Lecture 7 - RKHS

Tuesday, February 7, 2023 2:26 PM

today: RKHS

RKHS: (reproducing kernel Hilbert space)

$\Phi: \mathcal{X} \to \mathcal{H}$ s.t. $\langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = k(x, x')$ (important property of the RKHS)

$\mathcal{H}$ = RKHS

$\mathcal{H}$ is a space of functions from $\mathcal{X} \to \mathbb{R}$

Let $\tilde{\mathcal{H}} = \text{span}\{k(x,\cdot) : x \in \mathcal{X}\}$ $\qquad f_x \triangleq k(x,\cdot)$

$f_x(x') = k(x,x')$

e.g. $f \in \tilde{\mathcal{H}} \Rightarrow f(\cdot) = \sum_i \alpha_i^f k(x_i^f, \cdot)$ for some finite $\{x_i\}_{i=1}^n$, $\alpha_i \in \mathbb{R}$

α_i^f means α_i in expansion for f: $f = \sum_i \alpha_i k(x_i, \cdot)$

$\boxed{\text{"pre-Hilbert" space [inner product space] with } \langle f, g \rangle_{\tilde{\mathcal{H}}} \triangleq \sum_{i,j} \alpha_i^f \alpha_j^g \underbrace{k(x_i^f, x_j^g)}_{\langle k(x_i^f,\cdot),\, k(x_j^g,\cdot) \rangle_{\tilde{\mathcal{H}}}}}$

$\|f\|_{\tilde{\mathcal{H}}} \triangleq \sqrt{\langle f, f \rangle_{\tilde{\mathcal{H}}}}$

then RKHS $\mathcal{H} \triangleq$ completion($\tilde{\mathcal{H}}$) using $\|\cdot\|_{\tilde{\mathcal{H}}}$ as your norm

i.e. add all limit points of $\tilde{\mathcal{H}}$-Cauchy sequences to get $\mathcal{H}$

you could think $f = \sum_{i=1}^{\infty} \alpha_i k(x_i, \cdot)$

$f_n = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$

$\boxed{(*) \text{ "reproducing" property of } \mathcal{H}: \text{ for } f \in \mathcal{H} \quad \langle f, k(x,\cdot) \rangle_{\mathcal{H}} = f(x)}$

$k(x,\cdot) = \Phi(x)$ $\qquad$ (i.e. $\langle k(x,\cdot), k(x',\cdot) \rangle_{\mathcal{H}} = k(x,x')$)

$\langle k(x,\cdot), k(x',\cdot) \rangle_{\mathcal{H}} = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$

nice property of RKHS, fct. evaluation is a cts. operation in $\mathcal{H}$

mapping $E_x: \mathcal{H} \to \mathbb{R}$

$E_x f = f(x)$

$|f(x) - g(x)| = |\langle f - g, k(x,\cdot) \rangle_{\mathcal{H}}|$

$\leq \|f-g\|_{\mathcal{H}} \|k(x,\cdot)\|_{\mathcal{H}}$ i.e. E_x is Lipschitz cts. with $L = \|k(x,\cdot)\|_{\mathcal{H}}$

(*) this property is important to do statistics $\hat{f}_n \to f^*$

$\hat{f}_n \in \mathcal{H} \Rightarrow \|\hat{f}_n\|_{\mathcal{H}} < \infty$

can have: $f^* \notin \mathcal{H} \Rightarrow$ "$\|f\|_{\mathcal{H}} = \infty$"

but $f^* \in L_2 \Rightarrow \|f\|_{L_2} < \infty$

$\hat{f}_n \not\to$ anything in $\mathcal{H}$ $\|\cdot\|_{\mathcal{H}}$

but can still have $\hat{f}_n \xrightarrow{L_2} f^*$

representer's thm: says that (for $\mathcal{H}$ a RKHS)

$$\min_{f \in \mathcal{H}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2$$

is reached for $f^* = \sum_{i=1}^n \alpha_i^* k(x_i, \cdot)$

for some $\alpha_i^* \in \mathbb{R}$

$(x_i, y_i)_{i=1}^n$ training set

let $f_\alpha = \sum_{i=1}^n \alpha_i k(x_i,\cdot)$ $\alpha \in \mathbb{R}^n$

then $\|f_\alpha\|_{\mathcal{H}}^2 = \langle f_\alpha, f_\alpha \rangle_{\mathcal{H}} = \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) = \alpha^T K \alpha$

