Kernel Methods in Machine Learning - Homework

Tong ZHAO (tong.zhao@eleves.enpc.fr)

Exercise 1

(1) For $\forall n \in \mathbb{N}$, $x_1, \cdots, x_n \in \mathcal{X}$, $c_1, \cdots, c_n \in \mathbb{R}$, we have:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \cos(x_i - x_j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \Big(\cos x_i \cos x_j + \sin x_i \sin x_j \Big) \\
&= \sum_{i=1}^{n} c_i \cos x_i \Big(\sum_{j=1}^{n} \cos x_j \Big) + \sum_{i=1}^{n} c_i \sin x_i \Big(\sum_{j=1}^{n} \sin x_j \Big) \\
&= \| \sum_{i=1}^{n} c_i \cos x_i \|^2 + \| \sum_{i=1}^{n} c_i \sin x_i \|^2 \geq 0
\end{aligned}$$

So we conclude that the kernel $K(x, y) = \cos(x - y)$ is a positive definite kernel.

(2) For any $x, y \in \mathcal{X}$, $K(x, y) = \dfrac{1}{1 - x^T y}$ can be expanded by its Maclaurin Series: $\dfrac{1}{1 - x^T y} = \sum_{k=0}^{\infty} (x^T y)^k$. It converges since $\|x\|_2 < 1$ and $\|y\|_2 < 1$.

For $\forall n \in \mathbb{N}$, $x_1, \cdots, x_n \in \mathcal{X}$, $c_1, \cdots, c_n \in \mathbb{R}$, we have:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \frac{1}{1 - x_i^T x_j} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \sum_{k=1}^{\infty} (x_i^T x_j)^k \\
&= \sum_{k=1}^{\infty} \Big(\sum_{i=1}^{n} c_i^{1/k} x_i^T (\sum_{j=1}^{n} c_j^{1/k} x_j) \Big)^k \\
&= \sum_{k=1}^{\infty} \Big(\sum_{i=1}^{n} c_i^{1/k} x_i^T \Big)^{2k} \geq 0
\end{aligned}$$

So we conclude that the kernel $K(x, y) = \dfrac{1}{1 - x^T y}$ is positive definite.

(3) We consider a feature transformation $\Phi : \mathcal{X} \to [-1, 1]$ defined by $\Phi(A) = \mathbb{1}_A - P(A)$, where the indicator function $\mathbb{1}_A$ takes 1 on the set A and 0 otherwise. The inner product of this transformation between two set $A, B \subset \mathcal{A}$ is thus:

$$\begin{aligned}
\langle \Phi(A), \Phi(B) \rangle &= \int_{x \in \mathcal{A}} \big(\mathbb{1}_{x\in A} - P(A)\big)\big(\mathbb{1}_{x\in B} - P(B)\big) d\mu x \\
&= \int_{x \in \mathcal{A}} \mathbb{1}_{x\in A}\mathbb{1}_{x\in B} d\mu x - P(A) \int_{x \in \mathcal{A}} \mathbb{1}_{x\in B} d\mu x \\
&\quad - P(B) \int_{x \in \mathcal{A}} \mathbb{1}_{x\in A} d\mu x + P(A)P(B) \int_{x \in \mathcal{A}} d\mu x \\
&= P(A \cap B) - P(A)P(B) - P(B)P(A) + P(A)P(B) \\
&= P(A \cap B) - P(A)P(B)
\end{aligned}$$

Now we prove that the kernel $K(A, B) = P(A \cap B) - P(A)P(B) = \langle \Phi(A), \Phi(B) \rangle$ is positive definite. For $\forall n \in \mathbb{N}$, $A_1, \cdots, A_n \in \mathcal{A}$, $c_1, \cdots, c_n \in \mathbb{R}$, we have:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(A_i, A_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \left\langle \Phi(A_i), \Phi(A_j) \right\rangle \\
&= \left\langle \sum_{i=1}^{n} c_i \Phi(A_i), \sum_{j=1}^{n} c_j \Phi(A_j) \right\rangle \\
&= \left| \sum_{i=1}^{n} c_i \Phi(A_i) \right|^2 \geq 0
\end{aligned}$$

which gives us the conclusion.

