Séminaire TEST

Andrés SÁNCHEZ PÉREZ

November 27th, 2014

1 RKHS of vector-valued functions

1.1 Introduction

In our framework we deal with complicated, structured, high-dimensional data (images, texts, time series, graphs, distributions, permutations...) belonging to a set $\mathcal{X}$. Kernel methods is a collection of algorithms that study the more typical problems in machine learning (clustering, ranking, classification, etc.), without doing any assumption regarding the type of data.
The idea is to rely on a comparison or similarity function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and to represent the set of data points $\mathcal{S} = \{x_1, \ldots, x_n\}$ by the $n \times n$ matrix $[K]_{ij} = K(x_i, x_j)$.

Definition 1 *A positive definite (p.d.) kernel on the set $\mathcal{X}$ is a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ symmetric :*

$$K(x, x') = K(x', x)\,, \text{ for all } x, x' \in \mathcal{X}\ .$$

and which satisfies, for all $N \in \mathbb{N}, (x_1, \ldots, x_N) \in \mathcal{X}^N$ and $(a_1, \ldots, a_N) \in \mathbb{R}^N$:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K\left(x_i, x_j\right) \geq 0\ .$$

Equivalently, a kernel K is p.d. if and only if, for any $N \in \mathbb{N}$ and any set of points $(x_1, \ldots, x_N) \in \mathcal{X}^N$, the similarity matrix $[K]_{ij} = K(x_i, x_j)$ is positive semidefinite.
Kernel methods are algorithm that take such matrices as input.

Example 1 (The linear kernel) *Let $\mathcal{X} = \mathbb{R}^d$. The function $K : \mathcal{X}^2 \to \mathbb{R}$ is defined by*

$$K(x, x') = \langle x, x' \rangle_{\mathbb{R}^d}, \text{ for all } x, x' \in \mathcal{X}\ .$$

Lemma 1 *Let $\mathcal{X}$ be any set and $\phi : \mathcal{X} \to \mathbb{R}^d$. The function $K : \mathcal{X}^2 \to \mathbb{R}$, defined by*

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathbb{R}^d}, \text{ for all } x, x' \in \mathcal{X}\ ,$$

is p.d.

Theorem 1.1 *[1] K is a p.d. kernel on the set $\mathcal{X}$ if and only if there exists a Hilbert space $\mathcal{H}$ and a mapping $\phi : \mathcal{X} \to \mathcal{H}$, such that, for any $x, x' \in \mathcal{X}$:*

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}\ .$$

Definition 2 *Let $\mathcal{X}$ be a set and $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ be a class of functions forming a real Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The function $K : \mathcal{X}^2 \to \mathbb{R}$ is called a reproducing kernel (r.k.) of $\mathcal{H}$ if*

(i) $\mathcal{H}$ contains all functions of the form $K_x : y \mapsto K(x, y)$, for all $x \in \mathcal{H}$,

(ii) For all $x \in \mathcal{X}$ and $f \in \mathcal{H}$ the reproducing property holds : $f(x) = \langle f, K_x \rangle_{\mathcal{H}}$.

If a such r.k. exists, then $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS).

Theorem 1.2 *The Hilbert space $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ is a RKHS if and only if for any $x \in \mathcal{X}$, the mapping $F_x : \mathcal{H} \to \mathbb{R}$ defined by $F_x(f) = f(x)$ is continuous.*

Proof
Suppose that $\mathcal{H}$ is a RKHS, then a r.k. K exists. For any $x \in \mathcal{X}$, $f \in \mathcal{H}$

$$|f(x)| = |\langle f, K_x \rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} \|K_x\|_{\mathcal{H}} = \|f\|_{\mathcal{H}} K^{1/2}(x, x) \ .$$

If the mapping F_x is continuous we have that $F_x \in \mathcal{H}^*$. Riesz representation theorem implies that for any $x \in \mathcal{X}$ there exists $g_x \in \mathcal{H}$ such that $F_x = \langle \cdot, g_x \rangle_{\mathcal{H}}$. The function $K(x, y) = g_x(y)$ is a r.k. for $\mathcal{H}$.

