Algorithme des k moyennes

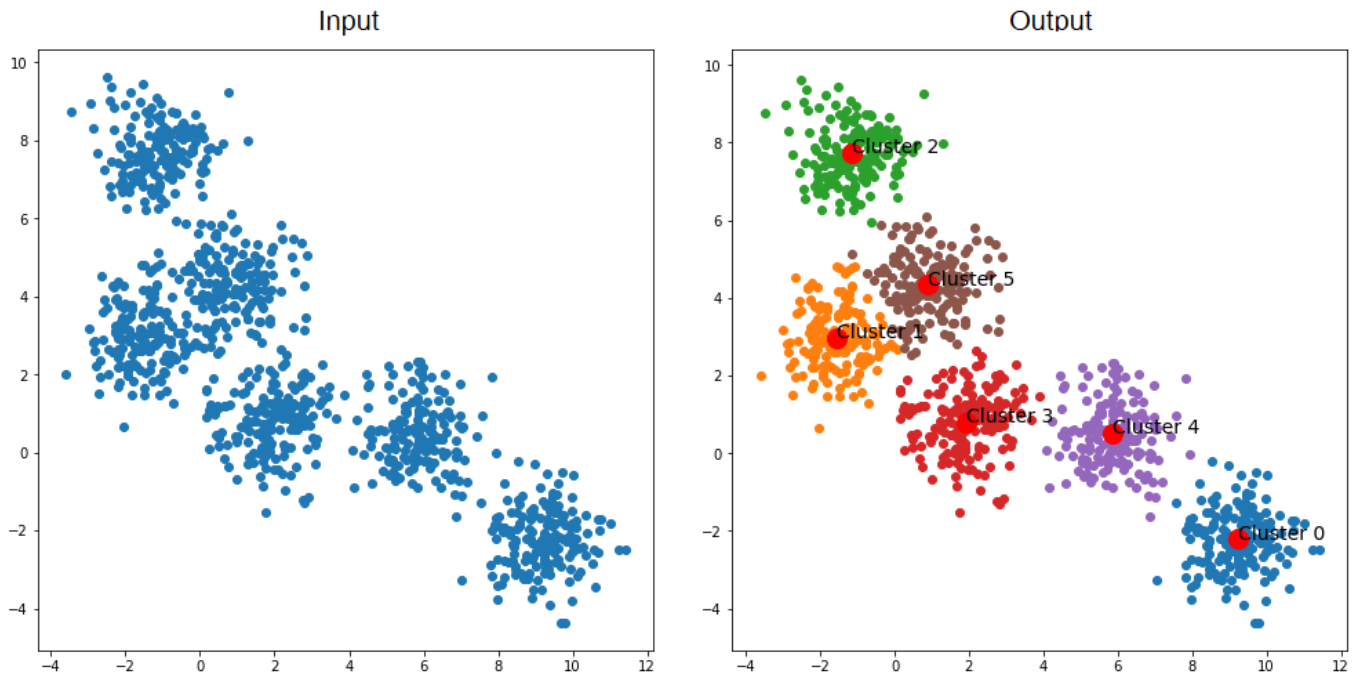
Fabrice Lembrez - PSI*

Lycée Pierre de Fermat

Plan

Le problème du partitionnement

Il s'agit de scinder les données en sous-groupes pertinents, mais on n'a pas d'information a priori, juste les données brutes.



Ici on a pris un cas où les "clusters" se voient bien, ce qui est rarement le cas en pratique.

Le problème du partitionnement

- Données multidimensionnelles ($\mathbb{R}^d$)
- à répartir en classes (ou groupes) aussi homogènes que possible
- selon des critères de distance mutuelle (distance ordinaire de $\mathbb{R}^d$)
- sans information a priori (non supervisé).

Nombre de classes a priori inconnu

Même le nombre de groupes pertinents est inconnu : on peut tenter une valeur k, une autre,... et chercher ce qui est mieux.

