Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 24

Yun Yang

- Non-parametric least squares

Estimation via constrained least-squares

- Standard non-parametric regression model:

$$y_i = f^*(x_i) + v_i, \quad v_i = \sigma w_i \sim \mathcal{N}(0, \sigma^2), \quad \text{for } i = 1, \ldots, n.$$

- We consider estimate f^* is by constrained least-squares

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \right\},$$

  where $\mathcal{F}$ is a suitably chosen subset of functions.

- Typically, we choose $\mathcal{F}$ to be a compact subset of some ambient function class $\mathcal{G}$, for example, a ball of radius R in some norm $\|\cdot\|_{\mathcal{G}}$.

- For computational reasons, it can be convenient to use regularized estimators of the form

$$\widehat{f} \in \operatorname*{argmin}_{f \in \mathcal{G}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{G}}^2 \right\}.$$

Example: Linear regression

- For a given vector $\theta \in \mathbb{R}^d$, define the function

$$f_\theta(x) = \langle \theta, x \rangle.$$

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

$$\mathcal{F}_\mathcal{C} = \left\{ f_\theta : \mathbb{R}^d \to \mathbb{R} \,\middle|\, \theta \in \mathcal{C} \right\}.$$

- Constrained least-square:

$$\widehat{\theta} \in \operatorname*{argmin}_{\theta \in \mathcal{C}} \left\{ \frac{1}{2n} \|y - X\theta\|_2^2 \right\}.$$

- Examples:
  - *Ridge regression*: $\mathcal{C} = \{\theta \in \mathbb{R}^d \mid \|\theta\|_2^2 \leq R_2\}$.
  - *Lasso*: $\mathcal{C} = \{\theta \in \mathbb{R}^d \mid \|\theta\|_1 \leq R_1\}$.

Example: Cubic smoothing spline

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

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

- This constraint can be understood as a Hilbert norm bound in a second-order Sobolev space.
- For this function class, the penalized non-parametric least squares estimate is given by

$$\widehat{f} \in \arg\min_{f} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \int_0^1 (f''(x))^2 \, dx \right\}.$$

- It can be shown that any minimizer f is a cubic spline.
- In the limit as $R \to 0$, the cubic spline fit $\widehat{f}$ becomes a linear function.

Example: Kernel ridge regression

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

$$\widehat{f} \in \underset{\|f\|_{\mathbb{H}} \leq R}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \right\}.$$

- In practice, its dual form, the penalized least-square estimator is commonly used

$$\widehat{f} \in \underset{f \in \mathbb{H}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathbb{H}}^2 \right\}.$$

- In particular, we assume $\mathbb{H}$ to be a *Reproducing kernel Hilbert space* (RKHS).

Reproducing kernel Hilbert space (RKHS)

Definition

A reproducing kernel Hilbert space $\mathbb{H}$ is a Hilbert space of real-valued functions on $\mathcal{X}$ such that for each $x \in \mathcal{X}$, the evaluation functional $L_x : \mathbb{H} \to \mathbb{R}$, $f \mapsto f(x)$ is bounded.

- When L_x is a bounded linear functional, the Riesz representation implies that there must exist some element R_x of the Hilbert space $\mathbb{H}$ such that

$$f(x) = L_x(f) = \langle f, R_x \rangle_{\mathbb{H}} \quad \text{for all } f \in \mathbb{H}.$$

- This element R_x of $\mathbb{H}$ is known as the representer of evaluation at $x \in \mathcal{X}$.
- The boundedness of R_x ensures that convergence of a sequence of functions in an RKHS implies pointwise convergence.

Examples

- The space of all linear functions $f_\beta(\cdot) = \langle \cdot, \beta \rangle$ over $\mathbb{R}^m$ under the inner product

$$\langle f_\beta, f_{\beta'} \rangle_{\mathbb{H}} = \langle \beta,\ \beta' \rangle.$$

  is an RKHS, whose representer of evaluation R_x is the function $R_x(z) = \langle x,\ z \rangle$.

