Chapitre 15 : Apprentissage Non Supervisé

26 février 2026

Plan du chapitre

Introduction à l'Apprentissage Non Supervisé

On dispose d'un échantillon $\mathscr{D}_n = \{x_1, \ldots, x_n\}$ de variables i.i.d. selon une loi inconnue P sur un espace $\mathscr{X} \subseteq \mathbb{R}^d$. Il n'y a **pas d'étiquettes** y.

Trois tâches fondamentales

1. **Estimation de densité** : Reconstruire la fonction de densité f génératrice des données.
2. **Partitionnement (Clustering)** : Diviser les données en groupes homogènes $C_1, \ldots, C_k$.
3. **Réduction de dimensionnalité** : Identifier une sous-variété $M \subset \mathscr{X}$ de dimension $m \ll d$ concentrant l'information.

La Malédiction de la Dimensionnalité

Rappel : En haute dimension, la géométrie défie l'intuition.

- Le volume d'une boule unité se concentre presque exclusivement près de sa surface (écorce).
- La distance entre deux points tirés uniformément tend à devenir constante (variance nulle).

Conséquence algorithmique

Les algorithmes basés sur la distance euclidienne stricte (comme les K-means) perdent leur sens géométrique si la dimension d n'est pas réduite au préalable.

Le problème des K-Means (K-Moyennes)

Le but est de minimiser la dispersion au sein des classes (Inertie intra-classe). On cherche une partition $\mathscr{C} = \{C_1, \ldots, C_k\}$ et des centres $\mu = \{\mu_1, \ldots, \mu_k\}$.

Critère de l'inertie intra-classe

$$W(\mathscr{C}, \mu) = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2$$

Ce problème d'optimisation est NP-difficile. On utilise l'algorithme itératif de **Lloyd** pour trouver un minimum local.

L'Algorithme de Lloyd et la Convergence

L'algorithme alterne deux étapes :

1. **Assignation** :
   $C_j^{(t)} = \{x \in \mathscr{D}_n : \|x - \mu_j^{(t)}\| \leq \|x - \mu_l^{(t)}\| \, \forall l \neq j\}$.
2. **Mise à jour** : $\mu_j^{(t+1)} = \frac{1}{|C_j^{(t)}|} \sum_{x \in C_j^{(t)}} x$ (calcul du barycentre).

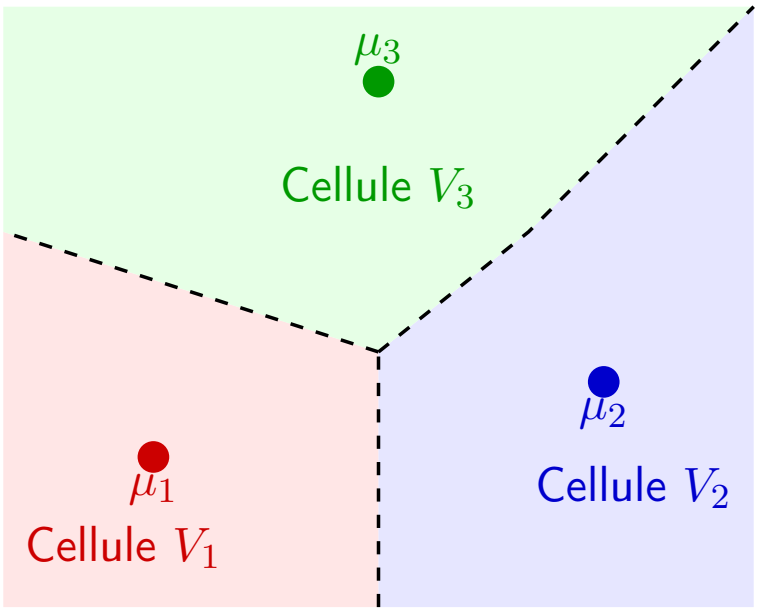
Théorème de convergence

L'algorithme converge vers un minimum local en un nombre **fini** d'itérations.

