Chapitre VII
Méthodes à noyaux
Cycle pluridisciplinaire d'études supérieures
Université Paris sciences et lettres

Les méthodes à noyaux permettent de transporter les entrées d'un problème dans un autre espace, souvent plus grand, parfois même de dimension infinie. Il existe deux raisons principales pour lesquelles cela peut être avantageux.

La première est que dans un espace plus grand, il peut être plus facile, par exemple dans un contexte de classification binaire, de classifier correctement les entrées à l'aide d'un hyperplan séparateur. Ainsi, cela permet d'utiliser les algorithmes donnant des prédicteurs linéaires et d'améliorer la qualité des prédicteurs obtenus.

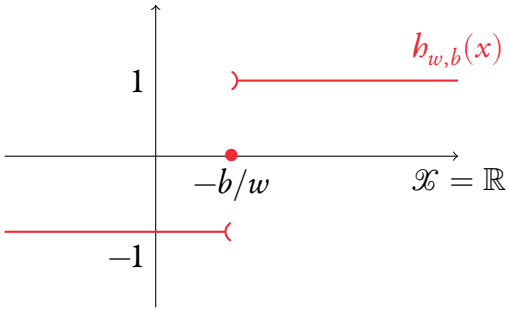
La seconde raison est que dans certains problèmes, l'ensemble d'entrées n'est pas un espace préhilbertien, ce qui rend impossible *a priori* l'utilisation des prédicteurs linéaires. Cependant, en utilisant une application qui va transporter les entrées dans un espace préhilbertien, il devient possible de considérer des prédicteurs linéaires dans ce nouvel espace.

I. Exemple introductif

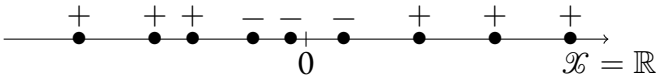
On considère un contexte de classification binaire en dimension 1 avec $\mathscr{X} = \mathbb{R}$ pour ensemble d'entrées et $\mathscr{Y} = \{-1, 1\}$ pour ensemble de sorties. La classe des classifieurs linéaires s'écrit :

$$\mathscr{F} = \{h_{w,b} \colon x \mapsto \operatorname{sign}(wx + b)\}_{\substack{w\in\mathbb{R} \\ b\in\mathbb{R}}},$$

dont l'allure, dans le cas $w > 0$, est représentée ci-après :



Si on imagine un échantillon $S = (x_i, y_i)_{i\in[n]}$ de la forme suivante :



on voit qu'il n'est pas linéairement séparable, et donc aucun classifieur linéaire n'est satisfaisant. L'idée clé est qu'on peut transporter les entrées dans un espace plus grand où les classes seront, on l'espère, plus facilement séparables. On considère le problème auxiliaire suivant :

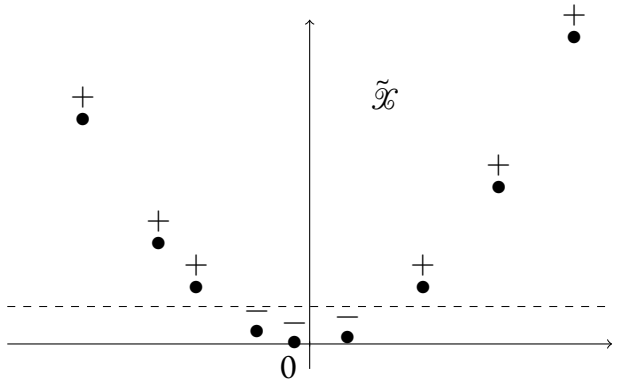
— $\tilde{\mathscr{X}} = \mathbb{R}^2$ un espace d'entrées augmenté (qu'on appellera *espace de redescription*) ;

— l'application :

$$\begin{array}{rccc} \psi : & \mathscr{X} & \longrightarrow & \tilde{\mathscr{X}} \\ & x & \longmapsto & (x, x^2) \end{array}$$

qu'on appellera *application de redescription*, pour transporter les entrées $(x_i)_{i\in[n]}$;

— l'échantillon $\tilde{S} = (\psi(x_i), y_i)_{i\in[n]}$.



