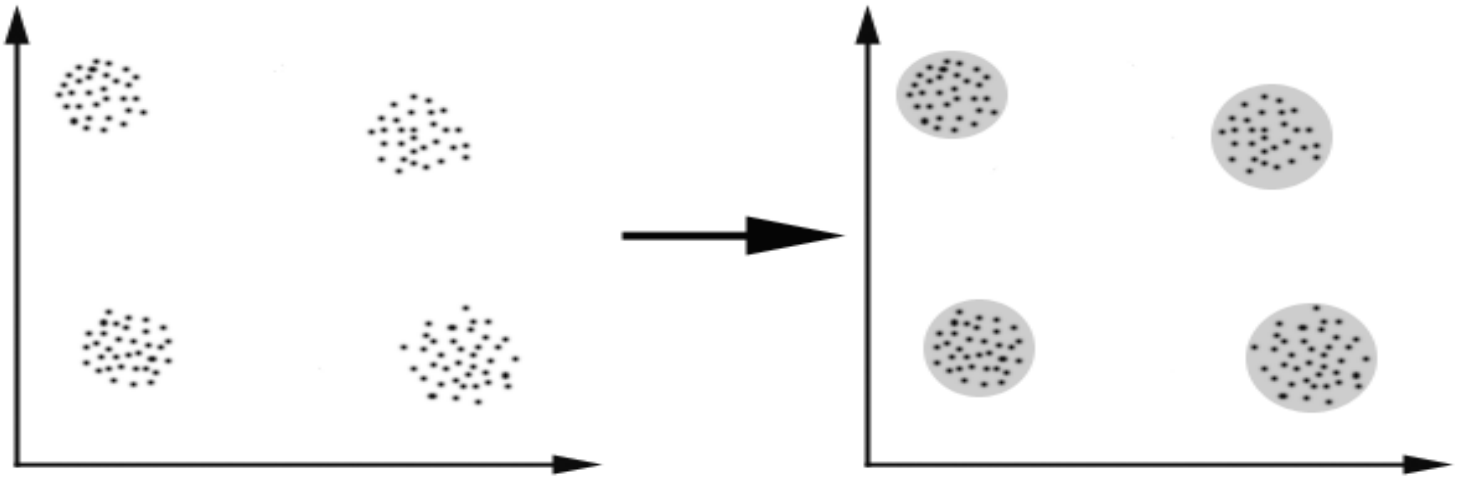
Clustering (Kmeans)

C’est quoi le clustering ?

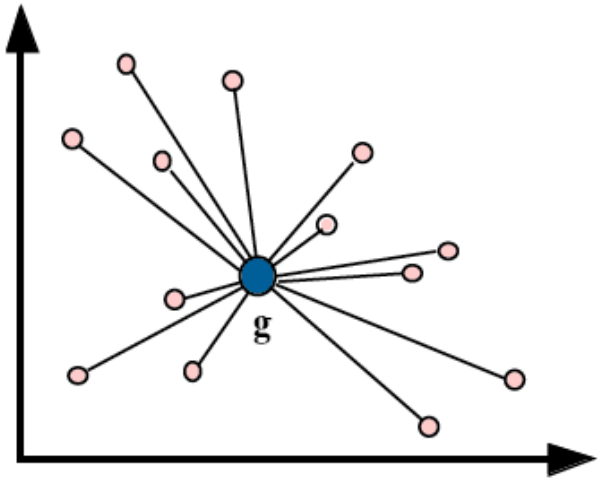


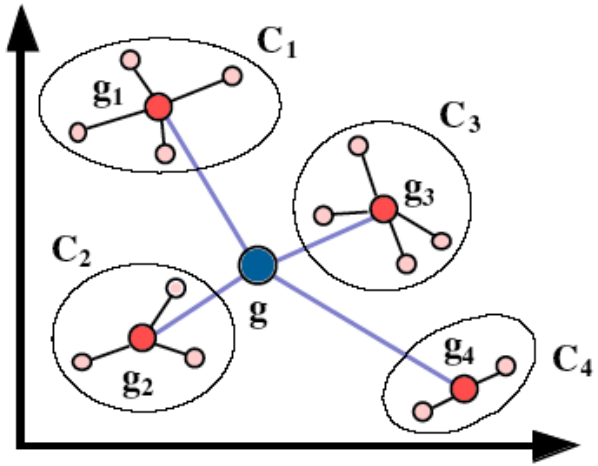
Trouver K clusters/ groupes/ensemble de données homogènes. (les données appartenant à des clusters différents sont dissimilaires)

Comment savoir si un regroupement est "correct" ?

- inertie (intra) d'un cluster = variance des points d'un même cluster

- inertie (inter) = variance des centres des clusters





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

K-means

K-means : construire une partition à K clusters

- Chaque cluster est associé à un centre (prototype)
- Chaque donnée est affectée au centre « *le plus proche* »
- Le nombre de clusters (K) est fixé a priori
- L'algorithme est simple

