Apprentissage statistique : modélisation descriptive et introduction aux réseaux de neurones (RCP208)

Classification automatique

Michel Crucianu
(prenom.nom@cnam.fr)
http://cedric.cnam.fr/vertigo/Cours/ml/

EPN05 Informatique
Conservatoire National des Arts & Métiers, Paris, France

24 octobre 2024

Plan du cours

Objectifs et utilisations de la classification automatique

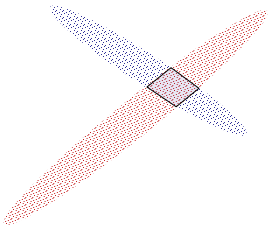
(*cluster analysis, clustering*)

- Objectif général : répartir un ensemble donné de N observations en groupes (catégories, classes, taxons, clusters) de façon à regrouper les observations similaires et à séparer les observations dissimilaires
  - Partitionnement des données, ou
  - Hiérarchie de groupes ($\rightarrow$ plusieurs partitionnements disponibles)
- Conditions :
  - Aucune information n'est disponible concernant l'appartenance de certaines données à certaines « classes »
  - Le nombre de groupes recherchés peut être connu *a priori* ou non
- Utilisations :
  - Mettre en évidence une structure (simple) dans un ensemble de données
  - Résumer un grand ensemble de données par les représentants des groupes

Typologie des méthodes de classification automatique

Choix méthode $\Leftarrow$ connaissance des données et de la nature des groupes recherchés

- Nature des données : numériques, catégorielles, mixtes
- Représentation des données :
  - Représentation vectorielle $\rightarrow$ définir centres de gravité, densités, intervalles, différentes distances $\Rightarrow$ complexité en général $O(N)$
  - Simple : seules sont disponibles les distances entre observations $\Rightarrow$ complexité $\geq O(N^2)$
- Groupes mutuellement exclusifs ou non ?
  - A quel groupe appartiennent les données entourées ?



- Nature des groupes :
  - Nets : une observation appartient ou n'appartient pas à un groupe
  - Flous : une observation peut appartenir à différents degrés à plusieurs groupes $\Rightarrow$ convergence souvent plus robuste de l'algorithme de classification
    - Groupes flous $\rightarrow$ nets : chaque observation affectée au groupe auquel elle appartient le plus

Typologie des méthodes de classification automatique (2)

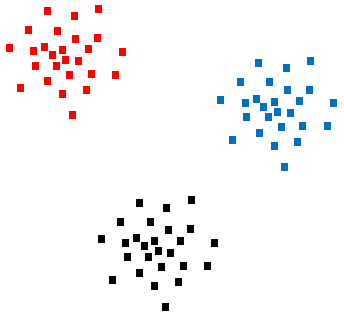
- Critère de regroupement (définition des groupes) : en général n'est pas explicite !

Généralités 4 / 25

Typologie des méthodes de classification automatique (2)

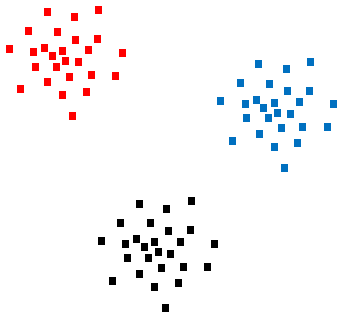
- Critère de regroupement (définition des groupes) : en général n'est pas explicite !

  - Ensembles compacts éloignés entre eux :

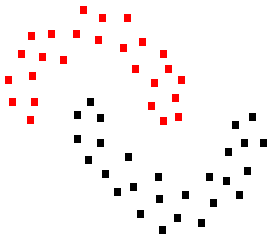


Typologie des méthodes de classification automatique (2)

- Critère de regroupement (définition des groupes) : en général n'est pas explicite !
  - Ensembles compacts éloignés entre eux :



  - Ensembles denses séparés par des régions moins denses :



Plan du cours

Centres mobiles : la méthode

- Ensemble $\mathcal{E}$ de N données décrites par p variables à valeurs dans $\mathbb{R}$
- Objectif : répartir les N données en k groupes disjoints $\mathcal{E}_1, \ldots, \mathcal{E}_k$ (inconnus a priori) en minimisant la somme des inerties intra-classe

$$\phi_{\mathcal{E}}(\mathcal{C}) = \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{E}_j} d^2(\mathbf{x}_i, \mathbf{m}_j) \tag{1}$$

avec

- $\mathcal{C} = \{\mathbf{m}_j, 1 \leq j \leq k\}$ l'ensemble des centres des k groupes
- d la distance dans $\mathbb{R}^p$ qui définit la nature des dissimilarités

