AI501 Homework 3

20200130 Yujun Kim

May 2023

1 Exercise #4.2

Note that the definition kernel is given by the existence of feature map. Existence of feature map implies positive definiteness of the map(p.65 in the textbook). In particular, every kernel is positive definite.

Anyway, let k_1, k_2 be positive definite kernel on $\mathcal{X}$ and let positive constants α, β be given. Let $x_1, \cdots, x_n \in \mathcal{X}$ and $a_1, \cdots, a_n \in \mathbb{R}$.

Then, $\sum_{i,j} a_i a_j k_1(x_i, x_j) \geq 0, \sum_{i,j} a_i a_j k_2(x_i, x_j) \geq 0$ so that

$$\sum_{i,j} a_i a_j (\alpha k_1(x_i, x_j) + \beta k_2(x_i, x_j)) = \alpha \sum_{i,j} a_i a_j k_1(x_i, x_j) + \beta \sum_{i,j} a_i a_j k_2(x_i, x_j) \geq 0$$

The sums are taken over $i = 1, \cdots n$ and $j = 1, \cdots, n$. Hence, $\alpha k_1 + \beta k_2$ is positive definite.

2 Exercise #4.3

(Step 1) We prove product of two positive definite kernel on $\mathcal{X}$ is positive definite kernel.

Let k_1, k_2 be two kernels with feature map $\phi_1 : \mathcal{X} \longrightarrow \mathcal{H}_1, \phi_2 : \mathcal{X} \longrightarrow \mathcal{H}_2$ respectively. i.e. $k_1(x,y) =< \phi_1(x), \phi_1(y) >_{\mathcal{H}_1}, k_2(x,y) =< \phi_2(x), \phi_2(y) >_{\mathcal{H}_2}$. Let $\mathcal{B}_1 = \{e_i\}_{i \in I}, \mathcal{B}_2 = \{e'_j\}_{j \in J}$ be orthonormal basis of $\mathcal{H}_1, \mathcal{H}_2$ respectively.

Given $x, y \in \mathcal{X}$, let $\phi_1(x) = (\alpha_i)_{i \in I}, \phi_1(y) = (\beta_i)_{i \in I}$ under representation by $\mathcal{B}_1$. In other words, $\phi_1(x) = \sum_{i \in I} \alpha_i e_i, \phi_1(y) = \sum_{i \in I} \beta_i e_i$. Note that α_i, β_i are zero except finite number of i.

Similarly, let $\phi_2(x) = (\gamma_j)_{j \in J}, \phi_2(y) = (\delta_j)_{j \in J}$ under representation by $\mathcal{B}_2$. In other words, $\phi_2(x) = \sum_{j \in J} \gamma_j e'_j, \phi_2(y) = \sum_{j \in J} \delta_j e'_j$. Note that γ_j, δ_j are zero except finite number of j.

Then, $k_1(x,y) = \sum_{i \in I} \alpha_i \beta_i, k_2(x,y) = \sum_{j \in J} \gamma_j \delta_j$.

$$k_1(x,y) k_2(x,y) = \sum_{i \in I} \alpha_i \beta_i \sum_{j \in J} \gamma_j \delta_j = \sum_{i \in I, j \in J} (\alpha_i \gamma_j)(\beta_i \delta_j)$$

Define $\varphi : \mathcal{X} \longrightarrow \mathbb{R}^{I \times J} =: V$ by $x \mapsto (\alpha_i \gamma_j)_{(i,j) \in I \times J}$, where α_i, γ_j are as defined above. φ is well defined by the definition of basis. Then, $\varphi(x) =$

$(\alpha_i\gamma_j)_{(i,j)\in I\times J}, \varphi(y) = (\beta_i\delta_j)_{(i,j)\in I\times J}$. Thus, $k_1(x,y)k_2(x,y) =< \varphi(x), \varphi(y) >_V$, where the inner product is the standard inner product. Components of $\varphi(x), \varphi(y)$ are zeros except finite indices and so the inner product is well defined. Hence, we proved k_1k_2 has a feature map. By the note on the first paragraph on Problem 1(Exercise #4.2), k_1k_2 is positive definite kernel.

