CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITÉ PARIS 13 NORD

Cours 6 :
Traitement de données quantitatives multivariées
-
Introduction au partitionnement de données

26 novembre 2012

Introduction

Généralités

Le partitionnement de données (data clustering en anglais) est une méthode statistique d'analyse des données qui a pour but de regrouper un ensemble de données en différents paquets homogènes, c'est à dire que les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité que l'on définit en introduisant des mesures de distance.

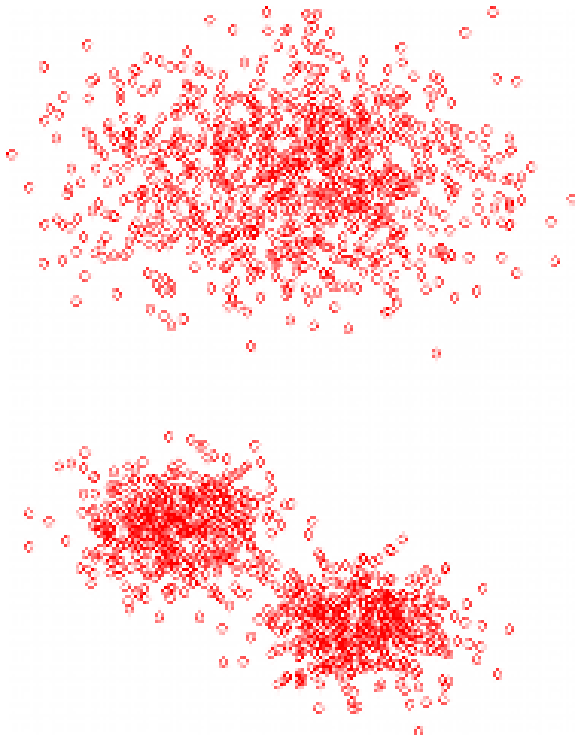
Généralités

Pour obtenir un bon partitionnement, il faut :

- Minimiser l'inertie (la variance) intra-classe pour obtenir des classes (= groupes ou *clusters*) les plus homogènes possibles.
- Maximiser l'inertie inter-classe afin d'obtenir des groupes bien différenciés.

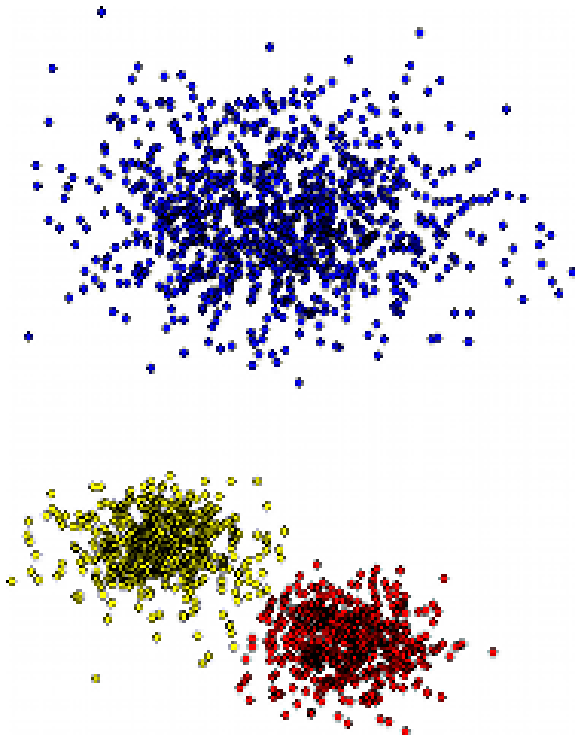
Exemple

On veut découvrir automatiquement les groupes de données similaires :



Exemple

On veut découvrir automatiquement les groupes de données similaires :



Méthode des K-Moyennes

Principe

La méthode des K-Moyennes est un cas particulier de la méthode des centres mobiles.

L'objectif principal de ces méthodes est de choisir un certain nombre de représentants (= centre ou prototypes) dans l'espace des données. Chaque prototype représente un groupe.

