STOR 767 Spring 2019 Hw4

Due on 02/27/2019 in Class

Zhenghan Fang

Remark. This homework focuses on splines and kernel methods.

Instruction.

- **Theoretical Part and Computational Part are respectively credited 60 points. At most 100 points in total will be accounted for this homework.**
- Submission of handwritten solution for the **Theoretical Part** of this homework is allowed.
- Please use **RMarkdown** to create a formatted report for the **Computational Part** of this homework.
- Some of the problems are selected or modified from the textbook [1].

Theoretical Part

1. (*15 pt*) (Textbook Ex. 5.1) Consider the following truncated power basis with 2 knots $\xi_1, \xi_2 \in \mathbb{R}$

   $$h_1(x) = 1, \quad h_2(x) = x, \quad h_3(x) = x^2, \quad h_4(x) = x^3, \quad h_5(x) = (x-\xi_1)_+^3, \quad h_6(x) = (x-\xi_2)_+^3. \quad (x \in \mathbb{R})$$

   Show that it represents a basis for a cubic spline with knots ξ_1, ξ_2.

   Let $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \theta_1 (x-\xi_1)_+^3 + \theta_2 (x-\xi_2)_+^3$. Then,

   $$f'(x) = \beta_1 + 2\beta_2 x + 3\beta_3 x^2 + 3\theta_1 (x-\xi_1)_+^2 + 3\theta_2 (x-\xi_2)_+^2$$
   $$f''(x) = 2\beta_2 + 6\beta_3 x + 6\theta_1 (x-\xi_1)_+ + 6\theta_2 (x-\xi_2)_+$$

   Then,

   $$f(\xi_1^-) = f(\xi_1^+)$$
   $$f'(\xi_1^-) = f'(\xi_1^+)$$
   $$f''(\xi_1^-) = f''(\xi_1^+)$$
   $$f(\xi_2^-) = f(\xi_2^+)$$
   $$f'(\xi_2^-) = f'(\xi_2^+)$$
   $$f''(\xi_2^-) = f''(\xi_2^+)$$

   Therefore, $\{h_1(x), ..., h_6(x)\}$ represents a basis for a cubic spline.

2. (*15 pt*) (Textbook Ex. 5.4) Consider the truncated power series representation for cubic splines with K-knots $\{\xi_k\}_{k=1}^K \subseteq \mathbb{R}$

$$f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \xi_k)_+^3. \quad (x \in \mathbb{R})$$

(a) Prove that the natural boundary conditions (*i.e.* f is linear on $(-\infty, \xi_1] \cup [\xi_K, +\infty)$) is equivalently to the coefficient constraints

$$\beta_2 = \beta_3 = \sum_{k=1}^{K} \theta_k = \sum_{k=1}^{K} \xi_k \theta_k = 0.$$

When $x \in (-\infty, \xi_1]$,

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

When $x \in [\xi_K, +\infty)$,

$$f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \xi_k)^3 = \left(\beta_3 + \sum_{k=1}^{K} \theta_k\right) x^3 - 3\left(\beta_2 + \sum_{k=1}^{K} \xi_k \theta_k\right) x^2 + ax + b$$

Therefore, the natural boundary conditions is equivalent to

$$\beta_2 = \beta_3 = \sum_{k=1}^{K} \theta_k = \sum_{k=1}^{K} \xi_k \theta_k = 0.$$

(b) Derive the natural cubic splines. That is, argue the set of basis functions (linearly independent, and in a special for of truncated basis)

$$N_1(x) = 1, \quad N_2(x) = x, \quad N_{k+2}(x) = d_k(x) - d_{K-1}(x) \quad (x \in \mathbb{R},\ 1 \leqslant k \leqslant K-2)$$

where

$$d_k(x) = \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k} \quad (x \in \mathbb{R},\ 1 \leqslant k \leqslant K-2)$$

can represent any truncated power basis under the constraints in (a).

For any $\alpha d_k(x)$, $\alpha \in \mathbb{R}$, $1 \leqslant k \leqslant K-1$,

$$\sum_{k=1}^{K} \theta_k = \frac{\alpha}{\xi_K - \xi_k} - \frac{\alpha}{\xi_K - \xi_k} = 0$$

For any $\alpha N_{k+2}(x)$, $\alpha \in \mathbb{R}$, $1 \leqslant k \leqslant K-1$,

$$\sum_{k=1}^{K} \xi_k \theta_k = \frac{\alpha \xi_k}{\xi_K - \xi_k} - \frac{\alpha \xi_K}{\xi_K - \xi_k} - \frac{\alpha \xi_{K-1}}{\xi_K - \xi_{K-1}} + \frac{\alpha \xi_K}{\xi_K - \xi_{K-1}} = 0$$

Therefore, $\{N_1(x), ..., N_K(x)\}$ can represent any truncated pawer basis under the constraints in (a).