On voit que l'échantillon $\tilde{S}$ est linéairement séparable. Autrement dit, il existe $\tilde{w} \in \mathbb{R}^2$ et $\tilde{b} \in \mathbb{R}$ tels que le classifieur linéaire $h_{\tilde{w},\tilde{b}}$ prédit correctement les exemples de $\tilde{S}$. Le prédicteur correspondant dans le problème initial est $h_{\tilde{w},\tilde{b}} \circ \psi$, qui n'est pas un prédicteur linéaire.

2. Espace de redescription

Quelques rappels sur les produits scalaires.

Définition. — *Produit scalaire.* — Soit E un espace vectoriel réel. Un *produit scalaire* est une application $\phi : E \times E \to \mathbb{R}$ telle que :

(i) Pour tout $x \in E$, les applications $y \mapsto \phi(x, y)$ et $y \mapsto \phi(y, x)$ sont linéaires,

(ii) $\forall x, y \in E,\ \phi(x, y) = \phi(y, x)$,

(iii) $\forall x \in E,\ \phi(x, x) = 0 \implies x = 0$,

(iv) $\forall x \in E,\ \phi(x, x) \geqslant 0$.

On note souvent $\phi(x, y) = \langle x, y \rangle$.

Définition. — *Espace préhilbertien.* — Un espace vectoriel réel muni d'un produit scalaire $(E, \langle \cdot, \cdot \rangle)$ est appelé *espace préhilbertien*. On note $\|x\| = \|x\|_E = \sqrt{\langle x, x \rangle}$ la norme associée.

Définition. — Soit $\mathcal{X}$ un ensemble quelconque et $(\tilde{\mathcal{X}}, \langle \cdot, \cdot \rangle_{\tilde{\mathcal{X}}})$ un espace préhilbertien.

(i) On appelle *application de redescription* (ou *feature map*) de $\mathcal{X}$ dans $\tilde{\mathcal{X}}$ toute application $\psi : \mathcal{X} \to \tilde{\mathcal{X}}$.

(ii) On appelle *noyau* associé à ψ l'application $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ définie par :

$$\forall x, x' \in \mathcal{X}, \quad K(x, x') = \langle \psi(x), \psi(x') \rangle_{\tilde{\mathcal{X}}}. \tag{$*$}$$

Définition. — Une application $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un *noyau* s'il existe un espace préhilbertien $\tilde{\mathcal{X}}$ tel que $(*)$ soit vérifié.

3. Apprentissage dans l'espace de redescription

On présente dans ce paragraphe le principe de l'apprentissage dans un espace de redescription et on le détaille dans l'exemple de la régression polynomiale.

Soit $\mathcal{X}$ et $\mathcal{Y}$ deux ensembles quelconques, $n \geqslant 1$ et $S = (x_i, y_i)_{i\in[n]} \in \mathcal{S}(\mathcal{X}, \mathcal{Y})$ un échantillon. Soit $\tilde{\mathcal{X}}$ un espace préhilbertien et $\psi\colon \mathcal{X} \to \tilde{\mathcal{X}}$ une application de redescription. On considère un *problème d'apprentissage auxiliaire* avec :

— $\tilde{\mathcal{X}}$ pour ensemble d'entrées,
— $\mathcal{Y}$ pour ensemble de sorties,
— $\tilde{S} = (\psi(x_i), y_i)_{i\in[n]}$ pour échantillon d'apprentissage,
— une classe de prédicteurs linéaires sur $\tilde{\mathcal{X}}$:

$$\tilde{\mathcal{F}} = \left\{\phi \circ g_{\tilde{w},\tilde{b}}\right\}_{\substack{\tilde{w}\in\tilde{\mathcal{X}}\\ \tilde{b}\in\mathbb{R}}},$$

pour une certaine fonction $\phi\colon \mathbb{R} \to \mathcal{Y}$.