L’algorithme

Choisir K centres initiaux

Repeter

// on crée K clusters

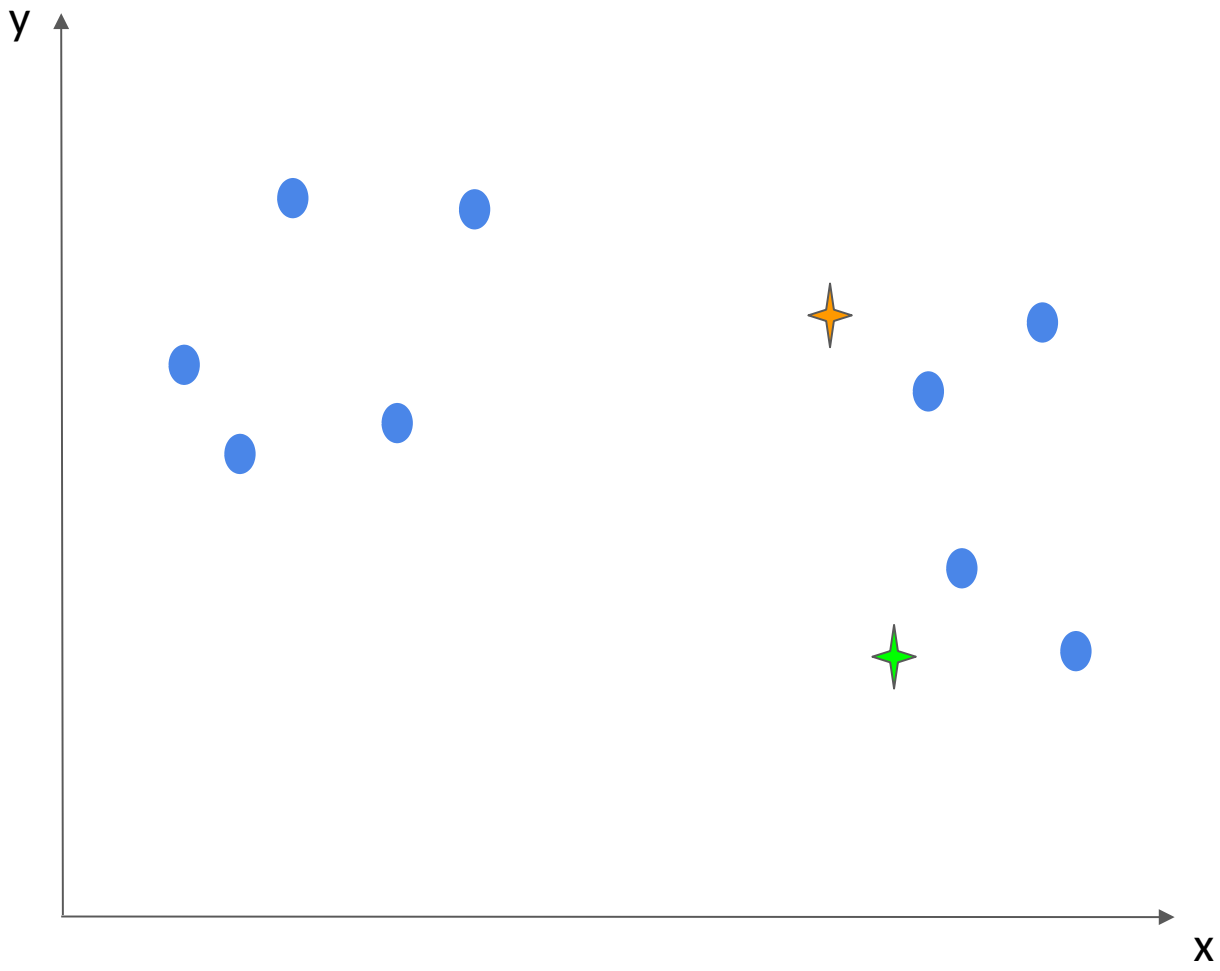
Affecter chaque donnée au centre le plus proche

