Mathematical Theory of Neural Network Models

Lecture 4: Reproducing Kernel Hilbert Space

July 9

Lecturer: Lei Wu *Scribe: Zhonglin Xie, Yuxuan Zhang*

1 Reproducing Kernel Hilbert Space (RKHS)

Kernel ridge regression: Given a set of data $\{(x_i, y_i)\}_{i=1}^n$, our aim is to learn a function $f(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$ which can best fit the target function f^* where $y_i = f^*(x_i)$.

Then what kind of target function is suitable for kernel ridge regression?

Example 1.1. *Assume that k is a linear kernel, such as $k(x, x') = <x, x'>$. Then function f generated by kernel k can only represent linear functions.*

Example 1.2. *Assume that k is a polynomial kernel, such as $k(x, x') = (<x, x'> + 1)^n$. Then function f generated by kernel k can represent all n-degree polynomials.*

When we specify a kernel k, the functions generated by k are only suitable for a certain kind of function. These functions form a space called Reproducing Kernel Hilbert Space(RKHS).

Definition 1.3. Let $\mathcal{H}$ be a Hilbert space, $\mathcal{H}$ is of function $\mathcal{X} \to \mathbb{R}$. $\mathcal{H}$ is said to be an RKHS if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which satisfies:

- $\forall x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{H}$
- **(reproducing property)** $\forall x \in \mathcal{X}$,$f \in \mathcal{H}$, $<f, k(\cdot, x)> = f(x)$

k is called a **reproducing kernel**.

The reproducing property is corresponding to the re-definition of value-taking operator.

Definition 1.4. A bi-variate function k is said to be a **positive definite function(PD function)** if it satisfies:

- $k(x, x') = k(x', x)$
- $\sum_{1 \le i,j \le n} \alpha_i \alpha_j k(x_i, x_j) \ge 0$

Definition 1.5. A bi-variate function k is said to be a **kernel** if there exists $\phi : \mathcal{X} \to \mathcal{H}$ where $\mathcal{H}$ is a Hilbert space, s.t. $k(x, x') = <\phi(x), \phi(x')>$. ϕ is called the feature map.

We should notice that a kernel must be a PD function. But is a PD function equivalent to a kernel ?

Claim 1.6. *A reproducing kernel is a kernel.*

Proof. If k is a reproducing kernel, from the reproducing property, we get:$<k(\cdot, x'), k(\cdot, x))> = k(x', x)$.
Then $k(\cdot,\cdot)$ could be the feature map of k. k is a kernel. □

Claim 1.7. *A kernel ⇔ a reproducing kernel ⇔ a PD function.*

To make this correct, we need to claim that a PD function is a reproducing kernel, which will be proved in next section.

2 Moore-Aronszajn Theorem

Given a PD function k, could we define an RKHS from k?

In fact, if we let $\mathcal{H}_0 = \{f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x), n \in \mathbb{N}_+\}$, then $\mathcal{H}_0$ is the set of functions which can fit kernel k. We hope there is a well-defined inner product in $\mathcal{H}_0$.

Theorem 2.1. *Let $f, g \in \mathcal{H}$, $f = \sum_{i=1}^{n} \alpha_i k(x_i, x), g = \sum_{j=1}^{n} \beta_j k(\tilde{x_j}, x)$, the inner product of f and g can be defined as:*

$$<f, g> = \sum_{i,j} \alpha_i \beta_j k(x_i, \tilde{x_j}) \tag{1}$$

Proof. The most essential thing is to prove that the norm induced by this inner product is positive definite(i.e. $\|f\|_2 = 0 \Rightarrow f = 0$).Because k is a PD function, we have that:

$$\|f\|_2^2 = <f, f> = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \alpha^T K \alpha \geq 0 \tag{2}$$

We want to prove that if $\|f\|_2 = 0$, then $\forall x \in \mathcal{X}, f(x) = 0$.
Let $\tilde{f} = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) + tk(x, \cdot)$, then we have:

$$\begin{aligned}
\|\tilde{f}\|_2^2 &= \sum_{1\leq i,j \leq n} \alpha_i \alpha_j k(x_i, x_j) + 2t \sum_{i=1}^{n} \alpha_i <k(x_i, \cdot), k(x, \cdot)> + t^2 <k(x, \cdot), k(x, \cdot)> \\
&= \|f\|_2^2 + 2t \sum_{i=1}^{n} \alpha_i k(x_i, x) + t^2 k(x, x) && (reproducing\ property) \\
&\geq 2t \sum_{i=1}^{n} \alpha_i k(x_i, x) + t^2 k(x, x) && (\|f\|_2^2 \geq 0) \\
&= 2tf(x) + t^2 k(x, x) \geq 0, \forall t \in \mathbb{R} && (k(x, x) \geq 0)
\end{aligned}$$