On construit un prédicteur $\hat{\tilde{f}} \in \tilde{\mathcal{F}}$. Le prédicteur correspondant dans le problème initial est alors :

$$\hat{f} = \hat{\tilde{f}} \circ \psi.$$

Remarque. — La classe de prédicteurs correspondante dans le problème initial est :

$$\mathcal{F} = \left\{\phi \circ g_{\tilde{w},\tilde{b}} \circ \psi\right\}_{\substack{\tilde{w}\in\tilde{\mathcal{X}}\\ \tilde{b}\in\mathbb{R}}}.$$

Exemple. — *Régression polynomiale.* — Le problème initial a pour ensembles d'entrées et de sorties $\mathcal{X} = \mathbb{R}$ et $\mathcal{Y} = \mathbb{R}$. Soit $m \geqslant 1$ et $\tilde{\mathcal{X}} = \mathbb{R}^m$ muni de son produit scalaire canonique. On considère l'application de redescription :

$$\begin{array}{rccc} \psi : & \mathbb{R} & \longrightarrow & \mathbb{R}^m \\ & x & \longmapsto & (x, x^2, \dots, x^m). \end{array}$$

Dans l'espace de redescription $\tilde{\mathcal{X}}$, on considère les prédicteurs linéaires :

$$\tilde{\mathcal{F}} = \left\{g_{\tilde{w},\tilde{b}}\colon \tilde{x} \mapsto \langle \tilde{w}, \tilde{x}\rangle + \tilde{b}\right\}_{\substack{\tilde{w}\in\mathbb{R}\\ \tilde{b}\in\mathbb{R}}}.$$

Pour $(\tilde{w}, \tilde{b}) \in \mathbb{R}^m \times \mathbb{R}$, le prédicteur correspondant à $g_{\tilde{w},\tilde{b}}$ dans le problème initial est $g_{\tilde{w},\tilde{b}} \circ \psi$, qui s'écrit :

$$\forall x \in \mathcal{X}, \quad (g_{\tilde{w},\tilde{b}} \circ \psi)(x) = \langle \tilde{w}, \psi(x)\rangle + \tilde{b} = \sum_{k=1}^{m} \tilde{w}_k x^k + \tilde{b}.$$

Par conséquent, la classe de prédicteurs correspondante dans le problème initial est :

$$\mathscr{F} = \left\{ x \mapsto \sum_{k=1}^{m} \tilde{w}_k x^k + \tilde{b} \right\}_{\substack{\tilde{w}\in\mathbb{R}^n \\ \tilde{b}\in\mathbb{R}}},$$

autrement dit celle des fonctions polynomiales de degré inférieur à m.

4. ASTUCE DU NOYAU

On présente dans ce paragraphe l'astuce du noyau qui permet l'apprentissage de prédicteurs linéaires dans l'espace de redescription sans avoir à manipuler explicitement cet espace de redescription qui peut être de dimension grande, voire infinie.

On se donne un cadre initial d'apprentissage avec $\mathscr{X}$ et $\mathscr{Y}$ des ensembles quelconques d'entrées et de sorties, $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$ une fonction de perte, $n \geqslant 1$ un entier et $S = (x_i, y_i)_{i\in[n]} \in \mathscr{S}(\mathscr{X}, \mathscr{Y})$ un échantillon d'apprentissage.

On considère également un cadre auxiliaire d'apprentissage dans un espace de redescription $\tilde{\mathscr{X}}$. Spécifiquement, soit $(\tilde{\mathscr{X}}, \langle \cdot, \cdot \rangle)$ un espace préhilbertien, $\psi : \mathscr{X} \to \tilde{\mathscr{X}}$ une application de redescription, K le noyau associé, et l'échantillon d'apprentissage :

$$\tilde{S} = (\psi(x_i), y_i)_{i\in[n]}.$$

Soit $\phi : \mathbb{R} \to \mathscr{Y}$ et on considère dans $\tilde{\mathscr{X}}$ les prédicteurs linéaires associés à ϕ :

$$\tilde{\mathscr{F}} = \{f_{w,b}\}_{\substack{w\in\tilde{\mathscr{X}} \\ b\in\mathbb{R}}} \quad \text{où} \quad f_{w,b}(x) = \phi(\langle w, x\rangle + b).$$

Soit $\lambda > 0$. On considère l'algorithme qui donne dans le problème initial le prédicteur : $\hat{f} = f_{\hat{w},\hat{b}} \circ \psi$ où :

$$(\hat{w}, \hat{b}) = \operatorname*{arg\,min}_{\substack{w\in\tilde{\mathscr{X}} \\ b\in\mathbb{R}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f_{w,b}(\psi(x_i))) + \frac{\lambda}{2} \left(\|w\|^2 + b^2 \right) \right\}, \qquad (*)$$

ce qui correspond à la minimisation du risque empirique avec régularisation Ridge dans le problème auxiliaire.