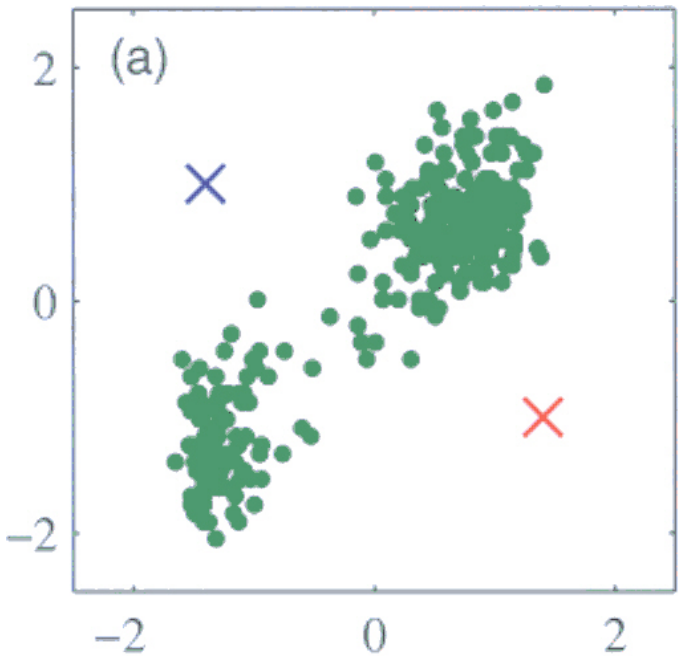
Ainsi à la fin du processus on associe chaque point de donnée à son prototype le plus proche, de façon à obtenir une segmentation des données en différents groupes homogènes.

Algorithme élémentaire

1. Choisir k centres initiaux au hasard (par exemple parmi les données).
2. Affecter chaque donnée x au groupe i dont le centre Ci est le plus proche de x.
3. Si aucun élément ne change de groupe alors arrêt.
4. Sinon calculer les nouveaux centres : pour tout i, Ci est le centre de gravité (le barycentre) des éléments du groupe i.
5. Aller en 2.

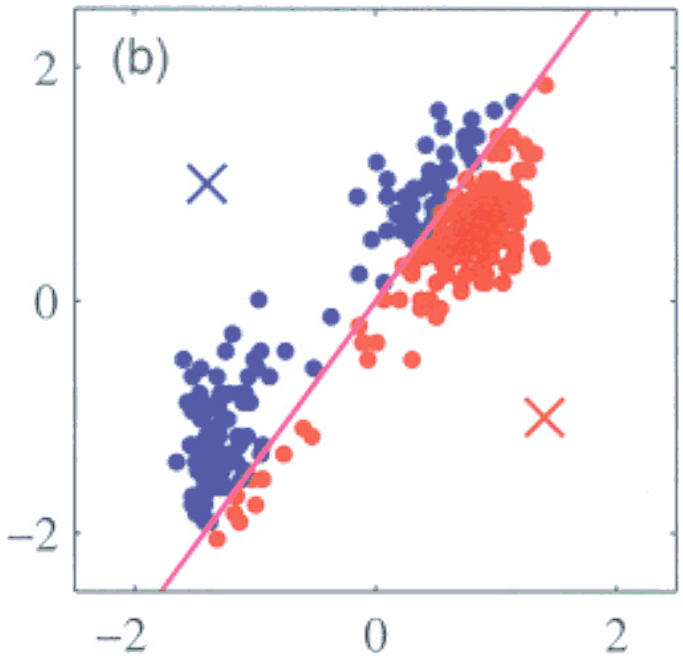
Algorithme élémentaire

Exemple :



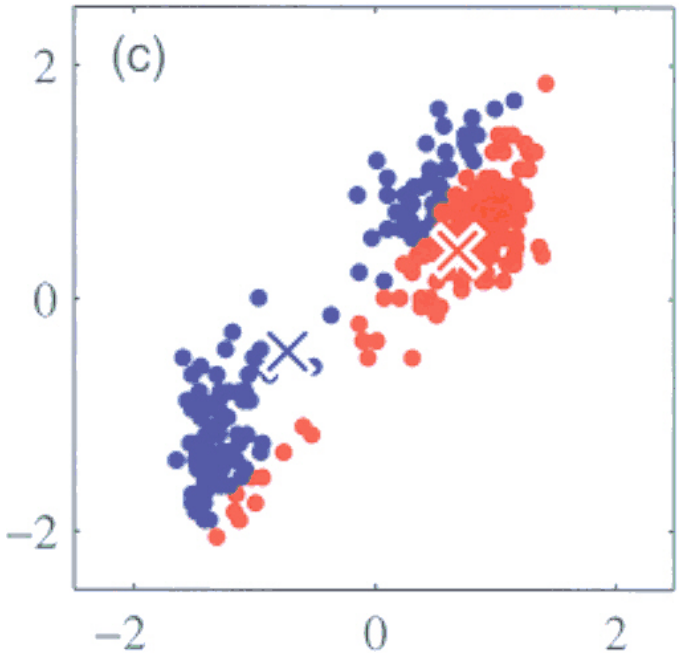
Algorithme élémentaire

Exemple :



