Lecture-08: Reproducing Kernel Hilbert Space (RKHS)

1 Reproducing Kernel Hilbert Space (RKHS)

Definition 1.1. For any PDS kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, we can define a kernel evaluation map $e_x : \mathcal{X} \to \mathbb{R}$ at a point $x \in \mathcal{X}$ by $e_x(x') \triangleq k(x, x')$ for all $x' \in \mathcal{X}$.

Definition 1.2. We can define a pre-Hilbert space $\mathbb{H}_0$ as the span of kernel evaluations defined in Definition 1.1, at finitely many elements of $\mathcal{X}$. That is,

$$\mathbb{H}_0 \triangleq \left\{ \sum_{x \in I} a_x e_x : \text{ finite } I \subseteq \mathcal{X}, a \in \mathbb{R}^I \right\} \subseteq \mathbb{R}^{\mathcal{X}}.$$

The completion of $\mathbb{H}_0$ is a complete Hilbert space denoted by $\mathbb{H} \triangleq \overline{\mathbb{H}_0}$ and called the *reproducing kernel Hilbert space* associated with kernel k.

Remark 1. Since $e_x \in \mathbb{R}^{\mathcal{X}}$, it follows that $\mathbb{H}_0 \subseteq \mathbb{H} \subseteq \mathbb{R}^{\mathcal{X}}$. We observe that $\mathbb{H}_0$ is dense in $\mathbb{H}$. By definition, we have $e_x \in \mathbb{H}$ for any $x \in \mathcal{X}$.

Definition 1.3. Then, we define a map $\langle \cdot, \cdot \rangle : \mathbb{H}_0 \times \mathbb{H}_0 \to \mathbb{R}$ defined for all $f, g \in \mathbb{H}_0$ such that $f = \sum_{x \in I} a_x e_x$ and $g = \sum_{y \in J} b_y e_y$, as

$$\langle f, g \rangle_{\mathbb{H}_0} \triangleq \sum_{x \in I} \sum_{y \in J} a_x b_y k(x, y) = \sum_{y \in J} b_y f(y) = \sum_{x \in I} a_x g(x).$$

Lemma 1.4. *The map $\langle \cdot, \cdot \rangle : \mathbb{H}_0 \times \mathbb{H}_0 \to \mathbb{R}$ defined in Definition 1.3 for any PDS kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is an inner product.*

Proof. We can verify that the map $\langle \cdot, \cdot \rangle : \mathbb{H}_0 \times \mathbb{H}_0 \to \mathbb{R}$ has the follow three properties.

1. *Symmetry*: By definition, $\langle \cdot, \cdot \rangle$ is symmetric.
2. *Bilinearity*: From symmetry, it suffices to show that $\langle \cdot, \cdot \rangle$ is linear in its first argument. Let $\alpha, \beta \in \mathbb{R}$ and $f, g, h \in \mathbb{H}_0$ such that $f = \sum_{z \in I} a_x e_x, g = \sum_{y \in J} b_y e_y, h = \sum_{z \in K} c_z e_z$. For simplicity, we assume that I and J are disjoint. We observe that $\alpha f + \beta g = \sum_{x \in I \cup J} (\alpha a_x \mathbb{1}_{\{x \in I\}} + \beta b_x \mathbb{1}_{\{x \in J\}}) e_x$. It follows that
$$\langle \alpha f + \beta g, h \rangle = \sum_{x \in I \cup J} \sum_{z \in K} (\alpha a_x \mathbb{1}_{\{x \in I\}} + \beta b_x \mathbb{1}_{\{x \in J\}}) c_z k(x, z) = \alpha \langle f, h \rangle + \beta \langle g, h \rangle .$$
3. *Positive semi-definiteness*: We will show that for any $f \in \mathbb{H}_0$ that can be written as $f = \sum_{x \in I} a_x e_x$, we have $\langle f, f \rangle \geqslant 0$. Recall that for any PDS kernel k and sample I, the associated gram matrix K is symmetric and positive semidefinite. It follows that for any column vector $a \in \mathbb{R}^I$, we have
$$\langle f, f \rangle = \sum_{x,y} a_x k(x, y) a_y = a^\top K a \geqslant 0.$$

It follows that $\langle \cdot, \cdot \rangle$ is an inner product on pre-Hilbert space $\mathbb{H}_0$. □

Theorem 1.5 (RKHS). *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel. Then, there exists a Hilbert space $\mathbb{H}$ and a mapping $\Phi : \mathcal{X} \to \mathbb{H}$ such that for all $x, x' \in \mathcal{X}$,*

$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathbb{H}} .$$

Furthermore, $\mathbb{H}$ has the following reproducing property, $h(x) = \langle (h(\cdot), k(x, \cdot) \rangle_{\mathbb{H}}$ for all $h \in \mathbb{H}$ and $x \in \mathcal{X}$.

