Quiz #2: Kernel Methods for Machine Learning

Problem 1

Let $\mathcal{X}$ be a set.
1. Give the definition of a positive definite (p.d.) kernel on $\mathcal{X}$.
2. If K_1 and K_2 are p.d. kernels on $\mathcal{X}$, show that $K = K_1 + K_2$ is p.d. on $\mathcal{X}$.
3. If K_1 is a p.d. kernel on $\mathcal{X}$ and $\lambda \in \mathbb{R}^+$, show that $K = \lambda K_1$ is p.d. on $\mathcal{X}$.
4. Are the following kernels p.d.? And why?

- For any $\mathcal{X}$:
$$\forall x, x' \in \mathcal{X}\,, \quad K_1(x, x') = C\,,$$
for a constant $C \in \mathbb{R}$.

- For $\mathcal{X} = \mathbb{R}$:
$$\forall x, x' \in \mathbb{R}\,, \quad K_2(x, x') = e^{x+x'}\,.$$

- For $\mathcal{X} = \mathbb{R}^+$:
$$\forall x, x' \in \mathbb{R}^+\,, \quad K_3(x, x') = \min(x, x')\,.$$

- For $\mathcal{X} = \mathbb{R}$:
$$\forall x, x' \in \mathbb{R}\,, \quad K_4(x, x') = \min(x, x')\,.$$

- For $\mathcal{X} = \mathbb{R}^+$:
$$\forall x, x' \in \mathbb{R}^+\,, \quad K_5(x, x') = \max(x, x')\,.$$

Solutions:
1. There are many answers to this question.

Definition 1. A p.d. kernel on a set $\mathcal{X}$ is a function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that is symmetric:

$$\forall x, x' \in \mathcal{X}\,, \quad K(x, x') = K(x', x)\,,$$

and that satisfies: $\forall n \in \mathbb{N}, \forall x_1, \ldots, x_n \in \mathcal{X}, \forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}$, it holds that

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

Or equivalently, $\forall n \in \mathbb{N}, \forall x_1, \ldots, x_n \in \mathcal{X}$, the Gram matrix $\mathbf{K}$ is a symmetric, positive semi-definite matrix.

Definition 2. Due to Aronszajn's theorem, a kernel is p.d. over $\mathcal{X}$ if and only if there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that, $\forall x, x' \in \mathcal{X}$:

$$K(x, x') = \Phi(x)^\top \Phi(x')\,.$$

2. It is trivial that K is symmetric. $\forall n \in \mathbb{N}, \forall x_1, \ldots, x_n \in \mathcal{X}, \forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}$,

$$\begin{aligned}\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K_1(x_i, x_j) + \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K_2(x_i, x_j) \\ &\geq 0\,,\end{aligned}$$

since K_1 and K_2 are p.d. kernels. K is therefore p.d. by definition.

3. It is trivial that K is symmetric. $\forall n \in \mathbb{N}, \forall x_1, \ldots, x_n \in \mathcal{X}, \forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}$,

$$\begin{aligned}\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) &= \lambda \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K_1(x_i, x_j) \\ &\geq 0\,,\end{aligned}$$

since K_1 is p.d. and $\lambda \geq 0$. K is therefore p.d. by definition.

4.

- K_1 is p.d. if and only if $C \geq 0$. By definition, it is trivial that K_1 is symmetric. $\forall n \in \mathbb{N}, \forall x_1, \ldots, x_n \in \mathcal{X}, \forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K_1(x_i, x_j) = C \left(\sum_{i=1}^{n} \alpha_i \right)^2 \begin{cases} \geq 0 & \text{if } C \geq 0\,, \\ \leq 0 & \text{if } C \leq 0\,.\end{cases}.$$

- K_2 is p.d. By definition, $K_2(x, x') = \Phi(x) \cdot \Phi(x')$ where $\Phi(x) = e^x$.

- K_3 is p.d. The symmetry is trivial. Now we show that, $\forall n \in \mathbb{N}, \forall x_1, \ldots, x_n \in \mathbb{R}^+$, the Gram matrix

$$\mathbf{K} = [\min(x_i, x_j)]_{i,j=1,\ldots,n}$$

is a positive semi-definite matrix. This is equivalent to showing that all the eigenvalues of $\mathbf{K}$ are non-negative, or equivalently that the determinants of all leading principle minors of $\mathbf{K}$ are non-negative. Without loss of generality, we may assume that $0 \le x_1 \le \cdots \le x_n$, we have

$$\mathbf{K} = \begin{bmatrix} x_1 & x_1 & \cdots & x_1 & x_1 \\ x_1 & x_2 & \cdots & x_2 & x_2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_1 & x_2 & \cdots & x_{n-1} & x_{n-1} \\ x_1 & x_2 & \cdots & x_{n-1} & x_n \end{bmatrix}.$$

