Reproducing Kernel Hilbert Space

Hwiyoung Lee

Department of Statistics, Florida State University

Overview

Nonparametric Regression

- Standard non parametric regression model :

$$Y_i = f(X_i) + \varepsilon_i, \qquad \varepsilon_i = \sigma w_i \sim N(0, \sigma^2)$$

- Estimate f by $\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - f(X_i))^2$

- Interpolating ? $\rightarrow$ Regularization !

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda P(f)$$

- Equivalent to choose $f \in \mathcal{F}$ minimize $\sum_{i=1} (Y_i - f(X_i))^2$, where $\mathcal{F} = \{f : P(f) \leq R\}$ for some $R > 0$

- Typically, $\mathcal{F}$ is a compact subset of some ambient function class $\mathcal{G}$
  ex : a ball of radius R in some norm $||\cdot||_{\mathcal{G}}$

$$\hat{f} = \underset{f \in \mathcal{G}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda ||f||_{\mathcal{G}}^2$$

Nonparametric Regression

Tikhnov Regularization

$$\frac{1}{n}\sum_{i=1}^{n} V\left(f(X_i), Y_i)\right) + \lambda||f||_{\mathcal{G}}^2 \qquad , \ \lambda > 0$$

- $V(f(x), y)$ is the loss function
- $||\cdot||_{\mathcal{G}}$ is the norm in the function space $\mathcal{G}$
- Powerful : does not need a specific algorithm, present a large class of algorithms
- General : By choosing V and $\mathcal{H}$ differently, we can derive many statistical methods, including linear regression and SVM
- We want to construct $\mathcal{G}$ so that it contains smooth functions
- RKHS is a good choice

Example 1 : Linear regression

For a given vector $\beta \in \mathbb{R}^d$, define the function $f_\beta(x) = \langle \beta, x \rangle$

For a compact set $\mathcal{C} \subset \mathbb{R}^d$, define

$$\mathcal{F}_\mathcal{C} = \{f_\beta : \mathbb{R}^d \to \mathbb{R} \mid \beta \in \mathcal{C}\}$$

Then the constrained lest square problem :

$$\underset{\beta \in \mathcal{C}}{\operatorname{argmin}} \frac{1}{2n} ||y - X\beta||_2^2$$

- Examples
  - Ridge : $\mathcal{C} = \{\beta \in \mathbb{R}^d \mid ||\beta||_2^2 \leq R\}$

$$\underset{\beta \in \mathcal{C}}{\operatorname{argmin}} \frac{1}{2n} ||y - X\beta||_2^2 + \lambda ||\beta||_2^2$$

  - Lasso : $\mathcal{C} = \{\beta \in \mathbb{R}^d \mid ||\beta||_1 \leq R\}$

$$\underset{\beta \in \mathcal{C}}{\operatorname{argmin}} \frac{1}{2n} ||y - X\beta||_2^2 + \lambda ||\beta||_1$$

Example 2 : Support Vector Machine

Suppose $Y_i \in \{-1, 1\}$

Linear SVM

$$\underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[1 - Y_i(\beta_0 + \beta^\top X_i)\right]_+ + \frac{\lambda}{2} ||\beta||_2^2$$

RKHS version SVM

$$\underset{f \in \mathcal{H}}{\operatorname{argmin}} \sum_{i=1}^{n} [1 - Y_i f(X_i)]_+ + \frac{\lambda}{2} ||f||_{\mathcal{H}}^2$$

Example 3 : Kernel Ridge Regression

- Let $\mathcal{H}$ be the Hilbert space equipped with the norm $||\cdot||_{\mathcal{H}}$
- For some radius $R > 0$, consider the constrained least square estimator

$$\hat{f} \in \underset{||f||_{\mathcal{H}} \leq R}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

- Dual form, the penalized least square estimator

$$\hat{f} \in \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{2n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda ||f||_{\mathcal{H}}^2$$

