Homework for the Course "Advanced Learninig Models"

Xhenis COBA
master MSIAM DS
xhenis.coba@etu.univ-grenoble-alpes.fr

Predrag PILIPOVIC
master MSIAM DS
predrag.pilipovic@grenoble-inp.org

1 Some kernels...

Show that the following kernels are positive definite:

1. **On $\mathcal{X} = \mathbb{R}$:**

$$\forall x, y \in \mathbb{R}, \;\; K(x, y) = \cos(x - y).$$

We will use Aronszajn's Theorem. This means we need to find map Φ and Hilbert space $\mathcal{H}$ such that $\Phi : \mathcal{X} \to \mathcal{H}$ is given by

$$\langle \Phi(x), \Phi(y) \rangle = K(x, y).$$

In our case if we choose $\Phi : \mathbb{R} \to [-1, 1]^2 \subset \mathbb{R}^2$ given by

$$\Phi(x) = \begin{bmatrix} \cos x \\ \sin x \end{bmatrix},$$

then we will have

$$\langle \Phi(x), \Phi(y) \rangle = \cos x \cos y + \sin x \sin y = \cos(x - y) = K(x, y).$$

And $[-1, 1]^2$ is Hilbert space, thus K is p.d.

2. **On $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^p \mid \|\mathbf{x}\|_2 < 1\}$:**

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, \;\; K(\mathbf{x}, \mathbf{y}) = \frac{1}{1 - \mathbf{x}^T \mathbf{y}}.$$

For this part we will use the fact from the lecture that scalar product is p.d. We will also need the fact that sum of two p.d. is again p.d. which is trivial consequence of definition. Also, we have seen in the course (and the previous homework) that product of two p.d. is also p.d. and limit of p.d. is again p.d. With this knowledge we can easily see that $K(\mathbf{x}, \mathbf{y})$ is p.d. because it can be written as

$$K(\mathbf{x}, \mathbf{y}) = \frac{1}{1 - \mathbf{x}^T \mathbf{y}} = \frac{1}{1 - \langle \mathbf{x}, \mathbf{y} \rangle} = \sum_{n=0}^{\infty} \langle \mathbf{x}, \mathbf{y} \rangle^n = \lim_{n \to \infty} \sum_{k=0}^{n} \langle \mathbf{x}, \mathbf{y} \rangle^k.$$

On the right hand side we have limits of sum of products of p.d. so, it is p.d. and thus K is p.d. Note here that we used geometric series, which is justified from the assumption of $\mathcal{X}$ and Cauchy-Schwarz inequality. Namely, we have

$$\langle \mathbf{x}, \mathbf{y} \rangle \leq \|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2 < 1,$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

3. **On $\mathcal{X} = \mathbb{N}$:**

$$\forall m, n \in \mathbb{N}, \;\; K(m,n) = (-1)^{m+n}.$$

Here, we can again use the Aronszajn's Theorem. We can define $\Phi : \mathbb{N} \to \{-1, 1\} \subset \mathbb{R}$ as

$$\Phi(n) = (-1)^n.$$

Then, obviously, we have

$$\langle \Phi(m), \Phi(n) \rangle = (-1)^m (-1)^n = (-1)^{m+n} = K(m,n).$$

4. **On $\mathcal{X} = \mathbb{R}^n$:**

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \;\; K(\mathbf{x}, \mathbf{y}) = \pi - \arccos\left(\frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}\right).$$

This time, we will use the $\arccos$ series representation, i.e.

$$\arccos x = \frac{\pi}{2} - \sum_{n=0}^{\infty} \underbrace{\frac{(2n)!}{4^n (n!)^2 (2n+1)}}_{C_n} x^{2n+1},$$

which converge for $|x| \leq 1$. In our case we have $\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} \leq 1$ because of the Cauchy-Schwarz inequality, so we can use the series representation. So, we have

$$K(\mathbf{x}, \mathbf{y}) = \frac{\pi}{2} + \sum_{n=0}^{\infty} C_n \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle^{2n+1} = \frac{\pi}{2} + \lim_{n \to \infty} \sum_{k=0}^{n} C_k \left\langle \frac{\mathbf{x}}{\|\mathbf{x}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right\rangle^{2n+1}.$$

Using the same argument as in the question 2, we have product of scalar products which is p.d. and thus, under the sum we have p.d., and the limit of the sum of p.d. which is again p.d. Finally, we have constant plus p.d. which is trivially p.d.

Or in another way: We try to prove that K satisfies the requirements of the definition of positive definite kernels. We can see that K is symmetric because

$$K(\mathbf{x}, \mathbf{y}) = \pi - \arccos\left(\frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}\right) = \pi - \arccos\left(\frac{\mathbf{y}^T \mathbf{x}}{\|\mathbf{y}\| \cdot \|\mathbf{x}\|}\right) = K(\mathbf{y}, \mathbf{x}).$$

For the positive definiteness, we have

$$\begin{aligned}
\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \left(\pi - \arccos\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|}\right)\right) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \pi - \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \arccos\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|}\right).
\end{aligned}$$

We know that

$$0 \leq \arccos\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|}\right) \leq \pi,$$

so

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \arccos\left(\frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|}\right) \leq \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j \pi$$

This means that $\sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$, and thus, K is a positive definite kernel.

