Lecture 7 - scribbles - RKHS

Tuesday, January 29, 2019 14:34

today: continue RKHS

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

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

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

↳ inner product of $\varphi(x_i)$ on data

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

finite dim opt.! thanks to representer's thm!

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

say $\mathcal{X}$ is finite i.e. $\{x_i\}_{i=1}^n$; form Gram matrix $(K)_{ij} \triangleq k(x_i, x_j)$ $n \times n$

k is a valid kernel $\Leftrightarrow K \succeq 0$

i.e. $z = \phi(x)$

K is a valid kernel $\Leftrightarrow K \succeq 0$

$$\mathcal{H} = \text{span}\{k(x_i, \cdot) : i = 1, \dots, n\} = \{K\alpha : \alpha \in \mathbb{R}^n\} \subseteq \mathbb{R}^n$$ "ℓ_2-view"

$f \in \mathcal{H}$ $\quad \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \in \mathbb{R}^n$

$f : \mathcal{X} \to \mathbb{R}$ is finite $\to$ just a vector

$\underset{\in \mathbb{R}^{n \times n}}{K} \succeq 0 \xrightarrow{\text{spectral thm.}} K = U \Lambda U^T$ with $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$

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

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

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

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

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

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

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

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

x-coord. of ψ_1 vector

"feature space view"

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

$$K = \Phi \Phi^T = \sum_{i=1}^n \lambda_i \psi_i \psi_i^T$$

$$k(x_i, x_j) = K_{ij} = \sum_{l=1}^n \lambda_l \psi_l(x_i) \psi_l(x_j)$$

note: $K \psi_j = \sum_i \lambda_i \psi_i \underbrace{\psi_i^T \psi_j}_{\langle \psi_i, \psi_j \rangle_{\ell_2} = \delta_{ij}} = \lambda_j \psi_j$

back to ℓ_2-view: $\mathcal{H} \subseteq \ell_2$; $v \in \mathcal{H} \Rightarrow v = K\alpha$ for some $\alpha \in \mathbb{R}^n$

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

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

so $\|v\|_{\mathcal{H}}^2 = \alpha_v^T K \alpha_v = v^T K^+ \cancel{K K^+} v$ check this? $K K^+ = U \Lambda \underbrace{U^T U}_{I_n} \Lambda^+ U^T = U\begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} U^T$

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

$= (U^T v)^T \Lambda^+ (U^T v)$ $\rightarrow$ $U^T v$ is projection of v on $\{\psi_i\}_{i=1}^n$ basis

so let $v = \sum_{j=1}^n \beta_j \psi_j$ i.e. $\beta_j = \langle v, \psi_j \rangle_{\mathcal{L}_2}$

so $\boxed{\|v\|_{\mathcal{H}}^2 = \sum_{j=1}^n \frac{\overbrace{\langle v, \psi_j \rangle}^{\beta_j \text{ representation}}{}_{\mathcal{L}_2}^2}{\lambda_j}}$

vs. $\|v\|_{\mathcal{L}_2}^2 = \sum_{j=1}^n \langle v, \psi_j \rangle_{\mathcal{L}_2}^2$

and $\|\psi_j\|_{\mathcal{H}}^2 = \frac{1}{\lambda_j}$

$V = \sum_{i=1}^d \beta_i \psi_i$

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

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

so orthonormal basis of $\mathcal{H}$ in $\mathcal{L}_2$ is $\{\sqrt{\lambda_j}\,\psi_j\}_{j=1}^d$
$\hookrightarrow \|\cdot\|_{\mathcal{H}}$

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

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

15h32

- higher coordinates are shrunk more? (since λ_j small)

generalization to ∞-dim space

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

$\mathcal{L}_2(\mathcal{X}) \triangleq \{ f : \mathcal{X} \to \mathbb{R} \mid \int_{\mathcal{X}} (f(x))^2 dx < \infty \}$

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

let k be a continuous psd kernel fct, (symmetric) note: it is

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

define $L_K : \mathcal{L}_2 \to \mathcal{L}_2$

s.t. $[L_K f](\cdot) \triangleq \int_{\mathcal{X}} k(x, \cdot) f(x) dx$

then can show that

L_K is a "compact self-adjoint positive" operator

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

finite version: when $K^T = K$

$\langle v, Kw \rangle = v^T K w = (v^T K^T) w = (Kv)^T w = \langle Kv, w \rangle$

and yields an orthonormal basis (countable) [Hilbert basis] (for $\mathcal{L}_2$) of e-functions for L_K $\{\psi_i\}_{i=1}^{\infty}$

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

i.e. $L_K \psi_i = \lambda_i \psi_i$

and we have $k(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x')$ [like $K = \Phi \Phi^T$ of before]

Mercer's thm.

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

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

"diagonalized representation"

[valid element of ℓ_2 because

$\sum_i (\Phi(x))_i^2 = \sum_i \lambda_i \psi_i(x) \psi_i(x)$

$= k(x, x) < \infty$

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)

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

b) $\mathcal{L}_2$ view ellipsoid in $\mathcal{L}_2$

[illegible] like [illegible])

b) $\mathcal{L}_2$ view

$$\mathcal{H} \in \mathcal{L}_2 : \quad \mathcal{H} = \left\{ f \in \mathcal{L}_2 : \sum_{i=1}^{\infty} \frac{\overbrace{\langle f, \psi_i \rangle_{\mathcal{L}_2}^2}^{\beta_i}}{\lambda_i} < \infty \right\}$$ ellipsoid in $\mathcal{L}_2$

$$\text{and } \langle f, g \rangle_{\mathcal{H}} \triangleq \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle_{\mathcal{L}_2} \langle g, \psi_i \rangle_{\mathcal{L}_2}}{\lambda_i}$$

(*) if k is "universal"

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

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

note: if $f \notin \mathcal{H}$; $\Rightarrow$ $\|h_n\|_{\mathcal{H}} \to \infty$

* non-parametric learning:

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

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

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

+ reg. property of ℓ + correct decrease of λ_n

$\Rightarrow$ consistency of $\hat{f}_n$ i.e. $\hat{f}_n \xrightarrow{\mathcal{L}_2} f^*$

e.g. SVM with RBF kernel is "universally consistent" when $\lambda_n \to 0$ at correct rate