// calcule les centres de K clusters

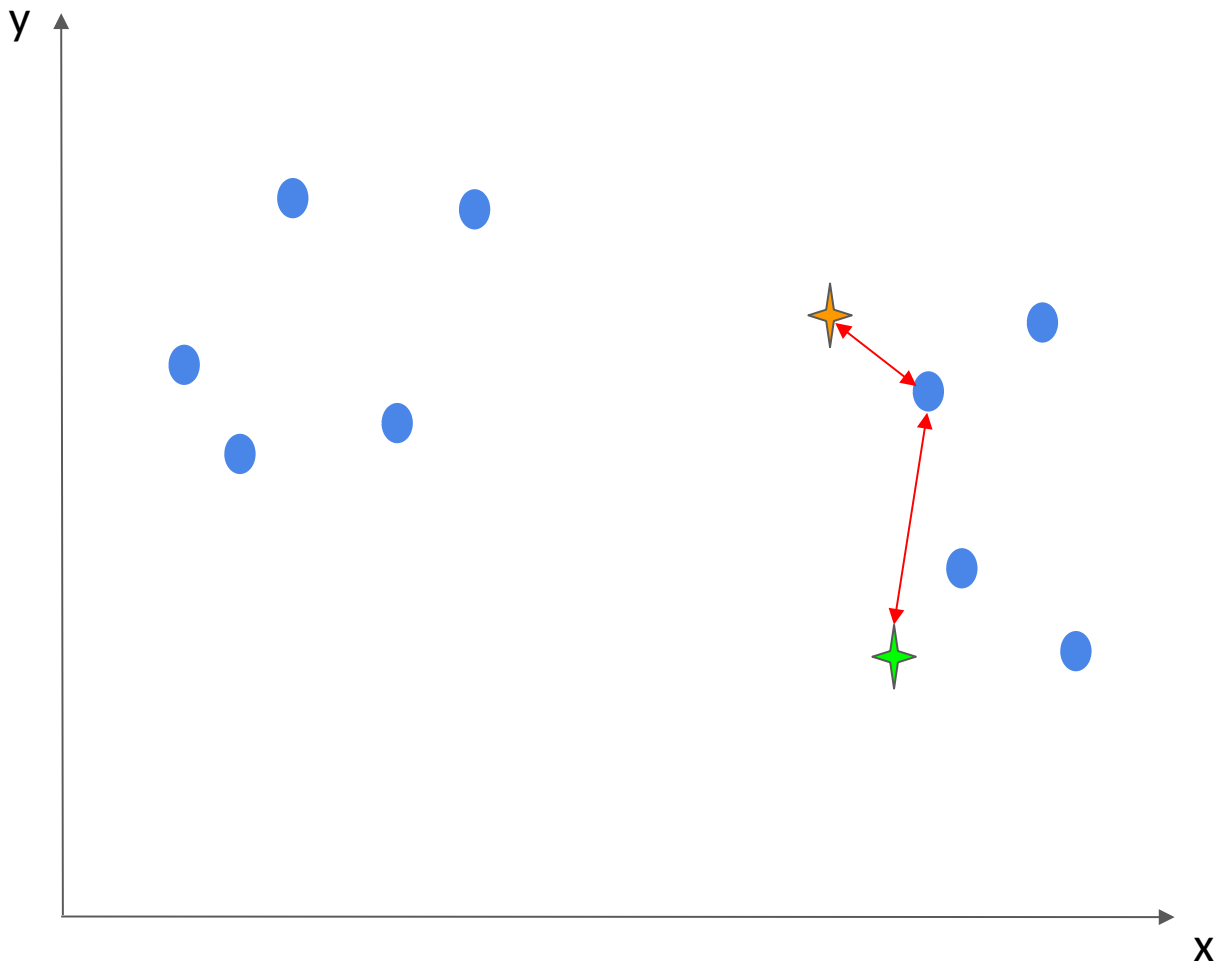
Mettre à jour les centres

Jusqu’a aucun centre n’a changé

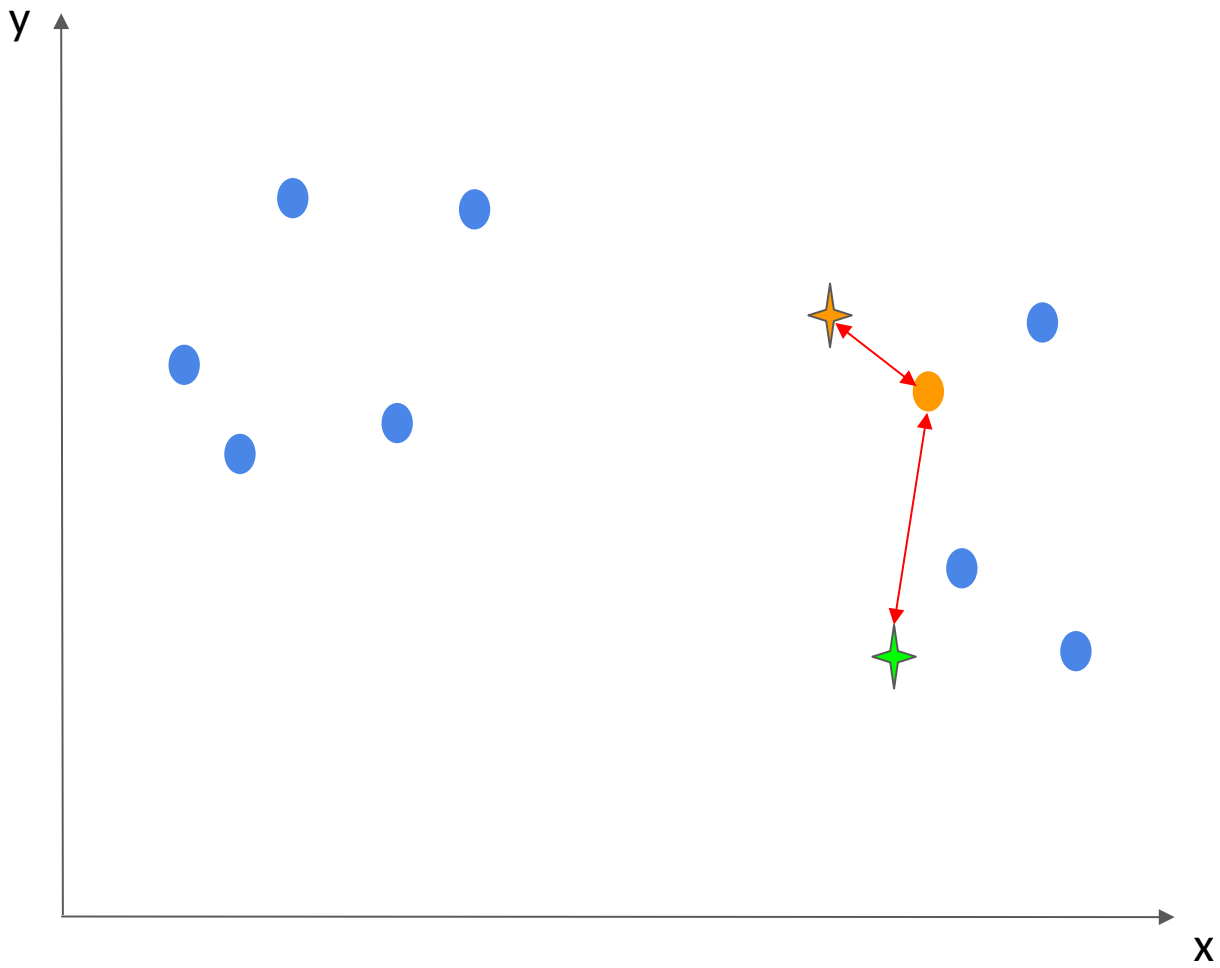
K-means - initialisation



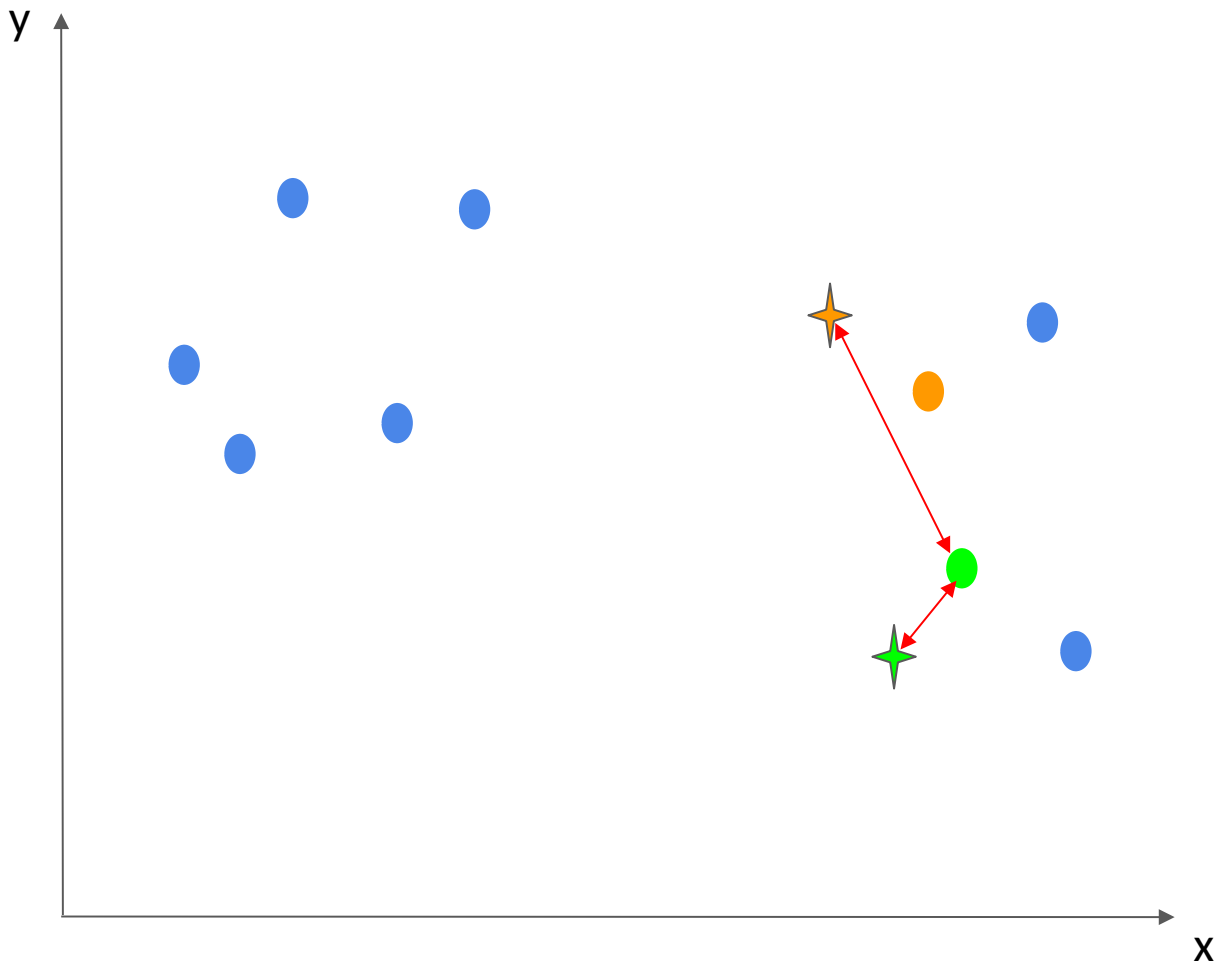
K-means - compute distances with prototypes