Remarque. — $(*)$ est un problème d'optimisation sur $\tilde{\mathcal{X}} \times \mathbb{R}$. La dimension de l'espace de redescription $\tilde{\mathcal{X}}$ peut être grande, voire infinie. Grâce à l'astuce du noyau présenté ci-dessous, on va pouvoir simplifier $(*)$ en un problème d'optimisation sur $\mathbb{R}^n \times \mathbb{R}$ et rendre ainsi la solution calculable.

Théorème. — *Théorème de représentation. — Soit* $(\hat{w}, \hat{b}) \in \tilde{\mathcal{X}} \times \mathbb{R}$ *solution de* $(*)$. *Alors,* $\hat{w} \in \mathrm{Vect}_{i\in[n]}\, \psi(x_i)$. *Autrement dit, il existe* $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ *tel que :*

$$\hat{w} = \sum_{i=1}^{n} \alpha_i \psi(x_i).$$

Démonstration. — Voir TD. □

Définition. — *Matrice de Gram.* — Soit $m \geqslant 1$ un entier et $(\tilde{\mathcal{X}}, \langle \cdot, \cdot \rangle)$ un espace préhilbertien. La *matrice de Gram* associée à la famille $(x_1, \dots, x_m) \in \tilde{\mathcal{X}}^m$ est définie par :

$$\left(\left\langle x_i, x_j \right\rangle \right)_{1\leqslant i,\, j\leqslant m}.$$

Notation. — On rappelle qu'on note $A^\top$ la transposée d'une matrice A. Si x est un vecteur de taille m, il est vu comme un vecteur colonne (c'est-à-dire une matrice de taille $m \times 1$) dans un contexte de calcul matriciel. $x^\top$ désigne alors la matrice de taille $1 \times m$ (horizontale) correspondante.

Corollaire. — *Astuce du noyau. — Soit* G *la matrice de Gram associée à la famille* $(\psi(x_i), \dots, \psi(x_n))$. *Soit* $(\hat{\alpha}, \hat{b})$ *défini par :*

$$(\hat{\alpha}, \hat{b}) = \operatorname*{arg\,min}_{\substack{\alpha\in\mathbb{R}^n \\ b\in\mathbb{R}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \phi(\alpha^\top G e_i + b)) + \frac{\lambda}{2}(\alpha^\top G \alpha + b^2) \right\}, \qquad (**)$$

où pour $i \in [n]$ *on note* e_i *le* i*-ème vecteur de la base canonique de* $\mathbb{R}^n$. *On pose :*

$$\hat{w} = \sum_{i=1}^{n} \hat{\alpha}_i \psi(x_i).$$

Alors, $(\hat{w}, \hat{b})$ *est solution de* $(*)$. *De plus,*

$$\forall x \in \mathcal{X}, \quad (f_{\hat{w},\hat{b}} \circ \psi)(x) = \phi\left(\sum_{i=1}^{n} \hat{\alpha}_i K(x_i, x) + \hat{b} \right).$$

Démonstration. — Soit $w \in \tilde{\mathcal{X}}$, $b \in \mathbb{R}$ et $\alpha \in \mathbb{R}^n$ tels que $w = \sum_{i=1}^n \alpha_i \psi(x_i)$. Notons alors :

$$\mathrm{F}(w,b) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, f_{w,b}(\psi(x_i))) + \frac{\lambda}{2}\left(\|w\|^2 + b^2\right),$$

et on peut écrire :

