remplace le centre de gravité G par celui de la classe G_1, G_2, ... et le poids $\frac{1}{n}$ par celui de la classe.

L'inertie totale d'un nuage n'est généralement pas égale à la somme des inerties des classes qui le composent, c'est-à-dire à l'inertie intraclasse (sauf dans le cas où les centres de gravité de toutes les classes sont confondus) car il faut prendre en compte également la dispersion des classes par rapport au centre de gravité du nuage. Il s'agit de l'*inertie interclasse* définie par

$\mathcal{I}_{inter} = \overline{p}_1 d_2(G_1, G)^2 + \overline{p}_2 d_2(G_2, G)^2 + \ldots + \overline{p}_k d_2(G_k, G)^2$, où $\overline{p}_j$ désigne le poids total de la classe Γ_j.

On montre le résultat suivant appelé *décomposition de Huygens :*

Théorème 1 *L'inertie totale d'un nuage de points composé de différentes classes disjointes deux à deux est la somme de son inertie intraclasse et de son inertie interclasse, c'est-à-dire :*

$$\mathcal{I}(\Gamma) = \mathcal{I}(\Gamma_1 \cup \Gamma_2 \cup \ldots \cup \Gamma_k) = \mathcal{I}_{intra} + \mathcal{I}_{inter}.$$

4 Classification hiérarchique ascendante

Pour classifier une population d'effectif n dont les individus sont numérotés 1, 2, ..., on considère cette population comme la réunion de n classes à un seul élément et on regroupe progressivement les classes deux à deux selon l'algorithme suivant :

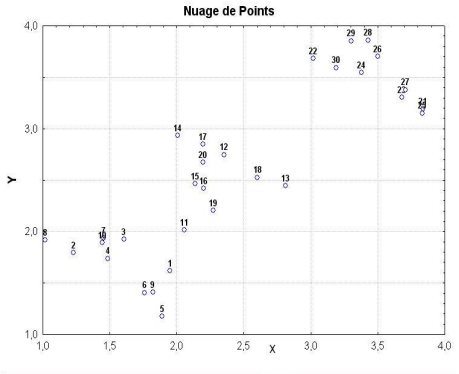
Etape 1 : Calculer la matrice des distances $\mathbb{D} = (d(M_i, M_j)_{0 \leq i \leq n, 0 \leq j \leq n}$ ou directement la matrice des distances de Ward des classes réduites aux points $\mathbb{W} = (\frac{p_i p_j}{p_i + p_j}(d(M_i, M_j)^2)_{1 \leq i \leq n, 1 \leq j \leq n}$.

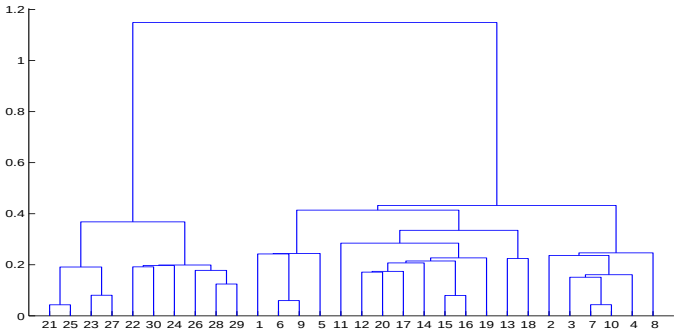
Etape 2 : Remplacer les deux individus de distance minimale par une classe (à 2 éléments) numérotée $n + 1$, qui sera représentée par le centre de gravité des individus et affectée de la somme des poids des individus.

Etape 3 : Calculer la perte d'inertie interclasse (ou gain d'inertie intraclasse) dû au regroupement précédent : il s'agit exactement de l'écart de Ward des deux individus regroupés.

Après ces trois étapes, la population compte alors $n - 1$ classes ($n - 2$ classes à un élément et une à 2 éléments). On peut donc recommencer à l'étape 1 en remplaçant "individus" par "classes" si nécessaire (et donc "distance entre individus" par "écarts entre classes"). Après $n - 1$ itérations, tous les individus sont regroupés en une classe unique.

On construit alors un arbre, appelé *dendrogramme* (voir dessin ci-dessus) de la façon suivante. On aligne sur l'axe horizontal des points représentant les différents individus et on les joint deux à deux, successivement, en suivant cet algorithme de classification hierarchique ascendante (commençant par les plus proches, etc...). On poursuit ainsi jusqu'à regroupement de tous les individus en une classe unique.





Pour plus de lisibilité, on pourra disposer les individus dans l'ordre dans lequel les regroupements ont été effectués. Le niveau (hauteur) de chaque noeud de l'arbre est, le plus souvent, choisi *proportionnel à la nouvelle d'inertie intra ;* en choisissant le rapport (inertie intra)/(inertie totale) ce niveau est zéro lorsque tous les individus sont séparés (en bas) et vaut 1 lorsqu'il sont tous réunis en une seule classe (en haut). En fait, on trace ce dendrogramme afin de visualiser le niveau où couper cet arbre pour réaliser la *meilleure* partition de l'ensemble initial. On peut comprendre qu'il sera optimal de couper le dendrogramme à un niveau où le regroupement entre classes conduit à une perte d'inertie inter maximale. On peut vérifier que *l'écart de Ward entre deux classes est en fait égal à la perte d'inertie inter (ou le gain d'inertie intra) que produirait la réunion de ces deux classes en une seule.* Le niveau des noeuds de l'arbre est donc facile à calculer à partir des écarts de Ward entre les classes.

5 Cas non euclidien : positionnement multidimentionnel

Il arrive souvent dans les situations concrètes que l'on dispose d'une matrice de distances entre les individus à classifier mais que cette distance (ou similarité) ne corresponde pas à la distance euclidienne entre des points repérés par leurs coordonnées (comme c'est le cas dans l'exercice 2). On peut néanmoins dans ce cas réaliser une classification mais l'interprétation donnée ci-dessus utilisant le nuage de points, les centres de gravités et l'écart de Ward n'a plus de sens. On peut néanmoins utiliser l'algorithme précédent mais avec une distance interclasse différente et en gardant les classes successives sans les représenter à travers leurs centre de gravité. L'ordonnée du dendrogramme est alors simplement la distance entre les deux classes regroupées. Les logiciels permettent cependant dans tous les cas de faire le choix de l'écart de Ward (c'est en général l'option par défaut) et on peut donc s'interroger pour savoir comment ils peuvent le faire en l'absence de coordonnées euclidiennes pour les individus. En fait il existe une méthode appelée *positionnement multidimentionnel* (en anglais MDS pour *multiDimentional Scaling*) qui permet de trouver, pour n'importe quelle matrice de distances de taille nxn un ensemble de n points repérés par leur coordonnées euclidiennes dont la matrice de distance est égale ou très proche de la matrice de distances donnée. Pour effectuer une classification automatique à partir d'une matrice de distances le logiciel fera donc, si besoin, précéder l'algorithme décrit ici par un MDS.