

Kernel methods in machine learning — Homework 1

Hugo Cisneros

March 13th, 2019

Exercise 1. Kernels

1.1 K is symmetric since

$$\forall x, y \in \mathbb{R}^2 \quad K(y, x) = \cos(-(x - y)) = \cos(x - y) = K(x, y)$$

and because for all sequences of (x_i) and (a_i) of \mathbb{R}

$$\begin{aligned} \sum_i \sum_j a_i a_j \cos(x_i - x_j) &= \sum_i \sum_j a_i a_j (\cos x_i \cos x_j + \sin x_i \sin x_j) \\ &= \left(\sum_i a_i \cos x_i \right)^2 + \left(\sum_i a_i \sin x_i \right)^2 \geq 0 \end{aligned}$$

K is p.d.

1.2 K is clearly symmetric and $\forall x, y \in \mathcal{X}$,

$$\begin{aligned} K(x, y) &= \frac{1}{1 - x^T y} \\ &= \lim_{n \rightarrow \infty} \sum_{k=0}^n (x^T y)^k \quad (\text{converges because by C-S } |x^T y| \leq \|x\|_2 \|y\|_2 < 1) \end{aligned}$$

Since all powers of the linear kernel are p.d. kernels, and a sum of p.d. kernels is a p.d. kernel, the above sum is a p.d. kernel for all n . Therefore, K is p.d. as a point-wise limit of a sequence of p.d. kernels.

1.3 The kernel K is symmetric because the intersection is a commutative operation.

$$\begin{aligned} \sum_{i=1}^n a_i a_j (P(A_i \cap A_j) - P(A_i)P(A_j)) &= \sum_{i=1}^n a_i a_j (\mathbb{E}[\mathbb{1}_{A_i \cap A_j}] - \mathbb{E}[\mathbb{1}_{A_i}]\mathbb{E}[\mathbb{1}_{A_j}]) \\ &= \mathbb{E} \left[\sum_{i=1}^n a_i a_j \mathbb{1}_{A_i} \mathbb{1}_{A_j} \right] - \sum_{i=1}^n \mathbb{E}[a_i \mathbb{1}_{A_i}] \mathbb{E}[a_j \mathbb{1}_{A_j}] \\ &= \mathbb{E} \left[\left(\sum_{i=1}^n a_i \mathbb{1}_{A_i} \right)^2 \right] - \left(\sum_{i=1}^n a_i \mathbb{E}[\mathbb{1}_{A_i}] \right)^2 \end{aligned}$$

By Jensen's inequality, the above quantity is positive because the function $\phi : (X_1, \dots, X_n) \mapsto (\sum_{i=1}^n a_i X_i)^2$ is convex. Therefore, K is **p.d.**

1.4 K is symmetric.

Let $x, y \in \mathcal{X}$ such that $0 < f(x)g(y) \leq f(y)g(x)$. The inequality implies $g(x) \neq 0$ and $g(y) \neq 0$. By dividing the inequality above and repeating the process for the opposite inequality, we obtain the following equality

$$\min(f(x)g(y), f(y)g(x)) = g(x)g(y) \min\left(\frac{f(x)}{g(x)}, \frac{f(y)}{g(y)}\right)$$

which holds for $g(x) \neq 0$ and $g(y) \neq 0$. Let $(x_i)_{i \in \mathbb{N}}$ a finite sequence of elements of \mathcal{X} and (a_i) a other sequence of scalars. We can ignore the terms for which $K(x_i, x_i) = 0$ in the sum since they do not account for the total and therefore assume $\forall x_i g(x_i) \neq 0$

$$\begin{aligned} \sum_{i,j} a_i a_j K(x_i, x_j) &= \sum_{i,j} a_i a_j \min(f(x_i)g(x_j), f(x_j)g(x_i)) \\ &= \sum_{i,j} a_i a_j g(x_i)g(x_j) \min\left(\frac{f(x_i)}{g(x_i)}, \frac{f(x_j)}{g(x_j)}\right) \\ &= \sum_{i,j} a_i a_j g(x_i)g(x_j) \int_0^{+\infty} \mathbb{1}_{t \leq \frac{f(x_i)}{g(x_i)}}(t) \mathbb{1}_{t \leq \frac{f(x_j)}{g(x_j)}}(t) dt \\ &= \int_0^{+\infty} \left(\sum_i a_i g(x_i) \mathbb{1}_{t \leq \frac{f(x_i)}{g(x_i)}}(t) \right)^2 dt \geq 0 \end{aligned}$$

K is p.d..

1.5

Exercise 2. RKHS

2.1 $\alpha K_1 + \beta K_2$ is p.d. as sum of p.d. kernels and because α and β are positive scalars. Let \mathcal{H}_1 and \mathcal{H}_2 be their respective RKHS.

