Homework 4
STATS205 (Spring 2021–2022)

Please structure your writeups hierarchically: convey the overall plan before diving into details. You should justify with words why something's true (by algebra, convexity, etc.). There's no need to step through a long sequence of trivial algebraic operations. Be careful not to mix assumptions with things which are derived. **Up to two additional points will awarded for especially well-organized and elegant solutions.**

Due date: Tuesday, May 24th, 2022.

Corrections/Clarifications of Monday May 23 in red.

Corrections/Clarifications of Tuesday May 24 in blue.

1. Kernel regression (*12 points*)

a. (*2 points*) **Kernel ridgeless regression with hinge features.** Suppose we are given n data points $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$ where $x^{(i)} \in \mathbb{R}$ and $y^{(i)} \in \mathbb{R}$.

Suppose the x-data are strictly sorted $x^{(1)} < x^{(2)} < \ldots < x^{(n)}$.

Consider the feature map $\phi : \mathbb{R} \to \mathbb{R}^m$ where $m = n$, based on so call *hinge features*. Namely, let $\phi(x) = (\phi_j(x))_{j=0}^{n-1}$, where

$$\phi_j(x) = (x - x^{(j)})_+, \qquad j = 1, \ldots, n-1.$$

with $(u)_+ \equiv max(0, u)$, and $\phi_0(x) = 1$.

Consider ridgeless kernel regression:

$$\begin{aligned} \min_{\theta \in \mathbb{R}^m} \quad & \|\theta\|_2^2 \\ \text{s.t.} \quad & \forall i \in \{1, \ldots, n\}, y^{(i)} = \theta^\top \phi(x^{(i)}) \end{aligned} \tag{1}$$

(1) Define the interpolating linear spline:

$$s_n(x) = \sum_{i=1}^{n-1} \left(y^{(i)} + \frac{(y^{(i+1)} - y^{(i)})}{(x^{(i+1)} - x^{(i)})}(x - x^{(i)}) \right) \cdot 1_{\{x \in [x^{(i)}, x^{(i+1)})\}}.$$

with knots at the n data points $(x^{(i)})_{i=1}^n$.

Define the matrix $\Phi = (\Phi_{i,j}, 1 \le i \le n, 0 \le j < n)$ by $\Phi_{ij} = \phi_j(x^{(i)})$. Denote the solution of (1) by $\theta^*_{1,n}$. Explain/Prove/Justify how,

$$(s_n(x^{(i)}) : 1 \le i \le n) = \Phi \theta^*_{1,n}$$

(2) Suppose more generally we define at $x \notin \{x^{(i)} : 1 \le i \le n\}$ that

$$s_n(x) \equiv \phi^T(x)\theta^*_{1,n} = (\theta^*_{1,n})_0 + \sum_{j=1}^{n-1} (\theta^*_{1,n})_n (x - x^{(j)})_+.$$

where $(\theta^*_{1,n})_j$ means: the j-th entry in vector $\theta^*_{1,n}$, What does this formula mean for $x < x^{(1)}$, and for $x \ge x^{(n)}$?

b. (*3 points*) Let $s_n(x)$ be again the interpolating spline in the last question. Suppose in addition that $x^{(j)} = (j-1)/(n-1)$, $j = 1, \ldots, n$.

(1) Prove/explain the formula, for $x \in [x^{(1)}, x^{(n)}]$, $x \neq x^{(j)}$ for $j \in \{1, \ldots, n\}$.

$$\frac{d}{dx} s_n(x) = \sum_{j=1}^{n-1} \theta_j 1_{\{x > x^{(j)}\}}.$$

You may use the formula

$$\frac{d}{dx} \max(0, x) = 1_{\{(0,\infty)\}}.$$

(2) What happens to $\frac{d}{dz} s_n(z)$ on each side of $x^{(i)}$?
(3) What happens to $\frac{d}{dz} s_n(z)$ for $x < x^{(1)}$?