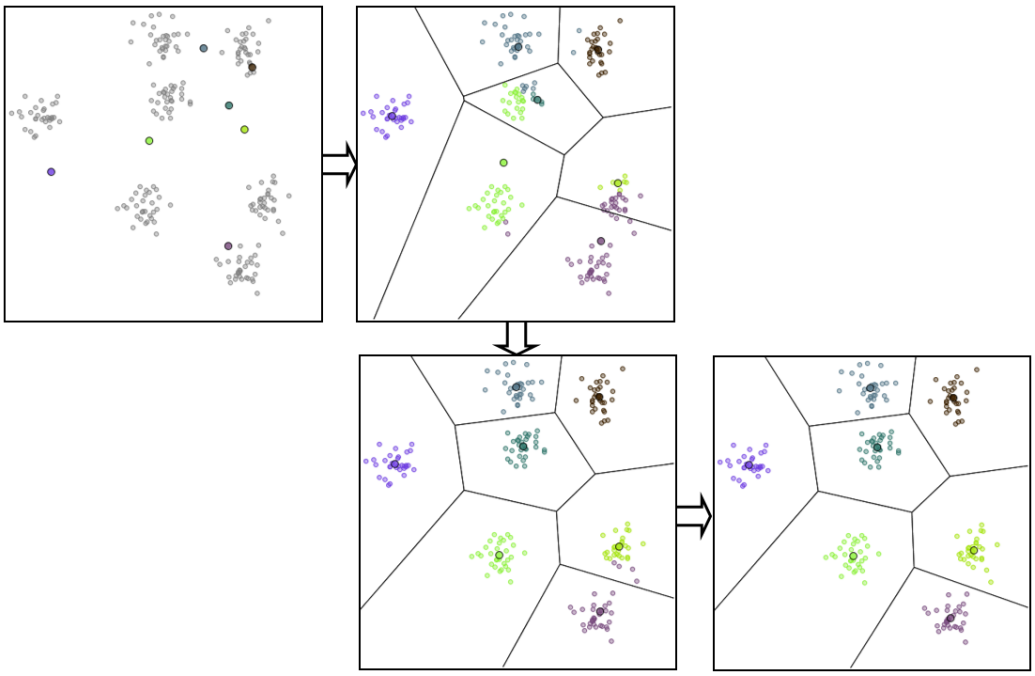
- La somme des inerties intra-classe peut s'écrire aussi

$$\phi_{\mathcal{E}}(\mathcal{C}) = \sum_{1 \leq l \leq N} d^2(\mathbf{x}_l, \mathbf{m}_{C(l)}) \tag{2}$$

où $C(l)$ est l'indice du groupe dont fait partie $\mathbf{x}_l$

Centres mobiles : illustration

(données issues de 7 lois normales bidimensionnelles, classification avec 7 centres)



Centres mobiles : l'algorithme

Data : Ensemble $\mathcal{E} = \{\mathbf{x}_i\}_{1 \leq i \leq N}$ de N données de $\mathbb{R}^p$

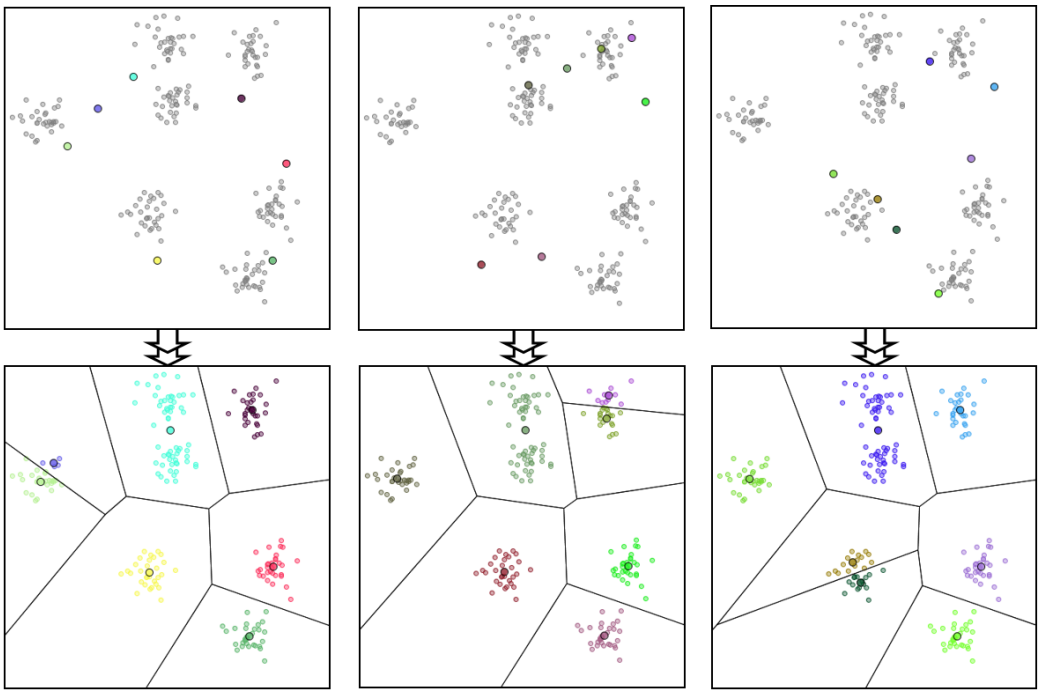
Result : k groupes (*clusters*) disjoints $\mathcal{E}_1, \ldots, \mathcal{E}_k$ et ensemble $\mathcal{C}$ de leurs centres