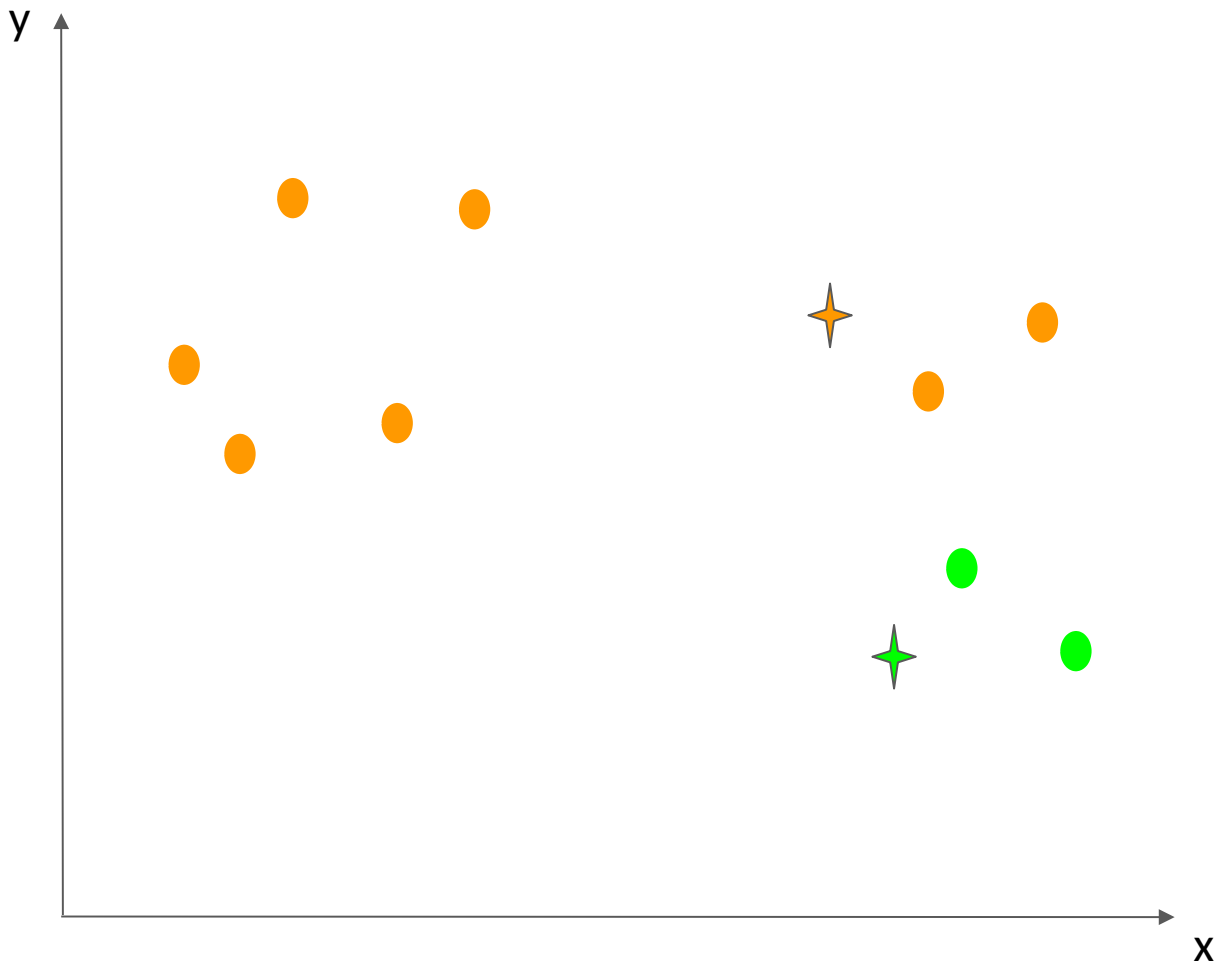
K-means : Assignment



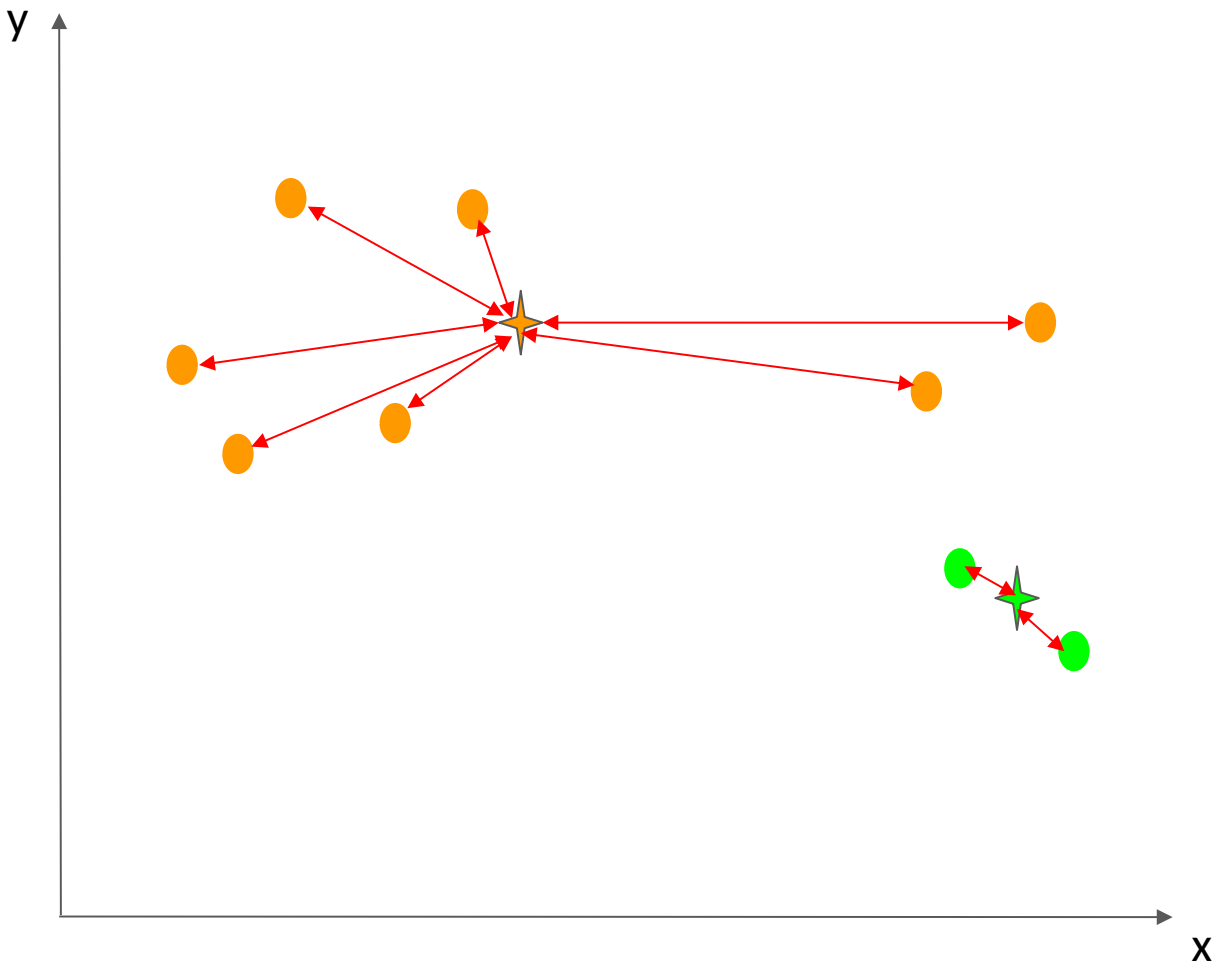
K-means : Assignment



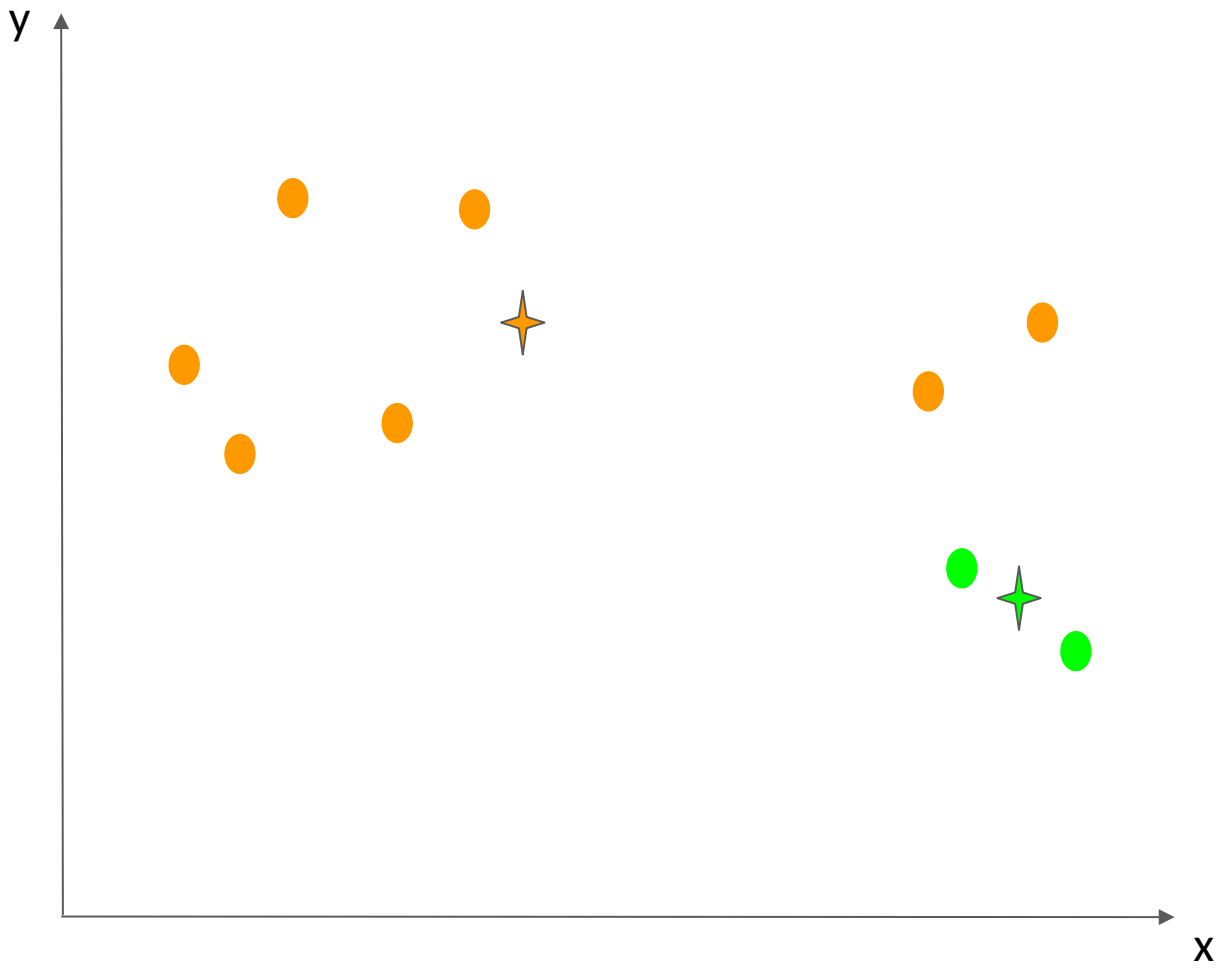
K-means : Assignment



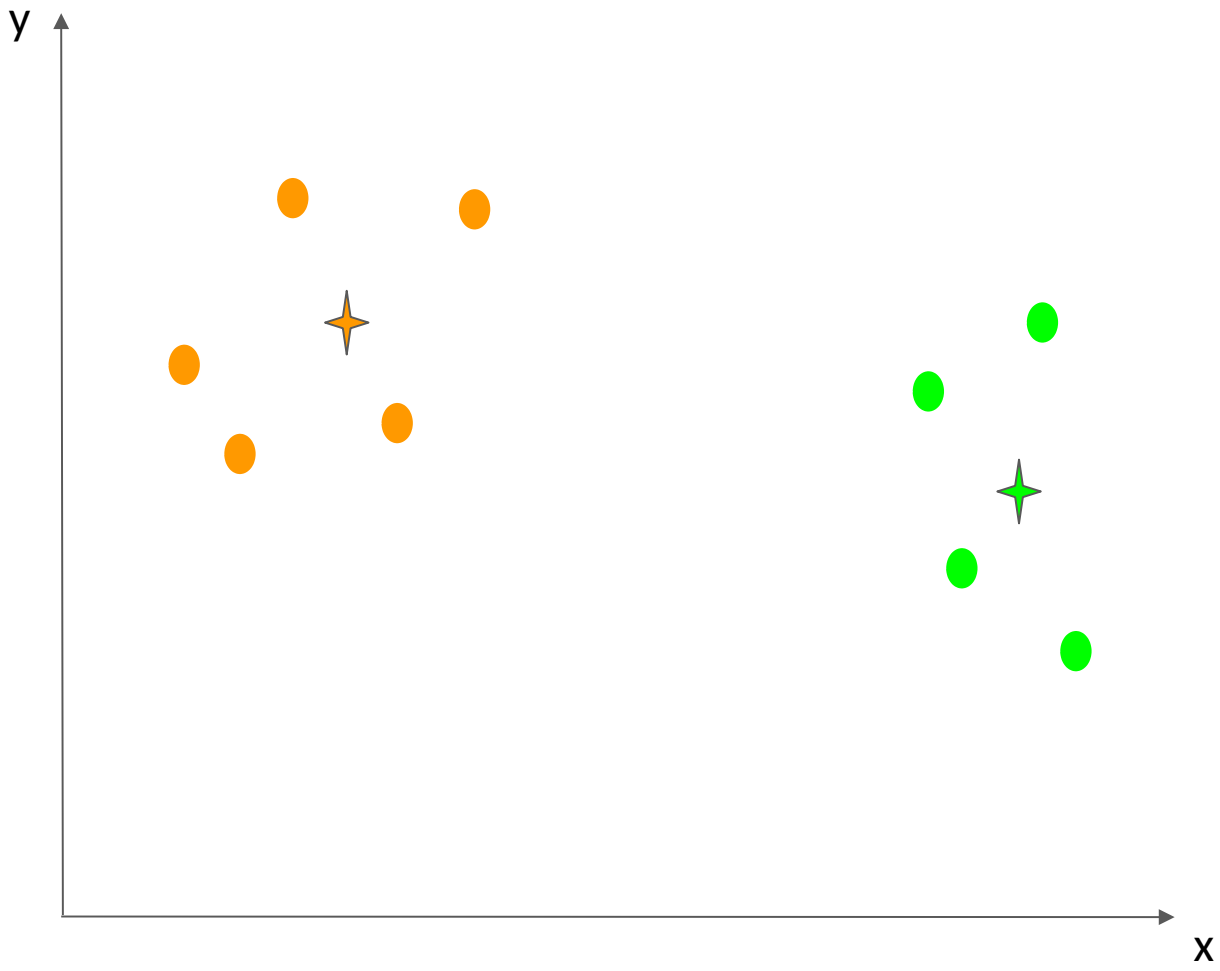
K-means - Reduce (prototype update)



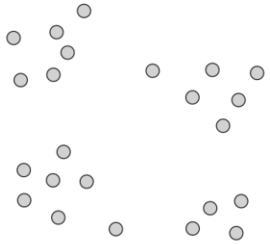
K-means : iteration 1



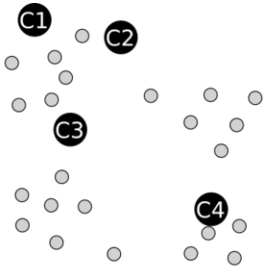
K-means : iteration 2



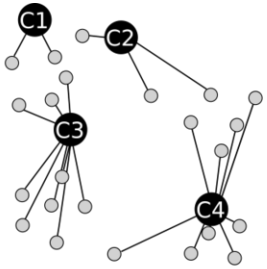
Déroulement algorithme