Proof. We define a feature map $\Phi : \mathcal{X} \to \mathbb{H}$ as $\Phi(x) \triangleq e_x$ for all $x \in \mathcal{X}$, where e_x is the kernel evaluation map defined in Definition 1.1 associated with PDS kernel k. It follows that $\Phi(x) \in \mathbb{H} \subseteq \mathbb{R}^{\mathcal{X}}$ from Remark 1. From definition, it follows that $[\Phi(x)](x') = k(x,x')$ for all $x' \in \mathcal{X}$. From the definition of inner product on pre-Gilbert space $\mathbb{H}_0$, we observe that for all $x, x' \in \mathcal{X}$,

$$\left\langle \Phi(x), \Phi(x') \right\rangle = \langle e_x, e_{x'} \rangle = k(x,x').$$

We can verify that the inner product $\langle \cdot, \cdot \rangle : \mathbb{H}_0 \times \mathbb{H}_0 \to \mathbb{R}$ has the following two additional properties.

1. *Reproducing property:* Consider a kernel evaluation map $e_{x'} \in \mathbb{H}$ and $f \in \mathbb{H}_0$ such that $f = \sum_{x \in I} a_x e_x$ for any finite $I \subseteq \mathcal{X}$ and $a \in \mathbb{R}^I$. Then,

$$\langle f, e_{x'} \rangle = \sum_{x \in I} a_x k(x,x') = \sum_{x \in I} a_x e_x(x') = f(x').$$

2. *Definiteness*: From the Cauchy-Schwarz inequality for inner products and reproducing property of $\mathbb{H}$, we observe that for any $f \in \mathbb{H}_0$ and $x \in \mathcal{X}$,

$$|f(x)|^2 = |\langle f, e_x \rangle|^2 \leqslant \langle f, f \rangle \langle e_x, e_x \rangle = (a^\top K a) k(x,x).$$

   It follows that $f(x)$ is bounded for any $f \in \mathbb{H}_0$ and $x \in \mathcal{X}$.

Since $\langle \cdot, \cdot \rangle$ is an inner product on $\mathbb{H}_0$ which is bounded, it follows that $\mathbb{H}_0$ is a pre-Hilbert space which can be made complete to form the Hilbert space $\mathbb{H} \triangleq \overline{\mathbb{H}}_0$, where $\mathbb{H}_0$ is dense in $\mathbb{H}$. □

1.1 Representer theorem

Observe that modulo the offset b, the hypothesis solution of SVMs can be written as a linear combination of the functions $k(x_i, \cdot)$, where x_i is a sample point. The following theorem known as the representer theorem shows that this is in fact a general property that holds for a broad class of optimization problems, including that of SVMs with no offset.

Theorem 1.6 (Representer). *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a PDS kernel with associated kernel evaluation map e_x for any $x \in \mathcal{X}$ and corresponding RKHS $\mathbb{H}$. Then, for any non decreasing function $G : \mathbb{R} \to \mathbb{R}$ and any loss function $L : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, the optimization problem*

$$\arg\min_{h \in \mathbb{H}} F(h) = \arg\min_{h \in \mathbb{H}} G(\|h\|_{\mathbb{H}}) + L(h(x_1), \dots, h(x_m)),$$

has a solution of the form $h^* = \sum_{i=1}^m \alpha_i e_{x_i}$. If G is strictly increasing, then any solution has this form.

Proof. Let $\mathbb{H}_1 = \text{span}(e_{x_i} : i \in [m])$. We can write the RKHS $\mathbb{H}$ as the direct sum of span of $\mathbb{H}_1$ and the orthogonal space $\mathbb{H}_1^\perp$, i.e. $\mathbb{H} = \mathbb{H}_1 \oplus \mathbb{H}_1^\perp$. Hence, any hypothesis $h \in \mathbb{H}$, can be written as $h = h_1 + h_1^\perp$. By the reproducing property, we have $h(x_i) = \langle h, e_{x_i} \rangle = \langle h_1, e_{x_i} \rangle = h_1(x_i)$ for all $i \in [m]$. Therefore, $L(h(x_1), \dots, h(x_m)) = L(h_1(x_1), \dots, h_1(x_m))$. Further, since G is non-decreasing $G(\|h_1\|_{\mathbb{H}}) \leqslant G(\sqrt{\|h_1\|_{\mathbb{H}}^2 + \|h_1^\perp\|_{\mathbb{H}}^2}) = G(\|h\|_{\mathbb{H}})$. It follows that $F(h_1) \leqslant F(h)$. Since G is strictly increasing, $F(h_1) < F(h)$ for all h such that $\|h_1^\perp\|_{\mathbb{H}} > 0$. Hence, any solution of the optimization problem must be in $\mathbb{H}_1$. □

Remark 2. Consider a PDS kernel k on input space $\mathcal{X}$, with associated kernel evaluation map e_x for each $x \in \mathcal{X}$ and RKHS $\mathbb{H}$. Representer theorem states that for any feature map $\Phi : \mathcal{X} \to \mathbb{H}$, it suffices to focus on its projection $(\Phi(x_1), \dots, \Phi(x_m)) = (\langle \Phi(x), e_{x_1} \rangle, \dots, \langle \Phi(x), e_{x_m} \rangle)$ on the subspace $\mathbb{H}_1$ spanned by $(e_{x_1}, \dots, e_{x_m})$ of unlabeled training sample $x \in \mathcal{X}^m$. We observe that the resulting map $\Phi(x_1), \dots, \Phi(x_m) : \mathcal{X} \to \mathbb{R}^m$.

