Clustering

Gilles Gasso

INSA Rouen - Département ASI
Laboratoire LITIS

27 septembre 2016

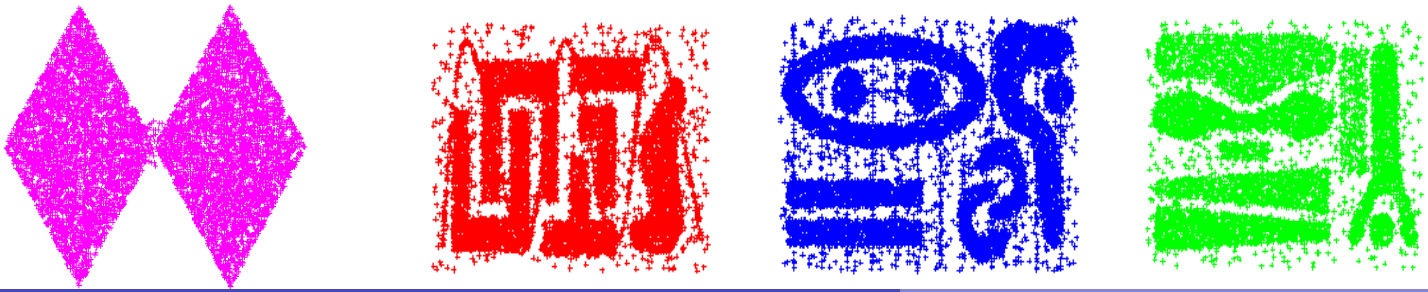
Plan

Introduction

Objectifs

- $\mathcal{D} = \{x_i \in \mathbb{R}^d\}_{i=1}^N$: ensemble de points décrits par d attributs.
- But : structuration des données en classes homogènes
  On cherche à regrouper les points en clusters ou classes tels que les données d'un cluster soient les plus similaires possibles
- Clustering $\equiv$ apprentissage non supervisé. C'est une technique d'exploration des données servant à résumer les informations sur les données ou à déterminer des liens entre les points.

Exemples de classes



Introduction

Domaines d'application

Domaine	Forme des données	Clusters
Text mining	Textes Mails	Textes proches Dossiers automatiques
Web mining	Textes et images	Pages web proches
BioInformatique	Gènes	Gènes ressemblants
Marketing	Infos clients, produits achetés	Segmentation de la clientèle
Segmentation d'images	Images	Zones homogènes dans l'image
Web log analysis	Clickstream	Profils utilisateurs

Problématiques

- Nature des observations : Données binaires, textuelles, numériques, arbres, ... ?
- Notion de similarité (ou de dissimilarité) entre observations
- Définition d'un cluster
- Evaluation de la validité d'un cluster
- Nombre de clusters pouvant être identifiés dans les données
- Quels algorithmes de clustering ?
- Comparaison de différents résultats de clustering

Proximité entre points

Mesure de la distance $D(x_1, x_2)$ entre 2 points x_1 et x_2

- Distance de Minkoswski : $D(x_1, x_2) = \left(\sum_{j=1}^{d} |x_{1,j} - x_{2,j}|^q \right)^{\frac{1}{q}}$
  - Distance Euclidienne correspond à $q = 2$:
    $D(x_1, x_2) = \sqrt{\sum_{j=1}^{d} (x_{1,j} - x_{2,j})^2} = \sqrt{(x_1 - x_2)^t (x_1 - x_2)}$
  - Distance de Manhattan ($q = 1$) : $D(x_1, x_2) = \sum_{j=1}^{d} |x_{1,j} - x_{2,j}|$
- Métrique liée à une matrice W définie positive :

$$D^2(x_1, x_2) = (x_1 - x_2)^\top W (x_1 - x_2)$$

  - Distance de Mahalanobis : $W = C^{-1}$ avec C=matrice de covariance des données

Proximité entre clusters

Mesure de la distance $D(x_1, x_2)$ entre 2 points x_1 et x_2 à valeurs discrètes

- Utiliser une matrice de contingence $A(x_1, x_2) = [a_{ij}]$
  - $x_1 = \begin{pmatrix} 0 & 1 & 2 & 1 & 2 & 1 \end{pmatrix}^\top$ et $x_2 = \begin{pmatrix} 1 & 0 & 2 & 1 & 0 & 1 \end{pmatrix}^\top$
  - $A(x_1, x_2) = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{pmatrix}$
- Distance de Hamming : nombre de places où les 2 vecteurs diffèrent :

$$D(x_1, x_2) = \sum_{i=1}^{d} \sum_{j=1, j \neq i}^{d} a_{ij}$$

  - Exemple : $D(x_1, x_2) = 3$

Notion de proximité (2)

Mesure de la distance $D(\mathcal{C}_1, \mathcal{C}_2)$ entre 2 classes $\mathcal{C}_1$ et $\mathcal{C}_2$

- plus proche voisin :

$$D_{\text{min}}(\mathcal{C}_1, \mathcal{C}_2) = \min \{D(x_i, x_j), x_i \in \mathcal{C}_1, x_j \in \mathcal{C}_2\}$$

- diamètre maximum :

$$D_{\text{max}}(\mathcal{C}_1, \mathcal{C}_2) = \max \{D(x_i, x_j), x_i \in \mathcal{C}_1, x_j \in \mathcal{C}_2\}$$

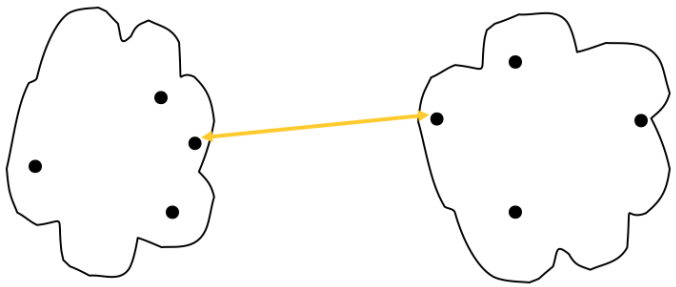
- distance moyenne :

$$D_{\text{moy}}(\mathcal{C}_1, \mathcal{C}_2) = \frac{\sum_{x_i \in \mathcal{C}_1} \sum_{x_j \in \mathcal{C}_2} D(x_i, x_j)}{n_1 n_2}$$

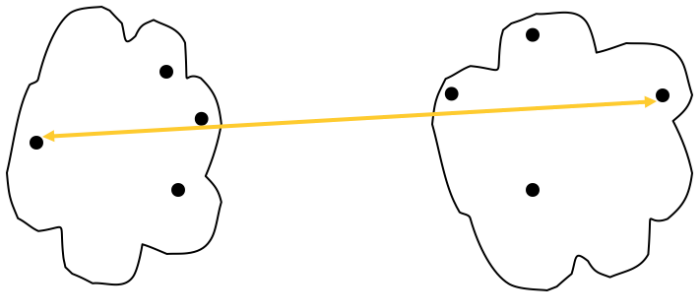
- distance de Ward : $D_{\text{Ward}}(\mathcal{C}_1, \mathcal{C}_2) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} D(\mu_1, \mu_2)$

Illustration

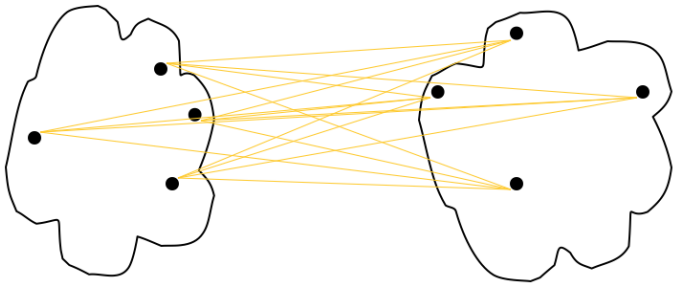
Distance min



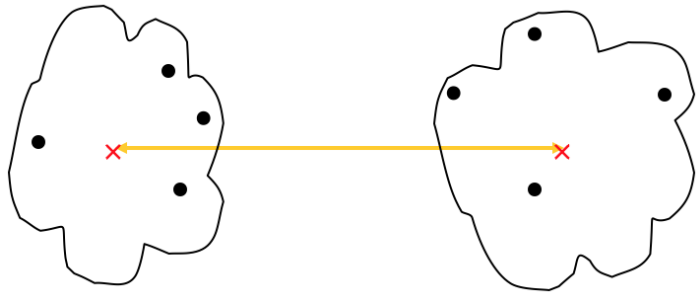
Diamètre maximum



Distance moyenne



Distance des centres de gravité



Qualité d'un clustering (2)

Caractéristiques d'un cluster

- Chaque cluster $\mathcal{C}_k$ est caractérisé par
  - Son centre de gravité : $\mu_k = \frac{1}{N_k} \sum_{i \in \mathcal{C}_k} x_i$ avec $N_k = \text{card}(\mathcal{C}_k)$
  - Son inertie : $J_k = \sum_{i \in \mathcal{C}_k} D^2(x_i, \mu_k)$

    L'inertie mesure la concentration des points autour de μ_k. Plus J_k est faible, plus petite est la dispersion des points autour de μ_k
  - Sa matrice de variance-covariance : $\Sigma_k = \sum_{i \in \mathcal{C}_k} (x_i - \mu_k)(x_i - \mu_k)^\top$

    Remarque : on a la propriété suivante $J_k = \text{trace}(\Sigma_k)$. L'inertie d'un cluster représente la variance des points de ce cluster

Inertie Intra-cluster

Inertie intra-cluster : $J_w = \sum_k \sum_{i \in \mathcal{C}_k} D^2(x_i, \mu_k) = \sum_{i \in \mathcal{C}_k} J_k$

Qualité d'un clustering (2)

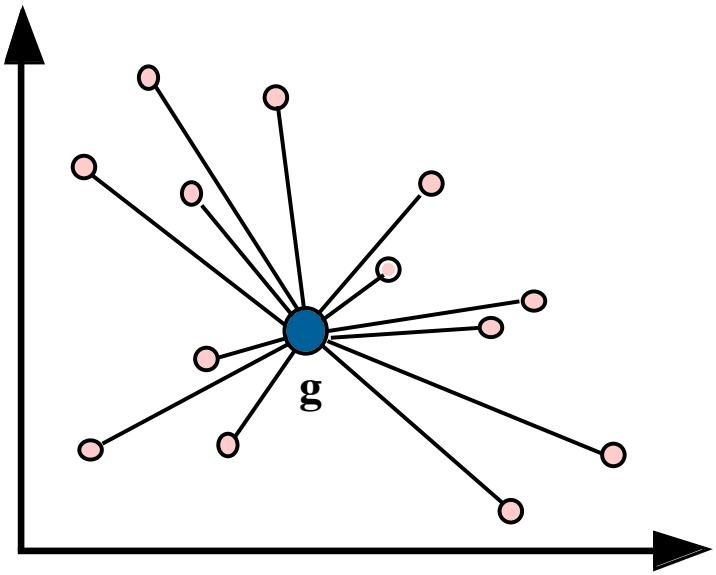
Inertie inter-cluster

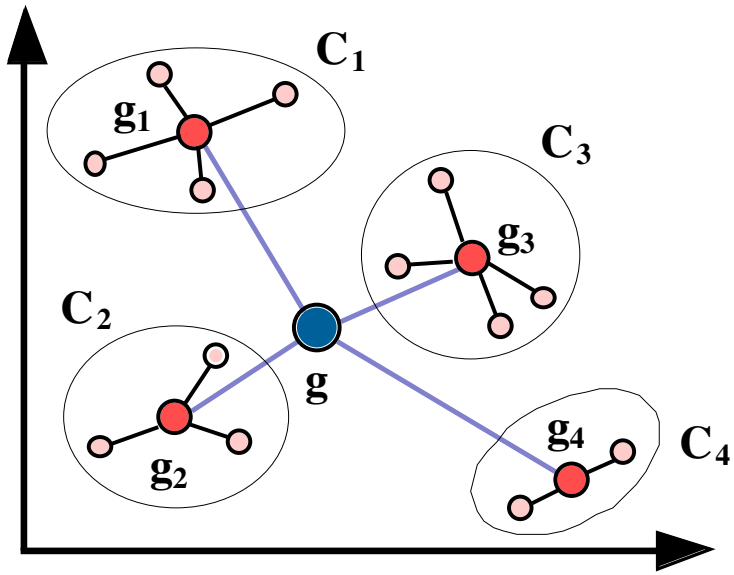
- Soit μ le centre de gravité du nuage de points : $\mu = \frac{1}{N} \sum_i x_i$
- Les centres de gravité des clusters forment eux aussi un ensemble de points caractérisé par
  - Inertie inter-cluster : $J_b = \sum_k N_k D^2(\mu_k, \mu)$

