Prof. Yiping Lu
IEMS 402 Statistical Learning
November 3, 2024

Homework 8: Reproducing Kernel Hilbert Space/Robust Learning

Question 1. (Hilbert Embedding of Probability) Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel with associated RKHS $\mathcal{H}$. Assume that $\mathcal{X}$ is compact. We call k *universal* if it is dense in $C(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$. That is, for any $\epsilon > 0$ and any continuous function $f : \mathcal{X} \to \mathbb{R}$, there exists a function $h \in \mathcal{H}$ such that $\sup_{x\in\mathcal{X}} |f(x) - h(x)| < \epsilon$.

Define $\varphi(x) = k(\cdot, x)$. (Thus $k(x, z) = \langle\varphi(x), \varphi(z)\rangle$, and $\varphi(x)$ is the representer of evaluation at x, i.e., $\langle h, \varphi(x)\rangle = h(x)$ for all $h \in \mathcal{H}$.) Let $\mathcal{P}$ be the collection of distributions on $\mathcal{X}$ for which $\mathbb{E}_P[\sqrt{k(X, X)}] < \infty$.

(a) Using the Riesz representation theorem for Hilbert spaces, argue that the mean mapping $\mu(P) := \mathbb{E}_P[\varphi(X)]$ exists and is a vector in $\mathcal{H}$. *Hint*: Letting $\|\cdot\|$ denote the norm on $\mathcal{H}$, the Riesz representation theorem for Hilbert spaces says that if $L : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional, meaning that $L(f) \le C \cdot \|f\|$ for some constant C, then there exists some $h_L \in \mathcal{H}$ such that $L(f) = \langle h_L, f\rangle$ for all $f \in \mathcal{H}$.

(b) Assume that $\mathcal{X}$ is compact and that k is universal. Show that the mean embedding

$$P \mapsto \mathbb{E}_P[\varphi(X)] = \int_{\mathcal{X}} \varphi(x)dP(x)$$

is one-to-one, that is, if $P \neq Q$ then $\mathbb{E}_P[\varphi(X)] \neq \mathbb{E}_Q[\varphi(X)]$.

(c) For distributions P and Q, show that

$$\sup_{f\in\mathcal{H},\|f\|\le 1} \{\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]\} = \sqrt{\mathbb{E}[k(X, X')] + \mathbb{E}[k(Z, Z')] - 2\mathbb{E}[k(X, Z)]},$$

where $X, X' \overset{i.i.d}{\sim} P$ and $Z, Z' \overset{i.i.d}{\sim} Q$.

Question 2. (Example of Kernel)

- Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a valid kernel function. Define

$$k_{\text{norm}}(x, z) := \frac{k(x, z)}{\sqrt{k(x, x)}\sqrt{k(z, z)}}.$$

  Is k_{norm} a valid kernel? Justify your answer.
- Consider the class of functions

$$\mathcal{H} := \{f : f(0) = 0, f' \in L^2([0, 1])\},$$

  that is, functions $f : [0, 1] \to \mathbb{R}$ with $f(0) = 0$ that are almost everywhere differentiable, where

$$\int_0^1 (f'(x))^2 dx < \infty.$$

  On this space of functions, we define the inner product by

$$\langle f, g\rangle = \int_0^1 f'(x)g'(x)dx.$$

  Show that $k(x, z) = \min\{x, z\}$ is the reproducing kernel for $\mathcal{H}$, so that it is (i) positive semidefinite and (ii) a valid kernel.

  (*My understanding:* By integral by parts, we have $\langle f, g\rangle_{\mathcal{H}} = \langle f, \Delta g\rangle_{\mathcal{L}_2}$ and $\Delta k(\cdot, z) = \delta_z$.)
- Consider the Sobolev space $\mathcal{F}_k$, which is defined as the set of functions that are $(k-1)$-times differentiable and have kth derivative almost everywhere on $[0, 1]$, where the kth derivative is square-integrable. That is, we define

$$\mathcal{F}_k := \left\{f : [0, 1] \mid f^{(k)}(x) \in L^2([0, 1])\right\}.$$

We define the inner product on $\mathcal{F}_k$ by

$$\langle f, g\rangle = \sum_{i=0}^{k-1} f^{(i)}(x)g^{(i)}(x) + \int_0^1 f^{(k)}(x)g^{(k)}(x)\,dx.$$

(a) Find the representer of evaluation for this Hilbert space, that is, find a function $r_x : [0,1] \to \mathbb{R}$ (defined for each $x \in [0,1]$) such that $r_x \in \mathcal{F}_k$ and

$$\langle r_x, f\rangle = f(x)$$

for all x.

(b) What is the reproducing kernel $k(x,z)$ associated with this space? (Recall that $k(x,z) = \langle r_x, r_z\rangle$ for an RKHS.)

Question 3. (φ-divergence DRO and Variance Regularization) Let $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be a convex function with $\varphi(1) = 0$. Then the φ-divergence between distributions P and Q defined on a space $\mathcal{X}$ is

$$D_\varphi(P\|Q) = \int \varphi\left(\frac{dP}{dQ}\right) dQ = \int_{\mathcal{X}} \varphi\left(\frac{p(x)}{q(x)}\right) q(x)d\mu(x),$$

where μ is any measure for which $P, Q \ll \mu$, and $p = \frac{dP}{d\mu}$, $q = \frac{dQ}{d\mu}$. Throughout this paper, we use $\varphi(t) = \frac{1}{2}(t-1)^2$, which gives the χ^2-divergence [45]. Given φ and a sample $X_1, \ldots, X_n$, we define the local neighborhood of the empirical distribution with radius ρ by

$$\mathcal{P}_n := \left\{\text{distributions } P \text{ such that } D_\varphi\left(P\|\hat{P}_n\right) \le \frac{\rho}{n}\right\},$$

where $\hat{P}_n$ denotes the empirical distribution of the sample, and our choice of $\varphi(t) = \frac{1}{2}(t-1)^2$ means that $\mathcal{P}_n$ consists of discrete distributions supported on the sample $\{X_i\}_{i=1}^n$. We then define the robustly regularized risk

$$R_n(\theta, \mathcal{P}_n) := \sup_{P \in \mathcal{P}_n} \mathbb{E}_P[\ell(\theta, X)] = \sup_P \left\{\mathbb{E}_P[\ell(\theta, X)] : D_\varphi(P\|\hat{P}_n) \le \frac{\rho}{n}\right\}.$$

Using convex duality please show that

$$R_n(\theta, \mathcal{P}_n) = \mathbb{E}_{\hat{P}_n}[\ell(\theta, X)] + \sqrt{\frac{2\rho}{n}\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2]}.$$

You can assume strong duality holds.

Further Reading: Connection between adversarial training and Wasserstein DRO `https://arxiv.org/abs/1710.10571`

Consider the DRO problem defined as:

$$R_n(\theta, \mathcal{P}_n) := \sup_{P \in \mathcal{P}_n} \mathbb{E}_P[\ell(\theta, X)] = \sup_P \left\{\mathbb{E}_P[\ell(\theta, X)] \,\middle|\, D_\varphi(P\|\hat{P}_n) \le \frac{\rho}{n}\right\},$$

where:

- θ represents the decision variables.
- $\mathcal{P}_n$ is the ambiguity set of probability distributions.
- $\ell(\theta, X)$ is the loss function.
- $D_\varphi(P\|\hat{P}_n)$ denotes the φ-divergence between distribution P and the empirical distribution $\hat{P}_n$.
- ρ controls the size of the uncertainty set.

Dual Formulation. To derive the dual, we utilize the definition of φ-divergence and convex duality. The φ-divergence is given by:

$$D_\varphi(P\|\hat{P}_n) = \int \varphi\left(\frac{dP}{d\hat{P}_n}(x)\right) d\hat{P}_n(x),$$

where $\varphi : \mathbb{R}_+ \to \mathbb{R}$ is a convex function satisfying $\varphi(1) = 0$.

Lagrangian Dualization. We can express the constrained optimization problem as its Lagrangian:

$$R_n(\theta, \mathcal{P}_n) = \sup_P \left\{ \mathbb{E}_P[\ell(\theta, X)] - \lambda \left(D_\varphi(P \| \hat{P}_n) - \frac{\rho}{n} \right) \right\},$$

where $\lambda \geq 0$ is the dual variable (Lagrange multiplier).

Substituting the expression for φ-divergence:

$$R_n(\theta, \mathcal{P}_n) = \sup_P \left\{ \int \ell(\theta, x) dP(x) - \lambda \left(\int \varphi \left(\frac{dP}{d\hat{P}_n}(x) \right) d\hat{P}_n(x) - \frac{\rho}{n} \right) \right\}.$$

Assuming that P is absolutely continuous with respect to $\hat{P}_n$, let $r(x) = \frac{dP}{d\hat{P}_n}(x)$. Then, the problem becomes:

$$R_n(\theta, \mathcal{P}_n) = \sup_{r(x) \geq 0} \left\{ \int \ell(\theta, x) r(x) d\hat{P}_n(x) - \lambda \left(\int \varphi(r(x)) d\hat{P}_n(x) - \frac{\rho}{n} \right) \right\}.$$

Rearranging terms:

$$R_n(\theta, \mathcal{P}_n) = \lambda \frac{\rho}{n} + \sup_{r(x) \geq 0} \int \left(\ell(\theta, x) r(x) - \lambda \varphi(r(x)) \right) d\hat{P}_n(x).$$

Optimizing over $r(x)$. For each x, the inner supremum can be solved independently:

$$\sup_{r(x) \geq 0} \left\{ \ell(\theta, x) r(x) - \lambda \varphi(r(x)) \right\}.$$

