Espaces de Hilbert à noyeaux reproduisants (RKHS)

1 Motivations

On considère un problème d'estimation :
$\forall i \in [\![1, n]\!], \quad y_i = f(x_i, \beta) \quad$ avec $\beta \in \mathbb{R}^p, \quad x_i \in \mathbb{R}^p, \quad y_i \in \mathbb{R}$.

$$\text{On pose : } X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times p} \text{ et } Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n.$$

- Estimation linéaire : $f(x, \beta) = x^T \beta$.
  On utilise la régularisation de Tikhonov : $\min\limits_{\beta \in \mathbb{R}^p} \frac{1}{2}\|X\beta - Y\|_2^2 + \frac{\lambda}{2}\|\beta\|_2^2$.

  $$\begin{aligned} \text{On a la CN1 (condition nécessaire d'ordre 1) : } \nabla f(\beta) = 0. \\ \Leftrightarrow X^T(X\beta - Y) + \lambda\beta = 0. \\ \Leftrightarrow X^T X\beta + \lambda\beta = X^T Y. \\ \Leftrightarrow (X^T X + \lambda I_p)\beta = X^T Y. \\ \Leftrightarrow \hat{\beta} = (X^T X + \lambda I_p)^{-1} X^T Y. \end{aligned}$$

  Donc la prédiction sera : $\hat{Y} = X\hat{\beta} = X(X^T X + \lambda I_p)^{-1} X^T Y$.
  $\to$ Complexité en $\mathcal{O}(p^2 n)$ ou $\mathcal{O}(n^2 p)$ (selon $X^T X$ ou XX^T).

- Estimation quadratique : $f(x, \beta, \theta) = x^T\theta x + x\beta = \sum_{k=0}^p x_k\beta_k + \sum_{k,l} x_k x_l \theta_{kl}$.
  2 inconnues : $\beta \in \mathbb{R}^p$ et $\theta \in \mathbb{R}^{p \times p}$.

  $$\text{Soit } \tilde{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \\ \theta_{11} \\ \vdots \\ \theta_{pp} \end{pmatrix} \in \mathbb{R}^{p+p^2} \quad , \quad \tilde{x}_i = \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^p \\ x_i^1 x_i^1 \\ \vdots \\ x_i^p x_i^p \end{pmatrix} \quad , \quad \tilde{X} = \begin{pmatrix} \tilde{x_1}^T \\ \vdots \\ \tilde{x_n}^T \end{pmatrix}.$$

  $\to$ Complexité en $\mathcal{O}(np^4)$ ou $\mathcal{O}(n^2 p^2)$ (selon $\tilde{X}^T\tilde{X}$ ou $\tilde{X}\tilde{X}^T$).

1. Est-ce qu'il existe une fonction qui permet de rester sur une résolution linéaire tout en augmentant la dimension du modèle ?
   $\rightarrow$ Oui, mais (quasi) jamais explicitée.

2. Peut-on s'épargner la complexité du modèle en gardant la complexité des données initiales ?
   $\rightarrow$ Oui, avec un noyau.

$$\text{Ici,}\quad \begin{array}{rccl} \phi_i : & \mathbb{R}^p & \longrightarrow & \mathbb{R}^{p+p^2} \\ & x_i & \longmapsto & \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^p \\ x_i^1 x_i^1 \\ \vdots \\ x_i^p x_i^p \end{pmatrix} = \tilde{x}_i \end{array}$$

Si on se ramène à $\tilde{x_i}^T \tilde{x_j} = (1 + x_i^T x_j)^2$, on n'a plus que des produits de $(x_i^T x_j)$, donc une complexité en $\mathcal{O}(np^2)$.

On définit une application $k : \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \mathbb{R}$ tel que $k(x_i, x_j) = (1 + x_i^T x_j)^2$.

Méthode générale :

On cherche à prédire $y \in \mathbb{R}$ avec un prédicteur $x \in \mathbb{R}^p$.