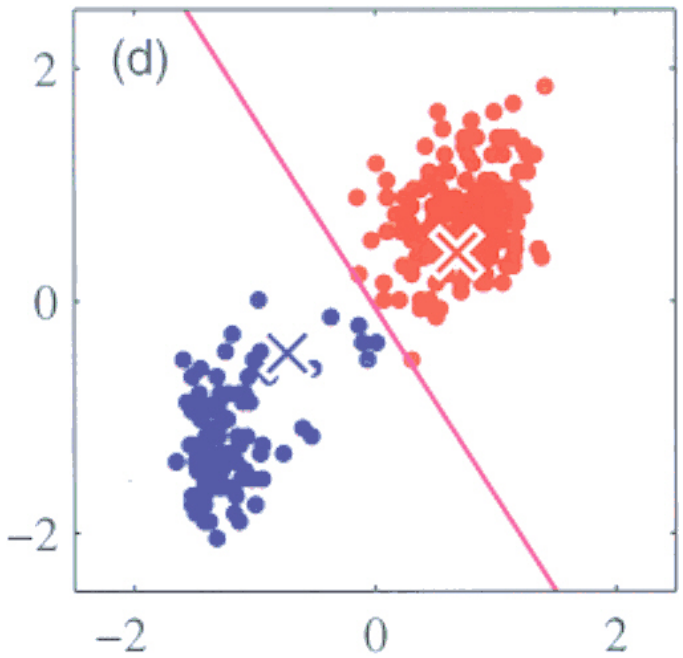
Algorithme élémentaire

Exemple :



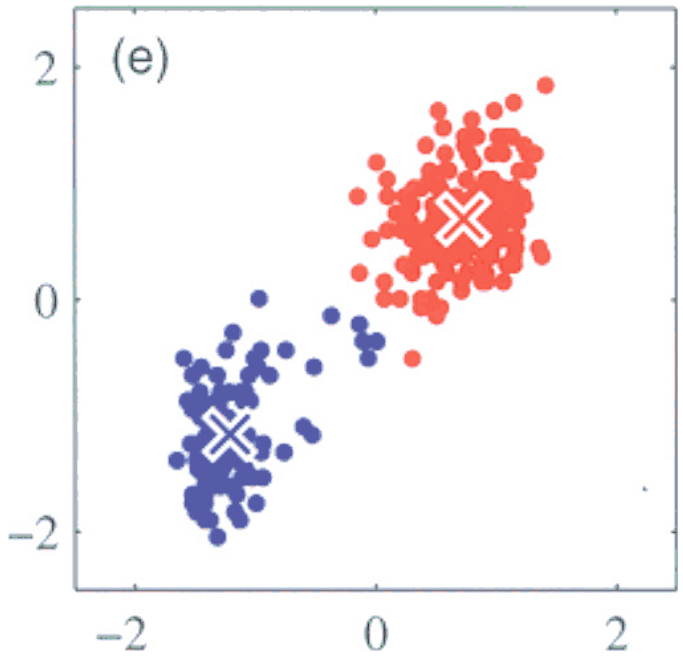
Algorithme élémentaire

Exemple :



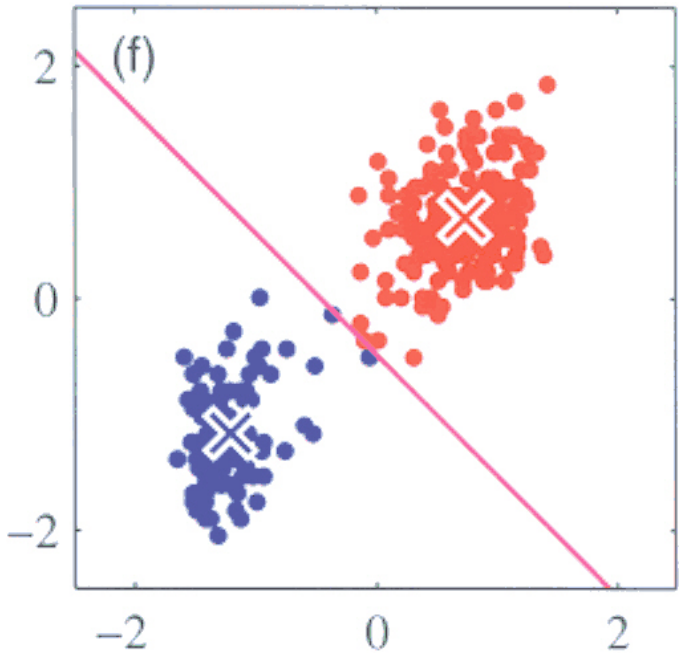
Algorithme élémentaire

Exemple :