1 Initialisation aléatoire des centres $\mathbf{m}_j$, $1 \leq j \leq k$;
2 **while** *centres non stabilisés* **do**
3 | Affectation de chaque donnée au groupe du centre le plus proche ;
4 | Remplacement des anciens centres par les centres de gravité des groupes ;
5 **end**

- $\phi_{\mathcal{E}}(\mathcal{C})$ diminue lors de chacune des deux étapes du processus itératif ; comme $\phi_{\mathcal{E}}(\mathcal{C}) \geq 0$, le processus itératif doit converger
- ... mais la solution obtenue sera un minimum *local*, dépendant de l'initialisation, souvent beaucoup moins bon que le minimum global

Centres mobiles : illustration (2)

(résultats avec 3 initialisations différentes)



→ Faire tourner l'algorithme plusieurs fois, à partir d'initialisations aléatoires différentes, ne garantit pas d'arriver à une bonne solution !

Centres mobiles : convergence

$\phi_{\mathcal{E}}(\mathcal{C})$ diminue de façon monotone non stricte à chaque étape de chaque itération :

1. Affectation de chaque donnée au groupe du centre le plus proche : $\mathbf{x}_i$ passe du groupe de centre $\mathbf{m}_p$ au groupe de centre $\mathbf{m}_q$ si $d^2(\mathbf{x}_i, \mathbf{m}_p) > d^2(\mathbf{x}_i, \mathbf{m}_q)$, donc

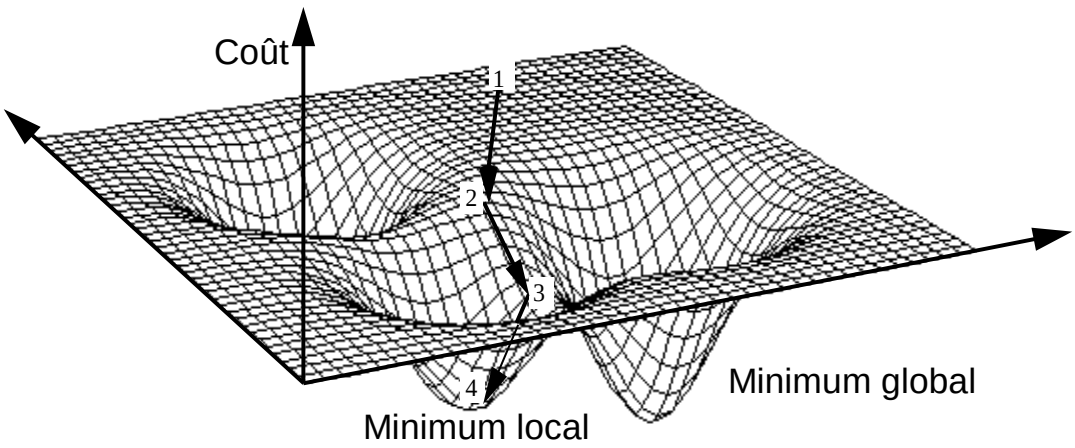
$$d^2(\mathbf{x}_i, \mathbf{m}_p) + \sum_{l \neq i} d^2(\mathbf{x}_l, \mathbf{m}_{C(l)}) > d^2(\mathbf{x}_i, \mathbf{m}_q) + \sum_{l \neq i} d^2(\mathbf{x}_l, \mathbf{m}_{C(l)})$$

