Chapitre 4a: Clustering

Abdelkrim El MOUATASIM
Professeur Habilité en Mathématique Appliquée
https://sites.google.com/a/uiz.ac.ma/elmouatasim/

FPO - SMI - S6

2018-2019

Plan

Organisation du cours

Organisation du cours

La problématique

- Regrouper les données en plusieurs groupes (=clusters) de manière à ce que chaque groupe soit homogène et se distingue des autres groupes.
- Contrairement à la classification où on dispose d'un ensemble d'apprentissage avec des classes connues, les clusters sont inconnus a priori.

Mesures de similarité et de distance

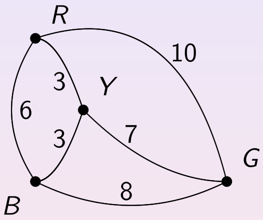
Soit $\mathbb{O}$ un ensemble d'objets. Une fonction $d : \mathbb{O} \times \mathbb{O} \longrightarrow \mathbb{R}$ définit une distance si elle satisfait les propriétés suivantes pour tout $x, y, z \in \mathbb{O}$:

$$
\begin{aligned}
& d(x, y) \geqslant 0 \\
& d(x, y) = 0 \ \text{ ssi } x = y \\
& d(x, y) = d(y, x) \\
& d(x, z) \leqslant d(x, y) + d(y, z)
\end{aligned}
$$

Exemple

$\mathbb{O} = \{R, B, G, Y\}$ et

d	R	B	G	Y
R	0	6	10	3
B	6	0	8	3
G	10	8	0	7
Y	3	3	7	0



Qu'est-ce que le centre d'un cluster ?

Soit $C \subseteq \mathbb{O}$. Qu'est-ce que le centre de C ?
Au moins deux définitions sont raisonnables :

1. Un objet m du cluster (i.e. $m \in C$) pour lequel $\sum_{x \in C} d(m, x)^2$ est minimal (on appelle m aussi medoïde ou médiane).
2. Un objet $c \in \mathbb{O}$, pas nécessairement dans C, pour lequel $\sum_{x \in C} d(c, x)^2$ est minimal (on appelle c aussi centroïde ou moyenne).

Exemple

Soit $C = \{R, B, G\}$.

$$\begin{aligned}
d(R,R)^2 + d(R,B)^2 + d(R,G)^2 &= 136 \\
d(B,R)^2 + d(B,B)^2 + d(B,G)^2 &= 100 \\
d(G,R)^2 + d(G,B)^2 + d(G,G)^2 &= 164
\end{aligned}$$

B est donc l'objet le plus central de C. Notez néanmoins :

$$d(Y,R)^2 + d(Y,B)^2 + d(Y,G)^2 = 67$$

Exemples en $\mathbb{R}^n$

Soient $\vec{x} = (x_1, x_2, \ldots x_n)$ et $\vec{y} = (y_1, y_2, \ldots y_n)$ deux points en $\mathbb{R}^n$.

Distance euclidienne : $d_{Eucl}(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

Distance Manhattan ou "city block" :

$$d_{Manh}(\vec{x}, \vec{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

Distance de Minkowski : Soit $q \in \mathbb{N}$, $q > 0$.

$$d_{Mink(q)}(\vec{x}, \vec{y}) = \sqrt[q]{\sum_{i=1}^{n} |x_i - y_i|^q}$$

Notez : $d_{Eucl}(\vec{x}, \vec{y}) = d_{Mink(2)}(\vec{x}, \vec{y})$ et $d_{Manh}(\vec{x}, \vec{y}) = d_{Mink(1)}(\vec{x}, \vec{y})$

Qu'est-ce qu'un cluster ?

Partitionner un ensemble $S \subseteq \mathbb{O}$ en plusieurs clusters.
Plusieurs charactérisations du concept "cluster" sont raisonnables. Par ex.

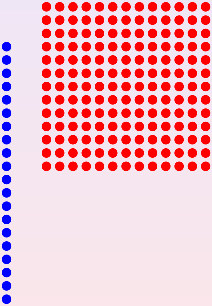
Centrisme Chaque objet est plus proche du centre de son propre cluster que de tout autre centre. Il suffit donc de spécifier les centres pour connaître les clusters.

