Mar 10^{th} - seminar

Mar 12^{th} - project presentations

1- page prop. 5 pages

—

Representer theorem:

$$\min_{h \in \mathcal{H} \to \text{RKHS}} \hat{R}_S(h) + \lambda \alpha(\|h\|_{\mathcal{H}}^2)$$

α : strictly increasing function on $\mathbb{R}^+$

Soln : $h(x) = \sum_{i=1}^{m} a_i \kappa(x_i, x)$

$S = \{x_i\}_{i \in [m]}$

Last time : • Proof of repr.

- Assuming $h(x) = \sum_{i=1}^{m} a_i \kappa(x_i, x)$, how to solve for $\{a_i\}_{i \in [m]}$

$a \in \mathbb{R}^m$ solves linear regression problem

$$a = (G^T G + \lambda I)^{-1} G Y$$

G : Gram matrix $\in \mathbb{R}^{m \times m}$

$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$ true values/labels

for the case : $\hat{R}_S(h) = \sum_{i=1}^{m} (y_i - h(x_i))^2$

$$\alpha(\|h\|_{\mathcal{H}}^2) = \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2$$

The regularization term ($\alpha(\|h\|_{\mathcal{H}}^2)$) decreases "capacity" of hypothesis class.

"Smallness" of function class

- $\mathcal{H}$: RKHS : inf. dim. for typical kernels.

Segue into linear regression

$$h(x) = w^T x \qquad \mathcal{H} = \{ w^T x : w \in \mathbb{R}^d \}$$

$$w = (X^T X)^{-1} X^T Y$$

$d = m \qquad X \in \mathbb{R}^{m \times d}$

$d < m$ capacity may be small

$d \gg m$ not well cond., $X^T X$ is not invertible!

upon adding Regularization :

$$w = (X^T X + \lambda I)^{-1} X^T Y$$

ℓ^2 regularization: $\| Y - X w \|^2 + \lambda \| w \|^2$

Regularization $\Rightarrow$ numerical stability

Effect of regularization in inf-dim regression

Form of Regularization : $\|h\|_{\mathcal{H}}^2 = \langle h, h \rangle_{\mathcal{H}}$

if $h = \mathcal{K}(x, \cdot)$

$$\mathcal{K}(x, x) = \langle h, h \rangle_{\mathcal{H}}$$

But WKT from Mercer's theorem,

$$\mathcal{K}(x, x) = \sum_i \lambda_i \psi_i(x) \psi_i(x)$$

where (λ_i, ψ_i) are eigenvalue-eigenfunction pairs of HS / kernel integral operator

$$T_{\mathcal{K}} f(x) = \int_{\mathcal{X}} \mathcal{K}(x, x') f(x') \, dx'$$

can replace $dx' \to d\mu(x')$

$\langle T_{\mathcal{K}} f, f \rangle \geq 0$ PD operator

$\uparrow$

L^2 inner prod.

$\lambda_i \geq 0$. $\sum \lambda_i < \infty$ (conseq. of Mercer's thm)

Interpretation of inner product on RKHS

For $f, g \in \mathcal{H}$ (RKHS),
define some operator Υ
so that

$$\langle f, g \rangle_{\mathcal{H}} = \langle \Upsilon f, \Upsilon g \rangle$$

L^2 inner product

Υ: regularization operator.

Regularization: $\|f\|_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}}$
$= \langle \Upsilon f, \Upsilon f \rangle$
$= \langle f, \Upsilon^* \Upsilon f \rangle$

Υ^*: adjoint of Υ

$\Upsilon^* \Upsilon$: always PD operator
$\langle \Upsilon^* \Upsilon f, f \rangle \geq 0$.

WLOG, consider Υ to be PD. in order to understand $\|f\|_{\mathcal{H}}^2$.

Takeaway:

$\mathcal{K} \Leftrightarrow$ RKHS $\Leftrightarrow \Upsilon$
PD $\quad \mathcal{H}$

Υ can be used to define PD $\mathcal{K}$.

Bochner's theorem: Any PD translation-invariant kernel can be written as the Fourier transform of a positive measure.

Symmetric
$\mathcal{K}(x, x') = \mathcal{K}(x', x)$

Translation-invariant
$\mathcal{K}(x, x') = f(x - x')$
Absolution position on image space does not matter.

e.g. $\mathcal{K}(x, x') = e^{-\frac{\|x - x'\|^2}{2\sigma^2}}$

B.T. says if $\mathcal{K}$ is PD
$\mathcal{K}(x) = (2\pi)^{-d/2} \int e^{-i\xi \cdot x} \rho(\xi) \, d\xi$

$\xi \to \rho(\xi)$ is positive. $\quad \rho(\xi) = \rho(-\xi) \geq 0$.
at all $\xi \in \mathbb{R}^d$

Fourier transform: operator on L^2

$$\mathcal{F}f(\xi) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} f(x) e^{-i\xi \cdot x} \, dx$$