2 Dual of the SVM with intercept

We recall the primal formulation of SVMs seen in the class

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2,$$

and its dual formulation

$$\max_{\alpha\in\mathbb{R}^n} 2\alpha^T \mathbf{y} - \alpha^T \mathbf{K}\alpha \text{ **such that** } 0 \leq y_i\alpha_i \leq \frac{1}{2\lambda n}, \text{ **for all** } i. \tag{1}$$

Consider the primal formulation of SVMs with intercept

$$\min_{\substack{f\in\mathcal{H}\\ b\in\mathbb{R}}} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i(f(x_i) + b)) + \lambda \|f\|_{\mathcal{H}}^2.$$

Can we still apply the representer theorem? Why? Derive the corresponding dual formulation by using Lagrangian duality. Provide a formulation in terms of $\alpha \in \mathbb{R}^n$ as in (1).

We can not use the representer theorem seen in the lecture on this problem because that theorem is non-parametric, which means that we are solving optimization problem over set of functions from $\mathcal{H}$. In this case, we have semi-parametric problem where we also include a parametric part, which means that we are solving optimization problem over set of functions $\mathcal{H} + \mathbb{R} = \{f + b \mid f \in \mathcal{H}, b \in \mathbb{R}\}$. We will provide generalized theorem [1], but we will not use it in this exercises.

Theorem 1 (**Semi-parametric representer theorem**). *Suppose that in addition to the assumptions of the representer theorem we are given a set of N real-valued functions $\{\psi_p\}_{p=1}^N$ on $\mathcal{X}$ with the property that the $n \times N$ matrix $[\psi_p(x_i)]_{ip}$ has rank N. Then any $\widetilde{f} = f + h$, with $f \in \mathcal{H}$ and $h \in \text{span}\{\psi_p\}$, minimizing*

$$\Psi(\widetilde{f}(x_1), \widetilde{f}(x_2), ..., \widetilde{f}(x_n), \|f\|_{\mathcal{H}}),$$

admits a representation of the form:

$$\widetilde{f}(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + \sum_{p=1}^{N} \beta_p \psi_p(x),$$

with $\beta_p \in \mathbb{R}$, for all $p = 1, 2, ..., M$.

Now, it can be easily seen that for a single $N = 1$ constant function $\psi_1(x) = b$, $b \in \mathbb{R}$, is used as an intercept. Thus, semi-parametric representer theorem would give us the representation for our problem. Here, we will obtain the result using Lagrangian duality. Let us rewrite the problem with the slack variables, like in the lecture

$$\min_{\substack{f\in\mathcal{H}\\ b\in\mathbb{R}\\ \xi\in\mathbb{R}^n}} \frac{1}{n}\sum_{i=1}^{n} \xi_i + \lambda \|f\|_{\mathcal{H}}^2,$$

such that $\xi_i \geq \max(0, 1 - y_i(f(x_i) + b))$ for all $i = 1, 2, ..., n$. This constraint can be written as $y_i(f(x_i) + b) \geq 1 - \xi_i$, and $\xi_i \geq 0$, for all $i = 1, 2, ..., n$. To cope with the primal problem, we need to formulate the Lagrangian and the solve the dual problem. Starting with the Lagrangian, we multiply each constraint with a variable (called Lagrange multiplier) and add them to the optimization function. The Lagrangian thus becomes

$$L(f, b, \xi, \alpha, \beta) = \lambda\|f\|^2 + \frac{1}{n}\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \Big[y_i\big(f(x_i) + b\big) - (1 - \xi_i) \Big] - \sum_{i=1}^{n} \beta_i \xi_i,$$

such that $\alpha_i, \beta_i \geq 0$, for all $i = 1, 2, ..., n$. Now, in the dual problem, we try to maximize the minimum of the Lagrangian function, i.e.

$$\max_{\alpha,\beta} \min_{f,b,\xi} L(f, b, \xi, \alpha, \beta).$$

To construct the dual problem, we need to determine the optimal f, b, and ξ in terms of the dual variables α and β. We achieve this by differentiating the constraints with respect to the primal variables, i.e. by taking the following gradients $\nabla_f L$, $\frac{\partial L}{\partial b}$ and $\nabla_\xi L$, and then setting them equal to 0. Using the fact that $f \in \mathcal{H}$, and $\mathcal{H}$ is RKHS, it means that $f(x) = \langle f, \Phi(x) \rangle_{\mathcal{H}}$, where $\Phi(x) = K_x$. Now, we can find $\nabla_f L$, and it is

$$\nabla_f L = 2\lambda f - \sum_{i=1}^{n} \alpha_i y_i K_{x_i} = 0,$$

from where we find $f = \frac{1}{2\lambda} \sum\limits_{i=1}^{n} \alpha_i y_i K_{x_i}$. Now, we can eliminate f from L by noticing that

$$f(x_j) = \frac{1}{2\lambda} \sum_{i=1}^{n} \alpha_i y_i K(x_i, x_j),$$

and

$$\|f\|_{\mathcal{H}}^2 = \frac{1}{4\lambda^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j).$$