Preuve : L'étape 1 diminue le critère W (points rapprochés). L'étape 2 diminue W (la moyenne minimise les moindres carrés). $W \geq 0$ forme une suite décroissante. Or, il n'existe qu'un nombre fini de partitions possibles de n points en k classes. L'algorithme stationne donc en temps fini.

Géométrie : La partition de Voronoï

D'un point de vue géométrique, l'étape d'assignation découpe l'espace $\mathscr{X}$ en cellules convexes définies par les hyperplans médiateurs entre les centroïdes.



Classification Ascendante Hiérarchique (CAH)

Contrairement aux K-Means, la CAH construit une **hiérarchie** de partitions emboîtées sans avoir besoin de fixer k à l'avance.
Elle fusionne itérativement les deux clusters les plus proches.

Formule de Lance-Williams

Permet de mettre à jour la distance entre un nouveau cluster fusionné $C_{i\cup j}$ et un autre cluster C_k de façon récursive :

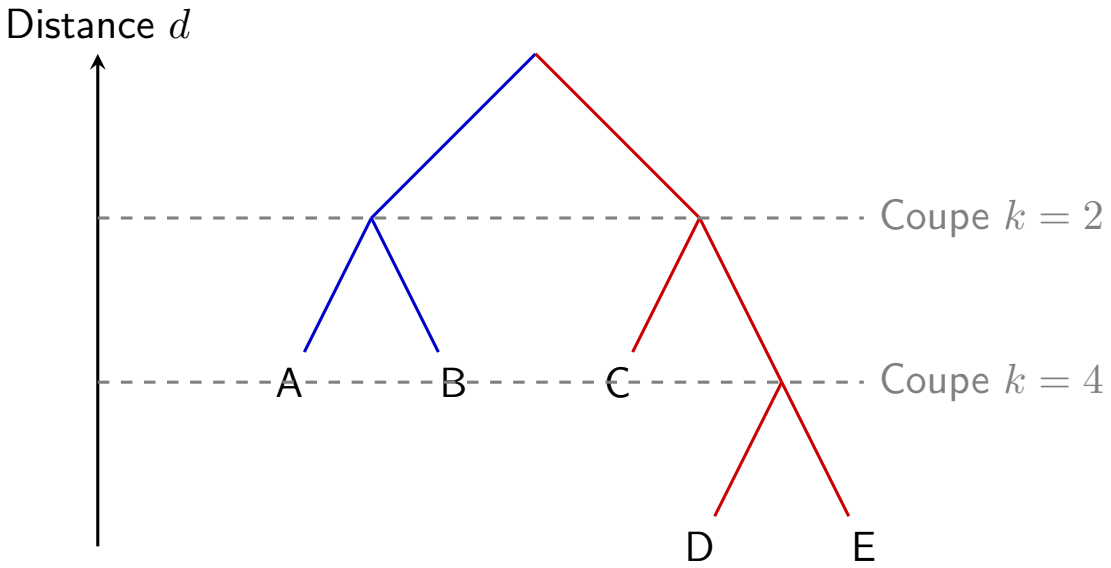
$$d(C_{i\cup j}, C_k) =$$

$$\alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)|$$

Les coefficients α, β, γ définissent le type de lien (Single linkage, Complete linkage, critère de Ward).

Ultramétrie et Dendrogramme

Un dendrogramme indexé définit de manière unique une distance particulière sur l'échantillon, appelée **ultramétrique**.



Inégalité triangulaire forte (Ultramétrique u)

Pour tout triplet $x, y, z \in \mathscr{X}$:

$$u(x, y) \leq \max(u(x, z), u(z, y))$$

Modèles de Mélange Gaussiens (GMM)

Les K-Means imposent des clusters sphériques de même rayon. Pour modéliser des densités plus complexes (ellipses), on utilise une approche probabiliste.

