Statistical learning homework 3

Dai Yuehao (1800010660@pku.edu.cn)

November 23, 2020

1 ESL 5.1

Assume that a cubic spline $f(x)$ with knots at ξ_1 and ξ_2 takes the form

$$f(x) = a_i x^3 + b_i x^2 + c_i x + d_i$$

with $i = 1, 2, 3$ in $(-\infty, \xi_1]$, $[\xi_1, \xi_2]$, $[\xi_2, +\infty)$ respectively, then

$$f_1(x) = f(x) - a_1 h_4(x) - b_1 h_3(x) - c_1 h_2(x) - d_1 h_1(x)$$

takes value 0 in $(-\infty, \xi_1]$, and $f_1'(\xi_1) = f_1''(\xi_1) = 0$, which means that $f_1(x) = (a_2 - a_1)x^3$ in $[\xi_1, \xi_2]$, then

$$f_2(x) = f_1(x) - (a_2 - a_1)h_5(x)$$

is twice differentiable in $[\xi_1, x_2]$ and takes value 0 in that interval, also satisfies $f_2'(\xi_2) = f_2''(\xi_2) = 0$, which means that

$$f_2(x) = (a_3 - 2a_2 + a_1)h_6(x)$$

in $[\xi_2, \infty)$, then let

$$f_3(x) = f_2(x) - (a_3 - 2a_2 + a_1)h_6(x)$$

we have $h_3(x) = 0$ in $\mathbb{R}$. Thereby we have

$$f(x) = a_1 h_4(x) + b_1 h_3(x) + c_1 h_2(x) + d_1 h_1(x) + (a_2 - a_1)h_5(x) + (a_3 - 2a_2 + a_1)h_6(x),$$

thereby we have already conclued the proof.

2 ESL 5.4

We hope that $f''(X) = 0$ when $X \leq \xi_1$ and $X \geq X_K$, since $F(X) = \sum_{j=0}^{3} X^j$ when $X \leq \xi_1$, one must have $\beta_2 = \beta_3 = 0$. Since

$$f(X) = \sum_{k=1}^{K} \theta_k (X - \xi_k)^3$$

when $X \geq \xi_K$, thus the coefficient of X^3 is $\sum_{k=1}^{K} \theta_k$, hence it must be 0. Similarly the coefficient of X^2 is $-3\sum_{k=1}^{K} \xi_k \theta_k$, hence it must be 0, too. Combine all of these we have

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

3 ESL 5.7

3.1 a

In terms of integration by parts we have

$$\int_a^b g''(x)h''(x)dx = g''(x)h'(x)\Big|_a^b - \int_a^b g'''(x)h'(x) = -\int_a^b g'''(x)h'(x)dx,$$

this is because g is natural cubic splines and is linear out of the boundary of samples and thus $g''(a) = g''(b) = 0$. Then since $g'''(x)$ is a constant in each interval $[x_j, x_{j+1}]$ and 0 in $[a, x_1]$ and $[x_N, b]$, we have

$$-\int_a^b g'''(x)h'(x)dx = -\sum_{j=1}^{N-1} g(x_j^+)\int_{x_j}^{x_{j+1}} h'(x)dx = -\sum_{j=1}^{N-1} g(x_j^+)[h(x_{j+1}) - h(x_j)] = 0$$

since $h(x_j) = 0$ for all $j = 1, \cdots, N$.

3.2 b

Now we have

$$\int_a^b [\tilde{g}''(x)^2 - g''(x)^2]dx = \int_a^b [\tilde{g}''(x) - g''(x)][\tilde{g}''(x) + g''(x)] = \int h''(x)[h''(x) + 2g''(x)]dx = \int_a^b h''(x)^2 dx \geq 0,$$

since $h(x)$ is a polynomial function with degree no more than 3, $h''(x)$ can be written as $h''(x) = a_j x + b_j$ in each interval, thus

$$\int_{x_j}^{x_{j+1}} h''(x)^2 dx = 0$$

if and only if $h''(x) = 0$ in $[x_j, x_{j+1}]$, hence we conclude the proof.

3.3 c

For each twice differential function f, there exists a cubic spline g such that $g(x_i) = f(x_i)$ and with the previous result

$$\int_a^b g''(t)dt \leq \int_a^b f''(t)dt,$$

thus g is better than f, hence the minimizer must be a cubic spline with knots at each of the x_i.