3. (*15 pt*) (Smoothing Splines, Textbook Ex. 5.7) Suppose $N \geqslant 2$, and that g is the natural cubic spline interpolant to the pairs $\{(x_i, z_i)\}_{i=1}^N$, with $-\infty < a < x_1 < \cdots < x_N < b < +\infty$, *i.e.* g as a cubic spline with knots $\{x_i\}_{i=1}^N$ satisfies

$$g(x_i) = z_i. \quad (1 \leqslant i \leqslant N)$$

Let $\widetilde{g} \in C^2[a, b]$ be any other twice continuously differentiable function supported on $[a, b]$ that interpolates the N pairs.

(a) Let $r = \widetilde{g} - g$ be the residual. Show that g'' is orthogonal to r'' in $L^2(\mathrm{d}\mathfrak{m})$ (denoting $\mathfrak{m}$ as the Lebesgue measure). That is,

$$\begin{aligned}\langle g'', r'' \rangle &:= \int_a^b g''(x) r''(x) \mathrm{d}x \\ &= -\sum_{j=1}^N g'''(x_j^+)[r(x_{j+1}) - r(x_j)] \quad \text{(by integration-by-part)} \\ &= 0.\end{aligned}$$

Hint. Note the stepwise constancies of g''' as a cubic spline function. Take care of the non-smoothness at knots in integration-by-part.

By integration-by-part,

$$\langle g'', r'' \rangle = \int_a^b g''(x) r''(x) \mathrm{d}x = g''(x) r'(x)|_a^b - \int_a^b g'''(x) r'(x) \mathrm{d}x$$

Because g is a natural spline, $g''(a) = g''(b) = 0$. Then,

$$\begin{aligned}\langle g'', r'' \rangle &= -\int_a^b g'''(x) r'(x) \mathrm{d}x \\ &= -\sum_{j=1}^{N-1} \int_{x_j}^{x_{j+1}} g'''(x) r'(x) \mathrm{d}x - \int_a^{x_1} g'''(x) r'(x) \mathrm{d}x - \int_{x_N}^b g'''(x) r'(x) \mathrm{d}x\end{aligned}$$

Because g''' is constant in $[x_j, x_{j+1}]$, and $g'''(x) = 0$ when $x \leqslant x_1$ and $x \geqslant x_N$,

$$\begin{aligned}\langle g'', r'' \rangle &= -\sum_{j=1}^{N-1} g'''(x_j^+) \int_{x_j}^{x_{j+1}} r'(x) \mathrm{d}x \\ &= -\sum_{j=1}^{N-1} g'''(x_j^+)[r(x_{j+1}) - r(x_j)] \\ &= 0 \quad (r(x_j) = 0, j = 1, \ldots N)\end{aligned}$$

(b) Show the Pythagorean identity

$$\|\widetilde{g}''\|_{L^2(\mathrm{d}\mathfrak{m})}^2 = \|g''\|_{L^2(\mathrm{d}\mathfrak{m})}^2 + \|r''\|_{L^2(\mathrm{d}\mathfrak{m})}^2 \geqslant \|g''\|_{L^2(\mathrm{d}\mathfrak{m})}^2,$$

and conclude that

$$\int_a^b \widetilde{g}''(x)^2 \mathrm{d}x \geqslant \int_a^b g''(x)^2 \mathrm{d}x,$$

with equality if and only if $g = \widetilde{g}$ *a.e.* $\mathfrak{m}$[1].

$$\begin{aligned}
\|\widetilde{g}''\|^2_{L^2(\mathrm{d}\mathfrak{m})} &= \int_a^b \widetilde{g}''(x)^2 \mathrm{d}x \\
&= \int_a^b (g''(x) + r''(x))^2 \mathrm{d}x \\
&= \int_a^b g''(x)^2 \mathrm{d}x + \int_a^b r''(x)^2 \mathrm{d}x \quad (\int_a^b r''(x) g''(x) \mathrm{d}x = 0) \\
&= \|g''\|^2_{L^2(\mathrm{d}\mathfrak{m})} + \|r''\|^2_{L^2(\mathrm{d}\mathfrak{m})} \geqslant \|g''\|^2_{L^2(\mathrm{d}\mathfrak{m})}
\end{aligned}$$

Therefore,

$$\int_a^b \widetilde{g}''(x)^2 \mathrm{d}x \geqslant \int_a^b g''(x)^2 \mathrm{d}x.$$