    L'inertie inter-cluster mesure "l'éloignement" des centres des clusters entre eux. Plus cette inertie est grande, plus les clusters sont bien séparés
  - Une matrice de covariance inter-cluster : $\Sigma_b = \sum_k (\mu_k - \mu)(\mu_k - \mu)^\top$
    Remarque : $J_b = \text{trace}(\Sigma_b)$

Bonne partition (1)

- Illustration (Bisson 2001) :





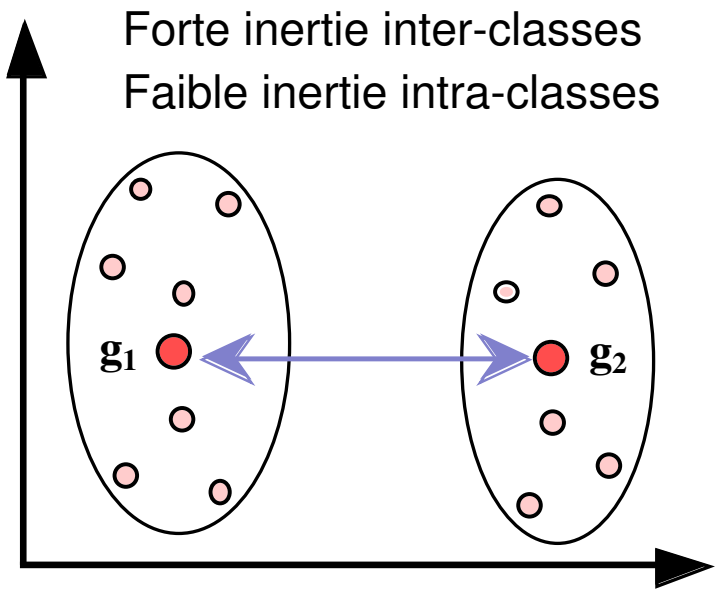
Inertie totale des points = Inertie Intra-cluster + Inertie Inter-cluster

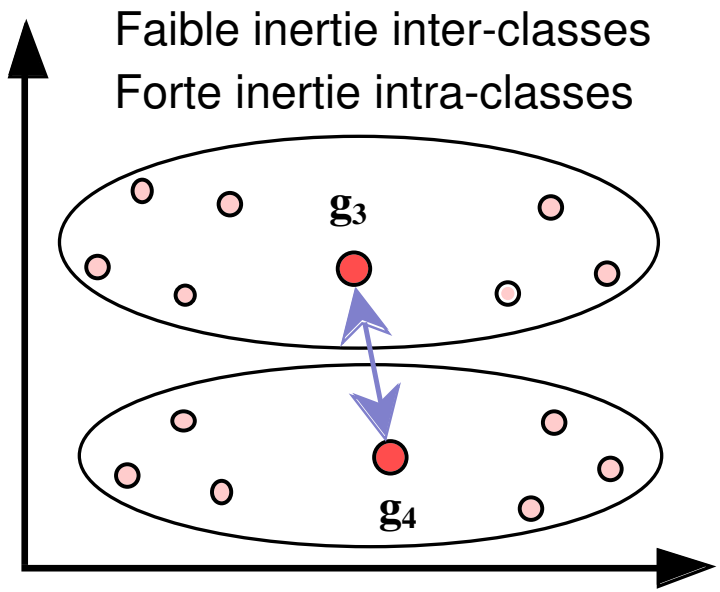
Comment obtenir une bonne partition ?

Il faut minimiser l'inertie intra-cluster et maximiser l'inertie inter-cluster

Bonne partition (2)

- Illustration (Bisson 2001) :





Approches de clustering

- Différentes approches sont possibles
- Clustering hiérachique
  - Clustering hiérarchique ascendant (CHA) et variantes
- Clustering par partitionnement
  - Algorithme des K-means
- Clustering par modélisation
  - Notion de modèles de mélange

CHA - principe

Principe

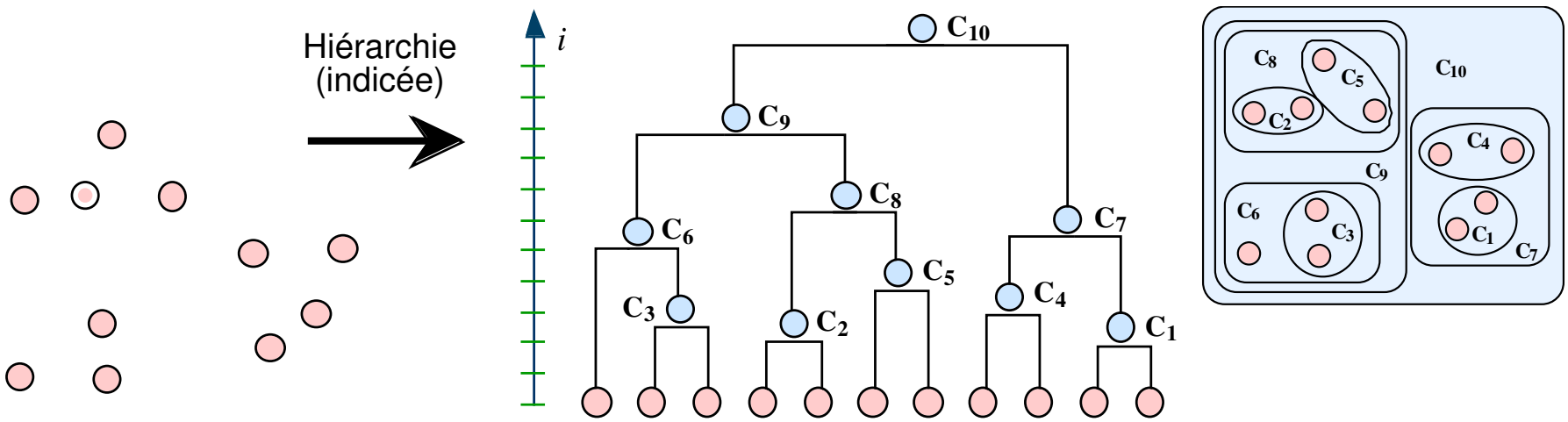
Chaque point ou cluster est progressivement "absorbé" par le cluster le plus proche.

Algorithme

- Initialisation :
  - Chaque point forme un cluster,
  - Calcul de la matrice de distance M entre chaque couple de clusters (ici les points)
- Répéter
  - Sélection dans M des deux clusters les plus proches $\mathcal{C}_I$ et $\mathcal{C}_J$
  - Fusion de $\mathcal{C}_I$ et $\mathcal{C}_J$ pour former un cluster $\mathcal{C}_G$
  - Mise à jour de M en calculant la distance entre C_G et les autres clusters
- Jusqu'à la fusion des 2 derniers clusters

CHA : principe

- Exemple (Bisson 2001)



- Schéma du milieu : dendrogramme = représentation des fusions successives
- Hauteur d'un cluster dans le dendrogramme = distance entre les 2 clusters avant fusion (sauf exception avec certaines mesures de similarité)...

CHA : métrique

Problème

Trouver l'ultramétrique (distance entre clusters) la plus proche de la métrique utilisée pour les points.

- Saut minimal (single linkage) basé sur la distance $D_{\text{min}}(\mathcal{C}_1, \mathcal{C}_2)$
  - tendance à produire des classes générales (par effet de chaînage)
  - sensibilité aux individus bruités.
- Saut maximal (complete linkage) basé sur la distance $D_{\text{max}}(\mathcal{C}_1, \mathcal{C}_2)$
  - tendance à produire des classes spécifiques (on ne regroupe que des classes très proches)
  - sensibilité aux individus bruités.

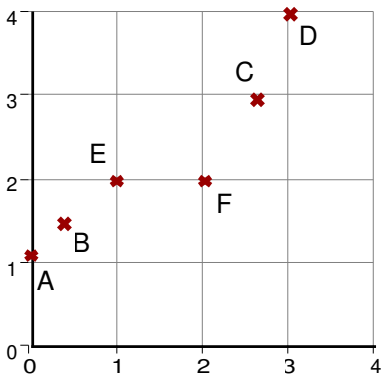
CHA : métrique

- Saut moyen basé sur la distance $D_{\text{moy}}(\mathcal{C}_1, \mathcal{C}_2)$
  - tendance à produire des classes de variance proche
- Barycentre basé sur la distance $D_{\text{Ward}}(\mathcal{C}_1, \mathcal{C}_2)$
  - bonne résistance au bruit

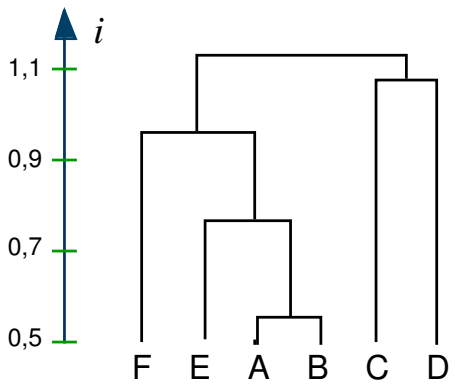
CHA : métrique

- Illustration : exemple (Bisson 2001)

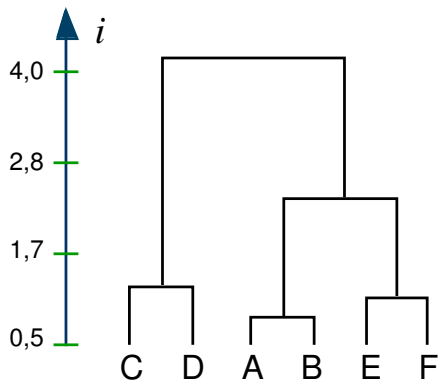
Données (métrique : dist. Eucl.)



Saut minimal



Saut maximal

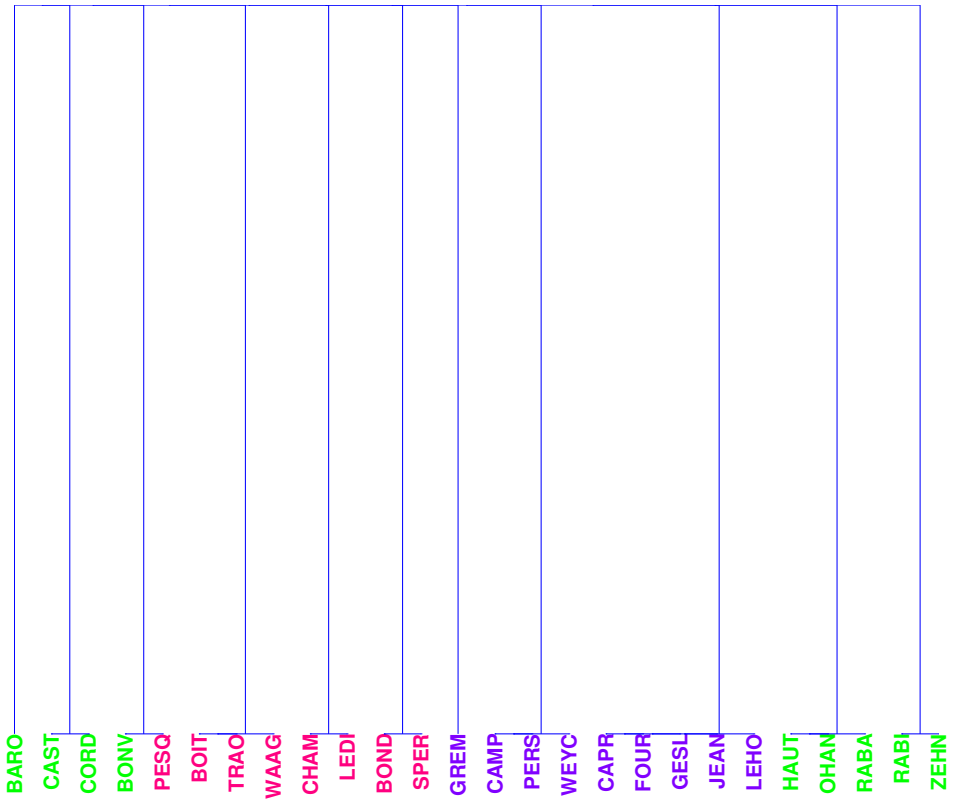


- Pas les mêmes résultats selon la métrique utilisée ...

