K-Means

TIPE - 25 NOVEMBRE 2016 - CPI2

DAVID RIGAUX, MEHDI DALAA, MANDRY MBUNDU & MAXIME LUNDQUIST

CONCEPT

- K-Moyennes est une méthode de partitionnement de données et un problème d'optimisation combinatoire.
- Méthode : Soit n un nombre de points et *k* un entier donné, le but de k-moyennes est de diviser les n-points en k-groupes Ces groupes sont appelés clusters.

MATHÉMATIQUES

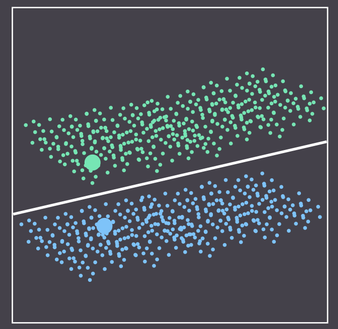
- Problème : K-moyennes doit minimiser la distance entre les points à l'intérieur de chaque cluster.

INFORMATIQUE

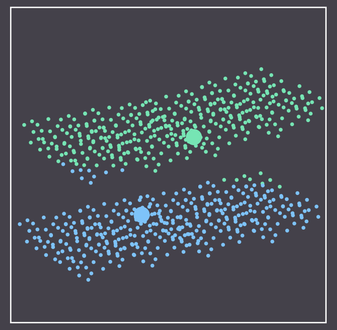
- Les k-moyennes sont utilisées par certains logiciels pour diviser un groupe hétérogène de données en sous-groupes plus homogènes.

CONCEPT

ETAPE 1

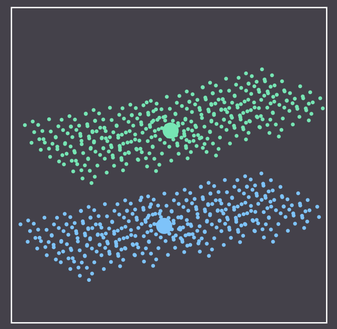


ETAPE 4

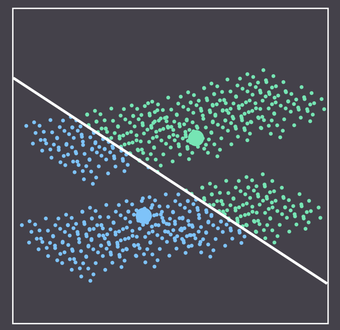


MATHÉMATIQUES

ETAPE 2

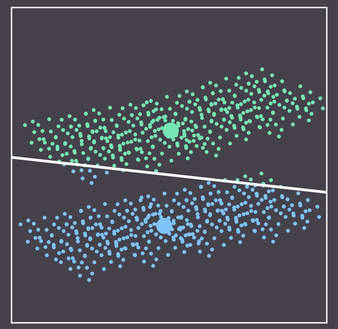


ETAPE 5

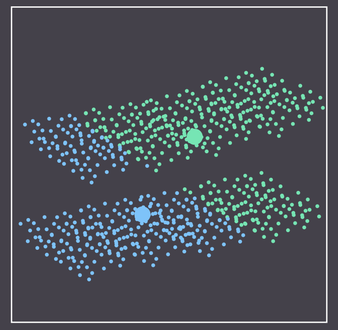


INFORMATIQUE

ETAPE 3



ETAPE 6



CONCEPT

MATHÉMATIQUES

INFORMATIQUE

- Étant donné un ensemble de points $(x_1, x_2, \ldots x_n)$, on cherche à partitioner les *n* points en *k* ensembles $S = \{S_1, S_2, \ldots, S_k\}$ $(k \leq n)$ en minimisant la distance entre les points à l'intérieur de chaque partition :
- où μ_i est la moyenne des points dans S_i

$$\arg_S \min \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2$$

DISTANCE, PAR EXEMPLE ICI, EUCLIDIENNE

CONCEPT

- Prendre un K constant, qui correspond au nombre clusters souhaité
- Initialiser ces K centroids aléatoirement
- Pour chaque point calculer le cluster le plus proche et créer des cluster avec une liste identique a celle des données mais qui a pour valeur le rang du cluster le plus proche

MATHÉMATIQUES

- Calculer la moyenne des coordonnées de chaque cluster
- Affecter cette moyenne aux coordonnées des centres de clusters

INFORMATIQUE

- Répéter cette opération jusqu'à ce que les coordonnées de chaque centroïdes à une variation inférieur à un certain seuil choisis par le développeur entre chaque répétition