Derive the condition for equality. If

$$\int_a^b \widetilde{g}''(x)^2 \mathrm{d}x = \int_a^b g''(x)^2 \mathrm{d}x,$$

then

$$\begin{aligned}
&\int_a^b r''(x)^2 \mathrm{d}x = \int_a^b \widetilde{g}''(x)^2 \mathrm{d}x - \int_a^b g''(x)^2 \mathrm{d}x = 0 \\
\Longrightarrow\ & r''(x) = 0 \\
\Longrightarrow\ & r(x) = 0 \quad (r(x_1) = r(x_N) = 0, \text{Liouville's Theorem})
\end{aligned}$$

Therefore, equality holds if and only if $g = \widetilde{g}$.

(c) Argue that the solution to the smoothing spline problem

$$\min_{f \in C^2[a,b]} \sum_{i=1}^N [y_i - f(x_i)]^2 + \lambda \int_a^b f''(x)^2 \mathrm{d}x \tag{1}$$

(for some $\lambda > 0$) must be a natural cubic spline with knots at $\{x_i\}_{i=1}^N$[2].

Hint. The solution $\widehat{f}_\lambda$ might not interpolate $\{(x_i, y_i)\}_{i=1}^N$, but one might consider g interpolates $\{(x_i, \widehat{f}_\lambda(x_i))\}_{i=1}^N$.

Let $\widehat{f}_\lambda(x)$ be the solution at a certain λ. Let $g(x)$ be the natural cubic spline with knots at $\{x_i\}_{i=1}^N$ that interpolates $\{(x_i, \widehat{f}_\lambda(x_i))\}_{i=1}^N$.

If $\widehat{f}_\lambda \neq g$, then

$$\int_a^b \widehat{f}''_\lambda(x)^2 \mathrm{d}x > \int_a^b g''(x)^2 \mathrm{d}x,$$

[1] A more profound result Liouville's Theorem from complex analysis, applied to PDE, dynamical system *etc.*, can be stated that any harmonic function (*i.e.* with second-order differential 0) being constant at boundary must remain constant over the entire domain.

[2] It suggests that the smoothing spline optimization over an infinite-dimensional functional space is actually a N-dimensional problem given the pairs $\{(x_i, y_i)\}_{i=1}^N$. We call $C^2[a,b]$ equipped with norm $f \mapsto \|f\|_{2,2} := \|f\|_2 + \|f''\|_2$ as the Sobolev space $W^{2,2}[a,b]$.

$$\sum_{i=1}^{N}[y_i - \widehat{f}_\lambda(x_i)]^2 = \sum_{i=1}^{N}[y_i - g(x_i)]^2 \quad (g(x_i) = \widehat{f}_\lambda(x_i))).$$

Then, $\widehat{f}_\lambda$ has a greater objective function value than g, which contradicts with that $\widehat{f}_\lambda$ is the minimal solution.

Therefore, $\widehat{f}_\lambda = g$.

4. (*15 pt*) (Leave-One-Out, Textbook Ex. 5.13) Let $\widehat{f}_\lambda$ be the solution to (1) given N pairs $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^N$ and $\lambda > 0$.

   (a) Suppose you augment the training sample with another pair $\left(x_0, \widehat{f}_\lambda(x_0)\right)$ and refit. Argue that the refitted solution remains unchanged.

   Let

   $$L_1(f) = \sum_{i=1}^{N}[y_i - f(x_i)]^2 + \lambda \int_a^b f''(x)^2 \mathrm{d}x,$$
   $$L_2(f) = [\widehat{f}_\lambda(x_0) - f(x_0)]^2.$$

   After augmentation, the new problem is

   $$\min_{f \in C^2[a,b]} L_1(f) + L_2(f).$$

   Because $\widehat{f}_\lambda$ minimizes $L_1(f)$ and $L_2(f)$ simultaneously, $\widehat{f}_\lambda$ minimizes $L_1(f) + L_2(f)$, i.e. $\widehat{f}_\lambda$ is the solution for the new problem.

   (b) Recall that with the smoother matrix $\mathbf{S}_\lambda = [s_{\lambda,ij}]_{N\times N}$, $\widehat{\boldsymbol{f}}_\lambda = \mathbf{S}_\lambda \boldsymbol{y}$ gives the vector of in-sample predictions. Show that the individual leave-one-out error

   $$e_i := y_i - \widehat{f}_\lambda^{(-i)}(x_i) = \frac{y_i - \widehat{f}_\lambda(x_i)}{1 - s_{\lambda,ii}} \quad (1 \leqslant i \leqslant N) \tag{2}$$

   where $\widehat{f}_\lambda^{(-i)}$ is prediction function based on sample $\mathcal{D}\backslash\{(x_i, y_i)\}$, hence prove the leave-one-out cross-validation (LOOCV) criteria[3]

   $$\mathrm{LOOCV}(\widehat{f}_\lambda) := \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{f}_\lambda^{(-i)}(x_i)\right)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \widehat{f}_\lambda(x_i)}{1 - s_{\lambda,ii}}\right)^2.$$

   Hint. Establish an equation by cooking up a refitting stable situation as in (a).