The RKHS of αK_1 is $\tilde{\mathcal{H}}_1 = \sqrt{\alpha} \mathcal{H}_1$ endowed with the inner product $\langle \cdot, \cdot \rangle_{\tilde{\mathcal{H}}_1} = \frac{1}{\alpha} \langle \cdot, \cdot \rangle_{\mathcal{H}_1}$. Because for $f = \sqrt{\alpha} f_1 \in \tilde{\mathcal{H}}_1$ and $x \in \mathcal{X}$, we have $\langle \alpha K_{1,x}, \sqrt{\alpha} f_1 \rangle_{\tilde{\mathcal{H}}_1} = \sqrt{\alpha} \langle K_{1,x}, f_1 \rangle_{\mathcal{H}_1} = \sqrt{\alpha} f_1(x) = f(x)$ and all $\alpha K_{1,x}$ are in $\tilde{\mathcal{H}}_1$.

From N. Aronszajn's "Theory of reproducing kernels", the RKHS of the sum of two kernels $\tilde{K}_1 = \alpha K_1$ and $\tilde{K}_2 = \beta K_2$ is the class \mathcal{H} of all functions $f = f_1 + f_2$ with $f_1 \in \tilde{\mathcal{H}}_1$ the RKHS of \tilde{K}_1 and $f_2 \in \tilde{\mathcal{H}}_2$ the RKHS of \tilde{K}_2 , endowed with the norm $\|f\|_{\mathcal{H}}^2 = \min_{f_1+f_2=f} \left(\|f_1\|_{\tilde{\mathcal{H}}_1}^2 + \|f_2\|_{\tilde{\mathcal{H}}_2}^2 \right)$ a minimum over all decompositions of f on $\tilde{\mathcal{H}}_1 + \tilde{\mathcal{H}}_2$.

2.2 K is symmetric because $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ is an inner product is symmetric and similarly, K is positive definite because $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ is an inner product.

A candidate RKHS for K is $\mathcal{H}_0 = \text{span} \{K_x, x \in \mathcal{X}\}$, endowed with the inner product

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i,j} \alpha_i \beta_j \langle \Psi(x_i), \Psi(x_j) \rangle_{\mathcal{F}}$$

where f and g were decomposed as $f = \sum_i \alpha_i K_{x_i}$ and $g = \sum_j \beta_j K_{x_j}$. The expression above does not depend on the decomposition of f and g .

This defines an inner product on \mathcal{H}_0 . With the same construction as in the proof of Aronszajn's theorem, we get a Hilbert space \mathcal{H} by extending \mathcal{H}_0 with the limits of Cauchy sequences.

Exercise 3. Sobolev spaces

3.1 \mathcal{H} is Hilbert:

\mathcal{H} is a vector space of functions. $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is a symmetric bilinear form verifying $\forall f, \langle f, f \rangle_{\mathcal{H}} \geq 0$.

Since f is absolutely continuous, it has a derivative almost everywhere and the following equality holds $\forall x \in [0, 1]$

$$\begin{aligned} |f(x)|^2 &= \left| f(0) + \int_0^x f'(u) du \right|^2 \\ &= \left| \int_0^x f'(u) du \right|^2 \quad (f(0) = 0) \\ &\leq x \cdot \int_0^1 f'(u)^2 du = x \cdot \langle f, f \rangle_{\mathcal{H}} \end{aligned}$$

$\langle f, f \rangle_{\mathcal{H}} \implies f = 0$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is therefore an inner product. **\mathcal{H} is a pre-Hilbert space with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ as inner product.**

Let (f_n) a Cauchy sequence of \mathcal{H} . (f'_n) is a Cauchy sequence of $L^2([0, 1])$ which is complete. Therefore it converges to a function $g \in L^2([0, 1])$.

Since for all (n, m) , $x \in [0, 1]$, $|f_n(x) - f_m(x)|^2 \leq x \cdot \|f_n - f_m\|_{\mathcal{H}}^2$, the sequence $f_n(x)$ is Cauchy for any x and converges to a real number $f(x)$. And since $f(x) = \lim_{n \rightarrow \infty} f_n(x) = \lim_{n \rightarrow \infty} \int_0^x f'_n(u) du = \int_0^x g(u) du$, f is absolutely continuous and $f' = g$ almost everywhere. Moreover, $f' \in L^2([0, 1])$ and $f(0) = \lim_{n \rightarrow \infty} f_n(0) = 0$.

Finally, $\|f_n - f\|_{\mathcal{H}} = \|f'_n - g\|_{L^2([0, 1])} \xrightarrow{n \rightarrow +\infty} 0$ and $f \in \mathcal{H}$. **\mathcal{H} is complete, therefore \mathcal{H} is a Hilbert space.**

Reproducing property:

We will now show that \mathcal{H} is the RKHS with corresponding kernel $K : (x, y) \rightarrow \min(x, y)$ on $[0, 1]^2$.

