Machine Learning

Answer Sheet for Homework 4

Da-Min HUANG

R04942045

Graduate Institute of Communication Engineering, National Taiwan University

November 8, 2015

Problem 1

Deterministic error is the difference between best $h^* \in H$ and f. If $H' \subset H$, then the complexity of H' is lower than H in general. Hence, in general, the deterministic error increases.

□

Problem 2

1.

$$H(10,0,3) \cap H(10,0,4) = \left\{ \sum_{i=0}^{2} w_q L_q(x) \right\} \cap \left\{ \sum_{i=0}^{3} w_q L_q(x) \right\} \tag{1}$$

$$= \left\{ \sum_{i=0}^{2} w_q L_q(x) \right\} = H_2 \tag{2}$$

2.

$$H(10,0,3) \cup H(10,1,4) = \left\{ \sum_{i=0}^{2} w_q L_q(x) \right\} \cup \left\{ \sum_{i=0}^{3} w_q L_q(x) + \sum_{i=4}^{10} L_q(x) \right\} \tag{3}$$

$$= \left\{ \sum_{i=0}^{3} w_q L_q(x) + \sum_{i=4}^{10} L_q(x) \right\} \tag{4}$$

3.

$$H(10,1,3) \cap H(10,1,4) = \left\{ \sum_{i=0}^{2} w_q L_q(x) + \sum_{i=3}^{10} L_q(x) \right\} \cap \left\{ \sum_{i=0}^{3} w_q L_q(x) + \sum_{i=4}^{10} L_q(x) \right\} \tag{5}$$

$$= \left\{ \sum_{i=0}^{2} w_q L_q(x) + \sum_{i=4}^{10} L_q(x) \right\} \tag{6}$$

4.

$$H(10,0,3) \cup H(10,0,4) = \left\{ \sum_{i=0}^{2} w_q L_q(x) \right\} \cup \left\{ \sum_{i=0}^{3} w_q L_q(x) \right\} \tag{7}$$

$$= \left\{ \sum_{i=0}^{3} w_q L_q(x) \right\} = H_3 \tag{8}$$

□

Problem 3

We have

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta \nabla E_{\text{aug}}(\mathbf{w}_t) \tag{9}$$

where $\nabla E_{\text{aug}}(\mathbf{w}_t)$ is

$$\nabla E_{\text{aug}}(\mathbf{w}_t) = \frac{\partial}{\partial \mathbf{w}_t^T} \left(E_{\text{in}}(\mathbf{w}_t) + \frac{\lambda}{N} \mathbf{w}_t^T \mathbf{w}_t \right) = \frac{\partial E_{\text{in}}(\mathbf{w}_t)}{\partial \mathbf{w}_t^T} + \frac{2\lambda}{N} \mathbf{w}_t \tag{10}$$

Hence, we have

$$\mathbf{w}_{t+1} \leftarrow \left(1 - \frac{2\eta\lambda}{N} \right) \mathbf{w}_t - \eta \nabla E_{\text{in}}(\mathbf{w}_t) \tag{11}$$

□

Problem 4

Since $\mathbf{w}_{\text{lin}}$ is the optimal solution for the plain-vanilla linear regression, we have

$$E_{\text{in}}(\mathbf{w}_{\text{lin}}) \leq E_{\text{in}}(\mathbf{w}_{\text{reg}}(\lambda)) \tag{12}$$

Also, $\mathbf{w}_{\text{reg}}(\lambda)$ is the optimal solution for $E_{\text{aug}}(\mathbf{w})$, we have

$$E_{\text{aug}}(\mathbf{w}_{\text{reg}}(\lambda)) \leq E_{\text{aug}}(\mathbf{w}_{\text{lin}}) \tag{13}$$

So, we have

$$E_{\text{in}}\left(\mathbf{w}_{\text{reg}}\left(\lambda\right)\right)+\frac{\lambda}{N}\mathbf{w}_{\text{reg}}^{T}\left(\lambda\right)\mathbf{w}_{\text{reg}}\left(\lambda\right)\leq E_{\text{in}}\left(\mathbf{w}_{\text{lin}}\right)+\frac{\lambda}{N}\mathbf{w}_{\text{lin}}^{T}\mathbf{w}_{\text{lin}} \tag{14}$$

$$0\leq E_{\text{in}}\left(\mathbf{w}_{\text{reg}}\left(\lambda\right)\right)-E_{\text{in}}\left(\mathbf{w}_{\text{lin}}\right)\leq\frac{\lambda}{N}\left(\left\|\mathbf{w}_{\text{lin}}\right\|^{2}-\left\|\mathbf{w}_{\text{reg}}\left(\lambda\right)\right\|^{2}\right),\ \forall\lambda \tag{15}$$

