Data Mining

Mohammed Fethi KHALFI

Fethi.Khalfi@yahoo.fr

Algorithme des centres mobiles (k means)

Université Djilali Liabés- Sidi Bel Abbes

k - means

K-moyennes

Principe de fonctionnement

Un cluster : un regroupement de donnée.

Regrouper tout ce qui se ressemble,

Apprentissage non supervisé

Données

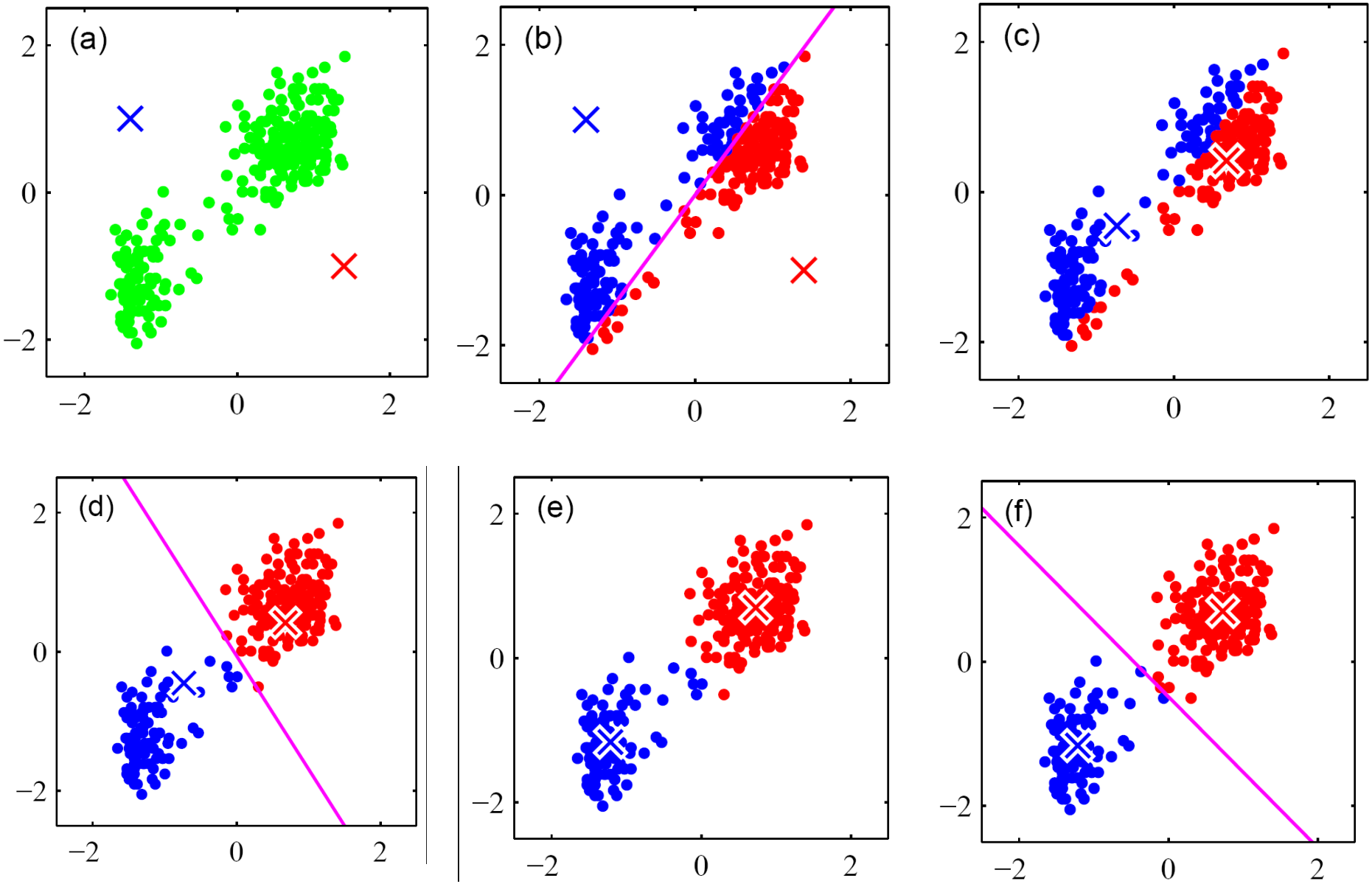
Clustering

K-Means : Définition

L'algorithme des centres mobiles vise à classer une population X en K classes. Cela se fait de manière automatique. Il est le mieux adapté aux très grands tableaux de données.

- Méthode des K-moyennes (*MacQueen'67)*
  - choisir K éléments initiaux "centres" des K groupes
  - placer les objets dans le groupe de centre le plus proche
  - recalculer le centre de gravité de chaque groupe
  - itérer l'algorithme jusqu'à ce que les objets ne changent plus de groupe
- Encore appelée méthode des centres mobiles

Description du problème :



Les distance les plus usuelles

Comment définir la distance ?

Il existe plusieurs fonctions de calcul de distance on choisit la fonction de distance en fonction des types de données qu'on manipule.

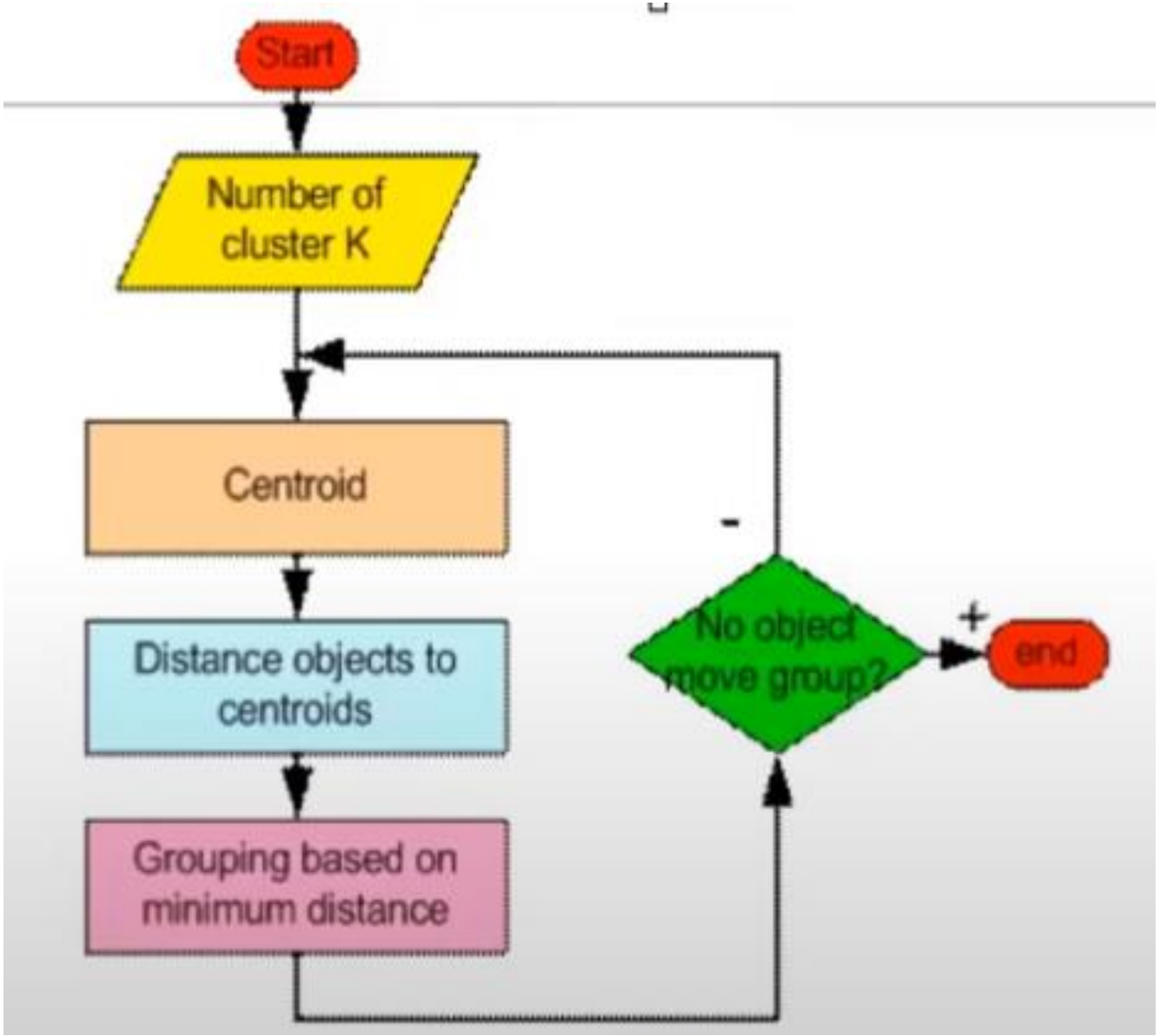
Pour les données **quantitatives** (exemple : poids, salaires, taille, etc.…) et du même type, **la distance euclidienne** est un bon candidat.

