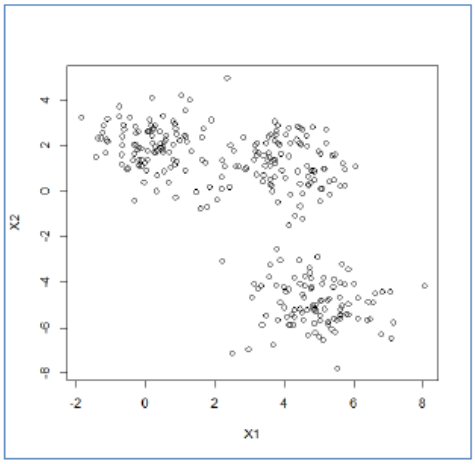
Clustering Algorithme Kmeans (Kmoyennes)

Clustering : Classification Automatique Non supervisée

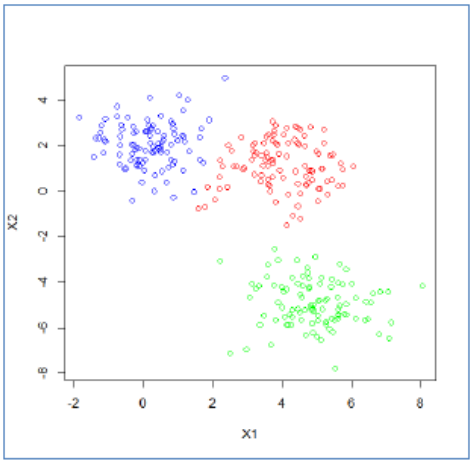
Parmi ces types, le Partitionnement. Il s'agit pour l'algorithme de mettre en évidence des groupes homogènes (naturels), qui se démarque les uns des autres.

Les individus d'un même groupe (cluster) se « ressemble » et sont différents des individus des autres groupes (clusters).

Schématiquement et dimension 2 (pour faciliter la compréhension, il s'agit de passer de la situation 1 à la situation 2



Situation 1



Situation 2

Deux paramètres (questions) à décider :

Le K : nombres de cluster

Et la mesure de distance à utiliser pour délimiter les clusters

Principales Caractéristiques :

- Fixer a priori le nombre de cluster : Mais peut être modifié en fonction de l'objectif fixé ou de contrainte. Reste un probleme ouvert.
- Définir une partition de départ des données : Souvent de manière aléatoire, mais peut faire l'objet d'une étude ou l'utilisation d'une méthode prenant en considération la répartition des individus.
- Réallocation. Déplacement des objets d'un groupe à l'autre pour obtenir une meilleure partition.

Algorithme K-Means : Lloyd 1957, Forgy 1965, MacQueen 1967

En 4 étapes :

1. Choisir k objets formant ainsi k clusters
2. (Ré)affecter chaque objet O au cluster C_i de centre M_i tel que dist(O, $M_{i)}$ est minimal

3. Recalculer Mi de chaque cluster (le barycentre)
4. Aller à l'etape 2 si on vient de faire une affectation

Avantages :

- Simplicité
- Scalabilité : Capacité à traiter les très grandes bases.
- Complexité linéaire par rapport au nombre d'observations

Inconvénients :

- La lenteur : plusieurs passages
- L'optimisation aboutit à un minimum local
- La solution dépend du choix initial des centres de clusters
- La solution peut dépendre de l'ordre des individus
- Ne traite que les variables numériques dans sa version de base.
- Résumer un cluster par le barycentre n'est pas toujours pertinent, voir K-Medoids

Exemple : Utilisation de l'algorithme Kmeans et de la distance d'Euclide pour le clustering des 8 données en 3 clusters : A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5 =(7,5),A6=(6,4), A7=(1,2), A8=(4,9).

On supposera que les centres initiaux sont A1, A4, et A7.

a.

	A1	A2	A 3	A4	A5	A6	A7	A8
C1(A1)	0 *	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
C2(A4)	$\sqrt{13}$	$\sqrt{18}$	$\sqrt{25}$ *	0 *	$\sqrt{13}$ *	$\sqrt{17}$ *	$\sqrt{52}$	$\sqrt{2}$ *
C3(A7)	$\sqrt{65}$	$\sqrt{10}$ *	$\sqrt{53}$	$\sqrt{52}$	$\sqrt{45}$	$\sqrt{29}$	0 *	$\sqrt{58}$

Pour A1=(2,10), et A4=(5,8), nous avons :

Dist(A1,A4) = $\sqrt{(2-5)^2+(10-8)^2}$ = $\sqrt{9+4}$ = $\sqrt{13}$

Le point de donnée sera selon l'algorithme Kmeans affecté au cluster de centre le plus proche.