Next, let's discuss it in two cases:

- if $k(x,x) = 0$, then obviously we have $f(x) = 0$.
- if $k(x,x) > 0$, then we have:

$$2tf(x) + t^2 k(x,x) = k(x,x)[t + \frac{f(x)}{k(x,x)}]^2 - \frac{f^2(x)}{k(x,x)} \geq 0 \tag{3}$$

  then:

$$-\frac{f^2(x)}{k(x,x)} \geq 0 \tag{4}$$

  Because $f^2(x) \geq 0$ and $k(x,x) \geq 0$, there must be $f(x) = 0$.

$\square$

So $\mathcal{H}$ is an inner product space, but it's incomplete, so it's not a Hilbert space. Let $\mathcal{H}_k = \overline{\mathcal{H}_0}$, then $\mathcal{H}_k$ is a complete Hilbert space.

Our question is that, is $\mathcal{H}_k$ an RKHS?

Claim 2.2. *$\mathcal{H}_k$ is an RKHS, k is the reproducing kernel of $\mathcal{H}_k$.*

Proof. To explain that $\mathcal{H}_k$ is an RKHS, we should prove that H_k satisfies the two conditions in 1.3:

- Because $\mathcal{H}_0 = \{f = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot), n \in \mathbb{N}_+\}$, then obviously $k(\cdot, x) \in \mathcal{H}$.
- We have that:

$$\begin{aligned} <f, k(\cdot,x)> &= \lim_{n\to+\infty} <\sum_{i=1}^{n} \alpha_i k(\cdot, x_i), k(\cdot, x)> & (5)\\ &= \lim_{n\to+\infty} \sum_{i=1}^{n} \alpha_i <k(\cdot, x_i), k(\cdot, x)> & (6)\\ &= \lim_{n\to+\infty} \sum_{i=1}^{n} \alpha_i k(x_i, x) & (7)\\ &= f(x) & (8)\end{aligned}$$

$\square$

To prove that equation (7) to (8) holds, we should notice that the **norm** operator is a bounded continuous operator on the definition of Hilbert space, and use the property:

$$\lim_{n\to+\infty} \|f(x) - \sum_{i=1}^{n} \alpha_i k(x_i, x)\| = 0 \tag{9}$$

You can refer to the reading materials for more infomation.

So we also prove that a PD function is a reproducing kernel. That is to say we prove 1.7.

A natural question is that, is there any other RKHS whose reproducing kernel is also k?

Theorem 2.3. *Given a PD funcion k, there exists an only RKHS $\mathcal{H}_k$ whose reproducing kernel is k.*

Proof. Existence has been proved before. So here we only prove the uniqueness.
If there exists $\mathcal{H}'_k$ and k is the reproducing kernel of $\mathcal{H}'_k$, then we have $\mathcal{H}_0 \subseteq \mathcal{H}'_k$.
And we already have $\mathcal{H}_k = \overline{\mathcal{H}_0}$, so $\overline{\mathcal{H}_0} = \mathcal{H}_k = \mathcal{H}'_k$. □

This method is graceful but too abstract. How to intuitively understand the definition of norm in RKHS?

3 Merce representation of RKHS

To simplify the problem, we add two constraints:

- $\mathcal{X}$ is compact.
- $k(x, x')$ is continuous.

So we have $\sup_{x,x'} k(x, x') < \infty$.

Definition 3.1. Let T_k be an integral operator:

$$\begin{gathered} T_k : L_2(\mathcal{X}, \mu) \to L2(\mathcal{X}, \mu) \\ T_k \circ f(x) = \int k(x, x')f(x')d\mu(x') \end{gathered} \tag{10}$$

where $L_2(\mathcal{X}, \mu) = \{f | \int f^2(x)d\mu(x) < \infty\}$, and $\mu(x)$ is a probability measure.

We can see that T_k is a compact operator and L_2 is a Hilbert space. And we have the spectral decomposition of T_k:

$$T_k \circ f = \sum_{j=1}^{\infty} \lambda_j {<} f, e_j {>} e_j \tag{11}$$

where $\{e_j\}$ is a set of orthonormal basis of $L_2(\mathcal{X}, \mu)$.

