Machine Learning Assignment 4

B08611035

December 2020

1 Deterministic Noise

(1) **c.**

$$f(x) = e^x,\ h(x) = wx,\ x \in [0, 2]$$

Thus, the area between the line $h(x) = wx$ and the curve $f(x) = e^x$ will be:

$$\int_0^2 (f(x) - h(x))^2 dx = \int_0^2 (e^x - wx)^2 dx$$

$$= \int_0^2 e^{2x} - 2e^x wx + w^2x^2 dx = \Big[\frac{1}{2}e^{2x} + \frac{1}{3}w^2x^3\Big]_0^2 - 2w\int_0^2 e^x x dx$$

$$= \frac{1}{2}e^4 + \frac{8}{3}w^2 - \frac{1}{2} - 2w\Big[xe^x - e^x\Big]_0^2$$

$$= \frac{1}{2}e^4 + \frac{8}{3}w^2 - 4we^2 + 2ew^2 - 2w = \frac{1}{2}e^4 + \frac{8}{3}w^2 - 2we^2 - 2w = A(w)$$

To find the best hypothesis, it's going to minimize $A(w)$:

$$\frac{dA(w)}{dw} = \frac{16}{3}w - 2e^2 - 2 = 0$$

$$\Rightarrow \frac{8}{3}w = e^2 + 1 \Rightarrow w = \frac{3}{8}(e^2 + 1)$$

To find the deterministic noise of best hypothesis, here are going to calculate $|f(x) - h(x)|$:

$$|f(x) - h(x)| = |e^x - wx| = |e^x - \frac{3}{8}(e^2 + 1)x|$$

$$= e^x - \frac{1}{8}(3e^2 + 3)x$$

2 Learning Curve

(2) **b.** 1

$$E_D[E_{in}(\mathcal{A}(\mathcal{D})] = E\Big[\frac{1}{N}\sum_{n=1}^{N} h^*(x_n) \neq y_N\Big] = E\Big([[h^*(x_n] \neq y_n]\Big)$$

$$E_D[E_{out}(\mathcal{A}(\mathcal{D})] = E\Big[\frac{1}{M}\sum_{m=1}^{M} h^*(\tilde{x_n}) \neq \tilde{y_N}\Big] = E\Big([[h^*(\tilde{x_n}] \neq \tilde{y_n}]\Big)$$

So $E_D[E_{in}$ and $E_D[E_{out}$ depend on its hypothesis instead of N and M. Because the data is independently and identically distribution, so if find a best hypothesis g^* that $E_{out}(g^*)$ is minimum, then

$$E_D[E_{in}(h^*) = E_D[E_{out}(g^*)$$

$$\Rightarrow E_D[E_{in}(h^*) = E_D[E_{out}(g^*) \leq E_D[E_{out}(h^*)$$

$$\Rightarrow E_D[E_{in}(\mathcal{A}(\mathcal{D})) = E_D[E_{out}(\mathcal{A}(\mathcal{D}))$$

3 Noisy Virtual Examples

(3) **d.**

$$X_h^T X_h = \Big([X^T\ 0] + [0\ X^T] + [0\ \varepsilon^T]\Big) \cdot \Big(\begin{bmatrix} X \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ X \end{bmatrix} + \begin{bmatrix} 0 \\ \varepsilon \end{bmatrix}\Big)$$

$$= X^T X + X^T X + X^T \varepsilon + \varepsilon^T X + \varepsilon^T \varepsilon$$

Thus the error of $X_h^T X_h$ is:

$$E(X_h^T X_h) = E(2X^T X) + E(X^T \varepsilon) + E(\varepsilon^T X) + E(\varepsilon^T \varepsilon)$$

$$= 2X^T X + N \cdot E(\varepsilon^2) = 2X^T X + N\sigma^2 \mathrm{I}_{d+1}$$

(4) **e.**

$$x_h^T y = \Big([X^T\ 0] + [0\ X^T] + [0\ \varepsilon^t]\Big) \begin{bmatrix} X \\ y \end{bmatrix}$$

$$= X^T y + X^T y + \varepsilon^T y$$

Thus, the error of $X_h^T y$ that $E(X_h^T y)$ would be

$$E(X_h^T y) = 2X^T y + yE(\varepsilon^T) = 2X^T y$$

4 Regularization

(5) **d.**

$$\frac{d}{dw}\left[\frac{1}{N}||zw-y||^2+\frac{\lambda}{N}w^Tw\right]$$

$$=\frac{2}{N}[z^Tzw-z^Ty]+\frac{\lambda}{N}\cdot 2w=0\Rightarrow(z^Tz+\lambda)w=z^Ty\Rightarrow w=(z^Tz+\lambda)^{-1}z^Ty$$

