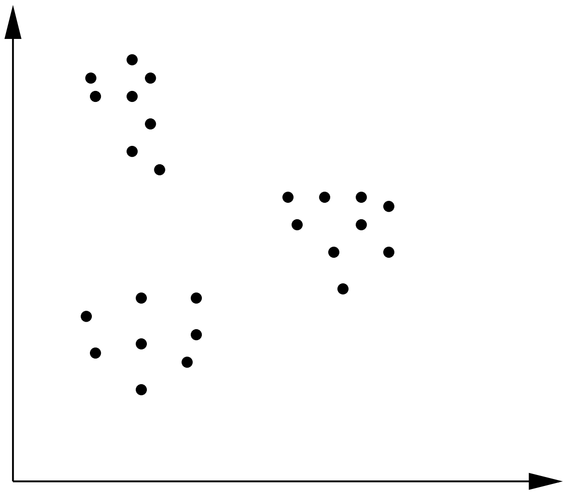
Apprentissage non-supervisé

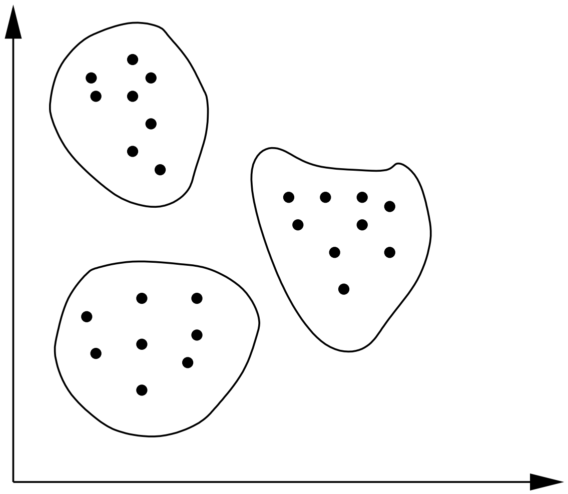
Algorithme de regroupement (“clustering”)

- Références:
  - `www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/`
  - `www.autonlab.org/tutorials/kmeans.html`
  - `www-db.stanford.edu/~ullman/cs345-notes.html`
- Introduction
- Algorithme k moyennes (k-means)
- Regroupement hiérarchique

Exemple



Exemple



Introduction

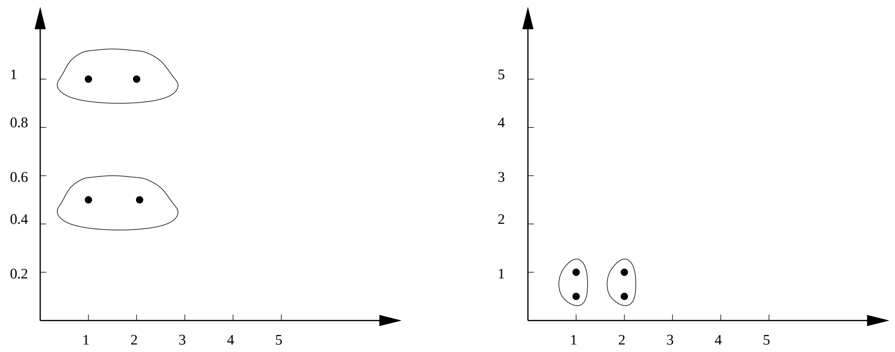
- Le but d’un algorithme de regroupement est de déterminer le groupement “intrinsèque” d’un ensemble de données.
- Qu’est-ce qu’un “bon” groupement ?
- Plein d’applications

Ingrédients de base

- Description des données: vecteurs
- Mesure de distance $d(\vec{x}, \vec{y})$
  - $d(\vec{x}, \vec{x}) = 0$
  - $d(\vec{x}, \vec{y}) \geq 0$
  - $d(\vec{x}, \vec{y}) \leq d(\vec{x}, \vec{z}) + d(\vec{z}, \vec{y})$
  - $d(\vec{x}, \vec{y}) = d(\vec{y}, \vec{x})$

Comment choisir la mesure de distance ?

- Problème: Les différentes composantes de la description ne sont pas forcément dans le même domaine
- Même si elles sont dans le même domaine, il y a un problème d'échelle.



