Lecture 7 - RKHS

Tuesday, January 28, 2020 14:29

today: continue RKHS in all their glory? :)

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 data

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 $K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$

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

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

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

I) start with finite matrices:

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

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

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

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

if k is a valid kernel ⇒ $K \succcurlyeq 0$

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

spectral thm.

because $K \succeq 0$ $\overset{\text{spectral thm.}}{\Rightarrow}$ $K = U \Lambda U^T$ ($n \times n$)

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

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

ie. $U = (\psi_1 \ \dots \ \psi_n)$

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

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

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

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

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

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

If we define $\phi(x) \triangleq \begin{pmatrix} \sqrt{\lambda_1}\,\psi_1(x) \\ \vdots \\ \sqrt{\lambda_d}\,\psi_d(x) \end{pmatrix} \in \mathbb{R}^d$ (x-coord of ψ_i vector)

"feature space pt. of view"

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

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

$$k(x_i, x_j) = k_{ij} = \sum_{\ell=1}^{n} \lambda_\ell \psi_\ell(x_i) \psi_\ell(x_j)$$

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: $H \subseteq \ell_2$; $v \in H \Rightarrow v = K\alpha$ for some $\alpha \in \mathbb{R}^n$

to get $\|v\|_H$, we compute $\alpha_v = K^\dagger v$ (pseudo-inverse)

so $\|v\|_H^2 = \alpha_v^T K \alpha_v = v^T K^\dagger K K^\dagger v$

$K = U \Lambda U^T$

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

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

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

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

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

ie. $v = \sum_{j=1}^{n} \beta_j \psi_j$ (β_j representation) ie. $\beta_j = \langle v, \psi_j \rangle_{\ell_2}$

so $$\|v\|_H^2 = \sum_{j=1}^{n} \frac{\langle v, \psi_j \rangle_{\ell_2}^2}{\lambda_j}$$

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

and $$\|\psi_j\|_H^2 = \frac{1}{\lambda_j}$$

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

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

λ_j

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

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

So orthonormal basis of $\mathcal{H}$ in ℓ_2 is $\{\sqrt{\lambda_j}\psi_j\}_{j=1}^{q}$
↳ $\|\cdot\|_{\mathcal{H}}$

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

↳ make an ellipsoid in ℓ_2

i.e. higher coordinates are shrunk more?
(since λ_j is smaller as j ↗)

15h31

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

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

$$\ell_2(\mathcal{X}) \triangleq \left\{ f : \mathcal{X} \to \mathbb{R} \;\middle|\; \int_{\mathcal{X}} (f(x))^2 dx < \infty \right\}$$

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

Let k be a continuous psd kernel fct. (note it is symmetric)
↳ with respect to standard norm on $\mathcal{X}$ & $\mathbb{R}$

define
$$L_k : \ell_2 \to \ell_2$$
$$\text{s.t. } [L_k f](\cdot) \triangleq \int_{\mathcal{X}} k(x, \cdot) f(x) dx$$

then can show that
L_k is a "compact selfadjoint positive" operator → $\langle f, L_k g \rangle_{\ell_2} = \langle L_k f, g \rangle_{\ell_2}$ $\forall f, g \in \ell_2$

and yields an (countable) orthonormal basis (for ℓ_2)
[Hilbert basis]
of e-functions for L_k $\{\psi_i\}_{i=1}^{\infty}$
with non-negative e-values $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$

finite version. when $k = k^T$

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

ie $L_k \psi_i = \lambda_i \psi_i$

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

Mercer's thm

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

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

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

a) feature space view of $\mathcal{H}$

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

here: $k(x,x') = \langle \Phi(x), \Phi(x') \rangle_{\ell_2}$ [valid element of ℓ_2 because $\sum_i (\Phi(x))_i^2 = \sum_i \lambda_i \psi_i(x)\psi_i(x) = k(x,x) < \infty$]

"diagonalized representation"

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

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

b) $\mathcal{H}$ view:

$\mathcal{H} \subset \mathcal{L}_2$: $\mathcal{H} = \{ f \in \mathcal{L}_2 : \sum_{i=1}^{\infty} \underbrace{\langle f, \psi_i \rangle_{\mathcal{L}_2}^2}_{\lambda_i} < \infty \}$ $\|f\|_{\mathcal{H}}^2 < \infty$ → ellipsoid in $\mathcal{L}_2$

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 $h_n \in \mathcal{H}_k$ s.t. $\|h_n - f\| \xrightarrow[n \to \infty]{\mathcal{L}_2} 0$

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

* non-parametric learning:

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

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

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

+ regularity property of ℓ + correct decrease of λ_n

$\Rightarrow$ 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