To better understand this, we can regard this spectral decomposition as eigenvalue decomposition of an infinite-dimension matrix, where T_k is an infinite-dimension matrix and f is an infinite-dimension vector.

So for every kernel k, we also have a decomposition:

$$k(x, x') = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(x') \tag{12}$$

From this decomposition, we can get the feature map:

$$\phi : x \to \begin{pmatrix} \sqrt{\lambda_1} e_1(x) \\ \sqrt{\lambda_2} e_2(x) \\ \dots \\ \sqrt{\lambda_n} e_n(x) \\ \dots \end{pmatrix} \in l^2 \tag{13}$$

Thus we have that:

$$k(x, x') = <\phi(x), \phi(x')> \tag{14}$$

Note that given a kernel k, the feature map of k is not unique. To better understand this, you can regard ϕ as a set of orthonormal basis, and the basis of a space is not unique.

The method in section 2 is to take $\{k(\cdot, x_i)\}$ as basis. Moreover, we can define a space $\mathcal{H}$ as follows:

Definition 3.2. Let kernel $k(x, x') = \sum_{j=1}^{\infty} \lambda_j e_j(x) e_j(x')$, we have:

$$\mathcal{H} = \{\sum_{j=1}^{\infty} \alpha_j e_j | \sum_{j=1}^{\infty} \frac{\alpha_j^2}{\lambda_j} < \infty\} \tag{15}$$

And

$$\begin{aligned} f = \sum_{j=1}^{\infty} \alpha_j e_j, g = \sum_{j=1}^{\infty} \beta_j e_j \\ <f, g> = \sum_{j=1}^{\infty} \frac{\alpha_j \beta_j}{\lambda_j} \end{aligned} \tag{16}$$

Claim 3.3. $\mathcal{H} = \mathcal{H}_k$

Proof. To explain that $\mathcal{H} = \mathcal{H}_k$, we should prove that k is the reproducing kernel of $\mathcal{H}$, using the uniqueness of RKHS, we obtian the result:

- Because $k(\cdot, x) = \sum_j \lambda_j e_j(x) e_j(\cdot)$

  $$\|k(\cdot, x)\|^2 = \sum_j \frac{(\lambda_j e_j(x))^2}{\lambda_j} = \sum_j \lambda_j e_j(x) e_j(x) = k(x, x) < \infty,$$

  then $k(\cdot, x) \in \mathcal{H}$.

- Suppose $f = \sum_j a_j e_j$, We have

  $$\begin{aligned} <f, k(\cdot, x)> &= \sum_j \frac{a_j \lambda_j e_j(x)}{\lambda_j} \\ &= \sum_j a_j e_j(x) \\ &= f(x). \end{aligned}$$

$\square$

The advantage of 3.2 is its intuitive. $\|f\|_{\mathcal{H}_k}^2 = \sum_j \frac{a_j^2}{\lambda_j}$ is similar to weighted L_2 norm. The role of introducing a kernel is to weight the basis, different e_j have different weight λ_j. The bigger the λ_j, the more important the e_j.

Theorem 3.4. *Consider the abstract KRR problem*

$$\hat{h}_n = \arg\min_{h \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} l(h(x_i), y_i) + \lambda \|h\|_{\mathcal{H}_k}^2,$$

This infinite dimensional optimization can be reduced to a finite dimension optimization, the solution is

$$\hat{h}_n(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x).$$

Proof. Review the problem in Lecture 1,

$$\hat{\beta}_n = \arg\min_{\beta \in \mathbb{R}^m} \|\Phi\beta - y\| + \lambda \|\beta\|^2 \tag{17}$$

We have

$$\hat{\beta}_n = \sum_{i=1}^{n} \alpha_i \phi(x_i),$$

Theorem 3.5 is similar to 17, we omit the details. Consider another problem which can't be solved in this method. $\square$

Theorem 3.5. *(Representer Theorem) Consider the abstract KRR problem*

$$\hat{h}_n = \arg\min_{h \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} l(h(x_i), y_i) + \lambda G(\|h\|_{\mathcal{H}_k}),$$

This infinite dimensional optimization can be reduced to a finite dimension optimization. The solution can be written as

$$\hat{h}_n(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

. Where $G(\cdot)$ is an increasing function, α_i relay on G.