(Step 2) Now we prove if k is positive definite kernel and a is nonnegative constant, ak is positive definite kernel. Let $\phi : \mathcal{X} \longrightarrow \mathcal{H}$ be the feature map for k. Then, $k(x,y) =< \phi(x), \phi(y) >_{\mathcal{H}}$. Thus, $ak(x,y) =< \sqrt{a}\phi(x), \sqrt{a}\phi(x) >_{\mathcal{H}}$. i.e.$\sqrt{a}\phi$ is the feature map of ak and so ak is positive definite kernel.

(Step 3) Sum of two positive definite kernel on $\mathcal{X}$ is a positive definite kernel.

Let k_1, k_2 be two positive definite kernel with feature map $\phi_1 : \mathcal{X} \longrightarrow \mathcal{H}_1, \phi_2 : \mathcal{X} \longrightarrow \mathcal{H}_2$ respectively. Let $\varphi : \mathcal{X} \longrightarrow \mathcal{H}_1 \times \mathcal{H}_2 =: \mathcal{H}$ by $\varphi(x) = (\phi_1(x), \phi_2(x))$. Then, we have a natural inherited inner product for the product space $\mathcal{H}$ given by $< (z_1; w_1), (z_2, w_2) >_{\mathcal{H}} =< z_1, z_2 >_{\mathcal{H}_1} + < w_1, w_2 >_{\mathcal{H}_2}$. Under this inner product, $\varphi(x)$ serve as a feature map for $k_1 + k_2$. i.e. sum of two positive definite kernel is a positive definite kernel.

(Step 4) Let k is a positive definite kernel. Let $p(t) = a_nt^n + \cdots a_0$ be a polynomial with nonnegative coefficients. We prove $p(k)$ is a positive definite kernel.

k^r is positive definite kernel all $r \in \mathbb{N}$, proved by using induction and step1. a_rk^r is positive definite kernel by step2. By induction and step3, $p(k)$ is a positive definite kernel.

3 Exercise #4.4

In step1 of Problem2(Exercise #4.3) above, we proved product of two positive definite kernel on same domain is positive definite kernel.

Let $\bar{x} = (x_1, \cdots, x_p), \bar{y} = (y_1, \cdots, y_p) \in \mathcal{X}$. Define $\bar{k}_i(\bar{x}, \bar{y}) := k_i(x_i, y_i)$ for $i = 1, \cdots p$. Then, $\bar{k}_1, \cdots, \bar{k}_p$ are positive definite kernel on $\mathcal{X}$. This is because if $\phi_i : \mathcal{X}_i \longrightarrow \mathcal{H}_i$ is the feature map for kernel k_i, then $\bar{\phi_i} : \mathcal{X} \longrightarrow \mathcal{H}_i$ defined by $\bar{\phi_i}(\bar{x}) = \phi_i(x_i)$ is a feature map for $\bar{k}_i$. By induction and step1 of the problem2, $\bar{k}_1\bar{k}_2 \cdots \bar{k}_p$ is a positive definite kernel. Since

$$k(\bar{x}, \bar{y}) = k_1(x_1, y_1)k_2(x_2, y_2) \cdots k_p(x_p, y_p) = \bar{k}_1(\bar{x}, \bar{y})\bar{k}_2(\bar{x}, \bar{y}) \cdots \bar{k}_p(\bar{x}, \bar{y})$$

we conclude that $k(\bar{x}, \bar{y})$ is a positive definite kernel.