Now, we have

$$L(b, \xi, \alpha, \beta) = -\frac{1}{4\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) + \frac{1}{n} \sum_{i=1}^{n} \xi_i - b \sum_{i=1}^{n} \alpha_i y_i - \sum_{i=1}^{n} \xi_i (\alpha_i + \beta_i) + \sum_{i=1}^{n} \alpha_i,$$

or in matrix form

$$L(b, \xi, \alpha, \beta) = -\frac{1}{4\lambda} \alpha^T \mathbf{G} \alpha + \frac{1}{n} \xi^T \mathbf{1} - b \alpha^T \mathbf{y} - \xi^T (\alpha + \beta) + \alpha^T \mathbf{1},$$

where $[\mathbf{G}]_{ij} = y_i y_j K(x_i, x_j)$, i.e. $\mathbf{G} = \mathbf{y}\mathbf{y}^T \mathbf{K}$. Taking derivative with respect to b we have

$$\frac{\partial L}{\partial b} = \alpha^T \mathbf{y} = 0.$$

Taking the derivative with respect to ξ_i we have

$$\nabla_\xi L = \frac{\mathbf{1}}{n} - (\alpha + \beta) = 0,$$

i.e. $\dfrac{\mathbf{1}}{n} = \alpha + \beta$. And finally, we can provide the Lagrangian dual problem by

$$\max_{\alpha, \beta} \min_{f, b, \xi} L(f, b, \xi, \alpha, \beta) = \max_{\alpha, \beta} \alpha^T \mathbf{1} - \frac{1}{4\lambda} \alpha^T \mathbf{G} \alpha,$$

such that $\alpha + \beta = \dfrac{\mathbf{1}}{n}$, $\alpha^T \mathbf{y} = 0$, $\alpha_i, \beta_i \geq 0$, for all $i = 1, 2, ..., n$. If $0 \leq \alpha_i \leq \frac{1}{n}$ for all $i = 1, 2, ..., n$, then the dual function takes finite values that depend only on α by taking $\beta_i = \frac{1}{n} - \alpha_i$. The dual problem is therefore equivalent to

$$\max_{\alpha \in \mathbb{R}^n} \alpha^T \mathbf{1} - \frac{1}{4\lambda} \alpha^T \mathbf{G} \alpha,$$

such that $\alpha^T \mathbf{y} = 0$, and $0 \leq \alpha_i \leq \frac{1}{n}$, for all $i = 1, 2, ..., n$, and where $\mathbf{G} = \mathbf{y}\mathbf{y}^T \mathbf{K}$.

3 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 $x, x', x'' \in \mathcal{X}$

(1) $K(x, x') = 1 \Leftrightarrow K(x', x) = 1$,

(2) $K(x, x') = K(x', x'') = 1 \Rightarrow K(x, x'') = 1$.

($\Rightarrow$) Let us assume that K is p.d. We want to prove that (1) and (2) hold. From symmetry of K we can conclude (1) immediately. we will prove (2) by assuming the opposite, i.e. there exist $x_1, x_2, x_3 \in \mathcal{X}$ such that

$$K(x_1, x_2) = K(x_2, x_3) = 1 \text{ and } K(x_1, x_3) = 0.$$

From the positive definiteness of K we know that for all $(a_1, a_2, a_3) \in \mathbb{R}^3$ it must be

$$\begin{aligned} 0 &\leq \sum_{i=1}^{3}\sum_{j=1}^{3} a_i a_j K(x_i, x_j) \\ &= a_1^2 K(x_1, x_1) + a_2^2 K(x_2, x_2) + a_3^2 K(x_3, x_3) \\ &+ 2a_1 a_2 K(x_1, x_2) + 2a_1 a_3 K(x_1, x_3) + 2a_2 a_3 K(x_2, x_3) \\ &= a_1^2 + a_2^2 + a_3^2 + 2a_1 a_2 + 2a_2 a_3 \\ &= (a_1 + a_2)^2 + (a_2 + a_3)^2 - a_2^2. \end{aligned}$$

If we choose $a_2 = 1$, $a_1 = a_3 = -1$, we have

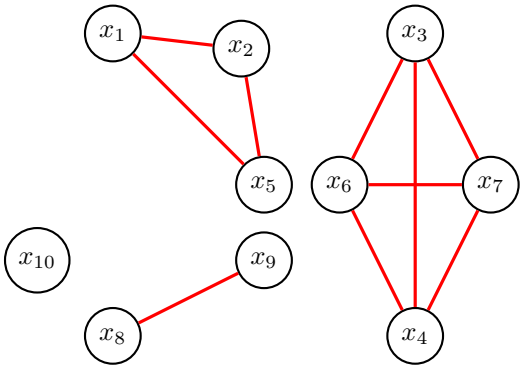
$$0 \leq 0^2 + 0^2 - 1^2 = -1,$$

which is contradiction.