2. Remplacement des anciens centres par les centres de gravité des groupes : si $\widetilde{\mathbf{m}}_j$ est l'ancien centre du groupe j et $\mathbf{m}_j$ le nouveau, alors

$$\begin{aligned}
\sum_{\mathbf{x}_i \in \mathcal{E}_j} d^2(\mathbf{x}_i, \widetilde{\mathbf{m}}_j) &= \sum_{\mathbf{x}_i \in \mathcal{E}_j} \|\mathbf{x}_i - \mathbf{m}_j + \mathbf{m}_j - \widetilde{\mathbf{m}}_j\|^2 \\
&= \sum_{\mathbf{x}_i \in \mathcal{E}_j} \|\mathbf{x}_i - \mathbf{m}_j\|^2 + \sum_{\mathbf{x}_i \in \mathcal{E}_j} \|\mathbf{m}_j - \widetilde{\mathbf{m}}_j\|^2 + 2\,(\mathbf{m}_j - \widetilde{\mathbf{m}}_j)^T \underbrace{\sum_{\mathbf{x}_i \in \mathcal{E}_j} (\mathbf{x}_i - \mathbf{m}_j)}_{=0} \\
&\geq \sum_{\mathbf{x}_i \in \mathcal{E}_j} \|\mathbf{x}_i - \mathbf{m}_j\|^2 \left(= \sum_{\mathbf{x}_i \in \mathcal{E}_j} d^2(\mathbf{x}_i, \mathbf{m}_j) \right)
\end{aligned}$$

Centres mobiles : convergence (2)

- Minimisation itérative d'une fonction de deux variables, différentiable :



- Contrairement au cas illustré ci-dessus, $\phi_{\mathcal{E}}(\mathcal{C})$ n'est pas différentiable (ni même continue) par rapport aux $\mathbf{m}_j$ $\Leftarrow$ un changement infinitésimal dans la position d'un centre peut provoquer un changement d'affectation de données aux centres et donc un changement significatif (non infinitésimal) de la valeur de $\phi_{\mathcal{E}}(\mathcal{C})$

K-means : l'algorithme *online* de [2]

- *K-means* de [2] est une variante *online* (non *batch*) de la méthode des centres mobiles ; souvent, *K-means* est utilisé comme synonyme des centres mobiles...

Data : Ensemble $\mathcal{E}$ de N données de $\mathbb{R}^p$
Result : k groupes (*clusters*) disjoints $\mathcal{E}_1, \ldots, \mathcal{E}_k$ et ensemble $\mathcal{C}$ de leurs centres

1 Initialisation aléatoire des centres $\mathbf{m}_j$, $1 \leq j \leq k$;
2 **while** *centres non stabilisés* **do**
3 Choix aléatoire d'une des données ;
4 Affectation de la donnée au groupe du centre le plus proche ;
5 Recalcul des centres pour le groupe que la donnée vient de rejoindre et celui qu'elle vient de quitter ;
6 **end**

- Recalcul du centre j rejoint par la donnée i : $\mathbf{m}_j = \frac{1}{n_j}(\widetilde{n}_j \widetilde{\mathbf{m}}_j + \mathbf{x}_i)$, avec $n_j = \widetilde{n}_j + 1$
- Recalcul du centre l quitté par la donnée i : $\mathbf{m}_l = \frac{1}{n_l}(\widetilde{n}_l \widetilde{\mathbf{m}}_l - \mathbf{x}_i)$, avec $n_l = \widetilde{n}_l - 1$
- Intermédiaire entre *batch* et *online* : à chaque itération un échantillon de b données $\rightarrow$ *mini-batch* (de taille b)

Initialisation *K-means* : *K-means++*

- Une bonne initialisation de l'algorithme *K-means*
  - permet d'obtenir une solution de meilleure qualité et
  - une convergence plus rapide (avec moins d'itérations) vers cette solution
- Parmi les nombreux algorithmes d'initialisation nous considérerons *K-means++* [1]
- Idée : choisir les centres successivement, suivant une loi non uniforme qui privilégie les candidats éloignés des centres déjà sélectionnés