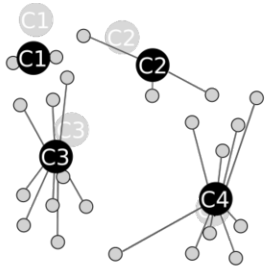
0a. Données d'entrée



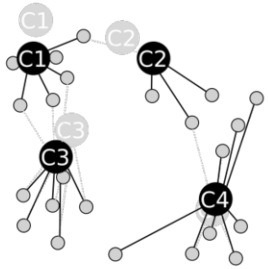
0b. intialisation



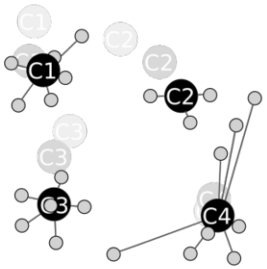
1a. assignation



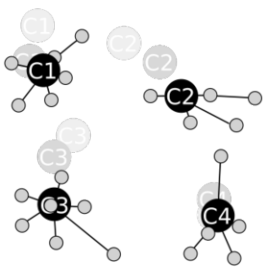
1b. calcul des points moyens



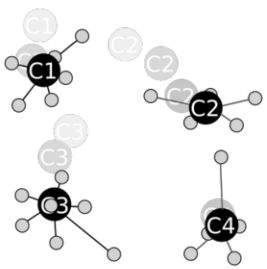
2a. assignation



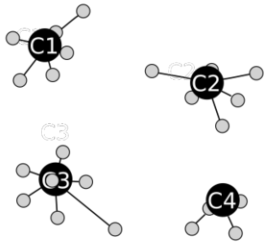
2b. calcul des points moyens



3a. assignation



3b. calcul des points moyens



4a. assignation

clusters stables (fin)