($\Leftarrow$) Let us assume that (1) and (2) are true. We want to prove that K is p.d. We can see that K is symmetric from (1). Indeed, if $K(x, y) = 1$, then $K(y, x) = 1$. On the other hand, using negation of (1) we see that $K(x, y) \neq 1 \Leftrightarrow K(y, x) \neq 1$, which means that $K(x, y) = K(y, x) = 0$. Now, we need to prove positive definiteness, i.e. for all $n \in \mathbb{N}$, all $(a_1, a_2, ..., a_n)$ and all $(x_1, x_2, ..., x_n)$

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

From the assumption $K(x, x) = 1$, and (1) and (2), we can obtain a relation of equivalence. Let us visualize our data as underacted graph as shown below.



Each pair of points $x, y \in \mathcal{X}$, for which we have $K(x, y) = 1$, will be connected in the graph, i.e. we have the graph (V, E), where $V = \mathcal{X}$ and $E = \{(x, y) \in \mathcal{X} \times \mathcal{X} \mid K(x, y) = 1\}$. Then, we have equivalence relation on the graph (V, E) given by K. Thus, we have M classes of equivalence over our graph (V, E). This means that we will cover all relations inside each class, and mutual relations within two different classes will be zero. If in class m we have N_m data points, then we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) = \sum_{m=1}^{M}\sum_{i=1}^{N_m}\sum_{j=1}^{N_m} a_i a_j K(x_i, x_j) = \sum_{m=1}^{M}\left(\sum_{i=1}^{N_m} a_i\right)^2 \geq 0.$$

Little remark. Last equation is not true in general, it is true when n is equal to the size of data set $\mathcal{X}$. For example, we can have $n = 4$ in our above example, and we can see that we are going to cover portion of two different classes, i.e. sub-graphs. Still, if we put N_m to be a number of points from $\{x_1, x_2, ..., x_n\}$ in class m, we will obtain the same result.

4 COCO

Given two sets of real numbers $X = (x_1, x_2, ..., x_n) \in \mathbb{R}^n$ and $Y = (y_1, y_2, ..., y_n) \in \mathbb{R}^n$, the covariance between X and Y is defined as

$$\text{cov}_n(X, Y) = \mathbb{E}_n[XY] - \mathbb{E}_n[X]\mathbb{E}_n[Y],$$

where

$$\mathbb{E}_n(U) = \frac{1}{n}\sum_{i=1}^{n} u_i.$$

The covariance is useful to detect linear relationships between X and Y. In order to extend this measure to potential nonlinear relationships between X and Y, we consider the following criterion:

$$C_n^K(X, Y) = \max_{f,g \in \mathcal{B}_K} \text{cov}_n(f(X), g(Y)),$$

where K is a positive definite kernel on $\mathbb{R}$, $\mathcal{B}_K$ is the unit ball of the RKHS of K, and $f(U) = (f(u_1), f(u_2), ..., f(u_n))$ for a vector $U = (u_1, u_2, ..., u_n)$.

1. **Express simply $C_n^K(X, Y)$ for the linear kernel $K(a, b) = ab$.**
   Since we are dealing with regular product on $\mathbb{R}$, the RKHS for this linear kernel is $\mathbb{R}$, and thus $\mathcal{B}_K = \{x \in \mathbb{R} \mid |x| \leq 1\}$. This means that for $f \in \mathcal{B}_K$ we have $f(x) = f \cdot x$, for $-1 \leq f \leq 1$. So, we have

$$\begin{aligned}
\text{cov}_n(f(X), g(Y)) &= \mathbb{E}_n[f(X)g(Y)] - \mathbb{E}_n[f(X)]\mathbb{E}_n[g(Y)] \\
&= \frac{1}{n}\sum_{i=1}^{n} f(x_i)g(y_i) - \left(\frac{1}{n}\sum_{i=1}^{n} f(x_i)\right)\left(\frac{1}{n}\sum_{i=1}^{n} g(y_i)\right) \\
&= \frac{fg}{n}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j\right) \\
&= \frac{fg}{n}\left(X^T Y - X^T U Y\right) \\
&= \frac{fg}{n} X^T (I - U) Y,
\end{aligned}$$

   where

$$U = \frac{1}{n}\begin{bmatrix} 1 & 1 & ... & 1 \\ 1 & 1 & ... & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & ... & 1 \end{bmatrix} = \frac{\mathbf{1} \cdot \mathbf{1}^T}{n}.$$

   Now, we want to maximize the previous expression over $f, g \in \mathcal{B}_K$, i.e. over $|f| \leq 1$, and $|g| \leq 1$. We will choose $f, g \in \{-1, +1\}$ depending on the sign of the rest of the expression, which means we will have

$$C_n^K(X, Y) = \max_{f,g \in \mathcal{B}_K} \text{cov}_n(f(X), g(Y)) = \max_{\substack{|f| \leq 1 \\ |g| \leq 1}} \frac{fg}{n} X^T (I - U) Y = \frac{1}{n}\left|X^T (I - U) Y\right|.$$

2. **For a general kernel K, express $C_n^K(X, Y)$ in terms of the Gram matrices of X and Y.**
   We would like to use the representer theorem, which means that we need to rewrite a maximization problem. Firstly, we have seen that

$$C_n^K(X, Y) = \max_{f,g \in \mathcal{B}_K} \frac{1}{n}\sum_{i=1}^{n} f(x_i)g(y_i) - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f(x_i)g(y_j).$$