Hence, we have $\left\|\mathbf{w}_{\text{lin}}\right\|\geq\left\|\mathbf{w}_{\text{reg}}\left(\lambda\right)\right\|$ if $\lambda>0$.
Since this inequality holds for all λ and $\left\|\mathbf{w}_{\text{lin}}\right\|$ is not a function of λ. We know that $\left\|\mathbf{w}_{\text{reg}}\left(\lambda\right)\right\|$ is a non-increasing function of λ for $\lambda\geq 0$.

□

Problem 5

For constant model with three points $A\left(-1,0\right)$, $B\left(\rho,1\right)$ and $C\left(1,0\right)$.

$$\frac{1}{3}\left(\underbrace{\left(0-\frac{1}{2}\right)^{2}}_{\text{leave }A}+\underbrace{\left(1-0\right)^{2}}_{\text{leave }B}+\underbrace{\left(0-\frac{1}{2}\right)^{2}}_{\text{leave }C}\right)=\frac{1}{2} \tag{16}$$

For linear model. Leave A, we get line $y=\dfrac{1}{\rho-1}\left(x-1\right)$; leave B, we get line $y=0$; leave C, we get line $y=\dfrac{1}{\rho+1}\left(x+1\right)$. So the error is

$$\frac{1}{3}\left(\left(0-\left(\frac{-2}{\rho-1}\right)\right)^{2}+\left(1-0\right)^{2}+\left(0-\frac{2}{\rho+1}\right)^{2}\right) \tag{17}$$

Then we have

$$\frac{1}{3}\left(\frac{4}{\rho^{2}-2\rho+1}+1+\frac{4}{\rho^{2}+2\rho+1}\right)=\frac{1}{2}\Rightarrow\rho=\pm\sqrt{9+4\sqrt{6}} \tag{18}$$

Since $\rho>0$, we have $\rho=\sqrt{9+4\sqrt{6}}$.

□

Problem 6

If the sender wants to make sure at least one person receives correct predictions on all 5 games. Then he should target at least 32 people at first game since there are half the number of people receive wrong prediction after each game and the sender can just ignore people who receives wrong prediction.

The sender sends

$$32 + 16 + 8 + 4 + 2 + 1 = 63 \text{ letters} \tag{19}$$

□

Problem 7

From the conclusion above, we have

$$1000 - 63 \times 10 = 370 \tag{20}$$

□

Problem 8

All mathematical function derivated before looking at data is hypothesis set of size 1.
Take positive ray hypothesis set for example, it can generate N possible models after given data. The function in this problem generated before given the data, so it must be some specific model instead of depending on some learning algorithm to learn from data. Hence, the hypothesis set is of size 1.

□

Problem 9

The Hoeffding bound is

$$2M \exp\left(-2\epsilon^2 N\right) = 2 \exp\left(-20000 \times (0.01)^2\right) = 2e^{-2} \approx 0.271 \tag{21}$$

□

Problem 10

The computation of Hoeffding bound is computated with data verified by $a(\mathbf{x})$. To improve the performance of $g(\mathbf{x})$, we should only give data sampled by $a(\mathbf{x})$.
Hence, if lucky enough, $a(\mathbf{x})$ AND $g(\mathbf{x})$ can improve the system.

□

Problem 11

To get optimal solution, consider the following equation.

$$\nabla \left(\sum_{n=1}^{N} \left(y_n - \mathbf{w}^T \mathbf{x}_n \right)^2 + \sum_{k=1}^{K} \left(\tilde{y}_k - \mathbf{w}^T \tilde{\mathbf{x}}_k \right)^2 \right) \tag{22}$$

$$= -2 \left(\sum_{n=1}^{N} \mathbf{x}_n \left(y_n - \mathbf{x}_n^T \mathbf{w} \right) + \sum_{k=1}^{K} \tilde{\mathbf{x}}_k \left(\tilde{y}_k - \tilde{\mathbf{x}}_k^T \mathbf{w} \right) \right) \tag{23}$$

$$= -2 \left(\left(\mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} \mathbf{w} \right) + \left(\tilde{\mathbf{X}}^T \tilde{\mathbf{y}} - \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \mathbf{w} \right) \right) \tag{24}$$

$$= \mathbf{0} \tag{25}$$

Hence, we have $\mathbf{w} = \left(\mathbf{X}^T \mathbf{X} + \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \right)^{-1} \left(\mathbf{X}^T \mathbf{y} + \tilde{\mathbf{X}}^T \tilde{\mathbf{y}} \right)$.