CHA : ASI4 Clustering

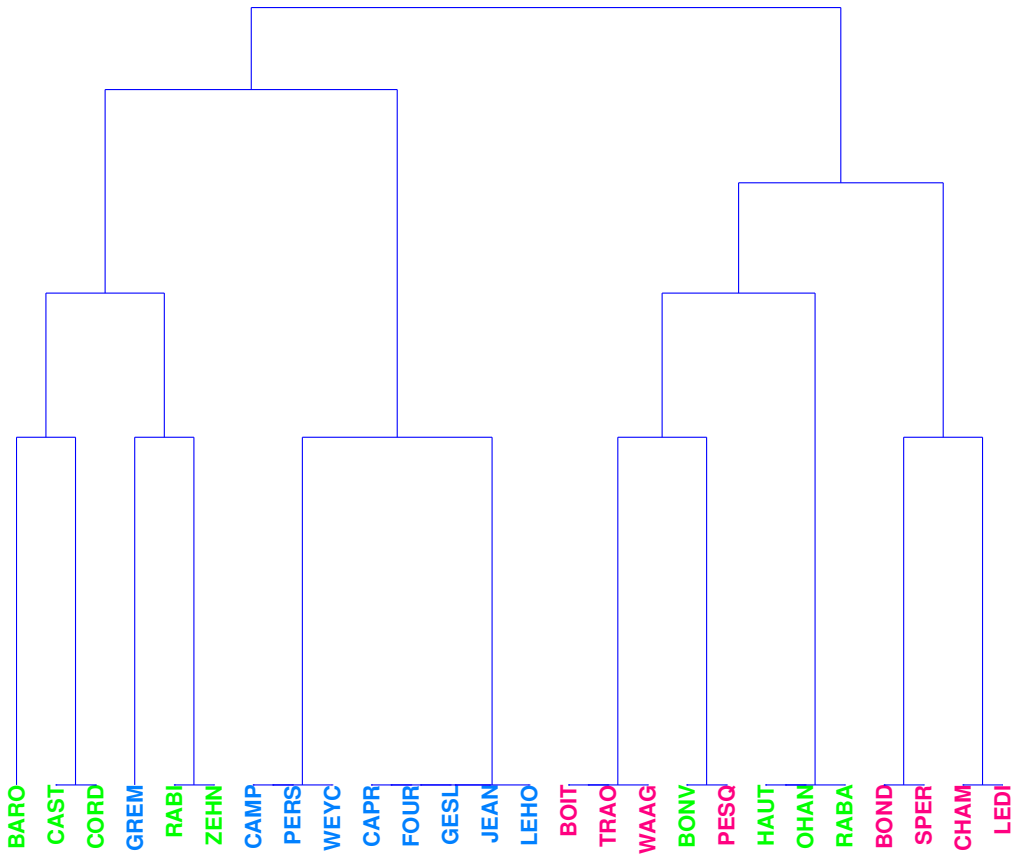
- 26 individus d'ASI4, 2003
- Données = 5 valeurs 0/1 (inscription en IR, MGPI, DM1, DM2, TIM)

Saut minimal



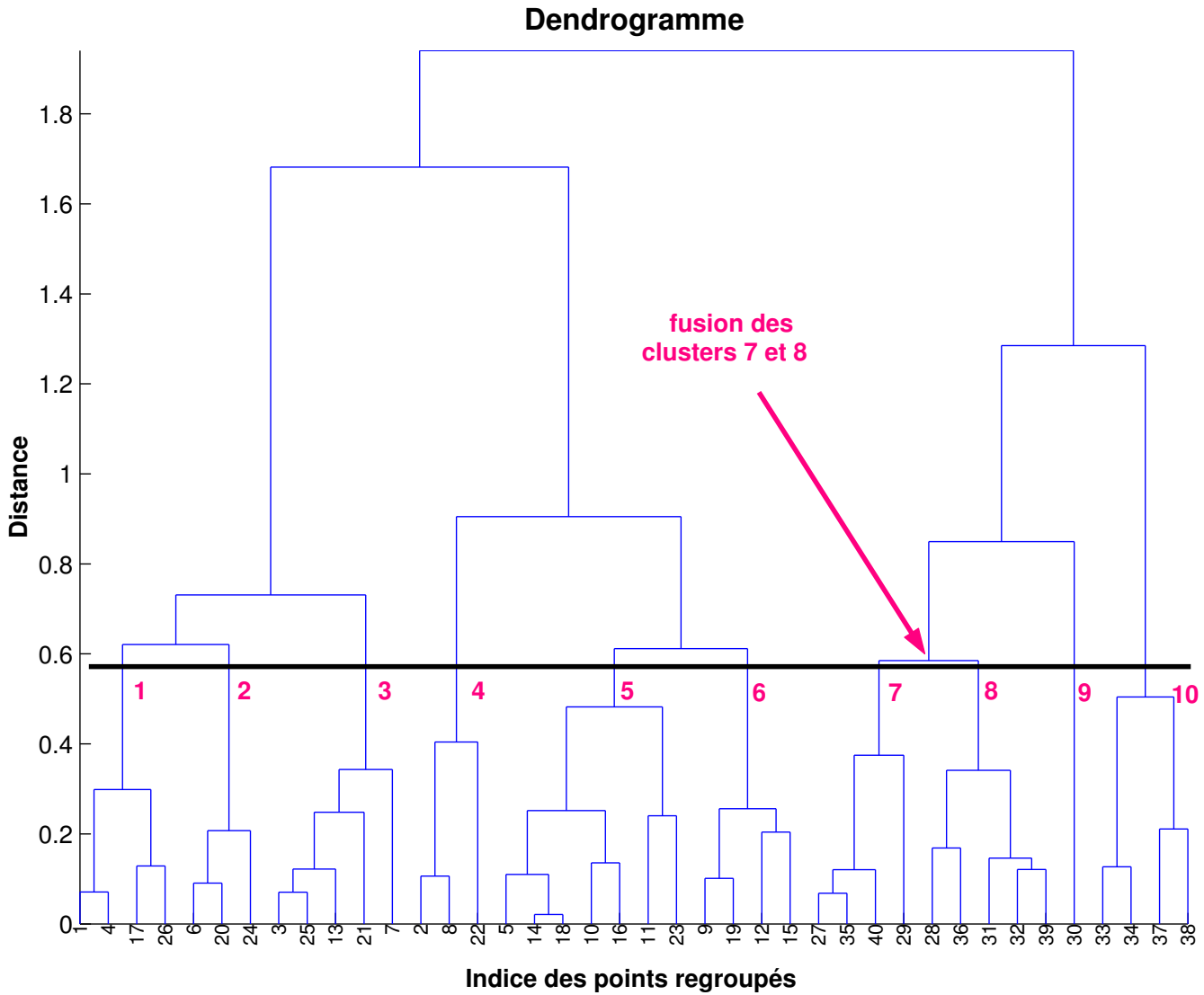
- Pas moyen de faire de sous-groupes, tout le monde est à une distance de 0 ou 1 des autres clusters.

CHA : clustering avec saut maximal données ASI4

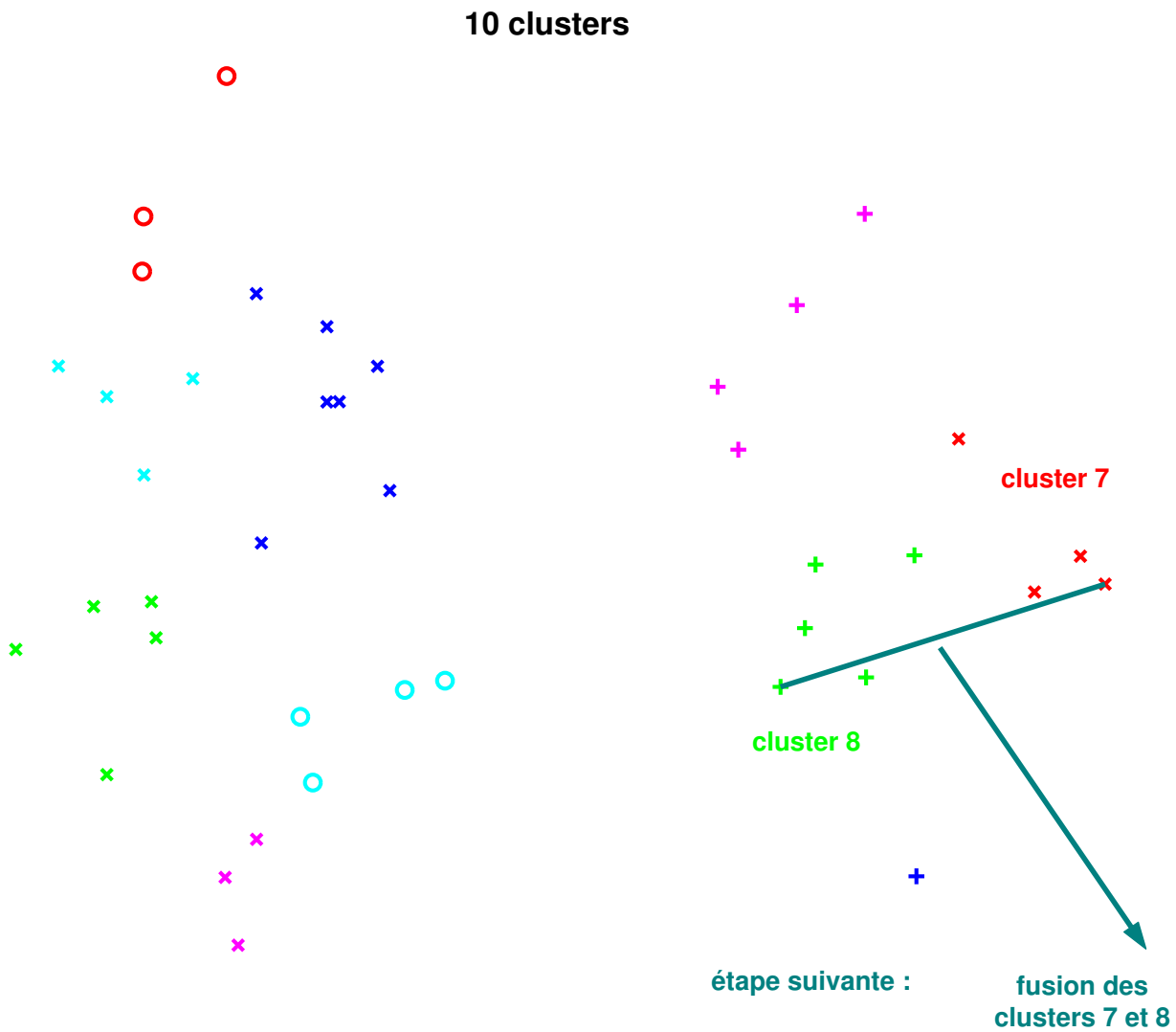


- En changeant de métrique, on observe plus de sous-regroupements
- Pour construire les clusters, on coupe l'arbre à la hauteur voulue