If we have a solution (f^*, g^*) of the previous maximization problem, then f^* is a solution of the maximization problem

$$\max_{f \in \mathcal{B}_K} \frac{1}{n} \sum_{i=1}^{n} f(x_i) g^*(y_i) - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(x_i) g^*(y_j).$$

Condition $f \in \mathcal{B}_K$ can be seen as maximizing over $f \in \mathcal{H}$, with respect to $\|f\|_{\mathcal{H}} \leq 1$, where $\mathcal{H}$ is corresponding RKHS for K. We can now, rewrite the previous problem as its dual

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} f(x_i) \left(\frac{1}{n} \sum_{j=1}^{n} g^*(y_j) - g^*(y_i) \right) + \lambda(\|f\|_{\mathcal{H}} - 1),$$

such that $\lambda \geq 0$. Now, we can use representer theorem and get that $f^* = \sum_{i=1}^{n} \alpha_i K_{x_i}$, for some $\alpha \in \mathbb{R}^n$. Using an f^* with this form, we apply the same reasoning to obtain that $g^* = \sum_{i=1}^{n} \beta_i K_{y_i}$, for some $\beta \in \mathbb{R}^n$. Now, we know that

$$f(x_j) = \sum_{i=1}^{n} \alpha_i K(x_i, x_j) = [K_X \alpha]_j.$$

Also,

$$\|f\|_{\mathcal{H}}^2 = \left\| \sum_{i=1}^{n} \alpha_i K_{x_i} \right\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) = \alpha^T K_X \alpha.$$

Similarly, we can derive for $g(y_j)$ and $\|g\|_{\mathcal{H}}^2$. Now, we can rewrite the covarience expression as

$$\begin{aligned}
\text{cov}_n(X, Y) &= \frac{1}{n} \sum_{i=1}^{n} f(x_i) g(y_i) - \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(x_i) g(y_j) \\
&= \frac{1}{n} \left(\sum_{i=1}^{n} [K_X \alpha]_i [K_Y \beta]_i - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} [K_X \alpha]_i [K_Y \beta]_j \right) \\
&= \frac{1}{n} \left((K_X \alpha)^T K_Y \beta - (K_X \alpha)^T U K_Y \beta \right) \\
&= \frac{1}{n} \alpha^T K_X (I - U) K_Y \beta.
\end{aligned}$$

This, means we have

$$C_n^K(X, Y) = \max_{\substack{\alpha^T K_X \alpha \leq 1 \\ \beta^T K_Y \beta \leq 1}} \frac{1}{n} \alpha^T K_X (I - U) K_Y \beta.$$

Having in mind that K_X and K_Y are symmetric positive semi-definite matrices, we know that they are invertible, and that they have invertible positive semi-definite square root. This means that we have $K_X = K_X^{1/2} K_X^{1/2}$, and thus

$$\alpha^T K_X \alpha = (K_X^{1/2} \alpha)^T K_X^{1/2} \alpha = \|K_X^{1/2} \alpha\|_{\mathcal{H}}.$$

Now, we can rewrite the previous optimization problem as

$$C_n^K(X, Y) = \max_{\substack{\|A\|_{\mathcal{H}} \leq 1 \\ \|B\|_{\mathcal{H}} \leq 1}} \frac{1}{n} A^T K_X^{1/2} (I - U) K_Y^{1/2} B,$$

where $A = K_X^{1/2} \alpha$, and $B = K_Y^{1/2} \beta$. Finally, using the property for the norm of the linear operator

$$\|A\| = \sup_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}\| \leq 1} \|A\mathbf{x}\| = \sup_{\|\mathbf{x}\| = 1} \|A\mathbf{x}\|,$$

we can calculate explicitly the previous expression, and the solution is given by

$$C_n^K(X, Y) = \frac{1}{n} \|K_X^{1/2} (I - U) K_Y^{1/2}\|_{\mathcal{H}}.$$

5 RKHS

1. **Let K_1 and K_2 be two positive definite kernels on a set $\mathcal{X}$, and α, β two positive scalars. Show that $\alpha K_1 + \beta K_2$ is positive definite, and describe its RKHS.**

   Let $\alpha, \beta \geq 0$ be two real numbers, and define $K = \alpha K_1 + \beta K_2$. The symmetry of K is immediate consequence from the symmetry of K_1 and $_2$. Moreover for $n \in \mathbb{N}$, real weights $(a_1, a_2, ..., a_n)$ and an observation $(x_1, x_2, ..., x_n) \in \mathcal{X}^n$, we have

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (\alpha K_1 + \beta K_2)(x_i, x_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha a_i a_j K_1(x_i, x_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} \beta a_i a_j K_2(x_i, x_j) \\ &= \alpha \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_1(x_i, x_j)}_{\geq 0} + \beta \underbrace{\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_2(x_i, x_j)}_{\geq 0} \geq 0,\end{aligned}$$

   and thus, K is positive definite kernel. Now, let us prove that $\mathcal{H}_K = \alpha\mathcal{H}_{K_1} + \beta\mathcal{H}_{K_2}$, i.e.