- In particular, we assume $\mathcal{H}$ to be a RKHS

Example 4 : Cubic Spline

- For some radius $R > 0$, consider the class of twice continuously differentiable functions $f : [0,1] \to \mathbb{R}$,

$$\mathcal{F}(R) := \{f : [0,1] \to \mathbb{R} \mid \int_0^1 (f''(x))^2 dx \leq R\}$$

- This constraint can be understood as a Hilbert norm bound in a second-order Sobolev space
- The penalized non-parametric least squares estimates is given by

$$\hat{f} \in \operatorname*{argmin}_f \frac{1}{2n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \int_0^1 (f''(x))^2 dx$$

- The minimizer f is a cubic spline
- As $R \to 0$, the cubic spline fit $\hat{f}$ becomes a liner function

Norm

$\mathcal{F}$: vector space over the field $\mathbb{K}$

Definition

Nonnegative function $||\cdot||_{\mathcal{F}} : \mathcal{F} \to \mathbb{K}$ is said to be a Norm on $\mathcal{F}$ if $\forall f, g \in \mathcal{F}, \alpha \in \mathbb{K}$

1. $||f||_{\mathcal{F}} = 0$ iff $f = 0$
2. $||f+g||_{\mathcal{F}} \leq ||f||_{\mathcal{F}} + ||g||_{\mathcal{F}}$
3. $||\lambda f||_{\mathcal{F}} = |\lambda| ||f||_{\mathcal{F}}$

- The norm induces a metric : $d(f,g) = ||f-g||_{\mathcal{F}}$
- Examples
  - $\mathcal{F} = \mathbb{R}^d$: $||X||_p = \left(\sum_{i=1}^d |X_i|^p\right)^{1/p}$, where $p \geq 1$
  - $\mathcal{F} = L_p$: $||X||_p = \left(\int |f(x)|^p d\mu\right)^{1/p}$, where $p \geq 1$

Note : $||f||_\infty = \sup_x |f(x)|$, and $||\cdot||_\infty = \lim_{p\to\infty} ||\cdot||_p$

Inner Product

Definition

A function $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \to \mathbb{K}$ is said to be an Inner product on $\mathcal{F}$

1. $\langle f, f \rangle \geq 0$, and $\langle f, f \rangle = 0$ iff $f = 0$
2. $\langle f, g \rangle = \langle g, f \rangle$
3. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle = \alpha_1 \langle f_1, g \rangle + \alpha_2 \langle f_2, g \rangle$

- Norm induced by the inner product : $||f||_{\mathcal{F}} = \langle f, f \rangle_{\mathcal{F}}^{1/2}$
- Some useful relations between norm and inner product
  1. $|\langle f, g \rangle| \leq ||f|| \cdot ||g||$
  2. $||f + g||^2 + ||f - g||^2 = 2||f||^2 + 2||g||^2$
  3. $4\langle f, g \rangle = ||f + g||^2 - ||f - g||^2$
- Examples
  - $\mathcal{F} = \mathbb{R}^d : \langle x, y \rangle = \sum x_i y_i$
  - $\mathcal{F} = \mathcal{C}([a, b]) : \langle f, g \rangle = \int_a^b f(x) g(x) dx$
  - $\mathcal{F} = \mathbb{R}^{d \times d} : \langle A, B \rangle = Tr(AB^\top)$

Cauchy Sequence

Definition

$\{f_n\}_{n=1}^{\infty}$ of $(\mathcal{F}, ||\cdot||_{\mathcal{F}})$ is said to be a Cauchy sequence if for every $\epsilon > 0$, $\exists N_\epsilon \in \mathbb{N}$ s.t $\forall n, m \geq N_\epsilon, ||f_n - f_m||_{\mathcal{F}} < \epsilon$

- Cauchy sequence is always bounded
- Convergent sequence is a Cauchy sequence

$$||f_n - f_m|| \leq ||f_n - f|| + ||f - f_m||$$

- Not every Cauchy sequence converges

  $1, 1.4, 1.414, 1.4142, \cdots$ is a cauchy seq in $\mathbb{Q}$, but does not converge ($\sqrt{2} \notin \mathbb{Q}$)