For $x \in [0, 1]$, the function $K_x = \min(x, \cdot)$ is differentiable except on the singleton $\{x\}$ which has a null measure, it is absolutely continuous. Its derivative is square integrable and $\min(x, 0) = 0$, **therefore K_x is in \mathcal{H} for all x .**

For any $x \in [0, 1]$ and $f \in \mathcal{H}$

$$\langle f, K_x \rangle_{\mathcal{H}} = \int_0^1 f'(u) K'_x(u) du = \int_0^x f'(u) du = f(x)$$

Therefore, **K is the r.k. of the RKHS \mathcal{H} .**

3.2 The above demonstration also shows that \mathcal{H} is a Hilbert space by remarking that

$$f(1) = \lim_{n \rightarrow \infty} f_n(1) = 1$$

for (f_n) Cauchy sequence of elements of \mathcal{H} .

We will now show that the r.k. corresponding to this Hilbert space is $K : (x, y) \mapsto \min(x, y) - xy$.

For $x \in [0, 1]$, the function K_x has a derivative everywhere except on $\{x\}$ and its derivative is square integrable. Moreover, $K_x(0) = 0$ and $K_x(1) = 0$, therefore **K_x is in \mathcal{H} .**

Let $x \in [0, 1]$ and $f \in \mathcal{H}$

$$\langle f, K_x \rangle_{\mathcal{H}} = \int_0^1 f'(u) K'_x(u) du = (1-x) \int_0^x f'(u) du - x \int_x^1 f'(u) du = f(x)$$

Therefore K is the r.k. of the RKHS \mathcal{H} .

3.3 (Note: we use the base Hilbert space \mathcal{H} from question 3.2)

\mathcal{H} is a pre-Hilbert space, and its completeness results from the fact that a Cauchy sequence of \mathcal{H} is also a Cauchy sequence in L^2 , it converges

By the theorem on Green kernels, we have that \mathcal{H} is a RKHS that admits as r.k. the Green function of the operator D^*D , where $D = \frac{d}{dx} + 1$. To find the r.k. of \mathcal{H} we need to solve in K

$$f(x) = \langle D^*DK_x, f \rangle_{L^2([0,1])} = \langle DK_x, Df \rangle_{L^2([0,1])} = \langle K_x, f \rangle_{\mathcal{H}}$$

Since $\forall x \in [0, 1]$

$$\begin{aligned} \int_0^1 K'_x(u) f'(u) + K_x(u) f(u) du &= [K'_x(u) f(u)]_0^1 - \int_0^1 K''_x(u) f(u) du + \int_0^1 K_x(u) f(u) du \\ &= \int_0^1 (K_x(u) - K''_x(u)) f(u) du \end{aligned}$$

We can then write $D^*D = 1 - \frac{d^2}{dx^2}$. Let $x \in [0, 1]$, the Green function of the operator D^*D solves the equation

$$(D^*D)G(x, t) = G(x, t) - G''(x, t) = \delta(x - t)$$

therefore, the solutions have the form $t \mapsto c_1 e^t + c_2 e^{-t}$. On $[0, x)$, the boundary conditions imply $c_1 = -c_2$ and on $(x, 1]$, $c'_1 e = -c'_2 e^{-1}$. We have the following general form of the Green function

$$G(x, t) = \begin{cases} A(x) \sinh(t) \\ B(x) \sinh(1-t) \end{cases}$$

The continuity condition at $t = x$ gives $A(x) \sinh(x) = B(x) \sinh(1-x)$ and the condition on the jump in derivative (obtained by integrating the differential equation between $x - \epsilon$ and $x + \epsilon$ with $\epsilon \rightarrow 0$) gives $A(x) \cosh(x) + \cosh(1-x) B(x) = -1$

Therefore, the solution is

$$\begin{aligned} K_x(t) &= G(x, t) = \begin{cases} \frac{1}{\sinh(1)} \sinh(1-x) \sinh(t) & \text{for } 0 \leq t < x \\ \frac{1}{\sinh(1)} \sinh(x) \sinh(1-t) & \text{for } x < t \leq 1 \end{cases} \\ &= \min \left(\frac{1}{\sinh(1)} \sinh(1-x) \sinh(t), \frac{1}{\sinh(1)} \sinh(x) \sinh(1-t) \right) \\ f(x) &= \langle D^*DG, f \rangle \\ &= \langle DG, Df \rangle \\ &= \int_0^x \frac{\sinh(1-x)}{\sinh 1} (\sinh(u) f(u) + \cosh(u) f'(u)) du + \\ &\quad \int_x^1 \frac{\sinh(x)}{\sinh 1} (\sinh(1-u) f(u) - \cosh(1-u) f'(u)) du \end{aligned}$$

Since all functions K_x are in \mathcal{H} , the kernel

$$K(x, y) = \min \left(\frac{1}{\sinh(1)} \sinh(1-x) \sinh(y), \frac{1}{\sinh(1)} \sinh(x) \sinh(1-y) \right)$$

is the r.k. of \mathcal{H} .

