Les reproducing kernel Hilbert spaces

Présentation du TER, supervisé par Christophe Giraud

Matthieu Denis

Université Paris Saclay

25 juin 2021

Contexte

- X un ensemble quelconque
- H un espace de Hilbert de fonctions réelles sur X
- $\forall x \in X,\ L_x$ une forme linéaire sur H :

$$L_x : f \mapsto f(x)\ \forall f \in H$$

Définitions

Définition : RKHS

L_x est bornée sur H, i.e :

$$\forall x \in X,\ \exists M_x > 0,\ \forall f \in H \text{ t.q } |L_x(f)| := |f(x)| \leq M_x ||f||_H$$

Définitions

Définition : Noyau / Kernel

Une fonction $k : X \times X \to \mathbb{R}$ est un noyau si

$$\exists \phi : X \to H \text{ t.q } k(x,y) = \langle \phi(x), \phi(y) \rangle_H \; \forall x, y \in X$$

Définitions

Définition : Noyau / Kernel

Une fonction $k : X \times X \to \mathbb{R}$ est un noyau si

$$\exists \phi : X \to H \text{ t.q } k(x, y) = \langle \phi(x), \phi(y) \rangle_H \; \forall x, y \in X$$

Définition : Noyau Reproduisant / Reproducing Kernel

Une fonction $k : X \times X \to \mathbb{R}$ est un noyau reproduisant de H si $\forall x \in X$, $f \in H$:

- $k(\cdot, x) \in H$
- $f(x) = \langle f, k(\cdot, x) \rangle_H$

Noyaux classiques

- $k(x, y) := \langle x, y \rangle$
- $k(x, y) := (\alpha\langle x, y \rangle + 1)^d, \ \alpha \in \mathbb{R}, \ d \in \mathbb{N}$
- $k(x, y) := \exp(||x - y||^2/(2\sigma^2)), \ \sigma > 0$
- $k(x, y) := \exp(||x - y||/\sigma), \ \sigma > 0$

Résultats importants

Propriété :

Noyau reproduisant $\Longrightarrow$ Noyau.

Résultats importants

Propriété :

Noyau reproduisant $\Longrightarrow$ Noyau.

Théorème :

H est un RKHS $\Longleftrightarrow \exists!$ noyau reproduisant de H.

Résultats importants

Propriété :
Noyau reproduisant $\Longrightarrow$ Noyau.

Théorème :
H est un RKHS $\Longleftrightarrow \exists!$ noyau reproduisant de H.

Théorème de Moore-Aronszaj :
Soit k un noyau. Alors $\exists!$ espace de Hilbert H de fonctions sur X pour lequel k est un noyau reproduisant.

Application au ML : Le Representer Theorem

Soit k un kernel sur X et soit H son RKHS associée. Posons $x_1, \cdots, x_n \in X$ notre training sample. Regardons le problème d'optimisation suivant :

$$\min_{f \in H} J(f) := E(f(x_1), \cdots, f(x_n)) + P(||f||_H^2)$$

Où P est une fonction croissante.
Alors si ce problème d'optimisation a (au-moins) une solution, il y a (au-moins) une solution de la forme

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

De plus, si P est strictement croissante, alors toute solution a cette forme.

Kernel Ridge Regression

Ici, $J(f) := \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda||f||_H^2$, prenons k un kernel sur X. Le representer theorem nous dit que la solution de ce problème (sous couvert d'existence) est nécessairement de la forme

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

Kernel Ridge Regression

Ici, $J(f) := \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda||f||_H^2$, prenons k un kernel sur X. Le representer theorem nous dit que la solution de ce problème (sous couvert d'existence) est nécessairement de la forme

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

$$\begin{aligned}
&\min_{f \in H} J(f) := \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda||f||_H^2 \\
\iff &\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n}(y_i - \sum_{j=1}^{n} \alpha_j \cdot k(x_i, x_j))^2 + \lambda \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) \\
\iff &\min_{\alpha \in \mathbb{R}^n} ||y - K\alpha||_2^2 + \lambda \alpha^T K \alpha \text{ avec } K_{ij} := k(x_i, x_j)
\end{aligned}$$

SVM

Dans le cadre d'une SVM, $J(f) := \frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i f(x_i)) + \frac{\lambda}{2}||f||_H^2$. Prenons un kernel k sur X. Encore une fois, le representer theorem nous dit que la seule solution (si elle existe) est sous la forme

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

SVM

Dans le cadre d'une SVM, $J(f) := \frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i f(x_i)) + \frac{\lambda}{2}||f||_H^2$. Prenons un kernel k sur X. Encore une fois, le representer theorem nous dit que la seule solution (si elle existe) est sous la forme

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

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} max(0, 1 - y_i \sum_{j=1}^{n} \alpha_j \cdot k(x_i, x_j)) + \frac{\lambda}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(x_i, x_j)$$

SVM

Dans le cadre d'une SVM, $J(f) := \frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i f(x_i)) + \frac{\lambda}{2}||f||_H^2$. Prenons un kernel k sur X. Encore une fois, le representer theorem nous dit que la seule solution (si elle existe) est sous la forme

