Chapitre 8

Classification automatique : introduction

La classification (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 appelé *méthode des centres mobiles*.

8.1 Distances entre individus d'une même population

Pour regrouper les individus qui se ressemblent (et séparer ceux qui ne se ressemblent pas), il faut 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, $\Gamma = \{M_i, i = 1, \ldots, n\}$ où 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

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

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

8.2 Ecarts entre classes

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 de nombreuses 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 $d_2(G_m, G_l)$. Mais la mesure que l'on utilise le plus souvent appelée *écart 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.

8.3 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 $\frac{1}{n}$,

$$\mathcal{I}(\Gamma) = \frac{1}{n}\left(d_2(M_1, G)^2 + d_2(M_2, G)^2 + \ldots + d_2(M_n, G)^2\right). \tag{8.1}$$

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 (8.1) 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.

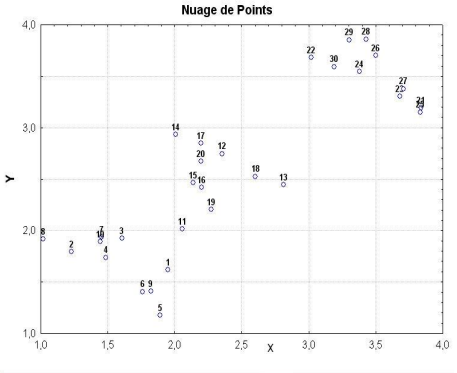
On montre le résultat suivant appelé *décomposition de Huygens :*

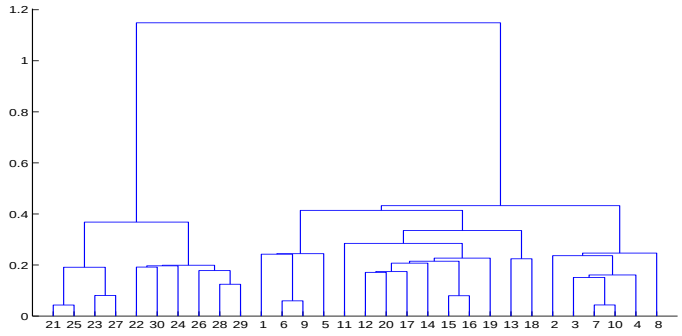
Théorème 8.1 *L'inertie totale d'un nuage de points composé de différentes classes disjointes deux à deux est la somme de son inertie intraclasse et de son inertie interclasse, c'est-à-dire :*

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

8.4 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.

8.5 Méthode des centres mobiles[1]

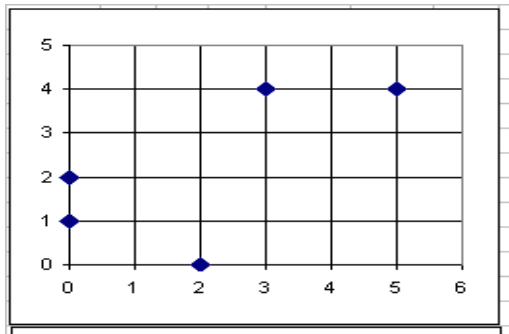
Cette méthode s'applique lorsque l'on sait à l'avance combien de classes on veut obtenir. Appelons k ce nombre. L'algorithme est le suivant :
Etape 0 : Pour initialiser l'algorithme, on tire au hasard k individus appartenant à la population, $C_1(0)$, $C_2(0)$, ..., $C_k(0)$: ce sont les k *centres* initiaux.
Etape 1 : On regroupe les individus autours de ces k centres de sorte à former k classes $\Gamma_1(0)$, $\Gamma_2(0)$, ..., $\Gamma_k(0)$ de la manière suivante : chaque classe $\Gamma_l(0)$ est constituée des points plus proches du centre $C_l(0)$ que des autres centres $\Gamma_m(0)$ pour $m \neq l$.
Etape 2 : On calcule alors les centres de gravité G_1, G_2, ... , G_k des k classes obtenues et on désigne ces points comme nouveaux centres $C_1(I) = G_1$, $C_2(I) = G_2$, ... , $C_k(I) = G_k$