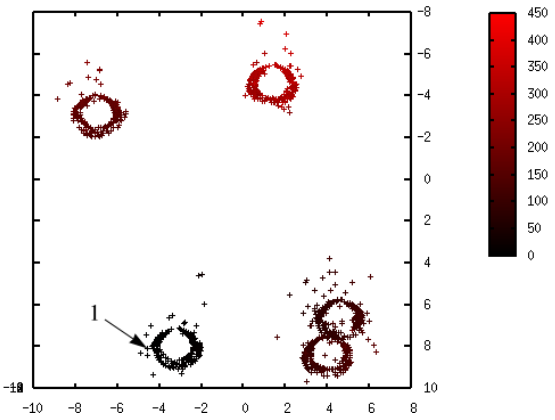
Data : Ensemble $\mathcal{E}$ de N données de $\mathbb{R}^p$; nombre souhaité de centres k
Result : $\mathcal{C} = \{\mathbf{c}_j, 1 \leq j \leq k\}$

1 $\mathcal{C} \leftarrow$ un $\mathbf{x}$ de $\mathcal{E}$ choisi au hasard ;
2 **while** $\|\mathcal{C}\| \leq k$ **do**
3 Sélectionner $\mathbf{x} \in \mathcal{E}$ avec la probabilité $\frac{d^2(\mathbf{x},\mathcal{C})}{\phi_{\mathcal{E}}(\mathcal{C})}$;
4 $\mathcal{C} \leftarrow \mathcal{C} \cup \{\mathbf{x}\}$;
5 **end**

- Notations : $d^2(\mathbf{x}, \mathcal{C}) = \min_{j=1,\ldots,t} d^2(\mathbf{x}, \mathbf{c}_j)$, $\phi_{\mathcal{E}}(\mathcal{C}) = \sum_{\mathbf{x} \in \mathcal{E}} d^2(\mathbf{x}, \mathcal{C})$

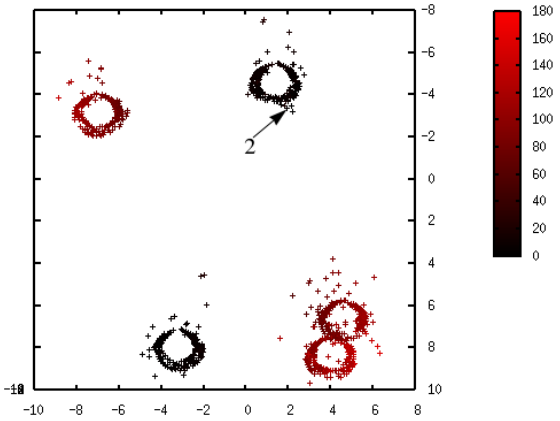
K-means++ : évolution des probabilités

(probabilité de sélection proportionnelle à $d^2(\mathbf{x}, \mathcal{C})$, représentée par la couleur rouge)



Après la sélection d'un point

$$\mathcal{C} = \left\{ \begin{pmatrix} -4,6 \\ 8,0 \end{pmatrix} \right\}$$

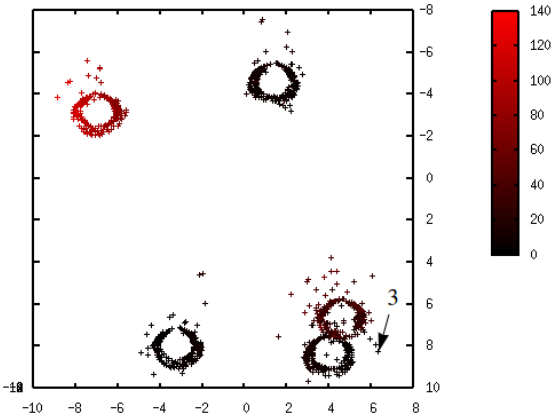


Après la sélection de 2 points

$$\mathcal{C} = \left\{ \begin{pmatrix} -4,6 & 2,15 \\ 8,0 & -3,45 \end{pmatrix} \right\}$$

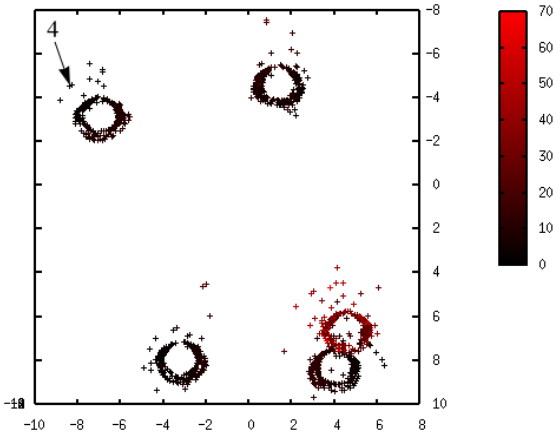
K-means++ : évolution des probabilités (2)

