Stats315a Problem Set 3

Due Date: Friday, March 22, 11:59pm

Question 3.1 (A semiparametric least squares model, 30 points)**:** Consider the model that predicts $\widehat{y}$ via

$$\widehat{y}_i = x_i^T \beta + f(x_i)$$

where $f : \mathbb{R}^p \to \mathbb{R}$ belongs to an RKHS with reproducing kernel $\mathsf{k} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. We have a sample $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ of size n, and solve the least-squares estimation problem

$$(\widehat{\beta}, \widehat{f}) = \underset{\beta, f}{\operatorname{argmin}} \left\{ \frac{1}{2} \|X\beta - f(X) - y\|_2^2 + \frac{\lambda_0}{2} \|\beta\|_2^2 + \frac{\lambda_1}{2} \|f\|^2 \right\}, \tag{3.1}$$

where $f(X) = [f(x_1) \ \cdots \ f(x_n)]^T \in \mathbb{R}^n$ denotes the vector of predictions of f and $\|f\|^2$ is the squared RKHS norm of f.

(a) If $K = [\mathsf{k}(x_i, x_j)]_{i,j \le n}$ is the Gram (Kernel) matrix, describe with a few words (literally) why problem (3.1) is equivalent to the problem

$$\underset{\beta \in \mathbb{R}^p, \alpha \in \mathbb{R}^n}{\operatorname{minimize}} \ \frac{1}{2} \|X\beta - K\alpha - y\|_2^2 + \frac{\lambda_0}{2} \|\beta\|_2^2 + \frac{\lambda_1}{2} \alpha^T K \alpha. \tag{3.2}$$

(b) Show that the minimizers for problem (3.2) satisfy the consistency conditions

$$H_{\lambda_0} \widehat{\beta} = X^T (y - \widehat{f})$$
$$S_{\lambda_1} \widehat{f} = y - X\widehat{\beta}$$

where $\widehat{f} = [\widehat{f}(x_1) \ \cdots \ \widehat{f}(x_n)]^T = K\widehat{\alpha}$ is the semiparametric part of the model. Give the matrices H_{λ_0} and S_λ. (You may assume that K is invertible.)

(c) Show that we may solve problem (3.2) via the block matrix inversion problem

$$\begin{bmatrix} H_{\lambda_0} & X^T K \\ X & K + \lambda_1 I \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \begin{bmatrix} X^T y \\ y \end{bmatrix}.$$

Question 3.2 (Fitting a semiparametric model, 50 points)**:** The datasets `adult_train.csv`, `adult_val.csv`, and `adult_test.csv` in the data directory contain random subsets of 2000 datapoints (each) from the Folktables package, with a full description available at https://github.com/socialfoundations/folktables. This consists of data with covariates for several categorical and numerical characteristics, including hours-per-week of work, educational attainment, and income. Treating income as the response, you will fit a semiparametric model as in Question 3.1.

For the non-income covariates, you should standardize the numerical covariates to have mean-zero and variance 1 across the data; for the non-numerical covariates, use a 1-hot encoding. (So if a categorical covariate has k distinct values, which may include missing, expand it into k positions in your vectors x with 1 in the position corresponding to the present category.) Note that this dataset has a few idiosyncrasies of which you ought to be aware: first, it is part of the census data from 1990 (updated through 1994), and so incomes were lower; it censors the highest income at 99999. You may ignore that censoring in your modeling. Second, we consider the following covariates in the model:

i. `hours_per_week`, a numerical covariate of the number of hours worked
ii. `age`, numerical, the age of the individual
iii. `workclass`, a binary variable of whether someone works in the private or public sector
iv. `education_num`, which is (related to) the number of years of education an individual has, with modifications, as 13 corresponds to completing a Bachelors, 10 some college, 9 finishing high school, among other strata.
v. `marital_status`, which is categorical
vi. `relationship`, which is categorical
vii. `race`, which includes mostly "White" and "Black" but three less common categories (which you may wish to group into "non-white-black")
viii. `sex`, which in this dataset is binary.

Use the Gaussian kernel function $\mathsf{k}(x, z) = \exp(-\frac{1}{2\tau^2} \|x - z\|_2^2)$, for $\tau > 0$ to be chosen, and regularization $\lambda_0 = 0$ to fit the model as in (3.2). Use the `adult_val.csv` data to perform held-out validation to choose the regularizer λ_1 and τ for the kernel, selecting values for each in the exponentially spaced range $\{2^{-2.5}, 2^{-2}, \ldots, 2^2, 2^{2.5}\} = \{2^{i/2}\}_{i=-5}^{5}$.

(a) What is the root-mean-square error on the data in `adult_test.csv` for the model you have selected?

