App. statistique non supervisé : la classification

C. HELBERT

Contexte :

- $X_1, ..., X_p$ sont des variables explicatives (descripteurs) observées sur n individus
- non supervisée : pas de variable de classement

Contexte :

- $X_1, ..., X_p$ sont des variables explicatives (descripteurs) observées sur n individus
- non supervisée : pas de variable de classement

Objectif du cours :

- partitionner l'espace individus/variables : rassembler les individus qui se ressemblent et/ou séparer ceux qui diffèrent

Contexte :

- $X_1, ..., X_p$ sont des variables explicatives (descripteurs) observées sur n individus
- non supervisée : pas de variable de classement

Objectif du cours :

- partitionner l'espace individus/variables : rassembler les individus qui se ressemblent et/ou séparer ceux qui diffèrent

Deux approches différentes :

- la classification hiérarchique
- les centres mobiles ("k means")

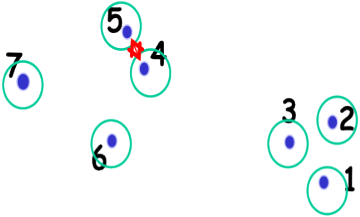
Plan

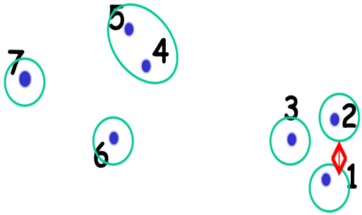
Classification hiérarchique

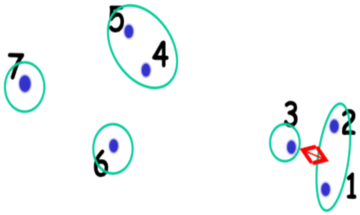
k means - centres mobiles

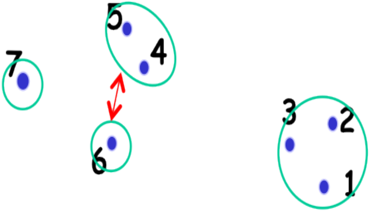
Deux types de classification hiérarchique : ascendante et descendante. Principe de la classification hiérarchique ascendante

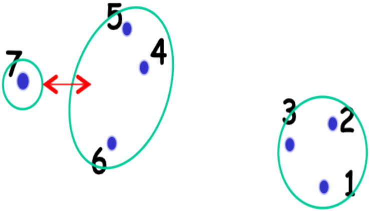
- Etat de départ : tous les individus sont séparés dans des classes distinctes (une classe = un individu)
- A chaque étape, on agrège les classes les plus proches.
- Arrêt de l'algorithme : tous les individus appartiennent à une seule classe.

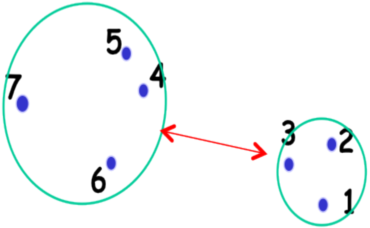

5
4
7
3
2
6
1

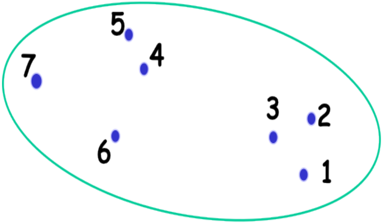

5
4
7
3
2
6
1

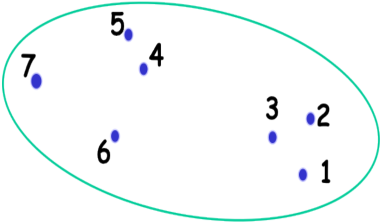

5
4
7
3
2
6
1

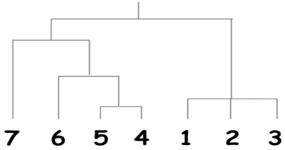

5
4
7
3
2
6
1


7
5
4
6
3
2
1


5
4
7
6
3
2
1


5
4
7
3
2
6
1


5
4
7
3
2
6
1


7 6 5 4 1 2 3

Deux types de classification hiérarchique : ascendante et descendante. Principe de la classification hiérarchique ascendante

- Etat de départ : tous les individus sont séparés dans des classes distinctes (une classe = un individu)
- A chaque étape, on agrège les classes les plus proches.
- Arrêt de l'algorithme : tous les individus appartiennent à une seule classe.