[1]Hors programme en 2007/2008

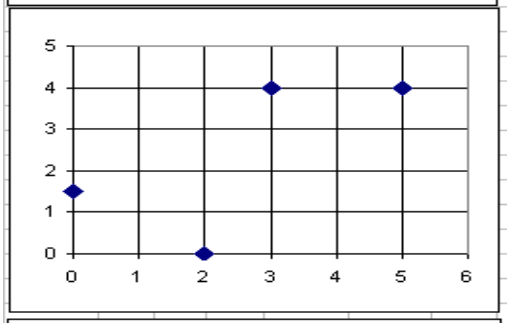
On répète les étapes 1 et 2 jusqu'à ce que le découpage en classes obtenu ne soit presque plus modifié par une itération suplémentaire. On peut montrer que la variance intra classe ne peut que décroître lorsque l'on passe d'un découpage en classes au suivant.

8.6 Exercices

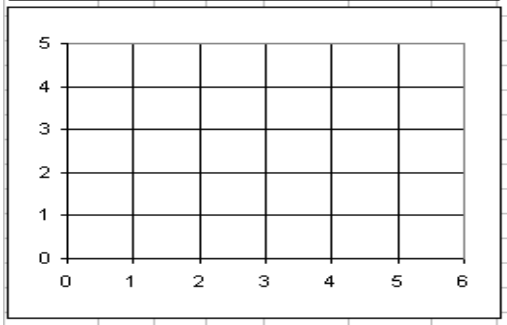
Exercice 1 : La succession des quatres dessins suivants correspond aux étapes successives d'une classification hierarchique ascendantes des cinq points $M_1(2,0)$, $M_2(0,1)$, $M_3(0,2)$, $M_4(3,4)$ et $M_5(5,4)$ progressivement regroupées en classes de deux ou trois points dont les centres de gravité sont notés G_6, G_7 et G_8. On suppose que les cinq points initiaux sont tous affectés du poids 1. La distance choisie pour cette classification, qui apparait dans les quatres matrices de distance, est l'écart de Ward.



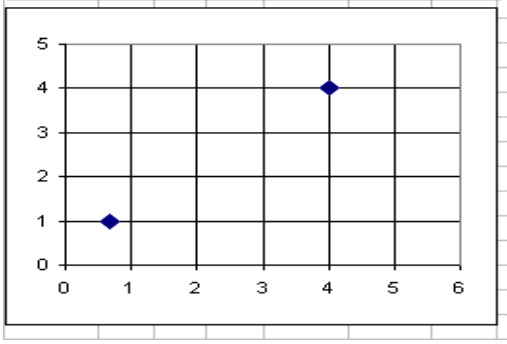
	M1	M2	M3	M4	M5
M1	0,00	2,50	4,00	8,50	12,50
M2	2,50	0,00	0,50	9,00	17,00
M3	4,00	0,50	0,00	6,50	14,50
M4	8,50	9,00	6,50		
M5	12,50			2,00	0,00



	M1	G6	M4	M5
M1	0,00	4,17	8,50	12,50
G6	4,17	0,00		20,83
M4	8,50		0,00	2,00
M5	12,50	20,83	2,00	0,00



	M1	G6	G7
M1	0,00	4,17	13,33
G6	4,17	0,00	22,25
G7	13,33	22,25	0,00



	G8	G7
G8	0,00	24,13
G7	24,13	0,00

1. Compléter le troisième dessin en y plaçant les trois points devant y figurer et indiquer sur les quatres dessins le nom des points.
2. Compléter les six distances manquantes dans les matrices de distances.
3. Préciser les coordonnées des points G_6, G_7 et G_8
4. Calculer les coordonnées du centre de gravité G_9 des cinq points.
5. Tracer un dendrogramme résumant cette classification.