Séparatisme Chaque objet est plus proche de tout objet de son propre cluster que de n'importe quel objet d'un autre cluster.

Atteignabilité Chaque objet appartient au même cluster que son voisin le plus proche.

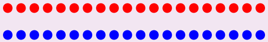
Exemples en $\mathbb{R}^2$

Le clustering suivant satisfait "atteignabilité" mais pas "séparatisme", ni "centrisme".



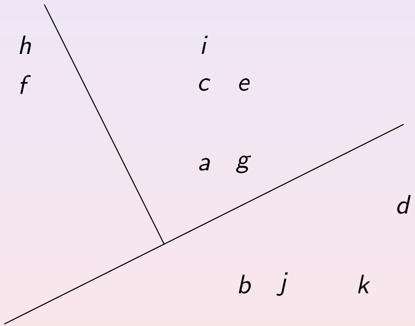
Exemples en $\mathbb{R}^2$

Le clustering suivant satisfait “centrisme” et “atteignabilité” mais pas “séparatisme”.

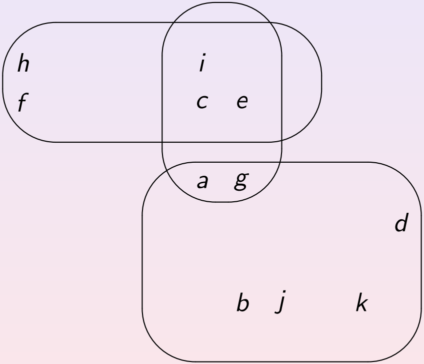


Le clustering suivant satisfait “centrisme” mais pas “atteignabilité”.

Les clusters disjoints



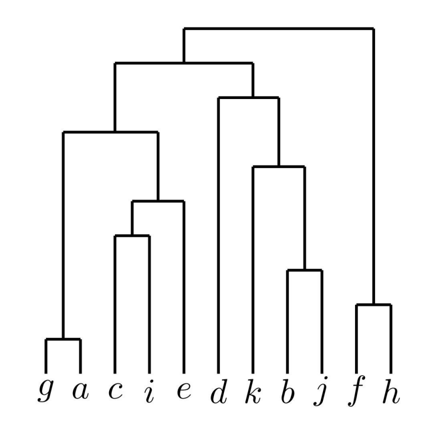
Les clusters pas forcément disjoints



Les clusters probabilistes

	1	2	3
a	0.3	0.4	0.3
b	0.1	0.2	0.7
c	0.4	0.5	0.1
$\vdots$		$\vdots$	

Dendrogram : une hiérarchie de clusters



Organisation du cours

Le clustering vu comme un problème d'optimisation

On souhaite partitionner un ensemble S en $k \geqslant 2$ clusters.
Soient $C_1, C_2, \ldots, C_k$ des clusters avec centres $c_1, c_2, \ldots, c_k$ respectivement. Définissons la dispersion intra-cluster comme :

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} d(c_i, x)^2$$

(SSE : Sum of the Squared Error)
Le but est de trouver un clustering avec une dispersion intra-cluster minimale.

Principe de kMeans clustering

Partitionner un ensemble S en k clusters.

1. Choisir les moyennes $m_1, m_2, \ldots, m_k$.
2. Attribuer tout objet de S à la moyenne la plus proche. Soient $C_1, C_2, \ldots, C_k$ les ensembles d'objets attribués respectivement à $m_1, m_2, \ldots, m_k$.
3. Ajuster les moyennes :

$$\begin{aligned} m_1 &:= \text{la moyenne de } C_1 \\ m_2 &:= \text{la moyenne de } C_2 \\ &\ldots \\ m_k &:= \text{la moyenne de } C_k \end{aligned}$$

