# lecture 7 - RKHS

Thursday, February 4, 2021 13:39

today: continue RKHS in all their glory! :)

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

$\Phi(x)$ $\quad$ $\Phi(x')$ $\quad$ $\langle \Phi(x), \Phi(x')\rangle = k(x,x')$

$\boxed{\text{representer's thm}}$: 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 $f^* = \sum_{i=1}^{n} \alpha_i^* k(x_i,\cdot)$

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

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

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: $(K)_{ij} \triangleq k(x_i,x_j)$ $\quad n\times n$ matrix

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

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

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

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

I) start with finite matrices

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

$f: \mathcal{X}\to\mathbb{R}$ $\quad \begin{pmatrix} f(x_1)\\ \vdots \\ f(x_n)\end{pmatrix} \in\mathbb{R}^n$ $\quad$ ("$\ell_2$ view")

↳ is finite ⇒ just a vector

$\mathcal{H} = \operatorname{span}\{k(x_i,\cdot) : i=1,\dots,n\} = \{K\alpha : \alpha\in\mathbb{R}^n\} \subseteq \mathbb{R}^n$ $\quad \boxed{\text{"}\ell^2\text{-view"}}$

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

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

if $k$ is a valid kernel (i.e. defines a inner product) $\Rightarrow K \succeq 0$ $\xrightarrow{\text{spectral thm.}}$ $\underset{n\times n}{K} = U \Lambda U^T$ where $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_1 \geq \lambda_2 \geq \dots \lambda_n \geq 0$

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

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

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

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

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

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

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

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

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

$\Rightarrow$ 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 $\psi_1$ vector

"feature space" pt. of view

$$k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathbb{R}^d} = \sum_{\ell=1}^{d} \lambda_\ell \psi_\ell(x_i) \psi_\ell(x_j)$$

note: $K \psi_i = \lambda_i \psi_i$

back to $\ell_2$-view: $\mathcal{H} \subseteq \ell_2$ $(\subseteq \mathbb{R}^n)$

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

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

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

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

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

$K = U \Lambda U^T$

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

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

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

$\Rightarrow U^T v$ is projection of $v$ onto $\{\psi_i\}_{i=1}^n$ basis

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

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

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

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

$$\langle v, v\rangle_{\mathcal{L}_2} = \sum_{i=1}^{n} \beta_i^2$$

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

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

makes an ellipsoid in $\mathcal{L}_2$

i.e. higher coordinates are shrunk more! (since $\lambda_i$'s is smaller as $i \nearrow$)

so orthonormal basis of $\mathcal{H}$ in $\mathcal{L}_2$ ($\|\cdot\|_{\mathcal{H}}$) is $\{\sqrt{\lambda_i}\,\psi_i\}_{i=1}^{d}$

14h45

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

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

+ Lebesgue measure on it

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

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

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

cts: with respect to standard norm on $\mathcal{X}$ & $\mathbb{R}$

$Kv = \sum_i K_{:i} v_i$

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

$L_K$ is a "compact self adjoint positive" operator

self adjoint: $\langle f, L_K g\rangle_{\mathcal{L}_2} = \langle L_K f, g\rangle_{\mathcal{L}_2}$

finite version

$\langle v, Kw\rangle = v^T K w$ ← use $K = K^T$

$= v^T K^T w$

$= (Kv)^T w$

$= \langle Kv, w\rangle$

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 $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \geq 0$

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

and ... $K(x, x') = \sum^{\infty} \lambda_i \psi_i(x) \psi_i(x')$ ... "... T" i.e. $L_K \psi_i = \lambda_i \psi_i$ $\qquad = \langle Kv, w \rangle$

<u>and</u> we have $\boxed{k(x,x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x')}$ $\qquad$ "$U \Lambda U^T$"

Mercer's thm.

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

a) <u>feature space view of $\mathcal{H}$</u> $\mathcal{H} \subseteq \ell_2$ $\qquad \Phi : \mathcal{X} \to \ell_2$ with $(\Phi(x))_i \triangleq \sqrt{\lambda_i}\, \psi_i(x)$

here: $k(x,x') = \langle \Phi(x), \Phi(x') \rangle_{\ell_2}$ $\qquad$ [valid element of $\ell_2$ because $\sum_i (\Phi(x))_i^2 = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)^2$

<u>"diagonalized representation"</u>

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

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

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

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

$$\boxed{\begin{aligned} &\mathcal{H} \subseteq \mathcal{L}_2 : \mathcal{H} = \Big\{ f \in \mathcal{L}_2 : \sum_{i=1}^{\infty} \frac{\overbrace{\langle f, \psi_i \rangle_{\mathcal{L}_2}^2}^{\beta_i}}{\lambda_i} < \infty \Big\} \\ &\langle f, g \rangle_{\mathcal{H}} \triangleq \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle_{\mathcal{L}_2} \langle g, \psi_i \rangle_{\mathcal{L}_2}}{\lambda_i} \end{aligned}}$$

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>

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

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

but $\mathcal{H}$ dense in $\mathcal{L}_2$

+ regularity property of $\mathcal{L}$ + decrease $\lambda_n$ at correct rate but $H$ dense in $\mathcal{L}_2$

+ regularity property of $\mathcal{J}$ + decrease $\lambda_n$ at correct rate

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

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

when $\lambda_n \to 0$ at correct rate