Gram matrix: $(K)_{i,j} \triangleq k(x_i, x_j)$ $n \times n$ matrix

$\hookrightarrow$ inner product of $\varphi(x_i)$ on data $K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$

$$\boxed{\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n \ell(y_i, \underbrace{\sum_j \alpha_j k(x_j, x_i)}_{f_\alpha(x_i)}) + \lambda \alpha^T K \alpha}$$

finite dim opt. (thanks to representer's thm ?)

getting a handle on $\mathcal{H}$: generalize diagonalization of matrices to ∞-dim

I) start with finite matrices

say $\mathcal{X}$ is finite e.g. $x_1, \dots, x_n$

$|\mathcal{X}| = n$

$f: \mathcal{X} \to \mathbb{R}$

$\hookrightarrow$ is finite $\Rightarrow$ just a vector $\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \in \mathbb{R}^n$ ("L_2 view")

$\mathcal{H} = \text{span}\{k(x_i,\cdot) : i=1,\dots,n\} = \{K\alpha : \alpha \in \mathbb{R}^n\} \subseteq \mathbb{R}^n$ $\boxed{\text{"}L_2\text{-view"}}$

let K be Gram matrix $(K)_{ij} \triangleq k(x_i, x_j)$
$n \times n$

if k is valid kernel $\Rightarrow K \succeq 0$

$\hookrightarrow$ spectral thm. $\Rightarrow$ $\underset{n\times n}{K} = U \Lambda U^T$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_1 \geq \lambda_2 \geq \dots \lambda_n \geq 0$

and U is orthonormal basis of $\mathbb{R}^n$

i.e. $U = \begin{pmatrix} | & & | \\ \psi_1 & \dots & \psi_n \\ | & & | \end{pmatrix}$

$U^T U = U U^T = I_n$ $\quad \langle \psi_i, \psi_j \rangle_{\ell_2} = \delta_{ij}$

we can let $\Phi = \Lambda^{1/2} U^T$

$\Rightarrow \Phi^T \Phi = K$

$$\Phi = \begin{pmatrix} - \sqrt{\lambda_1}\psi_1^T - \\ \vdots \\ - \sqrt{\lambda_d}\psi_d^T - \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$d = \text{rank}(K) \leq n$

$$\Phi = \begin{pmatrix} \Phi(x_1) & \dots & \Phi(x_n) \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$\Rightarrow$ if we define

$$\boxed{\Phi(x) = \begin{pmatrix} \sqrt{\lambda_1}\psi_1(x) \\ \vdots \\ \sqrt{\lambda_d}\psi_d(x) \end{pmatrix} \in \mathbb{R}^d}$$

x-coord. of ψ_1 vector

"feature space pt. of view"

$\langle \Phi(x_i), \Phi(x_j) \rangle_{\mathbb{R}^d} = k(x_i, x_j)$

$= \sum_{l=1}^{d} \lambda_l \psi_l(x_i) \psi_l(x_j)$

<u>back to ℓ_2 view:</u> $\mathcal{H} \subseteq \ell_2$ $(\subseteq \mathbb{R}^n)$

$v \in \mathcal{H} \Rightarrow v = K\alpha$ for some $\alpha \in \mathbb{R}^n$

to get $\|v\|_{\mathcal{H}}$, we compute $\alpha_v = K^+ v$ (pseudo-inverse)

so $\|v\|_{\mathcal{H}}^2 = \alpha_v^T K \alpha_v = v^T K^+ K K^+ v$

$= v^T U \Lambda^+ \underbrace{U^T U}_{I_n} \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} U^T v$

