

CLUSTERING

(Partitionnement)

n observations (pas d'étiquettes) \rightarrow former K groupes

1) Evaluation


1) Critères géométriques : on veut que les observations d'un même cluster soient proches
entre 2 clusters soient différentes

homogénéité

séparabilité

centroïde d'un cluster = barycentre

homogène



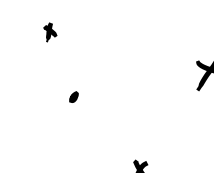
$$\vec{\mu}_k = \frac{1}{|C_k|} \sum_{\vec{x} \in C_k} \vec{x}$$

* homogénéité de C_k : "tightness"

$$T_k = \frac{1}{|C_k|} \sum_{\vec{x} \in C_k} d(\vec{x}, \vec{\mu}_k)$$

idéalement : PETIT

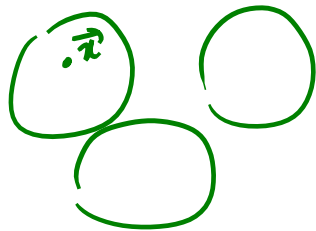
moins homogène



* séparabilité de C_k et C_l : $S_{kl} = d(\vec{\mu}_k, \vec{\mu}_l)$

idéalement : GRAND

Coefficient de silhouette = à quel point \vec{x} est "bien" dans son cluster



$$s(\vec{x}) = \frac{b(\vec{x}) - a(\vec{x})}{\max(a(\vec{x}), b(\vec{x}))}$$

$b(\vec{x})$: distance de \vec{x} aux autres points de son cluster (en moyenne)

$$b(\vec{x}) = \frac{1}{|C(\vec{x})|} \sum_{\vec{u} \in C(\vec{x})} d(\vec{x}, \vec{u})$$

$a(\vec{x})$: la plus petite valeur que prendrait $a(\vec{x})$ si \vec{x} était dans 1 autre cluster.

2) Comparaison à des étiquettes connues (quand on en a)

Problème = identification des numéros de cluster aux classes connues

Indice de Rand : compter le nombre de paires d'observations à la fois dans la même classe et dans le même cluster et à la fois dans 2 classes \neq et dans 2 clusters \neq .

index = 0

Pour $i = 1 \dots (n-1)$,

Pour $l = (i+1) \dots n$:

Si \vec{x}_i et \vec{x}_l ont la même étiquette & sont dans le même cluster :

index += 1

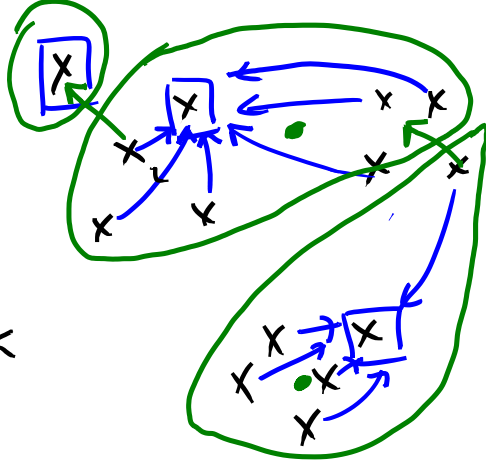
Si \vec{x}_i et \vec{x}_l n'ont pas la même étiquette & ne sont pas dans le même cluster

index += 1

return $\left\lfloor \frac{\text{index}}{n(n-1)/2} \right\rfloor$ entre 0 et 1

[2] K-moyennes (kmeans)

On fixe K (nombre de clusters) $K=3$



* il faut choisir K

* initialisation
aléatoire

↳ kmeans++ est déterministe

* kernel kmeans = version à noyaux

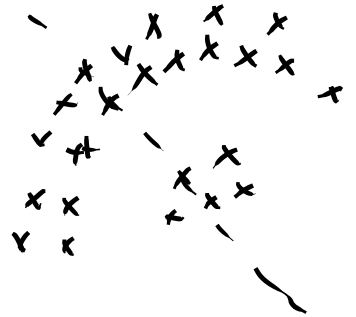
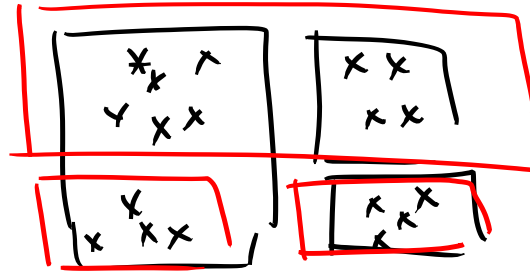
* clusters nécessairement convexes

* Initialisation:

choisir K observations au hasard
 $\equiv K$ centroïdes

→ * Associer chaque \vec{x} au centroïde
dont il est le plus proche

* Recalculer les centroïdes
Itérer jusqu'à stabilité



[3] Clustering par densité DBSCAN

voisinage de taille ϵ

on construit des clusters de
proche en proche



⊗ On ne précise pas k à l'avance

* Les points seuls dans leur cluster sont identifiés comme outliers

* Difficulté: choisir ϵ