4 Exercise #4.7

(a)

$$\begin{aligned} d_{\mathcal{X}}(x,y)^2 &= ||k(x,\cdot)-k(y,\cdot)||^2_{\mathcal{H}} \\ &=< k(x,\cdot)-k(y,\cdot), k(x,\cdot)-k(y,\cdot) > \\ &=< k(x,\cdot), k(x,\cdot) > + < k(y,\cdot), k(y,\cdot) > -2 < k(x,\cdot), k(y,\cdot) > \\ &= k(x,x)+k(y,y)-2k(x,y) \\ &= 2(1-k(x,y)) \end{aligned}$$

(b) It does not satisfy triangle inequality. $d_{\mathcal{X}}(x,y)+d_{\mathcal{X}}(y,z) \geq d_{\mathcal{X}}(x,z)$ is equivalent to $2(1-k(x,y))+2(1-k(y,z)) \geq 2(1-k(x,z))$ and hence equivalent to

$$1+k(x,z) \geq k(x,y)+k(y,z) \tag{1}$$

For $x \in \mathcal{X} = \mathbb{R}$, let $\phi(x) = \frac{1}{||(x,\sqrt{3})||}(x,\sqrt{3}) \in \mathbb{R}^2$. Let $k(x,y) =< \phi(x), \phi(y) >$. Then, k is a normalized kernel. $k(1,0) = k(-1,0) = \sqrt{3}/2$ and $k(1,-1) = 1/2$.

$$1+k(1,-1) = 1.5 < \sqrt{3} = k(1,0)+k(0,-1)$$

and sot (1) is not satisfied for $x = 1, y = 0, z = -1$. It follows that

$$d_{\mathcal{X}}(1,0)+d_{\mathcal{X}}(0,-1) < d_{\mathcal{X}}(1,-1)$$

and so the triangle inequality is not satisfied. $d\mathcal{X}$ is not a distance.

5 Exercise #4.8

(a)

$$\begin{aligned} ||\mu_\phi||^2_{\mathcal{H}} &= ||\frac{1}{n}\sum \phi(x_i)||^2_{\mathcal{H}} \\ &=< \frac{1}{n}\sum_i \phi(x_i), \frac{1}{n}\sum_j \phi(x_j) > \\ &= \frac{1}{n^2}\sum_{i,j} < \phi(x_i), \phi(x_j) > \\ &= \frac{1}{n^2}\sum_{i,j} k(x_i,x_j) \end{aligned}$$

(b)

$$\begin{aligned}
\sigma_\phi^2 &= \frac{1}{n}\sum_i ||\phi(x_i) - \mu_\phi||_{\mathcal{H}}^2 \\
&= \frac{1}{n}\sum \left(||\phi(x_i)||^2 - 2 < \phi(x_i), \mu_\phi > + ||\mu_\phi||^2\right) \\
&= \frac{1}{n}\left(\sum ||\phi(x_i)||^2\right) - 2 < \left(\frac{1}{n}\sum \phi(x_i)\right), \mu_\phi > + ||\mu_\phi||^2 \\
&= \frac{1}{n}Tr(K) - 2||\mu_\phi||^2 + ||\mu_\phi||^2 \\
&= \frac{1}{n}Tr(K) - ||\mu_\phi||^2
\end{aligned}$$

Note that $||\phi(x_i)||^2 = k(x_i, x_i)$ so that $\sum ||\phi(x_i)||^2 = \sum k(x_i, x_i) = Tr(K)$.

6 Exercise #4.9

By representer theorem minimizer of the given problem has represetation of the form

$$f(\cdot) = \sum_{j=1}^n \alpha_j k(x_j, \cdot)$$

Thus, the primal problem becomes

$$\min_{\alpha,\xi} \frac{1}{2}\alpha^T K\alpha + C\sum_{i=1}^n \xi_i^2$$

subject to $1 - y_i \sum_{j=1}^n \alpha_j k(x_j, x_i) \le \xi_i, \xi_i \ge 0$.

