Cours 9 : Classification automatique de données et théorie des graphes

L'inconvénient de la méthode de classification hiérarchique ascendante que nous avons étudiée est l'usage qui est fait de l'écart de Ward qui nécessite, à chaque étape, le calcul du centre de gravité des classes obtenues, calcul qui n'a de sens que si l'on dispose des coordonnées des points (dans un espace euclidien). L'écart de Ward n'est cependant pas la seule façon de mesurer la distance entre des classes, il y a beaucoup d'autres distances qui peuvent se calculer même si l'on n'a pas de coordonnées pour les points pourvu que l'on connaisse une "distance" entre chaque pair de point (matrice des distances). Parmi elles la *distance du plus proche voisin* (*single linkage* ou *nearest neighbor*) est particulièrement intéressante car elle est en relation directe avec une technique très utile de *théorie des graphes* qui est la recherche de l'arbre de longueur minimale (*Minimum Spanning Tree*).

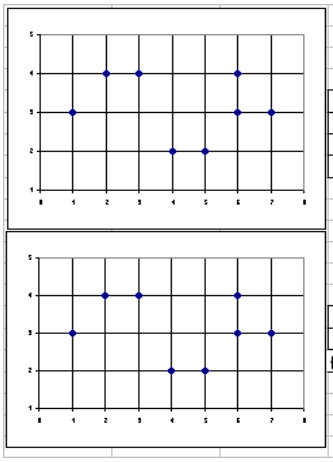
Agglomération au plus proche voisin :

La méthode de classification automatique par agglomération au plus proche voisin est une forme de classification hiérarchique ascendante : on part d'une matrice de "distances" entre les points, 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). On commence par former une classe avec les deux points les plus proches et, à l'étape n, on calcule les distances entre les classes deja formées en utilisant la formule

$$d(\Gamma_1, \Gamma_2) = \mathrm{Min}_{x_1 \in \Gamma_1, x_2 \in \Gamma_2} d(x_1, x_2)$$

puis on agglomère les deux classes les plus proches en une classe unique.

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.



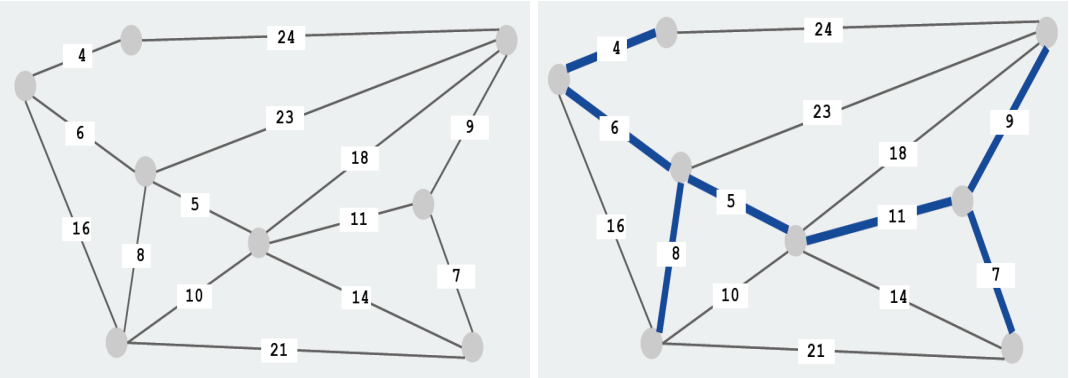
	Γ1	Γ2	Γ3
Γ1	0	9	5
Γ2	9	0	2
Γ3	5	2	0

	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.

Arbre de longueur minimale (MST) : La théorie des graphes est un chapitre important de l'informatique qui est utilisé dans beaucoup de domaines appliqués, de la bioinformatique à la recherche opérationnelle (optimisation de la production d'une entreprise) en passant par le marketing ou les reseaux électriques. Un *graphe valué* est un ensemble de points, qu'on appelle *sommets*, reliés entre eux par des *arrêtes* qui portent chacune une valeur qu'on appelle longueur ou poids de l'arrête. Lorsque toute pair de sommets est relié par une arrête, on dit que le graphe est *complet*. Un graphe valué peut aussi

être *orienté* comme les diagramme en points et flèches que nous avons associés aux chaines de Markov. Les graphe que nous étudions à présent ne sont pas orientés mais ils sont en général *connexe* ce qui sigifie qu'il existe un chemin entre chaque pair de sommets, composé d'une ou de plusieurs arrêtes. Un exemple de graphe valué est donné ci dessous (figure de gauche). Il n'est pas complet mais il est connexe.



Parmi les graphes les plus étudiés figurent les *arbres :* ce sont des *graphes connexes et sans cycles*. Un cycle est un chemin partant et aboutissant au même sommet sans emprunter deux fois la même arrête. L'une des propriété importante des arbres est qu'on peut toujours les représenter dans un plan sans que leurs arrêtes ne se recoupent.

Etant donné un graphe, on peut en extraire un *graphe partiel*, c'est-à-dire un graphe ayant les mêmes sommets mais moins d'arrêtes. Par exemple le graphe représenté en gras sur la figure de droite ci-dessus est un graphe extrait de celui représenté à gauche. Ce graphe partiel est un arbre et il est *couvrant* ce qui signifie qu'il contient tous les sommets du graphe initial. La *longueur* de cet arbre est la somme des poids de ses arrêtes. Dans cet exemple, la longueur de l'arbre est 50 ($4+6+8+5+11+9+7=50$).

On peut montrer qu'un graphe possède toujours un graphe partiel qui est un arbre couvrant de longueur minimale et il est même unique pourvu les poids de toutes les arrêtes soient différents. La recherche de cet arbre couvrant de longueur minimale (MST) est un problème classique d'informatique qui possède de nombreuses applications pratiques. Il nous intéresse ici car nous allons voir que c'est un problème équivalent à celui de la classification automatique par agglomération au plus proche voisin.

Il existe de nombreux algorithmes qui permettent de trouver cet arbre couvrant de longueur minimale. Nous indiquons ici l'un des plus utilisé, l'algorithme de Kruskal (1956).

On suppose que le graphe initial possède n sommets. On range les arrêtes par ordre de poid croissant. La première arrête de l'arbre est celle de poid minimal. L'arbre qui, à ce stade, ne comporte que 2 sommets et l'arrête qui les joint va être complété progressivement de la façon suivante. On ajoute à chaque étape l'arrête qui a le poid le plus petit parmi celles qui ne font pas encore partie de l'arbre sauf si le fait de l'ajouter crée un cycle. On interrompt la procédure lorsque l'arbre comporte $n-1$ arrêtes.

Ainsi dans l'exemple ci-dessus, où $n=8$, l'algorithme posera successivement les arrêtes de poid 4, 5, 6, 7, 8, puis 9. A ce stade, l'arrête suivante serait 10 mais elle créerait un cycle, on ne la pose donc pas et on pose alors la 11 qui sera la $n-1$ième.

La figure ci dessous illustre les liens qui existent entre l'arbre couvrant de longueur minimal et le dendrogramme dans le cas très simple d'un ensemble à 5 points.