$$\mathcal{H}_K = \{\alpha f_1 + \beta f_2 \mid f_i \in \mathcal{H}_{K_i}, i = 1, 2\}.$$

   It is sufficient to prove for the base case $\alpha = \beta = 1$, because of the linear property of vector spaces. Let us consider the orthogonal direct sum of the two Hilbert spaces

$$\mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2} = \{(f_1, f_2) \mid f_i \in \mathcal{H}_{K_1}, i = 1, 2\},$$

   with the inner product

$$\langle (f_1, f_2), (g_1, g_2) \rangle = \langle f_1, g_1 \rangle_{\mathcal{H}_{K_1}} + \langle f_2, g_2 \rangle_{\mathcal{H}_{K_2}}.$$

   This inner product will give us norm

$$\|(f_1, f_2)\|^2_{\mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2}} = \|f_1\|^2_{\mathcal{H}_{K_1}} + \|f_2\|^2_{\mathcal{H}_{K_2}}.$$

   Since $\mathcal{H}_{K_i}$, $i = 1, 2$ are both subspaces of the vector space of all functions on $\mathcal{X}$, the intersection $\mathcal{F}_0 = \mathcal{H}_{K_1} \cap \mathcal{H}_{K_2}$ is a well-defined vector space of functions on $\mathcal{X}$. Let us now introduce set

$$\mathcal{N} = \{(f, -f) \mid f \in \mathcal{F}_0\} \subset \mathcal{H}_{K_1} + \mathcal{H}_{K_2}.$$

   The first thing to be note is that $\mathcal{N}$ is a closed subspace. Let us prove this. If we assume that

$$\|(f_n, -f_n) - f(f, g)\|_{\mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2}} \to 0,$$

   then, by using the the norm on $\mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2}$, we get

$$\|f_n - f\|_{\mathcal{H}_{K_1}} \to 0 \text{ and } \| - f_n - g\|_{\mathcal{H}_{K_2}} \to 0,$$

   and hence, at each point we have $f(x) = -g(x)$. Now we can use the theorem of orthogonal projection and rewrite $\mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2} = \mathcal{N} + \mathcal{N}^\perp$, which yields for all $(f_1, f_2) \in \mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2}$

$$(f_1, f_2) = (f, -f) + (h_1, h_2),$$

   where $f \in \mathcal{F}_0$ and $(h_1, h_2) \perp \mathcal{N}$. Let us now introduce $\mathcal{H}$ as a vector space of functions of the form

$$\{f_1 + f_2 \mid f_i \in \mathcal{H}_{K_i}, i = 1, 2\}.$$

   We can define $\Gamma : \mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2} \to \mathcal{H}$ by

$$\Gamma(f_1, f_2) = f_1 + f_2.$$

   Obviously, the map Γ is a linear surjection, and it has kernel $\ker \Gamma = \mathcal{N}$. This means that $\Gamma : \mathcal{N}^\perp \to \mathcal{H}$ is a vector space isomorphism. If we endow $\mathcal{H}$ with the norm that comes

from this isomorphism, then $\mathcal{H}$ will be a Hilbert space. For this purpose, we need the orthogonal projection $P : \mathcal{H}_{K_1} \oplus \mathcal{H}_{K_2} \to \mathcal{N}^\perp$. For every $f = g_1 + g_2 \in \mathcal{H}$, we have

$$\begin{aligned}\|f\|^2_{\mathcal{H}} &= \|P(g_1, g_2)\|^2_{\mathcal{H}_{K_1}\oplus\mathcal{H}_{K_2}} = \min_{g\in\mathcal{F}_0} \|(g_1 + g, g_2 - g)\|^2_{\mathcal{H}_{K_1}\oplus\mathcal{H}_{K_2}} \\ &= \min_{\substack{f=f_1+f_2 \\ f_1\in\mathcal{H}_{K_1} \\ f_2\in\mathcal{H}_{K_2}}} \|(f_1, f_2)\|^2_{\mathcal{H}_{K_1}\oplus\mathcal{H}_{K_2}} \\ &= \min_{\substack{f=f_1+f_2 \\ f_1\in\mathcal{H}_{K_1} \\ f_2\in\mathcal{H}_{K_2}}} \left(\|f_1\|^2_{\mathcal{H}_{K_1}} + \|f_2\|^2_{\mathcal{H}_{K_2}}\right).\end{aligned}$$

Now, we have a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, and it remains to see that $\mathcal{H}$ is a RKHS of functions on $\mathcal{X}$ with reproducing kernel K. By computing the norm, we have seen that for any two functions $f = f_1 + f_2$, and $g = g_1 + g_2$ in $\mathcal{H}$ we will have that

$$\langle f, g\rangle_{\mathcal{H}} = \langle P(f_1, f_2), P(g_1, g_2)\rangle_{\mathcal{H}_{K_1}\oplus\mathcal{H}_{K_2}}.$$