(4) First of all, we prove that the min function of a non-negative functions f is a positive definite kernel. For $\forall n \in \mathbb{N}$, $x_1, \cdots, x_n \in \mathcal{X}$, $c_1, \cdots, c_n \in \mathbb{R}$, we have:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_f(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \min(f(x_i), f(x_j)) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \int_0^{+\infty} \mathbb{1}_{t \leq f(x_i)}(t) \mathbb{1}_{t \leq f(x_j)}(t) dt \\
&= \int_0^{+\infty} \Big(\sum_{i=1}^{n} c_i \mathbb{1}_{t \leq f(x_i)}(t) \Big) \Big(\sum_{j=1}^{n} c_j \mathbb{1}_{t \leq f(x_j)}(t) \Big) dt \\
&= \int_0^{+\infty} \Big(\sum_{i=1}^{n} c_i \mathbb{1}_{t \leq f(x_i)}(t) \Big)^2 \geq 0
\end{aligned}$$

Thus the kernel $K_{fg} = K_f K_g$ is also a positive definite kernel given f and g are non-negative. Then we show that the kernel $K(x,y) = \min(f(x)g(y), f(y)g(x))$ is positive definite.

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_f(x_i, x_j) K_g(x_i, x_j) \geq 0$$

which shows that $K(x,y)$ is positive definite.

(5) First of all, we show that the intersection kernel K_1 is positive definite. The indicator function I is defined by $I_A(a) = \mathbb{1}_{a\in A}$. Then for $\forall n \in \mathbb{N}$, $x_1, \cdots, x_n \in \mathcal{X}$, $c_1, \cdots, c_n \in \mathbb{R}$, we have:

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K_1(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \sum_{a\in E} I_{x_i}(a) I_{x_j}(a) \\
&= \sum_{a\in E}\sum_{i=1}^{n} c_i I_{x_i}(a) \big(\sum_{j=1}^{n} c_j I_{x_j}(a)\big) \\
&= \sum_{a\in E}\Big(\sum_{i=1}^{n} c_i I_{x_i}(a)\Big)^2 \\
&\geq 0
\end{aligned}$$

Now we consider the kernel $K_2(A,B) = \dfrac{1}{|A\cup B|} = \dfrac{1}{1 - |(E\setminus A)\cap(E\setminus B)|}$. By taking $A' = (E\setminus A)$, $B' = (E\setminus B)$, we have:

$$\begin{aligned}
K_2(A,B) &= \frac{1}{1 - |A'\cap B'|} \\
&= \sum_{k=0}^{\infty} |A'\cap B'|^k \quad \text{(positive definite kernels)}
\end{aligned}$$

So we conclude that K_2 is also positive definite by using the same argument in (2). We deduce that $K(A,B) = K_1(A,B)K_2(A,B) = \dfrac{|A\cap B|}{|A\cup B|}$ is also positive definite.

Exercise 2

(1) For $\forall n \in \mathbb{N}$, $x_1, \cdots, x_n \in \mathcal{X}$, we note the gram matrix associated with K_1 is $\mathbb{K}_1$, the one associated with K_2 is $\mathbb{K}_2$. Thus the gram matrix associated with the defined kernel is simply $\mathbb{K} = \alpha\mathbb{K}_1 + \beta\mathbb{K}_2$. For all vectors $c \in \mathbb{R}^n$, we have:

$$\begin{aligned} c^T\mathbb{K}c &= c^T\Big(\alpha\mathbb{K}_1 + \beta\mathbb{K}_2\Big)c \\ &= \alpha c^T\mathbb{K}_1 c + \beta c^T\mathbb{K}_2 c \\ &\geq 0 \quad (K_1, K_2 \text{ positive definite}) \end{aligned}$$

Its RKHS $\mathcal{H}$ is the sum of the two RKHSs of K_1 and K_2, which contains all the functions of the form: $f_x = K(x,\cdot) = \alpha K_1(x,\cdot) + \beta K_2(x,\cdot)$.