- The space $\mathcal{L}^2[0, 1]$ is not an RKHS.
- The first order Sobolev space $\mathbb{H}^1[0, 1] =$

$$\big\{ f : f(0) = 0,\ f \text{ is absolutely continuous with } f' \in L^2[0, 1] \big\}$$

  is an RKHS under the inner product

$$\langle f_1, f_2 \rangle_{\mathbb{H}} = \int_0^1 f_1'(x) f_2'(x)\, dx,$$

  whose representer of evaluation R_x is the function $R_x(z) = \min\{x,\ z\}$.

Examples

- More generally, consider the higher-order Sobolev space $\mathbb{H}^\alpha[0,1]$ of real-valued functions on $[0,1]$ that are α-times differentiable (almost everywhere), with $f^{(\alpha)}$ being Lebesgue integrable, and $f(0) = \cdots = f^{(\alpha)} = 0$.
- Define the inner-product

$$\langle f_1, f_2 \rangle_{\mathbb{H}} = \int_0^1 f_1^{(\alpha)}(x) f_2^{(\alpha)}(x)\, dx.$$

- $\mathbb{H}^\alpha[0,1]$ is an RKHS, and the representer of evaluation is

$$R_x(y) = \int_0^1 \frac{(x-z)_+^{\alpha-1}}{(\alpha-1)!} \frac{(y-z)_+^{\alpha-1}}{(\alpha-1)!}\, dz.$$

- This can been seen from the Taylor expansion formula

$$f(x) = \sum_{\ell=0}^{\alpha-1} f^{(\ell)}(0) \frac{x^\ell}{\ell!} + \int_0^1 f^{(\alpha)}(z) \frac{(x-z)_+^{\alpha-1}}{(\alpha-1)!}\, dz.$$

Kernel functions

Definition

Positive semidefinite kernel function A symmetric bivariate function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is positive semidefinite (PSD) if for all integers $n \geq 1$ and elements $\{x_i\}_{i=1}^n \subset \mathcal{X}$, the $n \times n$ matrix K with elements $K_{ij} := K(x_i, x_j)$ is positive semidefinite.

PSD kernel from representer of evaluation

- Define $\mathcal{K}$ via $\mathcal{K}(x, x') = \langle R_x, R_{x'} \rangle_{\mathbb{H}}$.
- $\mathcal{K}$ is symmetric.
- $\mathcal{K}$ is PSD: for any vector $\alpha \in \mathbb{R}^n$, we have

$$\alpha^T K \alpha = \sum_{j,k} \alpha_j \alpha_k \, \mathcal{K}(x_j, x_k) = \left\langle \sum_j \alpha_j R_{x_j}, \sum_k \alpha_k R_{x_k} \right\rangle_{\mathbb{H}}$$

$$= \left\| \sum_j \alpha_j R_{x_j} \right\|_{\mathbb{H}}^2 \geq 0.$$

Reproducing kernel

- Indeed, for any $x \in \mathbb{H}$, the object $\mathcal{K}(\cdot, x)$ can be identified with R_x as an element of our Hilbert space, since

$$\mathcal{K}(x', x) = \langle R_{x'}, R_x \rangle_{\mathbb{H}} = R_x(x'), \quad \text{for all } x \in \mathcal{X}.$$

- Consequently, the function $\mathcal{K}(\cdot, x)$ satisfies the *reproducing kernel property*, namely,

$$\langle \mathcal{K}(\cdot, x), f \rangle_{\mathbb{H}} = f(x), \quad \text{for all } f \in \mathbb{H}.$$

- Conversely, given a positive semidefinite kernel function $\mathcal{K}$, we can define an associated function space

$$\widetilde{\mathbb{H}} := \Big\{ f(\cdot) = \sum_{i=1}^{n} \alpha_i \, \mathcal{K}(\cdot, x_i) \, : \, n \in \mathbb{N} \text{ and } \{x_i\}_{i=1}^{n} \subset \mathcal{X} \Big\}.$$

Reproducing kernel

- Given two functions $f, \bar{f} \in \widetilde{\mathbb{H}}$, where $\bar{f} = \sum_{i=1}^{\bar{n}} \bar{\alpha}_i \, \mathcal{K}(\cdot, \bar{x}_i)$, their inner product is defined by

$$\langle f, \bar{f} \rangle_{\widetilde{\mathbb{H}}} = \sum_{j=1}^{n} \sum_{k=1}^{\bar{n}} \alpha_j \bar{\alpha}_k \, \mathcal{K}(x_j, \bar{x}_k).$$

- The PSD property of $\mathcal{K}$ implies $\|f\|_{\widetilde{\mathbb{H}}} \geq 0$, for all $f \in \widetilde{\mathbb{H}}$.
- We can define a Hilbert space $\mathbb{H}$ as the completion of $\widetilde{\mathbb{H}}$ with respect to the inner product $\| \cdot \|_{\widetilde{\mathbb{H}}}$.