Les distance les plus usuelles

- There are many possible distance measures
  - Euclidean Distance: $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
  - Manhattan Distance or City Block Distance :

$$\sum_{i=1}^{k}\left|x_i - y_i\right|$$

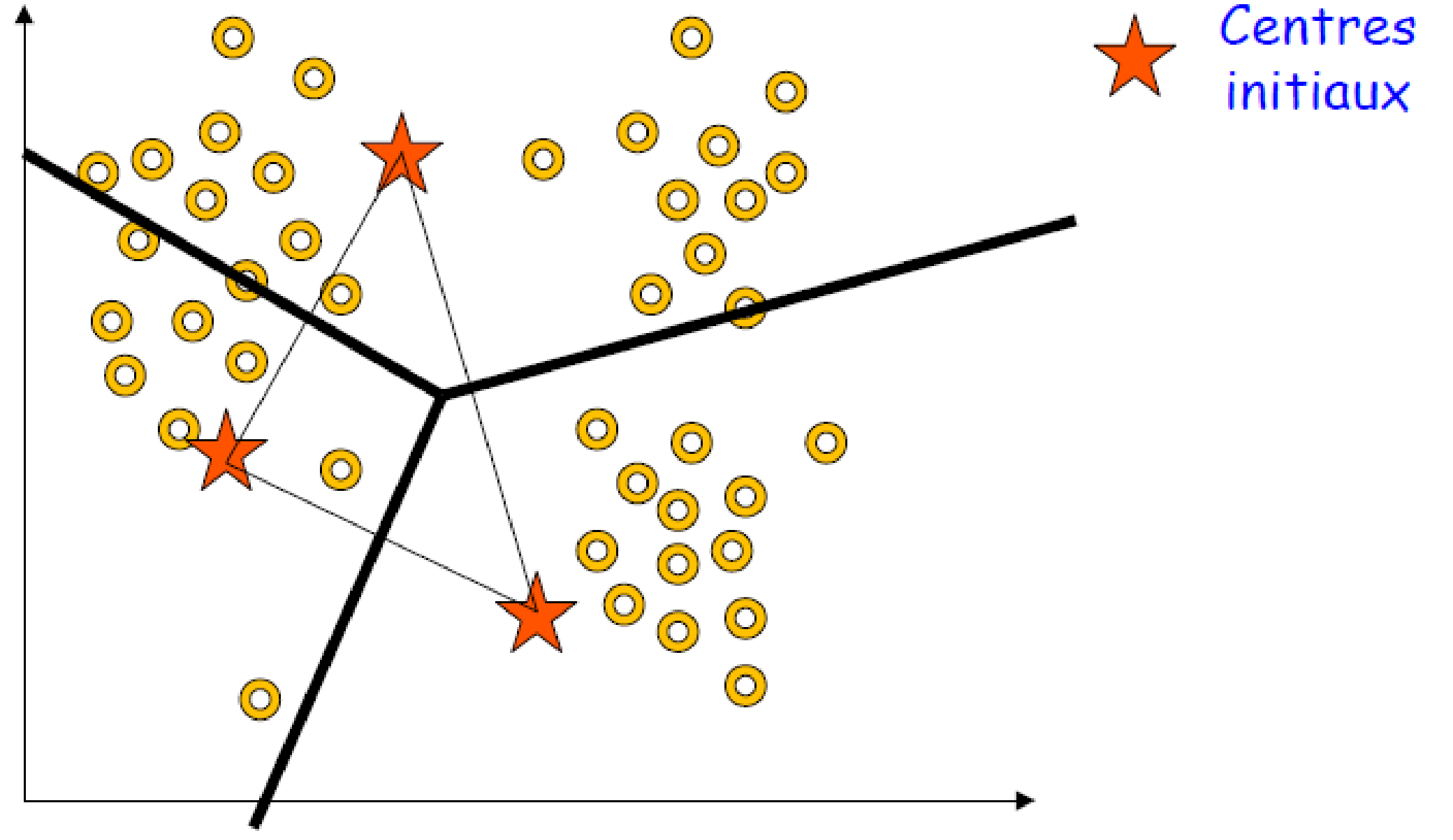
Algorithme K-means



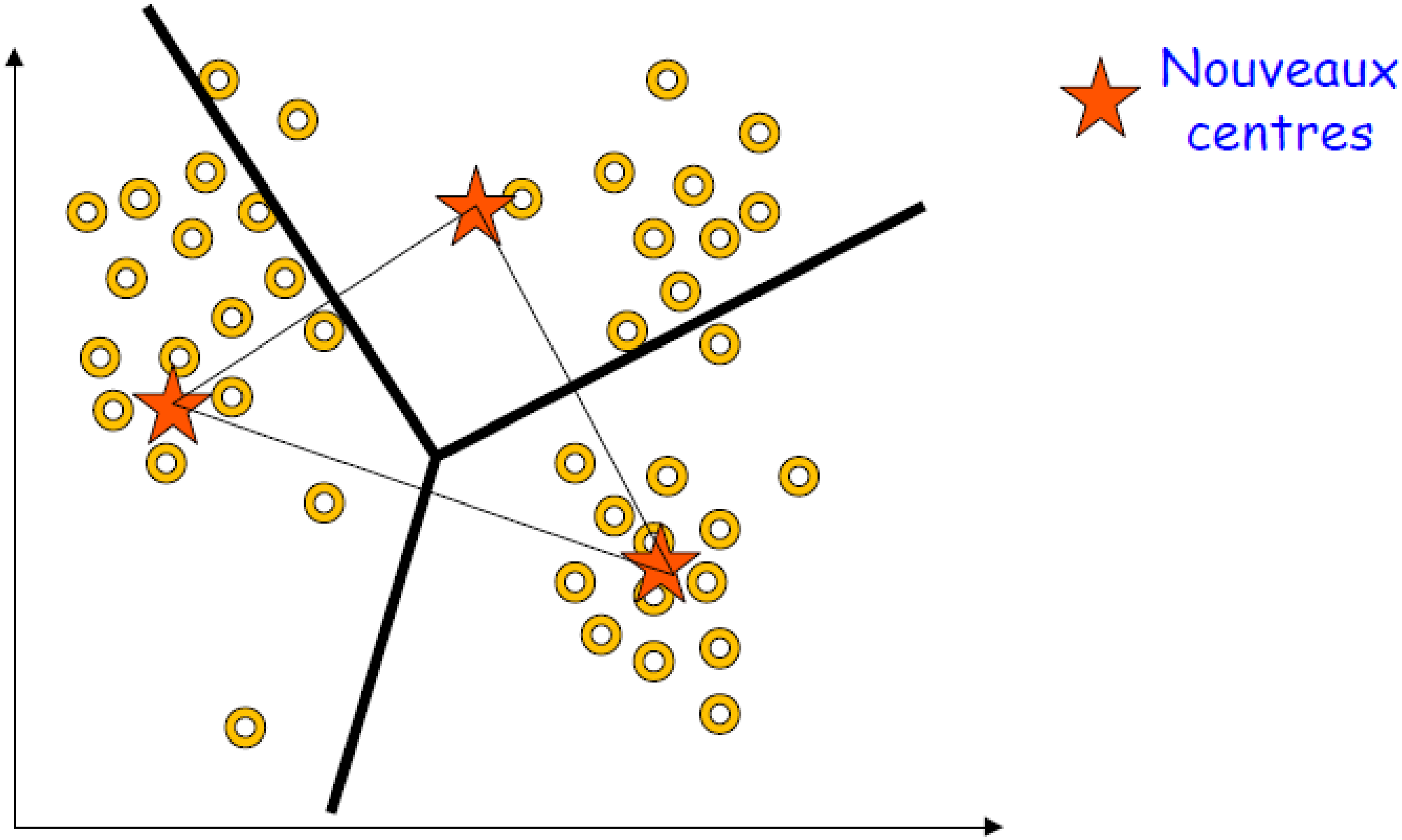
Algorithme

- Étapes:
  - fixer le nombre de clusters: k
  - choisir aléatoirement k tuples comme graines (centres)
  - assigner chaque tuple à la graine la plus proche
  - recalculer les k graines
  - tant que des tuples ont été changés
    - réassigner les tuples
    - recalculer les k graines
- C'est l'Algorithme le plus utilisé

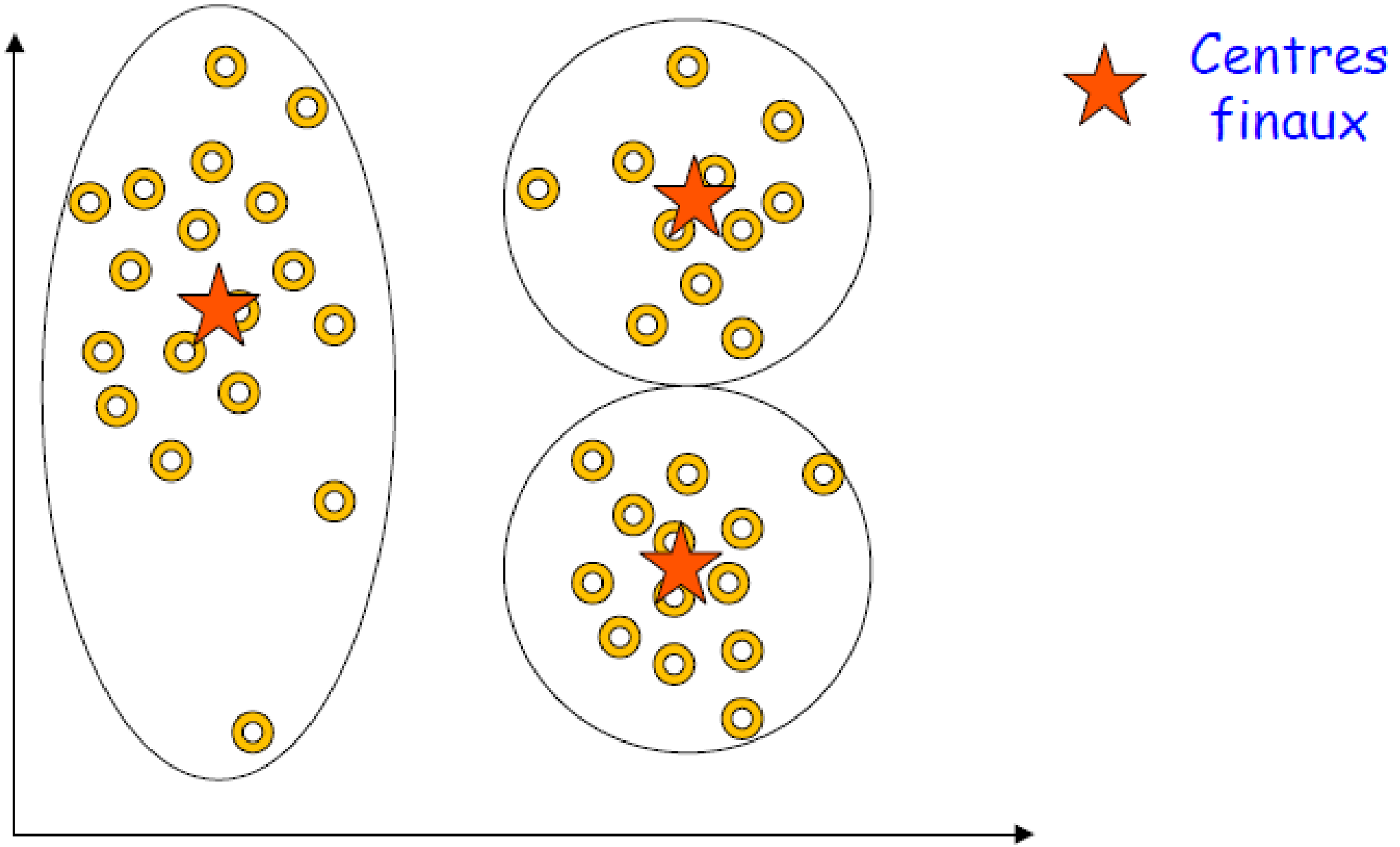
Illustration



Illustration



Illustration



Kmeans

- Soit l'ensemble D des entiers suivants :
- D= { 2, 5, 8, 10, 11, 18, 20 }

- On veut répartir les données de D en trois (3) clusters, en utilisant l'algorithme Kmeans. La distance d entre deux nombres a et b est calculée ainsi :
- d(a , b) = |a - b| (la valeur absolue de a moins b)
- - Appliquez Kmeans en choisissant comme centres initiaux des 3 clusters respectivement : 8, 10 et 11.

Kmeans

D= { 2, 5, 8, 10, 11, 18, 20 }

Les centres initiaux des 3 clusters respectivement sont :

C1={8}

C2={10}

C3={11}

- d(a , b) = |a - b|

μ1 = 8, **μ2 = 10,** **μ3 = 11**

Ptx	**2**	**5**	**8**	10	11	18	20
d(**μ1**)	6	3	0	2	3	10	12
d(**μ2**)	8	5	2	0	1	8	10
d(**μ3**)	9	6	3	1	0	7	9
Classe	C1	C1	C1	C2	C3	C3	C3

C1 = {2 ; 5 ; 8}, **C2={10},** **C3={11 ; 18 ; 20}**

Mise à jour des clusters :

Kmeans

R- estimation des centres de gravité :

$\mu 1 = (2+5+8)/3$

$\mu 2 = 10/1$

$\mu 3 = (11+18+20)/3$

$\mu 1 = 5$

$\mu 2 = 10$

$\mu 3 = 16.33$