Complete space

A space $\mathcal{F}$ is Complete if every cauchy sequence in $\mathcal{F}$ converges

- complete + norm = Banach space
- complete + inner product = Hilbert space

Exmple : L_p space

Let be $(X, \mathcal{A}, \mu)$ a measure space and $1 \leq p < \infty$. Then the L_p space consist of measurable functions $f : X \to \mathbb{R}$ such that

$$\int_X |f|^p d\mu < \infty$$

- L_p norm is defined by

$$||f||_p = \left(\int_X |f|^p d\mu \right)^{1/p}$$

- L_p is a Banach space $(1 \leq p \leq \infty)$
  - For $p = \infty$,

$$||f||_\infty := \operatorname*{ess\,sup}_X |f| = \inf \{C \geq 0 : |f(x)| \leq C \text{ for } a.e\}$$

- L_2 is a Hilbert space

$$\langle f, g \rangle = \int_X f(x)g(x)d\mu$$

Exmple : Hölder space

Let f be a function defined on $\Omega \subset \mathbb{R}^d$, $0 < \alpha \leq 1$, and $k \in \mathbb{Z}^+$

Definition

Functions whose kth order derivatives are Hölder continuous with α

$$\mathcal{C}^{k,\alpha} = \left\{ f \in C^k(\Omega) \mid [D^k f]_{\mathcal{C}^{0,\alpha}} < \infty \right\}$$

- Hölder continuous : $[f]_{\mathcal{C}^{0,\alpha}} := \sup_{x \neq y \in \Omega} \dfrac{|f(x) - f(y)|}{|x-y|^\alpha} < \infty$
- $\mathcal{C}^{0,1}$: Lipschitz space, $\mathcal{C}^{1,1}$: Bounded second derivatives
- Hölder space is a Banach space

$$||f||_{\mathcal{C}^{k,\alpha}(\Omega)} := ||f||_{\mathcal{C}^k} + \max_{|\beta|=k} [D^\beta f]_{\mathcal{C}^{0,\alpha}}$$

where $||f||_{\mathcal{C}^k} = \max_{|\beta| \leq k} \sup_{x \in \Omega} |D^\beta f(x)|$

- $\mathcal{C}^{k,\alpha} = \left\{ f \in C^k(\Omega) \mid ||f||_{\mathcal{C}^{k,\alpha}(\Omega)} < \infty \right\}$

Exmple : Sobolev space

Definition

The sobolev space of order m is defined by

$$W_{m,p} = \left\{ f \in L_p(\Omega) \ : \ D^j f \in L_p(\Omega), \ |j| \leq m \right\}$$

where $D^j f$ is a jth weak derivative of f

- $W_{m,p}$ is defined to be the subset of L_p s.t. f and weak derivatives upto order m have a finite L_p norm
- $W_{m,p}$ is a Banach space

$$||f||_{W_{m,p}} = \left(\sum_{|j| \leq m} \int_\Omega |D^j f|^p dx \right)^{1/p}, \quad p \in [1, \infty)$$

$$||f||_{W_{m,\infty}} = \sum_{|j| \leq m} \operatorname*{esssup}_{x \in \Omega} |D^j f|$$

- $W_m = W_{m,2}$ is a Hilbert space

$$\langle f, g \rangle = \sum_{|j| \leq m} \int_\Omega D^j f D^j g dx$$

Reference

Bhattacharya, R., and Patrangenaru, V. (2003). Large Sample Theory of Intrinsic and Extrinsic Sample Means on Manifolds. I. *Annals of Stat* **31** : 1-29

Sirovich, L., and Kirby, M. (1987) Low-dimensional procedure for the characterization of human faces. *J. Optical Society of Amer. A.* **4** : 519-524

Lin, L., Thomas, B. st., Zhu, H., and Dunson, D. (2017). Extrinsic Local Regression on Manifold-Valued Data *J. Amer. Statist. Assoc.* **112** : 1261-1273