Let $K_y^{(i)}(x) = K_i(x, y)$, so that $K_y^{(i)} \in \mathcal{H}_{K_i}$, $i = 1, 2$, is the kernel function. We can see that if $(f, -f) \in \mathcal{N}$, then

$$\left\langle (f, -f), (K_y^{(1)}, K_y^{(i)})\right\rangle_{\mathcal{H}_{K_1}\oplus\mathcal{H}_{K_2}} = \left\langle f, K_y^{(1)}\right\rangle_{\mathcal{H}_{K_1}} + \left\langle -f, K_y^{(2)}\right\rangle_{\mathcal{H}_{K_1}} = f(y) - f(y) = 0.$$

This means that $\left(K_y^{(1)}, K_y^{(2)}\right) \in \mathcal{N}^\perp$, for every $y \in \mathcal{X}$. Thus, for any $f = f_1 + f_2 \in \mathcal{H}$, we have

$$\begin{aligned}\left\langle f, K_y^{(1)} + K_y^{(2)}\right\rangle_{\mathcal{H}} &= \left\langle P(f_1, f_2), P\left(K_y^{(1)}, K_y^{(2)}\right)\right\rangle_{\mathcal{H}_{K_1}\oplus\mathcal{H}_{K_2}} \\ &= \left\langle P(f_1, f_2), \left(K_y^{(1)}, K_y^{(2)}\right)\right\rangle_{\mathcal{H}_{K_1}\oplus\mathcal{H}_{K_2}} \\ &= \left\langle f_1, K_y^{(1)}\right\rangle_{\mathcal{H}_{K_1}} + \left\langle f_2, K_y^{(2)}\right\rangle_{\mathcal{H}_{K_2}} \\ &= f_1(y) + f_2(y) = f(y).\end{aligned}$$

Thus, $\mathcal{H}$ is RKHS with reproducing kernel $K = K_1 + K_2$.

2. **Let $\mathcal{X}$ be a set and $\mathcal{F}$ be a Hilbert space. Let $\Psi : \mathcal{X} \to \mathcal{F}$, and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be:**

$$\forall x, x' \in \mathcal{X}, \;\; K(x, x') = \langle\Psi(x), \Psi(x')\rangle_{\mathcal{F}}.$$

Show that K is a positive definite kernel on $\mathcal{X}$, and describe its RKHS.
For $n \in \mathbb{N}$, real weights $(a_1, a_2, ..., a_n)$ and data $(x_1, x_2, ..., x_n) \in \mathcal{X}^n$ we have

$$\begin{aligned}\sum_{i=1}^n\sum_{j=1}^n a_i a_j K(x_i, x_j) &= \sum_{i=1}^n\sum_{j=1}^n a_i a_j \langle\Psi(x_i), \Psi(x_j)\rangle_{\mathcal{F}} \\ &= \left\langle \sum_{i=1}^n a_i\Psi(x_i), \sum_{j=1}^n a_j\Psi(x_j)\right\rangle_{\mathcal{F}} \\ &= \left\|\sum_{i=1}^n a_i\Psi(x_i)\right\|^2_{\mathcal{F}} \geq 0.\end{aligned}$$

This proved that K is p.d. Let $\mathcal{H}_0$ be the vector subspace of all functions from $\mathcal{X}$ to $\mathbb{R}$ spanned by $\{K_y\}_{y\in\mathcal{X}}$. We have seen in class that

$$\langle f, K_y\rangle_{\mathcal{H}_0} = f(y),$$

for all $y \in \mathcal{X}$, and $f \in \mathcal{H}_0$. We may complete the space, by taking equivalence classes of Cauchy sequences from $\mathcal{H}_0$ to obtain Hilbert space $\mathcal{H}$. We must show that every element

of $\mathcal{H}$ is actually a function on $\mathcal{X}$. To do this, let $h \in \mathcal{H}$ and $\{f_n\}_{n\in\mathbb{N}} \subset \mathcal{H}_0$ be a Cauchy sequence that converges to h. By the Cauchy-Schwartz inequality we have

$$|f_n(x) - f_m(x)| = |\langle f_n - f_m, K_x\rangle_{\mathcal{H}_0}| \leq \|f_n - f_m\|_{\mathcal{H}_0}\sqrt{K(x,x)}.$$

Hence, the sequence is pointwise Cauchy and we may define $h(x) = \lim_{n\to\infty} f_n(x)$. The usual argument shows that this value is independent of the particular Cauchy sequence chosen. If we let $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ to be the inner product on $\mathcal{H}$, then for h as above, we have

$$\langle h, K_y\rangle_{\mathcal{H}} = \lim_{n\to\infty}\langle f_n, K_y\rangle_{\mathcal{H}_0} = \lim_{n\to\infty} f_n(y) = h(y).$$

Thus, $\mathcal{H}$ is RKHS on $\mathcal{X}$ and since K_y is the reproducing kernel for the point y, we have that $K(x,y) = K_y(x)$ is the reproducing kernel for $\mathcal{H}$. This proved that if we add to the $\mathcal{H}_0$ the functions defined as the pointwise limits of Cauchy sequence, then $\mathcal{H}$ is RKHS for K. In our case, $\mathcal{H}_0$ is given with

$$\mathcal{H}_0 = \operatorname{span}\{K_y \mid y \in \mathcal{X}\} = \operatorname{span}\{\langle\Psi(y), \Psi(\cdot)\rangle_{\mathcal{F}} \mid y \in \mathcal{X}\} = \operatorname{span}\{\langle w, \Psi(\cdot)\rangle_{\mathcal{F}} \mid w \in \mathcal{F}\}.$$

