

Méthodes à noyau

Bassem Ben Hamed

Control, Optimization and Model Reduction in ML - CIMPA 2025

18 février 2025

Table des matières

- 1 Introduction
- 2 Noyaux symétriques définis positifs
- 3 Algorithmes basés sur les noyaux

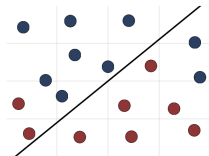
- 1 Introduction
- 2 Noyaux symétriques définis positifs
- 3 Algorithmes basés sur les noyaux

- Les méthodes à noyau étendent des algorithmes comme les SVMs pour définir des frontières de décision non linéaires.
- Elles reposent sur des noyaux définis positifs, qui induisent un produit scalaire dans un espace de Hilbert.
- La substitution du produit scalaire par un noyau positif permet une séparation linéaire dans un espace de dimension supérieure.

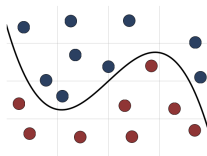
- Définition et propriétés des noyaux symétriques définis positifs.
- Extension des SVMs et garanties d'apprentissage basées sur la marge.

Classification linéaire et limitations

- Les SVMs permettent une séparation linéaire efficace et théoriquement justifiée.
- Dans certains cas, une séparation linéaire n'est pas possible dans l'espace d'entrée X .



(a)



(b)

Exemple où une séparation linéaire est impossible (a) et où une séparation non linéaire est nécessaire (b).

Projection dans un espace de Hilbert

- Une solution consiste à appliquer une transformation non linéaire $\Phi : X \rightarrow H$.
- L'espace H est de dimension plus élevée, où une séparation linéaire devient possible.

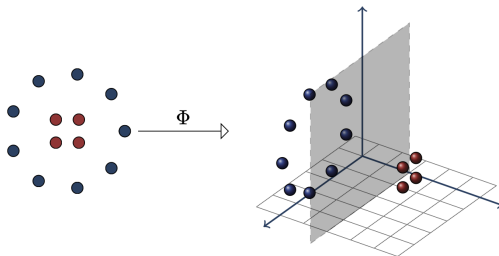


Illustration de la transformation dans un espace de Hilbert où la séparation devient linéaire.

Problème :

- L'espace H peut être très grand en pratique (ex. : classification de documents contenant 100K mots avec trigrams $\rightarrow 10^{15}$ dimensions).
- Le calcul des produits scalaires dans cet espace est coûteux.

Pourquoi cela fonctionne-t-il encore ?

- Les bornes de généralisation montrent que la performance des SVMs ne dépend pas de la dimension de H , mais du **marge** ρ et du nombre d'exemples m .
- Un bon ρ permet de bien généraliser même en haute dimension.

Solution :

- Utiliser les **méthodes à noyau**, qui permettent de contourner le calcul explicite dans l'espace de grande dimension.

Definition

Une fonction $K : X \times X \rightarrow \mathbb{R}$ est appelée un **noyau** sur X s'il existe une fonction $\Phi : X \rightarrow H$ vers un espace de Hilbert H tel que :

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle, \quad \forall x, x' \in X \quad (1)$$

Interprétation

- $K(x, x')$ représente un **produit scalaire** dans un espace de caractéristiques H .
- Comme un produit scalaire mesure la similarité entre deux vecteurs, $K(x, x')$ est souvent vu comme une **mesure de similarité** dans l'espace d'entrée X .

1. Efficacité

- Le calcul de $K(x, x')$ est souvent beaucoup plus rapide que celui de $\Phi(x)$ et du produit scalaire dans H .
- Exemples courants :

$$K(x, x') = O(N), \quad \langle \Phi(x), \Phi(x') \rangle = O(\dim(H))$$

avec $\dim(H) \gg N$, voire ∞ dans certains cas.

2. Flexibilité

- Pas besoin de définir ou de calculer explicitement Φ .
- Il suffit que K vérifie la condition de Mercer (voir Théorème suivant) pour garantir l'existence de Φ .
- Permet une grande liberté dans le choix des noyaux adaptés aux données.

