

Home Work 4

Machine Learning Foundations

R04323050

經濟碩三 陳伯駒

1.

The screenshot shows a web browser window displaying a Coursera quiz page. The browser's address bar shows the URL www.coursera.org/learn/ntumlone-algorithmicfoundations/exam/Xe849/zuo-4. The Coursera logo is in the top left, and a search bar is in the top center. The user's name '伯駒' is in the top right. The page has a navigation bar with 'Back to Week 8' and 'Lessons'. The main content area is titled 'Three Learning Principles' and lists four items: 'Occam's Razor' (10 min), 'Sampling Bias' (11 min), 'Data Snooping' (12 min), and 'Power of Three' (8 min). The 'Quiz: 作業四' (20 questions) is selected. The quiz results show a score of '200/200 points (100%)' with the message 'We keep your highest score. View Latest Submission'. A 'Take it again' button is visible. The bottom of the screen shows a macOS dock with various application icons.

2.

$$E_{aug}(\mathbf{w}) = E_{in}(\mathbf{w}) + \frac{\lambda}{N} \mathbf{w}^T \mathbf{w}$$

Take the derivatives of \mathbf{w} : $\nabla E_{aug}(\mathbf{w}) = \nabla E_{in}(\mathbf{w}) + \frac{2\lambda}{N} \mathbf{w}$

By the Gradient Descent:

$$\begin{aligned} \mathbf{w}_{t+1} &\leftarrow \mathbf{w}_t - \eta \cdot (\nabla E_{aug}(\mathbf{w})) \\ &= \mathbf{w}_t - \eta \cdot (\nabla E_{in}(\mathbf{w}_t) + \frac{2\lambda}{N} \mathbf{w}_t) \\ &= (1 - \frac{2\eta\lambda}{N}) \mathbf{w}_t - \eta \nabla E_{in}(\mathbf{w}_t) \end{aligned}$$

3.

Regularized Regression Problem:

$$\min_{\mathbf{w}} E_{in}(\mathbf{w}) = \frac{1}{N} (\mathbf{Z}\mathbf{w} - \mathbf{y})^T (\mathbf{Z}\mathbf{w} - \mathbf{y}) \quad \text{s.t.} \quad \mathbf{w}^T \mathbf{w} \leq C$$

① If \mathbf{w}_{lin} satisfies the constraints $\mathbf{w}^T \mathbf{w} \leq C$, then \mathbf{w}_{reg} is equivalent to \mathbf{w}_{lin} , thus

$$\|\mathbf{w}_{reg}\| = \|\mathbf{w}_{lin}\|.$$

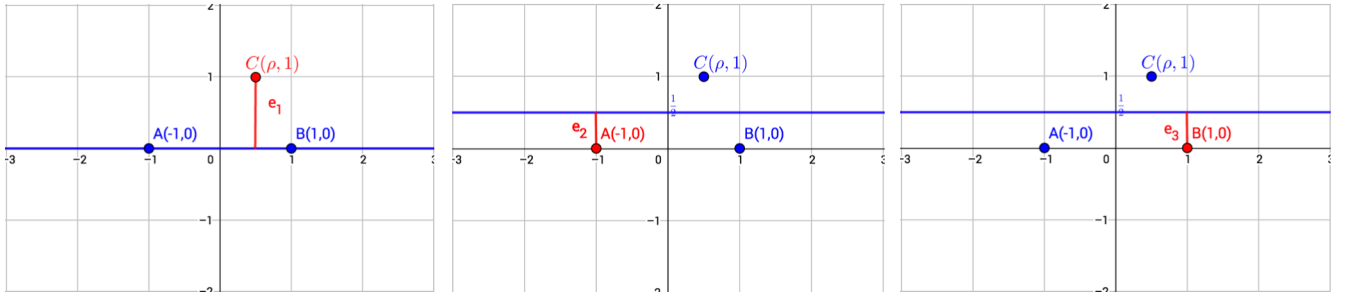
② If \mathbf{w}_{lin} does not satisfy the constraints $\mathbf{w}^T \mathbf{w} \leq C$, which means $\|\mathbf{w}_{lin}\|^2 > C$;

on the other hand, we know \mathbf{w}_{reg} satisfies the constraints, i.e $\|\mathbf{w}_{reg}\|^2 \leq C$. Hence,