(probabilité de sélection proportionnelle à $d^2(\mathbf{x}, \mathcal{C})$, représentée par la couleur rouge)



Après la sélection de 3 points

$$\mathcal{C} = \left\{ \begin{pmatrix} -4,6 & 2,15 & 6,32 \\ 8,0 & -3,45 & 8,22 \end{pmatrix} \right\}$$



Après la sélection de 4 points

$$\mathcal{C} = \left\{ \begin{pmatrix} -4,6 & 2,15 & 6,32 & -8,37 \\ 8,0 & -3,45 & 8,22 & -4,54 \end{pmatrix} \right\}$$

K-means : intérêt et limitations

- Intérêt (au-delà de la simplicité) :
  - Paramètre unique : valeur souhaitée pour le nombre de groupes
  - Faible complexité moyenne : $O(tkN)$ (avec t le nombre d'itérations)
- Limitations et solutions :
  - Données vectorielles uniquement (pour calculer les moyennes)
    → limitation levée dans des méthodes dérivées (ex. *k-medoids*)
  - Classes de forme sphérique (si la distance euclidienne usuelle est employée)
    → pour autres formes, on peut se servir de la distance de Mahalanobis (calculée par classe)
  - Dépendance des conditions initiales (car convergence vers minimum local)
    → initialisation évoluée (par ex. *K-means++*)
  - Sensibilité aux données aberrantes
    → fonctionnelle de coût robuste, estimation robuste des moyennes
  - Choix *a priori* difficile du nombre de classes
    → régularisation, sélection de modèle

Plan du cours

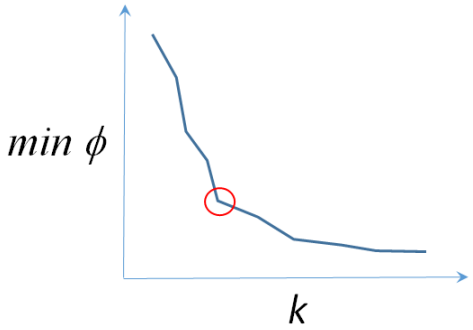
Méthode des *k-medoids*

- Objectif : traiter des données non vectorielles, pour lesquelles seule une métrique d est connue, tout en conservant la simplicité des centres mobiles
- *Medoid* d'un groupe = individu le plus « central » du groupe
- Le seul changement par rapport à *K-means* est le remplacement des centres de gravité par des *medoids*
- A chaque itération :
  1. Choix, pour chaque donnée, du *medoid* $\mathbf{m}_{C(l)}$ le plus proche : $C(l) = \arg\min_j d(\mathbf{x}_l, \mathbf{m}_j)$
  2. Constitution des groupes : $\mathcal{E}_j$ est constitué de tous les $\mathbf{x}_l$ qui sont plus proches de $\mathbf{m}_j$ que de tout autre *medoid*
  3. Recherche des *medoids* de ces (nouveaux) groupes : $\mathbf{m}_j = \arg\min_{\mathbf{x}_l \in \mathcal{E}_j} \sum_{\mathbf{x}_p \in \mathcal{E}_j} d(\mathbf{x}_l, \mathbf{x}_p)$
- Une initialisation de même nature que *K-means++* peut être employée
- Robustesse apportée par l'utilisation de *medoids* plutôt que de centres de gravité
- Mais complexité $O(N^2)$!

Plan du cours

Choix du nombre de groupes k

1. Méthode du « coude » : graphique des valeurs minimales atteintes par $\phi_{\mathcal{E}}(\mathcal{C})$ pour k croissant, choix de valeur de k avant un palier



2. Mise de la méthode de classification dans un cadre probabiliste et choix d'un critère d'information comme AIC (Akaike), BIC (Bayes), etc.
3. Stabilité des résultats : une valeur de k est meilleure si les groupes obtenus sont plus « stables » à l'initialisation aléatoire (voir par ex. [3])