$\mu 1$ = 5, **$\mu 2$** = 10, **$\mu 3$** = 16,3

Ptx	**2**	**5**	**8**	10	11	18	20
d(**$\mu 1$**)	3	0	3	5	6	13	15
d(**$\mu 2$**)	8	5	2	0	1	8	10
d(**$\mu 3$**)	14,33	11,33	8,33	6,33	5,33	2,33	4,33
Classe	C1	C1	C2	C2	C2	C3	C3

C1 = {2 ; 5 }, C2={8, 10, 11}, C3={18 ; 20}

Kmeans

R- estimation des centres de gravité :

$\mu 1 = (2+5)/2$

$\mu 2 = (8+10+11)/3$

$\mu 3 = (18+20)/2$

$\mu 1 = 3.5$

$\mu 2 = 9.66$

$\mu 3 = 19$

$\mu 1 = 3,5$, $\mu 2 = 9,66$, $\mu 3 = 19$

Ptx	**2**	**5**	**8**	10	11	18	20
d(μ1)	1,5	1,5	4,5	6,5	7,5	14,5	16,5
d(μ2)	7,66	4,66	1,66	0,33	1,33	8,34	10,34
d(μ3)	17	14	11	9	8	1	1
Classe	C1	C1	C2	C2	C2	C3	C3

C1 = {2 ; 5 }, C2={8, 10, 11}, C3={18 ; 20}

Stabilité : Les centres de gravité n'ont pas changé. L'algorithme s'arrête

Kmeans

- Utilisez l'algorithme k-means et la distance euclidienne pour regrouper les 8 exemples suivants en 3 clusters :

$A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9).$

- On considère comme centre de classes à l'initialisation les points A1, B1 et C1.

