AWZ ~ 2 different textbooks

Foundations of Machine Learning Mohri 2018

Learning with kernels Smola & Scholkopf 2002

—

Interpret kernel regression on RKHS

- closed form solution (finite-dim space)

$\operatorname{span}\{\mathcal{K}(x_i, \cdot)\}_{i \in [m]}$ $\qquad \{(x_i, y_i)\}_{i \in [m]} \equiv S$

- why does the kernel regression problem on RKHS become finite-dim ensured?

$$\min_{\substack{h \in \mathcal{H} \\ \downarrow \\ \text{RKHS}}} \frac{1}{m} \sum_{i=1}^{m} \ell(x_i, y_i, h(x_i)) + \lambda \|h\|_{\mathcal{H}}^2$$

Regularization term: reduces "capacity" of $\mathcal{H}$.

- "Smallness" of $\mathcal{H}$ comes from the compactness of $\Upsilon^*\Upsilon$ (regularization operator)

$\Upsilon^*\Upsilon \iff \mathcal{K}$ for PD kernels have a one-to-one correspondence

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

$\underbrace{\langle \Upsilon f, \Upsilon g \rangle}_{L^2 \text{ inner product}} = \underbrace{\langle f, g \rangle_{\mathcal{H}}}_{\text{RKHS product}}$

- $\Upsilon^*\Upsilon$ PD

$\equiv T_{\mathcal{K}}^{-1} \qquad T_{\mathcal{K}} g(x') = \int g(x)\, \mathcal{K}(x', x)\, dx$

- Green's function of $\Upsilon^*\Upsilon$

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

Last time: $G(x, x') = \mathcal{K}(x, x')$

- $\langle \Upsilon^*\Upsilon f, \mathcal{K}(x, \cdot) \rangle = f(x)$

$\int (\Upsilon^*\Upsilon f)(z)\, \mathcal{K}(x, z)\, dz = f(x)$

$\int (\Upsilon^*\Upsilon)(T_{\mathcal{K}} f)(z)\, \mathcal{K}(x, z)\, dz = T_{\mathcal{K}} f(x)$

$= \int f(z)\, \mathcal{K}(x, z)\, dz$

$\Rightarrow \int \left((\Upsilon^*\Upsilon)(T_{\mathcal{K}}) f(z) - f(z) \right) \mathcal{K}(x, z)\, dz = 0$

$(\Upsilon^*\Upsilon)\, T_{\mathcal{K}} = Id$ on $\mathcal{H}$.

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

space of functions on which $\|f\|_{\mathcal{H}}^2$ is minimized is "small"

e.g. Gaussian kernel $\rightarrow$ smooth functions (functions for which the Fourier coefficients decay rapidly) have small $\langle \Upsilon^*\Upsilon f, f \rangle = \|f\|_{\mathcal{H}}^2$

Questions: (1) what functions are "likely" to be found as minimizers?

(2) what are characteristics of the class of minimizers?

(2): Sample complexity of minimization problems

(1): Gaussian process interpretation of regression

: "Bayesian" perspective

Function Complexity

PAC bound (probably approximately correct) for finite hypothesis class:

$$\underset{\text{error}}{\text{generalization}} \leq \frac{1}{m} \log\left(\frac{|\mathcal{H}|}{\delta}\right)$$

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell(x, y, h(x)) \quad \text{w.h.p.}$$

Training error

$$\frac{1}{m} \sum_{(x,y)\in S} \ell(x, y, h(x))$$

Under the realizability assumption, training error $= 0$ for ERMs.

For an ERM $h \in \mathcal{H}$,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell(x, y, h(x)) \leq \varepsilon = \frac{1}{m} \log\frac{|\mathcal{H}|}{\delta}$$

with Prob. $\geq 1-\delta$
(over the randomness in training data)

Sample complexity: function $(\varepsilon, \delta) \to m(\varepsilon, \delta)$
s.t. if trained with at least $m(\varepsilon, \delta)$ samples, then, gen. error $R(h) < \varepsilon$ with prob. $\geq 1-\delta$.

if $\quad m = \frac{1}{\varepsilon} \log\frac{|\mathcal{H}|}{\delta}$

then, $R(h) < \varepsilon \quad$ with prob $\geq 1-\delta$.