Theorem

Soit $X \subset \mathbb{R}^N$ un ensemble compact et $K : X \times X \rightarrow \mathbb{R}$ une fonction continue et symétrique. Alors, K admet un développement uniformément convergent sous la forme :

$$K(x, x') = \sum_{n=0}^{\infty} a_n \phi_n(x) \phi_n(x')$$

avec $a_n > 0$ si et seulement si, pour toute fonction $c \in L^2(X)$, la condition suivante est satisfaite :

$$\int_X \int_X c(x) c(x') K(x, x') dx dx' \geq 0.$$

- Cette condition est importante pour garantir la convexité du problème d'optimisation pour des algorithmes tels que les SVM, assurant ainsi la convergence vers un minimum global. Une condition équivalente à celle de Mercer sous les hypothèses du théorème est que le noyau K soit positif défini symétrique (PDS).
- Cette propriété est en fait plus générale, car elle ne nécessite aucune hypothèse sur X .
- Dans la section suivante, on donne la définition de cette propriété, présentons plusieurs exemples courants de noyaux PDS, puis on montre que les noyaux PDS induisent un produit scalaire dans un espace de Hilbert et on prouve plusieurs propriétés de clôture générales pour les noyaux PDS.

Table des matières

- 1 Introduction
- 2 Noyaux symétriques définis positifs
- 3 Algorithmes basés sur les noyaux

Definition

Un noyau $K : X \times X \rightarrow \mathbb{R}$ est dit symétrique défini positif (PDS) si, pour tout ensemble $\{x_1, \dots, x_m\} \subset X$, la matrice $K = [K(x_i, x_j)]_{ij} \in \mathbb{R}^{m \times m}$ est semi-définie positive symétrique (SPSD).

K est SPSP si elle est symétrique et que l'une des deux conditions équivalentes suivantes est vérifiée :

- Les valeurs propres de K sont non négatives ;
- Pour tout vecteur colonne $c = (c_1, \dots, c_m)^\top \in \mathbb{R}^{m \times 1}$, on a

$$c^\top K c = \sum_{i,j=1}^m c_i c_j K(x_i, x_j) \geq 0. \quad (2)$$

Matrice de noyau (Gram matrix)

- Pour un échantillon $S = (x_1, \dots, x_m)$, $K = [K(x_i, x_j)]_{ij} \in \mathbb{R}^{m \times m}$ est appelé la matrice de noyau ou la matrice de Gram associée à K et à l'échantillon S .
- Terminologie : la matrice de noyau associée à un noyau défini positif est semi-définie positive. C'est la terminologie mathématique correcte. Cependant, il convient de noter que dans le contexte de l'apprentissage automatique, certains auteurs ont choisi d'utiliser à la place le terme "noyau défini positif" pour désigner une matrice de noyau définie positive ou ont utilisé de nouveaux termes comme "noyau semi-défini positif".

Exemple : Noyaux polynomiaux

Pour toute constante $c > 0$, un noyau polynomial de degré $d \in \mathbb{N}$ est défini sur \mathbb{R}^N par :

$$K(x, x') = (x \cdot x' + c)^d. \quad (3)$$

Par exemple, un noyau polynomial de degré 2 ($d = 2$) correspond au produit scalaire suivant :

$$\forall x, x' \in \mathbb{R}^N, K(x, x') = (x \cdot x' + c)^2.$$

Noyau polynomial de degré 2

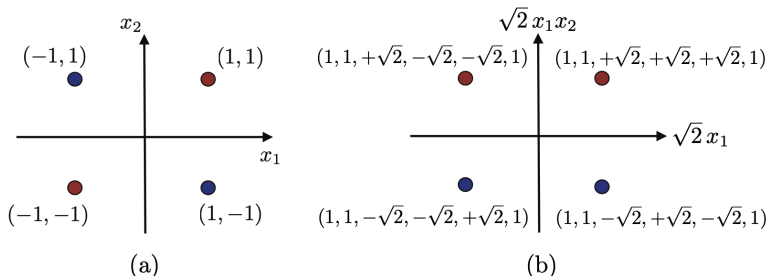
Pour un espace d'entrée de dimension $N = 2$, un noyau polynomial de degré 2 ($d = 2$) correspond au produit scalaire suivant dans un espace de dimension 6 :

$$\forall x, x' \in \mathbb{R}^2, K(x, x') = (x_1 x'_1 + x_2 x'_2 + c)^2 = \begin{bmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2}x_1 x_2 \\ \sqrt{2c}x_1 \\ \sqrt{2c}x_2 \\ c \end{bmatrix} \cdot \begin{bmatrix} x_1'^2 \\ x_2'^2 \\ \sqrt{2}x'_1 x'_2 \\ \sqrt{2c}x'_1 \\ \sqrt{2c}x'_2 \\ c \end{bmatrix}. \quad (4)$$

Ainsi, les caractéristiques associées à un noyau polynomial de degré 2 sont les caractéristiques originales (x_1, x_2) , ainsi que les produits de ces caractéristiques et la constante.