4 ESL 5.13

Let $f_1(x)$ denote the new smoothing spline trained by $N + 1$ samples $(x_0, \hat{f}(x_0)), (x_i, y_i), i = 1, \cdots N$, then $\hat{f}$ is the optimal solution of the optimization problem

$$\min_f \quad \mathrm{RSS}(f) = \sum_{i=1}^N \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 dt$$

and f_1 is the optimal solution of the optimization problem

$$\min_f \quad \mathrm{RSS}(f) = \left[f(x_0) - \hat{f}(x_0)\right]^2 + \sum_{i=1}^N \{y_i - f(x_i)\}^2 + \lambda \int \{f''(t)\}^2 \mathrm{d}t, \tag{4.0.1}$$

hence

$$\begin{aligned}
&\Big[f(x_0)-\hat f(x_0)\Big]^2+\sum_{i=1}^N\{y_i-f(x_i)\}^2+\lambda\int\{f''(t)\}^2\mathrm{d}t\\
\geq&\sum_{i=1}^N\{y_i-f(x_i)\}^2+\lambda\int\{f''(t)\}^2\mathrm{d}t\\
\geq&\sum_{i=1}^N\{y_i-\hat f(x_i)\}^2+\lambda\int\{\hat f''(t)\}^2\mathrm{d}t\\
=&\Big[\hat f(x_0)-\hat f(x_0)\Big]^2+\sum_{i=1}^N\{y_i-\hat f(x_i)\}^2+\lambda\int\{\hat f''(t)\}^2\mathrm{d}t
\end{aligned}$$

which yields that $\hat f$ is the optimal solution of problem(4.0.1), hence the refited smoothing spline is the one without the augmented sample.

To yield the formula (5.26), it is sufficient to verify that

$$y_i-\hat f_\lambda^{(-i)}(x_i)=\frac{y_i-\hat f_\lambda(x_i)}{1-S_\lambda(i,i)}.$$

We have

$$\begin{aligned}
y_i-\hat f_\lambda(x_i)&=y_i-\sum_{j=1}^N S_\lambda(i,j)y_j\\
&=y_i-S_\lambda(i,i)y_i-\left[S_\lambda(i,i)\hat f_\lambda^{(-i)}(x_i)+\sum_{j=1,j\neq i}^N S_\lambda(i,j)y_j-S_\lambda(i,i)\hat f_\lambda^{(-i)}(x_i)\right]\\
&=y_i-S_\lambda(i,i)y_i-[\hat f_\lambda^{(-i)}(x_i)-S_\lambda(i,i)\hat f_\lambda^{(-i)}(x_i)]\\
&=[1-S_\lambda(i,i)][y_i-\hat f_\lambda^{(-i)}(x_i)],
\end{aligned}$$

the key step is the equation

$$S_\lambda(i,i)\hat f_\lambda^{(-i)}(x_i)+\sum_{j=1,j\neq i}^N S_\lambda(i,j)y_j=\hat f_\lambda^{(-i)}(x_i), \tag{4.0.2}$$

this the direct consequence from the result we have proved: assume that $S_\lambda^{(-i)}$ is the smoother matrix trained without (x_i,y_i) but augmented by $(x_i,\hat f_\lambda^{(-i)}(x_i))$, since

$$S_\lambda^{(-i)}=\boldsymbol{N}(\boldsymbol{N}^T\boldsymbol{N}+\lambda\Omega_{\boldsymbol{N}})^{-1}\boldsymbol{N}^T$$

only depends on $\{x_i\}$ and λ, $S_\lambda^{(-i)}=S_\lambda$, hence equation (4.0.2) holds and we have concluded the proof.

5 ESL 5.15

5.1 a

In this case $\{\sqrt{\gamma_i}\phi(x)\}$ is a basis of $\mathcal{H}_K$ hence $\langle\phi_i,\phi_i\rangle=1/\gamma_i$, thus we have

$$\begin{aligned}
\langle K(\cdot,x_i),f\rangle&=\left\langle\sum_{j=1}^\infty\gamma_j\phi_j\phi_j(x_i),\sum_{j=1}^\infty c_j\phi_j(x_i)\right\rangle\\
&=\sum_{j=1}^\infty\gamma_jc_j\left\langle\phi_j,\phi_j\phi_j(x_i)f\right\rangle\\
&=\sum_{j=1}^\infty c_j\phi_j(x_i)=f(x_i)
\end{aligned}$$

5.2 b

From ESL 5.15(a) we let $f = K(\cdot, x_j)$ and immediately yield

$$\langle K(\cdot, x_i), K(\cdot, x_j)\rangle = K(x_i, x_j).$$

5.3 c

We have

$$\begin{aligned} J(g) = \langle g, g\rangle &= \left\langle \sum_{i=1}^N \alpha_i K(\cdot, x_i), \sum_{i=1}^N \alpha_i K(\cdot, x_i)\right\rangle \\ &= \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \left\langle K(\cdot, x_i), K(\cdot, x_j)\right\rangle \\ &= \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j K(x_i, x_j). \end{aligned}$$