(2) For $\forall n \in \mathbb{N}$, $x_1, \cdots, x_n \in \mathcal{X}$, $c_1, \cdots, c_n \in \mathbb{R}$, we have:

$$\begin{aligned} \sum_{i=1}^n \sum_{j=1}^n c_i c_j K(x_i, x_j) &= \sum_{i=1}^n \sum_{j=1}^n \langle c_i\Phi(x_i), c_j\Phi(x_j)\rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^n c_i\Phi(x_i), \sum_{j=1}^n c_j\Phi(x_j) \right\rangle_{\mathcal{H}} \\ &= \Big| \sum_{i=1}^n c_i\Phi(x_i)\Big|_{\mathcal{H}}^2 \geq 0 \end{aligned}$$

Its RKHS $\mathcal{H}$ is the set of all functions of the form: $\{f_x : \mathcal{X} \to \mathbb{R} | f_x(y) = \langle \Phi(x), \Phi(y)\rangle, x \in \mathcal{X}\}$.

Exercise 3

(1) First of all, we show that the bilinear form defines an inner product on $\mathcal{H}$, which means that $\langle f, f\rangle_{\mathcal{H}} = 0$ if and only if $f = 0$. Since f is absolutely continuous and $f(0) = 0$, we have:

$$f(x) = \int_0^x f'(t)dt = \int_0^1 f'(t)\mathbb{1}_{0\leq t\leq x}dt$$

By Cauchy-Schwartz inequality:

$$\begin{aligned}
|f(x)| &= |\int_0^1 f'(t)\mathbb{1}_{0\le t\le x}dt| \\
&\le \big(\int_0^1 f'(t)^2 dt\big)^{1/2}\big(\int_0^1 \mathbb{1}_{0\le t\le x}dt\big)^{1/2} \\
&= \|f\|\sqrt{x}
\end{aligned}$$

So $\|f\| = 0$ if and only if $f = 0$. What's more, $|f(x) \le \|f\|$ is hold.

Then we show that $\mathcal{H}$ is complete, which means that every cauchy sequence in $\mathcal{H}$ converges in $\mathcal{H}$. We take a Cauchy sequence $\{f_n\}$ in the well-defined norm, then $\{f_n'\}$ is a Cauchy sequence in $L^2[0,1]$, hence the limit $g = \lim_{n\to\infty} g_n \in L^2[0,1]$. We define a function $f = \lim_{n\to\infty} f_n$. Since $f(x) = \lim_{n\to\infty} f_n(x) = \lim_{n\to\infty}\int_0^x f_n'(t)dt = \int_0^x g(t)dt$, according to the absolutely continuity, we have $f' = g$ almost everywhere, hence $f' \in L^2[0,1]$. What's more, $f(0) = \lim_{n\to\infty} f_n(0) = 0$, so we conclude that $f \in \mathcal{H}$, and then we have $\mathcal{H}$ is a RKHS.

Now we find the reproducing kernel associated with $\mathcal{H}$.

Given a function $f \in \mathcal{H}$, we have $f(x) = \langle f, k_x\rangle = \int_0^1 f'(t)k_x'(t)dt$ and $f(x) = \int_0^1 f'(t)\mathbb{1}_{t\le x}dt$. Thus the kernel can be found by solving the following PDE:

$$\begin{cases} k_x'(t) = \mathbb{1}_{t\le x} \\ k_x(0) = 0 \end{cases}$$

Thus the solution is:

$$K(x,y) = k_y(x) = min(x,y)$$

(2) We proved in the previous exercise that $\mathcal{H}$ is a RKHS. Now we find its associated reproducing kernel.

