Feb 23 - 25, Zoom lectures

Mar 3rd – Zoom lecture

7th – in person

10th & 12th presentations

$\rightarrow$ Regression, RKHS, statistical learning theory $\rightarrow$ Current

$\rightarrow$ kernel methods, deep learning theory

$\rightarrow$ choice of kernels + SciML

$\rightarrow$ generative modeling + kernel

} Goal

Representer theorem:

Any minimizer of the "kernel regression" problem of the form

$$\min_{h \in \mathcal{H}} \underbrace{\hat{R}_S(h)}_{\text{(empirical risk)}} + F(\|h\|_{\mathcal{H}})$$

where $F: \mathbb{R}^+ \to \mathbb{R}$ is an increasing function

has the form

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

- In other words, h has a finite-dimensional representation in $\text{span}\{\kappa(x_i, \cdot)\}_{i=1}^m$ with $\{x_i\}_{i=1}^m$ being training points

- $\hat{R}_S(h) \equiv$ empirical risk $= \frac{1}{m}\sum_{i=1}^{m} \ell((x_i, y_i), h(x_i))$

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

- Regularized ERM: $\hat{R}_S(h) + F(\|h\|_{\mathcal{H}})$

- $\mathcal{H}$: RKHS associated with PD kernel κ.

Proof:

$$h = h_0 + h_\perp$$

where $h_0 \in \text{span}\{\kappa(x_i, \cdot)\}_{i=1}^m = H_0$

$h_\perp \in H_0^\perp$

- $\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell((x_i, y_i), h(x_i))$

$= \frac{1}{m}\sum_{i=1}^{m} \ell((x_i, y_i), \langle h, \kappa(x_i, \cdot)\rangle)$ ($\because$ reproducing property)

$= \frac{1}{m}\sum_{i=1}^{m} \ell((x_i, y_i), \langle h_0, \kappa(x_i, \cdot)\rangle)$

$= \frac{1}{m}\sum_{i=1}^{m} \ell((x_i, y_i), h_0(x_i))$

$(\langle h, \kappa(x_i, \cdot)\rangle = \langle h_0 + h_\perp, \kappa(x_i, \cdot)\rangle$

$= \langle h_0, \kappa(x_i, \cdot)\rangle + \cancel{\langle h_\perp, \kappa(x_i, \cdot)\rangle})$

$\Rightarrow \hat{R}_S(h) = \hat{R}_S(h_0)$

- $F(\|h\|_{\mathcal{H}})$ is monoto. increasing

$F(\|h\|_{\mathcal{H}}^2)$ is also increasing

($F \circ$ quadratic fun. on $[0, \infty)$ is increasing)

$F(\|h\|_H^2) = F(\|h_0\|_H^2 + \|h_\perp\|_H^2)$

$(\|h\|_{\mathcal{H}}^2 = \|h_0\|_H^2 + \|h_\perp\|_{\mathcal{H}}^2$ Pythagoras thm)

$\geq F(\|h_0\|_{\mathcal{H}}^2)$

($h_\perp = 0$ for minimizer)

if F is strictly increasing, then

$F(\|h\|_{\mathcal{H}}^2) > F(\|h_0\|_{\mathcal{H}}^2)$

! solution when F is strictly increasing

Value of loss at h_0 is smaller than value of loss at h, for every $h \in \mathcal{H}$.

How to use representation thm

WKT soln of

$$\min_{h \in \mathcal{H}} \hat{R}_S(h) + F(\|h\|^2)$$

has the form

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

$$\hat{R}_S(h) = \sum_{i=1}^{m} \ell\Big((x_i, y_i), \sum_{j=1}^{m} \alpha_j \kappa(x_i, x_j)\Big)$$

$$F(\|h\|^2) = \sum_{i,j=1}^{m} \alpha_i \alpha_j \kappa(x_i, x_j)$$

Can use any other regularizer, e.g. $\frac{\lambda}{2}\|\alpha\|^2$

Typical kernel regression:

$$\min_{\alpha \in \mathbb{R}^m} MSE(\alpha) + \frac{\lambda}{2}\alpha^T G \alpha$$

$$\|h(x)\|_H^2 = \langle h(x), h(x)\rangle_H$$

$$= \Big\langle \sum_{i=1}^{m} \alpha_i \kappa(x_i, x), \sum_{j=1}^{m} \alpha_j \kappa(x_j, x)\Big\rangle_H$$

$$= \sum_{j,i=1}^{m} \alpha_i \alpha_j \langle \kappa(x_i, x), \kappa(x_j, x)\rangle_H$$

$$= \alpha^T G \alpha$$

$$\underset{m \times m}{G_{ij}} = \kappa(x_i, x_j)$$

$$MSE(\alpha) = \frac{1}{m}\sum_{i=1}^{m} (y_i - h(x_i))^2$$

$$= \frac{1}{m}\sum_{i=1}^{m} \Big(y_i - \sum_{j=1}^{m} \alpha_j \kappa(x_i, x_j)\Big)^2$$

$$= \frac{1}{m}\sum_{i=1}^{m} \Big(y_i^2 + \Big(\sum_{j=1}^{m} \alpha_j \kappa(x_i, x_j)\Big)^2 - \sum_{j=1}^{m} 2 y_i \alpha_j \kappa(x_i, x_j)\Big)$$

$$\min_{\alpha} -\frac{2}{m}\sum_{i,j=1}^{m} y_i \alpha_j \kappa(x_i, x_j) + \frac{1}{m}\sum_{i=1}^{m}\Big(\sum_{j=1}^{m} \alpha_j \kappa(x_i, x_j)\Big)^2 + \lambda \alpha^T G \alpha$$

$$\min_{\alpha} \frac{1}{m}\|Y - G\alpha\|^2 + \lambda \alpha^T G \alpha$$

$$G[i, :]\alpha = h(x_i)$$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m.$$

Linear regression for $\alpha \in \mathbb{R}^m$.

