

Clustering : inertie, k-means, classification hiérarchique

Cours 6 – Modèles en biologie

Christophe Ambroise

L3 GBI – Université d'Évry

Norme, produit scalaire, Pythagore

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_p^2}, \quad \mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y}, \quad \|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x}$$

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + 2\mathbf{x} \cdot \mathbf{y} + \|\mathbf{y}\|^2$$

Pythagore : $\mathbf{x} \perp \mathbf{y} \Rightarrow \|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$.

Acquis des cours précédents

Centrage

Projection de $\mathbf{x}$ sur $\mathbf{1} : \bar{x} \mathbf{1}$.

Vecteur centré $\mathbf{x} - \bar{x} \mathbf{1}$: orthogonal à $\mathbf{1}$, soit $\sum_i (x_i - \bar{x}) = 0$.

Distances et réduction

Distance : positive, séparante, symétrique, inégalité triangulaire ; d_1, d_2, d_∞ sur $\mathbb{R}^p$.

Écart-type : $s_j = \frac{1}{\sqrt{n}} \|\mathbf{x}^j - \bar{x}_j \mathbf{1}\|$. Réduction : $x_{ij}^* = x_{ij}/s_j$, distances indépendantes des unités.

Hamming et matrice des distances

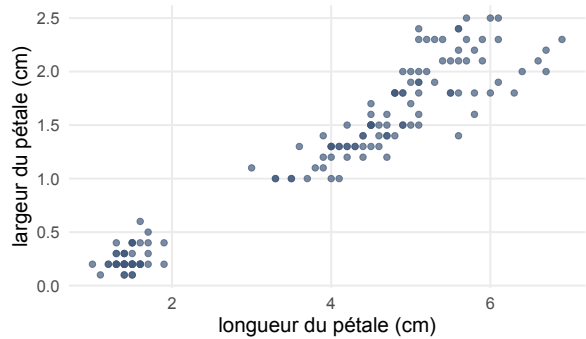
Hamming : nombre de positions différentes entre deux séquences.

	A	B	C	D	E
A	0	3	5	7	10
B	3	0	2	6	9
C	5	2	0	8	11
D	7	6	8	0	4
E	10	9	11	4	0

Matrice des distances D des espèces A–E : symétrique, diagonale nulle.

Le problème de classification

Groupes sans étiquette



150 fleurs d’iris, deux mesures. Combien de groupes ?

Iris d’Anderson (1935), rendues célèbres par Fisher (1936) ; jeu [iris](#) de R.

Partition en K classes

Un découpage des individus en K groupes non vides, sans recouvrement, qui couvrent tout l'ensemble.

$$C_1, \dots, C_K \neq \emptyset, \quad C_k \cap C_l = \emptyset \ (k \neq l), \quad C_1 \cup \dots \cup C_K = \{1, \dots, n\}$$

Nombre de partitions

Partitions de n individus en 2 classes : $2^{n-1} - 1$.

26 individus : **33 554 431** partitions. 150 fleurs en 3 classes : environ 6×10^{70} .

Conséquence

Un critère et un algorithme, pas une énumération.

Deux familles de méthodes

Partition directe : K fixé, une partition améliorée pas à pas (k-means).

Hiérarchie : des partitions emboîtées, de n classes à une seule (classification ascendante hiérarchique).

Inertie

Centre de gravité

Le point moyen du nuage : la moyenne des individus.

$$\mathbf{g} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \mathbf{x}_1 + \cdots + \frac{1}{n} \mathbf{x}_n$$

Une combinaison linéaire des individus. Coordonnée j : la moyenne $\bar{x}_j$ de la colonne j de $\mathbf{X}$.

Inertie et variances

$\|\mathbf{x}_i - \mathbf{g}\|^2 = \sum_j (x_{ij} - \bar{x}_j)^2$; échange des deux sommes :

$$I = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 = \sum_{j=1}^p s_j^2$$

Variance

s_j^2 , carré de l'écart-type : la **variance** de la variable j . L'inertie est la somme des variances.

Inertie d'un nuage

La moyenne des carrés des distances au centre de gravité.

$$I = \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{g}\|^2$$

Nuage resserré : inertie faible. Nuage étalé : inertie forte.