Comparaison de classifications

- Deux méthodes différentes, ou deux initialisations différentes pour une même méthode, produisent (approximativement) les mêmes groupes ?
  - Construction de classifications « consensuelles »
  - Évaluation de la stabilité (informe sur l'adéquation de la méthode aux données et même sur la présence de groupes dans les données)
- Parmi les propositions (voir par ex. [5], [4]) :
  - Indice de Rand ajusté : pour deux classifications $\mathcal{C}, \mathcal{C}'$,
    - n_{11} nombre de paires qui sont dans un même groupe suivant $\mathcal{C}$ et $\mathcal{C}'$
    - n_{00} nombre de paires qui sont dans des groupes différents suivant $\mathcal{C}$ et $\mathcal{C}'$
    - n_{10} nombre de paires dans un même groupe suivant $\mathcal{C}$ et groupes différents suivant $\mathcal{C}'$
    - n_{01} nombre de paires dans un même groupe suivant $\mathcal{C}'$ et groupes différents suivant $\mathcal{C}$

$$\mathcal{R}(\mathcal{C},\mathcal{C}') = \frac{2(n_{11}+n_{00})}{N(N-1)}, \quad 0 \leq \mathcal{R} \leq 1, \quad \mathcal{R}_{adj}(\mathcal{C},\mathcal{C}') = \frac{\mathcal{R} - E(\mathcal{R})}{\max(\mathcal{R}) - E(\mathcal{R})}$$

  - Utilisation de l'indice de Jaccard : $\mathcal{I}(\mathcal{C},\mathcal{C}') = \frac{n_{11}}{n_{11}+n_{10}+n_{01}}$ (une classification est définie comme l'ensemble des paires d'observations qui sont dans un même groupe, parmi toutes les paires possibles)
  - Information mutuelle normalisée, etc.

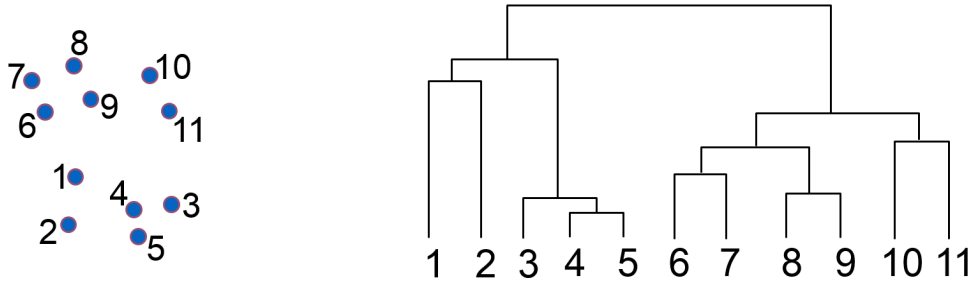
Validité de la classification

- L'algorithme converge vers un résultat quelles que soient les données, mais quelle est la validité de ce résultat ?
- Principales questions :
  - Y a-t-il réellement des regroupements « naturels » dans les données ?
  - Validation externe : les groupes identifiés sont-ils en accord avec nos (éventuelles) connaissances *a priori* du problème ?
    - → Ces connaissances ne sont pas nécessairement directement exploitables dans une fonctionnelle à minimiser.
  - Validation interne : les groupes identifiés sont-ils bien « ajustés » aux données ?
    - → Nombreux indices : statistique modifiée de Hubert (alignement entre distance et partition), indice Davies-Bouldin (rapport des inerties), silhouette, etc. Mais les propriétés des données (groupes plus ou moins séparables) ont un impact !
  - Validation relative : les résultats de la méthode A sont-ils meilleurs que les résultats de la méthode B ?
    - → Possibilités de sélection de modèle, utilisant les indices de validation interne (car sur les mêmes données seul l'ajustement compte), la stabilité (voir par ex. [3]), etc.

Plan du cours

Classification ascendante hiérarchique (CAH)

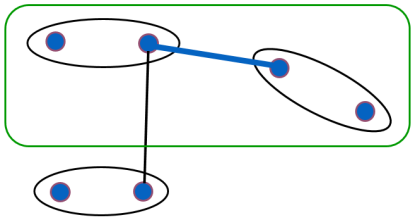
- Objectif : obtenir une hiérarchie de groupes, structure plus riche qu'un simple partitionnement
- Permet d'examiner l'ordre des agrégations de groupes, les rapports des similarités entre groupes, etc.
- Classification ascendante : procède par agrégation des données et des groupes



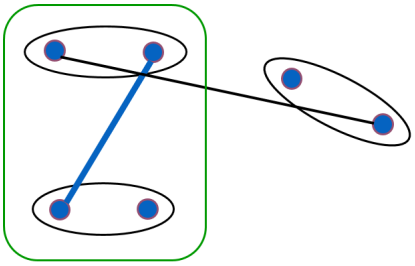
CAH : indices d'agrégation

- Sur la base de la distance entre données, $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, différents indices d'agrégation peuvent être utilisés pour mesurer la dissimilarité entre groupes :

$$\delta_s(h_p, h_q) = \min_{x_i \in h_p, x_j \in h_q} d_{\mathcal{X}}(x_i, x_j) \qquad \delta_s(h_p, h_q) = \max_{x_i \in h_p, x_j \in h_q} d_{\mathcal{X}}(x_i, x_j)$$