□

Problem 12

To minimize the formula, consider the following equation.

$$\nabla \left(\lambda \left\| \mathbf{w} \right\|^2 + \left\| \mathbf{X} \mathbf{w} + \mathbf{y} \right\|^2 \right) = 2\lambda \mathbf{w} + 2\mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{X}^T \mathbf{y} = \mathbf{0} \tag{26}$$

$$\Rightarrow \mathbf{w} = \left(\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{y} \tag{27}$$

Hence, we have $\tilde{\mathbf{X}} = \sqrt{\lambda} \mathbf{I}$, $\tilde{\mathbf{y}} = \mathbf{0}$.

□

Problem 13

$E_{\text{in}} = 0.05$ and $E_{\text{out}} = 0.045$.

□

Problem 14

There are three lambda with minimal E_{in}.

1. $\log_{10} \lambda = -10$, $E_{\text{in}} = 0.015$ and $E_{\text{out}} = 0.02$.
2. $\log_{10} \lambda = -9$, $E_{\text{in}} = 0.015$ and $E_{\text{out}} = 0.02$.
3. $\log_{10} \lambda = -8$, $E_{\text{in}} = 0.015$ and $E_{\text{out}} = 0.02$.

□

Problem 15

$\log_{10} \lambda = -7$, $E_{\text{in}} = 0.03$ and $E_{\text{out}} = 0.015$.

□

Problem 16

There are two lambda with minimal E_{in}.

1. $\log_{10} \lambda = -9$, $E_{\text{train}} = 0.0$, $E_{\text{val}} = 0.1$ and $E_{\text{out}} = 0.038$.
2. $\log_{10} \lambda = -8$, $E_{\text{train}} = 0.0$, $E_{\text{val}} = 0.05$ and $E_{\text{out}} = 0.025$.

□

Problem 17

There are eight lambda with minimal E_{in}.

1. $\log_{10} \lambda = -7$, $E_{\text{train}} = 0.033$, $E_{\text{val}} = 0.0375$ and $E_{\text{out}} = 0.021$.
2. $\log_{10} \lambda = -6$, $E_{\text{train}} = 0.033$, $E_{\text{val}} = 0.0375$ and $E_{\text{out}} = 0.021$.
3. $\log_{10} \lambda = -5$, $E_{\text{train}} = 0.033$, $E_{\text{val}} = 0.0375$ and $E_{\text{out}} = 0.021$.
4. $\log_{10} \lambda = -4$, $E_{\text{train}} = 0.033$, $E_{\text{val}} = 0.0375$ and $E_{\text{out}} = 0.021$.
5. $\log_{10} \lambda = -3$, $E_{\text{train}} = 0.033$, $E_{\text{val}} = 0.0375$ and $E_{\text{out}} = 0.021$.
6. $\log_{10} \lambda = -2$, $E_{\text{train}} = 0.033$, $E_{\text{val}} = 0.0375$ and $E_{\text{out}} = 0.021$.
7. $\log_{10} \lambda = -1$, $E_{\text{train}} = 0.033$, $E_{\text{val}} = 0.0375$ and $E_{\text{out}} = 0.022$.
8. $\log_{10} \lambda = 0$, $E_{\text{train}} = 0.033$, $E_{\text{val}} = 0.0375$ and $E_{\text{out}} = 0.028$.

□

Problem 18

The returned $\log_{10}(\lambda) = 0 \Rightarrow \lambda = 1$. $E_{\text{in}} = 0.035$ and $E_{\text{out}} = 0.020$.

□

Problem 19

The returned $\log_{10}(\lambda) = -8$ and $E_{\text{cv}} = 0.030$.

□

Problem 20

$E_{\text{in}} = 0.015$ and $E_{\text{out}} = 0.020$.

□

Reference

[1] Lecture Notes by Hsuan-Tien LIN, Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan.