Cours 7 : Classification automatique de données par la méthode hiérarchique ascendante.

La classification automatique (appelée clustering en anglais) est une méthode mathématique d'analyse de données : pour faciliter l'étude d'une population d'effectif important (animaux, plantes, malades, gènes, etc...), on regroupe les individus qui la forment 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 les unes des autres. 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 i-ème 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) ou encore des *dissimilarités* ou des é*carts*. 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

Considérons une matrice de "distances" entre les points du nuage, par exemple le carré de leur distance euclidienne s'il s'agit de points d'un espace euclidien ou toute autre mesure de distance entre les points. Supposons que le nuage $\Gamma = \{M_i, i = 1, \ldots, n\}$ est composé de plusieurs classes Γ_1,Γ_2,..., Γ_n, deux à deux disjointes. Pour mesurer la distance entre les deux classes Γ_l et Γ_m, il existe plusieurs façons de procéder. L'une des plus utilisée est la *distance au plus proche voisin*

$$d(\Gamma_l, \Gamma_m) = \text{Min}_{\,x\in\Gamma_l, y\in\Gamma_m} d(x, y).$$

3 Classification hiérarchique ascendante

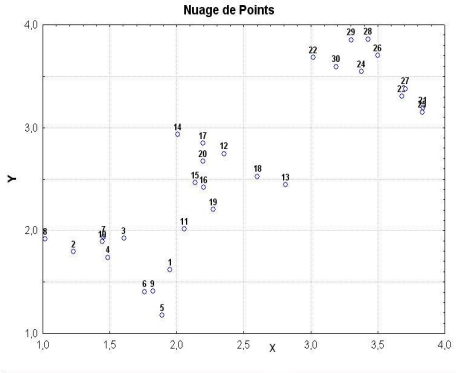
Pour classifier une population d'effectif n dont les individus sont numérotés 1, 2, ..., on considère cette population comme la réunion de n classes à un seul élément et on regroupe progressivement les classes deux à deux selon l'algorithme suivant :

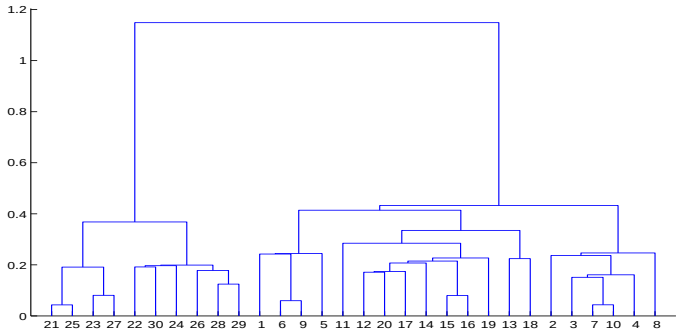
Etape 1 : Calculer la matrice des distances $\mathbb{D} = (d(M_i, M_j))_{1\leq i\leq n, 1\leq j\leq n}$.

Etape 2 : Remplacer les deux individus de distance minimale par une classe (à 2 éléments) notée Γ_1. 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, recalculer la matrice des distances entre ces $n-1$ classes, en remplaçant si nécessaire "distance entre individus" par "écarts entre classes", puis regrouper (étape 2) les deux classes de distance minimale en une seule. On appelle cela l'*aglomération au plus proche voisin.*

Après $n-1$ itérations, tous les individus seront regroupés en une classe unique.

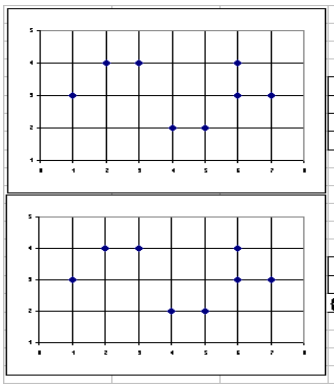




On peut construire alors un arbre, appelé *dendrogramme* (voir dessin ci-dessus) de la façon suivante. On aligne sur l'axe horizontal des points représentant les différents individus et on les joint deux à deux, successivement, en suivant cet algorithme de classification hierarchique ascendante (commençant 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. Deux points regroupés sont matérialisés par un petit "immeuble" du type de ceux représentés sur la figure et d'une hauteur égale à la distance séparant les points regroupés. Lorsque deux classes sont regroupées, on matérialise la nouvelle classe par un immeuble surplombant les deux immeubles regroupés d'une hauteur égale à l'écart entre les deux classes. Ce surplomb est donc d'autant plus grand que les classes regroupées sont éloignées.

On cherche ensuite à couper le dendrogramme au niveau où cela créé la meilleure répartition des points du nuage en classes bien distinctes entre elles. On peut comprendre qu'il ne sera pas optimal de couper le dendrogramme à un niveau où le regroupement s'est fait entre deux classes assez proches mais qu'au contraire on cherche à couper là où les classes regroupées étaient les plus éloignées.

Exemple : Voici par exemple une étape dans la classification d'un nuage de 8 points du plan ayant pour coordonnées $(1,3)$, $(2,4)$, $(3,4)$, $(4,2)$, $(5,2)$, $(6,3)$, $(6,4)$, $(7,3)$. On les suppose déja regroupés en trois classes $\Gamma_1 = \{(1,3)\,;(2,4)\,;(3,4)\}$, $\Gamma_2 = \{(4,2)\,;(5,2)\}$ et $\Gamma_3 = \{(6,3)\,;(6,4)\,;(7,3)\}$. Le calcul des distances entre les trois classes (distance entre les points les plus proches) montre que les plus proches sont Γ_2 et Γ_3. On les agglomère donc à l'étape suivante puis on calcule la distance entre la nouvelle classe ainsi formée $\{\Gamma_2, \Gamma_3\}$ et la classe rstante Γ_1.



	Γ1	Γ2	Γ3
Γ1	0	9	5
Γ2	9	0	2
Γ3	5	2	0

	Γ1	{Γ2,Γ3}
Γ1	0	5
{Γ2,Γ3}	5	0

Après avoir ainsi aggloméré l'ensemble des classes en une seule, on peut tracer le dendrogramme résumant les agglomérations successives, en choisissant de mettre en ordonnée la distance des deux classes regroupées.