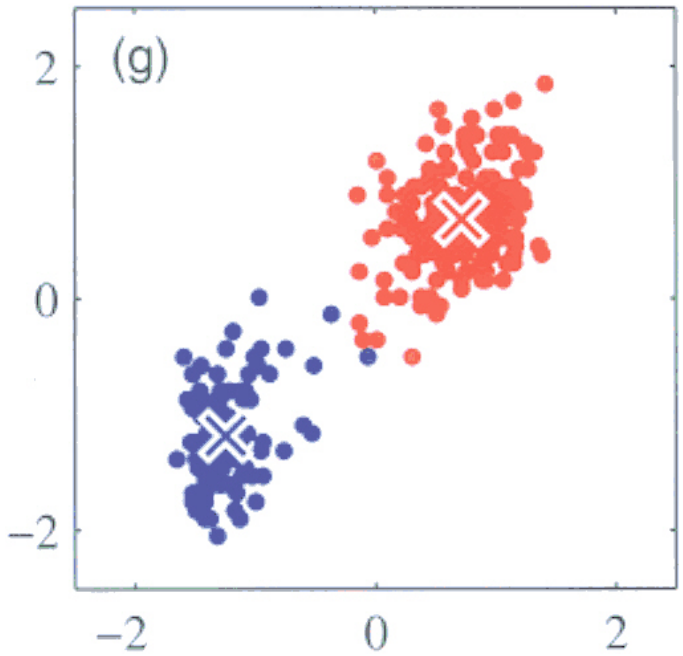
Algorithme élémentaire

Exemple :



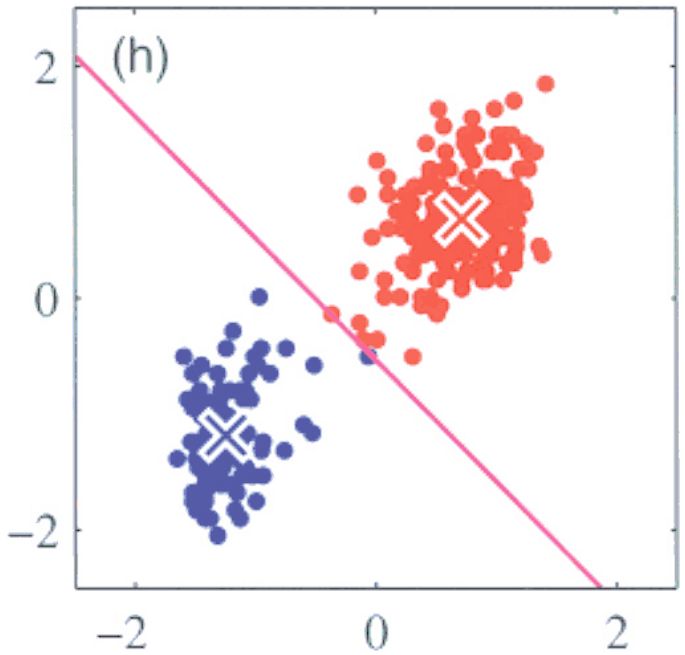
Algorithme élémentaire

Exemple :