On suppose que x suit une densité (avec $\theta_k = \{\mu_k, \Sigma_k\}$) :

$$p(x \mid \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

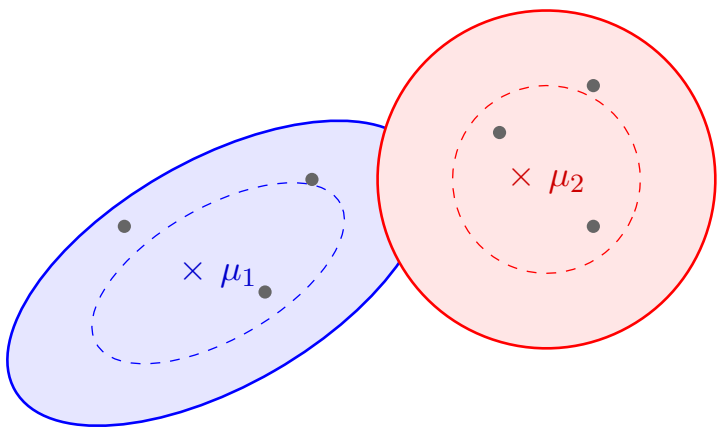


Figure – Ellipses de covariance d'un Modèle de Mélange Gaussien.

L'Algorithme EM (Expectation-Maximization)

On veut maximiser la log-vraisemblance incomplète : $\ell(\theta) = \sum_i \log \sum_z \pi_z p(x_i \mid z, \theta)$. La somme à l'intérieur du log rend le calcul direct impossible.

Preuve de la monotonie de l'EM (Inégalité de Jensen)

Pour toute distribution q sur la variable latente z :

$$\log p(x|\theta) = \log \sum_z q(z) \frac{p(x, z|\theta)}{q(z)} \geq \sum_z q(z) \log \frac{p(x, z|\theta)}{q(z)} \equiv \mathscr{L}(q, \theta)$$

- **Étape E (Expectation)** : On choisit $q(z) = p(z|x, \theta^{(t)})$. L'inégalité devient une **égalité** au point $\theta^{(t)}$.
- **Étape M (Maximization)** : On maximise $\mathscr{L}$ par rapport à θ.

Donc : $\ell(\theta^{(t+1)}) \geq \mathscr{L}(q, \theta^{(t+1)}) \geq \mathscr{L}(q, \theta^{(t)}) = \ell(\theta^{(t)})$

La vraisemblance est **strictement croissante** à chaque itération.

Le Clustering Spectral

Lorsque les clusters ne sont ni sphériques ni gaussiens (ex : lunes imbriquées), les approches classiques échouent.
Le clustering spectral utilise le **Laplacien de graphe**.

Soit W la matrice d'adjacence du graphe des similarités des données, et D la matrice diagonale des degrés ($D_{ii} = \sum_j W_{ij}$).

Les Matrices Laplaciennes

- $L = D - W$ (Laplacien combinatoire).
- $L_{sym} = I - D^{-1/2} W D^{-1/2}$ (Laplacien normalisé).

Algorithme : On extrait les k premiers vecteurs propres de L_{sym}, on plonge les données dans cet espace de dimension k, puis on y applique un simple K-Means. Cela résout une relaxation du problème de coupe minimale (*Normalized Cut*).

t-SNE et Divergence KL

Le **t-SNE** est conçu pour la visualisation en 2D/3D. Il aligne les probabilités de voisinage dans l'espace d'origine (P) avec celles dans l'espace réduit (Q).

Minimisation de la Divergence de Kullback-Leibler

$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

L'astuce (Loi de Student) : Dans l'espace cible, Q utilise une loi de Student à 1 degré de liberté (queues lourdes). Cela évite le *crowding problem* (l'effondrement de tous les points au centre) et repousse les clusters dissimilaires.

UMAP et Stabilité

UMAP (Uniform Manifold Approximation and Projection) : S'appuie sur la géométrie riemannienne et l'algèbre topologique. Il construit un complexe simplicial flou de la variété globale, puis l'optimise par Cross-Entropy. Plus rapide que t-SNE et préserve mieux la structure **globale** des données.

Théorie de la Stabilité du Clustering

Un algorithme est "stable" si deux échantillons d'une même loi P donnent des partitions similaires. Pour $n \to \infty$, la stabilité dépend de l'**unicité** du minimiseur du risque théorique $R(\mathscr{C})$. S'il existe des symétries (plusieurs minima globaux optimaux), l'algorithme est intrinsèquement **instable**.