Chapitre 10 : Méthodes à noyaux

26 février 2026

Plan du chapitre

Les limites de la séparabilité linéaire

Supposer qu'un problème de classification est séparable par un hyperplan dans son espace d'origine $\mathscr{X}$ est une hypothèse très forte, et le plus souvent fausse en pratique.

L'idée fondamentale (Espace de redescription)

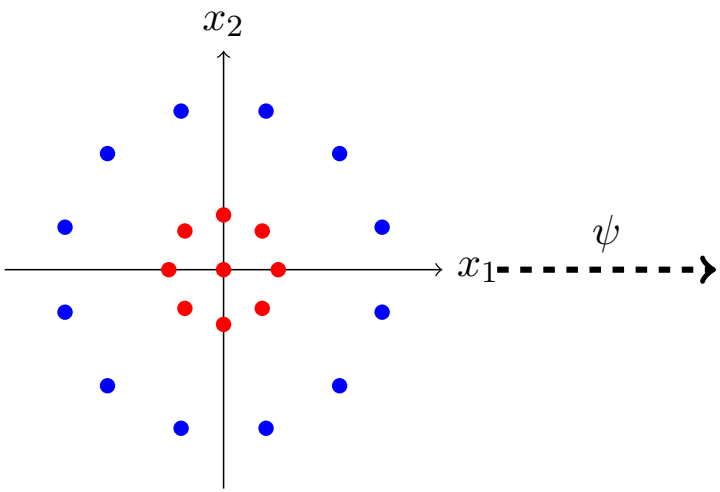
Plutôt que de chercher des frontières non-linéaires complexes dans l'espace d'origine $\mathscr{X}$, on **plonge** les données dans un espace de caractéristiques $\mathscr{F}$ de dimension supérieure via une application non-linéaire ψ.

$$\psi : \mathscr{X} \longrightarrow \mathscr{F}$$

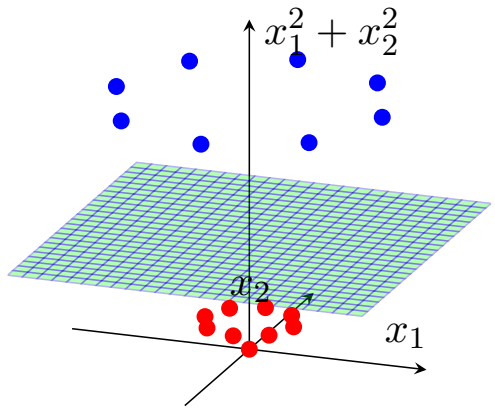
L'espoir est que les données deviennent **linéairement séparables** dans $\mathscr{F}$.

Illustration : Du cercle au plan

Considérons des données en 2D ($\mathscr{X} = \mathbb{R}^2$) réparties en cercles concentriques. Aucune droite ne peut les séparer. Appliquons la transformation $\psi(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$.



Espace d'origine $\mathscr{X}$



Espace de redescription $\mathscr{F}$

Le fléau algorithmique et l'Astuce du Noyau

Si la dimension de l'espace d'arrivée $\mathscr{F}$ est immense (voire infinie), le calcul explicite des coordonnées $\psi(\mathbf{x})$ et des produits scalaires $\langle \psi(\mathbf{x}), \psi(\mathbf{y}) \rangle_{\mathscr{F}}$ devient **informatiquement intraitable**.

Le Kernel Trick (L'Astuce du Noyau)

De nombreux algorithmes linéaires (SVM, ACP, Régression Ridge) ne dépendent des données **que** via leurs produits scalaires. Si nous pouvons définir une fonction K telle que :

$$K(\mathbf{x}, \mathbf{y}) = \langle \psi(\mathbf{x}), \psi(\mathbf{y}) \rangle_{\mathscr{F}}$$

Nous n'avons **jamais** besoin de calculer $\psi(\mathbf{x})$ explicitement ! L'évaluation se fait directement dans l'espace de départ $\mathscr{X}$.