consider the transformation that $z=xQ$, the weight expression could be rewrite as:

$$w=(Q^TQTQ^T+\lambda)^{-1}Q^TX^Ty=(T+\lambda)^{-1}Q^Tx^Ty$$

when $\lambda=0$,

$$v_i=\frac{1}{r_i}[q_{i0},q_{i1},...,q_{ii}]=\frac{1}{r_i}\cdot s$$

And when $\lambda>0$, $w=u=(T+\lambda)^{-1}Q^TX^Ty$:

$$u_i=\frac{1}{r_i+\lambda}[q_{i0},q_{i1},...,q_{ii}]x^Ty=\frac{1}{r_i+\lambda}\cdot s$$

Thus,

$$\frac{u_i}{v_i}=\frac{1}{r_i+\lambda}\cdot\frac{1}{\frac{1}{r_i}}=\frac{r_i}{r_i+\lambda}$$

(6) **a.**

$$\frac{d}{dw}\left[\frac{1}{N}\sum_{n=1}^{N}(wx_n-y_n)^2+\frac{\lambda}{N}w^2\right]=\frac{2}{\lambda}\sum_{n=1}^{N}(wx_n-y_n)x_n+\frac{2}{N}\lambda w=0$$

Turn the equation in to the format of x,y and λ for w:

$$w\left[\sum_{n=1}^{N}x_n^2+\lambda\right]=\sum_{n=1}^{N}x_ny_n$$

so that the optimal solution would be:

$$w^*=\frac{\sum x_ny_n}{\sum x_n^2+\lambda}$$

and with $C=(w^*)^2$, so that C will be

$$C=(w^*)^2=\left(\frac{\sum_{n=1}^{N}x_ny_n}{\sum_{n=1}^{N}x_n^2+\lambda}\right)$$

(7) **d.** $(y-0,5)^2$

$$\frac{d}{dy}\Big[\frac{1}{N}\sum_{n=1}^{N}(y-y_n)^2+\frac{2k}{N}\Omega(y)\Big]=\frac{2}{N}\sum_{n=1}^{N}(y-y_n)+\frac{2k}{N}\Omega'(y)=0$$

$$\Rightarrow Ny-\sum y_n+k\Omega'(y)=0\Rightarrow Ny+k\Omega'(y)=\sum_{n-1}^{N}y_n$$

Because we know that:

$$y=\frac{\sum y_n+k}{N+2k}$$

is the optimal solution, thus we can get the equation below:

$$\sum_{n=1}^{N}y_n=Ny+2ky-k$$

The from the two equation both related to $\sum y_n$, we can obtain some relationship like:

$$\Omega'(y)=2y-1\Rightarrow\Omega(y)=y^2-y+C$$

If we want indicated that Ω into some of square, then we cane take $C=1/4$ so that:

$$\Omega(y)=(y-0.5)^2-\frac{1}{4}+C\Rightarrow\Omega(y)=(y-0.5)^2$$

(8) **b.** $W^T\Gamma^2W$

$$\min\frac{1}{N}\sum_{n=1}^{N}(\tilde{\mathbf{w}}^T\Phi(x_n)-y_N)^2+\frac{\lambda}{N}(\mathbf{w}^{\tilde{\mathbf{T}}}\mathbf{w})$$

$$=\min_{\tilde{\mathbf{w}}\in\mathbb{R}^{N+1}}\frac{1}{N}\sum_{n=1}^{N}(\tilde{\mathbf{w}}^T\Gamma^{-1}x_n-y_N)^2+\frac{\lambda}{N}(\mathbf{w}^{\tilde{\mathbf{T}}}\mathbf{w})$$

Let $\mathbf{w}^T=\tilde{\mathbf{w}}^T\Gamma^{-1}$, and the above equation is equivalent to

$$\min\frac{1}{N}\sum_{n=1}^{N}(\mathbf{w}^Tx_n-y_n)^2+\frac{\lambda}{N}\Omega(\mathbf{w})$$

Find $\Omega(\mathbf{w})$ so that when using $w=\Gamma^{-1}\tilde{w}$ into it, $\Omega(w)$ could be came $\tilde{w}^T\tilde{w}$

$$w=\Gamma^{-1}\tilde{w}\Rightarrow\Gamma w=\tilde{w}$$

$$\Omega(w)=\tilde{w}^T\tilde{w}=\mathbf{w}^T\Gamma\Gamma\mathbf{w}=\mathbf{w}^T\Gamma^2\mathbf{w}$$

