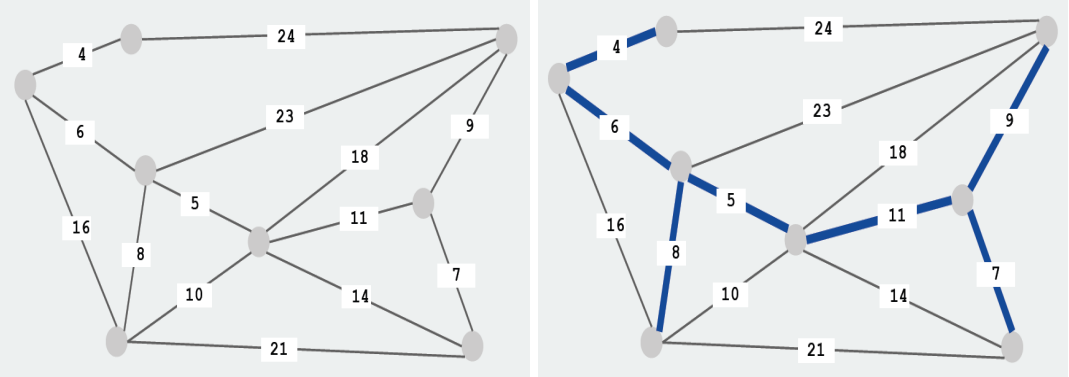
Cours 9 : Classification automatique de données et théorie des graphes

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.

1 Graphes

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.



2 Arbres

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 que tous les poids des arrêtes soient distincts. La recherche de cet arbre (que l'on appelle *Minimum Spanning Tree*) 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.

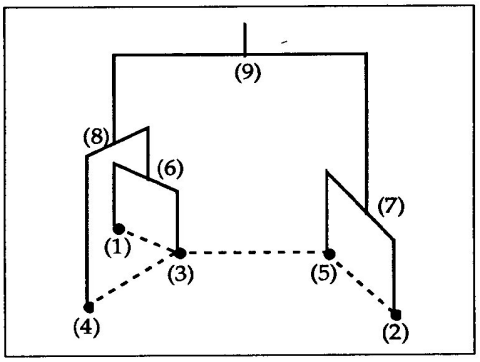
3 Recherche de l'arbre de longueur minimal (MST)

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 minimale et le dendrogramme dans le cas très simple d'un ensemble à 5 points.



On y voit en pointillé l'arbre couvrant de longueur minimale et au dessus le dendrogramme obtenu par une classification hiérarchique ascendante des 5 points du nuage, utilisant l'aglomération au plus proche voisin. Le choix d'une autre distance entre classes aurait pu produire une classifacation différente non necessairement compatible avec cet arbre. Mais on peut montrer que la distance du plus proche voisin (*single linkage* ou *nearest neighbor*) conduit toujours à un dendrogramme ayant les même ramifications que l'arbre de longueur minimal.