Kmeans

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to						Cluster	New Cluster
			2	10	5	8	1	2		
A1	2	10								
A2	2	5								
A3	8	4								
B1	5	8								
B2	7	5								
B3	6	4								
C1	1	2								
C2	4	9								

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Kmeans

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to 2	Distance to 10	Distance to 5	Distance to 8	Distance to 1	Distance to 2	Cluster	New Cluster
A1	2	10	0.00		3.61		8.06		1	
A2	2	5	5.00		4.24		3.16		3	
A3	8	4	8.49		5.00		7.28		2	
B1	5	8	3.61		0.00		7.21		2	
B2	7	5	7.07		3.61		6.71		2	
B3	6	4	7.21		4.12		5.39		2	
C1	1	2	8.06		7.21		0.00		3	
C2	4	9	2.24		1.41		7.62		2	

New Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to 2	Distance to 10	Distance to 6	Distance to 6	Distance to 1.5	Distance to 1.5	Cluster	New Cluster
A1	2	10	0.00		5.66		6.52		1	1
A2	2	5	5.00		4.12		1.58		3	3
A3	8	4	8.49		2.83		6.52		2	2
B1	5	8	3.61		2.24		5.70		2	2
B2	7	5	7.07		1.41		5.70		2	2
B3	6	4	7.21		2.00		4.53		2	2
C1	1	2	8.06		6.40		1.58		3	3
C2	4	9	2.24		3.61		6.04		2	1

New Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Kmeans

Initial Centroids:
A1: (2, 10)
B1: (5, 8)
C1: (1, 2)

Data Points			Distance to 2, 10	Distance to 5, 8	Distance to 1, 2	Cluster	New Cluster
A1	2	10	0.00	3.61	8.06	1	
A2	2	5	5.00	4.24	3.16	3	
A3	8	4	8.49	5.00	7.28	2	
B1	5	8	3.61	0.00	7.21	2	
B2	7	5	7.07	3.61	6.71	2	
B3	6	4	7.21	4.12	5.39	2	
C1	1	2	8.06	7.21	0.00	3	
C2	4	9	2.24	1.41	7.62	2	

New Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Current Centroids:
A1: (2, 10)
B1: (6, 6)
C1: (1.5, 3.5)

Data Points			Distance to 2, 10	Distance to 6, 6	Distance to 1.5, 1.5	Cluster	New Cluster
A1	2	10	0.00	5.66	6.52	1	1
A2	2	5	5.00	4.12	1.58	3	3
A3	8	4	8.49	2.83	6.52	2	2
B1	5	8	3.61	2.24	5.70	2	2
B2	7	5	7.07	1.41	5.70	2	2
B3	6	4	7.21	2.00	4.53	2	2
C1	1	2	8.06	6.40	1.58	3	3
C2	4	9	2.24	3.61	6.04	2	1

New Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

Kmeans

Current Centroids:
A1: (3, 9.5)
B1: (6.5, 5.25)
C1: (1.5, 3.5)

New Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3	9.5	6.5	5.25	1.5	3.5		
A1	2	10	1.12		6.54		6.52		1	1
A2	2	5	4.61		4.51		1.58		3	3
A3	8	4	7.43		1.95		6.52		2	2
B1	5	8	2.50		3.13		5.70		2	1
B2	7	5	6.02		0.56		5.70		2	2
B3	6	4	6.26		1.35		4.53		2	2
C1	1	2	7.76		6.39		1.58		3	3
C2	4	9	1.12		4.51		6.04		1	1

Current Centroids:
A1: (3.67, 9)
B1: (7, 4.33)
C1: (1.5, 3.5)

Data Points			Distance to						Cluster	New Cluster
			3.67	9	7	4.33	1.5	3.5		
A1	2	10	1.94		7.56		6.52		1	1
A2	2	5	4.33		5.04		1.58		3	3
A3	8	4	6.62		1.05		6.52		2	2
B1	5	8	1.67		4.18		5.70		1	1
B2	7	5	5.21		0.67		5.70		2	2
B3	6	4	5.52		1.05		4.53		2	2
C1	1	2	7.49		6.44		1.58		3	3
C2	4	9	0.33		5.55		6.04		1	1

Conclusion

K-moyennes : Avantages

Relativement extensible dans le traitement d'ensembles de taille importante
Relativement efficace.

K-moyennes : Inconvénients

Besoin de spécifier k, le nombre de clusters, a priori
Incapable de traiter les données bruitées (noisy).