Homework 2

Advanced Topics in Statistical Learning, Spring 2023

Due Friday March 3 at 5pm

1 Properties of RKHS regression [16 points]

In this exercise, we will work out some facts about RKHS regression.

(a) First, let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be an arbitrary kernel. Recall that this means there exists a feature map $\phi : \mathcal{X} \to \mathcal{H}$, and a Hilbert space $\mathcal{H}$ with inner product $\langle \cdot\,, \cdot \rangle_{\mathcal{H}}$, such that for any $x, y \in \mathcal{X}$,

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}.$$

Prove that k is positive semidefinite, meaning, it is symmetric, and for any $n \geq 1$ and $x_1, \ldots, x_n \in \mathcal{X}$, if we define a matrix $K \in \mathbb{R}^{n \times n}$ to have entries $K_{ij} = k(x_i, x_j)$, then [3 pts]

$$a^{\mathsf{T}} K a \geq 0, \quad \text{for all } a \in \mathbb{R}^n.$$

Hint: express $a^{\mathsf{T}} K a$ in terms of the feature map ϕ.

(b) Henceforth, suppose that k is the reproducing kernel for $\mathcal{H}$ (and hence $\mathcal{H}$ is an RKHS). Recall that this means the following two properties are satisfied:

(a) for any $x \in \mathcal{X}$, the function $k(\cdot, x)$ is an element of $\mathcal{H}$;

(b) for any function $f \in \mathcal{H}$ and $x \in \mathcal{X}$, it holds that $\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$.

Let f be a function of the form

$$f(x) = \sum_{i=1}^{n} \beta_i k(x, x_i),$$

for coefficients $\beta_1, \ldots, \beta_n \in \mathbb{R}$. Show that [2 pts]

$$\|f\|_{\mathcal{H}}^2 = \sum_{i,j=1}^{n} \beta_i \beta_j k(x_i, x_j).$$

(c) Let h be any function (in $\mathcal{H}$) that is orthogonal (with respect to $\langle \cdot\,, \cdot \rangle_{\mathcal{H}}$) to the linear space of functions of the form in part (b). Prove that [3 pts]

$$h(x_i) = 0, \quad i = 1, \ldots, n,$$

and

$$\|f + h\|_{\mathcal{H}} \geq \|f\|_{\mathcal{H}}, \quad \text{with equality iff } h = 0.$$

(d) Argue that for any $\lambda > 0$, the infinite-dimensional RKHS optimization problem

$$\underset{f}{\text{minimize}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2$$

(where the minimization is implicitly over $f \in \mathcal{H}$) has a unique solution of the form in part (b), and we can rewrite it as [2 pts]

$$\underset{\beta}{\text{minimize}}\ \|Y - K\beta\|_2^2 + \lambda\beta^\mathsf{T} K\beta.$$

for the matrix $K \in \mathbb{R}^{n\times n}$ with entries $K_{ij} = k(x_i, x_j)$.

For the uniqueness part, you may assume may assume that k is a positive definite kernel (strictly), so that K is positive definite matrix (strictly).

Hint: let $g = f + h$ and use the results in part (c) to argue that g has a larger criterion value, unless $h = 0$. Use part (b) to complete the reduction to finite-dimensional form.

(e) Finally, we establish a cool fact about leave-one-out cross-validation (LOOCV) in RKHS regression problems. Recall that in general, the LOOCV error of an estimator $\hat{f}$ is defined as

$$\text{CV}(\hat{f}) = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{f}^{-i}(x_i))^2,$$

where $\hat{f}^{-i}$ is the estimator trained on all but the i^{th} pair (x_i, y_i). Prove the following shortcut formula for LOOCV with an RKHS regression estimator $\hat{f}$: [6 pts]

$$\frac{1}{n}\sum_{i=1}^n (y_i - \hat{f}^{-i}(x_i))^2 = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}\right)^2,$$

where $S = K(K + \lambda I)^{-1}$ is the smoother matrix for the RKHS estimator (so that the vector of fitted values is given by $\hat{Y} = SY$).