5.4 d

Since $\rho(x)$ is orthogonal to each $K(\cdot, x_i)$ hence $\langle K(\cdot, x_i), \rho\rangle = 0$, we have

$$\begin{aligned} J(\tilde{g}) &= \langle g+\rho, g+\rho\rangle \\ &= J(g) + J(\rho) + 2\sum_{i=1}^N \alpha_i \left\langle K(\cdot, x_i), \rho\right\rangle = J(g) + J(\rho) \geq J(g), \end{aligned}$$

on the other hand we have

$$\tilde{g}(x_i) = \langle K(\cdot, x_i), g+\rho\rangle = \langle K(\cdot, x_i), g\rangle = g(x_i),$$

thereby

$$\sum_{i=1}^N L(y_i, \tilde{g}(x_i)) + \lambda J(\tilde{g}) = \sum_{i=1}^N L(y_i, g(x_i)) + \lambda J(g) + \lambda J(\rho) \geq \sum_{i=1}^N L(y_i, g(x_i)) + \lambda J(g),$$

with equaility if and only if $J(\rho) = 0$, which is equivelant as $\rho = 0$.

6 ESL 6.2

From the standard linear regression we have

$$\hat{f}(x_0) = b(x_0)^T(\boldsymbol{B}^T\boldsymbol{W}(x_0)\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{W}(x_0)\boldsymbol{Y} = \sum_{i=1}^N l_i(x_0) y_i,$$

thus

$$b(x_0)^T(\boldsymbol{B}^T\boldsymbol{W}(x_0)\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{W}(x_0)\boldsymbol{B} = b(x_0)^T = \begin{pmatrix}1 & x_0\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^N l_i(x_0) & \sum_{i=1}^N x_i l_i(x_0)\end{pmatrix}$$

hence

$$\sum_{i=1}^N l_i(x_0) = 1, \quad \sum_{i=1}^N (x - x_i) l_i(x_0) = x\sum_{i=1}^N l_i(x_0) - \sum_{i=1}^N x_i l_i(x_0) = x_0 - x_0 = 0.$$

For $j \in \{1, \cdots, k\}$, similarly we still have

$$x_0^j = \sum_{i=1}^N l_i(x_0) x_i^j,$$

then

$$b_j(x_0) = \sum_{k=0}^{j}\sum_{i=1}^{N}(-1)^k C_j^k x_i^k x_0^{j-k} l_i(x_0) = \sum_{k=0}^{j}(-1)^k C_j^k x_0^j = 0.$$

Now we have

$$\begin{aligned}
\mathbb{E}[\hat{f}(x_0) - f(x_0)] &= \sum_{i=1}^{N} l_i(x_0) f(x_i) - f(x_0) \\
&= \left[f(x_0)\sum_{i=1}^{N} l_i(x_0) - f(x_0)\right] + \sum_{j=1}^{\infty}\frac{1}{j!}f^{(j)}(x_0)\sum_{i=1}^{N}(x_i - x_0)^j l_i(x_0) \\
&= \sum_{j=k+1}^{\infty}\frac{1}{j!}f^{(j)}(x_0)\sum_{i=1}^{N}(x_i - x_0)^j l_i(x_0)
\end{aligned}$$

which means that the bias only depends on terms higher than $k+1$.

7 ESL 6.3

Assume that

$$\boldsymbol{B}_q = \begin{pmatrix}\boldsymbol{1} & \boldsymbol{x} & \cdots & \boldsymbol{x}^q\end{pmatrix}$$

consider $q < k$, we have

$$l(x_0)^T = b(x_0)^T(\boldsymbol{B}^T\boldsymbol{W}(x_0)\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{W}(x_0),$$

since $||l(x_0)||^2 = l(x_0)^T l(x_0)$, we need to varify that

$$\begin{aligned}
&b_k(x_0)^T(\boldsymbol{B}_k^T\boldsymbol{W}(x_0)\boldsymbol{B}_k)^{-1}\boldsymbol{B}_k^T\boldsymbol{W}(x_0)\boldsymbol{W}(x_0)\boldsymbol{B}_k(\boldsymbol{B}_k^T\boldsymbol{W}(x_0)\boldsymbol{B}_k)^{-1}b_k(x_0) \\
\geq &b_q(x_0)^T(\boldsymbol{B}_q^T\boldsymbol{W}(x_0)\boldsymbol{B}_q)^{-1}\boldsymbol{B}_q^T\boldsymbol{W}(x_0)\boldsymbol{W}(x_0)\boldsymbol{B}_q(\boldsymbol{B}_q^T\boldsymbol{W}(x_0)\boldsymbol{B}_p)^{-1}b_q(x_0).
\end{aligned} \tag{7.0.1}$$