Exercice 2 : Soient $M_1 = (1, 0)$, $M_2 = (0, 1)$ et $M_3 = (3, 1)$ trois points du plan.

1. Calculer les matrices des distances du nuage formé de ces trois points en utilisant successivement la distance euclidienne d_2 puis les distances d_1 et d_∞.
2. On ajoute au nuage précédent les deux points $M_4 = (4, 2)$ et $M_5 = (4, 3)$. Décrire les étapes successives d'une classification hiérarchique ascendante en calculant notamment les coordonnées et poids des classes obtenues par regroupement et la perte d'inertie intraclasse à chaque regroupement.
3. En déduire le dendrogramme. Quelle coupure suggérez-vous ?

Exercice 3 : *(Sujet inspiré d'un article de John Hartshorne, paru dans le journal de la "British Ecological Society")*

Un laboratoire d'écologie étudie les espèces micro-animales (larves, ..) présentes dans les rivières et les étangs. Il réalise, dans 6 sites de rivière, notés $R1$, $R2$, $R3$, $R4$, $R5$ et $R6$, et 3 sites d'étangs, notés $E1$, $E2$ et $E3$, des prélèvements répétés qui lui permettent d'avancer une liste des espèces présentes dans chacun de ces sites et de repérer les espèces présentes dans plusieurs sites à la fois. La matrice suivante contient, pour chaque paire de sites A et B, le nombre d'espèces communes aux 2 sites. Ainsi on y lit par exemple que 11 espèces sont présentes au site $R1$ et qu'**il** y a 7 espèces présentes à la fois au site $R1$ et au site $R2$.

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$	$E1$	$E2$	$E3$
$R1$	11	7	4	6	6	7	4	4	3
$R2$	7	15	8	8	9	6	3	3	2
$R3$	4	8	13	7	7	4	2	3	2
$R4$	6	8	7	15	7	6	6	8	6
$R5$	6	9	7	7	12	4	3	5	4
$R6$	7	6	4	6	4	10	6	5	5
$E1$	4	3	2	6	3	6	13	10	9
$E2$	4	3	3	8	5	5	10	15	11
$E3$	3	2	2	6	4	5	9	11	12

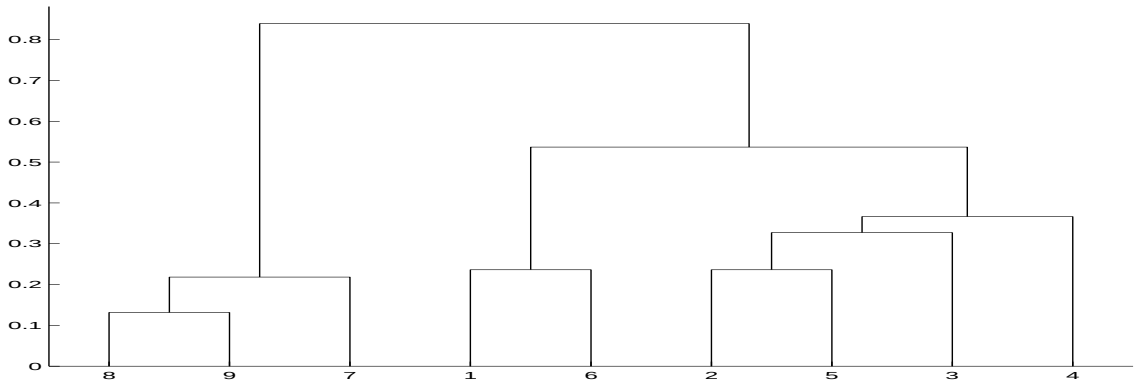
On se propose de regrouper les 9 sites en trois ou quatre classes composées de sites où ce sont pratiquement les mêmes espèces qui sont présentes. Pour réaliser cette classification, on propose de mesurer la distance entre deux sites A et B par la formule

$$d(A, B) = \frac{n_A + n_B - 2n_{AB}}{n_A + n_B}$$

où n_A (resp. n_B) désigne le nombre d'espèces présentes au site A (resp. au site B) et n_{AB} le nombre d'espèces en commun entre les sites A et B. On obtient la matrice des distances suivante :

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$	$E1$	$E2$	$E3$
$R1$	0	0,462	0,666	0,538	0,478	0,334	0,666	0,692	0,74
$R2$	0,462	0	0,428	.....	0,334	0,52	0,786	0,8	0,852
$R3$	0,666	0,428	.....	.....	0,44	0,652	0,846	0,786	0,84
$R4$	0,538	0,466	.....	0	0,482	0,52	0,572	0,466	0,556
$R5$	0,478	0,334	0,44	0,482	0	0,636	0,76	0,63	0,666
$R6$	0,334	0,52	0,652	0,52	0,636	0	.....	.....	0,546
$E1$	0,666	0,786	0,846	0,572	0,76	0,478	.....	.....	0,28
$E2$	0,692	0,8	0,786	0,466	0,63	0,6	.....	.....	0,186
$E3$	0,74	0,852	0,84	0,556	0,666	0,546	0,28	0,186	0

1. Compléter les coefficients manquants de cette matrice.
2. Préciser quels sont les deux sites les plus proches ainsi que les deux sites les plus éloignés.
3. La classification conduit au dendrogramme représenté ci-dessous. Décrire la composition des classes de la partition qui vous semble la plus appropriée.



4. Un autre choix de distance entre les sites aurait-il pu conduire à une partition différente ? Pourquoi n'a-t-on pas choisi la distance euclidienne ?

Exercice 4 :
1. En choisissant un nuage de trois points alignés sur l'axe des x regroupés en deux classes, calculer l'inertie totale, l'inertie intraclasse et l'inertie interclasse. Vérifier le théorème de Huygens dans cet exemple.
2. En considérant cette fois trois points du plan non nécessairement alignés, montrer le théorème de Huygens (on pourra utiliser le fait que leurs projections sur les deux axes de coordonnées vérifient le théorème).

Exercice 5 : Soit $\Gamma := \{M_i = (x_i, y_i), i = 1, \ldots, n\}$ un nuage de points du plan, chacun étant pondéré d'un poids $\frac{1}{n}$.
1. Quelle formule donne les coordonnées (x, y) du centre de gravité G du nuage en fonction de x_i, y_i et n ?
2. En utilisant votre calculette, vérifier sur quelques exemples de nuages la "transitivité" du centre de gravité, c'est-à-dire le fait que pour calculer les coordonnées de G on peut, lorsque le nuage est la réunion de deux classes Γ_1 et Γ_2, calculer d'abord les centres de gravité G_1 et G_2 des deux classes puis calculer le centre de gravité de G_1 et G_2 affectés de leurs poids respectifs.

Exercice 6 : On considère les 6 points $M_1 = (-2, 3)$, $M_2 = (-2, 1)$, $M_3 = (-2, -1)$, $M_4 = (2, -1)$, $M_5 = (2, 1)$ et $M_6 = (1, 0)$.
1. En supposant que les deux premiers points M_1 et M_2 sont les centres initiaux, décrire par une succession de dessins, l'algorithme des centres mobiles en représentant les centres, les classes, les nouveaux centres ... jusqu'à stabilisation de l'algorithme. On calculera au passage si nécessaire les coordonnées des centres.

2. Recommencer en choisissant différemment les centres initiaux. Obtient-on la même classification ?

Exercice 7 : Classifier les points du nuage précédent par une classification hiérarchique ascendante et représenter le dendrogramme (à noter que lorsqu'on doit regrouper les deux points les plus proches et qu'il existe deux couples de points satisfaisant cette condition, on convient de choisir les deux points dont les numéros sont les plus petits).