"Kernel methods in machine learning" Final Homework Due March 31st, 2021, 11:59pm

Julien Mairal and Jean-Philippe Vert

Exercice 1. Positive definiteness
Are the following kernels, defined on $\mathcal{X} = \mathbb{R}_+$ positive definite?
$K_1(x, x') = 2^{x-x'}$;
$K_2(x, x') = 2^{x+x'}$;
$K_3(x, x') = 2^{xx'}$;
$K_4(x, x') = \log(1 + xx')$;
$K_5(x, x') = \max(x, x')$
$K_6(x, x') = \min(f(x)g(x'), f(x')g(x))$ where f, g are non-negative functions.
You need to provide short proofs.

Exercice 2. Kernels encoding equivalence classes.
Consider a similarity measure $K : \mathcal{X} \times \mathcal{X} \to \{0, 1\}$ with $K(x, x) = 1$ for all x in $\mathcal{X}$. Prove that K is p.d. if and only if, for all x, x', x'' in $\mathcal{X}$,

- $K(x, x') = 1 \Leftrightarrow K(x', x) = 1$, and
- $K(x, x') = K(x', x'') = 1 \Rightarrow K(x, x'') = 1$.

Exercice 3. Kernel mean embedding
Let us consider a Borel probability measure P of some random variable X on a compact set $\mathcal{X}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous, bounded, p.d. kernel and $\mathcal{H}$ be its RKHS. The kernel mean embedding of P is defined as

the function

$$\begin{aligned}\mu(P) : \mathcal{X} &\to \mathbb{R}\\ y &\mapsto \mathbb{E}_{X\sim P}[K(X,y)].\end{aligned}$$

1. Show that $\mu(P)$ is in $\mathcal{H}$ and that $\mathbb{E}_{X\sim P}[f(X)] = \langle f, \mu(P)\rangle_{\mathcal{H}}$ for any $f \in \mathcal{H}$.
   Remark: If P and Q are two Borel probability measures, then

   $$\mu(P) = \mu(Q) \ \ \textit{implies} \ \ \{\mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{X\sim Q}[f(X)] \ \ \textit{for all} \ \ f \in \mathcal{H}\}\,.$$

   When $\mathcal{H}$ is dense in the space of continuous bounded functions on $\mathcal{X}$, this relation is sufficient to show that $P = Q$. Hence, the kernel mean embedding (single point in the RKHS!) carries all information about the distribution. We call such kernels "universal". It is possible to show that the Gaussian kernel is universal.

2. Consider the empirical distribution

   $$P_{\mathcal{S}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i},$$

   where $\mathcal{S} = \{x_1, \ldots, x_n\}$ is a finite subset of $\mathcal{X}$ and δ_{x_i} is a Dirac distribution centered at x_i. Show that

   $$\mathbb{E}_{\mathcal{S}}[\|\mu(P) - \mu(P_{\mathcal{S}})\|_{\mathcal{H}}] \leq \frac{4\sqrt{\mathbb{E}K(X,X)}}{\sqrt{n}},$$

   where $\mathbb{E}_{\mathcal{S}}$ is the expectation by randomizing over the training set (each x_i is a r.v. distributed according to P). Remember that you are allowed to (and you should!) use any existing result from the slides.

3. Consider the quantity

   $$MMD(\mathcal{S}_1, \mathcal{S}_2) = \|\mu(P_{\mathcal{S}_1}) - \mu(P_{\mathcal{S}_2})\|^2_{\mathcal{H}}$$

   for two sets $\mathcal{S}_1 = (x_1, \ldots, x_n)$ and $\mathcal{S}_2 = (y_1, \ldots, y_m)$. Show that

   $$MMD(\mathcal{S}_1, \mathcal{S}_2) = \left(\sup_{\|f\|_{\mathcal{H}}\leq 1}\left\{\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \frac{1}{m}\sum_{j=1}^{m} f(y_j)\right\}\right)^2,$$

and give a formula for this quantity in terms of kernel evaluations only. *Remark: this is called the maximum mean discrepancy criterion, which can be used for statistical testing (are $\mathcal{S}_1$ and $\mathcal{S}_2$ coming from the same distribution?).*

4. We consider $\mathcal{X} = \mathbb{R}^d$ and the *normalized* Gaussian kernel with bandwidth σ: $K(x, y) = \sigma^{-d} \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$. For any two sets $\mathcal{S}_1$ and $\mathcal{S}_2$, show that $MMD(\mathcal{S}_1, \mathcal{S}_2)$ is a decreasing function of σ.

Exercice 4. Properties of the dot-product kernel

Consider the dot-product kernel on the sphere $K_1 : \mathbb{S}^{p-1} \times \mathbb{S}^{p-1} \rightarrow \mathbb{R}$ such that for all pair of points x, x' in $\mathbb{S}^{p-1}$ (unit sphere of $\mathbb{R}^p$),

$$K_1(x, x') = \kappa(\langle x, x' \rangle),$$

where $\kappa : [-1, 1] \rightarrow \mathbb{R}$ is an infinitely differentiable function that admits a polynomial expansion on $[-1, 1]$:

$$\kappa(u) = \sum_{i=0}^{+\infty} a_i u^i, \tag{1}$$

where the a_i's are real coefficients and the sum above is always converging.

1. Show that if all coefficients a_i are non-negative and $\kappa \neq 0$, then K_1 is p.d.

2. If K_1 is p.d., show that the homogeneous dot-product kernel $K_2 : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ is also p.d..

$$K_2(x, x') = \begin{cases} \|x\| \|x'\| \kappa\left(\frac{\langle x, x' \rangle}{\|x\| \|x'\|}\right) & \text{if } \|x\| \neq 0 \text{ and } \|x'\| \neq 0 \\ 0 & \text{otherwise} \end{cases}.$$

   Remark: it is in fact possible to show that all coefficients a_i need to be non-negative for the positive definiteness to hold for all dimension p, but we do not ask for a proof of this result, which is due to Shoenberg, 1942.

3. Assume that all coefficients a_i are non-negative (K_1 is thus p.d.) and that $\kappa(1) = \kappa'(1) = 1$. Let $\mathcal{H}$ be the RKHS of K_1 and consider its RKHS mapping $\varphi : \mathbb{S}^{p-1} \to \mathcal{H}$ such that $K_1(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$ for all x, x' in $\mathbb{S}^{p-1}$. Show that:

$$\forall x, x' \in \mathbb{S}^{p-1}\,, \quad \|\varphi(x) - \varphi(x')\|_{\mathcal{H}} \leq \|x - x'\|.$$

4. Find an explicit feature map $\psi : \mathbb{S}^{p-1} \to \ell^2$, where ℓ^2 is the Hilbert space of real-valued sequences (see definition on slide 240), such that for all x, y in $\mathbb{S}^{p-1}$

$$K_1(x, y) = \langle \psi(x), \psi(y) \rangle_{\ell_2}.$$

Hint: remember that $\langle x, y \rangle^2 = \langle xx^\top, yy^\top \rangle_F$, where $\langle ., \rangle_F$ is the Frobenius inner-product. You may want to use the tensor product notation $x^{\otimes 2} = xx^\top$ and its generalization for degrees higher than 2.

5. Let us assume that you have found an explicit feature map ψ in the previous question. Remember from one of our previous homeworks that the RKHS $\mathcal{H}$ of K_1 can be characterized by

$$\mathcal{H} = \{f_w : w \in \ell_2\} \quad \text{such that} \quad f_w : x \mapsto \langle w, \psi(x) \rangle_{\ell_2},$$

with

$$\|f_w\|^2_{\mathcal{H}} = \inf_{w' \in \ell_2} \left\{ \|w'\|^2_{\ell_2} : f_w = f_{w'} \right\}.$$

Consider then a function $g_z : \mathbb{S}^{p-1} \to \mathbb{R}$ of the form

$$g_z : x \mapsto \sigma(\langle z, x \rangle)$$

with z in $\mathbb{S}^{p-1}$ and σ admits a polynomial expansion $\sigma(u) = \sum_{i=0}^{+\infty} b_i u^i$. Could you find a sufficient condition on z and on the coefficients b_i for g_z to be in $\mathcal{H}$?

Remark: g_z can be interpreted as a one-layer neural network function. We could ask you to do the same analysis for the homogeneous kernel K_2, but this would be unnecessary technical for this homework which is already too long. This being said, if you found it too short, we're happy to see your analysis of K_2 and the type of functions g_z you will consider.