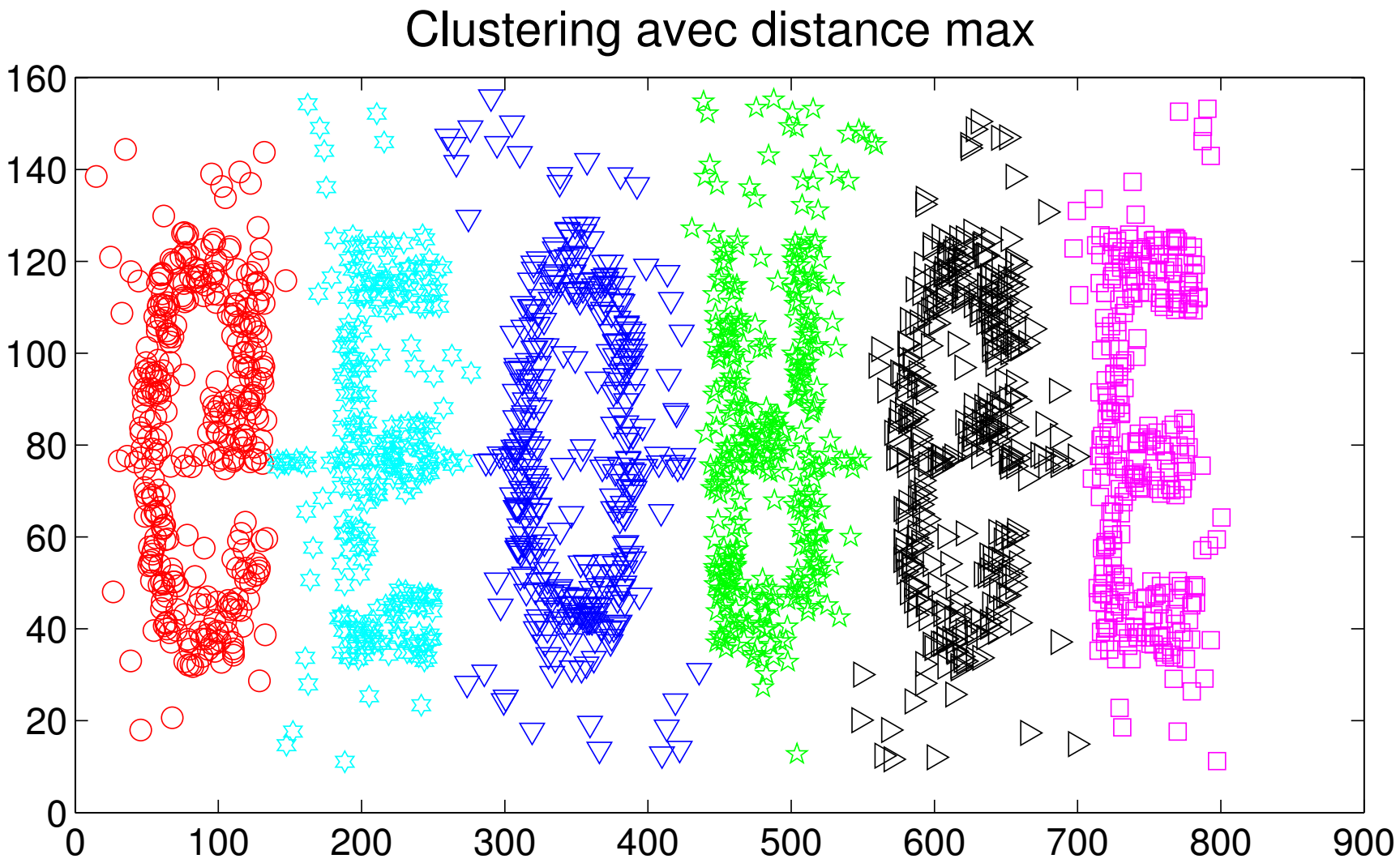
CHA : Autre exemple de saut maximal



CHA : Autre exemple de saut maximal

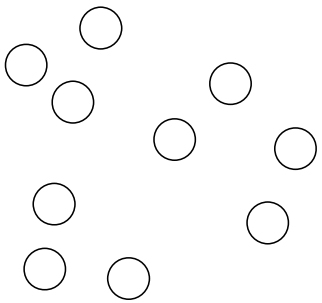


CHA : exemple sur les données george.dat

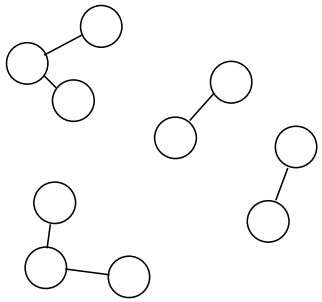


Extension du CHA : CHAMELEON

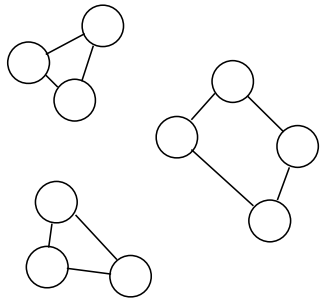
- Une méthode de classification hiérarchique ascendante plus évoluée
- Principe = estimer la densité intra-cluster et inter-cluster à partir du graphe des k plus proches voisins



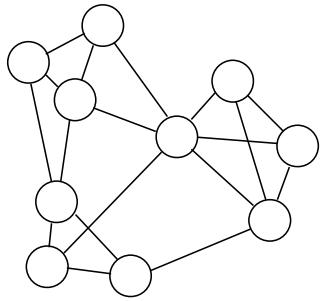
données



graphe 1-ppv



2-ppv

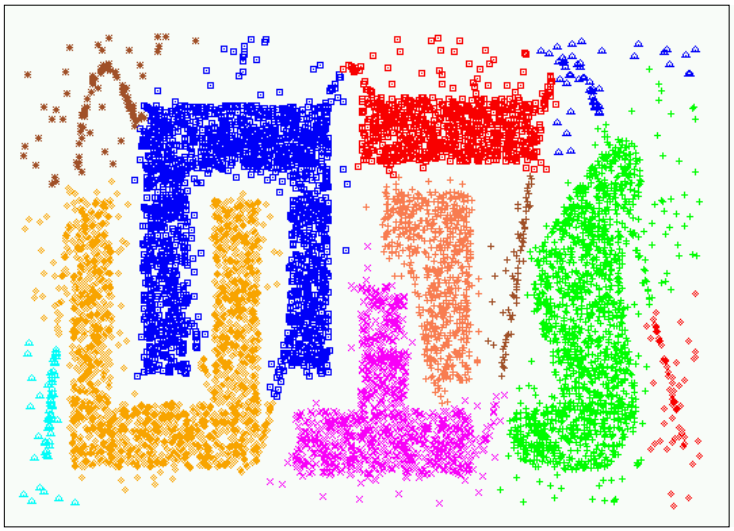


3-ppv

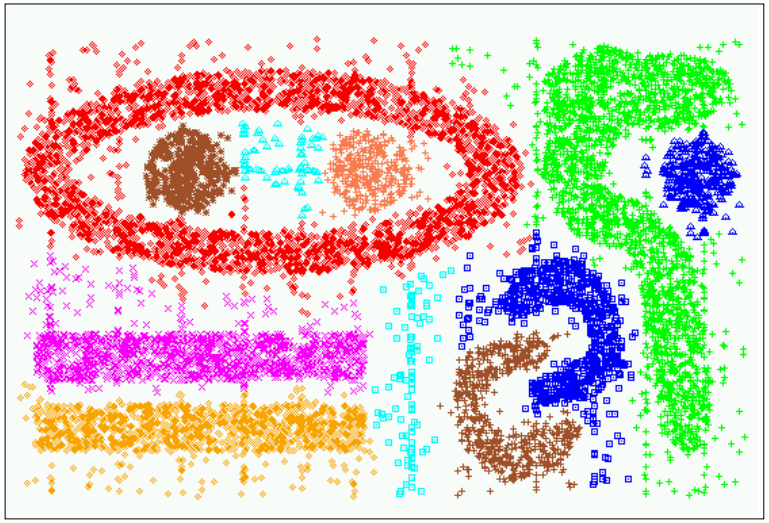
CHAMELEON : principe de l'algorithme

- Deux phases
  - Trouver des sous-clusters initiaux
    - en partitionnant le graphe k-ppv en *m* partitions "solides" (ou la distance entre les points est minimisée)
  - Fusionner dynamiquement les sous-clusters
    - en fonction de la densité inter-cluster et de la densité intra-cluster

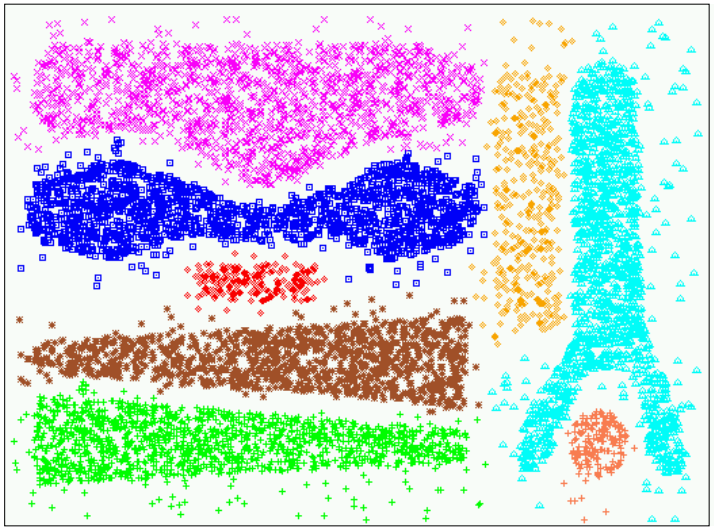
CHAMELEON : résultats



DS3



DS4



DS5

Approches de clustering

- Différentes approches sont possibles
- Clustering hiérachique
  - Clustering hiérarchique ascendant (CHA) et variantes
- Clustering par partitionnement
  - Algorithme des K-means
- Clustering par modélisation
  - Notion de modèles de mélange

Clustering par partitionnement

Objectifs

- N données $\mathcal{D} = \{x_i \in \mathbb{R}^d\}_{i=1,\cdots,N}$ disponibles
- Recherche d'une partition en $K < N$ clusters $\mathcal{C}_k$ des données

Approche directe

- Construire toutes les partitions possibles
- Evaluer la qualité de chaque clustering clustering et retenir la meilleure partition

Souci

Problème NP difficile car le nombre de partitions possibles augmente exponentiellement $\#\text{Clusters} = \frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}C_k^K k^N$. Pour $N = 10$ et $K = 4$, on a 34105 partitions possibles !

Clustering par partitionnement

Solution plus pratique

- Minimisation de l'inertie intra-classe

$$J_w = \sum_{k=1}^{K} \sum_{i \in C_k} D^2(x_i, \mu_k)$$

- Eviter l'énumération exhaustive de toutes les partitions possibles
- Utilisation d'une approche heuristique donnant au moins une bonne partition et pas nécessairement la partition optimale au sens de J_w

Clustering par partitionnement

Un algorithme très connu : algorithmes des K-means (ou centres mobiles)

- Si on connait les centres de gravité $\mu_k, k = 1, \cdots, K$, on affecte un point x_i à un cluster et un seul. Le point est affecté au cluster $\mathcal{C}_\ell$ dont le centre μ_ℓ est le plus proche
- Connaissant les clusters $\mathcal{C}_k, k = 1, \cdots, K$, on estime leur centre de gravité
- On alterne ainsi l'affectation des points aux clusters et l'estimation des centres de gravité jusqu'à convergence du critère

K-Means : algorithme

- Initialiser les centres $\mu_1, \cdots \mu_K$
- Répeter
  - Affectation de chaque point à son cluster le plus proche

$$\mathcal{C}_\ell \leftarrow x_i \quad \text{tel que} \quad \ell = \arg \min_k D(x_i, \mu_k)$$

  - Calculer le critère $J_w = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}_k} D^2(x_i, \mu_k)$
  - Recalculer le centre μ_k de chaque cluster

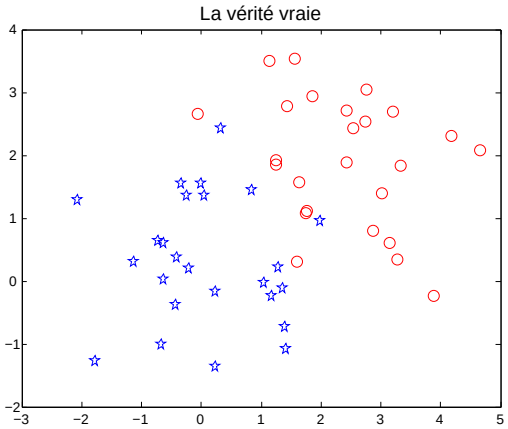
$$\mu_k = \frac{1}{N_k} \sum_{i \in \mathcal{C}_k} x_i \quad \text{avec} \quad N_k = \text{card}(\mathcal{C}_k)$$