Les nouveaux clusters sont :

1 : {A1}, 2 : {A3, A4, A5, A6, A8}, 3 : {A2,A7}

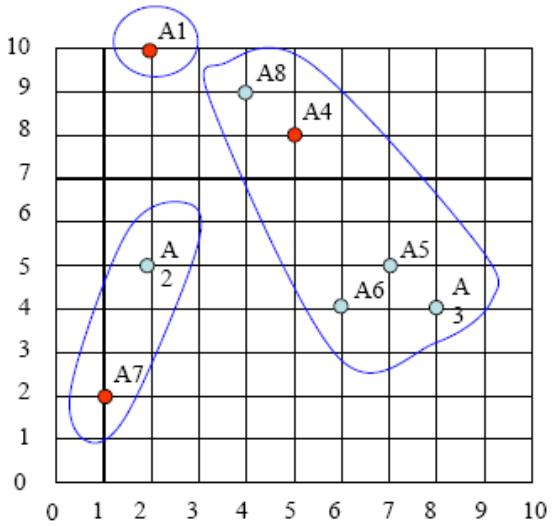
b.

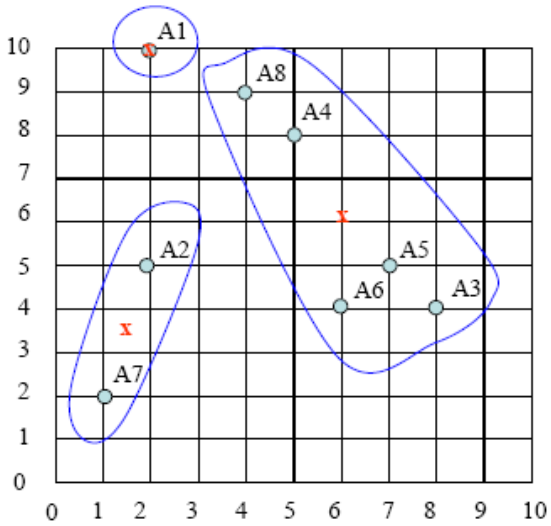
Les centres des nouveaux clusters :

C1= (2,10)

C2= ((8+5+7+6+4)/5,(4+8+5+4+9)/5)=(6,6)

C3=((2+1)/2,(5+2)/2)= (1.5,3.5)





On continue :

	A1	A2	A 3	A4	A5	A6	A7	A8
C1								
C2								
C3								

Les nouveaux clusters sont : 1 : {}, 2 : {}, 3 : {}

De centres :

C1= ()

C2=()

C3=()

Et encore une fois :

	A1	A2	A 3	A4	A5	A6	A7	A8
C1								
C2								
C3								

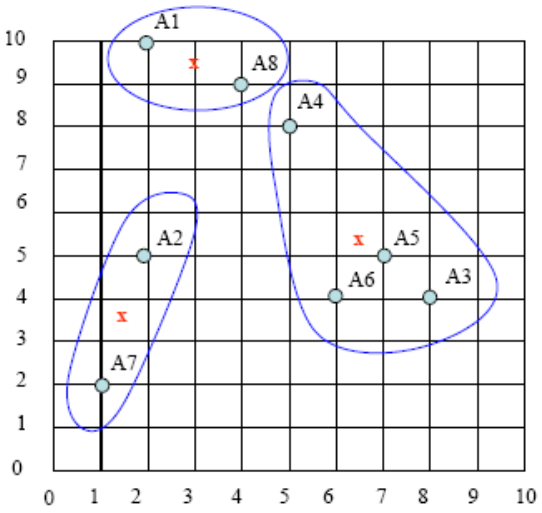
Les nouveaux clusters : 1 : {}, 2 : {}, 3 : {}

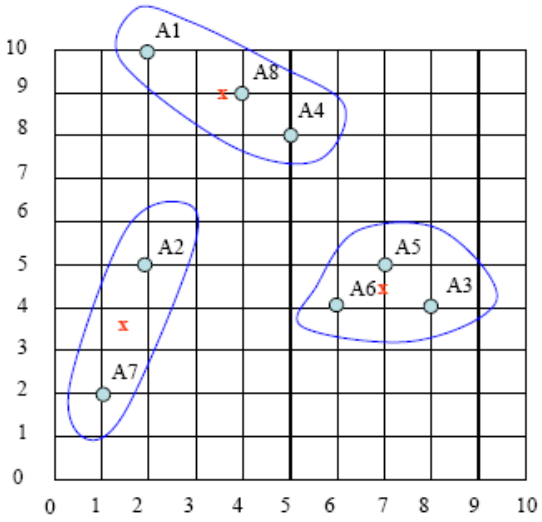
De centres :

C1= ()

C2=()

C3=()


A1
A8
A4
A2
A5
A6
A3
A7
0 1 2 3 4 5 6 7 8 9 10


A1
A8
A4
A2
A5
A6
A3
A7
0 1 2 3 4 5 6 7 8 9 10