c. (*3 points*)
(1) Let $M = (M_{i,j} : 0 \leq j < n, i = 1, \ldots, n)$: $M_{i,0} = 0, \quad i = 1, \ldots, n$, $M_{n,j} = 0$, $j = 1, \ldots, n$.

$$M_{i,j} = 1_{\{x_i \geq x_j\}}, \qquad 1 \leq j < n, 1 \leq i < n.$$

Prove/explain that

$$\frac{d}{dz} s_n(z)|_{z \downarrow x^{(i)}} = (M\theta)_i, \qquad i=1,\ldots,n-1.$$

(2) Consider the penalization

$$P(\theta) = (n-1)^{-1} \|M\theta\|_2^2 = \frac{1}{n-1} \sum_{i=1}^{n} (M\theta)_i^2$$

Explain the relationship of $P(\theta)$ to

$$Q(s_n) = \int_0^1 (\frac{d}{dz} s_n(z))^2 dz.$$

(3) Consider this variation on ridgeless kernel regression:

$$\begin{aligned} \min_{\theta \in \mathbb{R}^m} \quad & P(\theta) \\ \text{s.t.} \quad & \forall i \in \{1, \ldots, n\}, y^{(i)} = \theta^\top \phi(x^{(i)}) \end{aligned} \tag{2}$$

What is the relation between the solution θ_2^*, say, of (2) and the solution θ_1^*, say, of (1).

d. (*5 points*) **Penalized Kernel regression.** In the same setting as above, consider the Penalized Kernel regression defined as

$$\min_{\theta \in \mathbb{R}^m} \; n^{-1} \cdot \|y - \Phi\theta\|_2^2 + \lambda \cdot P(\theta). \tag{3}$$

(1) Show that the optimal solution $\hat{\theta}_{\lambda,P}$ for the program (3) is given by

$$\hat{\theta}_{\lambda,P} = (\Phi^T \Phi + \lambda M^T M)^{-1} \Phi^T y\,, \tag{4}$$

where M is the $n \times n$ matrix defined earlier.
(3) Evaluate and present $(M^T M)_{k,\ell}$ explicitly for $0 \leq k, \ell \leq 2$.
(4) Write an R code for computing the solution $\theta_{\lambda,P}$ of (3) and returning $f_{\lambda,P} = \Phi\hat{\theta}_{\lambda,P}$. Test your code.
(5) Suppose you instead use the software

`ss(x,y, m=1, lambda=<SPECIFY>,all.knots=TRUE)`

where `<SPECIFY>` is set to whatever value of λ you use in (4). How will this compare to the solution to (4).

(6) What will happen if the data obey $y^{(i)} = x^{(i)}$ for $i = 1, \ldots, n$, and you solve the above problem i.e. what will $f_{\lambda,P}$ be. (explicit formula possible).

2. Comparing smoothers (*13 points*)

a. (*2 points*) Get the paper 'A review of spline function procedures in R' by Perperoglou et al.

`https://bmcmedresmethodol.biomedcentral.com/track/pdf/10.1186/s12874-019-0666-3.pdf`

Also get the dataset `Triceps Skinfold Thickness` that is used in that article, esp see Figure 5.

`https://mfp.imbi.uni-freiburg.de/book#dataset_tables`

Also get the code from Supplemental File of the above article.

Alternatively look in our Canvas website for all these files, inside folder: `Canvas>Files>HW4`.

b. (*3 points*) Replicate Figure 5 in the above paper.

c. (*8 points*) Figure 5 shows - in the lower two panels - several spline functions, with different explicit choices of equivalent degrees of freedom (edf=4 in lower left, edf=10 in lower right).

Design and conduct a leave-one-out cross-validation study giving, for each type of smoother in those two lower panels, the best choice of edf [equivalent degrees of freedom] from among the two given (4 and 10) as well as the additional possibilities 6 8 12 and 20.

Show your study's code and plot the raw data as well as for each type of smoother under study, the instance having the selected smoothing parameters.