Hint: prove that for each i,

$$\hat{f}^{-i}(x_i) = \frac{1}{1 - S_{ii}}[\hat{f}(x_i) - S_{ii}y_i].$$

The desired result will follow by rearranging, squaring both sides, and summing over $i = 1, \ldots, n$. There are different ways to establish the above display; one nice way is as follows. Consider solving the RKHS regression problem on

$$(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n),$$

and consider solving it on

$$(x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_i, \tilde{y}_i), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n),$$

where $\tilde{y}_i = \hat{f}^{-i}(x_i)$. Argue that these should produce the same solutions. Derive the zero gradient condition for optimality (differentiate the criterion and set it equal to zero) for each problem, and use these to solve for $\tilde{y}_i = \hat{f}^{-i}(x_i)$.

2 Sub-Gaussian maximal inequalities [12 points]

In this exercise, we will derive tail bounds on maxima of sub-Gaussian random variables X_i, $i = 1, \ldots, n$. Suppose each X_i has mean zero and variance proxy σ^2. (We assume nothing about their dependence structure.)

(a) Prove that for any $\lambda \in \mathbb{R}$, [3 pts]

$$\exp\left(\lambda\mathbb{E}\Big[\max_{i=1,\ldots,n} X_i\Big]\right) \le ne^{\sigma^2\lambda^2/2}.$$

(b) Rearrange the result in part (a), then choose a suitable value of λ to show that [2 pts]

$$\mathbb{E}\Big[\max_{i=1,\ldots,n} X_i\Big] \le \sigma\sqrt{2\log n}.$$

(c) Prove that for any $\lambda \geq 0$, [2 pts]

$$\mathbb{P}\Big(\max_{i=1,\dots,n} X_i \geq \lambda \Big) \leq n e^{-\lambda^2/(2\sigma^2)}.$$

Hint: use the fact that $\mathbb{P}(X_i \geq \lambda) \leq e^{-\lambda^2/(2\sigma^2)}$, for any $\lambda \geq 0$, which you can view as a consequence of the tail bound for sub-Gaussian averages when $n = 1$.

(d) Reparametrize the result in part (c) to show that for any $t > 0$, [1 pt]

$$\mathbb{P}\Big(\max_{i=1,\dots,n} X_i \geq \sigma\sqrt{2(\log n + t)} \Big) \leq e^{-t}.$$

(e) Now, we turn the question of the role of dependence: do correlations between X_i, $i = 1, \dots, n$ make their maximum stochastically larger or smaller? Conduct (and include the results of) a small simulation in order to inform your answer. [4 pts]

3 Risk analysis for the constrained lasso [12 points]

This exercise explores simple in-sample and out-of-sample risk bounds for the lasso. Assume that we observe i.i.d. $(x_i, y_i) \in [-M, M]^d \times \mathbb{R}$, $i = 1, \dots, n$, where each

$$y_i = x_i^\mathsf{T} \beta_0 + \epsilon_i,$$

and each ϵ_i is sub-Gaussian with mean zero and variance proxy σ^2. Consider the constrained-form lasso estimator $\hat\beta$, which solves

$$\underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2 \ \text{ subject to } \ \|\beta\|_1 \leq t,$$

where $Y = (y_1, \dots, y_n) \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times d}$ is the predictor matrix (whose i^{th} row is $x_i \in \mathbb{R}^d$), and $t \geq 0$ is a tuning parameter.

(a) Prove that the lasso estimator, with $t = \|\beta_0\|_1$, has in-sample risk satisfying [3 pts]

$$\frac{1}{n}\mathbb{E}\|X\hat\beta - X\beta_0\|_2^2 \leq 8M\sigma\|\beta_0\|_1\sqrt{\frac{2\log(2d)}{n}},$$

where the expectation is taken over the training data (x_i, y_i), $i = 1, \dots, n$.

Hint: follow the strategy that we used in lecture to derive the "slow" rate for the constrained lasso. Note that this was for fixed X, so you will need to condition on X here. Then apply the result from Q2 part (b). You may use the fact that for $\epsilon = (\epsilon_1, \dots, \epsilon_n) \in \mathbb{R}^n$, a vector of i.i.d. sub-Gaussians with mean zero and variance proxy σ^2, and an arbitrary fixed vector $a \in \mathbb{R}^n$, the random variable $a^\mathsf{T} epsilon$ is mean zero sub-Gaussian with variance proxy $\sigma^2\|a\|_2^2$.

