Lecture 7 - RKHS

Tuesday, January 30, 2024 9:35 AM

today: continue RKHS in all their glory! :)

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

$\Phi(x)$ $\Phi(x')$

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

"reproducing property"

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

"$\Phi(x)$"

analog: $\langle w, \Phi(x)\rangle$

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

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

$n\times n$ matrix

Gram matrix: $(K)_{ij} \triangleq k(x_i,x_j) = \langle \Phi(x_i), \Phi(x_j)\rangle_{\mathcal{H}}$

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

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

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

$$f_\alpha(x) = \sum_j \alpha_j k(x_j,x)$$

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

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

I) start with finite matrices

say X is ~~finite~~ e.g. $x_1,\dots,x_n$

$|X| = n$

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

↳ is fine ⇒ just a vector $\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n)\end{pmatrix} \in \mathbb{R}^n$ ("ℓ_2 view")

$(K\alpha)_i = f_\alpha(x_i) = \sum_j \alpha_j k(x_i,x_j) = (K\alpha)_i$

↳ vector $(f(x_1), \dots, f(x_n))$

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

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

if k is a valid kernel (i.e. defines a inner product) $\Rightarrow K \succeq 0$ (psd)

$\text{diag}(\lambda_1, \dots, \lambda_n)$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq 0$

spectral thm. $\Rightarrow K = U \Lambda U^T$ ($n \times n$)

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 $\tilde{\Phi} \triangleq \Lambda^{1/2} U^T$

$\Rightarrow K = \tilde{\Phi}^T \tilde{\Phi}$

$\tilde{\Phi} = \begin{pmatrix} - \sqrt{\lambda_1}\psi_1^T - \\ \vdots \\ - \sqrt{\lambda_d}\psi_d^T - \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ $\quad d = \text{rank}(K) \leq n$

$\Phi : \mathcal{X} \to \mathbb{R}^d$

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

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

$\Rightarrow$ $\Phi(x) = \begin{pmatrix} \sqrt{\lambda_1}\psi_1(x) \\ \vdots \\ \sqrt{\lambda_d}\psi_d(x) \end{pmatrix} \in \mathbb{R}^d$ (x-coord. of ψ_i vector)

"feature space" pt. of view

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

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

back to ℓ_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 $\alpha_v = K^{\dagger} v$ ($K^{\dagger}$: pseudo-inverse)

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

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

$= (U^T v)^T \Lambda^{\dagger} (U^T v)$

$K = U \Lambda U^T$

$K^{\dagger} = U \Lambda^{\dagger} U^T$

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

$\|v\|_{\mathcal{H}}^2 = \sum_{j=1}^{d} \frac{\langle v, \psi_j \rangle_{\ell_2}^2}{\lambda_j}$ (ℓ_2 representation)

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

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

more generally, $\|v\|_{\mathcal{H}}^2 = \sum_{j=1}^{n} \frac{\langle v, \psi_j \rangle_{\ell_2}^2}{\lambda_j}$

$= +\infty$ if any $\langle v, \psi_j \rangle \neq 0$ for $j > d$

for $j > d$

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

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

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

$\|\cdot\|_{\mathcal{H}}$

so orthonormal basis of $\mathcal{H}$ in ℓ_2 is $\{\sqrt{\lambda_i}\psi_i\}_{i=1}^{d}$

$\langle v, v \rangle_{\ell_2} = \sum_{i=1}^{n} \beta_i^2$

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

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

↳ makes an ellipsoid in ℓ_2

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

Oh 47

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

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

+ Lebesgue measure on it

$\ell_2(\mathcal{X}) \triangleq \{ f : \mathcal{X} \to \mathbb{R} \mid \int_{\mathcal{X}} (f(x))^2 dx < \infty \}$

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

Let k be a cts. psd kernel fct. (note: it is symmetric)

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

define $L_k : \ell_2 \to \ell_2$ $\quad k(\cdot, x)$

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

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

then can show that

L_k is a "compact self-adjoint positive" operator

and yields an (countable) orthonormal basis [Hilbert basis] (for ℓ_2)

$\langle f, L_k g \rangle_{\ell_2} = \langle L_k f, g \rangle_{\ell_2} \quad \forall f, g$

finite version: $\langle v, Kw \rangle = v^T K w = v^T K^T w = \langle Kv, w \rangle$ (use $K = K^T$)

of eigenfunctions 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 $k(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x')$ "$U \Lambda U^T$"

cond we have $K(x,x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x')$ — Mercer's thm

" $U \Lambda U^T$ like $K = \tilde{\Phi}^T \tilde{\Phi}$ of before

a) feature space $\mathcal{H} \subseteq \ell_2$

view of $\mathcal{H}$ $\quad \Phi : \mathcal{X} \to \ell_2$ 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"

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

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

(do not know what $\mathcal{H} \subseteq \ell_2$ looks like though) β_i $\|f\|_{\mathcal{H}}^2 < \infty$

b) ℓ_2 view

$\mathcal{H} \subseteq \ell_2$: $\mathcal{H} = \{ f \in \ell_2 : \sum_{i=1}^{\infty} \underbrace{\langle f, \psi_i \rangle_{\ell_2}}_{\lambda_i} < \infty \}$ → ellipsoid in ℓ_2

$\langle f, g \rangle_{\mathcal{H}} \triangleq \sum_{i=1}^{\infty} \underbrace{\langle f, \psi_i \rangle_{\ell_2} \langle g, \psi_i \rangle_{\ell_2}}_{\lambda_i}$

(*) if K is "universal"

$\Rightarrow \mathcal{H}_K$ is dense in ℓ_2 i.e for any $f \in \ell_2$

$\exists$ a sequence $h_n \in \mathcal{H}_K$ s.t. $\|h_n - f\|_{\ell_2} \xrightarrow{n \to \infty} 0$

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

(*) non-parametric learning:

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

$f^* \triangleq \underset{f \in \ell_2}{\operatorname{argmin}} \mathbb{E}\ell(Y, f(X))$ perhaps $f^* \notin \mathcal{H}$

but $\mathcal{H}$ dense in ℓ_2

→ regularity property ℓ + decrease λ_n at correct rate

$\Rightarrow$ consistency of $\hat{f}_n$ i.e. $\hat{f}_n \xrightarrow[n \to \infty]{\ell_2} f^*$ $\quad \|\hat{f}_n - f^*\|_{\ell_2}$

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