Proof. Consider decomposition

$$h = h^{\perp} + h^{//},$$

Where $h^{//} \in span(\{k(x_i, x)\}_{i=1}^n)$,$\left\langle h^{\perp}, k(x_i, x) \right\rangle = 0$, then

$$h(x_i) = \langle h, k(\cdot, x_i) \rangle = \left\langle h^{\perp} + h^{//}, k(\cdot, x_i) \right\rangle = \left\langle h^{//}, k(\cdot, x_i) \right\rangle = h^{//}(x_i)$$

then optimal solution belong to$span(\{k(x_i, x)\}_{i=1}^n)$. $\square$

4 Error analysis of kernel methods

In Random Feature Model, we use

$$f(x;a) = \frac{1}{m}\sum_{i=1}^{m} a_i\phi(x;b_i^0) \quad b_i^0 \sim \pi(\cdot) \tag{18}$$

to approch the target function, where ϕ is randomly generated, the variable to learn is a. Consider the Empirical Risk

$$J_\lambda(a) = L_n(a) + \lambda\|a\|^2$$
$$\hat{a}_{n,\lambda} = \arg\min J_\lambda(a)$$

Then we have the priori estimate of population risk

$$L(\hat{(a)}_n, \lambda) \lesssim \frac{\|f^*\|_1}{n^\alpha} + \frac{\|f^*\|_2}{m^\beta}$$

m is the number of features, when $m \to \infty$, $f(x;a) = \int a(\omega)\phi(x;\omega)\,\mathrm{d}\pi(\omega)$, we can use 18 to approching $f(x;a)$.

Definition 4.1.

$$\mathcal{H} = \{f = \int a(\omega)\phi(x;\omega)\,\mathrm{d}\pi(\omega) | \int a^2(\omega)\,\mathrm{d}\pi(\omega) < \infty\}$$
$$\langle f, g\rangle = \int a_f(\omega)a_g(\omega)\,\mathrm{d}\pi(\omega)$$

Then $\mathcal{H}$ is a RKHS, and $k(x,x') = \int \phi(x,\omega)\phi(x',\omega)\,\mathrm{d}\pi(\omega)$ is the reproducing kernel of $\mathcal{H}$, where ϕ is bounded.

Proof. We verify that $\mathcal{H}$ satisfies 1.3.

- $k(x,\cdot) = \int \phi(x,\omega)\phi(x,\omega)\,\mathrm{d}\pi(\omega)$, then

$$\int \phi^2(x,\omega)\,\mathrm{d}\pi(\omega) < \infty.$$

- $\langle f, k(x,\cdot)\rangle = \int a(x,\omega)\phi(x,\omega)\,\mathrm{d}\pi(\omega) = f(x)$.

Thus $\mathcal{H} = \mathcal{H}_k$. □

We know

$$generalization\,error = approximation\,error + estimate\,error,$$

and approximation property: $\forall f \in \mathcal{H}_k, \delta > 0$, with probability $1-\delta$ over the sampling of feature there exists $a \in \mathbb{R}^m, s.t.$

$$\mathbb{E}_x(\frac{1}{m}\sum_{k=1}^{m} a_k\phi(x;\omega_k^0) - f(x))^2 \lesssim \frac{\|f^*\|_{\mathcal{H}_k}}{m} + \frac{\sqrt{\log(1/\delta)}}{m}.$$

Claim 4.2. *If we calculate the expectations of ω_k^0, we have*

$$\mathbb{E}_{\omega_k^0}\mathbb{E}_x(\frac{1}{m}\sum_{k=1}^{m} a_k\phi(x;\omega_k^0) - f(x))^2 \lesssim \frac{\|f^*\|_{\mathcal{H}_k}}{m}.$$

Proof. We remeber $\omega^0 = \{\omega_k^0\}_{k=1}^m$, and

$$Z(\omega^0) = \sqrt{\mathbb{E}_x(\frac{1}{m}\sum_{k=1}^{m} a_k\phi(x;\omega_k^0) - f(x))^2}.$$

Then

$$\mathbb{E}_{\omega^0}(Z^2(\omega^0)) = \mathbb{E}_{\omega^0}\mathbb{E}_x(\frac{1}{m}\sum_{k=1}^{m} a_k\phi(x;\omega_k^0) - f(x))^2,$$