Inerties intra-classe et inter-classe

Classe C_k : effectif n_k , centre de gravité $\mathbf{g}_k$. Dispersion dans les classes, et dispersion des centres de classe.

$$W = \frac{1}{n} \sum_{k=1}^K \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2, \quad B = \frac{1}{n} \sum_{k=1}^K n_k \|\mathbf{g}_k - \mathbf{g}\|^2$$

Théorème de Huygens

L'inertie totale est la somme de l'inertie intra-classe et de l'inertie inter-classe.

$$I = W + B$$

Vrai pour toute partition du nuage.

Exemple en dimension 1

Quatre valeurs 1, 2, 6, 7 ; classes {1, 2} et {6, 7}.

$$g = 4, \quad g_1 = 1,5, \quad g_2 = 6,5$$

$$I = \frac{1}{4} (9 + 4 + 4 + 9) = 6,5$$

$$W = \frac{1}{4} (0,25 + 0,25 + 0,25 + 0,25) = 0,25$$

$$B = \frac{1}{4} (2 \times 6,25 + 2 \times 6,25) = 6,25$$

Vérification : $6,5 = 0,25 + 6,25$.

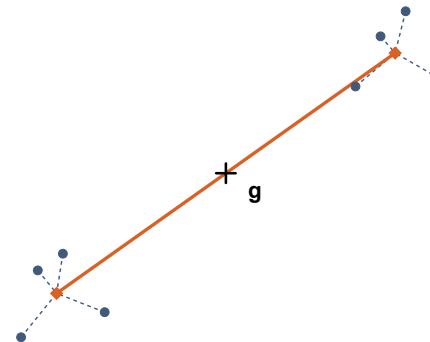
Démonstration de Huygens

Pour $i \in C_k$: $\mathbf{x}_i - \mathbf{g} = (\mathbf{x}_i - \mathbf{g}_k) + (\mathbf{g}_k - \mathbf{g})$.

$$\begin{aligned} \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}\|^2 &= \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2 + n_k \|\mathbf{g}_k - \mathbf{g}\|^2 \\ &\quad + 2(\mathbf{g}_k - \mathbf{g}) \cdot \underbrace{\sum_{i \in C_k} (\mathbf{x}_i - \mathbf{g}_k)}_{=0} \end{aligned}$$

Terme croisé nul, comme dans Pythagore : les écarts au centre d'une classe se compensent. Somme sur k , division par n . $\square$

Huygens en image



Pointillés bleus : écarts intra-classe (W). Traits orange : écarts des centres de classe au centre $\mathbf{g}$ (B).

Critère de classification

Inertie totale I fixée par les données :

$$\text{minimiser } W \iff \text{maximiser } B$$

Part d'inertie expliquée

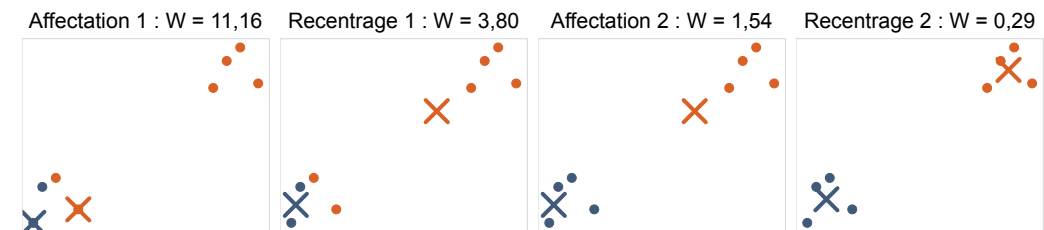
$B/I \in [0, 1]$: part de la dispersion portée par les écarts entre classes.
Exemple en dimension 1 : $6,25/6,5 \approx 96\%$.

Algorithme des k-means

1. Choisir K centres.
2. **Affectation** : chaque individu rejoint le centre le plus proche.
3. **Recentrage** : chaque centre devient le centre de gravité de sa classe.
4. Répéter 2 et 3 jusqu'à ce que la partition ne change plus.

Algorithme des k-means

k-means pas à pas



Huit points, $K = 2$, deux centres de départ placés à gauche. Croix : les centres.
 W décroît à chaque étape.