Define the convex conjugate (Legendre-Fenchel transform) of $\lambda\varphi(r)$ as:

$$\varphi_\lambda^*(s) = \sup_{r \geq 0} \left\{ sr - \lambda \varphi(r) \right\}.$$

Thus, the dual problem becomes:

$$R_n(\theta, \mathcal{P}_n) = \lambda \frac{\rho}{n} + \int \varphi_\lambda^* \left(\ell(\theta, x) \right) d\hat{P}_n(x).$$

Choice of φ-Divergence. To obtain a variance regularization, we choose the φ-divergence corresponding to the chi-squared divergence, which is defined as:

$$\varphi(r) = \frac{1}{2}(r-1)^2.$$

Its convex conjugate is:

$$\varphi_\lambda^*(s) = \sup_{r \geq 0} \left\{ sr - \lambda \cdot \frac{1}{2}(r-1)^2 \right\}.$$

To compute $\varphi_\lambda^*(s)$, take the derivative with respect to r and set it to zero:

$$\frac{d}{dr} \left(sr - \frac{\lambda}{2}(r-1)^2 \right) = s - \lambda(r-1) = 0 \implies r = 1 + \frac{s}{\lambda}.$$

Substituting back:

$$\varphi_\lambda^*(s) = s \left(1 + \frac{s}{\lambda} \right) - \frac{\lambda}{2} \left(\frac{s}{\lambda} \right)^2 = s + \frac{s^2}{\lambda} - \frac{s^2}{2\lambda} = s + \frac{s^2}{2\lambda}.$$

Thus:

$$\varphi_\lambda^*(s) = s + \frac{s^2}{2\lambda}.$$

Expressing as Variance Regularization. Substituting $\varphi_\lambda^*(s)$ back into the dual formulation:

$$R_n(\theta, \mathcal{P}_n) = \lambda \frac{\rho}{n} + \int \left(\ell(\theta, x) + \frac{\ell(\theta, x)^2}{2\lambda} \right) d\hat{P}_n(x).$$

Simplifying, we get:

$$R_n(\theta, \mathcal{P}_n) = \lambda \frac{\rho}{n} + \mathbb{E}_{\hat{P}_n}[\ell(\theta, X)] + \frac{1}{2\lambda} \mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2].$$

Optimizing over λ. To obtain the tightest possible bound, we optimize the expression with respect to the dual variable $\lambda > 0$. Consider the function:

$$f(\lambda) = \lambda\frac{\rho}{n} + \frac{1}{2\lambda}\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2].$$

Taking the derivative of $f(\lambda)$ with respect to λ and setting it to zero:

$$\frac{df}{d\lambda} = \frac{\rho}{n} - \frac{1}{2\lambda^2}\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2] = 0.$$

Solving for λ:

$$\frac{\rho}{n} = \frac{1}{2\lambda^2}\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2] \implies \lambda^2 = \frac{1}{2} \cdot \frac{\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2]}{\rho/n} = \frac{n}{2\rho}\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2],$$

$$\lambda = \sqrt{\frac{n}{2\rho}\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2]}.$$

Substituting Optimal λ Back. Substituting the optimal λ back into the expression for $R_n(\theta, \mathcal{P}_n)$:

$$R_n(\theta, \mathcal{P}_n) = \sqrt{\frac{n}{2\rho}\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2]} \cdot \frac{\rho}{n} + \mathbb{E}_{\hat{P}_n}[\ell(\theta, X)] + \frac{1}{2\sqrt{\frac{n}{2\rho}\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2]}}\mathbb{E}_{\hat{P}_n}[\ell(\theta, X)^2].$$

Simplifying each term:

$$\lambda\frac{\rho}{n} = \sqrt{\frac{n}{2\rho}\mathbb{E}[\ell^2]} \cdot \frac{\rho}{n} = \sqrt{\frac{\rho}{2n}\mathbb{E}[\ell^2]},$$

$$\frac{1}{2\lambda}\mathbb{E}[\ell^2] = \frac{1}{2\sqrt{\frac{n}{2\rho}\mathbb{E}[\ell^2]}}\mathbb{E}[\ell^2] = \sqrt{\frac{\rho}{2n}\mathbb{E}[\ell^2]}.$$

Therefore:

$$R_n(\theta, \mathcal{P}_n) = \sqrt{\frac{\rho}{2n}\mathbb{E}[\ell^2]} + \mathbb{E}[\ell] + \sqrt{\frac{\rho}{2n}\mathbb{E}[\ell^2]} = \mathbb{E}[\ell] + 2\sqrt{\frac{\rho}{2n}\mathbb{E}[\ell^2]} = \mathbb{E}[\ell] + \sqrt{\frac{2\rho}{n}\mathbb{E}[\ell^2]}.$$

References

Northwestern University