Dans un premier temps on le fixe à une valeur donnée k

Une quantité à minimiser

On cherche à partitionner les données sous la forme $(S_1, \ldots, S_k)$ en minimisant les carrés des distances intra-blocs : on définit la distorsion

$$D = \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{(x,y)\in S_i^2} \|x - y\|^2$$

Une quantité à minimiser

On cherche à partitionner les données sous la forme $(S_1, \ldots, S_k)$ en minimisant les carrés des distances intra-blocs : on définit la distorsion

$$\boxed{D = \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{(x,y)\in S_i^2} \|x - y\|^2}$$

Argument classique de géométrie : introduire l'isobarycentre μ_i des éléments de S_i

$$\sum_{(x,y)\in S_i^2} \|x - y\|^2 = \sum_{(x,y)\in S_i^2} \|(x - \mu_i) + (\mu_i - y)\|^2$$

en développant ce carré, les produits scalaires disparaissent et

$$\sum_{(x,y)\in S_i^2} \|x - y\|^2 = 2|S_i| \sum_{(x,y)\in S_i^2} \|x - \mu_i\|^2$$

Une quantité à minimiser

On cherche à partitionner les données sous la forme $(S_1, \ldots, S_k)$ en minimisant les carrés des distances intra-blocs : on définit la distorsion

$$D = \sum_{i=1}^{k} \frac{1}{|S_i|} \sum_{(x,y)\in S_i^2} \|x - y\|^2$$

On arrive à

$$D = 2 \sum_{i=1}^{k} \sum_{x\in S_i} \|x - \mu_i\|^2$$

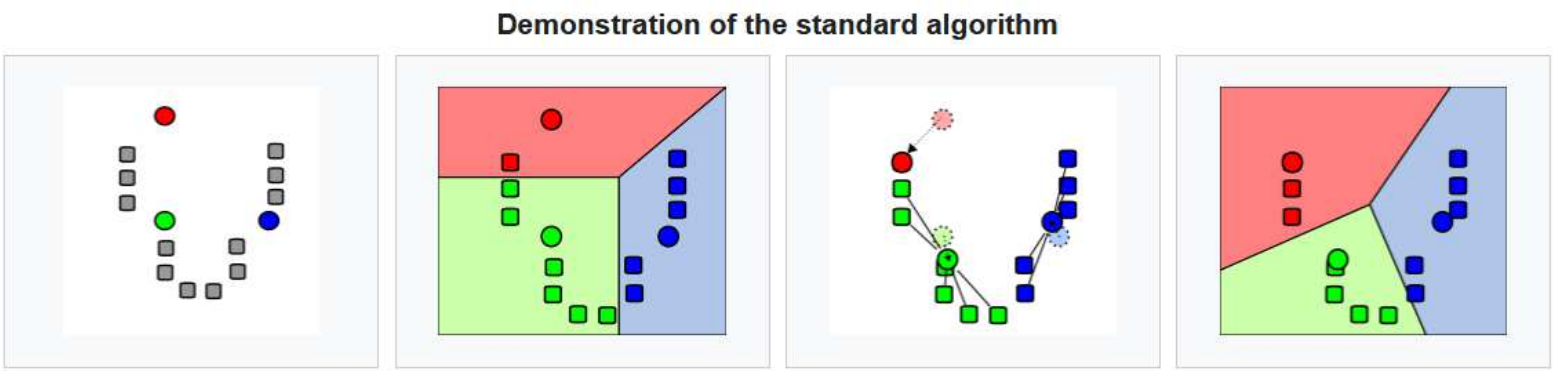
qu'on peut également interpréter en termes de variance.

On va voir un algorithme qui permet de trouver des minima locaux (pas toujours globaux) de la distorsion.

Plan

Algorithme des k-moyennes

- Initialisation : choix de k données particulières (voir plus tard) qui seront nos "centres"
- Succession d'étapes
  - affectation de chaque donnée à un centre "optimal"
  - recentrage en calculant de nouveaux isobarycentres
- et ceci jusqu'à ce que la situation n'évolue plus : pas de réaffectation.