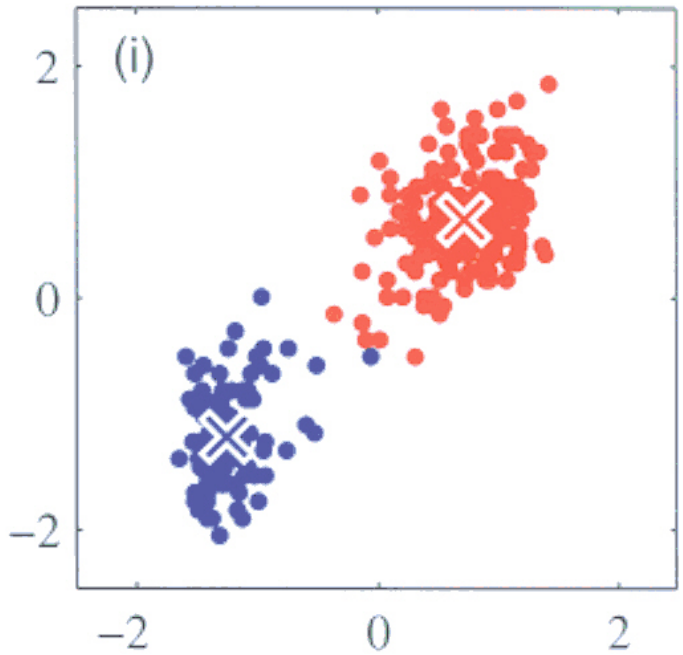
Algorithme élémentaire

Exemple :



Algorithme élémentaire

Exemple :



Problemes

Choix de la mesure de distance (métrique)

Ce choix est très important, avec des mesures de distance différentes on obtient des résultats différents !

Le plus souvent on utilise la distance euclidienne :

$$\sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

avec n le nombre de variables.

Problemes

Instabilité

Le résultat final est fortement dépendant de l'initialisation des centre. Pour remédier à ça on lance l'algorithme plusieurs fois avec une initialisation aléatoire et on ne garde que le meilleur résultat.

Choix du nombre de groupes

Le nombre de groupes obtenu en sortie de l'algorithme doit être choisi par l'utilisateur. Or en général il n'est pas connu ! On lance donc généralement l'algorithme plusieurs fois avec des choix différents pour le nombre de groupes et on ne garde que le meilleur résultat.

Mesure de performance

On a vu qu'il est nécessaire de relancer l'algorithme des K-Moyennes de nombreuses fois et de garder le meilleur résultat. Il nous faut donc une estimation de la qualité de la segmentation.

Indice de Davies-Bouldin

Soit s_i la distance moyenne des données du groupe i à leur prototype c_i, cet indice sélectionne la segmentation qui maximise la distance entre groupe et minimise la variance intra-groupe. C'est un des plus utilisées.

$$DB(K) = \frac{1}{K} \sum_{i=1}^{K} max_{j \neq i} \frac{s_i + s_j}{dist(c_i, c_j)}$$

Algorithme Effectif

1. Choisir k entre 1 et k_{max}.
2. Lancer 50 fois l'algorithme élémentaire de K-Moyennes et garder la meilleure segmentation S_k selon l'indice de Davies-Bouldin.
3. Si toute les valeurs de k on été testé, retenir parmi les S_k la meilleure segmentation S selon Davies-Bouldin, sinon aller en 1.

Méthode Ascendante Hiérarchique

Principe

Créer à chaque étape une partition obtenue en agrégeant 2 à 2 les éléments les plus proches (élément = donnée ou groupe de données). L'algorithme fournit une hiérarchie de partitions : arbre contenant l'historique de la classification et permettant de retrouver n-1 partitions.

- Nécessité de se munir d'une métrique (distance euclidienne, ...).
- Nécessité de fixer une règle pour agréger une donnée ou un groupe de donnée avec un autre groupe : le critère d'agrégation.

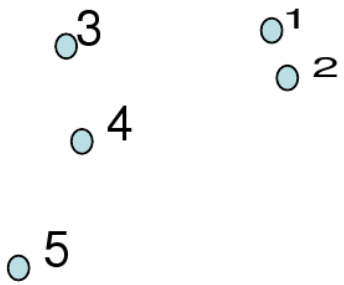
Algorithme

1. Calculer la matrice de distance entre les n éléments et regrouper les 2 éléments les plus proches.
2. Si toutes les données ne sont pas regroupées en un seul groupe, retourner en 1.
3. Sinon construire le dendrogramme (arbre hiérarchique) et utiliser un critère de qualité (Davies-Bouldin,...) pour choisir la coupure la plus pertinente.
4. En déduire une segmentation des données.