4. Goto 2.

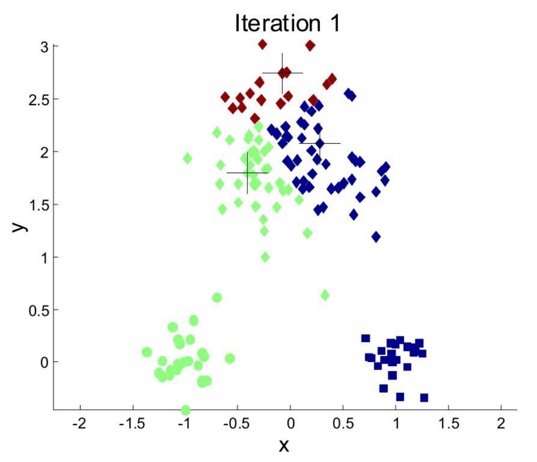
kMeans : algorithme

Algorithm 8.1 Basic K-means algorithm.

1: Select K points as initial centroids.
2: **repeat**
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: **until** Centroids do not change.

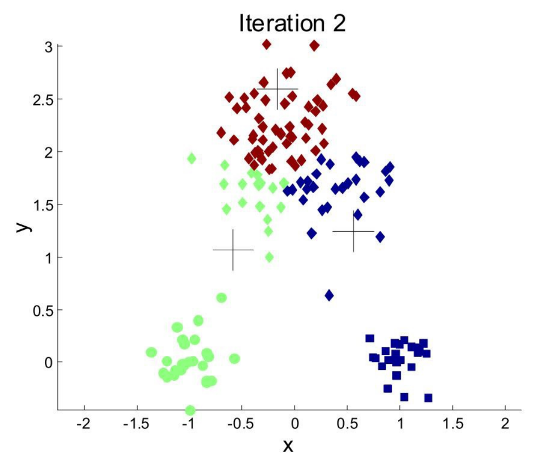
Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

kMeans : Example



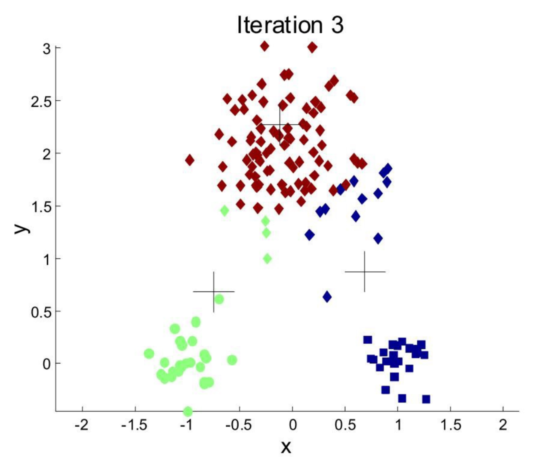
Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

kMeans : Example



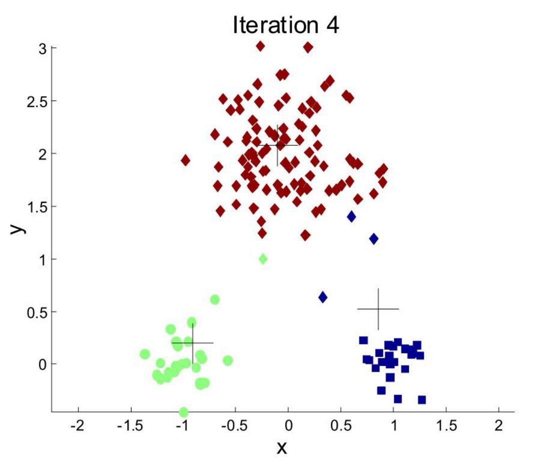
Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

kMeans : Example



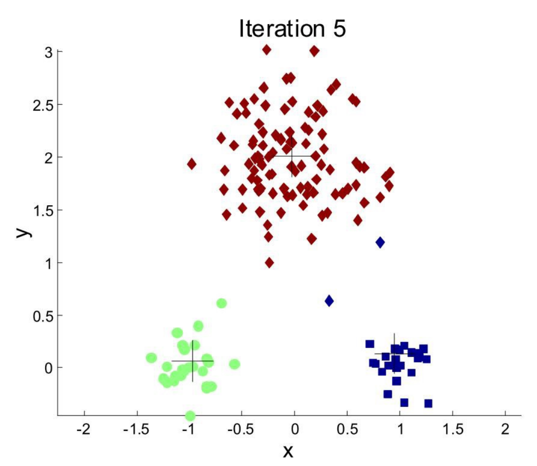
Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