Kernel regression: $\mathcal{H}$: RKHS

vacuous bound

Rademacher complexity:
(one of many notions
to represent complexity
of function class)

Given $S \equiv \{(x_i, y_i)\}_{i \in [m]}$

$$\mathrm{Rad}_S(\mathcal{H}) = \frac{1}{m} \mathbb{E}_{\sigma} \sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i \, h(x_i)$$

$$\sigma = [\sigma_1, \ldots, \sigma_m]^T \in \mathbb{R}^m$$

σ_i iid Rademacher RV

$$\sigma_i = \begin{cases} +1 & \text{with Pr. } 1/2 \\ -1 & \text{with Pr } 1/2 \end{cases}$$

$\mathrm{Rad}(\mathcal{H}) \uparrow \implies$ can represent noisy functions

Rademacher complexity of functions learned in kernel regression on RKHS

$$f(x) = \alpha \cdot \begin{bmatrix} \mathcal{K}(x_1, x) \\ \vdots \\ \mathcal{K}(x_m, x) \end{bmatrix} = \sum_{i=1}^{m} \alpha_i \mathcal{K}(x_i, x)$$

$$= \alpha \cdot \Phi(x)$$

$\lambda \alpha^T G \alpha \leq \lambda \|\alpha\|^2 \|G\|$

$$\mathcal{H} = \{ x \to \alpha \cdot \Phi(x) : \|\alpha\| < \Lambda \}$$

$Tr(G) < r$

Reg. term: $\lambda \alpha^T G \alpha$

$S = (x_i, y_i)_{i=1}^m$

$$Rad_S(\mathcal{H}) = \frac{1}{m} \mathop{E}_{\sigma} \sup_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i)$$

$$= \frac{1}{m} \mathop{E}_{\sigma} \sup_{\|\alpha\| < \Lambda} \sum_{i=1}^{m} \sigma_i \alpha \cdot \phi(x_i)$$

$$= \frac{1}{m} \mathop{E}_{\sigma} \sup_{\|\alpha\| < \Lambda} \alpha \cdot \sum_{i=1}^{m} \sigma_i \phi(x_i)$$

$$= \frac{\Lambda}{m} \mathop{E}_{\sigma} \left\| \sum_{i=1}^{m} \sigma_i \phi(x_i) \right\| \quad \text{(Cauchy Schwarz)}$$

$$\leq \frac{\Lambda}{m} \left(\mathop{E}_{\sigma} \left\| \sum_{i=1}^{m} \sigma_i \phi(x_i) \right\|^2 \right)^{1/2}$$

$[\sigma_1, \ldots, \sigma_m]$

(Jensen's inequality $EX^2 \geq (EX)^2$)

$$\left(\mathop{E}_{\sigma} \left\| \sum_{i=1}^{m} \sigma_i \phi(x_i) \right\|^2 = \mathop{E}_{\sigma} \sum_{i=1}^{m} \sigma_i^2 \|\phi(x_i)\|^2 + \mathop{E}_{\sigma} \sum \sigma_i \sigma_j \phi(x_i) \phi(x_j) \right.$$

$$\left. \mathop{E}_{\sigma} \sigma_i \sigma_j = 0 \right)$$

σ_i & σ_j are independent

$$\leq \frac{\Lambda}{m} \left(\mathop{E}_{\sigma} \sum_{i=1}^{m} \sigma_i^2 \|\phi(x_i)\|^2 \right)^{1/2}$$

$$= \frac{\Lambda}{m} \left(\sum_{i=1}^{m} \|\phi(x_i)\|^2 \right)^{1/2} = \frac{\Lambda}{m} \|G\|_F$$

Recall

$\phi(x_i) = [\mathcal{K}(x_i, x_1), \ldots, \mathcal{K}(x_i, x_m)]$

$\|\phi(x_1)\|^2 = ?$

$$Rad_S(\mathcal{H}) \leq \frac{\Lambda}{m} \|G\|_F$$

Thm:

Form of generalization bound:

$$R(h) \leq \hat{R}_S(h) + Rad_S(\mathcal{H}) + 3\sqrt{\frac{\log 2/\delta}{2m}}$$

gen error

with Prob. $\geq 1 - \delta$.

Proof: Mcdiarmid's inequality

$$R(h) - \hat{R}_S(h) \leq Rad_S(\mathcal{H}) + 3\sqrt{\frac{\log 2/\delta}{2m}}$$

generalization gap

$$\leq \frac{\Lambda}{m} \|G\|_F + 3\sqrt{\frac{\log 2/\delta}{2m}}$$

$$\frac{\Lambda}{m} \|G\|_F = \varepsilon$$

if eigenvalues of G are $\uparrow$, $Rad(\mathcal{H}) \uparrow$.

(which converge to eigenvalues of $T_{\mathcal{K}}$ as $m \to \infty$)