(b) Assume that the estimate $\widehat{f}$ is sufficiently consistent that solving

$$\widehat{\beta} = \operatorname*{argmin}_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{f}(x_i) - x_i^T \beta)^2$$

is equivalent to the "oracle" solution

$$\widehat{\beta}^{\text{oracle}} = \operatorname*{argmin}_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - f^\star(x_i) - x_i^T \beta)^2,$$

where $(\beta^\star, f^\star) = \operatorname{argmin}_{\beta, f} \mathbb{E}[(y - f(x) - x^T\beta)^2] + \lambda \|f\|^2$, using the notation of problem 3.1. Using this, give a sandwich covariance estimate, computable from the data, for the covariance in the approximation

$$\widehat{\beta} - \beta^\star \stackrel{\cdot}{\sim} \mathsf{N}(0, \widehat{\Sigma}). \tag{3.3}$$

(c) For the preceding covariance, give a 95% confidence interval for the component $\beta_j^\star$ associated to the `sex` variable.

(d) For the preceding covariance, give a 95% confidence interval for the variable corresponding to being `married`.

Question 3.3 (Reproducing Kernel Hilbert Spaces, 20 points)**:** In this question we explicate some of the conditions required for a symmetric $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to be a valid kernel function. Recall that K is a valid kernel if for all sets of points $\{x_i\}_{i=1}^n \subset \mathcal{X}$, the Gram matrix

$$G := [K(x_i, x_j)]_{i,j=1}^n \in \mathbb{R}^{n \times n}$$

is positive semidefinite, that is, $G \succeq 0$. An equivalent statement is that $K(x, z) = \langle \phi(x), \phi(z) \rangle$ for some feature mapping ϕ and inner product $\langle \cdot, \cdot \rangle$.

(a) Let K_1, K_2 be valid kernel functions. Show that $K_1 + K_2$ is a valid kernel.

(b) Let K_1 be a kernel on $\mathbb{R} \times \mathbb{R}$ and let K_2 be a kernel on $\mathbb{R} \times \mathbb{R}$. Define the "direct sum" kernel $K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ by

$$K((x_1, x_2), (z_1, z_2)) = K_1(x_1, z_1) + K_2(x_2, z_2).$$

Show that K is a valid kernel.

(c) Let K_1, K_2 be valid kernel functions. Show that $K_1 \cdot K_2$, that is, the function $K(x, z) = K_1(x, z)K_2(x, z)$ is a valid kernel.

Question 3.4 (A direct sum Hilbert space, 20 points)**:** Let $\mathcal{X}_1, \ldots, \mathcal{X}_d$ be arbitrary spaces (for example, each could be just a copy of $\mathbb{R}$), and let $\mathcal{X}^d = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$ be their Cartesian product. (So $x \in \mathcal{X}^d$ has the form $x = (x_1, \ldots, x_d)$ for $x_j \in \mathcal{X}_j$.) Suppose that K_i is the reproducing kernel for a Hilbert space $\mathcal{H}_i$ of functions from spaces $\mathcal{X}_i \to \mathbb{R}$, where $\mathcal{H}_i$ has inner product $\langle \cdot, \cdot \rangle_i$. That is, $\langle K(x, \cdot), f \rangle_i = f(x)$ for any $f \in \mathcal{H}_i$ and $x \in \mathcal{X}_i$. Let $\mathcal{F}$ be the space of functions mapping $\mathcal{X}^d \to \mathbb{R}$ of the form

$$f(x) = \sum_{j=1}^{d} f_j(x_j),$$

where $f_j \in \mathcal{H}_j$. Define the direct sum inner product for $f, g \in \mathcal{F}$ by

$$\langle f, g \rangle = \sum_{j=1}^{d} \langle f_j, g_j \rangle_j,$$

noting that if $f \in \mathcal{F}$, then the reproducing property becomes $\langle f, K_j(x_j, \cdot) \rangle = \langle f_j, K_j(x_j, \cdot) \rangle_j = f_j(x_j)$, and for $K = \sum_{j=1}^{d} K_j$ we have the coordinate-wise reproducing inner product

$$\langle f, K(x, \cdot) \rangle = \sum_{j=1}^{d} \langle f_j, K_j(x_j, \cdot) \rangle = \sum_{j=1}^{d} f_j(x_j) = f(x).$$

(a) Write $\|f\|^2 = \langle f, f \rangle$ in terms of the norms $\|h\|_{\mathcal{H}_i}^2 := \langle h, h \rangle_i$, defined for $h \in \mathcal{H}_i$.

(b) Now you will demonstrate a variant of the representer theorem specialized to such direct sums. Consider the problem

$$\underset{f \in \mathcal{F}}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) + \frac{\lambda}{2} \|f\|^2, \tag{3.4}$$

where $\lambda > 0$, $\|\cdot\|$ is the norm from part (a), and $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$ is some loss function. Show that it is no loss of generality to assume that the minimizer of this problem takes the form

$$f(x) = \sum_{j=1}^{d} \sum_{i=1}^{n} \alpha_{ij} K_j(x_i, x),$$

and rewrite the problem (3.4) as an nd-dimensional optimization problem.