Inverse Fourier transform

$$f(x) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} (\mathcal{F}f)(\xi) e^{i\xi \cdot x} \, d\xi$$

Connection with regularization
Want Υ s.t.
$\langle \Upsilon^* \Upsilon f, f \rangle = \langle \Upsilon f, \Upsilon f \rangle$
$= \|f\|_{\mathcal{H}}^2$

Green's function (typically used in (numerical) PDE theory) $\quad G: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

$$\langle \Upsilon^* \Upsilon G(x, \cdot), f \rangle = f(x)$$

L^2 inner product

(In PDEs, solns are written in span of $G(x, \cdot)$ since the diff operator takes $G(x, \cdot)$ to δ_x)

Ansatz: (for construction of $\Upsilon^* \Upsilon$)
Define

$$\langle \Upsilon f, \Upsilon g \rangle = (2\pi)^{-d/2} \int \frac{\overline{\mathcal{F}(f)(\xi)} \, \mathcal{F}(g)(\xi)}{\rho(\xi)} \, d\xi$$

Some +ve density $\rho(\xi)$.

What is an ansatz for $G(x, \cdot)$ of $\Upsilon^* \Upsilon$?

$$\langle \Upsilon^* \Upsilon G(x, \cdot), f \rangle = f(x)$$

$$\langle \Upsilon^* \Upsilon G(x, \cdot), f \rangle = \langle \Upsilon G(x, \cdot), \Upsilon f \rangle$$
$$= (2\pi)^{-d/2} \int \frac{\overline{\mathcal{F}G(x, \cdot)(\xi)} \, \mathcal{F}(f)(\xi)}{\rho(\xi)} \, d\xi \quad \to (1)$$

Given a positive density, we can define a PD kernel by B.T.

$$\mathcal{K}(x - x') = \mathcal{F}(\rho)(x - x')$$
$$= (2\pi)^{-d/2} \int e^{-i(x - x') \cdot \xi} \rho(\xi) \, d\xi$$

Set $G(x, x') = \mathcal{K}(x' - x)$
To check that this is indeed the Green's function for our $\Upsilon^* \Upsilon$, sub. in (1),

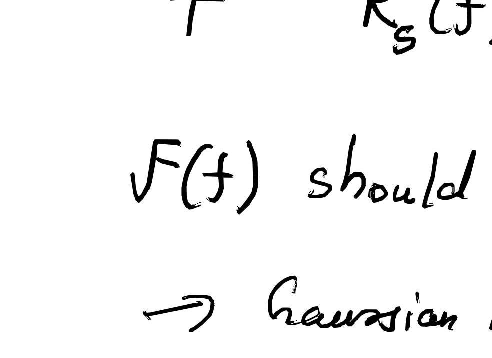
$$\langle \Upsilon^* \Upsilon G(x, \cdot), f \rangle$$
$$= (2\pi)^{-d/2} \int e^{ix \cdot \xi} \frac{\cancel{\overline{\rho(\xi)}} \, \mathcal{F}(f)(\xi)}{\cancel{\rho(\xi)}} \, d\xi$$
$$= f(x)$$

Takeaway: Given ρ, one can define PD kernel using B.T.
and a corresponding regularization operator $\Upsilon^* \Upsilon$
s.t. $\langle \Upsilon^* \Upsilon f, f \rangle = \|f\|_{\mathcal{H}}^2$

where
$$\langle \Upsilon^* \Upsilon f, g \rangle = \langle \Upsilon f, \Upsilon g \rangle$$
$$= (2\pi)^{-d/2} \int \frac{\overline{\mathcal{F}f(\xi)} \, \mathcal{F}(g)(\xi)}{\rho(\xi)} \, d\xi$$

E.g.
Gaussian kernel $\quad e^{-\frac{(x - x')^2}{2\sigma^2}}$
$\rho(\xi) = \sigma e^{-\xi^2 / 2\sigma^2}$

Reg. term
$$\|f\|_{\mathcal{H}}^2 = (2\pi)^{-1/2} \int \frac{\overline{\mathcal{F}f(\xi)} \, \mathcal{F}(f)(\xi)}{\rho(\xi)} \, d\xi$$



if $\mathcal{F}(f)(\xi) \to 0$ rapidly, then $\|f\|_{\mathcal{H}}^2$ can be small.

For a minimizer f
of $\hat{R}_S(f) + \lambda \|f\|_{\mathcal{H}}^2$

$\mathcal{F}(f)$ should ~~have~~ be rapidly decaying
$\to$ Gaussian kernels pick smooth functions

Laplace kernel: can pick non smooth functions

$$e^{-\frac{\|x - x'\|}{2\sigma^2}}$$

$\hat{R}_S(h) + \lambda \|h\|_{\mathcal{H}}^2$ can be interpreted as a functional on $\mathcal{H}$.

Suppose this is a continuous functional on a compact subset of $\mathcal{H}$.

Then, the inverse exists and is continuous.