Which yields that our RKHS is given by

$$\mathcal{H}_\Psi = \{f : X \to \mathbb{R} \mid (\exists w \in \mathcal{F})(\forall x \in \mathcal{X})\ f(x) = \langle w, \Psi(x)\rangle_{\mathcal{F}}\}.$$

3. **Prove that for any p.d. kernel K on a space $\mathcal{X}$, a function $f : \mathcal{X} \to \mathbb{R}$ belongs to the RKHS $\mathcal{H}$ with kernel K if and only if there exists $\lambda > 0$ such that**

$$K(x,x') - \lambda f(x)f(x')$$

is p.d.

($\Rightarrow$) For $n \in \mathbb{N}$, $(a_1, a_2, ..., a_n) \in \mathbb{R}^n$ and $(x_1, x_2, ..., x_n) \in \mathcal{X}^n$ we can define $g = \sum_{i=1}^n a_i K_{x_i}$. Since $f \in \mathcal{H}$, it can be represented as $f(x) = \langle f, K_x\rangle_{\mathcal{H}}$. Then, we have

$$\begin{aligned}
\sum_{i=1}^n\sum_{j=1}^n a_i a_j f(x_i)f(x_j) &= \left(\sum_{i=1}^n a_i f(x_i)\right)^2 \\
&= \left(\sum_{i=1}^n a_i\langle f, K_{x_i}\rangle_{\mathcal{H}}\right)^2 \\
&= \left\langle f, \sum_{i=1}^n a_i K_{x_i}\right\rangle_{\mathcal{H}}^2 \\
&= \langle f, g\rangle_{\mathcal{H}}^2 \leq \|f\|_{\mathcal{H}}^2\|g\|_{\mathcal{H}}^2 \ \text{(Cauchy-Schwarz)} \\
&= \|f\|_{\mathcal{H}}^2 \sum_{i=1}^n\sum_{j=1}^n a_i a_j K(x_i, x_j).
\end{aligned}$$

This means, we have proved that

$$\sum_{i=1}^n\sum_{j=1}^n a_i a_j \left(K(x_i,x_j) - \frac{1}{\|f\|_{\mathcal{H}}^2} f(x_i)f(x_j)\right) \geq 0,$$

i.e. $K(x,x') - \lambda f(x)f(x')$ is p.d. for $\lambda = \frac{1}{\|f\|_{\mathcal{H}}^2}$.

($\Leftarrow$) Let us assume now that $K(x,x') - \lambda f(x)f(x')$ is p.d. Using the second problem in this exercises we know that $\widetilde{K}$, defined by $\widetilde{K}(x,x') = f(x)f(x') = \langle f(x), f(x')\rangle_{\mathbb{R}}$, is also p.d. Using the first problem in this exercises we have

$$\mathcal{H}_K = \mathcal{H}_{K-\lambda\widetilde{K}} + \mathcal{H}_{\lambda\widetilde{K}} \supset \mathcal{H}_{\lambda\widetilde{K}} = \mathcal{H}_{\widetilde{K}}.$$

The inclusion is there because $\mathcal{H}$s are vector spaces, and for fixing zero vector in $\mathcal{H}_{K-\lambda\widetilde{K}}$ we will get $\mathcal{H}_{\lambda\widetilde{K}}$. The last equality is true because if for every $x \in \mathcal{X}$ it is

$$f(x) = \left\langle f, \lambda\widetilde{K}\right\rangle,$$

then we have

$$\underbrace{\lambda f(x)}_{g(x)} = \left\langle \lambda f, \widetilde{K} \right\rangle,$$

which yields $\mathcal{H}_{\lambda\widetilde{K}} = \mathcal{H}_{\widetilde{K}}$. Note: this does not mean that the Hilbert spaces are the same, they are not, because the norms will be different. But the sets will be the same. Finally, the only thing to be noticed is that $f \in \mathcal{H}_{\widetilde{K}}$, and thus $f \in \mathcal{H}_K$. This is true because pre-Hilbert space $\mathcal{H}_0$ from the Moore theorem is spanned by functions $K_x = f(x)f$, i.e.

$$\mathcal{H}_0 = \operatorname{span}\{\widetilde{K}_x \mid x \in \mathcal{X}\} = \operatorname{span}\{f(x)f \mid x \in \mathcal{X}\} = \operatorname{span}\{\alpha f \mid \alpha \in \mathbb{R}\}.$$

Having in mind that $\mathcal{H}_0$ is just the one-dimensional space spanned by f, and since finite dimensional spaces are automatically complete, $\mathcal{H}_{\widetilde{K}}$ is just the span of f. Thus, $f \in \mathcal{H}_{\widetilde{K}}$.

References

[1] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, pages 416–426, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.