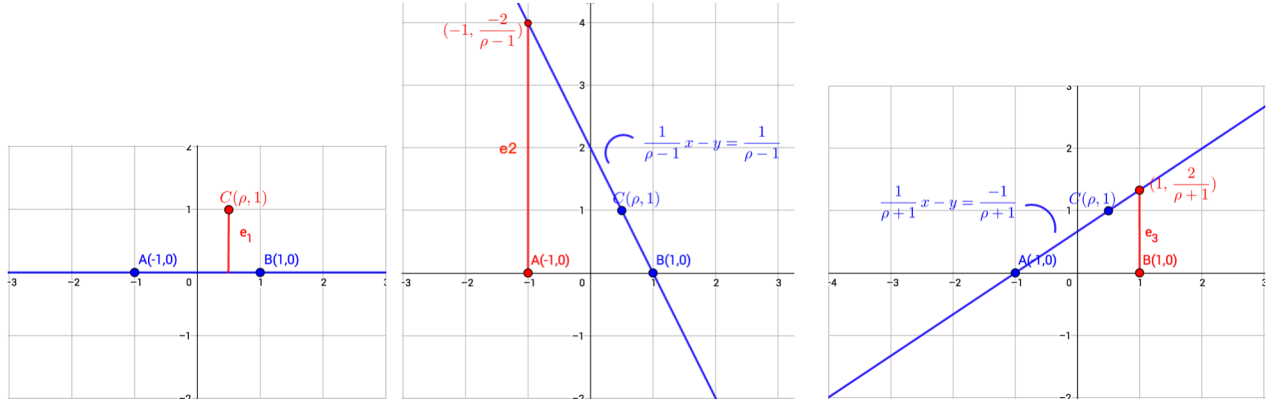
$$\|\mathbf{w}_{reg}\| < \|\mathbf{w}_{lin}\|.$$

By ①、②, we have $\|\mathbf{w}_{reg}(\lambda)\| \leq \|\mathbf{w}_{lin}\|$ for all $\lambda > 0$.

4.



$$\therefore E_{loocv}(h_0) = \frac{1}{3} \cdot (1^2 + (\frac{1}{2})^2 + (\frac{1}{2})^2) = \frac{1}{2}$$



$$\therefore E_{loocv}(h_1) = \frac{1}{3} \cdot (1^2 + (\frac{-2}{\rho-1})^2 + (\frac{2}{\rho+1})^2)$$

$$\text{Let } E_{loocv}(h_0) = E_{loocv}(h_1) \Rightarrow \rho = \sqrt{9 + 4\sqrt{6}}$$

5.

$$E_{in} = \frac{1}{N+K} (\mathbf{w}^T \mathbf{x} \mathbf{x}^T \mathbf{w} - 2 \mathbf{w}^T \mathbf{x}^T \mathbf{y} + \mathbf{y}^T \mathbf{y} + \mathbf{w} \tilde{\mathbf{x}}^T \tilde{\mathbf{x}} \mathbf{w} - 2 \tilde{\mathbf{w}}^T \tilde{\mathbf{x}} \tilde{\mathbf{y}} + \tilde{\mathbf{y}}^T \tilde{\mathbf{y}})$$

$$\therefore \nabla E_{in}(\mathbf{w}) = \frac{1}{N+K} (\mathbf{x}^T \mathbf{x} \mathbf{w} - \mathbf{x}^T \mathbf{y} + \tilde{\mathbf{x}}^T \tilde{\mathbf{x}} \mathbf{w} - \tilde{\mathbf{x}}^T \tilde{\mathbf{y}})$$

$$\text{F.O.C. } \nabla E_{in}(\mathbf{w}) = 0 \Rightarrow \mathbf{w}^* = (\mathbf{x}^T \mathbf{x} + \tilde{\mathbf{x}}^T \tilde{\mathbf{x}})^{-1} (\mathbf{x}^T \mathbf{y} + \tilde{\mathbf{x}}^T \tilde{\mathbf{y}}).$$

6.