■

Theorem 1.3 *If $\mathcal{H}$ is a RKHS, the it has a unique r.k. and conversely, a function K can be the r.k. of at most one RKHS.*

Theorem 1.4 *A function $K : \mathcal{X}^2 \to \mathbb{R}$ is p.d. if and only if it is a r.k.*

Theorem 1.5 (Representer Theorem) *Let $\mathcal{X}$ be a set endowed with a p.d. kernel K, $\mathcal{H}_K$ the corresponding RKHS, and $\mathcal{S} = \{x_1, \ldots, x_n\}$ a finite set of points in $\mathcal{X}$. Let $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ be a function of $n + 1$ variables, strictly increasing with respect to the last variable. Then, any solution to the optimization problem :*

$$\min_{f \in \mathcal{H}_K} \Psi \left(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}_K} \right)$$

admits a representation of the form :

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

Proof
Consider $\mathcal{H}_K^{\mathcal{S}}$, the linear span in $\mathcal{H}_K$ of the vectors K_{x_i}, i.e.,

$$\mathcal{H}_K^{\mathcal{S}} = \left\{ f \in \mathcal{H}_K : f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x), (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n \right\} \ .$$

$\mathcal{H}_K^{\mathcal{S}}$ is a closed subspace of $\mathcal{H}_K$, therefore any $f \in \mathcal{H}_K$ can be uniquely decomposed as $f = f_{\mathcal{S}} + f_{\perp}$ where $f_{\mathcal{S}} \in \mathcal{H}_K^{\mathcal{S}}$ and $f_{\perp} \in (\mathcal{H}_K^{\mathcal{S}})^{\perp}$.
Since $\mathcal{H}_K$ is a RKHS and for all $i = 1, \ldots, n$, $K_{x_i} \in \mathcal{H}_K^{\mathcal{S}}$, we have $f_{\perp}(x_i) = \langle f_{\perp}, K(x_i, \cdot) \rangle_{\mathcal{H}_K} = 0$. Hence, $f(x_i) = f_{\mathcal{S}}(x_i)$. Note also that $\|f\|_{\mathcal{H}_K}^2 = \|f_{\mathcal{S}}\|_{\mathcal{H}_K}^2 + \|f_{\perp}\|_{\mathcal{H}_K}^2$. These facts imply that

$$\Psi \left(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}_K} \right) \geq \Psi \left(f_{\mathcal{S}}(x_1), \ldots, f_{\mathcal{S}}(x_n), \|f_{\mathcal{S}}\|_{\mathcal{H}_K} \right) \ ,$$

with equality if and only if $f_{\perp} = 0$. The optimum of Ψ belongs necessarily to $\mathcal{H}_K^{\mathcal{S}}$.

■

1.2 Regression

Given n pairs $(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$, the regression problem consists in find a function $f : \mathcal{X} \to \mathbb{R}$ to predict y by $f(x)$. The prediction error is quantified by $(y - f(x))^2$. We need to fix a set of functions $\mathcal{H}$.
The least-square regression amounts to solve :

$$\widehat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \ .$$

But this problem is numerically unstable and the risk of overfitting is high in case of large $\mathcal{H}$.
Take $\mathcal{H} = \mathcal{H}_K$ the RKHS associated to a p.d. kernel K on $\mathcal{X}$. Let us regularize the functional to be minimized as in the following

$$\widehat{f} = \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|^2_{\mathcal{H}_K} \right\} \ . \tag{1}$$

We prevent overfitting by penalizing the non-smooth functions.

Theorem 1.6 *The minimizing problem (1) admits a solution $\widehat{f}$ with the following expansion*

$$\widehat{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$$

where $\boldsymbol{\alpha} = [\alpha_1 \ldots \alpha_n]'$ satisfies $\boldsymbol{\alpha} = (K + \lambda nI)^{-1}\boldsymbol{y}$, with K the Gram matrix ($K_{ij} = K(x_i, x_j)$) $\boldsymbol{y} = [y_1 \ldots y_n]'$.