- Afin d'appliquer l'approche de régularisation de Tikhonov, on va chercher $\phi : \mathbb{R}^p \longrightarrow \mathbb{R}^d$ avec d à préciser ($d \in \{\mathbb{N}, \infty\}$).
- On va appliquer la régularisatiion aux données transformées $(\phi(x_i))_{i \in [\![1,n]\!]}$ comme prédicteur.
- On va être amené à construire un noyau $K \in \mathcal{M}_p(\mathbb{R})$ avec $K_{ij} = \langle \phi(x_i), \phi(x_j) \rangle$.
  $\rightarrow$ On va chercher $\begin{array}{rccl} k : & \mathbb{R}^p \times \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & (x_i, x_j) & \longmapsto & \langle \phi(x_i), \phi(x_j) \rangle \end{array}$

2 Espaces de Hilbert à noyeau reproduisant

Soit $\mathcal{H}$ un espace de Hilbert (muni du produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$) et $X \subset \mathcal{H}$.

Propriété

Soit $\phi : X \longrightarrow \mathcal{H}$.
On définit $k : \begin{array}{ccc} X \times X & \longrightarrow & \mathbb{R} \\ (x, x') & \longmapsto & \langle \phi(x), \phi(x') \rangle_{\mathcal{H}} \end{array}$

Soit K la matrice définie par :
$\forall i, j \in [\![1, n]\!], \quad K_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$.

Alors K est symétrique positive.

- Soit $\alpha \in \mathbb{R}^n$, $(x_i) \in X^n$.

On a : $\alpha^T K \alpha = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$

$$= \sum_{i,j} \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$$

$$= \langle \sum_{i}^{n} \alpha_i \phi(x_i), \sum_{j}^{n} \alpha_j \phi(x_j) \rangle_{\mathcal{H}}$$

$$= ||\mathcal{Z}||_{\mathcal{H}}^2 \geq 0 \qquad \text{avec } \mathcal{Z} = \sum_{i}^{n} \alpha_i \phi(x_i)$$

Définition

Un *noyau défini positif* (ou *noyau*) est une application symétrique : $k : X \times X \longrightarrow \mathbb{R}$ telle que $\forall (x_i) \in X^n$, la matrice $K \in \mathcal{M}_n(\mathbb{R})$ définie par $K_{ij} = k(x_i, x_j)$ est symétrique positive.

Propriété

Soit $(k_i)_{i \in J}$ une famille de noyaux.

i) Soit $(\alpha_i, \alpha_j) \in \mathbb{R}_+^2$. Alors $\alpha_i k_i + \alpha_j k_j$ est un noyau.

ii) Si $\exists \lim_{p \to +\infty} k_p(x, x'), \forall (x, x') \in X^2$, alors k la limite simple de $(k_p)_{p \in J}$ est un noyau.

iii) Le produit (terme à terme) $k_i k_j$ est un noyau.
$k(x, x') = k_i(x, x') k_j(x, x')$

Exercice :

1. Noyeau linéaire : $X = \mathbb{R}^p$ et $\forall (x, x') \in X^2, \quad k(x, x') = x^T x'$.
2. Noyeau polynomial : $X = \mathbb{R}^p$ et $\forall (x, x') \in X^2, \quad k(x, x') = (1 + x^T x')^d$.
3. Noyeau exponentiel : $X = \mathbb{R}^p$ et $\forall (x, x') \in X^2, \quad k(x, x') = \exp(-\frac{||x - x'||^2}{2\sigma^2})$. avec $\sigma > 0$.

Montrer que les noyeaux polynomial et exponentiel sont des noyaux.

Correction :

1. Noyeau par définition.

2. Par produit de noyaux.

3. $k(x,x') = \exp(-\frac{||x-x'||^2}{2\sigma^2}) = \exp(-\frac{||x||^2}{2\sigma^2})\exp(\frac{x^Tx'}{\sigma^2})\exp(-\frac{||x'||^2}{2\sigma^2}) = k_1(x,x')k_2(x,x')$. avec $k_1(x,x') = \exp(-\frac{||x||^2}{2\sigma^2})\exp(-\frac{||x'||^2}{2\sigma^2})$ et $k_2(x,x') = \exp(\frac{x^Tx'}{\sigma^2})$.

   $k_1(x,x') = \langle\phi(x),\phi(x')\rangle$ avec $\phi(x) = \exp(-\frac{||x||^2}{2\sigma^2})$. k_1 noyeau (proposition 1).
   $k_2(x,x') = \sum_{i=0}^{+\infty}\frac{1}{i!}\left(\frac{x^Tx'}{\sigma^2}\right)^i$. k_2 noyeau (proposition 2).