Exemples de distances

La distance de Minkowski (pour dimension d)

$$d_p(\vec{x}, \vec{y}) = (\sum_{k=1}^{d} |x_k - y_k|^p)^{\frac{1}{p}}$$

- Pour $p = 1$: distance de Manhattan
- Pour $p = 2$: distance euclidienne

Algorithme k-moyennes (k-means)

- Un nombre de groupe ("cluster") est fixé a priori
- On fixe k barycentres initiaux
- qui doivent être définis d'une façon "intelligente".
- Le résultat dépend du choix initial
- On associe chaque point au barycentre le plus proche
- Cela définit un premier groupement
- On recalcule k nouveaux barycentre (moyenne des points qui lui sont associés), etc.

Algorithme k-moyennes

1. Placer k points dans l'espace représenté par les objets à regrouper. Ces points constituent les barycentres initiaux.
2. Associer à chaque point le groupe avec le barycentre le plus proche
3. Recalculer les positions des k barycentres
4. Répéter les étapes 2 et 3 jusqu'à-ce-que les barycentres ne changent plus
5. Le résultat est un groupement

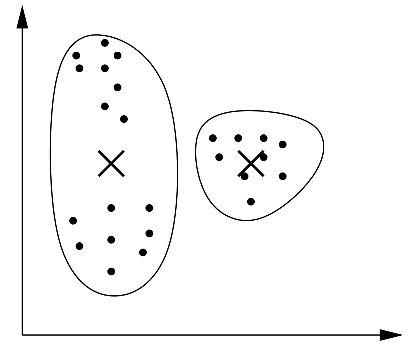
Remarques

- Comment trouver les k points initiaux ?
- Le résultat en dépend.
- Il peut arriver que l'ensemble des points les plus proche d'un barycentre est vide.
- Le résultat dépend de k. Comment le choisir ?

Comment trouver le bon k ?

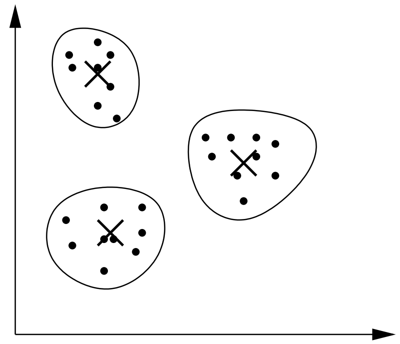
- Essayer plusieurs k en commençant avec une petite valeur en l'incrémentant
- Calculer la distance moyenne avec le barycentre
- La moyenne tombe rapidement jusqu'au "bon" k, ensuite elle ne change plus beaucoup.

Exemple



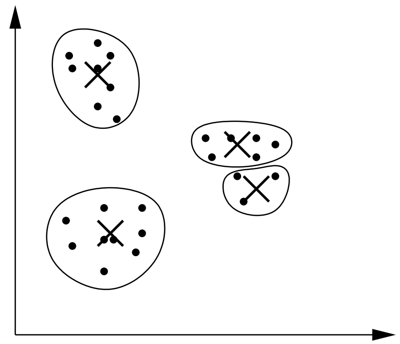
$k = 2$: moyenne des distances très grande

Exemple



$k = 3$: moyenne des distances plus petite

Exemple



$k = 4$: moyenne des distances ne diminue plus beaucoup

Algorithme de regroupement hiérarchique

1. Chaque point est considéré comme un groupe. Donc si on a n points au départ, on a n groupes. On calcule une distance entre tous les paires de groupes.

2. Trouver le paire de groupe le plus proche et regrouper ces deux groupes

3. Calculer les distances entre le nouveau groupe et les autres.

4. Répéter étapes 2 et 3 jusqu'à-ce-qu'il ne reste qu'un groupe avec n éléments.

Mesures de distance

- lien simple: la distance entre deux groupes est la distance la plus courte entre un membre d'un groupe et un membre de l'autre.

- lien complet: la distance entre deux groupes est la distance la plus grande entre un membre d'un groupe et un membre de l'autre.

- lien moyen: la distance entre deux groupes est la distance moyenne de toutes les distances entre un membre d'un groupe et un membre de l'autre.

Exemple: Regrouper des villes européennes selon leurs distances

	Francfort	Marseille	Milan	Paris	Vienne
Francfort		1004	725	592	726
Marseille			587	809	1414
Milan				850	887
Paris					1285
Vienne					