- Tant que $\| \Delta\mu \| > \epsilon$ ou $\|J_w\| > \epsilon_2$

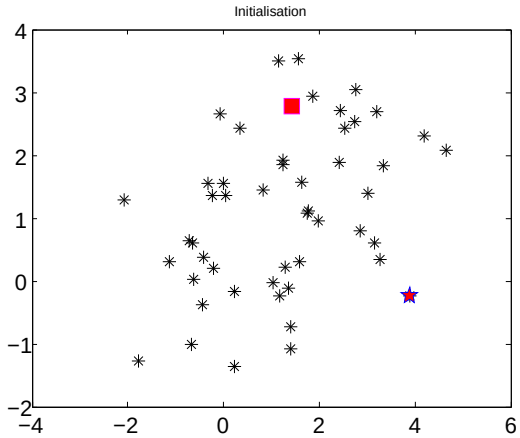
K-Means : illustration

Clustering en $K = 2$ classes

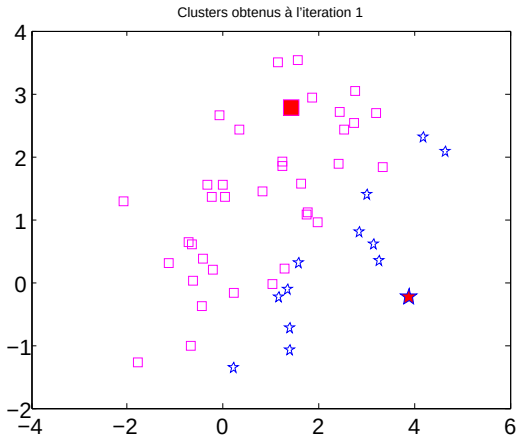
Données



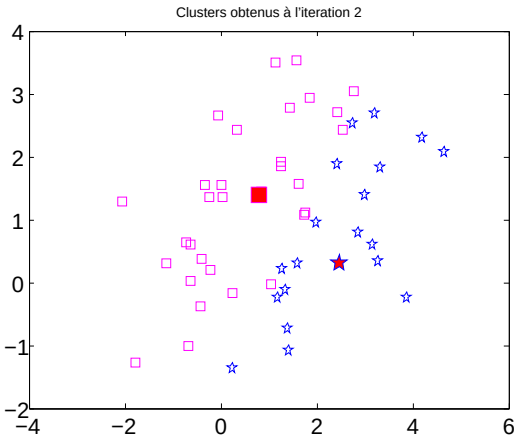
Initialisation



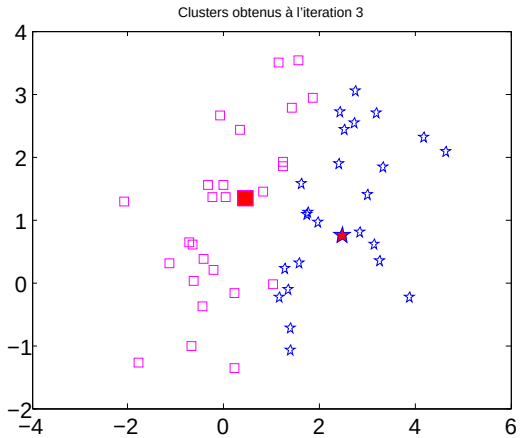
Itération 1



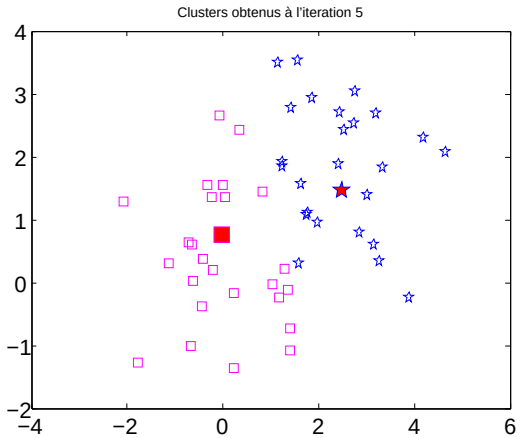
Itération 2



Itération 3



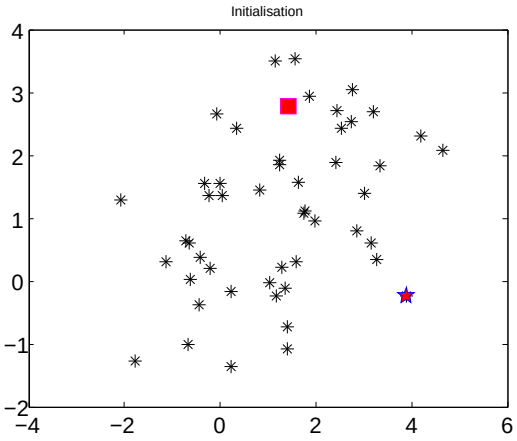
Itération 5



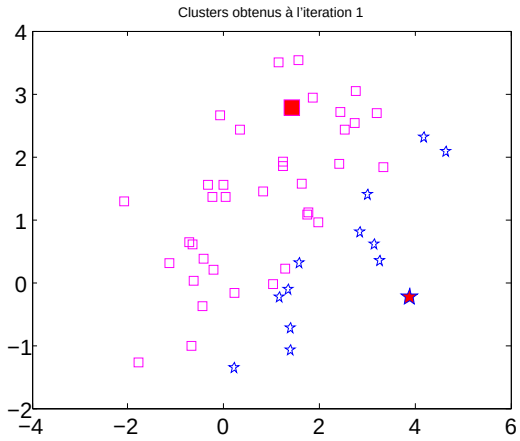
K-Means : illustration évolution critère J_w