2 Empirical kernel map

Advantages of working with a kernel is that no explicit definition of a feature map Φ is needed. Following are the advantages of working with explicit feature map Φ.

(i) For primal method in various optimization problems.
(ii) To derive an approximation based on Φ.
(iii) Theoretical analysis where Φ is more convenient.

Definition 2.1 (Empirical kernel map). Given an unlabeled training sample $x \in \mathcal{X}^m$ and a PDS kernel k, the associated *empirical kernel map* $E : \mathcal{X} \to \mathbb{R}^m$ is a feature mapping defined as $E \triangleq \begin{bmatrix} e_{x_1} & \dots & e_{x_m} \end{bmatrix}^\top$ such that for all $y \in \mathcal{X}$

$$E(y) = \begin{bmatrix} e_{x_1}(y) \\ \vdots \\ e_{x_m}(y) \end{bmatrix} = \begin{bmatrix} k(x_1,y) \\ \vdots \\ k(x_m,y) \end{bmatrix}.$$

Remark 3. The empirical kernel map evaluated at a point $y \in \mathcal{X}$ is the vector of k-similarity measure of y with each of the m training points.

Remark 4. For any $i \in [m]$, we have $E(x_i) = K^\top e_i = Ke_i$, where e_i is the ith unit vector. Hence, $\langle E(x_i), E(x_j) \rangle = \langle Ke_i, Ke_j \rangle = \langle e_i, K^2 e_j \rangle$. That is, the kernel matrix associated with the empirical kernel map E is K^2.

Definition 2.2. Let $K^\dagger$ denote the pseudo-inverse of the gram matrix K and let $(K^\dagger)^{\frac{1}{2}}$ denote the SPSD matrix whose square is $K^\dagger$. We define a feature map $F : \mathcal{X} \to \mathbb{R}^m$ using the empirical kernel map E and the matrix $(K^\dagger)^{\frac{1}{2}}$ for all $y \in \mathcal{X}$, as

$$F(y) \triangleq (K^\dagger)^{\frac{1}{2}} E(y).$$

Remark 5. Using the identity $KK^\dagger K = K$, we see that

$$\langle F(x_i), F(x_j) \rangle = \left\langle (K^\dagger)^{\frac{1}{2}} E(x_i), (K^\dagger)^{\frac{1}{2}} E(x_j) \right\rangle = \left\langle Ke_i, K^\dagger K e_j \right\rangle = \langle e_i, Ke_j \rangle.$$

Thus, the kernel matrix associated to map F is K.

Remark 6. For the feature mapping $G : \mathcal{X} \to \mathbb{R}^m$ defined by $G(x) \triangleq K^\dagger E(x)$ for all $x \in \mathcal{X}$, we check that the

$$\langle G(x_i), G(x_j) \rangle = \left\langle K^\dagger E(x_i), K^\dagger E(x_j) \right\rangle = \left\langle Ke_i, K^\dagger e_j \right\rangle = \left\langle e_i, KK^\dagger e_j \right\rangle.$$

Thus, the kernel matrix associated to map G is $KK^\dagger$.

3 Kernel-based algorithms

We can generalize SVMs in the input space $\mathcal{X}$ to the SVMs in the feature space $\mathbb{H}$ mapped by the feature mapping Φ. Recall that $k(y,z) = \langle \Phi(y), \Phi(z) \rangle_{\mathbb{H}}$ for all $y, z \in \mathcal{X}$, and hence the gram matrix K generated by the kernel map k and the unlabeled training sample $x \in \mathcal{X}^m$ suffices to describe the SVM solution completely.

Definition 3.1 (Hadamard product). We define Hadamard product of two vectors $x, y \in \mathbb{R}^m$ as $x \circ y \in \mathbb{R}^m$ such that $(x \circ y)_i \triangleq x_i y_i$ for all $i \in [m]$.

Remark 7. We can write the dual problem for non-separable training data in this high dimensional space $\mathbb{H}$ as

$$\max_{\alpha} \mathbf{1}^\top \alpha - \frac{1}{2} (\alpha \circ y)^\top K (\alpha \circ y)$$
$$\text{subject to: } 0 \leqslant \alpha \leqslant C \text{ and } \alpha^\top y = 0.$$

The solution hypothesis h can be written as $h(x) = \text{sign}\left(\sum_{i=1}^m \alpha_i y_i k(x_i, x) + b\right)$, where $b = y_i - (\alpha \circ y)^\top K e_i$ for all x_i such that $0 < \alpha_i < C$.