(b) For i.i.d. mean zero random variables Z_i, $i = 1, \dots, n$ that lie almost surely in $[a, b]$, prove that for any $t \in \mathbb{R}$, [2 pts]

$$\mathbb{E}\left[\exp\left(\frac{t}{n} \sum_{i=1}^n Z_i \right) \right] \leq e^{t^2(b-a)^2/(8n)}.$$

Hint: you may use the fact that each Z_i is sub-Gaussian with variance proxy $(b - a)/2$, and the hint about linear combinations of sub-Gaussians from part (a).

(c) Let Σ denote the predictor covariance matrix, that is, $\Sigma = \mathbb{E}[x_0 x_0^\mathsf{T}]$ for a draw x_0 from the predictor distribution. Let $\hat\Sigma = X^\mathsf{T} X/n$ be the empirical covariance matrix, and let $V = \hat\Sigma - \Sigma$. Prove that [2 pts]

$$\mathbb{E}\Big[\max_{j,k=1,\dots,d} |V_{jk}| \Big] \leq M^2\sqrt{\frac{2\log(2d^2)}{n}}.$$

Hint: apply part (b) to the entries of V.

(d) Prove that the lasso estimator, with $t = \|\beta_0\|_1$, has out-of-sample risk satisfying [5 pts]

$$\mathbb{E}[(x_0^\mathsf{T}\hat{\beta} - x_0^\mathsf{T}\beta_0)^2] \le 8M\sigma\|\beta_0\|_1\sqrt{\frac{2\log(2d)}{n}} + 4M^2\|\beta_0\|_1^2\sqrt{\frac{2\log(2d^2)}{n}},$$

where the expectation is taken over the training data (x_i, y_i), $i = 1, \dots, n$ and an independent draw x_0 from the predictor distribution.

Hint: first, argue that the in-sample risk and out-of-sample risk can be written as

$$\mathbb{E}\big[(\hat{\beta} - \beta_0)^\mathsf{T}\hat{\Sigma}(\hat{\beta} - \beta_0)\big] \quad \text{and} \quad \mathbb{E}\big[(\hat{\beta} - \beta_0)^\mathsf{T}\Sigma(\hat{\beta} - \beta_0)\big],$$

respectively. (Note that the expectations above are each taken with respect to the training samples (x_i, y_i), $i = 1, \dots, n$ only—there is nothing else that is random.) Next, argue that

$$(\hat{\beta} - \beta_0)^\mathsf{T}\Sigma(\hat{\beta} - \beta_0) - (\hat{\beta} - \beta_0)^\mathsf{T}\hat{\Sigma}(\hat{\beta} - \beta_0) \le \sum_{j,k=1}^{d} |(\hat{\beta} - \beta_0)_j||(\hat{\beta} - \beta_0)_k||V_{jk}|,$$

where recall $V_{jk} = (\hat{\Sigma} - \Sigma)_{jk}$. Then do a little bit more algebra to bound the right-hand side above and apply the previous parts of this question to conclude the result.

(e) The bound derived in this question for the out-of-sample risk is always larger than that for the in-sample risk (by nature of its construction). As a bonus, investigate: can the out-of-sample risk of the lasso be lower than the in-sample risk? Use a simulation, a pointer to an experiment or result in the literature, or any means of answering that you believe provides a convincing argument.

4 Γ-minimaxity of ridge regression [10 points]

This exercise will explore a (strong, finite-sample) notion of minimax optimality of ridge regression. Assume that $X \in \mathbb{R}^{n \times d}$ is a fixed, arbitrary predictor matrix, and consider the following model:

$$\begin{aligned} (\epsilon, \beta_0) &\sim F^n \times G^d, \\ Y &= X\beta_0 + \epsilon. \end{aligned} \tag{1}$$

Here, F and G are distributions on $\mathbb{R}$ that have mean zero and variance $\sigma^2 > 0$ and $r^2/d \ge 0$, respectively. Abbreviate $\gamma = (F, G)$, and let Γ denote the set of all such pairs of distributions (F, G) that satisfy these moment conditions.