Algorithme

Exemple :

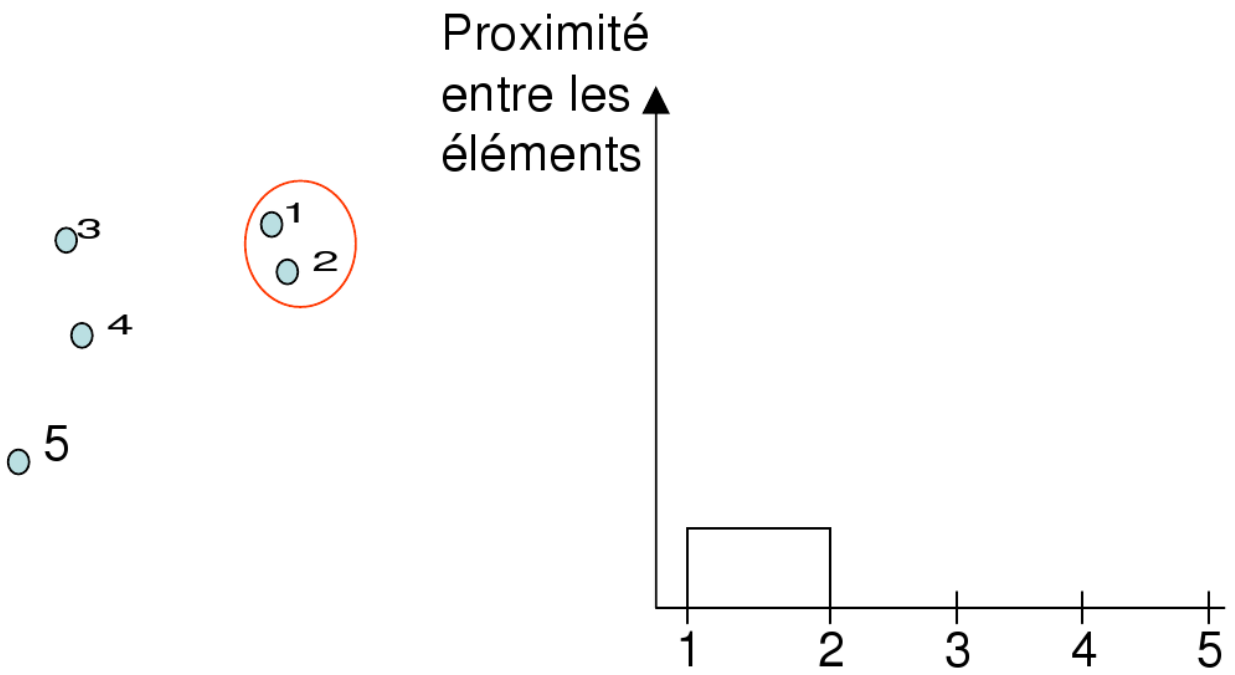
Etape 1 : n individus / n classes



Algorithme

Exemple :

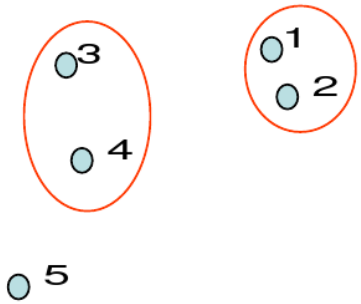
Etape 2 : n -1 classes

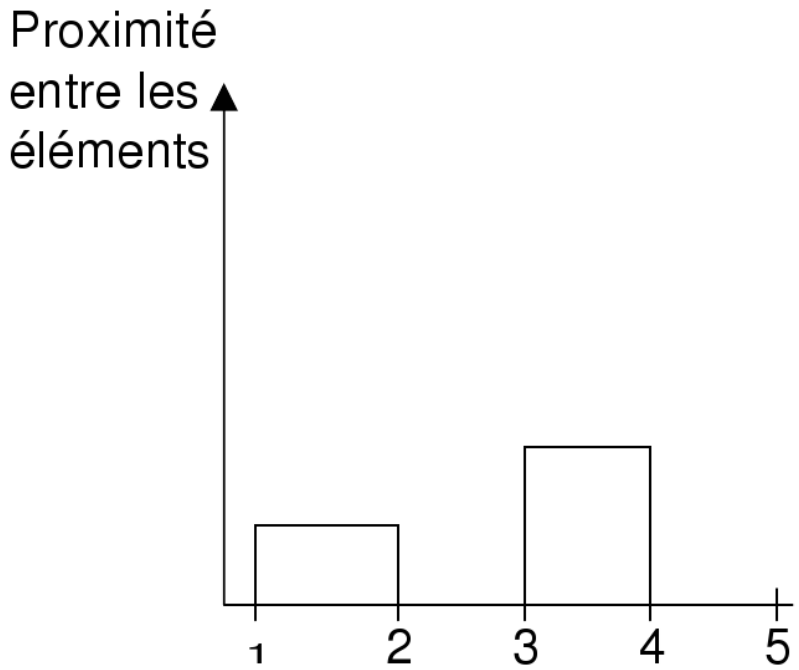


Algorithme

Exemple :