Proof
Observe that $[\widehat{f}(x_1) \ldots \widehat{f}(x_n)]' = K\boldsymbol{\alpha}$ and that $\|\widehat{f}\|^2_{\mathcal{H}_K} = \boldsymbol{\alpha}' K \boldsymbol{\alpha}$. The problem is equivalent to

$$\widehat{f} = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \left\{ \frac{1}{n} (K\boldsymbol{\alpha} - \boldsymbol{y})' (K\boldsymbol{\alpha} - \boldsymbol{y}) + \lambda \boldsymbol{\alpha}' K \boldsymbol{\alpha} \right\} \ .$$

Since K is positive-semidefinite the expression between brackets is convex and differentiable on $\boldsymbol{\alpha}$. Its minimum can be determined by setting the gradient to zero :

$$K\left[(K + \lambda nI)\boldsymbol{\alpha} - \boldsymbol{y}\right] = 0 \ .$$

We obtain then that $(K + \lambda nI)\boldsymbol{\alpha} - \boldsymbol{y} \in \ker(K)$. Because $K + \lambda nI$ is positive definite and commutes with K we get the following chain of equivalences

$$\begin{aligned} (K + \lambda nI)\boldsymbol{\alpha} - \boldsymbol{y} &\in \ker(K) \Leftrightarrow \\ \boldsymbol{\alpha} - (K + \lambda nI)^{-1}\boldsymbol{y} &\in \ker(K) \Leftrightarrow \\ \boldsymbol{\alpha} = (K + \lambda nI)^{-1}\boldsymbol{y} &+ \boldsymbol{\epsilon}, \text{with}\, K\boldsymbol{\epsilon} = 0 \ . \end{aligned}$$

Suppose that f and f' are generated by $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ such that $\boldsymbol{\alpha} = \boldsymbol{\alpha}' + \boldsymbol{\epsilon}$ with $K\boldsymbol{\epsilon} = 0$. Then,

$$\left\|f - f'\right\|^2_{\mathcal{H}_K} = (\boldsymbol{\alpha} - \boldsymbol{\alpha}')' K (\boldsymbol{\alpha} - \boldsymbol{\alpha}') = 0 \ .$$

Thus, one solution to the problem is given by the claimed expression with $\boldsymbol{\alpha} = (K + \lambda nI)^{-1}\boldsymbol{y}$.

■

2 Proximal methods

2.1 Introduction

Proximal algorithms is a class of optimization algorithms, useful to solve convex, nonsmooth, constrained and large-scale problems.

Definition 3 (Proximal operator) *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a closed proper convex function, which means that its epigraph*

$$\mathbf{epi} f = \{(x,t) \in \mathbb{R}^n \times \mathbb{R} : f(x) \leq t\} \ ,$$

is a nonempty closed convex set. The effective domain of f is

$$\mathbf{dom} f = \{x \in \mathbb{R}^n : f(x) < +\infty\} \neq \emptyset \ ,$$

i.e., the set of points for which f takes on finite values. The proximal operator $\mathbf{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ *of f is defined by*

$$\mathbf{prox}_f(v) = \arg\min_x \left\{ f(x) + \frac{1}{2} \|x - v\|_2^2 \right\}$$

We will often encounter the proximal operator of the scaled function λf, with parameter $\lambda > 0$. It is

$$\mathbf{prox}_{\lambda f}(v) = \arg\min_x \left\{ f(x) + \frac{1}{2\lambda} \|x - v\|_2^2 \right\}$$

Evaluating $\mathbf{prox}_f$ involves solving a convex optimization problem. This is done via standard methods like BFGS (Broyden–Fletcher–Goldfarb–Shanno), but very often has an analytical solution or simple specialized linear-time algorithm.

Example 2 *Let I_C an indicator function of a convex set, i.e.*

$$I_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C \end{cases}$$

The proximal operator of I_C is the projection : $\mathbf{prox}_{\lambda I_C}(v) = \Pi(v) = \arg\min_{x \in C} \|x - v\|_2$.

