第五次作业

Ext 5.9

Ex. 5.9 Derive the Reinsch form $\mathbf{S}_\lambda = (\mathbf{I} + \lambda\mathbf{K})^{-1}$ for the smoothing spline.

$$\begin{aligned} \mathrm{S} &= \mathbf{N}(\mathbf{N}^T\mathbf{N} + \lambda\mathbf{\Omega}_N)^{-1}\mathbf{N}^T \\ &= \mathbf{N}(\mathbf{N}^T(\mathbf{I} + \lambda(\mathbf{N}^T)^{-1}\mathbf{\Omega}_N\mathbf{N}^{-1})\mathbf{N})^{-1}\mathbf{N}^T \\ &= (\mathbf{I} + \lambda\mathbf{K})^{-1} \end{aligned}$$

Ext 5.13

Ex. 5.13 You have fitted a smoothing spline $\hat{f}_\lambda$ to a sample of N pairs (x_i, y_i). Suppose you augment your original sample with the pair $x_0, \hat{f}_\lambda(x_0)$, and refit; describe the result. Use this to derive the N-fold cross-validation formula (5.26).

Follow the notation in textbook, we define $\hat{f}_\lambda^{-i}(x_i)$ denotes the prediction which doesn't use $\{x_i, y_i\}$ doing the fit.

And we use this lemma without proof(Ext 7.3 (a)):

$$\hat{f}_\lambda^{(-i)}(x_i) = \frac{1}{1 - S_\lambda(i,i)} \sum_{j \neq i} S_\lambda(i,j) y_j.$$

so

$$\hat{f}_\lambda^{(-i)}(x_i) = \sum_{j \neq i} S_\lambda(i,j) y_j + S_\lambda(i,i) \hat{f}_\lambda^{(-i)}(x_i).$$

recall that :

$$\hat{f}_\lambda(x_i) = \sum_{j=1}^{n} S_\lambda(i,j)y_j,$$

We bring into the equation and with appropriate additions and subtractions we can get：

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

which is exactly what we want.

Ext 5.15

Ex. 5.15 This exercise derives some of the results quoted in Section 5.8.1. Suppose $K(x,y)$ satisfying the conditions (5.45) and let $f(x) \in \mathcal{H}_K$. Show that

(a)

(a) $\langle K(\cdot, x_i), f\rangle_{\mathcal{H}_K} = f(x_i)$.

We use the properties of inner products to expand, noting that the different bases are orthogonal to each other.

so:

$$\begin{aligned}
\langle K(\cdot, y), f\rangle_{\mathcal{H}_K} &= \left\langle \sum_{i=1}^{\infty}(\gamma_i\phi_i(x))\phi_i(y), \sum_{i=1}^{\infty} c_i\phi_i(x)\right\rangle \\
&= \sum_{i=1}^{\infty} \frac{c_i\lambda_i\phi_i(y)}{\lambda_i} \\
&= f(y).
\end{aligned}$$

(b)

(b) $\langle K(\cdot, x_i), K(\cdot, x_j)\rangle_{\mathcal{H}_K} = K(x_i, x_j)$.

just let $f = K(\cdot, x_j)$, and we can easily get the equation.

(c)

(c) If $g(x) = \sum_{i=1}^{N} \alpha_i K(x, x_i)$, then

$$J(g) = \sum_{i=1}^{N}\sum_{j=1}^{N} K(x_i, x_j)\alpha_i\alpha_j.$$

Suppose that $\tilde{g}(x) = g(x) + \rho(x)$, with $\rho(x) \in \mathcal{H}_K$, and orthogonal in $\mathcal{H}_K$ to each of $K(x, x_i),\ i = 1, \ldots, N$. Show that

Use the result in (b) and we get:

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

(d)

(d)

$$\sum_{i=1}^{N} L(y_i, \tilde{g}(x_i)) + \lambda J(\tilde{g}) \geq \sum_{i=1}^{N} L(y_i, g(x_i)) + \lambda J(g) \qquad (5.74)$$

with equality iff $\rho(x) = 0$.

If $\forall i, \langle \rho, K(x, x_i) \rangle = 0$, we get

$$\lambda J(\tilde{g}) = \lambda J(g) + \lambda \|\rho\|_{\mathcal{H}_K}^2 \geq \lambda J(g).$$

Moreover, we get that:

$$\begin{aligned}\tilde{g}(x_i) &= \langle K(\cdot, x_i), \tilde{g} \rangle_{\mathcal{H}_K} \\ &= \langle K(\cdot, x_i), g + \rho \rangle_{\mathcal{H}_K} \\ &= \langle K(\cdot, x_i), g \rangle_{\mathcal{H}_K},\end{aligned}$$

thus,

$$L(y_i, \tilde{g}(x_i)) = L(y_i, g(x_i))$$

so the loss just depends on the data space.

Ext 5.16

Ex. 5.16 Consider the ridge regression problem (5.53), and assume $M \geq N$. Assume you have a kernel K that computes the inner product $K(x, y) = \sum_{m=1}^{M} h_m(x) h_m(y)$.

(a)

(a) Derive (5.62) on page 171 in the text. How would you compute the matrices $\mathbf{V}$ and $\mathbf{D}_\gamma$, given K? Hence show that (5.63) is equivalent to (5.53).

By definition of the kernel K,we have

$$K(x,y)=\sum_{m=1}^{M}h_m(x)h_m(y)=\sum_{i=1}^{\infty}\gamma_i\phi_i(x)\phi_i(y).$$

Multiply each summand above by $\phi_k(x)$ and calculate $\langle K(x,y),\phi_k(x)\rangle$,

$$\sum_{m=1}^{M}\langle h_m(x),\phi_k(x)\rangle h_m(y)=\sum_{i=1}^{\infty}\langle\phi_i(x),\phi_k(x)\rangle\phi_i(y)=\gamma_k\phi_k(y).$$

consider that all ϕ_i are orthogonal and unit.