Let us first show that $\det(\mathbf{K}) \ge 0$. In fact,

$$\begin{aligned} \det(\mathbf{K}) &= \det \begin{bmatrix} x_1 & 0 & \cdots & 0 & 0 \\ x_1 & x_2 - x_1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_1 & x_2 - x_1 & \cdots & x_{n-1} - x_{n-2} & 0 \\ x_1 & x_2 - x_1 & \cdots & x_{n-1} - x_{n-2} & x_n - x_{n-1} \end{bmatrix} \\ &= x_1 \prod_{i=2}^{n} (x_i - x_{i-1}), \end{aligned}$$

where the determinant of $\mathbf{K}$ remains the same when we sequentially subtract the $(n-1)$-th from the n-th column, then subtract the $(n-2)$-th column from the $(n-1)$-th column, ..., until finally we subtract the first column from the second column. Since we have assumed that $0 \le x_1 \le \cdots \le x_n$, we know $\det(\mathbf{K}) \ge 0$.

Using mathematical induction on all the leading principle minors of $\mathbf{K}$, we know $\mathbf{K}$ is a positive semi-definite matrix. Therefore K_3 is p.d.

- K_4 is not p.d. Similarly to the reasoning for K_3, we know that, $\forall x_1 \le \cdots \le x_n$, $\det(\mathbf{K}) = x_1 \prod_{i=2}^{n}(x_i - x_{i-1})$, which can be negative if $x_1 < 0$. Alternatively, you may reason with a counterexample using a particular set of x_i's.

- K_5 is not p.d. Similarly to the reasoning for K_3, we know that, $\forall x_1 \geq \cdots \geq x_n \geq 0$, $\det(\mathbf{K}) = x_1 \prod_{i=2}^{n}(x_i - x_{i-1})$, which can be negative if n is an even number. Alternatively, you may reason with a counterexample using a particular set of x_i's.

Problem 2

Let K be a p.d. kernel on a set $\mathcal{X}$, and $\Phi : \mathcal{X} \to \mathcal{F}$ a mapping to a Hilbert space $\mathcal{F}$ (i.e., a "feature space") such that

$$\forall x, x' \in \mathcal{X}\,, \quad K(x, x') = \Phi(x)^\top \Phi(x')\,.$$

Let $d_K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be the distance in the feature space, i.e.,

$$\forall x, x' \in \mathcal{X}\,, \quad d_K(x, x') = \|\, \Phi(x) - \Phi(x')\, \|\,.$$

1. For any $x, x' \in \mathcal{X}$, show that we can compute $d_K(x, x')$ using K only (i.e., without Φ).
2. Application: take $\mathcal{X} = \mathbb{R}$ and $K(x, x') = e^{-(x-x')^2}$, compute $d_K(1, 2)$.
3. Show that $-d_K^2$ is *conditionally positive definite*, that is: $\forall n \in \mathbb{N}$, $\forall x_1, \ldots, x_n \in \mathcal{X}$, $\forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that $\sum_{i=1}^n \alpha_i = 0$, it holds that

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j d_K(x_i, x_j)^2 \leq 0\,.$$

4. Given a set of n points $\mathcal{S} = (x_1, \ldots, x_n) \in \mathcal{X}^n$, let $m_{\mathcal{S}}$ be their barycenter in the feature space, i.e.,

$$m_{\mathcal{S}} = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i)\,.$$

- Show that the function $K_{\mathcal{S}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as

$$\forall x, x' \in \mathcal{X}\,, \quad K_{\mathcal{S}}(x, x') = (\Phi(x) - m_{\mathcal{S}})^\top (\Phi(x') - m_{\mathcal{S}})$$

  is a p.d. kernel on $\mathcal{X}$.

- For any $x, x' \in \mathcal{X}$, express $K_{\mathcal{S}}(x, x')$ using only the kernel K (i.e., without Φ or m).

- Let $\mathbf{K}$ and $\mathbf{K}_{\mathcal{S}}$ be the Gram matrices of K and $K_{\mathcal{S}}$ on $\mathcal{S}$ (i.e., the $n \times n$ matrices such that $[\mathbf{K}]_{ij} = K(x_i, x_j)$ and $[\mathbf{K}_{\mathcal{S}}]_{ij} = K_{\mathcal{S}}(x_i, x_j)$). Find an $n \times n$ matrix $\mathbf{A}$ such that

$$\mathbf{K}_{\mathcal{S}} = \mathbf{A}\mathbf{K}\mathbf{A}\,.$$

<u>**Solutions**</u>:

1. By definition,

$$\begin{aligned} d_K(x, x') &= \sqrt{\| \Phi(x) - \Phi(x') \|^2} \\ &= \sqrt{(\Phi(x) - \Phi(x'))^\top (\Phi(x) - \Phi(x'))} \\ &= \sqrt{K(x,x) + K(x',x') - 2K(x,x')}\,. \end{aligned} \tag{1}$$

2. By (1), we have

$$d_K(1,2) = \sqrt{e^{-(1-1)^2} + e^{-(2-2)^2} - 2e^{-(1-2)^2}} = \sqrt{2 - 2e^{-1}}\,.$$

3. By (1) and K p.d., $\forall n \in \mathbb{N}$, $\forall x_1, \ldots, x_n \in \mathcal{X}$, $\forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}$ such that $\sum_{i=1}^n \alpha_i = 0$, it holds that

$$\begin{aligned} &\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j d_K(x_i, x_j)^2 \\ =& \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j (K(x_i,x_i) + K(x_j,x_j) - 2K(x_i,x_j)) \\ =& \underbrace{\left(\sum_{j=1}^n \alpha_j\right)}_{=0} \left(\sum_{i=1}^n \alpha_i K(x_i,x_i)\right) + \underbrace{\left(\sum_{i=1}^n \alpha_i\right)}_{=0} \left(\sum_{j=1}^n \alpha_j K(x_j,x_j)\right) - 2\underbrace{\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i,x_j)}_{\geq 0} \\ \leq& 0\,. \end{aligned}$$

4.

- Denote by $\Phi_{\mathcal{S}} : \mathcal{X} \to \mathcal{F}$ the mapping defined by $\Phi_{\mathcal{S}}(x) = \Phi(x) - m_{\mathcal{S}}$, we have

$$\forall x, x' \in \mathcal{X}\,, \quad K_{\mathcal{S}}(x, x') = \Phi_{\mathcal{S}}(x)^\top \Phi_{\mathcal{S}}(x')\,.$$

  Therefore, $K_{\mathcal{S}}$ is p.d. by definition.

- Plugging the definition of $m_{\mathcal{S}}$ into $K_{\mathcal{S}}$, we have

$$K_{\mathcal{S}}(x,x') = \left(\Phi(x) - \frac{1}{n}\sum_{i=1}^{n}\Phi(x_i)\right)^{\top}\left(\Phi(x') - \frac{1}{n}\sum_{j=1}^{n}\Phi(x_j)\right)$$
$$= K(x,x') - \frac{1}{n}\sum_{j=1}^{n}K(x,x_j) - \frac{1}{n}\sum_{i=1}^{n}K(x',x_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(x_i,x_j)\,.$$

- Let $\mathbf{\Phi}, \mathbf{\Phi}_{\mathcal{S}}$ be the feature matrix corresponding to $\mathbf{K}, \mathbf{K}_{\mathcal{S}}$ respectively, i.e.:

$$\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\top}\,, \quad \mathbf{K}_{\mathcal{S}} = \mathbf{\Phi}_{\mathcal{S}}\mathbf{\Phi}_{\mathcal{S}}^{\top}\,,$$

  where $\mathbf{\Phi} = (\Phi(x_1)|\ldots|\Phi(x_n))^{\top}$ whose row vectors consist of the feature vectors of $x_1, \ldots, x_n$, and similarly for $\mathbf{\Phi}_{\mathcal{S}}$. Denote by $\mathbf{1}$ the $n \times n$ matrix of 1's, it is easy to verify that

$$\mathbf{\Phi}_{\mathcal{S}} = \mathbf{\Phi} - \frac{1}{n}\mathbf{1}\mathbf{\Phi} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\right)\mathbf{\Phi}\,.$$

  Denote by

$$\mathbf{A} = \mathbf{I} - \frac{1}{n}\mathbf{1} = \begin{bmatrix} 1-\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \ddots & \vdots \\ -\frac{1}{n} & \cdots & 1-\frac{1}{n} \end{bmatrix}_{n\times n}\,,$$

  we have $\mathbf{A}^{\top} = \mathbf{A}$ and

$$\mathbf{K}_{\mathcal{S}} = \mathbf{\Phi}_{\mathcal{S}}\mathbf{\Phi}_{\mathcal{S}}^{\top} = \mathbf{A}\mathbf{\Phi}\mathbf{\Phi}^{\top}\mathbf{A} = \mathbf{A}\mathbf{K}\mathbf{A}\,.$$