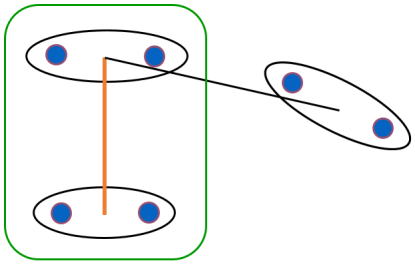
lien minimum (*single linkage*)



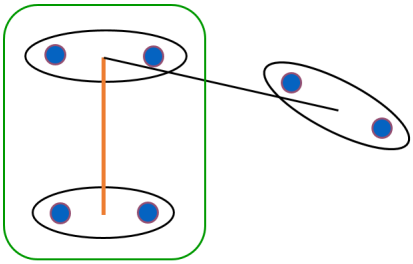
lien maximum (*complete linkage*)

CAH : indices d'agrégation (2)

$$\delta_s(h_p, h_q) = \frac{1}{\|h_p\| \cdot \|h_q\|} \sum_{x_i \in h_p, x_j \in h_q} d_{\mathcal{X}}(x_i, x_j) \qquad \delta_s(h_p, h_q) = \frac{\|h_p\| \cdot \|h_q\|}{\|h_p\| + \|h_q\|} d^2_{\mathcal{X}}(\mathbf{m}_p, \mathbf{m}_q)$$



lien moyen (*average linkage*)



indice de Ward (données vectorielles !)

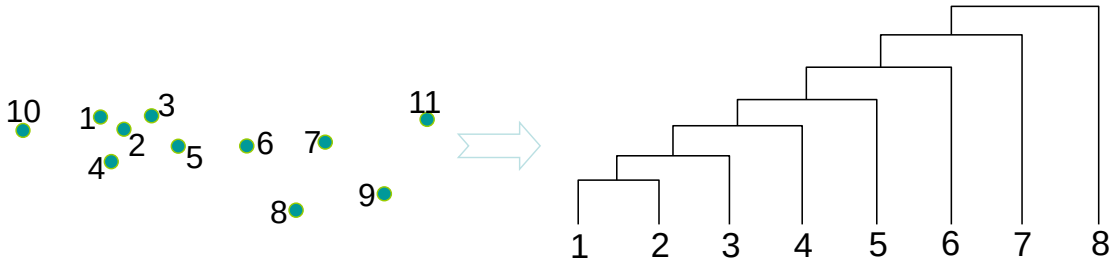
CAH : algorithme, mise en œuvre

Data : Ensemble $\mathcal{E}$ de N données de $\mathcal{X}$ muni de la distance $d_{\mathcal{X}}$
Result : Hiérarchie de groupes (dendrogramme)
1 Chaque donnée définit un groupe ;
2 **while** *nombre de groupes* > 1 **do**
3 Calcul indices d'agrégation entre tous les groupes issus de l'itération précédente ;
4 Regroupement des 2 groupes ayant la plus petite valeur de l'indice d'agrégation ;
5 **end**

- Complexité algorithmique $O(N^2 \log N)$!
- N élevé : application de *K-means* avec k élevé (mais $k \ll N$), ensuite application de la CAH sur les groupes obtenus par *K-means*

CAH : effet des différents indices

- Indice du lien minimum :
  - Permet de s'approcher d'un critère de regroupement basé sur la densité
  - Peut facilement créer des arbres en escalier, déséquilibrés et peu exploitables :



- Indice du lien maximum, indice de Ward : tiennent compte de la compacité des groupes résultants, arbres plus équilibrés

Références I

D. Arthur and S. Vassilvitskii.

K-means++ : The advantages of careful seeding.

In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

J. B. MacQueen.

Some methods for classification and analysis of multivariate observations.

In L. M. L. Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

O. Shamir and N. Tishby.

Stability and model selection in k-means clustering.

Machine Learning, 80(2) :213–243, 2010.

N. X. Vinh, J. Epps, and J. Bailey.

Information theoretic measures for clusterings comparison : Variants, properties, normalization and correction for chance.

J. Mach. Learn. Res., 11 :2837–2854, Dec. 2010.

S. Wagner and D. Wagner.

Comparing Clusterings – An Overview.

Technical Report 2006-04, Universität Karlsruhe (TH), 2007.