   Let $g_\lambda^{(-i)}$ be the prediction function based on sample $(\mathcal{D}\backslash\{(x_i, y_i)\}) \cup \{(x_i, \widehat{f}_\lambda^{(-i)}(x_i))\}$. By conclu-

[3]It generalizes the influence measures as discussed in STOR 664, where $\mathbf{S}_\lambda = \mathbf{H} = [h_{ij}]_{N\times N}$ is the hat matrix, h_{ii} is called the **leverage**, (2) is the **DFFITS** for the i-th observation, and its square proportionates to the **Cook's distance**.

sion from (a), $g_\lambda^{(-i)} = \widehat{f}_\lambda^{(-i)}$. The smoother matrix of $g_\lambda^{(-i)}$ is $\mathbf{S}_\lambda$. Therefore,

$$\begin{aligned} g_\lambda^{(-i)}(x_i) &= \sum_{j\neq i} s_{\lambda,ij} y_j + s_{\lambda,ii} \widehat{f}_\lambda^{(-i)}(x_i) \\ &= \sum_{j=1}^{N} s_{\lambda,ij} y_j - s_{\lambda,ii} y_i + s_{\lambda,ii} \widehat{f}_\lambda^{(-i)}(x_i) \\ &= \widehat{f}_\lambda(x_i) - s_{\lambda,ii} y_i + s_{\lambda,ii} \widehat{f}_\lambda^{(-i)}(x_i) \end{aligned}$$

Then,

$$\begin{aligned} &\widehat{f}_\lambda^{(-i)}(x_i) = g_\lambda^{(-i)}(x_i) = \widehat{f}_\lambda(x_i) - s_{\lambda,ii} y_i + s_{\lambda,ii} \widehat{f}_\lambda^{(-i)}(x_i) \\ \Longrightarrow\ &\widehat{f}_\lambda^{(-i)}(x_i) = \frac{\widehat{f}_\lambda(x_i) - s_{\lambda,ii} y_i}{1 - s_{\lambda,ii}} \\ \Longrightarrow\ &y_i - \widehat{f}_\lambda^{(-i)}(x_i) = \frac{y_i - \widehat{f}_\lambda(x_i)}{1 - s_{\lambda,ii}} \end{aligned}$$

Computational Part

1. (*20 pt*) **Simulation**

   Consider the simulation setup as in Section 5.5.2

   $$X \sim \textbf{Uniform}[0,1] \perp\!\!\!\perp \epsilon \sim \mathcal{N}(0,1), \quad f(X) := \frac{\sin(12(X+0.2))}{X+0.2}, \quad Y = f(X) + \epsilon$$

   with $N = 100$ randomly generated training sample. Fit polynomial regression, B-spline, natural cubic spline, smoothing spline and local polynomial regression with various kernels. Use cross-validation to tune any parameters. Compare their performances on a 10,000 test set.

2. (*20 pt*) **Zip Code Digit Data** (Textbook Ex. 6.12)

   Write a computer program to perform a local linear discriminant analysis. At each query point x_0, the training data $\{(x_i, y_i)\}_{i=1}^N$ receive weights $\{K_\lambda(x_0, x_i)\}_{i=1}^n$ from a weighting kernel K_λ, and return a weighted least-square discriminant prediction. Try out your program on the *Zip Code Digits Data* to discriminate 3's and 8's with various kernel functions.

   Hint. R package `kernlab` might be helpful to get various kernel functions and compute their linear form and quadratic form efficiently. Parameters indexing the kernel function should be tuned via cross-validation.

3. (*20 pt*) **Phoneme Recognition Data**

   One can use splines not only to increase flexibility of the functional modeling, but also to reduce the flexibility. In Section 5.2.3, natural cubic splines are used to simplify the input signals which have strong positive autocorrelation. Reproduce the analysis as is done in the textbook and also work on Ex. 5.5. The *Phoneme Recognition Data* are available at https://web.stanford.edu/~hastie/ElemStatLearn/datasets/phoneme.data. Report both the R code and the results.

References

[1] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Springer-Verlag, https://web.stanford.edu/~hastie/Papers/ESLII.pdf, second edition, 2009. 1