Etape 3 : n -2 classes

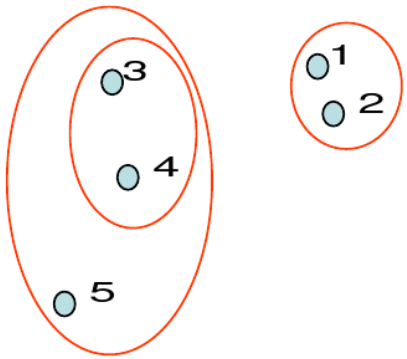


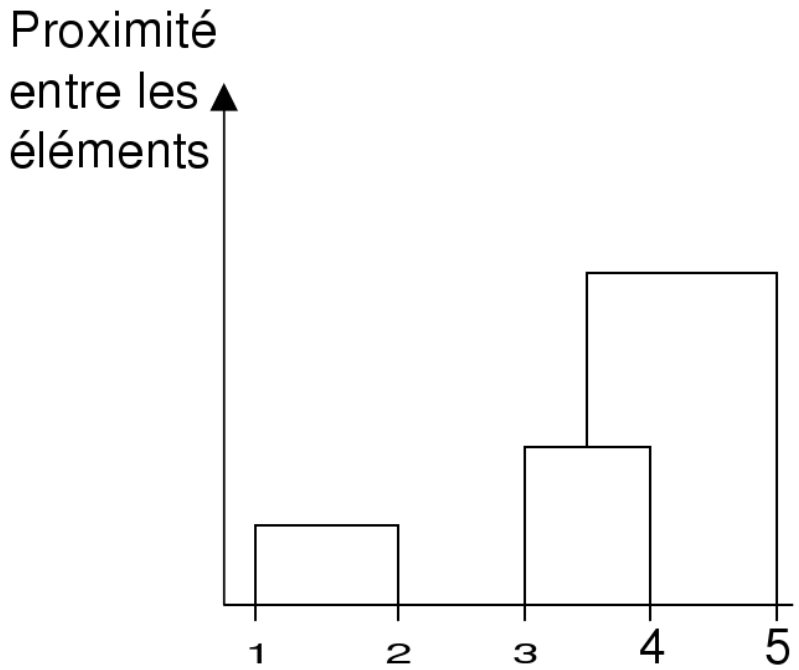


Algorithme

Exemple :

Etape 4 : n -3 classes

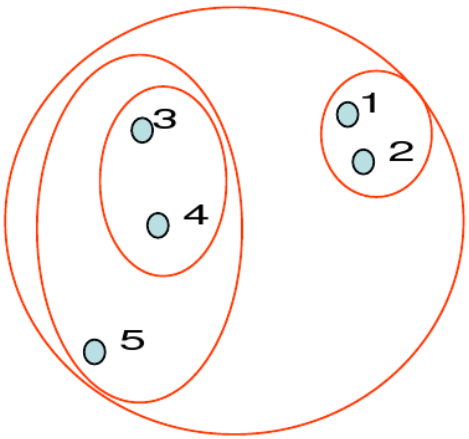


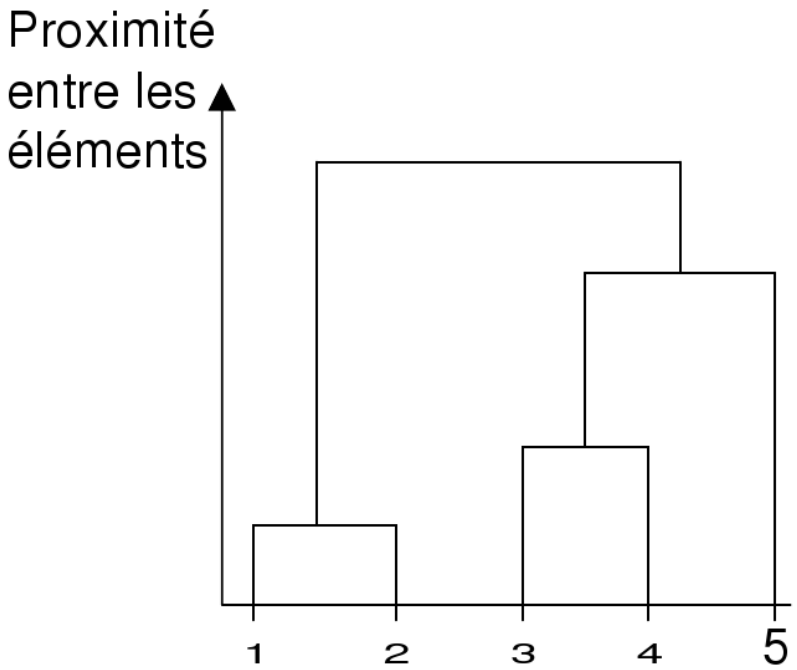


Algorithme

Exemple :