Given a function $f \in \mathcal{H}$, we have:

$$
\begin{aligned}
f(x) = \langle f, k_x \rangle &= \int_0^1 f'(t)k_x'(t)dt \\
&= f(t)k_x'(t)\Big|_0^1 - \int_0^1 f(t)k_x''(t)dt \\
&= -\int_0^1 f(t)k_x''(t)dt
\end{aligned}
$$

What's more, we have $f(x) = \int_0^1 f(t)\delta_x dt$. So the kernel function can be found by solving the following PDE:

$$
\begin{cases}
k_x''(t) = -\delta_x \\
k_x(0) = k_x(1) = 0
\end{cases}
$$

Thus the solution is:

$$
K(x,y) = k_y(x) = \begin{cases}
(1-y)x & \text{x} \leq \text{y} \\
(1-x)y & \text{x} > \text{y}
\end{cases}
$$

(3) Given the new inner product, we have:

$$
\begin{aligned}
|f(x)|^2 &= |\int_0^1 f'(t)\mathbb{1}_{0\leq t\leq x}dt|^2 \\
&\leq \big(\int_0^1 f'(t)^2 dt\big)\big(\int_0^1 \mathbb{1}_{0\leq t\leq x}dt\big) \\
&\leq \big(\int_0^1 f'(t)^2 dt + \int_0^1 f(t)^2 dt\big)\big(\int_0^1 \mathbb{1}_{0\leq t\leq x}dt\big) \\
&= \|f\|^2 x
\end{aligned}
$$

Again we have $\|f\| = 0$ if and only if $f = 0$ and $|f(x) \leq \|f\|$ is bounded. We have already proved that $\mathcal{H}$ is complete thus we deduce that $\mathcal{H}$ is also a RKHS.

Now we find the reproducing kernel associated with $\mathcal{H}$.

$$
\begin{aligned}
f(x) = \langle f, k_x \rangle &= \int_0^1 f(t)k_x(t) + f'(t)k_x'(t)dt \\
&= \int_0^1 f(t)k_x(t)dt + \int_0^1 k_x'(t)d(f(t)) \\
&= \int_0^1 f(t)k_x(t)dt + f(t)k_x'(t)\Big|_0^1 - \int_0^1 f(t)k_x''(t)dt \\
&= \int_0^1 f(t)\big(k_x(t) - k_x''(t)\big)dt
\end{aligned}
$$

So the kernel function can be found by solving the following ODE:

$$
\begin{cases}
k_x(t) - k_x''(t) = \delta_x \\
k_x(0) = k_x(1) = 0
\end{cases}
$$

Exercise 4

(a) Given that l_y is a convex loss function, the constrained problem is a convec problem in f for which the strong duality holds. In particular f solves the problem if and only if it solves for some dual parameter λ the unconstrained problem:

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

and complimentary slackness holds.

By using the Representer Theorem, the optimal solution f^* admits a representation of the form: $f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$. Suppose that K is the gram matrix of x, the optimization problem can be written as:

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

where $R : \mathbb{R}^n \to \mathbb{R}$ is a function decided by the choice of l.

(b) The Fenchel-Legendre transform R^* is:

$$
\begin{aligned}
R^*(u) &= \sup_{x\in\mathbb{R}^n} x^T u - R(x) \\
&= \sup_{x\in\mathbb{R}^n} \sum_{i=1}^{n}(x_i u_i) - \frac{1}{n}\sum_{i=1}^{n} l_{y_i}(x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} l^*_{y_i}(u_i)
\end{aligned}
$$

Since y takes value from $\{-1, 1\}$, $l^*_y(x)$ can be written as $l^*(xy)$. Thus we have:

$$
R^*(u) = \frac{1}{n}\sum_{i=1}^{n} l^*(y_i u_i)
$$

(c) We define a lagrange variable $\beta \in \mathbb{R}^n$. Let $g(\alpha) = \lambda\alpha^T K\alpha$ and then the dual problem of (3) is written as:

$$
\begin{aligned}
l(\beta) &= \inf_{u\in\mathbb{R}^n, \alpha\in\mathbb{R}^n} R(u) + g(\alpha) + \beta^T(u - K\alpha) \\
&= \inf_{u\in\mathbb{R}^n} \big(R(u) + \beta^T u\big) + \inf_{\alpha\in\mathbb{R}^n} \big(g(\alpha) - (K^T\beta)^T\alpha\big) \\
&= -\sup_{u\in\mathbb{R}^n} \big(-\beta^T u - R(u)\big) - \sup_{\alpha\in\mathbb{R}^n} \big((K^T\beta)^T\alpha - g(\alpha)\big) \\
&= -R^*(-\beta) - g^*(K^T\beta)
\end{aligned}
$$

Now we calculate g^*:

$$
g^*(y) = \sup_{x\in\mathbb{R}^n} x^T y - \lambda x^T K x
$$

The optimal x^* is found when the gradient of $f(x) = x^T y - \lambda x^T K x$ equals to 0, i.e.

$$
\frac{\partial f(x)}{\partial x} = y - 2\lambda K x = 0
$$

Thus $x^* = \frac{1}{2\lambda}K^{-1}y$. We put x^* in g^* and we have:

$$g^*(y) = \frac{1}{4\lambda} y^T K^{-1} y$$

So the dual problem of (3) is:

$$-R^*(-\beta) - g^*(K^T\beta) = -\frac{1}{n}\sum_{i=1}^{n} l_{y_i}^*(-\beta_i) - \frac{1}{4\lambda}\beta^T K \beta$$

We conclude that the dual problem of (3) is:

$$\max_{\beta \in \mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} l^*(y_i\beta_i) - \frac{1}{4\lambda}\beta^T K \beta$$

TODO: explain how to find primal solution after solving the dual problem!!!

(d) We first calculate the Fenchel transform for logistic loss:

$$l^*(yu) = l^*(z) = \sup_{x \in \mathbb{R}} xz - \log\left(1 + e^{-x}\right)$$

The gradient of the function $f(x) = xu - \log(1 + e^{-x})$ is:

$$\frac{\partial f(x)}{\partial x} = z + \frac{e^{-x}}{1 + e^{-x}} = z + 1 - \frac{1}{1 + e^{-x}} = 0$$

So we have $x = \log(z+1) - \log(-z)$ when $z \in (-1, 0)$ and:

$$l^*(z) = z\log(z+1) - z\log(-z) + \log(z+1) = (z+1)\log(z+1) - z\log(z)$$

Thus the dual can be written as:

$$\max_{\beta \in \mathbb{R}^n} \frac{1}{n}\sum_{i=1}^{n} \Big((y_i\beta_i + 1)\log(y_i\beta_i + 1) - y_i\beta_i\log(y_i\beta_i)\Big) - \frac{1}{4\lambda}\beta^T K \beta$$

$$s.t. \quad 0 > y_i\beta_i > -1 \qquad \forall i \in \{1, \cdots, n\}$$

Now we consider the squared hinge loss.

$$l^*(yu) = l^*(z) = \sup_{x \in \mathbb{R}} xz - (1-x)_+^2$$

When $1 - x < 0$, $l^*(z) = \sup_{x \in \mathbb{R}} xz$. When $z > 0$, $l^*(z) = \infty$, otherwise $l^*(z) = z$.

When $1 - x \geq 0$, $l^*(z) = \sup_{x \in \mathbb{R}} xz - (1-x)^2$. When $z > 0$, $l^*(z) = z$, otherwise $l^*(z) = z + \frac{z^2}{4}$.

We conclude that the Fenchel conjugate of squared hinge loss is $l^*(z) = (z + \frac{z^2}{4})\mathbb{1}_{z \leq 0}$. Thus the dual problem of (3) can be written as:

$$\max_{\beta \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \left(y_i \beta_i + \frac{y_i^2 \beta_i^2}{4} \right) - \frac{1}{4\lambda} \beta^T K \beta$$
$$s.t. \quad y_i \beta_i \leq 0 \qquad \forall i \in \{1, \cdots, n\}$$