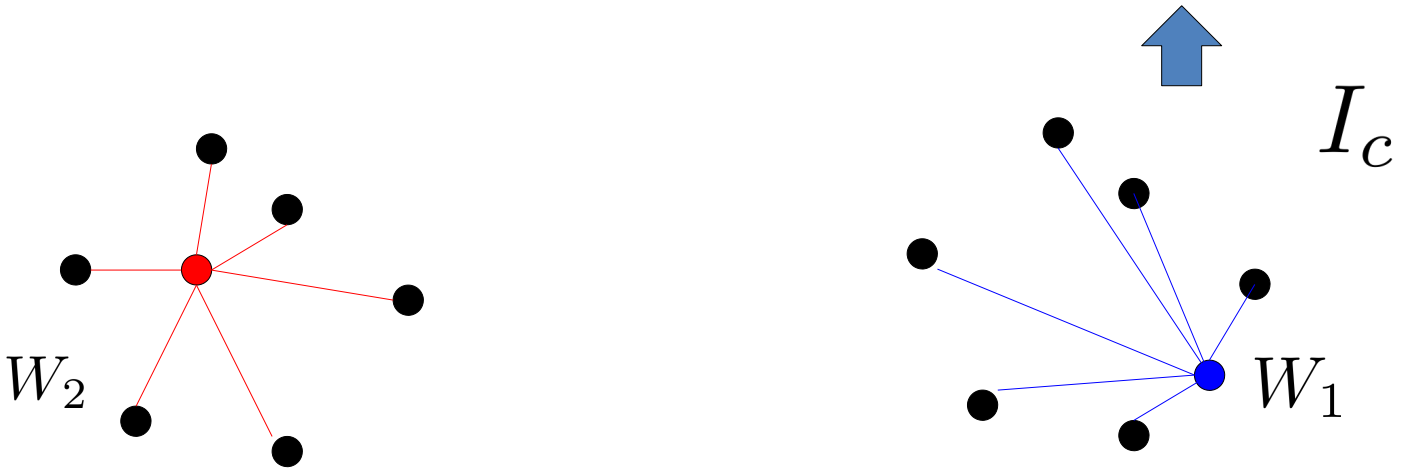
K-means : minimisation de l'inertie

ϕ (x) : fonction d'affectation de la donnée *x* à un cluster *c*

W = {w_1,..,w_k} : ensemble des centres des clusters

I (W, ϕ) : inertie globale

$$I(W,\phi) = \sum_{\mathbf{x}_i} ||\mathbf{x}_i - \mathbf{w}_{\phi(\mathbf{x}_i)}||^2 = \sum_{c} \sum_{\mathbf{x}_i \in P_c} ||\mathbf{x}_i - \mathbf{w}_c||^2$$



Calcul des centres

Le centre W_c d'un cluster c est la valeur qui minimise la variance des données x_i du cluster

$$\mathbf{w}_c = \frac{\sum_{\mathbf{x}_i \in P_c} \mathbf{x}_i}{|P_c|}$$

Choix des centres initiaux

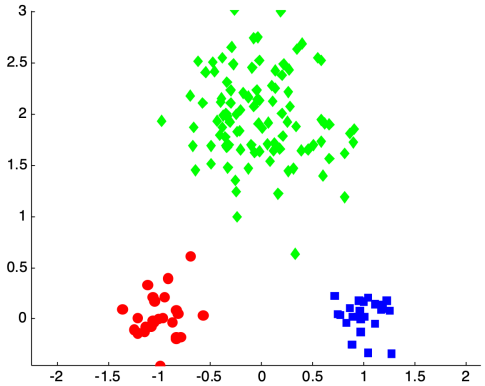
- aléatoirement dans l'intervalle de définition des x_i

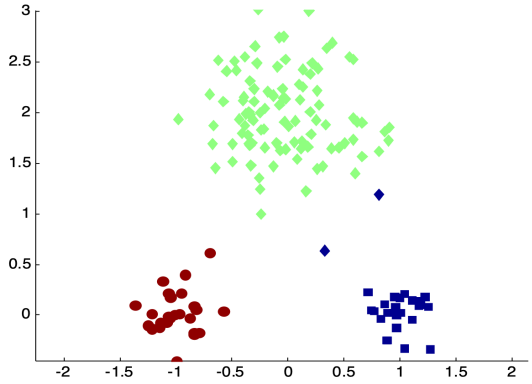
- aléatoirement dans l'ensemble des x_i

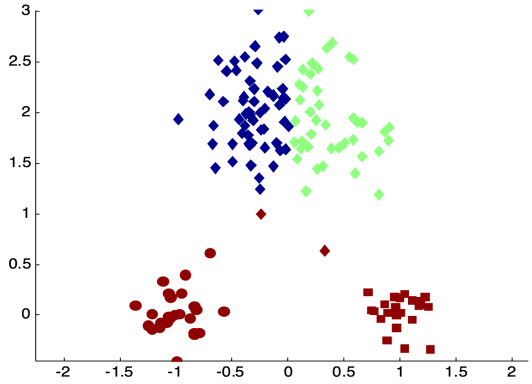
Des initialisations différentes peuvent mener à des clusters différents (problèmes de minima locaux)

méthode **générale pour obtenir des clusters "stables" =**
formes fortes, on répète l'algorithme n fois

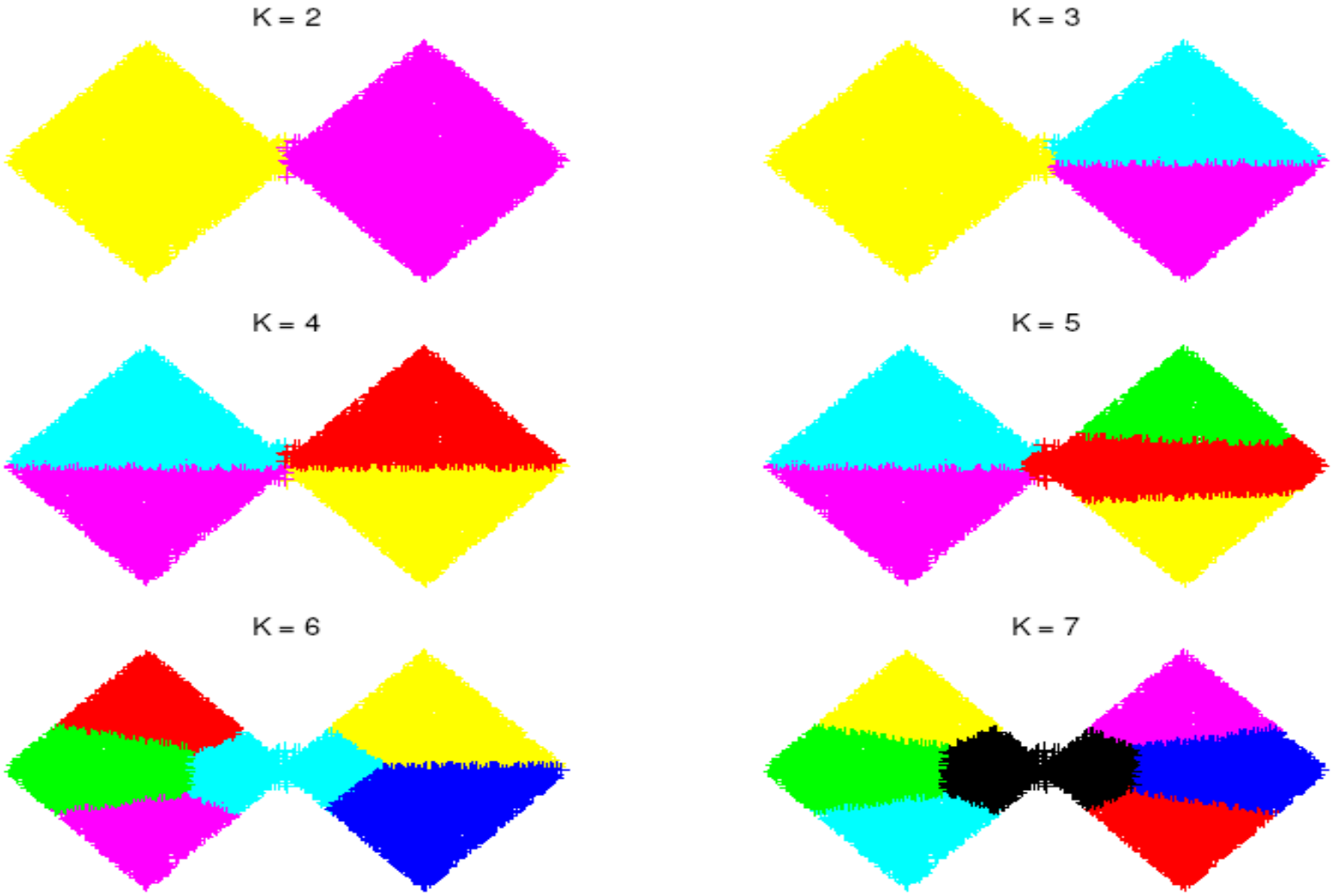
K-Means : résultats à partir de centres initiaux différents