Example 3 *Let $f(x) = (1/2)\boldsymbol{x'Px} + \boldsymbol{q'x} + \boldsymbol{r}$, then* $\mathbf{prox}_{\lambda f}(v) = (I + \lambda P)^{-1}(v - \lambda q)$

Theorem 2.1 (Fixed point) *The point x^* minimizes f if and only if $x^* = \mathbf{prox}_f(x^*)$.*

Proof
We assume for convenience that f is subdifferentiable on its domain, though the result is true in general.
If x^* minimizes f, i.e., then

$$f(x) + \frac{1}{2} \|x - x^*\|_2^2 \geq f(x^*) = f(x^*) + \frac{1}{2} \|x^* - x^*\|_2^2$$

Therefore, x^* also minimizes $f(x) + (1/2)\|x - x^*\|_2^2$.
Suppose now that x^* minimizes $f(x) + (1/2)\|x - x^*\|_2^2$. Using the subdifferential characterization of the function we get that $0 \in \partial f(x^*)$ where

$$\partial f(x) = \{y : f(z) \geq f(x) + y'(z - x), \forall z \in \mathbf{dom} f\} \ ,$$

and then x^* also minimizes f.

■

Lemma 2 *The following properties hold*

(i) *Separable sum : if f is block separable, so $f(x) = \sum_{i=1}^{N} f(x_i)$ then $(\mathbf{prox}_f(v))_i = \mathbf{prox}_{f_i}(v_i)$, for $i = 1, \ldots, N$. This is the key to parallel / distributed proximal algorithms.*

(ii) *Post composition : let $f, \phi : \mathbb{R}^n \rightarrow \mathbb{R}$ and $(a, b) \in \mathbb{R}_+ \times \mathbb{R}$. If $f(x) = a\phi(x) + b$ then $\mathbf{prox}_{\lambda f}(v) = \mathbf{prox}_{a\lambda\phi}(v)$.*

(iii) *Pre composition : let $f, \phi : \mathbb{R}^n \rightarrow \mathbb{R}$ and $(a, b) \in \mathbb{R}^* \times \mathbb{R}$. If $f(x) = \phi(ax + b)$ then $\mathbf{prox}_{\lambda f}(v) = (1/a)(\mathbf{prox}_{a^n\lambda\phi}(av + b) - b)$.*

(iv) *Affine addition : let $f, \phi : \mathbb{R}^n \rightarrow \mathbb{R}$ and $(a, b) \in \mathbb{R}^n \times \mathbb{R}$. If $f(x) = \phi(x) + a'x + b$ then $\mathbf{prox}_{\lambda f}(v) = \mathbf{prox}_{\lambda\phi}(v - \lambda a)$.*

Definition 4 *Let $\alpha \in (0, 1)$. An operator $T : \mathrm{dom}T = D \rightarrow D$ is nonexpansive if for all $x, y \in D$*

$$||Tx - Ty|| \leq ||x - y|| \ ,$$

and α-averaged if $T = (1 - \alpha)Id + \alpha R$ for some nonexpansive operator $R : \mathrm{dom}R = D \rightarrow D$. The class of α-averaged operators on D is denoted by $\mathcal{A}(\alpha)$. In particular, $\mathcal{A}(1/2)$ is the class of firmly nonexpansive operators.

Lemma 3 $\mathbf{prox}_f$ *is firmly nonexpansive.*

Lemma 4 *[3, Lemma 2.3] Suppose that $B : D \rightarrow D$ and $\beta \in (0, +\infty)$ satisfy $\beta B \in \mathcal{A}(1/2)$, and let $\gamma \in (0, 2\beta)$. Then, $I - \gamma B \in \mathcal{A}(\gamma/(2\beta))$.*