Deux types de classification hiérarchique : ascendante et descendante. Principe de la classification hiérarchique ascendante

- Etat de départ : tous les individus sont séparés dans des classes distinctes (une classe = un individu)
- A chaque étape, on agrège les classes les plus proches.
- Arrêt de l'algorithme : tous les individus appartiennent à une seule classe.

Avantage : construction d'une suite de partitions imbriquées les unes dans les autres, i.e. un arbre de classification ou **dendrogramme**.

Deux types de classification hiérarchique : ascendante et descendante. Principe de la classification hiérarchique ascendante

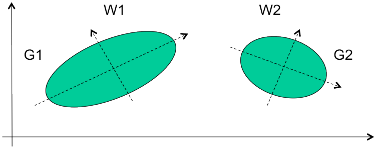
- Etat de départ : tous les individus sont séparés dans des classes distinctes (une classe = un individu)
- A chaque étape, on agrège les classes les plus proches.
- Arrêt de l'algorithme : tous les individus appartiennent à une seule classe.

Avantage : construction d'une suite de partitions imbriquées les unes dans les autres, i.e. un arbre de classification ou **dendrogramme**. ordonnée du dendrogramme = valeur du critère d'agrégation

Besoin : critère d'agrégation entre classes + distance entre individus.

Les distances usuelles pour les **variables quantitatives** :

- distance euclidienne $d(x,x')^2 = \sum_{j=1}^{p}(x_j - x'_j)^2$
- distance de Mahalanobis $d_k(x,x')^2 = {}^t(x-x')W_k^{-1}(x-x')$



- distance de Minkowski $d(x,x')^q = \sum_{j=1}^{p}(x_j - x'_j)^q$

Les distances usuelles pour les **variables qualitatives** : mesures de dissimilarité (basées sur les concordances et discordances) entre deux individus en codage disjonctif complet.

Un critère d'agrégation = critère qui quantifie la proximité entre classes.

Deux types de critères :

- critères basés sur la notion de distance
- critères basés sur la notion d'inertie

agrégation sur notion de distance

Les critères de distances les plus utilisés sont :

- **single linkage (saut minimum)** :

$$D(C_1, C_2) = \min_{x_i \in C_1, x_j \in C_2} d(x_i, x_j)$$

- **complete linkage (saut maximum)** :

$$D(C_1, C_2) = \max_{x_i \in C_1, x_j \in C_2} d(x_i, x_j)$$

- **group average (moyenne)** :

$$D(C_1, C_2) = \frac{1}{n_1 + n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} d(x_i, x_j)$$

où $n_1 = \#C_1$ et $n_2 = \#C_2$

agrégation sur notion de variance p=1

Rappel. Soit X une v.a. quantitative. Soit G une v.a. de bernoulli

$$Var(X) = \mathbb{E}(Var(X|G)) + Var(\mathbb{E}(X|G))$$

- $\mathbb{E}(Var(X|G)) = pVar(X|G=1) + (1-p)Var(X|G=0)$
- $Var(\mathbb{E}(X|G)) =$
  $p\left(\mathbb{E}(X|G=1) - \mathbb{E}(X)\right)^2 + (1-p)\left(\mathbb{E}(X|G=0) - \mathbb{E}(X)\right)^2$

Estimations :

- $\mathbb{E}(Var(X|G)) = \frac{n_1}{n}\frac{\sum_{i=1}^{n_1}(x_i - m1)^2}{n_1} + \frac{n_2}{n}\frac{\sum_{j=1}^{n_2}(x_j - m2)^2}{n_2}$
  $=>$ variance intra
- $Var(\mathbb{E}(X|G)) = \frac{n_1}{n}(m_1 - m)^2 + \frac{n_2}{n}(m_2 - m)^2$ variance inter

agrégation sur notion d'inertie p >1

Application à la classification hierarchique ascendante

- $p >> 1$ extension de la notion de variance : notion d'inertie. Inertie = somme des variances
- Passage de q classes à $q-1$ classes. Parmi les $\frac{q(q-1)}{2}$ regroupements possibles, on choisit celui qui entraine la hausse de l'inertie intra la plus faible.

$$D(C_1, C_2) = \frac{n_1+n_2}{n}\frac{\sum(x_i-m_{12})^2}{n_1+n_2} - \{\frac{n_1}{n}\frac{\sum(x_i-m_1)^2}{n_1} + \frac{n_2}{n}\frac{\sum(x_i-m_2)^2}{n_2}\}$$

Cette méthode est très populaire car elle produit un graphe particulièrement interprétable à différents niveaux.

Attention

- La suite de partitions obtenue est très sensible aux données
- Cet algorithme suppose que les données proviennent d'un modèle hiérarchique : ce n'est pas toujours vrai.