Formalisme Mathématique : La Matrice de Gram

Définition (Noyau semi-défini positif)

Une fonction $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ est un noyau symétrique semi-défini positif (SDP) si :

1. $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$ pour tout $\mathbf{x}, \mathbf{y} \in \mathscr{X}$.
2. Pour tout entier N, pour tout ensemble d'exemples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathscr{X}$, et pour tout vecteur $\mathbf{c} \in \mathbb{R}^N$:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$$

La matrice symétrique $\mathbf{K}$ de terme général $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ est appelée **Matrice de Gram**. Elle contient toute la géométrie (distances et angles) du jeu de données dans l'espace $\mathscr{F}$.

Théorème de Moore-Aronszajn

Toute fonction K ne correspond pas à un produit scalaire. Le théorème suivant est le socle de l'analyse fonctionnelle pour l'apprentissage à noyaux.

Théorème (Moore-Aronszajn, 1950)

Pour toute fonction $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ symétrique semi-définie positive, il existe un **unique** (à isométrie près) espace de Hilbert $\mathscr{F}$ et une application $\psi : \mathscr{X} \to \mathscr{F}$ tels que :

$$K(\mathbf{x}, \mathbf{y}) = \langle \psi(\mathbf{x}), \psi(\mathbf{y}) \rangle_{\mathscr{F}} \quad \forall \mathbf{x}, \mathbf{y} \in \mathscr{X}$$

L'espace $\mathscr{F}$ est appelé **Espace de Hilbert à Noyau Reproduisant (RKHS)**. Ce théorème garantit mathématiquement que dès qu'on s'assure que K est une matrice SDP, le plongement ψ existe, même si on ne le connaît pas !

Les Noyaux Classiques

En pratique, on choisit des noyaux génériques pré-établis par la théorie :

- **Noyau Linéaire** : $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$
  (Correspond à l'espace d'origine, aucune redescription).
- **Noyau Polynomial** : $K(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + c)^p$
  (L'espace $\mathscr{F}$ contient tous les monômes de degré $\leq p$).
- **Noyau Gaussien (RBF)** : $K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$
  (L'espace $\mathscr{F}$ est de dimension infinie !).

Preuve : Dimension infinie du Noyau RBF

Prenons $\mathscr{X} = \mathbb{R}$ et montrons que le noyau Gaussien correspond à un espace de dimension infinie. Soit $\sigma = 1$.

$$K(x,y) = \exp\left(-\frac{(x-y)^2}{2}\right) = \exp\left(-\frac{x^2}{2}\right)\exp\left(-\frac{y^2}{2}\right)\exp(xy)$$

Développons $\exp(xy)$ en série de Taylor :

$$K(x,y) = \exp\left(-\frac{x^2}{2}\right)\exp\left(-\frac{y^2}{2}\right)\sum_{n=0}^{\infty}\frac{(xy)^n}{n!}$$

En identifiant $K(x,y) = \langle \psi(x), \psi(y) \rangle_{\ell_2}$, on trouve les coordonnées de $\psi(x)$ dans un espace de dimension infinie :

$$\psi(x)_n = \frac{1}{\sqrt{n!}}\exp\left(-\frac{x^2}{2}\right)x^n \quad \text{pour } n \in \mathbb{N}$$

Le calcul du produit scalaire dans $\mathbb{R}^\infty$ nous coûte l'évaluation d'une simple exponentielle dans $\mathbb{R}$!

Rappel : SVM dans l'espace primal vs dual

Dans le Chapitre 11, le problème dual du SVM s'écrivait :

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

En appliquant la transformation ψ, on remplace simplement le produit scalaire par notre fonction noyau K :

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

La fonction de décision finale pour un nouveau point $\mathbf{x}$ devient :

$$f(\mathbf{x}) = \sum_{i \in \text{V.S.}} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b$$

La non-linéarité est gérée avec une élégance mathématique absolue.