kMeans : Example



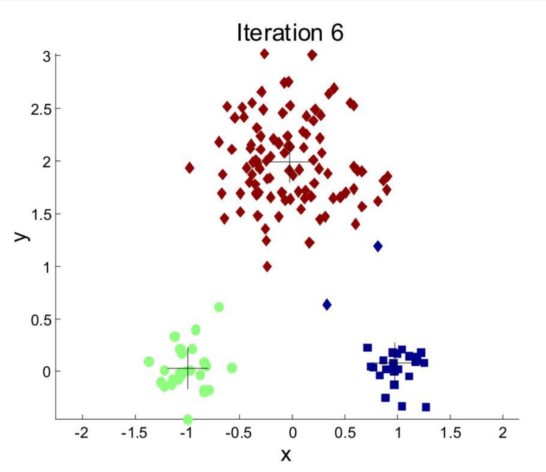
Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

kMeans : Example



Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

kMeans : Example



Source: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar: *Introduction to Data Mining*. Addison Wesley, 2006

Discussion

- Comment déterminer k ?
- kMeans garantit “centrisme” (voir transparent 11), mais pas “atteignabilité”.

Résultat si on démarre avec a, b :

Une meilleure solution est $\{\{a, b\}, \{c, d\}\}$.

kMedoids clustering

Partitionner un ensemble S en k clusters.

1. Choisir les medoïdes $m_1, m_2, \ldots, m_k \in S$.
2. Chercher $m_j \in \{m_1, m_2, \ldots, m_k\}$ et $p \in S \backslash \{m_1, m_2, \ldots, m_k\}$ tel que remplacer m_j par p améliore le clustering.
3. Goto 2.

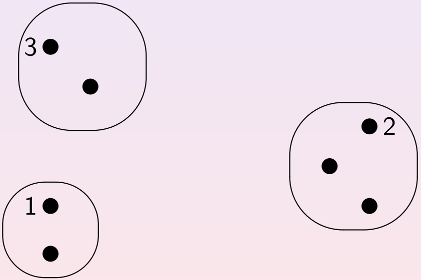
Farthest First

Pour un objet o et en ensemble S d'objets, définissons $d(o, S) := \min\{d(o, p) \mid p \in S\}$.

- Choisir un point m_1.
- Choisir pour m_2 le point le plus éloigné de m_1.
- Choisir pour m_3 le point le plus éloigné de $\{m_1, m_2\}$.
- Choisir pour m_4 le point le plus éloigné de $\{m_1, m_2, m_3\}$.
- ...
- Choisir pour m_k le point le plus éloigné de $\{m_1, m_2, \ldots, m_{k-1}\}$.

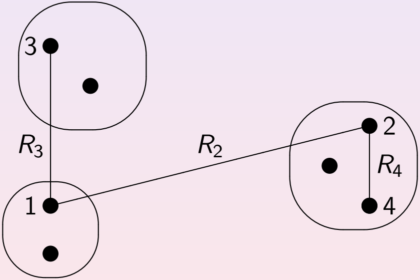
Farthest First $k = 3$

- Le premier point est choisi au hasard.

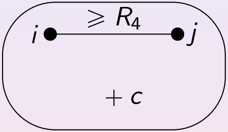


Farthest First $k = 3$: garantie de performance

- Le rayon de chaque cluster est $\leqslant R_4$.



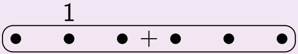
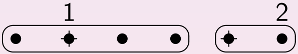
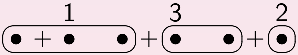
- Tout 3-clustering contiendra un cluster regroupant deux points parmi $\{1, 2, 3, 4\}$ (pigeon hole principle). Soit c le centre de ce cluster.
- $i, j \in \{1, 2, 3, 4\}$ et $i \neq j$.