Theorem

Given any reproducing kernel Hilbert space $\mathbb{H}$ of functions $f : \mathcal{X} \to \mathbb{R}$, there exists a unique positive semidefinite kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Conversely, given any positive semidefinite kernel function $\mathcal{K}$, we can define an RKHS in which $\mathcal{K}$ acts as the representer of evaluation.

Example: Kernel ridge regression, continued

- Recall the KRR estimator

$$\widehat{f} \in \arg\min_{f \in \mathbb{H}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathbb{H}}^2 \right\}.$$

- Let $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote the reproducing kernel associated with the RKHS $\mathbb{H}$.

Theorem (Representer theorem)

Any solution $\widehat{f}$ of the KRR optimization problem takes the form

$$\widehat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i \, \mathcal{K}(\cdot, x_i).$$

Example: Kernel ridge regression, continued

- Define the empirical kernel matrix $K \in \mathbb{R}^{n\times n}$, with $K_{ij} = n^{-1}\,\mathcal{K}(x_i, x_j)$, and recall

$$\widehat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i \, \mathcal{K}(\cdot, x_i).$$

- Then, we can write

$$(\widehat{f}(x_1), \ldots, \widehat{f}(x_n))^T = \sqrt{n}\, K\, \widehat{\alpha},$$

where $\widehat{\alpha} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)^T$.

- Solving the KRR optimization problem is inequivalent to solving the following quadratic programming

$$\widehat{\alpha} \in \operatorname*{argmin}_{\alpha\in\mathbb{R}^n} \left\{ \frac{1}{2n} \|y - \sqrt{n}\, K\, \alpha\|_2^2 + \lambda_n \underbrace{\alpha^T K \alpha}_{\|f\|_{\mathbb{H}}^2} \right\}.$$

Example: Convex regression

- Now suppose $f^* : \mathcal{C} \to \mathbb{R}$ is known to be a convex function over its domain $\mathcal{C}$, where $\mathcal{C}$ is some convex and open subset of $\mathbb{R}^d$.
- It is natural to consider the least-squares estimator with a convexity constraint,

$$\widehat{f} \in \underset{f \text{ is convex}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \right\}.$$

- Although this optimization problem is infinite-dimensional, we can convert it to an equivalent finite-dimensional problem.
- The convexity constraint implies there exist sub-gradient vectors $\{\widetilde{z}_i\}_{i=1}^n$, such that for all $i = 1, \ldots, n$,

$$f(x) \geq f(x_i) + \langle \widetilde{z}_i, x - x_i \rangle \quad \text{for all } x \in \mathcal{C}.$$

Example: Convex regression

- Since the cost function depends only on the values $\widetilde{y}_i = f(x_i)$, the optimum does not depend on the function behavior elsewhere.
- It suffices to solve the optimization problem

$$\min_{\{(\widetilde{y}_i, \widetilde{z}_i)\}_{i=1}^n} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \widetilde{y}_i)^2$$

such that $\quad \widetilde{y}_j \geq \widetilde{y}_i + \langle \widetilde{z}_i, x_j - x_i \rangle \quad$ for all $i, j = 1, \ldots, n$.

- An optimal solution $\{(\widehat{y}_i, \widehat{z}_i)\}_{i=1}^n$ can be used to define an estimate $\widehat{f} : \mathcal{C} \to \mathbb{R}$ via

$$\widehat{f}(x) = \max_{i=1,\ldots,n} \left\{ \widehat{y}_i + \langle \widehat{z}_i, x - x_i \rangle \right\}.$$

- $\widehat{f}$ is convex, and by the feasibility of the solution $\{(\widehat{y}_i, \widehat{z}_i)\}_{i=1}^n$, we are guaranteed that $\widehat{f}(x_i) = \widehat{y}_i$.