Clustering en $K = 2$ classes

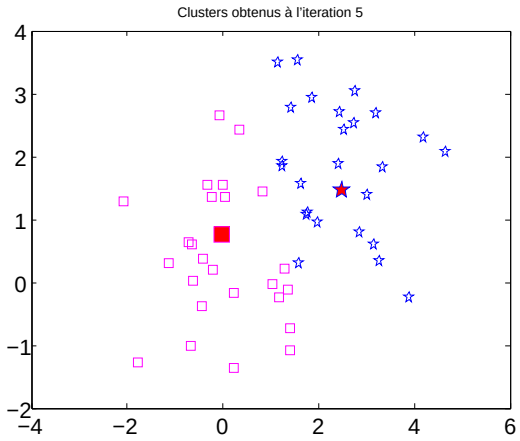
Initialisation

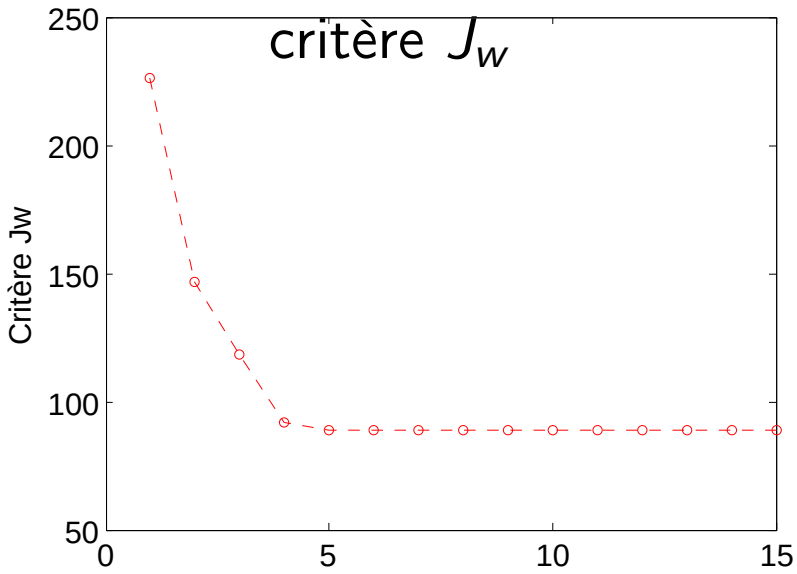


Itération 1



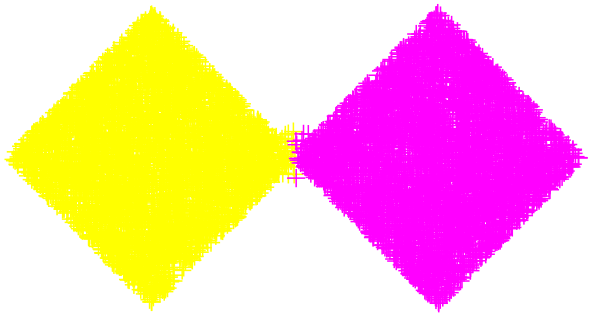
Itération 5



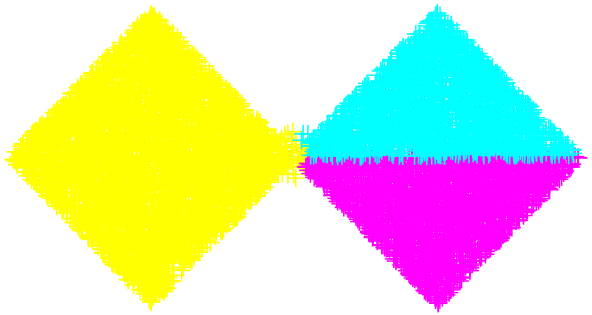


K-Means : exemple

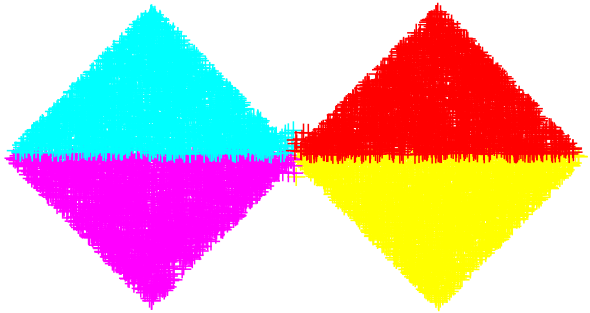
K = 2



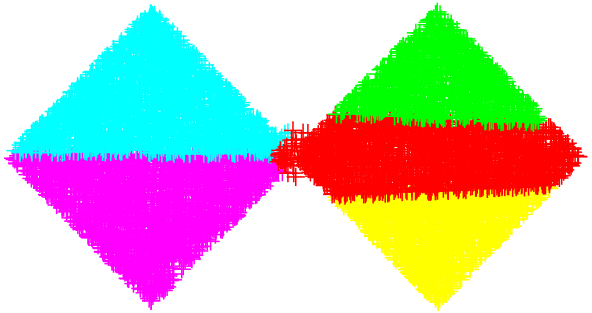
K = 3



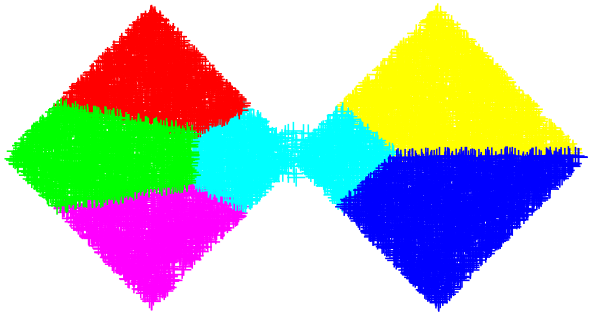
K = 4



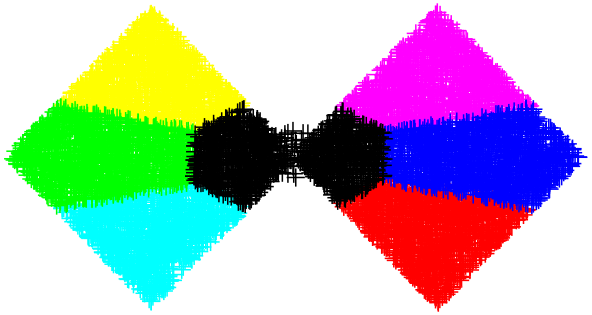
K = 5



K = 6

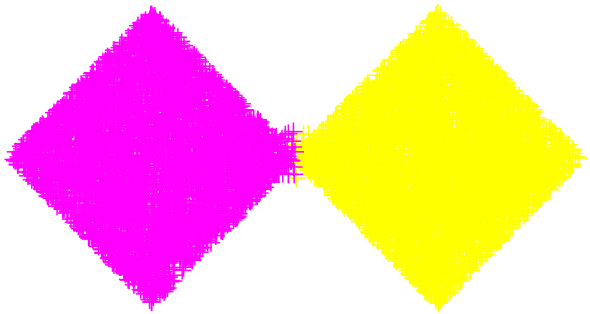


K = 7

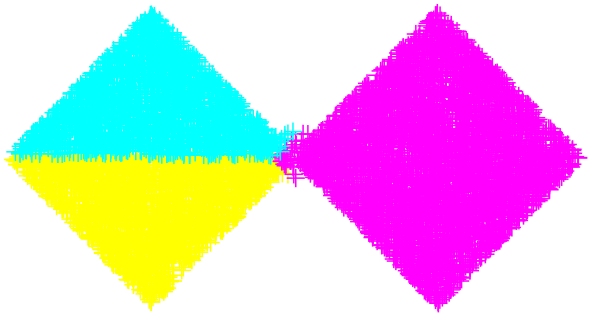


K-Means : exemple

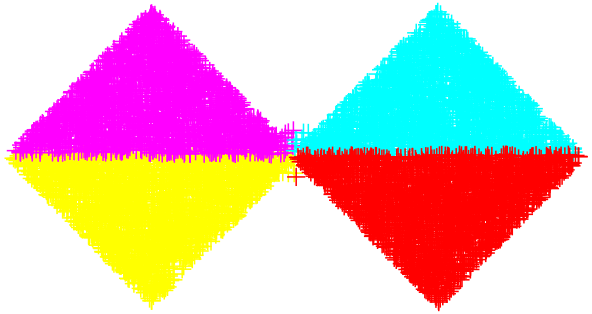
K = 2



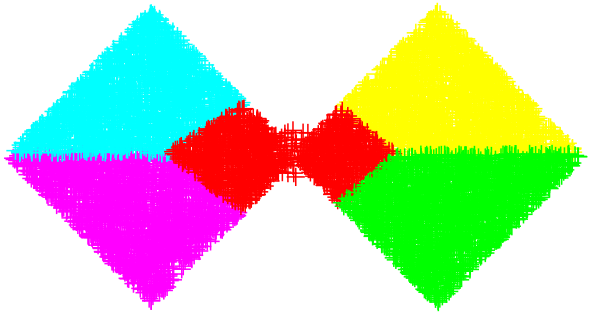
K = 3



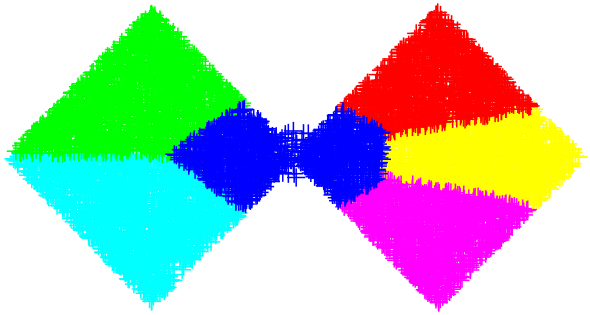
K = 4



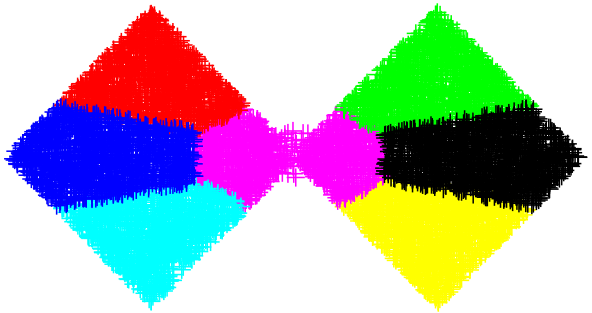
K = 5



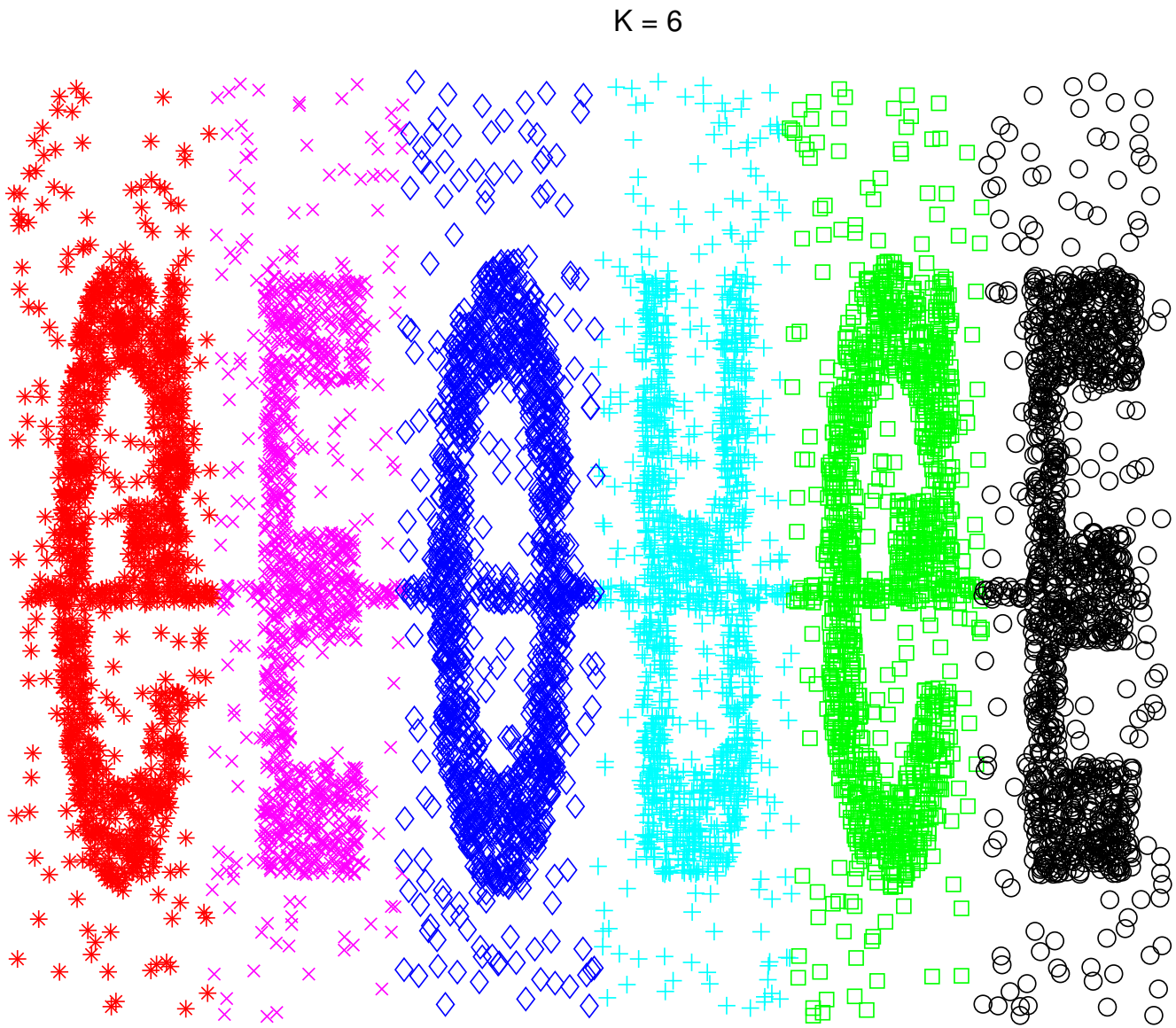
K = 6



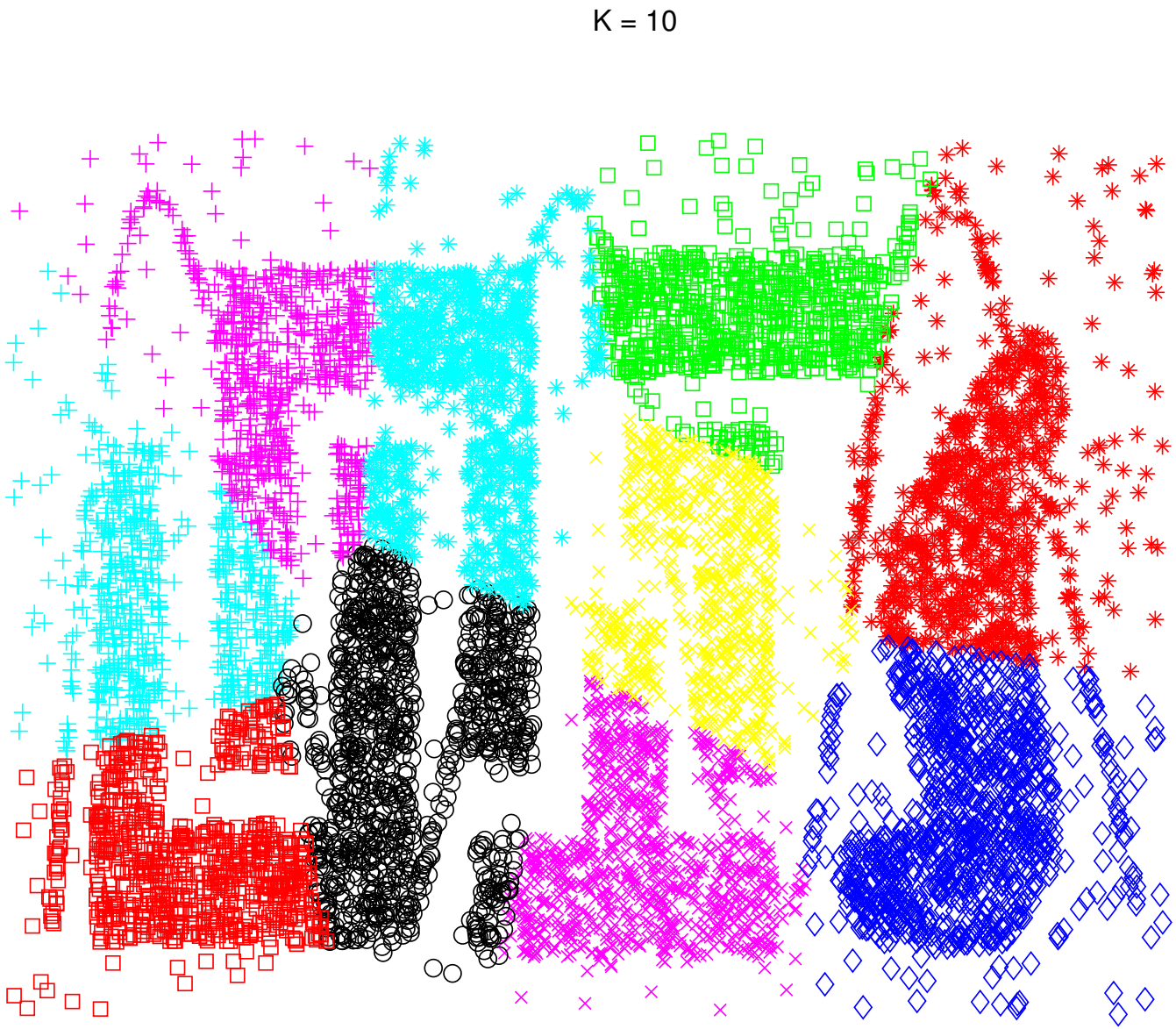
K = 7



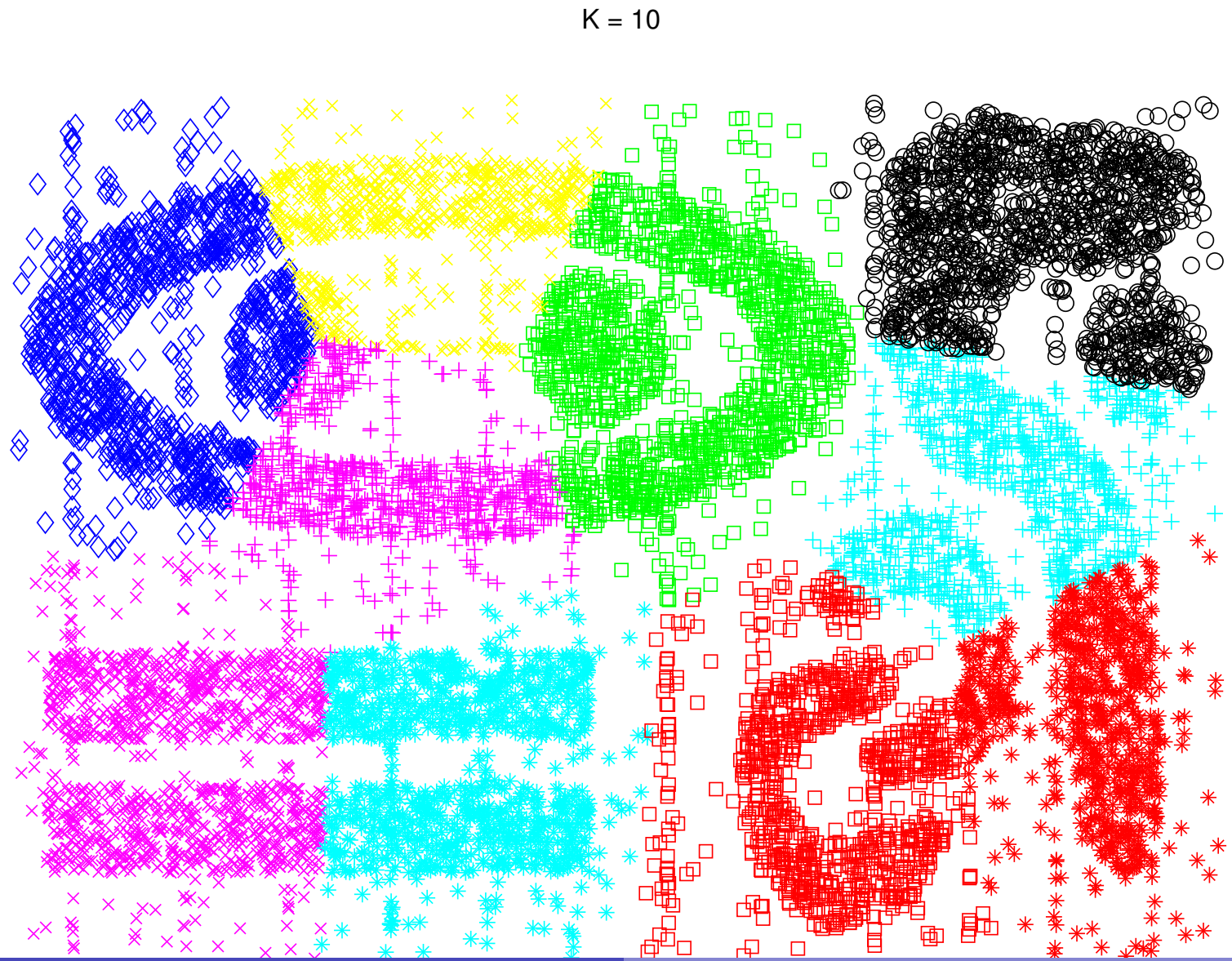
K-Means : exemple



K-Means : exemple



K-Means : exemple



K-Means : Remarques et problèmes

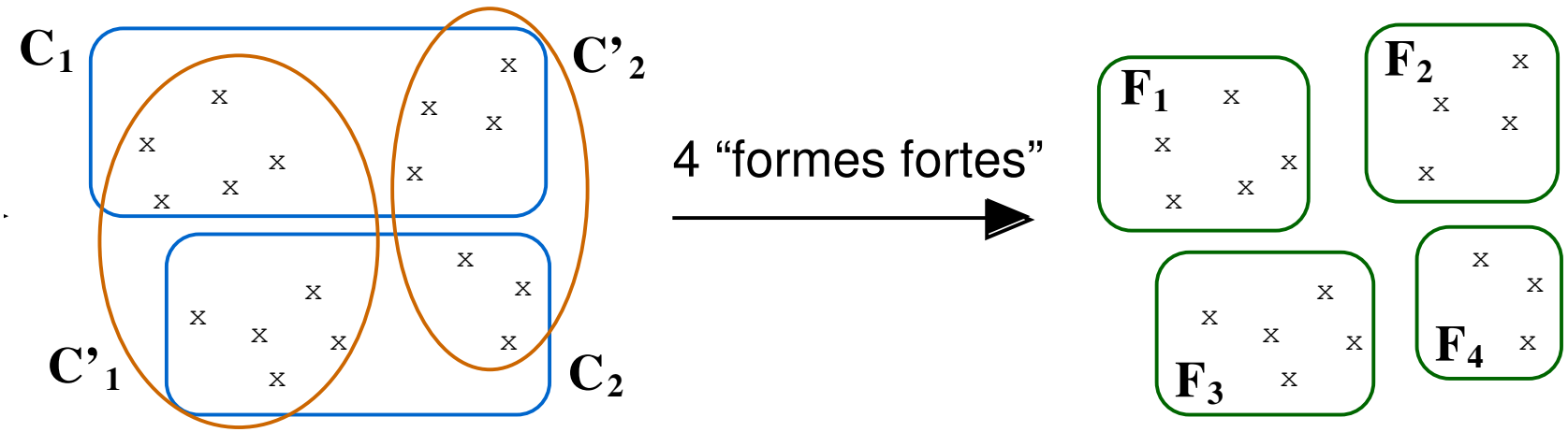
- On montre qu'à chaque itération de l'algorithme le critère J_w diminue.