Affectation et inertie intra-classe

Centres fixés : chaque individu rejoint le centre le plus proche.

Chaque terme $\|\mathbf{x}_i - \mathbf{c}_k\|^2$ diminue ou reste égal : W ne peut que diminuer.

Convergence des k-means

L'algorithme s'arrête après un nombre fini d'itérations.

W décroît strictement tant que la partition change
nombre de partitions fini

Démonstration : aucune partition ne revient deux fois, et elles sont en nombre fini.

□

W joue le rôle d'une **fonction de Lyapunov** : une quantité qui décroît à chaque changement et ne peut décroître indéfiniment.

Scott E. Page, *The Model Thinker*, ch. 16. Départage des égalités : un individu à égale distance de deux centres garde sa classe.

Recentrage et inertie intra-classe

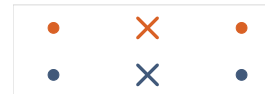
Dans une classe, le centre de gravité est le point le plus proche de tous ses individus au sens des carrés.

$$\sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{c}\|^2 = \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2 + n_k \|\mathbf{g}_k - \mathbf{c}\|^2 \geq \sum_{i \in C_k} \|\mathbf{x}_i - \mathbf{g}_k\|^2$$

Même calcul que Huygens, avec un point $\mathbf{c}$ quelconque. Remplacer $\mathbf{c}_k$ par $\mathbf{g}_k$: W diminue ou reste égal.

Minimum local

Partition stable : $W = 4$



Partition optimale : $W = 0,25$

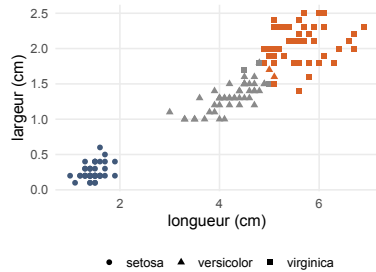


Départ haut/bas : partition stable, mais pas optimale.

Remède

Plusieurs départs ; garder la partition de plus petit W .

k-means sur les iris



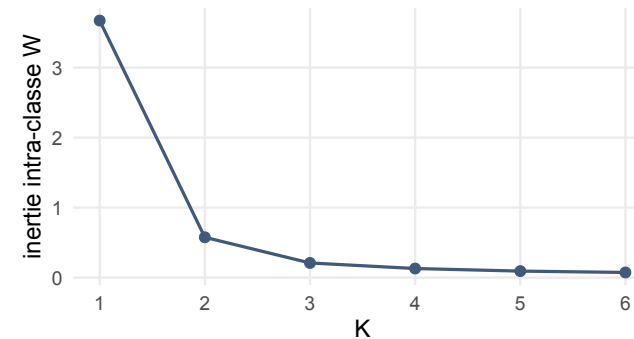
$K = 3$, 25 départs.

	setosa	versicolor	virginica
classe 1	50	0	0
classe 2	0	48	4
classe 3	0	2	46

144 fleurs sur 150 regroupées avec leur espèce, sans que l'espèce soit fournie.

$B/I \approx 94\%$.

Choix du nombre de classes



W décroît toujours quand K augmente. Un **coudé** vers $K = 2$ ou 3 : au-delà, peu de gain.

Classification ascendante hiérarchique

Classification ascendante hiérarchique

1. Partir de n classes, un individu par classe.
2. Fusionner les deux classes les plus proches.
3. Répéter jusqu'à une seule classe : $n - 1$ fusions.

Une distance entre individus ne suffit pas : il faut une distance entre **classes**.

Distance entre deux classes

Trois façons de résumer les distances entre les individus de deux classes : la plus petite, la plus grande, la moyenne.

saut minimal : $\min_{i \in C, j \in C'} d_{ij}$

saut maximal : $\max_{i \in C, j \in C'} d_{ij}$

moyenne : $\frac{1}{|C| |C'|} \sum_{i \in C} \sum_{j \in C'} d_{ij}$

Critère de Ward : fusionner les deux classes qui augmentent le moins W .

Fusions suivantes

Fusion de A et BC, à la distance 3 :

	ABC	D	E
ABC	0	6	9
D	6	0	4
E	9	4	0

Fusion de D et E à 4, puis de ABC et DE à 6.