We know $f(x) = \mathbb{E}_{\omega_k^0}(a(\omega_k^0)\phi(x;\omega_k^0))$, we remember $\triangle_k^0(x) = a_k\phi(x;\omega_k^0) - f(x)$, then

$$\mathbb{E}_{\omega^0}(Z^2(\omega^0)) = \mathbb{E}_{\omega^0}\mathbb{E}_x(\frac{1}{m}\sum_{k=1}^{m}\triangle_k^0)^2 = \mathbb{E}_x\mathbb{E}_{\omega^0}(\frac{1}{m}\sum_{k=1}^{m}\triangle_k^0)^2,$$

since $\mathbb{E}_{\omega^0}(\triangle_k^0) = 0$, we write

$$\mathbb{E}_x\mathbb{E}_{\omega^0}(\frac{1}{m}\sum_{k=1}^{m}\triangle_k^0)^2 = \mathbb{E}_x\mathbb{E}_{\omega^0}(\frac{1}{m^2}\sum_{i,j=1}^{m}\triangle_i^0\triangle_j^0) = \frac{1}{m}\mathbb{E}_x\mathbb{E}_{\omega_k^0}(\triangle_k^0)^2,$$

we have

$$\mathbb{E}_{\omega_k^0}(\triangle_k^0)^2 = a_k^2(\omega_k^0)\phi^2(x;\omega_k^0) - f^2(x),$$

then

$$\frac{1}{m}\mathbb{E}_x\mathbb{E}_{\omega_k^0}(\triangle_k^0)^2 \le \frac{1}{m}\mathbb{E}_x\mathbb{E}_{\omega_k^0}(a_k^2\phi^2(x;\omega_k^0)),$$

sppouse $|\phi| < B$, we get

$$\frac{1}{m}\mathbb{E}_x\mathbb{E}_{\omega_k^0}(a_k^2\phi^2(x;\omega_k^0)) \lesssim \frac{1}{m}\mathbb{E}_{\omega_k^0}a_k^2(\omega_k^0) = \frac{\|f^*\|_{\mathcal{H}_k}^2}{m}.$$

It's easy to verify $Z(\omega^0)$ is a continous function of ω^0 when a is bounded. By McDiamid's inequality, with probability $\ge 1-\delta$

$$Z(\omega^0) \le \mathbb{E}_{\omega^0}(Z(\omega^0)) + \sqrt{\frac{\log(1/\delta)}{m}} \tag{19}$$

$$\le \sqrt{\mathbb{E}_{\omega^0}Z^2(\omega^0)} + \sqrt{\frac{\log(1/\delta)}{m}} \tag{20}$$

$$= \frac{\|f\|_{\mathcal{H}_k}}{\sqrt{m}} + \sqrt{\frac{\log(1/\delta)}{m}}. \tag{21}$$

□

Claim 4.3. *Consider estimation error, with probability* $\geq 1-\delta$

$$|L(\hat{a}_n) - L_n(\hat{a}_n)| \leq \sup_{\|a\|\leq C} |L(a) - L_n(a)| \leq 2Rad(\mathcal{F}_C) + \sqrt{\frac{\log(1/\delta)}{n}}, \tag{22}$$

where $\mathcal{F}_C = \{\frac{1}{m}\sum_{k=1}^m a_k \phi(\cdot;\omega_k^0) | \|a\| \leq C\}$, *and*

$$Rad(\mathcal{F}_C) \lesssim \frac{1}{\sqrt{n}}. \tag{23}$$

Proof. The proof of 22 is established in previous lecture, we will focus on 23, suppose $|\phi| \leq B$

$$\begin{aligned}
nRad(\mathcal{F}_C) &= \mathbb{E}_\xi \sup_{\|a\|\leq C} \sum_{i=1}^n \xi_i \sum_{k=1}^m a_k \phi(x;\omega_k^0) && (24)\\
&= \mathbb{E}_\xi \sup_{\|a\|\leq C} \left\langle a, \sum_{i=1}^n \xi_i \phi(x;\omega_k^0) \right\rangle && (25)\\
&\leq \mathbb{E}_\xi \sqrt{\sum_{k=1}^m a_k^2} \sqrt{(\sum_{i=1}^n \xi_i \phi(x;\omega_k^0))^2} && (26)\\
&\leq C \sqrt{\mathbb{E}_\xi (\sqrt{(\sum_{i=1}^n \xi_i \phi(x;\omega_k^0))^2}^{\,2}\ )} && (27)\\
&= C \sqrt{\mathbb{E}_\xi (\sum_{i=1}^n \xi_i \phi(x;\omega_k^0))^2} && (28)\\
&\leq C\sqrt{nB^2 \mathbb{E}_{\xi_i} \xi_i^2} && (29)\\
&= CB\sqrt{n}. && (30)
\end{aligned}$$

□

References