- L'algorithme converge au moins vers un minimum local de J_w
- Convergence rapide
- Initialisation des μ_k :
  - aléatoirement dans l'intervalle de définition des x_i
  - aléatoirement dans l'ensemble des x_i
- Des initialisations différentes peuvent mener à des clusters différents (problèmes de minima locaux)

K-Means : Remarques et problèmes

- Méthode **générale** pour obtenir des clusters "stables" = formes fortes
  - On répète l'algorithme r fois
  - On regroupe ensemble les x_i qui se retrouvent toujours dans les mêmes clusters.
- Choix du nombre de clusters : problème difficile
  - Fixé a priori (exple : on veut découper une clientèle en K segments)
  - Chercher la meilleure partition pour différents $K > 1$ et chercher "un coude" au niveau de la décroissance de $J_w(K)$
  - Imposer des contraintes sur le volume ou la densité des clusters obtenus

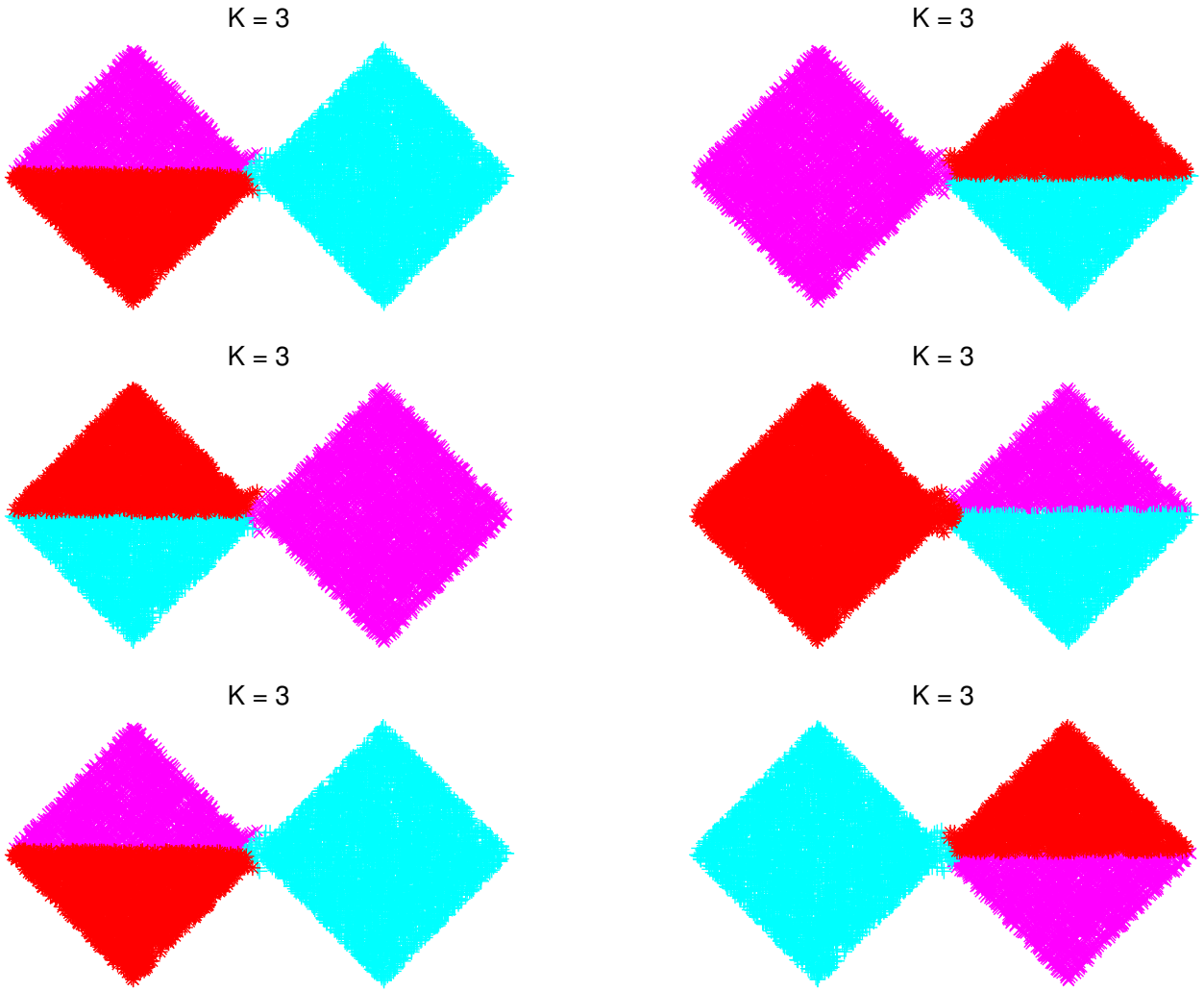
K-Means : formes fortes

- Illustration (Bisson 2001) :



K-Means : formes fortes

- K-Means répété 6 fois



K-Means : formes fortes

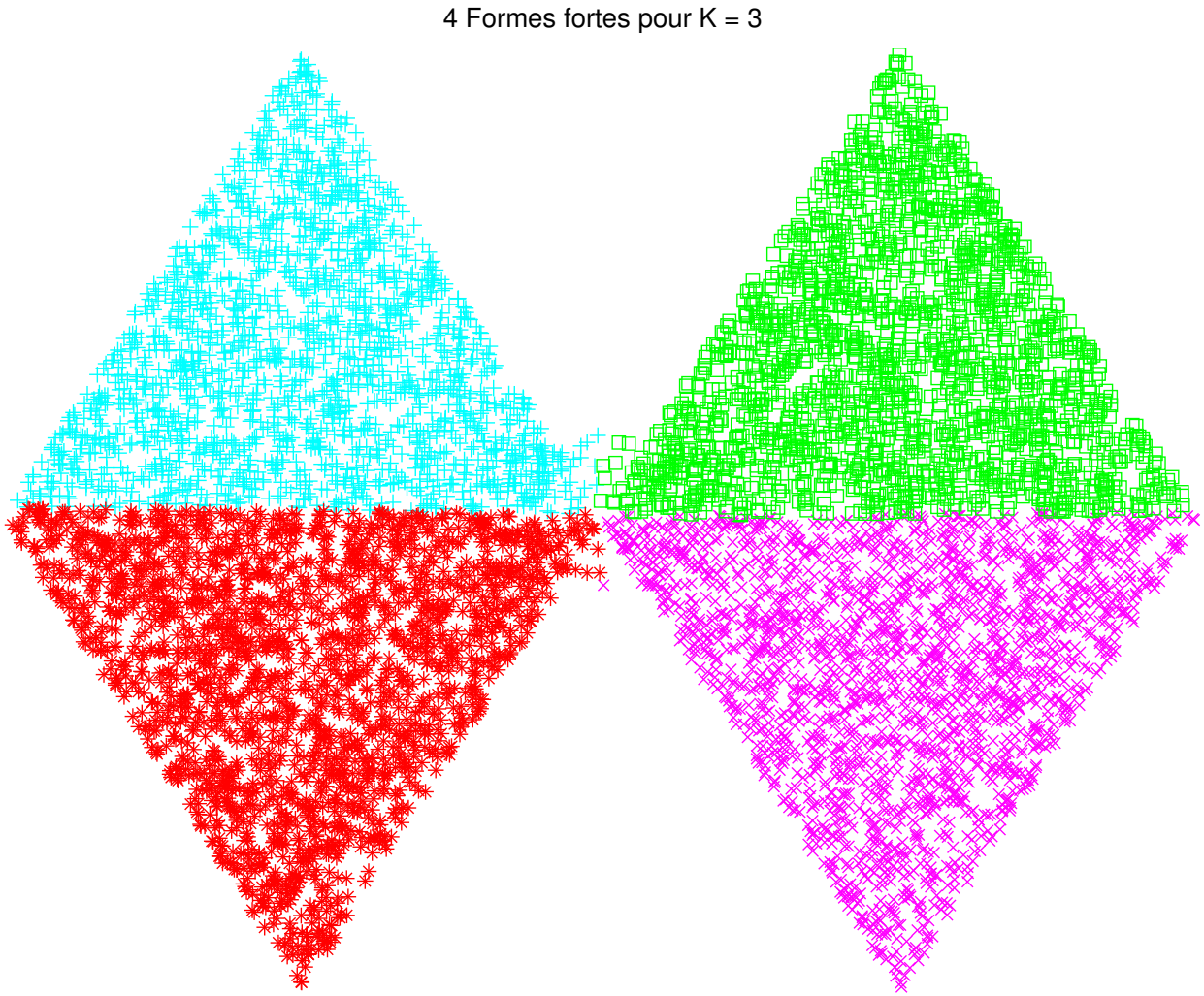
- On trouve 5 regroupements de points différents :