We can think of F and G as describing a noise and prior distribution, respectively: the components of ϵ are i.i.d. from F, and those of β_0 are i.i.d. from G. We measure risk according to:

$$\mathrm{Risk}(\hat{\beta}; \gamma) = \mathbb{E}\|\hat{\beta} - \beta_0\|_2^2,$$

where the expectation is over all that is random: ϵ, β_0, Y drawn from (1). To be specific, this is the *Bayes* risk of an estimator $\hat{\beta}$, though we'll often just call it risk for short. The notation $\mathrm{Risk}(\hat{\beta}; \gamma)$ emphasizes the dependence of the risk on γ.

(a) Prove that the ridge regression estimator,

$$\hat{\beta} = \underset{\beta}{\mathrm{argmin}} \; \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2,$$

for any fixed tuning parameter value $\lambda > 0$, has risk [4 pts]

$$\mathrm{Risk}(\hat{\beta}; \gamma) = \frac{\sigma^2}{n}\mathrm{tr}\big[\lambda^2\alpha(\hat{\Sigma} + \lambda I)^{-2} + \hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-2}\big],$$

where $\hat{\Sigma} = (X^\mathsf{T}X)/n$ and $\alpha = r^2 n/(\sigma^2 d)$. Hint: use a bias-variance decomposition.

(b) In general, consider estimation of a functional $\theta(P)$, given i.i.d. draws from P, with respect to some metric d. Let Γ let be class of distributions γ, where each $\gamma = (P, \pi)$, with P specifying the distribution of the data, and π specifying the distribution of the functional $\theta(P)$. We can think of P as the likelihood and π as the prior. Then the Γ-minimax risk is defined as:

$$\inf_{\hat{\theta}} \sup_{\gamma \in \Gamma} \mathbb{E}_\gamma \big[d(\theta(P), \hat{\theta})\big].$$

The expectation here is with respect to γ, and hence is a Bayes risk. Note that if for each $\gamma = (P, \pi)$, the prior is a point mass at a single value of $\theta(P)$, then Γ-minimax risk reduces to the usual notion of minimax risk defined in lecture.

Let $\hat{\theta}^B$ be the Bayes estimator with respect to some likelihood-prior pair $\gamma_0 = (P_0, \pi_0)$, where $\gamma_0 \in \Gamma$. Prove that if its Bayes risk is constant as we vary $\gamma \in \Gamma$, then it is Γ-minimax optimal, [3 pts]

$$\sup_{\gamma \in \Gamma} \mathbb{E}_\gamma \big[d(\theta(P), \hat{\theta}^B)\big] = \inf_{\hat{\theta}} \sup_{\gamma \in \Gamma} \mathbb{E}_\gamma \big[d(\theta(P), \hat{\theta})\big].$$

Hint: suppose not, and obtain a contradiction to the fact that $\hat{\theta}^B$ is Bayes.

(c) Returning to our regression problem setting, prove that ridge regression is the Bayes estimator with respect to a particular instantiation of $\gamma = (F, G)$, and use the previous parts to establish that it is Γ-minimax for the class Γ defined in part (a). [3 pts]

(d) As a bonus: extend the previous result to the case of out-of-sample Bayes prediction risk. That is, instead of (1), consider

$$\begin{aligned} (\epsilon, \beta_0, x_0) &\sim F^n \times G^d \times Q, \\ Y &= X\beta_0 + \epsilon, \end{aligned} \tag{2}$$

where F, G are as before, and now Q is a distribution on $\mathbb{R}^d$ with mean zero and covariance $\Sigma \succ 0$. Abbreviate $\gamma = (F, G, Q)$, and let Γ denote the set of all such triplets of distributions (F, G, Q) that meet the specified moment conditions. While the training predictors X are still fixed, Q specifies the distribution of the test predictor x_0 used to measure risk, defined as:

$$\mathrm{Risk}(\hat{\beta}; \gamma) = \mathbb{E}[(x_0^\mathsf{T} \hat{\beta} - x_0^\mathsf{T} \beta_0)^2],$$

where the expectation is over everything that is random: $\epsilon, \beta_0, x_0, Y$ drawn from (2). For this model and new definition of risk, prove that ridge regression is still Γ-minimax optimal. Along the way, you will find it useful to prove that the out-of-sample Bayes prediction risk of ridge is

$$\mathrm{Risk}(\hat{\beta}; \gamma) = \frac{\sigma^2}{n} \mathrm{tr}\big[\lambda^2 \alpha (\hat{\Sigma} + \lambda I)^{-2} \Sigma + \hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-2} \Sigma\big].$$