Plan

Quelles garanties ?

On peut montrer que l'algorithme converge (du moins quand on travaille effectivement avec la norme euclidienne). Mais

- le nombre d'étapes est souvent limité, mais parfois vraiment long

- la complexité globale reste élevée de toute façon

- surtout on tombe couramment sur des extrema locaux peu intéressants (cf TP).

Quelles garanties ?

On peut montrer que l'algorithme converge (du moins quand on travaille effectivement avec la norme euclidienne). Mais

- le nombre d'étapes est souvent limité, mais parfois vraiment long

- la complexité globale reste élevée de toute façon

- surtout on tombe couramment sur des extrema locaux peu intéressants (cf TP).

Il y a différentes pistes d'amélioration un peu lourdes, comme faire tourner plusieurs fois l'algorithme et prendre la distorsion la plus faible...

Plan

Stratégies d'initialisation

Vous n'avez pas à les connaître...

- le plus simple (Forgy) : prendre k données au hasard

- un autre départ aléatoire : affecter chacun des points au hasard et attaquer par l'étape de recentrage.

En termes qualitatifs on voit quels défauts peuvent présenter ces choix.

Stratégies d'initialisation

Vous n'avez pas à les connaître...

- le plus simple (Forgy) : prendre k données au hasard

- un autre départ aléatoire : affecter chacun des points au hasard et attaquer par l'étape de recentrage.

En termes qualitatifs on voit quels défauts peuvent présenter ces choix.

Etudié en TP : l'algorithme k-means++ (2007) propose un choix initial adapté à la géométrie

Pour le TP : l'idée de k-means++

Pour initialiser,

- choisir une première donnée D1 au hasard

- choisir la seconde D2 en la cherchant plutôt loin de la première
  loi de probabilité pondérée par la distance au carré

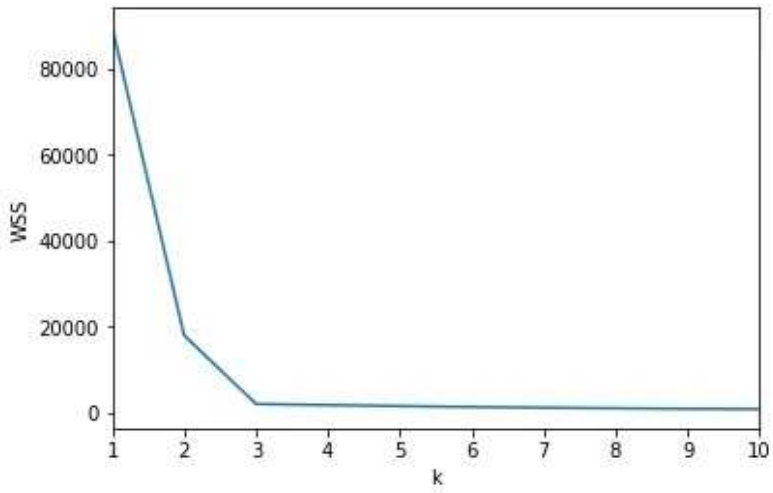
- choisir la troisième en la cherchant plutôt loin des deux premières
  loi de probabilité : la plus petite distance au carré (avec D1 ou D2)

- etc

Plan

Trouver le bon k

Une approche naïve consiste à tester les valeurs de k les unes après les autres, à mesurer la distorsion.
La distorsion a tendance à diminuer, d'abord très fortement, puis très peu.
On peut utiliser la valeur de k qui provoque la séparation entre ces deux comportements ("méthode du coude" car le graphe fait un coude).



Après k=3, plus d'amélioration notable.

Il existe aussi des critères quantitatifs plus précis pour estimer la qualité du partitionnement et aider notamment à choisir le bon k : on verra en TP le coefficient de silhouette.