- Rmk: we are solving for function $h \in \mathcal{H}$ that takes $\mathcal{X}$ to scalars.

Inf-dim opt problem $\rightarrow$ finite-dim $\boxed{\text{convex}}$ optim. in dim = # training pts

and superficially, inde. of dim($\mathcal{X}$).

- Compare with linear reg.

$$\mathcal{H} = \{x \to w^T x : w \in \mathbb{R}^d\}$$

$$\min_{w} \frac{1}{m}\|Y - Xw\|^2 + \lambda\|w\|^2$$

$$w_{opt} = \Big(\frac{1}{m}X^T X + \lambda I\Big)^{-1} X^T Y$$

$$X \in \mathbb{R}^{m \times d} \quad \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

- Prob. opt for α:

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{m}\|Y - G\alpha\|^2 + \lambda \alpha^T G \alpha$$

Solve:

$$\nabla_\alpha\Big(\frac{1}{m}(Y - G\alpha)^T(Y - G\alpha) + \lambda \alpha^T G \alpha\Big) = 0$$

$$\Rightarrow -\frac{1}{m}(Y - G\alpha)^T G + \lambda \alpha^T G = 0$$

$$\lambda \alpha^T G = \frac{1}{m}(Y - G\alpha)^T G$$

$$\lambda \alpha^T G = \frac{1}{m}Y^T G - \frac{1}{m}\alpha^T G^T G$$

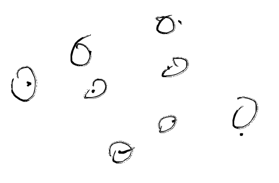
$$\alpha^T\Big(\lambda G + \frac{1}{m}G^T G\Big) = \frac{1}{m}Y^T G$$

$$\Big(\frac{1}{m}G^T G + \lambda G\Big)\alpha = \frac{1}{m}G Y$$

$$\alpha = \Big(\frac{1}{m}G^T G + \lambda G\Big)^{-1}\frac{1}{m}G Y$$

- why ~~are~~ is kernel regression effective?

o/p: $h(x) = \sum_{i=1}^{m} \alpha_i \kappa(x_i, x)$



Feature maps

RKHS feature map :

$$\boxed{x \rightarrow \kappa(x, \cdot)}$$

$$\phi^{RKHS}(x) = \kappa(x, \cdot)$$

Mercer map

$$x \rightarrow \{\sqrt{\lambda_j}\ \psi_j(x)\}_{i \in \mathbb{N}}$$

where $\lambda_j \in \mathbb{R}^+$ were eigenvalues of T_κ (H-S operator)

and $\psi_j \in L^2(\mathcal{X}, \mu)$ are corresponding eigenfunctions

$$T_\kappa f(x) = \int f(y)\ \kappa(x, y)\ d\mu(y)$$

Last time : T_κ is compact on $L^2(\mathcal{X}, \mu)$ and when κ is sym PD kernel, $\lambda_i \in \mathbb{R}^+$

$$\kappa(x, x') = \sum_{j \in \mathbb{N}} \lambda_j\ \psi_j(x)\ \psi_j(x')$$

Want features s.t.

$$\kappa(x, x') = \sum_{j=0}^{\infty} \sqrt{\lambda_j}\psi_j(x) \cdot \sqrt{\lambda_j}\ \psi_j(x')$$

$$= \langle \phi^{mer}(x), \phi^{mer}(x') \rangle_{\ell^2}$$

$$\phi^{mer}(x) = \{\sqrt{\lambda_j}\ \psi_j(x)\}_{j \in \mathbb{N}}$$

Recall : finite-dim. Mercer map (over fixed data pts)

( ) $\{\sqrt{\lambda_j}\ v_j\}_{j \in [m]}$ v_j were rows of eigenvectors of Gram matrix

λ_j are eigenvalues of Gram matrix

$$\kappa(x, x') = \langle \phi^{mer}(x), \phi^{mer}(x') \rangle_{\ell^2(\mathcal{X})}$$

$$= \langle \phi^{RKHS}(x), \phi^{RKHS}(x') \rangle_{\mathcal{H}}$$

If eigenvalues of T_κ decay rapidly,

$$\phi^{mer}(x) = [\sqrt{\lambda_1}\ \psi_1(x), \sqrt{\lambda_2}\ \psi_2(x) \ldots]$$

Even if $\sup_{x \in \mathcal{X}} |\psi_j(x)| < \infty$ is false,

$\sup_{x \in \mathcal{X}} |\sqrt{\lambda_j}\ \psi_j(x)| < \infty$ can be true

Then, we may think of $\mathcal{H}$ as a finite-dim. space

Even if $\mathcal{H}$ is inf-dime., regularization with $\|\ \|_{\mathcal{H}}$ makes opt prob finite dimensional

Empirical kernel map

$$x \rightarrow [\ \kappa(x_1, x), \ldots\ \kappa(x_m, x)]$$

$$\in \mathbb{R}^m$$

(dis. of RKHS map)

- Generalization bounds based on Rademacher complexity
(Next time; prepare by reading bounds for finite hypothesis classes)

- Rademacher comp - one way of measuring size of hypothesis class.