(9) **b.**

First of that:

$$\frac{d}{dw}\Big[\frac{1}{N}(\mathbf{w^T x^T x w} - \mathbf{2 w^T x^T} y + y^T y + \frac{\lambda}{N} B\mathbf{w^T w}\Big]$$

$$= \frac{2}{N}(\mathbf{x^T x w} - \mathbf{x^T} y) + \frac{2}{N}\lambda B\mathbf{w} = 0$$

$$\Rightarrow \mathbf{w} = (\mathbf{x^T x} + \lambda B)^{-1}\mathbf{x}^T y$$

Adding k virtual examples to training data set:

$$\frac{d}{dw}\frac{1}{N+K}\Big[\sum_{n=1}^{N}(\mathbf{w^T x_n} - y_n)^2 + \sum_{n=1}^{K}(\mathbf{w^T \tilde{x_n}} - \tilde{y_n})^2\Big]$$

$$= \frac{2}{N+K}\Big[\mathbf{x^T x w} - \mathbf{x^T} y + \mathbf{\tilde{x}^T \tilde{x} w} - \mathbf{\tilde{x}^T}\tilde{y}\Big] = 0$$

$$\Rightarrow \mathbf{w} = (\mathbf{x^T x} + \mathbf{\tilde{x}^T \tilde{x}})^{-1}(\mathbf{x^T} y + \mathbf{\tilde{x}^T}\tilde{y})$$

thus we can infer that $\tilde{y} = 0$, and:

$$\mathbf{\tilde{x}^T \tilde{x}} = \lambda B = \sqrt{\lambda}\sqrt{B}\sqrt{\lambda}\sqrt{B}$$

$$\mathbf{\tilde{x}^T} = \mathbf{\tilde{x}} = \sqrt{\lambda}\sqrt{B}$$

5 Leave-one-out

(10) **e.** 1

The leave-one-out cross validation error of constant $\mathcal{A}_{majority}$ is like:

$$E_{loocv}(\mathcal{A}_{majority}) = \frac{1}{2N}[e_1 + e_2 + e_3 + ... + e_N + \tilde{e_1} + \tilde{e_2} + ... + \tilde{e_N}]$$

for N positive and negative respectively. e_i: i-th positive example for test, thus $N-1$ positive and N negative ones for training.

$$e_i = [[h_i(x_i) \neq y_i]] = 1$$

Similarly,

$$\tilde{e_i} = [[h_i(\tilde{x_i} \neq \tilde{y_i})]] = 1$$

Thus, the leave-one-out cross validation error could be:

$$E_{loocv}(\mathcal{A}_{majority}) = \frac{1}{2N}[2N] = 1$$

(11) **c. 2/N**

The leave-one-out cross validation error for the decision stump would be like:

$$E_{loocv} = \frac{1}{N}(e_1 + e_2 + ... + e_N)$$

$$e_i = 0 \text{ for } i \geq 4$$

$$\Rightarrow 0 \ \leq E_{loocv} \leq \frac{2}{N} = \frac{1}{N}\sum(0 + 1 + 1 + 0 + ... + 0)$$

For the tightest bound, here choose $2/N$ as the answer.

(12) **e.** $\sqrt{81 + 36\sqrt{6}}$

For the three given data points: $(x_1, y_1) = (3, 0), (x_2, y_2) = (\rho, 2)$ and $(x_3, y_3) = (-3, 0)$ for $\rho \geq 0$. And for the constant hypothesis, the leave-one-out cross validation error is:

$$E_{loocv} = \frac{1}{3}[(1-0)^2 + (0-2)^2 + (1-0)^2] = 2$$

and for the linear model hypothesis that:

$$\frac{h_1(x) - 2}{2 - 0} = \frac{x - \rho}{\rho + 3} \Rightarrow (\rho + 3)h_1(x) - 2\rho - 6 = 2x - 2\rho$$

$$h_1(x) = \frac{2}{\rho + 3}x + \frac{6}{\rho + 3}$$

$$\frac{h_2(x) - 2}{2 - 0} = \frac{x - 3}{3 + 3} \Rightarrow h_2(x) = 0$$

$$\frac{h_3(x) - 2}{2 - 0} = \frac{x - \rho}{\rho - 3} \Longrightarrow (\rho - 3)h_3(x) - 2\rho + 6 = 2x - 2\rho$$

$$h_3(x) = \frac{2}{\rho - 3}x - \frac{6}{\rho - 3}$$

The leave-one-out cross validation of linear model is equal to the one of constant model hypothesis:

$$E_{loocv}(\text{linear}) = E_{loocv}(\text{constant})$$

$$\frac{1}{3}[(h_1(x) - y_1)^2 + (h_2(x) - y_2)^2 + (h_3(x) - y_3)^2] = \frac{1}{3}[(\frac{12}{\rho + 3})^2 + (\frac{-12}{\rho - 3})^2 + 4] = 2$$

Then solving the equation

$$72(\rho - 3)^2 + 72(\rho + 3)^2 = (\rho - 3)^2(\rho + 3)^2 = (\rho^2 - 6\rho + 9)(\rho^2 - 6\rho + 9)$$

$$\rho^4 - 162\rho^2 = 1215 \Rightarrow (\rho^2 - 81) = 7776 \Rightarrow (\rho^2 - 81) = \pm 36\sqrt{6}$$

$$\rho^2 = 81 + 36\sqrt{6} \Rightarrow \rho = \sqrt{81 + 36\sqrt{6}}$$

(13) **d.** $1/k$

Due to the data are independently and identically distribute from the distribution:

$$(x_i, y_i) \sim^{i.i.d} P$$

$$\text{Var}_{\mathcal{D}_{x,y}}[E_{val}(h)] = \text{Var}_{\mathcal{D}_{x,y}}[\frac{1}{k}\sum_{i=1}^{k} err(h(x_i), y_i)] = \frac{1}{k^2}\text{Var}_{\mathcal{D}_{x,y}}[\sum_{i=1}^{k} err(h(x_i), y_i)]$$

$$= \frac{1}{k^2}\sum_{i=1}^{k}\text{Var}_{\mathcal{D}_{x,y}}(err(h(x_i), y_i))$$

$$= \frac{1}{k^2} \cdot k \cdot \text{Var}_{\mathcal{D}_{x,y}}[err(h(x), y)] = \frac{1}{k}\text{Var}_{\mathcal{D}_{x,y}}[err(h(x), y)]$$

6 Learning Principles

(14) **c.** 2/64

For four vertices of rectangle in $\mathbb{R}$. Can not find a line to obtain the combination like (with clockwise marked from top left) *oxox* and *xoxo*. Thus,

$$\min E(\mathbf{w}) = \frac{1}{4}\sum_{i=1}^{4}[[h(x_i) \neq y_i]] = \frac{1}{4}$$

for the above two case. And except the two case above, we can find a line to perfectly separate the data, so that the $\min E(\mathbf{w}) = 0$. Thus,

$$\mathbb{E}_{y_1,y_2,y_3,y_4}\Big(\min_{\mathbf{w}\in\mathbb{R}^{2+1}} E_{in}(\mathbf{w})\Big) = \frac{1}{16}\Big[\min E_{in}(\mathbf{w})|_{\text{case 1 to 16}}\Big]$$

$$= \frac{1}{16}\Big[\frac{1}{4} + \frac{1}{14}\Big] = \frac{2}{64}$$

(15) **a.**

$$E_{out}(g) = \frac{1}{N}\sum_{i=1}^{N}[[g(x_i) \neq y_i]]$$

$$\frac{1}{N}\sum_{g(x_i)=1, y_i=-1}[[g(x_i) \neq y_i]] + \frac{1}{N}\sum_{g(x_i)=-1, y_i=1}[[g(x_i) \neq y_i]]$$

$$= P(g(x_i) = 1, y_i = -1) + P(g(x_i) = -1, y_i = 1)$$

$$= P(y_i = -1)P(g(x_i) = 1|y_i = -1) + P(y_i = 1)P(g(x_i) = -1|y_i = 1) = (1-p)\varepsilon_- + p \cdot \varepsilon_+$$

$$E_{out}(g_c) = 1 - p$$

$$E_{out}(g) = E_{out}(g_c) \Rightarrow (1-p)\varepsilon_- + p \cdot \varepsilon_+ = 1 - p$$

$$p \cdot (\varepsilon_+ - \varepsilon_- + 1) = 1 - \varepsilon_- \Rightarrow p = \frac{1 - \varepsilon_-}{\varepsilon_+ - \varepsilon_- + 1}$$

7 Programming*

(16) **b.** -2

(17) **a.** -4

(18) **e.** 0.14

(19) **d.** 0.13

(20) **c.** 0.12