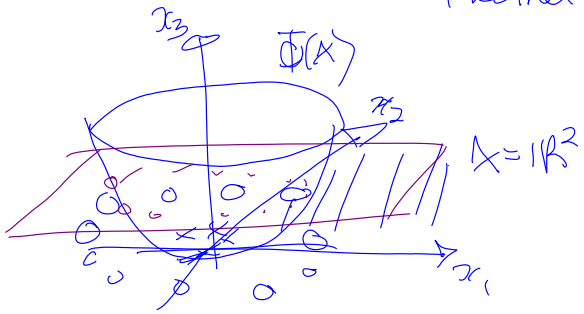
Lecture 7 - scribbles - RKHS

Tuesday, February 13, 2018
14:31

today: more on RKHS

RKHS:

motivation: generalize $\langle w, \varphi(x) \rangle$ to higher dimensional space
+ kernel trick $\langle \varphi(x), \varphi(y) \rangle = k(x,y)$



$\mathcal{X} = \mathbb{R}^2$

$\Phi : \mathcal{X} \to \mathbb{R}^3$

$\Phi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{pmatrix}$

$\langle \Phi(x), \Phi(y) \rangle_{\mathbb{R}^3} = (\langle x, y \rangle_{\mathbb{R}^2})^2 = k(x,y)$

Representer's thm says that: for $\mathcal{H}$ RKHS produced by a kernel K i.e. $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.t. symmetric & psd.

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

is reached for $f^* = \sum_{i=1}^n \alpha_i^* k(x_i, \cdot)$ i.e. $\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^n l\left(y_i, \underbrace{\sum_{j=1}^n k(x_j, x_i) \alpha_j}_{f(x_i)}\right) + \lambda \alpha^T K \alpha$

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

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

$\langle k(x_i, \cdot), k(x_j, \cdot) \rangle_{\mathcal{H}} = k(x_i, x_j)$

for any finite $\{x_i\}_{i=1}^m$ and $\alpha \in \mathbb{R}^m$ and any m

$\alpha^T K \alpha \geq 0$

$(K)_{ij} = k(x_i, x_j)$

"Gram matrix on $\{x_i\}_{i=1}^m$"

RKHS:

RKHS:

$\Phi: \mathcal{X} \to \mathcal{H}$ s.t. $\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = k(x,y)$

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

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

"pre-Hilbert" space [inner product space]

with $\langle f, g \rangle_{\tilde{\mathcal{H}}} = \sum_{i,j} \alpha_i^f \alpha_j^g k(x_i^f, x_j^g)$ $\quad \|f\|_{\tilde{\mathcal{H}}} \triangleq \sqrt{\langle f, f \rangle_{\tilde{\mathcal{H}}}}$

then RKHS $\mathcal{H}$ is = completion($\tilde{\mathcal{H}}$) using $\|\cdot\|_{\tilde{\mathcal{H}}}$ as a norm

ie. add all limit points of $\tilde{\mathcal{H}}$-Cauchy sequence to $\mathcal{H}$

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

$\{x_i\}_{i=1}^m$

by definition, we have that $\langle f, K(x,\cdot) \rangle_{\mathcal{H}} = f(x)$

$\Downarrow$

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

$E_x(f) = f(x)$

$\Downarrow$

$|f(x) - g(x)|$
$= |\langle f - g, k(x,\cdot) \rangle|$
C.S. $\leq \|f - g\|_{\mathcal{H}} \|k(x,\cdot)\|_{\mathcal{H}}$

ie. E_x is Lipschitz continuous with $L = \|k(x,\cdot)\|_{\mathcal{H}}$

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

K a valid kernel $\Rightarrow K \succcurlyeq 0$

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

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

$K \in \mathbb{R}^{n \times n}$, $K \succcurlyeq 0$ $\overset{\text{spectral thm.}}{\Rightarrow}$ $K = U \Lambda U^T$ 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 $\mathbb{R}^n$

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

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

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

$\Phi = \begin{pmatrix} \sqrt{\lambda_1} \psi_1^T \\ \end{pmatrix}$ row here: $\sqrt{\lambda_i}(\psi_i(x_1) \cdots \psi_i(x_n))$

suppose $\lambda_{d+1}=0$ $\Rightarrow$ $\begin{pmatrix} \sqrt{\lambda_d}\psi_{d1} \\ 0 \\ 0 \end{pmatrix}$

if define $\Phi(x) \triangleq \begin{pmatrix} \sqrt{\lambda_1}\psi_1(x) \\ \vdots \\ \sqrt{\lambda_d}\psi_d(x) \end{pmatrix} \in \mathbb{R}^d$ ← "feature space view"

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

$K = \Phi\Phi^T = \sum_{j=1}^{n} \lambda_j \psi_j \psi_j^T$

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

$\mathcal{H} \cong \mathbb{R}^n$ when $\mathcal{X}$ = finite set

$K\psi_j = \sum_{i=1}^{n} \lambda_i \psi_i \underbrace{\psi_i^T \psi_j}_{\langle \psi_i, \psi_j \rangle_{\mathbb{R}^n} \; \langle \psi_i, \psi_j \rangle_{\mathcal{H}}} \Rightarrow K\psi_j = \lambda_j \psi_j$

back to $\mathcal{H}$-view: $\mathcal{H} \subset \mathcal{H}$; $v \in \mathcal{H} \Rightarrow v = \sum_{j=1}^{n} \alpha_j K(x_j, \cdot)$ for some α

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

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

so $\|v\|_{\mathcal{H}}^2 = \alpha_v^T K \alpha_v = v^T K^{\dagger} v = (U^T v)^T \Lambda^{\dagger} (U^T v)$ $\rightsquigarrow$ $U^T v$ is projection of v on $\{\psi_j\}_{j=1}^n$ basis

so let $v = \sum_j \beta_j \psi_j$ i.e. $\beta_j = \langle v, \psi_j \rangle_{\mathcal{H}}$; then $\boxed{\|v\|_{\mathcal{H}}^2 = \sum_{j=1}^{n} \frac{\langle v, \psi_j \rangle_{\mathcal{H}}^2}{\lambda_j}}$ β_j = representation

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

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

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

$\mathcal{H}$ (outside space) view of $\mathcal{H}$: $\mathcal{H} \subset \mathbb{R}^n$ (i.e. $\mathcal{H}$) say $v \in \mathcal{H}$ ($v \in \mathrm{Im}(K)$) i.e. $v = \sum_{j=1}^{d} \beta_j \psi_j$

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

but $\langle v, v \rangle_{\mathcal{H}} = \sum_{j=1}^{d} \beta_j^2 \langle \psi_j, \psi_j \rangle_{\mathcal{H}} = \sum_{j=1}^{d} \frac{\beta_j^2}{\lambda_j}$ thus $\|v\|_{\mathcal{H}} \le 1$ makes ellipsoid in $\mathcal{H}$

feature space view : $\mathcal{H} \subseteq \mathbb{R}^d$ $\text{span}(\Phi(x))$ $\langle v, v \rangle_{\mathcal{H}} = \langle v, v \rangle_{\mathbb{R}^d}$

ie. already diagonalized

(with small coordinates for j s.t. λ_j is small)

$|X| = n$

$\mathbb{R}^X \cong \mathbb{R}^n$

generalization to ∞-dim space:

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

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

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

Let k be a continuous psd kernel function

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

when $k = k^T$

$\langle v, kw \rangle = v^T k w = (v^T k^T) w = \langle kv, w \rangle$

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

s.t. $[L_k f](\cdot) \triangleq \int_X k(x, \cdot) f(x) dx$ → $\langle f, Lkg \rangle_{\mathcal{L}_2} = \langle Lkf, g \rangle_{\mathcal{L}_2}$ $\forall f, g \in \mathcal{L}_2$

then can show that

L_k is a "compact self-adjoint positive" operator

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

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

ie. $L_k \psi_j = \lambda_j \psi_j$

Mercer's thm

and we have $k(x,y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(y)$ [like the $K = \Phi \Phi^T$ of before?]

feature space $\mathcal{H} \subseteq \ell_2$ $\Phi : X \to \ell_2$ with $(\Phi(x))_i \triangleq \sqrt{\lambda_i} \psi_i(x)$ $\left(\sum_{i=1}^{\infty} (\Phi(x))_i^2 = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)^2 \right)$

of $\mathcal{H}$

i.e. if we identify $k(x,\cdot) \in \ell_2$ as $\phi(x) \in \ell_2$

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

where $k(x,y) = \langle \phi(x), \phi(y) \rangle_{\ell_2}$ "diagonalized representation"

ℓ_2 now

(do not know what is $\mathcal{H} \subset \ell_2$ though)

$$\mathcal{H} \subset \ell_2 : \quad \mathcal{H} = \left\{ f \in \ell_2 : \sum_{i=1}^{\infty} \frac{\langle f, \psi_i \rangle_{\ell_2}}{\lambda_i^2} \langle \psi_i, \cdot \rangle < \infty \right\}$$

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

Structured prediction setup:

$S(\underset{\in \mathcal{X}}{\cdot}, y) \in \mathcal{H}_y$ for each y

before, had $S(x,y;w)$

now, we have $S(x,y;\vec{f})$ $\quad \vec{f} = (f_{y_1}, \dots, f_{y_K})$

$\in \mathcal{H}_{y_1}$

i.e. $\vec{f} : \mathcal{X} \to \mathbb{R}^K$

$S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

consider $\ell(x,y,S) \triangleq \frac{1}{2k} \sum_{\tilde{y}} \left(S(x,\tilde{y}) + \ell(y,\tilde{y}) \right)^2$

$K = |\mathcal{Y}|$

can be thought of as generalization of squared loss for binary classification to multiclass $\to (1 - y_i \langle w, \phi(x_i) \rangle)^2$

$$\ell_{q_x}(x,S) \triangleq \mathbb{E}_{q(y|x)} \left(\frac{1}{2k} \sum_{\tilde{y}} S(x,\tilde{y})^2 + 2 S(x,\tilde{y}) \ell(y,\tilde{y}) + \text{constant} \right)$$

$$= \frac{1}{2k} \sum_{\tilde{y}} \left(S(x,\tilde{y})^2 + 2 S(x,\tilde{y}) \underbrace{\mathbb{E}_{q(y|x)} [\ell(y,\tilde{y})]}_{\ell_{q_x}(\tilde{y})} + \text{cst.} \right)$$

$\ell_{q_x}(\tilde{y})$

Suppose s is unconstrained:

$$\min_s \ell_{q_x}(x;s) \rightarrow 2s(\tilde{y}) + 2\,\ell_{q_x}(\tilde{y}) = 0 \quad \forall \tilde{y}$$

$$\Rightarrow s^*(\tilde{y}) = -\ell_{q_x}(\tilde{y})$$

$$\operatorname*{argmax}_{\tilde{y}} s^*(x,\tilde{y}) = \operatorname*{argmin}_{\tilde{y}} \ell_{q_x}(\tilde{y}) \quad \text{i.e. you predict optimally}$$

so ℓ here is consistent

↳ i.e. $s^* \in \operatorname{argmin}_{\text{over all } s} \ell(s) \Rightarrow$ that $L(h_{s^*}) = \min_{\text{all } h} L(h)$

$$h_s(x) \triangleq \operatorname*{argmax}_{\tilde{y}} s(x,\tilde{y})$$