From ESL 6.2 we know that

$$\boldsymbol{B}_k^T l_k(x_0) = b_k(x_0),$$

conseder the optimization problem

$$\begin{aligned}
\min_{l_k(x_0)} \quad & \frac{1}{2} l_k(x_0)^T\boldsymbol{W}(x_0)^{-1} l_k(x_0) \\
\text{s.t.} \quad & \boldsymbol{B}_k^T l_k(x_0) = b_k(x_0)
\end{aligned} \tag{7.0.2}$$

this is a convex problem and the Slater's condition is satisfied, and since the Lagrange function of problem (7.0.2) is

$$L(l_k(x_0), \lambda) = \frac{1}{2} l_k(x_0)^T\boldsymbol{W}(x_0)^{-1} l_k(x_0) - \lambda^T\left[\boldsymbol{B}_k^T l_k(x_0) - b_k(x_0)\right],$$

then we have

$$\frac{\partial L}{\partial l_k^*(x_0)} = \frac{\partial L}{\partial \lambda^*} = 0,$$

which yields

$$l_k(x_0)^* = \boldsymbol{W}(x_0)\boldsymbol{B}_k(\boldsymbol{B}_k^T\boldsymbol{W}(x_0)\boldsymbol{B}_k)^{-1} b_k(x_0),$$

hence $l_k(x_0)$ is the optimal solution to problem (7.0.2). Now consider another optimization problem

$$\begin{aligned}
\min_{l_q(x_0)} \quad & \frac{1}{2} l_q(x_0)^T\boldsymbol{W}(x_0)^{-1} l_q(x_0) \\
\text{s.t.} \quad & \begin{pmatrix}\boldsymbol{B}_q^T \\ \boldsymbol{0}\end{pmatrix} l_q(x_0) = \begin{pmatrix} b_q(x_0) \\ \boldsymbol{0}\end{pmatrix}
\end{aligned} \tag{7.0.3}$$

since

$$\boldsymbol{B}_k^T = \begin{pmatrix} \boldsymbol{B}_q^T \\ \boldsymbol{B}_{k-q}^T \end{pmatrix}, \quad b_k(x_0) = \begin{pmatrix} b_q(x_0) \\ b_{k-q}(x_0) \end{pmatrix}$$

where

$$\boldsymbol{B}_{k-q} = \begin{pmatrix} \boldsymbol{x}^{p+1} & \cdots & \boldsymbol{x}^k \end{pmatrix}, \quad b_{k-q}(x_0)^T = \begin{pmatrix} x_0^{p+1} & \cdots & x_0^k \end{pmatrix},$$

we know that each feasible solution to problem (7.0.2) is also feasible to problem (7.0.3), hence the optimal value of (7.0.3) is less than or equal to the optimal value of (7.0.2), hence

$$l_q(x_0)^T \boldsymbol{W}(x_0)^{-1} l_q(x_0) \leq l_k(x_0)^T \boldsymbol{W}(x_0)^{-1} l_k(x_0).$$

Notes[1]: The conclusion of this exercise seems to be wrong. We can construct a counterexample as follows:

$$b_0(x_0) = \begin{pmatrix} 1 \end{pmatrix}, \quad b_1(x_0) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \quad \boldsymbol{B}_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{B}_1 = \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}, \quad \boldsymbol{W}(x_0) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},$$

here $N = 2$, then we can calculate that

$$l_0(x_0)^T = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \end{pmatrix} \quad l_1(x_0)^T = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

where $||l_0(x_0)||^2 = 5/9 > ||l_1(x_0)||^2 = 1/2$. The promal conclusion holds only when $W(x_0) = k\boldsymbol{I}$.

8 ESL 6.5

Let $y_{ij} = 1$ if $g_i = j$ else 0, then the local log-likelihood function can be written as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} K_\lambda(x_0, x_i) \left\{ \sum_{j=1}^{J} y_{ij}\beta_j - \log\left[1 + \sum_{j=1}^{J-1} \exp \beta_j \right] \right\},$$

take the derivative of β_j and let it be 0 we have

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N} K_\lambda(x_0, x_i) \left[y_{ij} - \frac{\exp(\beta_j)}{1 + \sum_{j=1}^{J-1} \exp \beta_j} \right] = \sum_{i=1}^{N} K_\lambda(x_0, x_i)(y_{ij} - P\{g_{x_0} = j\}) = 0,$$

hence

$$P\{g_{x_0} = 1\} = \sum_{i=1}^{N} y_i \frac{K_\lambda(x_0, x_i)}{\sum_{i=1}^{N} K_\lambda(x_0, x_i)}$$

which is exactly the result of smoothing each class separately under a Nadaraya-Watson kernel smoother.

[1] The counterexample is implemented in MATLAB