$= (U^T v)^T \Lambda^+ (U^T v)$

$K = U \Lambda U^T$

$K^+ = U \Lambda^+ U^T$

$K K^+ = U \Lambda \underbrace{U^T U}_{I_n} \Lambda^+ U^T$

$= U \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} U^T$

$\hookrightarrow U^T v$ is projection of v onto $\{\psi_j\}_{j=1}^n$ basis

i.e. $v = \sum_{j=1}^{n} \beta_j \psi_j$ i.e. $\beta_j = \langle v, \psi_j \rangle_{\ell_2}$
$(U^T v)_j$

$$\boxed{\|v\|_{\mathcal{H}}^2 = \sum_{j=1}^{d} \frac{\overbrace{\langle v, \psi_j \rangle_{\ell_2}}^{\beta_j \text{ representation}}{}^2}{\lambda_j}}$$

vs. $\|v\|_{\ell_2} = \sum_{j=1}^{n} \langle v, \psi_j \rangle_{\ell_2}^2$

<u>and</u> $\boxed{\|\psi_j\|_{\mathcal{H}}^2 = \frac{1}{\lambda_j} \text{ for } j \leq d}$

"$\|\psi_j\|_{\mathcal{H}} = \infty$" for $j > d$

$\langle v, v \rangle_{\ell_2} = \sum_{j=1}^{n} \beta_j^2$

$\langle v, v \rangle_{\mathcal{H}} = \sum_{j=1}^{d} \beta_j^2$

for $j > d$

$$\boxed{\text{So orthonormal basis of } \mathcal{H} \text{ (with } \|\cdot\|_{\mathcal{H}}\text{) in } \mathcal{L}_2 \text{ ie. } \{\sqrt{\lambda_i}\,\psi_i\}_{i=1}^{d}}$$

$$\langle v, v\rangle_{\mathcal{H}} = \sum_{i=1}^{d} \frac{\beta_i^2}{\lambda_i}$$

thus $\|v\|_{\mathcal{H}} \leq 1$

↳ thus makes ellipsoid in $\mathcal{L}_2$

ie. higher coordinates are shrunk more ? (since λ_i's is smaller as $i \nearrow$)

15h50

II) generalization to ∞-dim $\mathcal{H}$

suppose $\mathcal{X}$ is a compact space (e.g. $\mathcal{X} = [0,1]$)

+ Lebesgue measure on it

$$\langle f, g\rangle_{\mathcal{L}_2} \triangleq \int_{\mathcal{X}} f(x) g(x)\, dx$$

$$\mathcal{L}_2(\mathcal{X}) \triangleq \left\{ f: \mathcal{X} \to \mathbb{R} \;\middle|\; \int_{\mathcal{X}} (f(x))^2 dx < \infty \right\}$$

$$\ell_2 \triangleq \left\{ (\alpha_i)_{i=1}^{\infty} \text{ s.t. } \sum_i \alpha_i^2 < \infty \right\}$$

let K be a cts psd. kernel fct. (note: it is symmetric)

↳ with respect to standard on $\mathcal{X}$ & $\mathbb{R}$

$$K\alpha = \sum_i \alpha_i k(x_i, \cdot)$$

$$Kv = \sum_i k_i v_i$$

define
$$\boxed{\begin{aligned} &L_K: \mathcal{L}_2 \to \mathcal{L}_2 \\ &\text{s.t. } [L_K f](\cdot) \triangleq \int_{\mathcal{X}} k(x, \cdot) f(x)\, dx \end{aligned}}$$

then can show that

$$\langle f, L_K g\rangle_{\mathcal{L}_2} = \langle L_K f, g\rangle_{\mathcal{L}_2} \quad \forall f, g$$