Plan

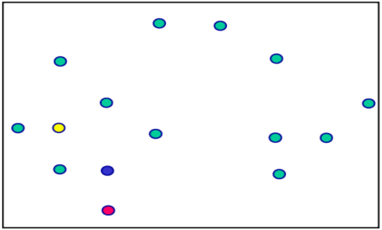
Objectif : partitionner les individus $E = \{x_1, ..., x_n\}$ en K classes, K étant connu.

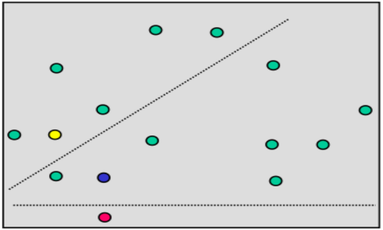
étape 0 Choix de K points parmi les n points de l'ensemble (tirage sans remise). Ces points sont appelés les noyaux et sont notés : $\{e_{01}, ..., e_{0K}\}$. Soit $\{C_{01}, ..., C_{0K}\}$ la partition de l'espace associée (diagramme de Voronoï)

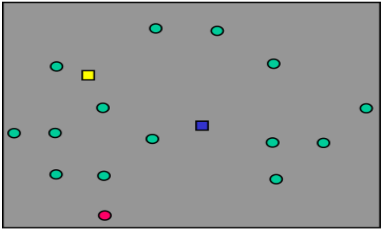
$$C_{0k} = \{x_i \in E, \forall j \in \{1, ..., K\}, d(x_i, e_{0k}) \leq d(x_i, e_{0j})\}$$

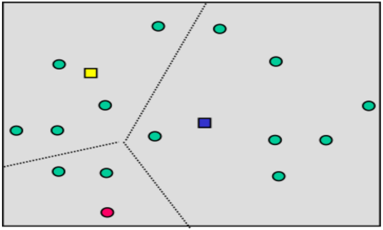
étape m A chaque étape, on met à jour les noyaux et la partition.

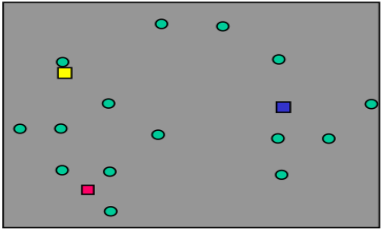
arrêt La partition n'évolue plus (partitions identiques entre deux étapes successives.

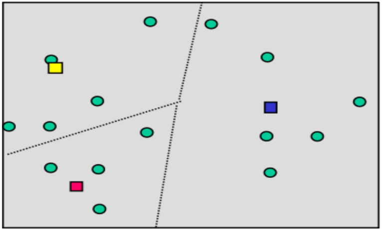


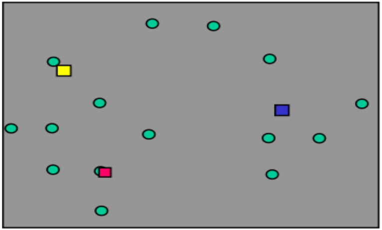


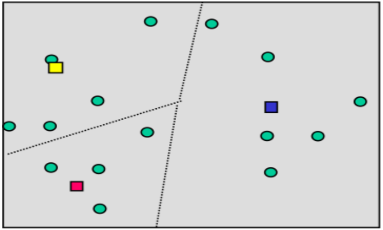


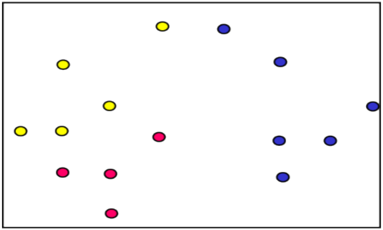




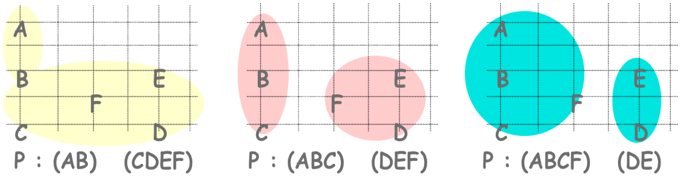








- Algorithme d'optimisation local => sensibilité à l'initialisation
- Différentes évaluations de l'algorithme : partitions résultantes départagées suivant un critère de l'inertie



- **Formes Fortes** : groupes stables d'objets, toujours classés ensemble
- **Formes Faibles** : objets rattachés tantôt à un groupe, tantôt à un autre