Etape 5 : n -4 = 1 classe

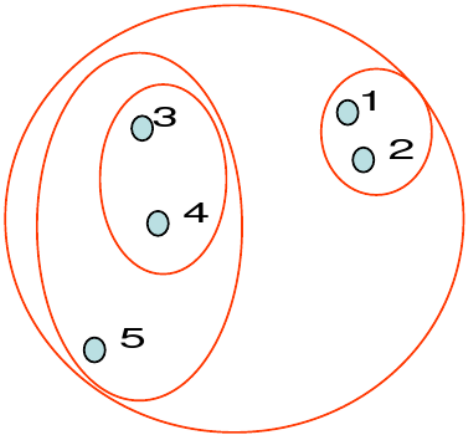


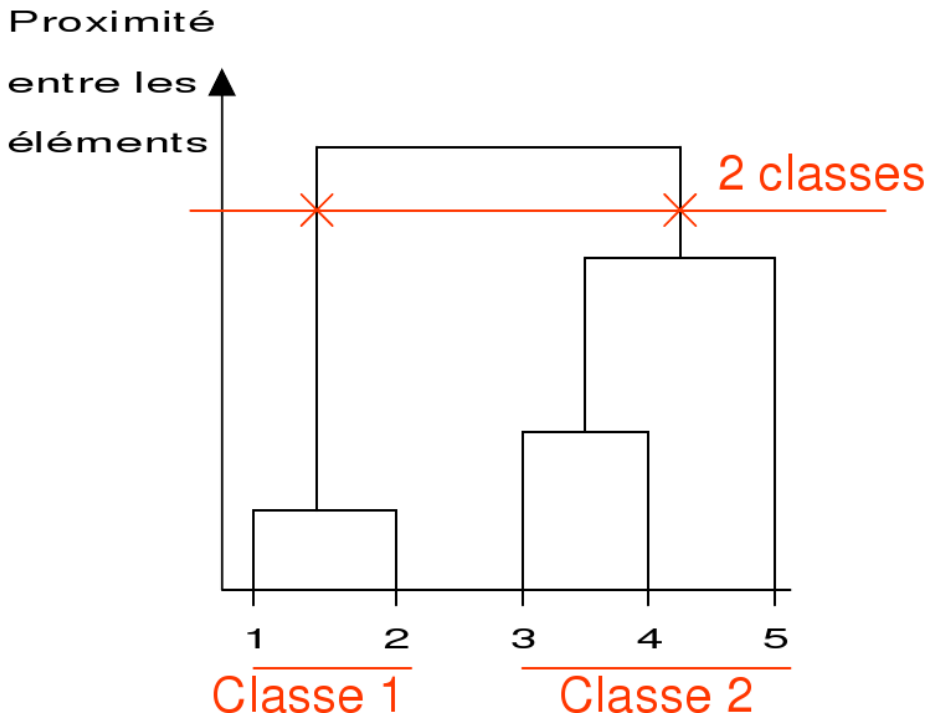


Algorithme

Exemple :

Etape 5 : n -4 = 1 classe





Problèmes

Choix de la règle d'agrégation

Ce choix est très important, avec des règles différentes on obtient des résultats différents ! Il existe de nombreuses règles possibles :

- Distance entre les barycentres des deux éléments.
- Distance entre les deux données les plus proches des deux éléments.
- Distance entre les deux données les plus éloignées des deux éléments.
- Distance moyenne entre une données d'un élément et une donnée de l'autre élément.

Conclusion

Conclusion

Il existe de très nombreuses méthodes de segmentation des données. Les résultats obtenus dépendent :

- De l'algorithme utilisé (K-Moyennes, Méthodes Ascendantes selon la règle d'agrégation, Méthodes Descendantes, ...).
- De la métrique (distance Euclidienne, distance de Manhattan, distance de Minkowski, ...).
- De l'indice de performance (Davies-Bouldin, Calinski-Harabatz, Silhouette, ...).

Cependant plus les groupes sont compacts et bien séparés, plus les différentes méthodes aurons tendance à donner les mêmes résultats.