Théorème de Moore-Aronszajn

Soit k un noyeau, $k : X \times X \longrightarrow \mathbb{R}$.
Alors il existe un espace de Hilbert $\mathcal{H}$ et une application $\phi : X \longrightarrow \mathcal{H}$ tel que $\forall (x,x') \in X^2, \quad k(x,x') = \langle\phi(x),\phi(x')\rangle_{\mathcal{H}}$.

▶ Soit $\mathcal{H} = \{f : X \to \mathbb{R}$ tel que $\exists n \in \mathbb{N}, \exists(\alpha_j)_{j\in[\![1,n]\!]} \in \mathbb{R}^n$, $\exists(x_j)_{j\in[\![1,n]\!]} \in X^n, \quad f(\cdot) = \sum_{j=1}^n \alpha_j k(\cdot,x_j)\}$.

Texte manquant.

3 Applications

3.1 Régularisation de Tikhonov

Le problème s'écrit :

$$\text{(P)} : \quad \min_{\beta\in\mathbb{R}^p} \frac{1}{2}||X\beta - Y||_2^2 + \frac{\lambda}{2}||\beta||_2^2 \tag{1}$$

Soit le RKHS de noyau reproduisant le noyau linéaire.

$f \in \mathcal{H}$: $\sum_{i=1}^m \alpha_i k(\cdot,x_i) = k(\cdot,\sum_{i=1}^m \alpha_i x_i)$ (par linéarité du noyau).
$f(x) = x^T\tilde{x}$ avec $\tilde{x} = \sum_{i=1}^m \alpha_i x_i$.

$f(\tilde{x}) = \tilde{x}^T\tilde{x} = ||\tilde{x}||_2^2$.

$$\begin{aligned}
\text{et } ||f||_{\mathcal{H}}^2 &= \langle f,f\rangle_{\mathcal{H}} \\
&= \langle\sum_{i=1}^m \alpha_i k(\cdot,x_i), \sum_{j=1}^m \alpha_j k(\cdot,x_j)\rangle_{\mathcal{H}} \\
&= \langle k(\cdot,\tilde{x}), k(\cdot,\tilde{x})\rangle_{\mathcal{H}} \\
&= k(\tilde{x},\tilde{x}) \\
&= ||\tilde{x}||_2^2
\end{aligned}$$

Donc (P) s'écrit :

$$(\text{P}) : \quad \min_{f \in \mathcal{H}} \frac{1}{2} \sum_{i=1}^{m} (f(x_i) - y_i)^2 + \frac{\lambda}{2} ||f||_{\mathcal{H}}^2 \tag{2}$$

D'après le théorème du représentant, $\hat{f} \in \text{Vect}(k(\cdot, x_i))_{i \in [\![1,m]\!]}$.

$\hat{f} = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i)$.

On pose $K = [k(x_i, x_j)]_{i,j \in [\![1,m]\!]^2} \in \mathcal{M}_m(\mathbb{R})$.

$$\begin{aligned} \text{On a : } ||\hat{f}||_{\mathcal{H}}^2 &= \langle \sum_{i=1}^{m} \alpha_i k(\cdot, x_i), \sum_{j=1}^{m} \alpha_j k(\cdot, x_j) \rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \\ &= \alpha^T K \alpha \end{aligned}$$

et $f(x_i) = \sum_{j=1}^{m} \alpha_j k(x_i, x_j) = [K\alpha]_i$.

D'où (P') $\Leftrightarrow \min_{\alpha \in \mathbb{R}^m} \frac{1}{2} ||Y - K\alpha||_2^2 + \frac{\lambda}{2} \alpha^T K \alpha$.

$$\begin{aligned} \text{On a la CN1 : } \nabla f(\alpha) &= 0 \\ &\Leftrightarrow -K^T(Y - K\alpha) + \lambda K \alpha = 0 \\ &\Leftrightarrow K^T K \alpha + \lambda K \alpha = K^T Y \\ &\Leftrightarrow (K^T K + \lambda)\alpha = K^T Y \\ &\Leftrightarrow \hat{\alpha} = (K^T K + \lambda K)^{-1} K^T Y \end{aligned}$$

Et la prédiction : $K\hat{\alpha} = K(K^T K + \lambda K)^{-1} K^T Y$.