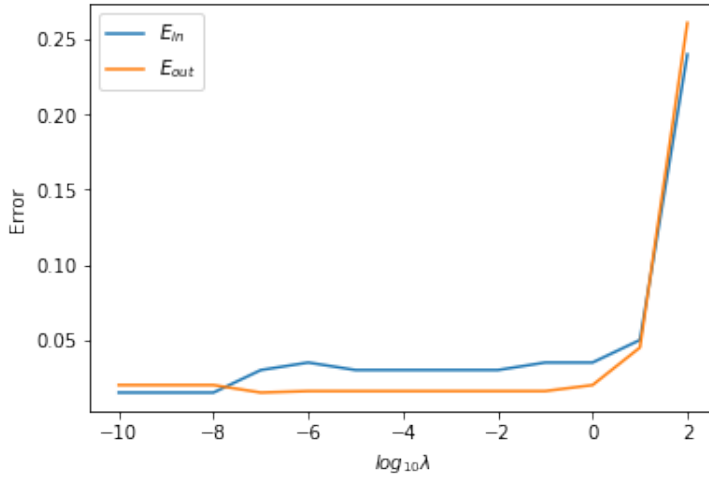
The condition of \mathbf{w}_{reg} given in question is exact the Regularized Regression Problem.

By P.10 in slides 14, we know the optimal solution of regularized regression is :

$$\mathbf{w}_{reg}^* = (\mathbf{x}^T \mathbf{x} + \lambda \mathbf{I})^{-1} \mathbf{x}^T \mathbf{y}$$

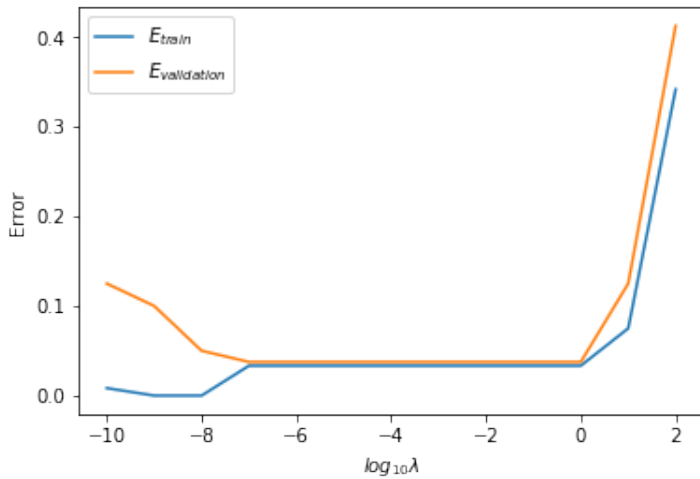
$$\therefore \text{Let } \tilde{\mathbf{x}} = \sqrt{\lambda} \mathbf{I}, \tilde{\mathbf{y}} = \mathbf{0} \text{ then } \mathbf{w}^* = \mathbf{w}_{reg}^*.$$

7.



In the left figure, we can observe that if we minimize E_{in} by choosing $\log_{10}\lambda = 10$, then E_{in} is close enough to E_{out} . This result corresponds to the learning goal: $E_{in} \approx E_{out}$ and E_{in}, E_{out} are small enough.

8.



We can observe the $E_{validation} \leq E_{train}, \forall \log_{10}\lambda$. It is quite intuitive since we split the original into two parts "train & validation" and tune the parameter to minimize the error in train. The model on train is tend to fit on the training set, the error on validation tend to be higher consequently.

9.

