Reproducing kernels

Recap:

Def$^{\underline{n}}$ We say $k: \Omega \times \Omega \to \mathbb{R}$ is a reproducing kernel for a Hilbert space $\mathcal{H}$ (of functions $f: \Omega \to \mathbb{R}$) if

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}} \qquad \text{for all } f \in \mathcal{H}.$$

"Therefore, $k(\cdot, x)$ is a "delta" function for $\mathcal{H}$."

Def$^{\underline{n}}$: If a Hilbert space $\mathcal{H}$, has a reproducing kernel, then $\mathcal{H}$ is called a RKHS.

(Every $\mathcal{H}$ such that $L_x : f \to f(x)$ is such that $|L_x(f)| \le M \|f\|_{\mathcal{H}}$ is a RKHS. Here, $M < \infty$.)

RKHS induced by a pos. def. matrix

Given a square complex matrix $P = (p_{ij})_{i,j=1}^{n}$. If we let $X = \{1, 2, \ldots, n\}$ then $k: X \times X \to \mathbb{C}$ can be defined as $k(i,j) = p_{ij}$.

We can ~~define~~ identify $f: X \to \mathbb{C}$ with $(f(1), \ldots, f(n))$.

$$\text{Let} \quad \mathcal{H}_0 = \text{Span}\{ k(\cdot, x) : x \in X \}$$

$$= \text{Span}\{ P(\cdot, j) : j = 1, \ldots, n\}$$

$$= \text{Column space of } P.$$

Now, suppose that P is pos. def. so K is a pos. def. kernel.

If k is reproducing:

$$k_x(y) = k(y,x) = \langle k(\cdot,x), k(\cdot,y)\rangle_H$$
$$\Vert$$
$$p_j(i) = P(i,j) = \langle Pe_j, e_i\rangle = \langle P^{\frac{1}{2}}e_j, P^{\frac{1}{2}}e_i\rangle$$

p_j = jth col of P

Thus, $\langle f, g\rangle_H = \cancel{\langle \alpha, \beta\rangle}$ ~~where~~

$$= \langle \sum_{i=1}^{n} \alpha_i k(\cdot,i), \sum_{j=1}^{n} \beta_j k(\cdot,j)\rangle$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \bar{\beta}_j \langle k(\cdot,i), k(\cdot,j)\rangle$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \bar{\beta}_j\, e_j^* P^{\frac{1}{2}}P^{\frac{1}{2}} e_i$$

$$= \underline{\beta}^* P \underline{\alpha}$$

Examples of pos. def. kernels

Linear kernels $\quad K(\underline{x},\underline{y}) = \underline{x}^T\underline{y}$.

poly. kernels $\quad K(\underline{x},\underline{y}) = (\underline{x}^T\underline{y} + r)^n \quad r \geq 0$.

Gaussian $\quad K(\underline{x},\underline{y}) = e^{-\frac{\|\underline{x}-\underline{y}\|^2}{2\sigma^2}} \quad \sigma > 0$

Laplacian $\quad K(\underline{x},\underline{y}) = e^{-\alpha(\|\underline{x}-\underline{y}\|)} \quad \alpha > 0$.

Sobolev space kernel: Let k, d be integers, $k > \frac{d}{2}$

Matérn kernels: $K(\underline{x}, \underline{y}) = \|\underline{x} - \underline{y}\|_2^{k-\frac{d}{2}} \, \mathcal{K}_{d/2-k}(\|x-y\|_2)$

This reproducing kernel for $W_2^k(\mathbb{R}^d)$. That is, functions on $\mathbb{R}^d$ that are k-times weakly differentiable with

$$\|f\|_{W_2^k}^2 = \sum_{j=0}^{k} \|\underbrace{D^j}_{\text{diff op.}} f\|_{L^2}^2 \,.$$

The representer theorem:

- Given "training examples": $x_1, \dots, x_n \in \Omega$
- Given "training values": $y_1, \dots, y_n \in \mathbb{R}$
- Given arbitrary error function: $E : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$.

$$E = E\big((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))\big)$$

measures how well ~~you did on~~ f does on training set.

- Strictly monotonically increasing real-valued func.

$$g : [0, \infty) \to \mathbb{R}.$$

Then any minimizer of

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \; E\big((x_1, y_1, f(x_1)), \dots, (x_n, y_n, f(x_n))\big) + g(\|f\|_{\mathcal{H}})$$

$\mathcal{H}$: RKHS with kernel $K : \Omega \times \Omega \to \mathbb{R}$

has the representation:

$$f^*(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) \qquad \text{where } \alpha_i \in \mathbb{R},\ 1 \le i \le n.$$

<u>Proof</u>: Let $f \in \mathcal{H}$. Then,

$$f^*(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + v$$

where $\langle v, K(\cdot, x_i) \rangle_H = 0$, for $1 \le i \le n$.

Therefore,

$$f(x_j) = \Big\langle \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) + v,\ K(\cdot, x_j) \Big\rangle_H \qquad \text{reproducing property}$$

$$= \sum_{i=1}^{n} \alpha_i \langle K(\cdot, x_i), K(\cdot, x_j) \rangle_H$$

Thus, E in minimization is independent of v.

$$g(\|f\|) = g\Big(\Big\| \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) + v \Big\| \Big)$$

$$= g\Big(\Big(\Big\| \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) \Big\|^2 + \|v\|^2 \Big)^{\frac{1}{2}} \Big)$$

$$\ge g\Big(\Big\| \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) \Big\| \Big)$$

Consequently, best to pick $v = 0$. That is

$$f^*(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i).$$

Suppose

$$E = \frac{1}{2}\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \quad \text{and} \quad g(\|f\|_{\mathcal{H}}) = \lambda \|f\|_{\mathcal{H}}^2$$

for $\lambda > 0$.

Then,

$$f^* = \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 + g(\|f\|_{\mathcal{H}})$$

Equivalent to:

$$\underline{\alpha}^* = \underset{\alpha_1, \ldots, \alpha_n \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \alpha_j K(x_i, x_j)\right)^2 + \lambda \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i K(x_i, x_j) \alpha_j .$$

$$= \underset{\underline{\alpha} \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2} \| \underline{y} - A \underline{\alpha} \|_2^2 + \lambda \underline{\alpha}^T A \underline{\alpha}$$

$$= \underset{\underline{\alpha} \in \mathbb{R}^n}{\operatorname{argmin}} \left[\frac{1}{2} \underline{y}^T \underline{y} - \underline{y}^T A \underline{\alpha} + \frac{1}{2} \underline{\alpha}^T A^T A \underline{\alpha} + \lambda \underline{\alpha}^T A \underline{\alpha} \right]$$

$$= \underset{\underline{\alpha} \in \mathbb{R}^n}{\operatorname{argmin}} \left[\underline{\alpha}^T \left(\frac{1}{2} A^T A + \lambda A \right) \underline{\alpha} - \underline{y}^T A \underline{\alpha} \right]$$

Therefore,

$$\boxed{\left(A^T A + \lambda A \right) \underline{\alpha}^* = A \underline{y}}$$

An infinite-dimensional problem can be solved via a linear system!