Lecture 8: Representer theorems in RKHSs

In this lecture we will re-visit representer theorems which we proved earlier in the general Hilbert space setting.

Recall: (Rep. Thm)

Suppose $(\mathcal{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$ is a Hilbert space & $R: \mathbb{R} \to \mathbb{R}$ is non-decreasing. Then if

$$J(h) = L(\langle h_1, h\rangle, \ldots, \langle h_n, h\rangle) + R(\|h\|)$$

has a minizer then there exists at least one minizer $h^* \in \operatorname{span}\{h_1, \ldots, h_n\}$.

8.1 Representer theorems & formulae in RKHSs

Since RKHSs satisfy the reproducing property

$$f \in \mathcal{H} \qquad f(x) = \langle f, K(x, \cdot)\rangle$$

Then we can immediately obtain a very useful RKHS version of the rep. thm.

Thm Let $\mathcal{H}$ be an RKHS with kernel $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ & consider $J: \mathcal{H} \to \mathbb{R}$

$$J(h) = L(h(x_1), \ldots, h(x_n)) + R(\|h\|)$$

where $R: \mathbb{R} \to \mathbb{R}$ is non-decreasing. If J has a minimizer then there exists at least one minimizer h^* s.t.

$$h^* = \sum_{i=1}^{n} \alpha_i^* K(x_i, \cdot)$$

Since we now have an ansatz for h^* we can now compute h^* in two ways:

first observe that

$$\|h^*\|^2 = \langle h^*, h^* \rangle = \left\langle \sum_i \alpha_i^* K(x_i, \cdot), \sum_j \alpha_j^* K(x_j, \cdot) \right\rangle$$

$$= \sum_{ij} \alpha_i^* \alpha_j^* K(x_i, x_j) = (\underline{\alpha}^*)^T K(X, X) \underline{\alpha}^*$$

Furthermore

$$h^*(x) = \sum_{j=1}^{n} \alpha_j^* K(x_j, x) = K(x, X)\,\underline{\alpha}^*$$

where we introduced the vector field

Row vector

$$K(x, X) = \big(K(x_1, x), \cdots, K(x_n, x)\big) \in H^{\otimes n}$$

we then obtain the following Corollary

Corollary 1: Every vector $\underline{\alpha}^* \in \mathbb{R}^n$ that is a minimizer of

(*)
$$J(\underline{\alpha}) = L\big(K(x_1, X)\underline{\alpha}, \dots, K(x_n, X)\underline{\alpha}\big) + R\big(\underline{\alpha}^T K(X, X)\underline{\alpha}\big)$$

is associated with a minimizer h^* of $J(h)$ of the form $h^* = K(x, X)\underline{\alpha}^*$.

Observe that (*) can be easily implemented numerically & often solved using off-the shelf optimization algorithms.

In the setting when $K(X,X)$ is invertible we can characterize h^* in yet another way which proves to be quite useful in practice. Define

$$x_i \in X, \qquad z_i^* := h^*(x_i) = K(x_i, X)\, \underline{\alpha}^*$$

Then we have for $\underline{z}^* := (z_1, \ldots, z_n)^T$

$$\underline{z}^* = K(X,X)\, \underline{\alpha}^*$$

$$\Rightarrow \underline{\alpha}^* = K(X,X)^{-1} \underline{z}^*$$

This implies immediately that

$$\|h^*\|^2 = (\underline{\alpha}^*)^T K(X,X)\, \underline{\alpha}^*$$

$$= (\underline{z}^*)^T K(X,X)^{-1} \underline{z}^*$$

& so by substitution in (*) we obtain the Corollary:

Corollary 2:

Every vector $z^* \in \mathbb{R}^n$ that is a minizer of

$$J(\underline{z}) = L(\underline{z}) + R(\underline{z}^T K(X,X)\underline{z})$$

is associated with a minimizer h^* of $J(h)$ given by the formula

$$(**) \quad h^*(x) = K(x, X) K(X,X)^{-1} \underline{z}^*$$

The above Corollary is what is often referred to as "the rep. theorem" in ML literature on kernel methods & (**) is often called the representer formula.

8.2 Application to supervised learning.

Consider input & output spaces $\mathcal{X}$ & $\mathcal{Y}$ respectively. The goal of supervised

learning is to approximate/learn a function $f^t: \mathcal{X} \to \mathcal{Y}$ given a "training data set" of the form $\{(x_i, y_i)\}_{i=1}^n$.
For simplicity, let us assume $\mathcal{Y} \equiv \mathbb{R}$ so that the $\underline{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ & suppose X is a Banach space. We will derive an approx to f^t within an RKHS.

(Step 1) Pick a kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
a simple choice is the RBF kernel

$$K(x, x') = \exp\left(-\gamma \|x - x'\|_{\mathcal{X}}^2\right)$$

(Step 2) Pick regularization term

$$R(\|f\|) = \frac{\lambda}{2}\|f\|^2, \quad \lambda > 0$$

RKHS norm!

(Step 3) Pick loss (mean squared error)

$$L(f) = \frac{1}{2N}\sum_{j=1}^{n} |f(x_i) - y_i|^2$$

(Step 3) Formulate opt. problem

$$\underset{f \in H_k}{\text{minimize}} \; J(f) := L(f) + \frac{\lambda}{2}\|f\|^2$$

(Step 3) Apply rep. thm.

$$h^* = K(\cdot, X) K(X,X)^{-1} \underline{z}^*$$

where

$$\underline{z}^* = \underset{\underline{z} \in \mathbb{R}^n}{\text{argmin}} \; \frac{1}{2N}\|\underline{z} - \underline{y}\|^2 + \frac{\lambda}{2} \underline{z}^T K(X,X)^{-1} \underline{z}$$

Since the functional is quadratic solve using first order opt. cond.

$$\underline{z}^* - \underline{y} + N\lambda K(X,X)^{-1} \underline{z}^* = 0$$

$$\Longrightarrow \underline{z}^* = \left(K(X,X) + N\lambda I\right)^{-1} K(X,X) \underline{y}$$

sub. back into rep. formula

$$h^*(x) = K(x,X) \, K(X,X)^{-1} \left(K(X,X) + N\lambda I\right)^{-1} K(X,X) \, \underline{y}$$

In practice, $K(X,X)$ can become very ill-posed. So, we regularize it using a nugget term:

$$K(X,X)^{-1} \longrightarrow \left(K(X,X) + \sigma^2 I\right)^{-1}$$

where $\sigma^2 > 0$ is a small parameter.