K-Means : résultats selon le nombre de clusters



Indice de qualité

Propriétés requises

détermination du nombre de clusters optimal

(variations de l'indice en fonction du nb de clusters)

- compacité des clusters

- séparabilité des clusters

Différents types d'indice de qualité

- **Indices internes** l'évaluation des regroupements obtenus se fait de façon non supervisée à partir des seules données utilisées pour construire les clusters.

Nécessité d'indices de qualité internes spécifiques.

- **Indices externes** l'évaluation du clustering se fait à partir d'informations externes sur les objets, par exemple une étiquette de classe.

On est alors ramené aux indices d'une évaluation supervisée.

Indices relatifs L'évaluation se fait en comparant les résultats de plusieurs clusters ou clusterings, à partir d'indices externes ou internes.

Inerties within et between

***I_{tot}*, l'inertie totale du nuage** est la somme des carrés des distances des objets au centre du nuage.

$$I_{tot} = \sum_{i=1}^{n} \| x_i - c \|^2 = \sum_{i=1}^{n} d^2(x_i, c)$$

Décomposition : $I_{tot} = W + B$

W mesure la cohésion des clusters par la somme des carrés des distances de chaque objet à son centre de cluster
→ **le plus petit souhaité**

$$W = \sum_{i=1}^{n} \| x_i - c(x_i) \|^2 = \sum_{i=1}^{n} d^2(x_i, c(x_i))$$

B mesure la séparation des clusters par la somme des carrés des distances entre centres → **le plus grand souhaité**

$$B = \sum_{k=1}^{K} n_k \| c_k - c \|^2 = \sum_{k=1}^{K} n_k d^2(c_k, c)$$

Ratio B/W et indice CH

Ces mesures intègrent *B* et *W* en un seul indice, pénalisé par le nombre de classes, noté *K*, dans le cas de *CH*

$$\text{ratio } B/W = \frac{B}{W} \qquad CH = \frac{(n-K)B}{(K-1)W}$$

- plus les classes sont compactes et séparables, plus *B* est grand, *W* petit et donc ratio *B/W* et *CH* grand.
- *CH* pénalise d'autant plus le ratio *B/W* que *K* est grand, ce qui facilite la détermination du nombre de classe optimal
- *CH* est à maximiser, compris entre 0 et + ∞, sensible au bruit

DB, Indice de Davies-Bouldin

L'indice DB repose sur la notion de similarité de 2 clusters

Similarité de 2 clusters

- dispersion d'un cluster k : δ_k
- dissimilarité de 2 clusters : $\Delta_{kk'}$
- similarité des clusters k et k' : $$R_{kk'} = \frac{\delta_k + \delta_{k'}}{\Delta_{kk'}}$$

Définition de DB

- dispersion du cluster k → δ_k = distance moyenne des objets du cluster k au centre du cluster
- dissimilarité de 2 clusters → $\Delta_{kk'}$ = distance du centre du cluster k au centre du cluster k'
- formule de DB : $$DB = \frac{1}{K} \sum_{k=1}^{K} Max_{k' \neq k} \{R_{kk'}\}$$

SIL, indice Silhouette

Silhouette d'un objet i

$$SIL(i)_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

***a_i*, distance moyenne de l'objet i aux objets de son cluster → Compacité**

$$a_i = \frac{1}{n(C(x_i)) - 1} \sum_{x_j \in C(x_i)} d(x_i, x_j)$$

***b_i*, minimum des distances moyenne de l'objet i aux objets de chacune des autres classes**
→ Séparabilité

$$b_i = \underset{k' \neq k(x_i)}{Min} \left\{ \frac{1}{n_{k'}} \sum_{x_j \in C_{k'}} d(x_i, x_j) \right\}$$

Silhouette d'un cluster C_k ou d'un clustering C
= moyenne des silhouettes concernées, SIL(C_k) ou SIL(C)

SIL est compris entre -1 est 1, à maximiser. Très bons résultats lors des expériences d'Arberlaitz 13.

Accuracy, Rappel, Précision, F-mesure

Procédure

Si on dispose d'objets étiquetés, il est possible d'appliquer les procédures d'évaluation habituelles en supervisé.

- clustering sur LS, sans tenir compte de l'étiquette
- prédiction à partir de l'étiquette majoritaire dans le cluster ou de règles adaptées au déséquilibre des classes
- calcul sur TS des indices de qualité habituels à partir de la matrice de confusion booléenne qui croise prédit et réel.
- ule accuracy qui est inadaptée au cas de

Matrice de confusion, Accuracy, Rappel,

Accuracy :
p = *(TP + TN)/n = (a+d)/n*

Precision :
p = *TP/(TP + FP) = a/(a+c)*

Recall :
r = *TP/(TP + FN) = a/(a+b)*

Predict. Actual	Class 1	Class 2	**total**
Class 1	a (TP)	b (FN)	a+b
Class 2	c (FP)	d (TN)	c+d
total	a+c	b+d	n

F1-mesure : compromis entre *p* and *r*, défini comme la moyenne harmonique de *p* et *r*. Identique à l'indice d'association de *Czekanowski* (ou indice de Dice) entre étiquette prédite et étiquette réelle.

F1-measure = 2TP/(2TP+FN+FP) = 2a/(2a+b+c)