Illustration de la séparation avec un noyau polynomial

Soit un ensemble de données simple en deux dimensions qui n'est pas linéairement séparable. Ce problème est connu sous le nom de problème XOR.



La projection de ces points dans l'espace bi-dimensionnel défini par les troisième et quatrième coordonnées, illustre comment les données deviennent séparables dans un espace de dimension plus élevée grâce à l'utilisation du noyau polynomial.

Exemple : Noyaux gaussiens

Pour toute constante $\sigma > 0$, un noyau gaussien ou fonction de base radiale (RBF) est défini sur \mathbb{R}^N par :

$$\forall x, x' \in \mathbb{R}^N, K(x, x') = \exp\left(-\frac{\|x' - x\|^2}{2\sigma^2}\right). \quad (5)$$

- Les noyaux gaussiens sont parmi les noyaux les plus fréquemment utilisés dans les applications.
- En utilisant le développement en série de l'exponentielle, on peut réécrire l'expression de K comme suit :

$$\forall x, x' \in \mathbb{R}^2, K(x, x') = \sum_{n=0}^{\infty} \frac{(x \cdot x')^n}{n! \sigma^{2n}}.$$

- Cela montre que les noyaux gaussiens sont des combinaisons linéaires positives de noyaux polynomiaux de tous les degrés $n \geq 0$.

Exemple : Noyaux sigmoïdes

Pour toute constante réelle $a, b \geq 0$, un noyau sigmoïde est défini sur \mathbb{R}^N par :

$$\forall x, x' \in \mathbb{R}^N, K(x, x') = \tanh(a(x \cdot x') + b). \quad (6)$$

- L'utilisation des noyaux sigmoïdes avec les SVMs conduit à un algorithme étroitement lié aux algorithmes d'apprentissage basés sur des réseaux neuronaux simples, qui sont également souvent définis via une fonction sigmoïde.
- Lorsque $a < 0$ ou $b < 0$, le noyau n'est pas PDS et le réseau neuronal correspondant ne bénéficie pas des garanties de convergence de l'optimisation convexe.

Espace de Hilbert à noyaux reproduisants

Une propriété cruciale des noyaux PDS, consiste à induire un produit scalaire dans un espace de Hilbert. La démonstration fait appel au lemme suivant.

Lemma (Inégalité de Cauchy-Schwarz pour les noyaux PDS)

Soit K un noyau PDS. Alors, pour tout $x, x' \in X$,

$$K(x, x')^2 \leq K(x, x)K(x', x'). \quad (7)$$

Démonstration : Considérons la matrice

$$\mathbf{K} = \begin{bmatrix} K(x, x) & K(x, x') \\ K(x', x) & K(x', x') \end{bmatrix}.$$

Puisque K est PDS, la matrice \mathbf{K} est semi-définie positive, ce qui implique que son déterminant est non-négatif :

$$\det(K) = K(x, x)K(x', x') - K(x, x')^2 \geq 0. \quad \square$$

Theorem

Soit $K : X \times X \rightarrow \mathbb{R}$ un noyau PDS. Alors, il existe un espace de Hilbert H (voir définition A.2) et une application Φ de X vers H telle que :

$$\forall x, x' \in X, K(x, x') = \langle \Phi(x), \Phi(x') \rangle. \quad (8)$$

De plus, H possède la propriété suivante, connue sous le nom de propriété reproduisante :

$$\forall h \in H, \forall x \in X, h(x) = \langle h, K(x, \cdot) \rangle. \quad (9)$$

H est appelé un espace de Hilbert reproduisant les noyaux (RKHS) associé à K .

Démonstration : Tu sais bien qu'il faut le faire !