Soit r le rayon de ce cluster. Évidemment, $r \geqslant d(c, i)$ et $r \geqslant d(c, j)$. Puisque $d(i, c) + d(c, j) \geqslant d(i, j) \geqslant R_4$, $\boxed{r \geqslant R_4/2}$.

- Tout 3-clustering contiendra donc un cluster avec rayon $\geqslant R_4/2$.
- Les clusters trouvés pas Farthest First ont tous un rayon qui est au pire deux fois ce rayon minimal.

Exemple

	Rayon maximal Farthest First	Rayon maximal Optimal(+)
	4	2.5
	2	1
	1	0.5

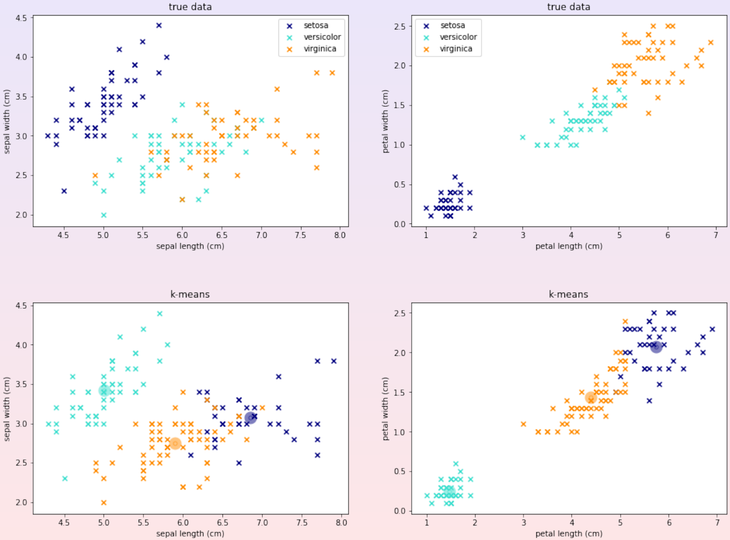
Les algorithmes de clustering

- *Clustering* : Se rapproche d'un problème de classification (sans labels)
- Ces algorithmes cherchent à rassembler les exemples en cluster
- À la différence des arbres de décision, le choix du cluster ne s'effectue pas par une suite de décisions simple, mais en déterminant la plus petite distance possible dans l'espace des variables
- Utilisés pour la segmentation d'utilisateurs/marchés dans le commerce en ligne mais aussi en génétique
- Il faut (dans la majorité des cas) définir au préalable le nombre de clusters à construire (hyperparamètre du modèle)

K-means

- Sépare les données en K clusters C d'égale variance (dispersion)
- Soit les *centroïdes*, la moyenne des échantillons dans chaque cluster
- *K-means* modifie la position des *centroïdes* jusqu'à trouver la valeur qui minimise l'écart moyens des échantillons vis-à-vis de son cluster correspondant
- Critère de minimisation : **inertie** : $I = \sum_{k=0}^{K} \sum_{x \in C_k} ||x - \mu_k||^2$
- La performance du *K-means* est fortement dépendante de son initialisation *(Solution : plusieurs initialisation → moyenne)*
- Le *K-means* ne fonctionne pas avec des variables catégorielles (il existe des adaptations). Il est préférable de normaliser les variables.

Clustering : Iris dataset, K-means result



Mean Shift

- Cherche les zones de fortes densité en modifiant itérativement la position de *centroïdes*
- Des *centroïdes* de rayons R sont aléatoirement initialisés
- Ils sont déplacés vers la région de plus haute densité (nombres de points dans le rayon R)
- On continue jusqu'à maximiser la densité de chaque *centroïde*
- Plusieurs *centroïdes* dans une zone : celui avec la plus haute densité est conservé
- l'ensemble du dataset est labelisé (plus petite distance)
- Le *Mean Shift* détermine le nombre optimal de clusters pour la valeur du rayon choisie (R)

Clustering : Iris dataset, Mean Shift result