Now, the dual problem becomes

$$\max_{\lambda,\gamma} g(\lambda, \gamma)$$

subject to $\lambda \ge 0, \gamma \ge 0$, where

$$\begin{aligned}
g(\lambda, \gamma) &= \min_{\alpha,\xi}\{\frac{1}{2}\alpha^T K\alpha + C\sum_{i=1}^n \xi_i^2 + \sum_{i=1}^n \lambda_i(1 - y_i\sum_{j=1}^n \alpha_j k(x_j, x_i) - \xi_i) - \sum_{i=1}^n \gamma_i\xi_i\} \\
&= \min_{\alpha,\xi}\{\frac{1}{2}\alpha^T K\alpha + \sum_{i=1}^n \lambda_i + (C\xi_i - \lambda_i - \gamma_i)\xi_i - r^T K\alpha\}
\end{aligned}$$

with $r = [y_1\lambda_1, \cdots, y_n\lambda_n]^T$

By FIFO on α, $K\alpha = Kr$ so that $\alpha = r$. By FIFO on ξ,

$$2C\xi - \lambda - \gamma = 0 \Rightarrow \lambda + \gamma = 2C\xi$$

$$
\begin{aligned}
g(\lambda,\gamma) &= \lambda^T \mathbb{1} + (C\xi - \gamma - \lambda)^T \xi - \frac{1}{2} r^T K r \\
&= \lambda^T \mathbb{1} - (\frac{\lambda+\gamma}{2})^T \frac{1}{2C}(\lambda+\gamma) - \frac{1}{2} r^T K r \\
&= \sum_i \lambda_i - \frac{1}{4C}\sum_i (\lambda_i + \gamma_i)^2 - \sum_{i,j} \lambda_i \lambda_j y_i y_j k(x_i, x_j)
\end{aligned}
$$

Compared to 1-SVM, we have additional term starting with coefficient $\frac{1}{4C}$. However, the problem is still quadratic which makes it easy to solve.

7 Exercise #4.10

Let $h(t) = log(1+e^t)$. Then $h'(t) = e^t/(1+e^t)$. Now, the original problem becomes

$$\min_{f\in\mathcal{H}_k} \frac{1}{2}||f||^2 + C\sum_i h(-y_i f(x_i))$$

By representer theorem, the minimizater has representation of the form $f(\cdot) = \sum_j \alpha_j k(x_j, \cdot)$. Thus, $f(x_i) = \sum_j \alpha_j k(x_j, x_i)$. Hence, we have

$$\min_{\alpha,\xi} \frac{1}{2}\alpha^T K\alpha + C\sum_i h(\xi_i)$$

with respect to $\xi_i = -y_i f(x_i) = -y_i \sum_j \alpha_j k(x_j, x_i)$. Now The Lagrangian is

$$
\begin{aligned}
L(\alpha,\xi,\lambda) &= \frac{1}{2}\alpha^T K\alpha + C\sum_i h(\xi_i) + \sum_i \lambda_i(-\xi - y_i \sum_j \alpha_j k(x_j, x_i)) \\
&= \frac{1}{2}\alpha^T K\alpha + C\sum_i h(\xi_i) - \lambda^T\xi - r^T K\alpha
\end{aligned}
$$

where $r = [\lambda_1 y_1, \cdots, \lambda_n y_n]^T$. Dual function is

$$g(\lambda) = \min_{\alpha,\xi} L(\alpha,\xi,\lambda)$$

By FIFO for α, we have $K\alpha - Kr = 0 \Rightarrow \alpha = r$. By $\partial L/\partial\xi_i = 0$, we have

$$Ch'(\xi_i) = \lambda_i \Rightarrow C\xi_i/(1+\xi_i) = \lambda_i$$

Thus, $\xi_i = \lambda_i/(C-\lambda_i)$. Hence,

$$g(\lambda) = \sum_i \left[Ch(\frac{\lambda_i}{C-\lambda_i}) - \frac{\lambda_i^2}{C-\lambda_i} \right] - \frac{1}{2} r^T K r$$

The dual problem is

$$\max_\lambda g(\lambda)$$

To solve dual problem, we have to use FIFO.

$$\partial g/\partial \lambda_i = Ch'(\frac{\lambda_i}{C-\lambda_i}) \times \frac{C}{(C-\lambda_i)^2} - \frac{\lambda_i(2-\lambda_i)}{(C-\lambda_i)^2} - 2y_i\lambda_i \sum_j y_j\lambda_j k(x_i, x_j)$$