Kernel PCA (ACP à noyau) - 1/3

L'Analyse en Composantes Principales (ACP) classique réduit la dimension de manière linéaire. Le **Kernel PCA** permet d'extraire des composantes principales non-linéaires.

Soit la matrice de covariance dans l'espace $\mathscr{F}$ (en supposant les données $\psi(\mathbf{x}_i)$ centrées) :

$$\mathbf{C} = \frac{1}{N} \sum_{i=1}^{N} \psi(\mathbf{x}_i)\psi(\mathbf{x}_i)^T$$

On cherche les vecteurs propres $\mathbf{v} \in \mathscr{F}$ tels que $\mathbf{C}\mathbf{v} = \lambda\mathbf{v}$.
Or, tout vecteur propre $\mathbf{v}$ appartient à l'espace engendré par les données. Il s'écrit donc comme combinaison linéaire :

$$\mathbf{v} = \sum_{i=1}^{N} \alpha_i \psi(\mathbf{x}_i)$$

Kernel PCA (ACP à noyau) - 2/3

Substituons $\mathbf{v}$ dans l'équation aux valeurs propres :

$$\left(\frac{1}{N}\sum_{j=1}^{N}\psi(\mathbf{x}_j)\psi(\mathbf{x}_j)^T\right)\left(\sum_{i=1}^{N}\alpha_i\psi(\mathbf{x}_i)\right)=\lambda\sum_{i=1}^{N}\alpha_i\psi(\mathbf{x}_i)$$

Multiplions à gauche par $\psi(\mathbf{x}_k)^T$ pour tout k :

$$\frac{1}{N}\sum_{j=1}^{N}\underbrace{\psi(\mathbf{x}_k)^T\psi(\mathbf{x}_j)}_{K_{kj}}\sum_{i=1}^{N}\alpha_i\underbrace{\psi(\mathbf{x}_j)^T\psi(\mathbf{x}_i)}_{K_{ji}}=\lambda\sum_{i=1}^{N}\alpha_i\underbrace{\psi(\mathbf{x}_k)^T\psi(\mathbf{x}_i)}_{K_{ki}}$$

En notation matricielle (avec $\mathbf{K}$ la matrice de Gram de taille $N\times N$) :

$$\frac{1}{N}\mathbf{K}^2\boldsymbol{\alpha}=\lambda\mathbf{K}\boldsymbol{\alpha}\implies\mathbf{K}\boldsymbol{\alpha}=N\lambda\boldsymbol{\alpha}$$

Le problème se réduit à trouver les vecteurs propres $\boldsymbol{\alpha}$ de la matrice de Gram $\mathbf{K}$!

Kernel PCA (ACP à noyau) - 3/3

Le problème du centrage

Dans la dérivation précédente, nous avons supposé que $\frac{1}{N}\sum \psi(\mathbf{x}_i) = 0$. Comme on ne calcule jamais $\psi(\mathbf{x}_i)$, on ne peut pas les centrer explicitement.

L'astuce consiste à "centrer la matrice de Gram" algébriquement. Soit $\mathbf{1}_N$ la matrice $N \times N$ dont tous les éléments valent $1/N$. La matrice de Gram centrée $\tilde{\mathbf{K}}$ se calcule par :

$$\tilde{\mathbf{K}} = \mathbf{K} - \mathbf{1}_N\mathbf{K} - \mathbf{K}\mathbf{1}_N + \mathbf{1}_N\mathbf{K}\mathbf{1}_N$$

Algorithme KPCA complet :

1. Construire la matrice de Gram $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$.
2. Centrer la matrice pour obtenir $\tilde{\mathbf{K}}$.
3. Calculer les vecteurs propres $\boldsymbol{\alpha}$ de $\tilde{\mathbf{K}}$.
4. La projection d'un nouveau point $\mathbf{x}$ sur la k-ème composante est : $z_k = \sum_{i=1}^{N} \alpha_{k,i} K(\mathbf{x}, \mathbf{x}_i)$.