Exercise 4. Duality

4.a The problem is a convex problem for which strong duality holds because it is strictly feasible (e.g the point $f = 0$ satisfies the inequality constraint) and thus verifies Slaters constraint qualification. Therefore, there exist a dual parameter such that the problem is equivalent to

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^n \ell_{y_i}(f(x_i)) + \lambda \|f\|_{\mathcal{H}_K}$$

According to the representer theorem, the above problem has a solution of the form $\forall x \in \mathcal{X}$

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

The optimization problem can be re-written as

$$\min_{\alpha \in \mathbb{R}^n} R(K\alpha) + \lambda \alpha^\top K\alpha$$

where

$$R : \mathbb{R}^n \rightarrow \mathbb{R}$$

$$([K\alpha]_1, \dots, [K\alpha]_n) \mapsto \frac{1}{n} \sum_{i=1}^n \ell_{y_i}([K\alpha]_i)$$

4.b From the definition

$$\begin{aligned} R^*(u) &= \sup_{x \in \mathbb{R}^n} (x^\top u - R(x)) \\ &= \sup_{x \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n (nx_i u_i - \ell_{y_i}(x_i)) \\ &= \frac{1}{n} \sum_{i=1}^n \ell_{y_i}^*(nu_i) \end{aligned}$$

4.c The Lagrangian of the problem is

$$L(u, \alpha, \mu) = R(u) + \lambda \alpha^\top K\alpha + \mu^\top (K\alpha - u)$$

The dual function is given by $g(\mu) = \inf_{u, \alpha} L(u, \alpha, \mu)$

$$g(\mu) = R^*(\mu) + \inf_{\alpha} (\lambda \alpha^\top K\alpha + \mu^\top K\alpha)$$

Since the gradient w.r.t α of $\lambda \alpha^\top K\alpha + \mu^\top K\alpha$ is the expression $2\lambda K\alpha + K\mu$, an α minimizing this quantity is

$$\alpha^* = -\frac{\mu}{2\lambda}$$

And

$$g(\mu) = R^*(\mu) - \frac{1}{4\lambda} \mu^\top K\mu$$

The dual problem is then

$$\max_{\mu \in \mathbb{R}^n} g(\mu) = R^*(\mu) - \frac{1}{4\lambda} \mu^\top K\mu$$

The problem being strongly convex, for a μ^* maximizing g , a solution to the original problem $\min_{\alpha \in \mathbb{R}^n} (R(K\alpha) + \lambda \alpha^\top K\alpha)$ can be deduced with $\alpha^* = -\frac{\mu^*}{2\lambda}$

4.d

- For $\ell_y(u) = \log(1 + e^{-yu})$, the function ℓ_y^* is written

$$\ell_y^*(u) = \sup_{x \in \mathbb{R}} (xu - \log(1 + e^{-yx}))$$

By setting the derivative of the term in the sup to 0, we get $e^{-yx} = \frac{-u}{u+y}$ and

$$\ell_y^*(u) = \frac{u}{y} \log \left(\frac{u+y}{-u} \right) - \log \left(\frac{y}{u+y} \right) + \begin{cases} \chi_{[-1,0]}(u) & \text{if } y = 1 \\ \chi_{[0,1]}(u) & \text{if } y = -1 \end{cases}$$

The dual problem is therefore written

$$\begin{aligned} \max_{\mu \in \mathbb{R}^n} \quad & g(\mu) = \frac{1}{n} \sum_{i=1}^n \ell_{y_i}^*(n\mu_i) - \frac{1}{4\lambda} \mu^\top K \mu \\ \text{s.t} \quad & -1 \leq y_i \mu_i \leq 0 \quad \forall i \end{aligned}$$

- For $\ell_y(u) = \max(0, 1 - yu)^2$, the function ℓ_y^* is written

$$\begin{aligned} \ell_y^*(u) &= \sup_{x \in \mathbb{R}} (xu - \max(0, 1 - yx)^2) \\ &= \frac{u}{y} \left(1 + \frac{u}{2y} \right) - \max \left(0, \frac{u}{2y} \right)^2 \end{aligned}$$

The resulting dual problem is

$$\max_{\mu \in \mathbb{R}^n} \quad g(\mu) = \sum_{i=1}^n \left[\frac{\mu_i}{y_i} \left(1 + \frac{n\mu_i}{2y_i} \right) - \max \left(0, \frac{\mu_i}{2y_i} \right)^2 \right] - \frac{1}{4\lambda} \mu^\top K \mu$$