Espace de Hilbert reproduisant les noyaux (RKHS)

- L'espace de Hilbert H défini dans la démonstration du théorème pour un noyau PDS K est appelé l'espace de Hilbert reproduisant les noyaux (RKHS) associé à K .
- Un espace de Hilbert H tel qu'il existe $\Phi : X \rightarrow H$ avec $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$ pour tous $x, x' \in X$ est appelé un espace de caractéristiques associé à K , et Φ est appelé un mappage de caractéristiques.
- La norme induite par le produit scalaire dans l'espace de caractéristiques H est notée $\| \cdot \|_H$, avec $\|w\|_H = \langle w, w \rangle$ pour tout $w \in H$.
- Les espaces de caractéristiques associés à K ne sont généralement pas uniques et peuvent avoir des dimensions différentes. En pratique, lorsque l'on fait référence à la dimension de l'espace de caractéristiques associé à K , cela désigne soit la dimension de l'espace de caractéristiques basé sur un mappage explicite, soit celle du RKHS associé à K .

Rôle des noyaux PDS dans l'apprentissage

- Le Théorème précédent implique que les noyaux PDS peuvent être utilisés pour définir implicitement un espace de caractéristiques ou des vecteurs de caractéristiques.
- Ainsi, dans le contexte de l'apprentissage avec des noyaux PDS et pour un espace d'entrée fixe, le problème de recherche de caractéristiques utiles est remplacé par celui de la recherche de noyaux PDS utiles.
- Le choix approprié d'un noyau PDS pour une tâche sera donc crucial dans la pratique.

Table des matières

- 1 Introduction
- 2 Noyaux symétriques définis positifs
- 3 Algorithmes basés sur les noyaux

Extension des SVMs avec noyaux PDS

On a noté que le problème dual d'optimisation des SVMs et la solution obtenue ne dépendent pas directement des vecteurs d'entrée, mais uniquement des produits scalaires.

- Un noyau PDS définit implicitement un produit scalaire.
- On peut donc généraliser les SVMs en remplaçant chaque produit scalaire $x \cdot x'$ par $K(x, x')$.

Cela mène à la formulation générale du problème d'optimisation des SVMs avec noyaux PDS :

$$\max_{\alpha} \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (10)$$

sous contraintes :

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^m \alpha_i y_i = 0, \quad i \in [m].$$

Solution du SVM avec noyaux PDS

L'hypothèse h résultante est donnée par :

$$h(x) = \text{sgn} \left(\sum_{i=1}^m \alpha_i y_i K(x_i, x) + b \right). \quad (11)$$

Avec :

$$b = y_i - \sum_{j=1}^m \alpha_j y_j K(x_j, x_i), \quad \text{pour tout } x_i \text{ tel que } 0 < \alpha_i < C.$$

Conclusion :

- L'utilisation des noyaux permet d'étendre les SVMs à des espaces de grande ou infinie dimension sans calculer explicitement les transformations des données.
- Le choix du noyau est crucial pour la performance du modèle.

Formulation vectorielle du problème d'optimisation

En utilisant la matrice de noyau \mathbf{K} associée au noyau K pour l'échantillon d'apprentissage (x_1, \dots, x_m) , nous pouvons réécrire le problème d'optimisation sous forme vectorielle :

$$\max_{\alpha} \quad \frac{1}{2} \mathbf{1}^\top \alpha - (\alpha \circ y)^\top \mathbf{K} (\alpha \circ y) \quad (12)$$

sous contraintes :

$$0 \leq \alpha \leq C, \quad \alpha^\top y = 0.$$

Remarque :

- $\alpha \circ y$ est le produit d'Hadamard (produit élément par élément) des vecteurs α et y .
- La solution en notation vectorielle est la même que précédemment, avec :

$$b = y_i - (\alpha \circ y)^\top \mathbf{K} e_i, \quad \text{pour tout } x_i \text{ tel que } 0 < \alpha_i < C.$$

SVMs avec noyaux PDS : une généralisation clé

- L'utilisation des noyaux PDS permet une extension naturelle des SVMs.
- L'extension est cruciale car elle permet une **projection implicite non linéaire** des points d'entrée dans un espace de grande dimension où une séparation à large marge est recherchée.

Extensions aux autres algorithmes

- La même approche peut être appliquée à d'autres domaines :
 - Régression
 - Classement (ranking)
 - Réduction de dimension
 - Clustering