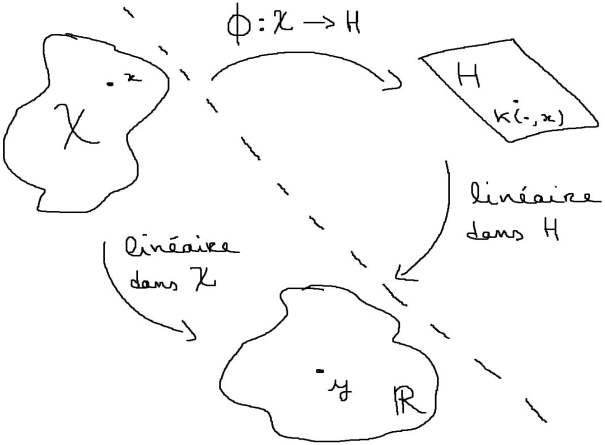
$$f = \sum_{i=1}^{n} \alpha_i \cdot k(\cdot, x_i)$$

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} max(0, 1 - y_i \sum_{j=1}^{n} \alpha_j \cdot k(x_i, x_j)) + \frac{\lambda}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(x_i, x_j)$$

On peut montrer que le dual de ce problème est :

$$\min_{\gamma \in \mathbb{R}^n} -\sum_{i=1}^{n} \gamma_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_i \gamma_j y_i y_j k(x_i, x_j) \text{ t.q } 0 \leq \gamma_i \leq \frac{1}{n\lambda},\ \alpha_i = y_i \gamma_i \forall i$$

Visualisation



Apparition des RKHS dans le cas d'un réseau de neurone simple

$\Phi : (\mathbb{R}^m \times \mathbb{R}^{m\times m} \times \mathbb{R}^m) \times \mathbb{R} \to \mathbb{R}$ combinaison d'applications linéaires, sans non linéarités intermédiaires :

$$\Phi((\beta, A, u), x) := \frac{1}{m^\alpha}\beta^T\left(\frac{1}{m^\gamma}A\right)ux$$

Apparition des RKHS dans le cas d'un réseau de neurone simple

$\Phi : (\mathbb{R}^m \times \mathbb{R}^{m\times m} \times \mathbb{R}^m) \times \mathbb{R} \to \mathbb{R}$ combinaison d'applications linéaires, sans non linéarités intermédiaires :

$$\Phi((\beta, A, u), x) := \frac{1}{m^\alpha}\beta^T \left(\frac{1}{m^\gamma}A\right) ux$$

On initialise $\theta^0 := (\beta^0, A^0, u^0)$ de manière standarde :
$\forall i, j \in \{1, \cdots, m\},\ u_i^0, A_{ij}^0, \beta_i^0 \sim_{iid} N(0,1)$

Apparition des RKHS dans le cas d'un réseau de neurone simple

Posons une fonction de perte $F : \mathbb{R} \to \mathbb{R}$ t.q $F'(0) \neq 0$. Lorsque $\alpha < 1$, on a pour m grand :

$$||\nabla_{\beta/u/A} F(\Phi(\theta^{t+1}, x)) - \nabla_{\beta/u/A} F(\Phi(\theta^{t}, x))|| \ll ||\nabla_{\beta/u/A} F(\Phi(\theta^{t}, x))||$$

Apparition des RKHS dans le cas d'un réseau de neurone simple

Posons une fonction de perte $F : \mathbb{R} \to \mathbb{R}$ t.q $F'(0) \neq 0$. Lorsque $\alpha < 1$, on a pour m grand :

$$||\nabla_{\beta/u/A} F(\Phi(\theta^{t+1}, x)) - \nabla_{\beta/u/A} F(\Phi(\theta^t, x))|| \ll ||\nabla_{\beta/u/A} F(\Phi(\theta^t, x))||$$

$$F(\Phi(\theta^T, x)) = F(\Phi(\theta^0, x)) + \langle \theta^T - \theta^0, \nabla_\theta F(\Phi(\theta^0, x)) \rangle + \mathcal{O}(m^{-1/2})$$

Apparition des RKHS dans le cas d'un réseau de neurone simple

Posons une fonction de perte $F : \mathbb{R} \to \mathbb{R}$ t.q $F'(0) \neq 0$. Lorsque $\alpha < 1$, on a pour m grand :

$$||\nabla_{\beta/u/A} F(\Phi(\theta^{t+1}, x)) - \nabla_{\beta/u/A} F(\Phi(\theta^{t}, x))|| \ll ||\nabla_{\beta/u/A} F(\Phi(\theta^{t}, x))||$$

$$F(\Phi(\theta^{T}, x)) = F(\Phi(\theta^{0}, x)) + \langle \theta^{T} - \theta^{0}, \nabla_{\theta} F(\Phi(\theta^{0}, x)) \rangle + \mathcal{O}(m^{-1/2})$$

On apprend donc un modèle linéaire relatif aux features $\nabla_{\theta} F(\Phi(\theta^{0}, x))$, c'est-à-dire qu'après la transformation $x \to \nabla_{\theta} F(\Phi(\theta^{0}, x))$, on est linéaire. On fait donc face à un RKHS de noyau (par définition)

$$k(x, y) = \langle \nabla_{\theta} F(\Phi(\theta^{0}, x)), \nabla_{\theta} F(\Phi(\theta^{0}, y)) \rangle \xrightarrow{LGN} \mathbb{E}[\langle \nabla_{\theta} F(\Phi(\theta^{0}, x)), \nabla_{\theta} F(\Phi(\theta^{0}, y)) \rangle]$$