(a). $E_{loocv} = \frac{1}{N} \cdot (e_1 + e_2 + \dots + e_N)$, where $N = 1126 + 1126 = 2252$

Suppose we select the validation instances " $x_1 : +, x_2 : +, \dots, x_{1126} : +$ " corresponding to " $e_1, e_2, \dots, e_{1126}$ "; and select the validation instances " $x_{1127} : -, x_{1128} : -, \dots, x_{2252} : -$ " corresponding to " $e_{1127}, e_{1128}, \dots, e_{2252}$ ".

$\mathcal{A}_{majority}$: Always predicts the majority class.

e_1 : 1 instance with + as validation; 2251 instances as train: $\begin{cases} 1125 \text{ with } + \\ 1126 \text{ with } - \Rightarrow \text{Majority} \end{cases}$

$\therefore e_1 = 1$. Similarly, $e_2 = e_3 = \dots = e_{1126} = 1$

e_{1127} : 1 instance with - as validation; 2251 instances as train: $\begin{cases} 1126 \text{ with } + \Rightarrow \text{Majority} \\ 1125 \text{ with } - \end{cases}$

$\therefore e_{1127} = 1$. Similarly, $e_{1128} = e_{1129} = \dots = e_{2252} = 1$

Hence, $E_{loocv}(\mathcal{A}_{majority}) = \frac{1}{2252} \cdot (1 + 1 + \dots + 1) = \frac{1}{2252} \cdot (2252) = 1$

$\mathcal{A}_{minority}$: Always predicts the minority class.

e_1 : 1 instance with + as validation; 2251 instances as train: $\begin{cases} 1125 \text{ with } + \Rightarrow \text{minority} \\ 1126 \text{ with } - \end{cases}$

$\therefore e_1 = 0$. Similarly, $e_2 = e_3 = \dots = e_{1126} = 0$

e_{1127} : 1 instance with - as validation; 2251 instances as train: $\begin{cases} 1126 \text{ with } + \\ 1125 \text{ with } - \Rightarrow \text{minority} \end{cases}$

$\therefore e_{1127} = 0$. Similarly, $e_{1128} = e_{1129} = \dots = e_{2252} = 0$

Hence, $E_{loocv}(\mathcal{A}_{minority}) = \frac{1}{2252} \cdot (0 + 0 + \dots + 0) = \frac{1}{2252} \cdot (0) = 0$

Therefore, we will choose $\mathcal{A}_{minority}$ based on E_{loocv} .

(b). We follow the same strategy of selecting instances as validation in (a). Suppose we

have the instances $\{y_1, y_2, \dots, y_N\}$, let $\bar{y} = \sum_{i=1}^N \frac{y_i}{N}$.

$$e_1 = \left(y_1 - \frac{y_2 + y_3 + \dots + y_N}{N-1} \right)^2 = \left[y_1 - \frac{(\bar{y} \cdot N - y_1)}{N-1} \right]^2 = \left(\frac{y_1 \cdot N - y_1 - \bar{y} \cdot N + y_1}{N-1} \right)^2 =$$

$$\begin{aligned}
& \left[\frac{N(y_1 - \bar{y})}{N-1} \right]^2. \\
e_2 &= \left(y_2 - \frac{y_1 + y_3 + y_4 + \dots + y_N}{N-1} \right)^2 = \left[y_2 - \frac{(\bar{y} \cdot N - y_2)}{N-1} \right]^2 = \left[\frac{N(y_2 - \bar{y})}{N-1} \right]^2. \\
& \vdots \\
e_N &= \left[\frac{N(y_N - \bar{y})}{N-1} \right]^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
E_{loocu} &= \frac{1}{N} \cdot (e_1 + e_2 + \dots + e_N) \\
&= \frac{1}{N} \cdot \left\{ \left(\frac{N}{N-1} \right)^2 [(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \dots + (y_N - \bar{y})^2] \right\} \\
&= \left(\frac{N}{N-1} \right)^2 \cdot \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2 = \left(\frac{N}{N-1} \right)^2 \cdot Var(y_n)
\end{aligned}$$

\therefore the scale factor is $\left(\frac{N}{N-1} \right)^2$