Hauteurs des fusions : 2, 3, 4, 6.

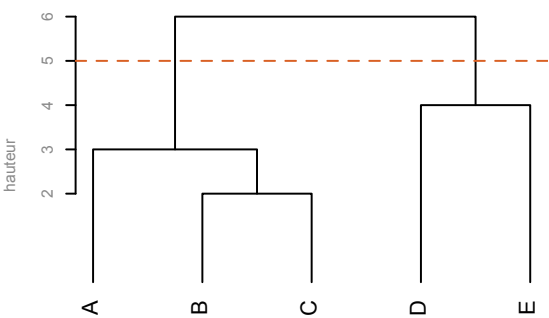
Première fusion

Plus petit terme hors diagonale : $d(B, C) = 2$. Fusion de B et C.

Saut minimal : $d(\{B, C\}, X) = \min(d(B, X), d(C, X))$.

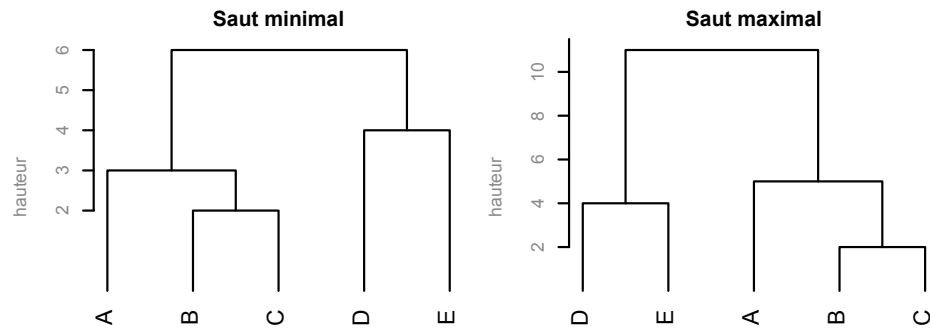
	A	BC	D	E
A	0	3	7	10
BC	3	0	6	9
D	7	6	0	4
E	10	9	4	0

Dendrogramme



Hauteur d'une fusion : distance entre les deux classes fusionnées. Coupe à 5 : deux classes, $\{A, B, C\}$ et $\{D, E\}$.

Saut minimal et saut maximal



Même matrice, hauteurs différentes : 2, 3, 4, 6 contre 2, 4, 5, 11. En trois classes : $\{A, B, C\}$, $\{D\}$, $\{E\}$ contre $\{A\}$, $\{B, C\}$, $\{D, E\}$.

37

UPGMA

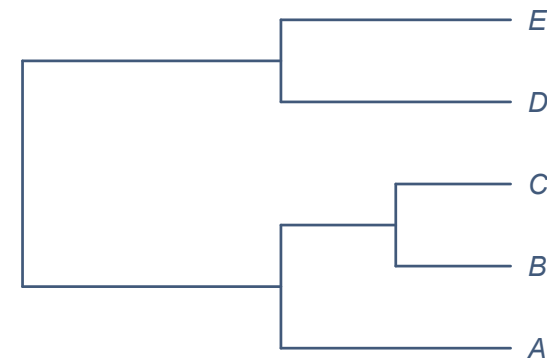
Classification hiérarchique au saut moyen : la méthode UPGMA de la phylogénie.

Espèces A–E : fusion de B et C à 2 ; puis, à la même distance 4, fusion de A avec BC ($\frac{3+5}{2}$) et de D avec E.

Dernière fusion : $\frac{1}{6} (7 + 10 + 6 + 9 + 8 + 11) = 8,5$.

UPGMA : *Unweighted Pair Group Method with Arithmetic mean*.

Arbre UPGMA des espèces A–E



Un arbre enraciné construit à partir des seules distances entre séquences.

39

40

Horloge moléculaire

UPGMA suppose un taux de mutation constant sur toutes les branches : toutes les feuilles à la même distance de la racine.

Taux variables selon les lignées : d'autres méthodes de reconstruction (neighbor-joining, maximum de vraisemblance), hors de ce cours.

Séquences A–E : illustratives (C5).