(c) Consider an extension of the previous part in which we model predictions of a response $y \in \mathbb{R}$ given $x \in \mathbb{R}^d$ as

$$\widehat{y}_{\theta,f}(x) = \theta_0 + x^T \theta + \sum_{j=1}^d f_j(x_j).$$

Show that for $\lambda_0 \ge 0, \lambda_1 > 0$, it is no loss of generality assume that the minimizers (in f) of the problem

$$\underset{\theta \in \mathbb{R}^{d+1}, f \in \mathcal{F}}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^n \ell(\widehat{y}_{\theta,f}(x_i), y_i) + \lambda_0 \cdot \text{reg}(\theta) + \lambda_1 \|f\|^2 \tag{3.5}$$

take the form $f(x) = \sum_{j=1}^d \sum_{i=1}^n \alpha_{ij} K_j(x_i, x)$.

Question 3.5 (ℓ_1-regularization and forward-selection, 20 points)**:** Consider a forward-selection- or boosting-type procedure for predicting targets y from $x \in \mathcal{X}$, where at iteration k we have a feature mapping $\phi^k : \mathcal{X} \to \{-1, 1\}^k$, $\phi^k(x) = (\phi_1(x), \ldots, \phi_k(x))$, and we wish to add a new feature $\phi_{k+1} : \mathcal{X} \to \{-1, 1\}$. At iteration k, our predictive model is thus

$$f_k(x) = \langle \theta^k, \phi^k(x) \rangle = \sum_{j=1}^k \theta_j \phi_j(x).$$

We assume we are minimizing a loss $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, convex in its first argument, so that this new feature should (approximately) minimize

$$\frac{1}{n} \sum_{i=1}^n \ell(f_k(x_i) + \theta_{k+1} \phi_{k+1}(x_i), y_i)$$

jointly in $\theta_{k+1} \in \mathbb{R}$ and ϕ_{k+1}.

At each iteration, we conduct a hypothesis test to assess whether to add a prospective new feature ϕ_{k+1}. Say that the null at iteration $k+1$ is that

$$H_{0,k+1} : \underset{\theta}{\text{argmin}} \left\{ \mathbb{E}[\ell(f_k(x) + \theta \phi_{k+1}(x), y)] \right\} = 0$$

(where the expectation is over (x, y) drawn from the population being sampled).

(a) Show that the null $H_{0,k+1}$ equivalent to the equality

$$\mathbb{E}[\ell'(f_k(x), y) \phi_{k+1}(x)] = 0,$$

where $\ell'(t, y) = \frac{\partial}{\partial t} \ell(t, y)$.

(b) Ignoring the issue that f_k depends on the sample, an approximation to the preceding condition is that

$$\frac{1}{n} \sum_{i=1}^n \ell'(f_k(x_i), y_i) \phi_{k+1}(x_i) \stackrel{\cdot}{\sim} \mathsf{N}\left(0, \frac{1}{n} \left(\frac{1}{n} \sum_{i=1}^n \ell'(f_k(x_i), y_i)^2 \right) \right) \tag{3.6}$$

(because $\phi_{k+1}(x_i)^2 = 1$ for each x_i). Give an (approximate) level $1 - \alpha$ test of $H_{0,k+1}$ using the approximation (3.6), that is, test whether $\theta^\star_{k+1} = 0$.

(c) Suppose we are given the potential new feature mapping ϕ_{k+1} and choose the value θ_{k+1} as

$$\theta_{k+1} = \operatorname*{argmin}_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(f_k(x_i) + \theta\phi_{k+1}(x_i), y_i) + \lambda|\theta| \right\}.$$

Give the value $\lambda > 0$ such that $\theta_{k+1} \neq 0$ if and only if your test from part (b) rejects that $\theta^\star_{k+1} = 0$.

(d) Assume now that $\ell(t, y)$ has M-Lipschitz continuous derivative in t, or, equivalently, that $\ell''(t, y) \leq M$ for all t. Show that with your value λ from part (c), the alternative update

$$\theta_{k+1} = \operatorname*{argmin}_{\theta} \left\{ \left(\frac{1}{n} \sum_{i=1}^{n} \ell'(f_k(x_i), y_i)\phi_{k+1}(x_i) \right) \cdot \theta + \frac{M}{2}\theta^2 + \lambda|\theta| \right\},$$

which arises by upper bounding ℓ with a quadratic, satisfies $\theta_{k+1} \neq 0$ if and only if your test from part (b) rejects that $\theta^\star_{k+1} = 0$.

(e) Let $\ell(t, y) = \log(1 + e^{t-y}) + \log(1 + e^{y-t})$ be a smooth robust regression loss. Give $M = \sup_{t \in \mathbb{R}} \ell''(t, y)$.