$$\begin{aligned}
\mathrm{F}(w,b) &= \frac{1}{n}\sum_{i=1}^n \ell(y_i, \phi(\langle w, \psi(x_i)\rangle + b)) + \frac{\lambda}{2}\left(\|w\|^2 + b^2\right) \\
&= \frac{1}{n}\sum_{i=1}^n \ell\left(y_i, \phi\left(\sum_{k=1}^n \alpha_k \langle \psi(x_k), \psi(x_i)\rangle + b\right)\right) \\
&\qquad + \frac{\lambda}{2}\left(\left\langle \sum_{i=1}^n \alpha_i \psi(x_i), \sum_{i=1}^n \alpha_i \psi(x_i)\right\rangle + b^2\right) \\
&= \frac{1}{n}\sum_{i=1}^n \ell(y_i, \phi(\alpha^\top \mathrm{G} e_i + b)) + \frac{\lambda}{2}\left(\alpha^\top \mathrm{G}\alpha + b^2\right).
\end{aligned}$$

Notons :

$$\mathrm{H}(\alpha,b) = \frac{1}{n}\sum_{i=1}^n \ell(y_i, \phi(\alpha^\top \mathrm{G} e_i + b)) + \frac{\lambda}{2}\left(\alpha^\top \mathrm{G}\alpha + b^2\right).$$

Soit $(\hat{w}', \hat{b}')$ une solution de $(*)$. D'après le théorème de représentation, il existe $\hat{\alpha}' \in \mathbb{R}^n$ tel que $\hat{w}' = \sum_{i=1}^n \hat{\alpha}'_i \psi(x_i)$. On a alors :

$$\mathrm{F}(\hat{w}, \hat{b}) = \mathrm{H}(\hat{\alpha}, \hat{b}) \leqslant \mathrm{H}(\hat{\alpha}', \hat{b}') = \mathrm{F}(\hat{w}', \hat{b}').$$

où l'inégalité est valable car $(\hat{\alpha}, \hat{b})$ est solution de $(**)$ par hypothèse. On a donc $\mathrm{F}(\hat{w}, \hat{b}) \leqslant \mathrm{F}(\hat{w}', \hat{b}')$. Or $(\hat{w}', \hat{b}')$ est solution de $(*)$. $(\hat{w}, \hat{b})$ est donc également solution de $(*)$.

Enfin, pour $x \in \mathcal{X}$,

$$\begin{aligned}
(f_{\hat{w},\hat{b}} \circ \psi)(x) &= \phi\left(\langle \hat{w}, \psi(x)\rangle + \hat{b}\right) \\
&= \phi\left(\left\langle \sum_{i=1}^n \hat{\alpha}_i \psi(x_i), \psi(x)\right\rangle + \hat{b}\right) \\
&= \phi\left(\sum_{i=1}^n \hat{\alpha}_i \mathrm{K}(x_i, x) + \hat{b}\right).
\end{aligned}$$

□

Remarque. — Si on sait calculer K, grâce à l'astuce du noyau, on peut calculer le prédicteur $f_{\hat{w},\hat{b}} \circ \psi$ ainsi que ses prédictions $(f_{\hat{w},\hat{b}} \circ \psi)(x)$ sans avoir à manipuler ni ψ, ni aucun vecteur dans $\tilde{\mathcal{X}}$.

5. Caractérisation des noyaux

Définition. — Soit $m \geqslant 1$ un entier. Une matrice réelle $\mathrm{M} = (\mathrm{M}_{ij})_{1 \leqslant i,j \leqslant m}$ est *(semi-définie) positive* si :

(i) $\forall i, j \in [m],\ \mathrm{M}_{ij} = \mathrm{M}_{ji}$;

(ii) $\forall u \in \mathbb{R}^m,\ u^\top \mathrm{M} u \geqslant 0$.

Théorème. — *Caractérisation des noyaux. — Soit $\mathcal{X}$ un ensemble quelconque et $\mathrm{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Les deux propositions suivantes sont équivalentes :*

(i) Il existe un espace préhilbertien $\tilde{\mathcal{X}}$ et une application $\psi : \mathcal{X} \to \tilde{\mathcal{X}}$ tels que K *est le noyau associé.*

(ii) Pour tout $m \geqslant 1$, et $(x_1, \dots, x_m) \in \mathcal{X}^m$, la matrice $(\mathrm{K}(x_i, x_j))_{1 \leqslant i,j \leqslant m}$ est semi-définie positive.

Démonstration. — Voir le TD pour le sens direct. Le sens indirect est admis. □