	F_1	F_2	F_3	F_4	F_5
N_i	2040	1940	49	2042	1929

- F_3 n'est pas représentatif
- F_1, F_1, F_4 et F_5 sont les formes fortes
- on peut recalculer les clusters à partir des centres des formes fortes

K-Means : formes fortes

- K-Means répété 6 fois



K-Means séquentiels

Adaptation des K-means lorsque les exemples arrivent au fur et à mesure

- Initialiser $\mu_1, \cdots \mu_K$
- Initialiser $n_1, \cdots n_K$ à 0
- Répeter
  - acquérir x
  - affectation du point au cluster le plus proche

$$\mathcal{C}_\ell \leftarrow x \quad \text{tel que} \quad \ell = \arg\min_k D(x, \mu_k)$$

  - incrémenter n_ℓ
  - recalculer le centre μ_ℓ de ce cluster

$$\mu_\ell = \mu_\ell + \frac{1}{n_\ell}(x - \mu_\ell)$$

Remarque

Si on dispose d'une partition initiale, on utilisera les centres des clusters et on initialisera $n_k = \text{card}(\mathcal{C}_k), k = 1, \cdots, K$

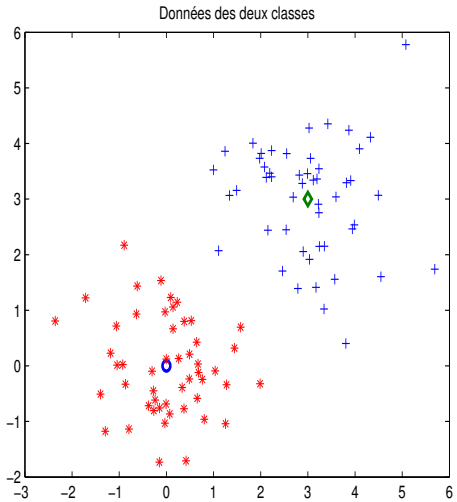
Approches de clustering

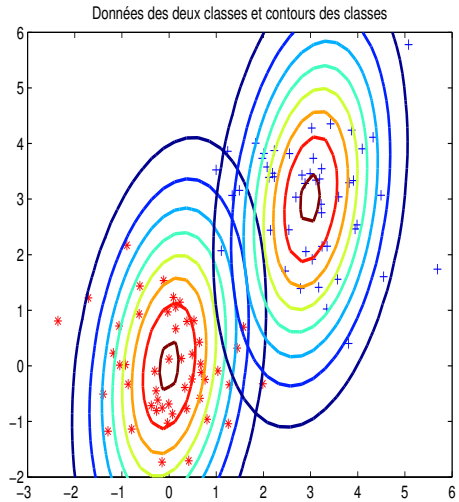
- Différentes approches sont possibles
- Clustering hiérachique
  - Clustering hiérarchique ascendant (CHA) et variantes
- Clustering par partitionnement
  - Algorithme des K-means
- Clustering par modélisation
  - Notion de modèles de mélange

Modèle de mélange pour le clustering

Introduction par l'exemple

- Considérons N données $\{x_i \in \mathbb{R}^d\}_{i=1,\dots,N}$ formant deux classes





- On veut trouver le modèle statistique des données.
- On constate que pour modéliser les données, il faut deux distributions gaussiennes

Clustering par modélisation statistique (1)

Introduction par l'exemple

- Dans la classe 1, les données suivent une loi normale $\mathcal{N}(\mu_1, \Sigma_1)$ Dans la classe 2, x suit $\mathcal{N}(\mu_2, \Sigma_2)$

- Loi marginale de x

$$\begin{aligned} f(x) &= f(x, \mathcal{C}_1) + f(x, \mathcal{C}_2) \\ &= f(x/\mathcal{C}_1)\Pr(\mathcal{C}_1) + f(x/\mathcal{C}_2)\Pr(\mathcal{C}_2) \text{ Th de Bayes} \\ f(x) &= \pi_1 f(x/\mathcal{C}_1) + \pi_2 f(x/\mathcal{C}_2) \end{aligned}$$

avec $$f(x/\mathcal{C}_1) \equiv \mathcal{N}(\mu_1, \Sigma_1), \quad f(x/\mathcal{C}_2) \equiv \mathcal{N}(\mu_2, \Sigma_2)$$

Clustering par modélisation statistique (2)

Introduction par l'exemple

- Loi marginale de x

$$f(x) = \pi_1 f(x/\mathcal{C}_1) + \pi_2 f(x/\mathcal{C}_2)$$

- π_1 et π_2 désignent la probabilité a priori que X relève resp. de la classe $\mathcal{C}_1$ et $\mathcal{C}_2$. Remarque : on a $\pi_1 + \pi_2 = 1$

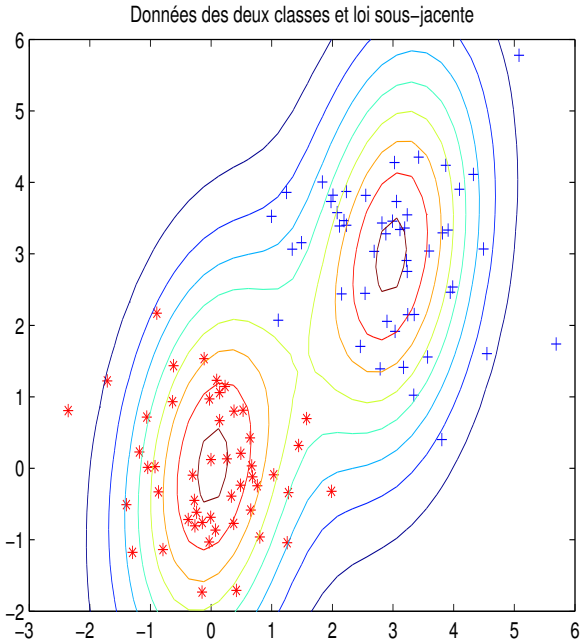
- $f(x/\mathcal{C}_1)$ et $f(x/\mathcal{C}_2)$ désignent la densité conditionnelle de x respectivement à $\mathcal{C}_1$ et $\mathcal{C}_2$

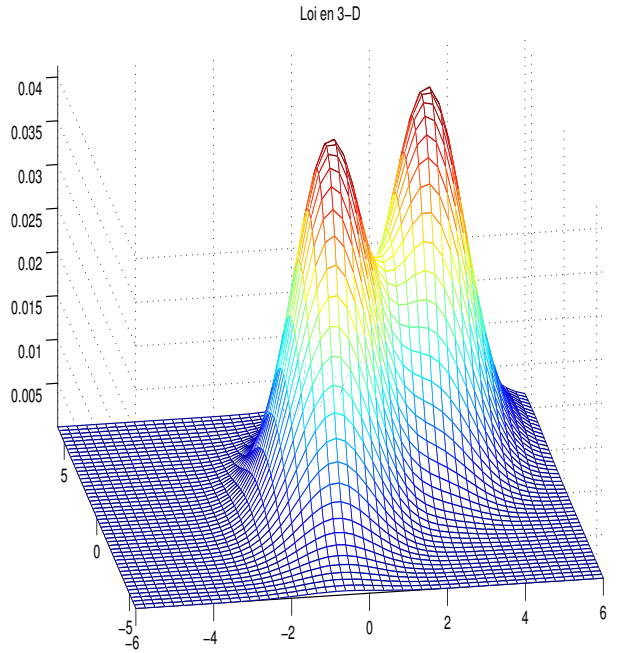
- $f(x)$ est entièrement déterminé par la connaissance de π_j, μ_j, Σ_j, $j \in \{1, 2\}$. On l'appelle modèle de mélange de densités

Modèle de mélange : illustration

- Modèle de mélange gaussien

$$f(x) = \pi_1 \mathcal{N}(x; \mu_1, \Sigma_1) + \pi_2 \mathcal{N}(x; \mu_2, \Sigma_2) \quad \text{avec} \quad \pi_1 + \pi_2 = 1$$





Du modèle de mélange au clustering (1)

Intérêt du modèle statistique pour faire du clustering ?

- Si on connait le modèle de mélange, on connait les probabilités à priori π_j et les lois conditionnelles $f(x/\mathcal{C}_j)$, $j \in \{1, 2\}$
- D'après le théorème de Bayes, on en déduit les probabilités a posteriori d'appartenance du point x à $\mathcal{C}_1$ et $\mathcal{C}_2$

$$\Pr(\mathcal{C}_1/x) = \frac{\pi_1 f(x/\mathcal{C}_1)}{f(x)}$$

$$\Pr(\mathcal{C}_2/x) = \frac{\pi_2 f(x/\mathcal{C}_2)}{f(x)}$$

- Remarque : on a $\Pr(\mathcal{C}_1/x) + \Pr(\mathcal{C}_2) = 1$

Du modèle de mélange au clustering (2)

Affectation des points aux clusters

- Affectation probabiliste
- Le point x est affecté à la classe de plus grande probabilité a posteriori

$$\begin{aligned} \mathcal{C}_1 &\leftarrow x \quad \text{si} \quad \Pr(\mathcal{C}_1/x) > \Pr(\mathcal{C}_2)/x \\ \mathcal{C}_2 &\leftarrow x \quad \text{sinon} \end{aligned}$$

- Ceci se généralise aisément pour $K > 2$ clusters

Du modèle de mélange au clustering (3)

Comparaison avec K-means

	K-means	Approche modélisation
Modèle	-	Modèle de mélanges $f(x) = \sum_{k=1}^{K} \pi_k f(x/\mathcal{C}_k)$
Paramètres à estimer	Centres des clusters $\mu_k,\ k = 1, \cdots, K$	Proba a priori π_k Paramètres des lois $f(x/\mathcal{C}_k)$
Critère optimisé	Variance intra-classe	Log-Vraisemblance
Indicateur d'affectation	$d(x, \mu_k)$	Proba a posteriori $\Pr(\mathcal{C}_k/x)$
Règle d'affectation de x	Cluster dont le centre est le plus proche	Cluster de plus grande proba a posteriori

Conclusion

- Clustering : apprentissage non-supervisé
- Regroupement des données en classes (clusters) homogènes
- Propriétés d'un bon clustering : identifier des classes très homogènes (faible variance intra-classe) et bien séparées (forte variance inter-classe)
- Il existe des critères composites permettant d'évaluer la qualité d'un clustering
- Le nombre de clusters est un hyper-paramètre qui dépend de l'application mais peut être déterminé sur la base de ces critères composites
- Algorithmes : CHA, K-means, mélange de distributions et bien d'autres

Quelques Toolboxes

- Python :
  `http://scikit-learn.org/stable/modules/clustering.html`
- R :
  - `https://cran.r-project.org/web/packages/fastcluster/index.html`
  - `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html`
- Matlab
  `http://fr.mathworks.com/help/stats/cluster-analysis.html`