Let $g_{km}=\langle h_m(x),\phi_k(x)\rangle$ and calculate $\langle K(x,y),\phi_l(y)\rangle$, we get

$$\sum_{m=1}^{M}g_{km}h_m(y)=\gamma_k\phi_k(y),$$

$$\sum_{m=1}^{M}g_{km}\langle h_m(y),\phi_l(y)\rangle=\gamma_k\langle\phi_k(y),\phi_l(y)\rangle,$$

$$\sum_{m=1}^{M}g_{km}g_{lm}=\gamma_k\delta_{k,l}$$

Let $\mathbf{G}M=\{g_{nm}\}\in\mathbb{R}^{M\times N}$, i.e.

$$\mathbf{G}_M\mathbf{G}_M^T=\operatorname{diag}\{\gamma_1,\gamma_2,\dots,\gamma_M\}=\mathbf{D}_\gamma.$$

Let $\mathbf{V}^T=\mathbf{D}_\gamma^{-\frac{1}{2}}\mathbf{G}_M$, we have,

$$\mathbf{V}\mathbf{V}^T\mathbf{G}_M^T=\mathbf{G}_M^T\mathbf{D}_\gamma^{-1}\mathbf{G}_M=\mathbf{I}_N$$

So the three equation before can be rewrite as:

$$
\begin{gathered}
G_M h(x) = \mathbf{D}_\gamma \phi(x) \\
\mathbf{V}\mathbf{D}_\gamma^{-\frac{1}{2}} \mathbf{G}_M h(x) = \mathbf{V}\mathbf{D}_\gamma^{-\frac{1}{2}} \mathbf{D}_\gamma \phi(x) \\
h\left(x\right) = \mathbf{V}\mathbf{D}_\gamma^{\frac{1}{2}}.
\end{gathered}
$$

Finally we can show that (5.62) equiv (5.53)

$$
\begin{aligned}
& \min_{\{\beta_m\}_1^M} \sum_{i=1}^{N} \left(y_i - \sum_{m=1}^{M} \beta_m h_m(x_i) \right)^2 + \lambda \sum_{m=1}^{M} \beta_m^2 \\
&= \min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T h(x_i))^2 + \lambda \beta^T \beta \\
&= \min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T \mathbf{V}\mathbf{D}_\gamma^{\frac{1}{2}} \phi(x_i))^2 + \lambda \beta^T \beta \\
&= \min_{c} \sum_{i=1}^{N} (y_i - c^T \phi(x_i))^2 + \lambda (\mathbf{V}\mathbf{D}_\gamma^{\frac{1}{2}} c)^T \mathbf{V}\mathbf{D}_\gamma^{\frac{1}{2}} c \\
&= \min_{c} \sum_{i=1}^{N} (y_i - c^T \phi(x_i))^2 + \lambda c^T c \mathbf{D}_\gamma^{-1} \\
&= \min_{\{c_j\}_1^\infty} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{\infty} c_j \phi_j(x_i) \right)^2 + \lambda \sum_{j=1}^{\infty} \frac{c_j^2}{\gamma_j},
\end{aligned}
$$

(b)

(b) Show that

$$\begin{aligned}\hat{\mathbf{f}} &= \mathbf{H}\hat{\boldsymbol{\beta}} \\ &= \mathbf{K}(\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y},\end{aligned} \tag{5.75}$$

where $\mathbf{H}$ is the $N \times M$ matrix of evaluations $h_m(x_i)$, and $\mathbf{K} = \mathbf{H}\mathbf{H}^T$ the $N \times N$ matrix of inner-products $h(x_i)^T h(x_j)$.

Recall that we define $\hat{\beta}$ as follows:

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta^T h(x_i))^2 + \lambda \beta^T \beta.$$

By using derivate and get the zero point, we get:

$$\hat{\beta} = (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}$$
$$\hat{\mathbf{f}} = \mathbf{H}(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{y}.$$

By the identity of Woodbury matrix, we have

$$(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1} = \frac{1}{\lambda}\mathbf{I} - \frac{1}{\lambda}\mathbf{I}\mathbf{H}^T\left(\mathbf{I} + \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{H}\cdot\frac{1}{\lambda}\mathbf{I}.$$

thus,

$$\begin{aligned}\hat{\mathbf{f}} &= \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\mathbf{y} - \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\left(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{H}\mathbf{H}^T\mathbf{y} \\ &= \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\left[\mathbf{I} - (\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{H}^T\right]\mathbf{y} \\ &= \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\left[(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T)^{-1}(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T) - (\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{H}^T\right]\mathbf{y} \\ &= \frac{1}{\lambda}\mathbf{H}\mathbf{H}^T\left[(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T)^{-1}\lambda\mathbf{I}\right]\mathbf{y} \\ &= \mathbf{H}\mathbf{H}^T(\lambda\mathbf{I} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{y} \\ &= \mathbf{K}(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}.\end{aligned}$$

(c)

(c) Show that

$$\begin{aligned}\hat{f}(x) &= h(x)^T\hat{\boldsymbol{\beta}} \\ &= \sum_{i=1}^{N} K(x, x_i)\hat{\boldsymbol{\alpha}}_i \end{aligned} \tag{5.76}$$

and $\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$.

Obvious from (b)

(d)

(d) How would you modify your solution if $M < N$?

$\mathbf{K} + \lambda\mathbf{I}$ is invertible as long as $\lambda \neq 0$, else we have

$$\hat{\mathbf{f}} = \mathbf{H}\hat{\beta} = \mathbf{H}(\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{y} = \mathbf{y}.$$