Theorem 2.2 (Krasnosel’skiĭ–Mann algorithm) *[2, Theorem 5.14] Let D be a nonempty closed convex subset of H, let $R : D \rightarrow D$ be a nonexpansive operator such that $\mathrm{Fix}R \neq \emptyset$, let $(\lambda_n)_{n\in\mathbb{N}}$ be a sequence in $[0, 1]$ such that $\sum_{n\in\mathbb{N}} \lambda_n(1 - \lambda_n) = +\infty$, and let $x_0 \in D$. Set, for all $n \in \mathbb{N}$*

$$x_{n+1} = (1 - \lambda_n)\, x_n + \lambda_n R x_n \ .$$

Then the following hold :

(i) *$(x_n)_{n\in\mathbb{N}}$ is Fejér monotone with respect to $\mathrm{Fix}R$, i.e., for all $x \in R$ and $n \in \mathbb{N}$, $||x_{n+1} - x|| \leq ||x_n - x||$.*

(ii) *$(Tx_n - x_n)_{n\in\mathbb{N}}$ converges strongly to 0.*

(iii) *$(x_n)_{n\in\mathbb{N}}$ converges weakly to a point in $\mathrm{Fix}R$.*

2.2 Proximal methods

Theorem 2.3 (Proximal minimization algorithm convergence) *Assume that $\arg\min f \neq \emptyset$. The proximal minimization algorithm, defined iteratively by*

$$\begin{aligned} x^{(0)} &\in \mathbb{R}^n \ , \\ x^{(k+1)} &= \mathbf{prox}_f\left(x^{(k)}\right) \ , \end{aligned}$$

converges.

Consider now the problem

$$\min \{f(x) + g(x)\} ,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, f is differentiable and g possibly nonsmooth.

Theorem 2.4 (Proximal gradient algorithm convergence) *Assume that ∇f is Lipschitz continuous with constant L. The proximal gradient algorithm, defined iteratively by*

$$\begin{array}{rcl} x^{(0)} & \in & \mathbb{R}^n , \\ x^{(k+1)} & = & \mathbf{prox}_{\lambda^{(k)} g}\left(x^{(k)} - \lambda^{(k)} \nabla f\left(x^{(k)}\right)\right) , \end{array}$$

converges if $\lambda^{(k)} = \lambda \in (0, 1/L)$.

Proof (Sketch)
A point x^* minimizes $f + g$ if and only if

$$0 \in \nabla f(x^*) + \partial g(x^*)$$

If and only if, for any $\lambda > 0$ the following equivalent statement hold :

$$\begin{array}{rcl} 0 & \in & \lambda \nabla f(x^*) + \lambda \partial g(x^*) \\ 0 & \in & \lambda \nabla f(x^*) - x^* + x^* + \lambda \partial g(x^*) \\ (I - \lambda \nabla f)(x^*) & \in & (I + \lambda \partial g)(x^*) \\ x^* & \in & (I + \lambda \partial g)^{-1} (I - \lambda \nabla f)(x^*) \\ x^* & = & \mathbf{prox}_{\lambda g}(x^* - \lambda \nabla f(x^*)) \end{array}$$

The operator $(I + \lambda \partial g)^{-1} (I - \lambda \nabla f)$ is averaged as composition of averaged operators [3, Lemma 2.2].

■

Theorem 2.5 (Accelerated proximal gradient algorithm convergence) *Assume that ∇f is Lipschitz continuous with constant L. The accelerated proximal gradient algorithm, defined iteratively by*

$$\begin{array}{rcl} x^{(0)} & \in & \mathbb{R}^n, \omega^{(0)} = 0 , \\ y^{(k+1)} & = & x^{(k)} + \omega^{(k)}\left(x^{(k)} - x^{(k-1)}\right) \\ x^{(k+1)} & = & \mathbf{prox}_{\lambda^{(k)} g}\left(y^{(k+1)} - \lambda^{(k)} \nabla f\left(y^{(k+1)}\right)\right) , \end{array}$$

converges if $\lambda^{(k)} = \lambda \in (0, 1/L)$ and $\omega^{(k)} = k/(k+3)$.

Références

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68 :337–404, 1950.

[2] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces.* CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011. With a foreword by Hédy Attouch.

[3] Patrick L. Combettes. Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5-6) :475–504, 2004.