DS-GA 3001.009 Applied Statistics: Homework #8 Solutions

Due on Thursday, December 14, 2023

Please hand in your homework via Gradescope (entry code: RKXJN2) before 11:59 PM.

1. Find the natural cubic spline $f(x)$ with $f(0) = 0$, $f(1) = 2$, and $f(2) = 3$. Specifically, you should specify the coefficients $(a_0, a_1, a_2, a_3, b_0, b_1, b_2, b_3)$ such that

$$f(x) = \begin{cases} g(x) = a_3x^3 + a_2x^2 + a_1x + a_0 & \text{if } x \in [0, 1], \\ h(x) = b_3(x-2)^3 + b_2(x-2)^2 + b_1(x-2) + b_0 & \text{if } x \in [1, 2], \end{cases}$$

where $g(1) = h(1)$, $g'(1) = h'(1)$, $g''(1) = h''(1)$, and $f''(0) = f''(2) = 0$ (boundary conditions for the natural spline). Provide details of how you arrive at your answer.

Solution: By $f(0) = 0$ and $f''(0) = 0$ we conclude that $a_0 = a_2 = 0$. Similarly, by $f(2) = 3$ and $f''(2) = 0$, we have $b_0 = 3, b_2 = 0$. The rest of the conditions gives

$$\begin{cases} a_3 + a_1 = 2 \\ -b_3 - b_1 + 3 = 2 \\ 3a_3 + a_1 = 3b_3 + b_1 \\ 6a_3 = -6b_3 \end{cases} \implies \begin{cases} a_1 = 9/4 \\ a_3 = -1/4 \\ b_1 = 3/4 \\ b_3 = 1/4 \end{cases}.$$

Consequently,

$$f(x) = \begin{cases} (9x - x^3)/4 & \text{if } 0 \le x \le 1, \\ (x^3 - 6x^2 + 15x - 2)/4 & \text{if } 1 \le x \le 2. \end{cases}$$

2. In wavelet shrinkage, the soft and hard thresholding estimator aim to minic the *ideal truncated estimator*. Consider a one-dimensional Gaussian location model $y \sim \mathcal{N}(\theta, \sigma^2)$ with known σ; there is an (unknown) upper bound τ of $|\theta|$, i.e. $|\theta| \le \tau$.

   (a) For the MLE $\widehat{\theta}_1(y) = y$, compute the worst-case MSE $\max_{|\theta|\le\tau} \mathbb{E}_\theta[(\widehat{\theta}_1(y) - \theta)^2]$.

   (b) For the zero estimator $\widehat{\theta}_2(y) \equiv 0$, compute the worst-case MSE $\max_{|\theta|\le\tau} \mathbb{E}_\theta[(\widehat{\theta}_2(y) - \theta)^2]$.

   (c) The ideal truncated estimator assumes that θ is known, but forces the learner to use either $\widehat{\theta}_1(y)$ or $\widehat{\theta}_2(y)$. In other words, the learner finds a subset $R = R(\sigma) \subseteq \mathbb{R}$ based on the knowledge of σ, and uses

$$\widehat{\theta}(y) = \begin{cases} \widehat{\theta}_1(y) & \text{if } \theta \in R, \\ \widehat{\theta}_2(y) & \text{if } \theta \notin R. \end{cases}$$

   Which choice of R minimizes the worst-case MSE $\max_{|\theta|\le\tau} \mathbb{E}_\theta[(\widehat{\theta}(y) - \theta)^2]$? The resulting estimator is known as the ideal truncated estimator. What is the worst-case MSE for the ideal truncated estimator?

Solution:

(a) Since $\mathbb{E}_\theta[(y-\theta)^2] = \sigma^2$ for all θ, the worst-case MSE is σ^2.

(b) Since $\mathbb{E}_\theta[(0-\theta)^2] = \theta^2$, the worst-case MSE is $\max_{|\theta|\le\tau}\{\theta^2\} = \tau^2$.

(c) By (a) and (b), it is clear that

$$\mathbb{E}_\theta[(\widehat{\theta}(y)-\theta)^2] = \begin{cases} \sigma^2 & \text{if } \theta \in R, \\ \theta^2 & \text{if } \theta \notin R. \end{cases}$$

It is then clear that the learner should prefer the MLE if and only if $|\theta| \ge \sigma$. In other words, we choose $R = \{\theta : |\theta| \ge \sigma\}$, so that the ideal truncated estimator is $\widehat{\theta}^{\text{ITE}}(y) = y\mathbb{1}(|\theta| \ge \sigma)$. The worst-case MSE is

$$\max_{|\theta|\le\tau} \mathbb{E}_\theta[(\widehat{\theta}^{\text{ITE}}(y)-\theta)^2] = \min\{\sigma^2, \tau^2\}.$$

3. In class we have shown that the local polynomial fit $\widehat{f}_k(x_0)$ of degree k for $f(x_0)$ takes the form $\widehat{f}_k(x_0) = \sum_{i=1}^n w_k(x_0, x_i)y_i$. In this problem we aim to show that

$$\sum_{i=1}^n w_k(x_0, x_i)^2 \le \sum_{i=1}^n w_{k+1}(x_0, x_i)^2,$$

and therefore the variance of the fit becomes larger when one increases the polynomial degree. The proof relies on the following result in linear algebra:

Lemma 1. *For any positive definite matrix M with a block-wise form*

$$M = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix},$$

where A and C are symmetric square matrices, the matrix

$$M^{-1} - \begin{bmatrix} A^{-1} & \\ & O \end{bmatrix}$$

is positive semi-definite, where O is the all-zero matrix.