L_K is a "compact self adjoint positive" operator

[finite reason

$$\langle v, Kw\rangle = v^T K w \quad \text{if } K = K^T$$
$$= v^T K^T w$$
$$= (Kv)^T w$$
$$= \langle Kv, w\rangle$$

and yields an (countable) orthonormal basis (for $\mathcal{L}_2$)

[Hilbert basis]

of e-functions for L_K $\{\psi_i\}_{i=1}^{\infty}$

with non-negative e-values $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$

ie. $L_K \psi_i = \lambda_i \psi_i$

and we have
$$\boxed{k(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x')}$$

Mercer's thm.

"$K = U \Lambda U^T$"

like $K = \Phi^T \Phi$

cf before

$i=1$

Mercer's thm.

like $K = \Phi^T \Phi$ of before

a) <u>feature space view of $\mathcal{H}$</u> $\mathcal{H} \subseteq \ell_2$

$\Phi : \mathcal{X} \to \ell_2$ with $(\Phi(x))_i \triangleq \sqrt{\lambda_i}\, \psi_i(x)$

here : $k(x,x') = \langle \Phi(x), \Phi(x') \rangle_{\ell_2}$

"<u>diagonalized representation</u>"

[valid element of ℓ_2 because $\sum_i (\Phi(x))_i^2 = \sum_i \lambda_i \psi_i(x)^2 = k(x,x) < \infty$

here identify $k(x,\cdot) \in \mathcal{L}_2$ as $\Phi(x) \in \ell_2$

(do not ~~what $\mathcal{H} \subseteq \ell_2$ looks like though~~)

$\mathcal{H} = \ell_2$?

b) <u>$\mathcal{L}_2$ view</u>

$$\boxed{\begin{gathered}\mathcal{H} \subseteq \mathcal{L}_2 \;:\; \mathcal{H} = \left\{ f \in \mathcal{L}_2 : \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle_{\mathcal{L}_2}^2}{\lambda_i} < \infty \right\} \\ \langle f, g \rangle_{\mathcal{H}} \triangleq \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle_{\mathcal{L}_2} \langle g, \psi_i \rangle_{\mathcal{L}_2}}{\lambda_i}\end{gathered}}$$

$\|f\|_{\mathcal{H}}^2 < \infty$

$\rightsquigarrow$ ellipsoid in $\mathcal{L}_2$

(*) if K is "universal"

$\Rightarrow$ $\mathcal{H}_K$ is dense in $\mathcal{L}_2$ i.e. for any $f \in \mathcal{L}_2$

$\exists$ a sequence $h_n \in \mathcal{H}_K$ s.t. $\|h_n - f\|_{\mathcal{L}_2} \xrightarrow{n\to\infty} 0$

<u>note</u>: if $f \notin \mathcal{H}$ $\Rightarrow$ $\|h_n\|_{\mathcal{H}} \to \infty$

* <u>non-parametric learning</u>

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{H}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i))}_{\substack{n\to\infty \\ \mathbb{E}\,\mathcal{L}(f)}} + \lambda_n \|f\|_{\mathcal{H}}^2$$

$$f^* \triangleq \operatorname*{argmin}_{f \in \mathcal{L}_2} \mathbb{E}\, \mathcal{L}(Y, f(X))$$ perhaps $f^* \notin \mathcal{H}$

but $\mathcal{H}$ dense in $\mathcal{L}_2$
+ regularity property of $\mathcal{L}$ + decrease λ_n <u>at correct rate</u>

$\Rightarrow$ consistency of $\hat{f}_n$ i.e. $\hat{f}_n \xrightarrow[n\to\infty]{\mathcal{L}_2} f^*$

$\Rightarrow$ consistency of $\hat{f}_n$ i.e. $\hat{f}_n \xrightarrow[n\to\infty]{L_2} f^*$

e.g. SVM with RBF kernel is "universally consistent"

when $\lambda_n \to 0$ at correct rate