(a) Use the lemma and the matrix-form expression of $w_k(x_0, x_i)$ derived in class, prove that for the box kernel $K(x) = \mathbb{1}(|x| \le 1/2)$ and any bandwidth parameter $h > 0$, it holds that $\sum_{i=1}^n w_k(x_0, x_i)^2 \le \sum_{i=1}^n w_{k+1}(x_0, x_i)^2$.

(b) (*Bonus 5 points*) Prove the lemma.

Solution:

(a) By the lecture note, the weight vector $w_k = (w_k(x_0, x_1), \cdots, w_k(x_0, x_n))$ takes the form of

$$w_k = v_k (X_k^\top D X_k)^{-1} X_k^\top D,$$

where

$$\begin{aligned} v_k &= \begin{bmatrix} 1 & x_0 & \cdots & x_0^k \end{bmatrix} \in \mathbb{R}^{1\times(k+1)}, \\ X_k &= \begin{bmatrix} 1 & x_1 & \cdots & x_1^k \\ 1 & x_2 & \cdots & x_2^k \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^k \end{bmatrix} \in \mathbb{R}^{n\times(k+1)}, \\ D &= \operatorname{diag}(K_h(x_0 - x_1), \cdots, K_h(x_0 - x_n)) \in \mathbb{R}^{n\times n}. \end{aligned}$$

Note that D does not depend on k, and

$$v_{k+1} = \begin{bmatrix} v_k & x_0^{k+1} \end{bmatrix}, \quad X_{k+1} = \begin{bmatrix} & x_1^{k+1} \\ & x_2^{k+1} \\ X_k & \vdots \\ & x_n^{k+1} \end{bmatrix}.$$

As a result,

$$\begin{aligned} \sum_{i=1}^n w_k(x_0, x_i)^2 = w_k w_k^\top &= v_k (X_k^\top D X_k)^{-1} (X_k^\top D^2 X_k)(X_k^\top D X_k)^{-1} v_k^\top \\ &= \frac{1}{h} v_k (X_k^\top D X_k)^{-1} v_k^\top, \end{aligned}$$

where the last step is due to $K_h(x)^2 = K_h(x)/h$ everywhere, so that $D^2 = D/h$. By the recursive structure between X_k and X_{k+1}, it is easily seen that

$$X_{k+1}^\top D X_{k+1} = \begin{bmatrix} X_k^\top D X_k & \star \\ \star & \star \end{bmatrix}.$$

Therefore, by the lemma,

$$\begin{aligned} \sum_{i=1}^n w_k(x_0, x_i)^2 &= \frac{1}{h} v_k (X_k^\top D X_k)^{-1} v_k^\top \\ &= \frac{1}{h} \begin{bmatrix} v_k & x_0^{k+1} \end{bmatrix} \begin{bmatrix} (X_k^\top D X_k)^{-1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} v_k & x_0^{k+1} \end{bmatrix}^\top \\ &\le \frac{1}{h} \begin{bmatrix} v_k & x_0^{k+1} \end{bmatrix} (X_{k+1}^\top D X_{k+1})^{-1} \begin{bmatrix} v_k & x_0^{k+1} \end{bmatrix}^\top \\ &= \frac{1}{h} v_{k+1} (X_{k+1}^\top D X_{k+1})^{-1} v_{k+1}^\top \\ &= \sum_{i=1}^n w_{k+1}(x_0, x_i)^2. \end{aligned}$$

(b) Applying Gauss elimination to M gives

$$M = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} = \begin{bmatrix} I & O \\ B^\top A^{-1} & I \end{bmatrix} \begin{bmatrix} A & O \\ O & C - B^\top A^{-1} B \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ O & I \end{bmatrix}.$$

Taking inversion at both sides gives

$$\begin{aligned} M^{-1} &= \begin{bmatrix} I & -A^{-1}B \\ O & I \end{bmatrix} \begin{bmatrix} A^{-1} & O \\ O & (C - B^\top A^{-1} B)^{-1} \end{bmatrix} \begin{bmatrix} I & O \\ -B^\top A^{-1} & I \end{bmatrix} \\ &= \begin{bmatrix} A^{-1} & \\ & O \end{bmatrix} + \begin{bmatrix} -A^{-1}B \\ I \end{bmatrix} (C - B^\top A^{-1} B)^{-1} \begin{bmatrix} -A^{-1}B & I \end{bmatrix}. \end{aligned}$$

The second term takes the form of $U^\top V U$ with $V = (C - B^\top A^{-1} B)^{-1}$ being PSD (since M is), and is consequently a PSD matrix.

4. Coding: we will use the Doppler example in "Donoho, D.L. and Johnstone, I.M. (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika, 81, 425–455" to visualize and test the performances of the nonparametric estimators we learned in class. In this problem we will implement the following estimators:

   (a) Nadaraya–Watson estimator, with a data-driven bandwidth;

   (b) local polynomial regressors, with $d \in \{1, 20\}$;

   (c) cubic smoothing and regression splines;

   (d) Fourier projection estimator;

   (e) wavelet (soft and hard) thresholding estimators.

   Based on inline instructions, fill in the missing codes in `https://tinyurl.com/2aesjazk`. Be sure to submit a pdf with your codes, outputs, and colab link.

   Solution: see `https://tinyurl.com/27eaz5xe`.