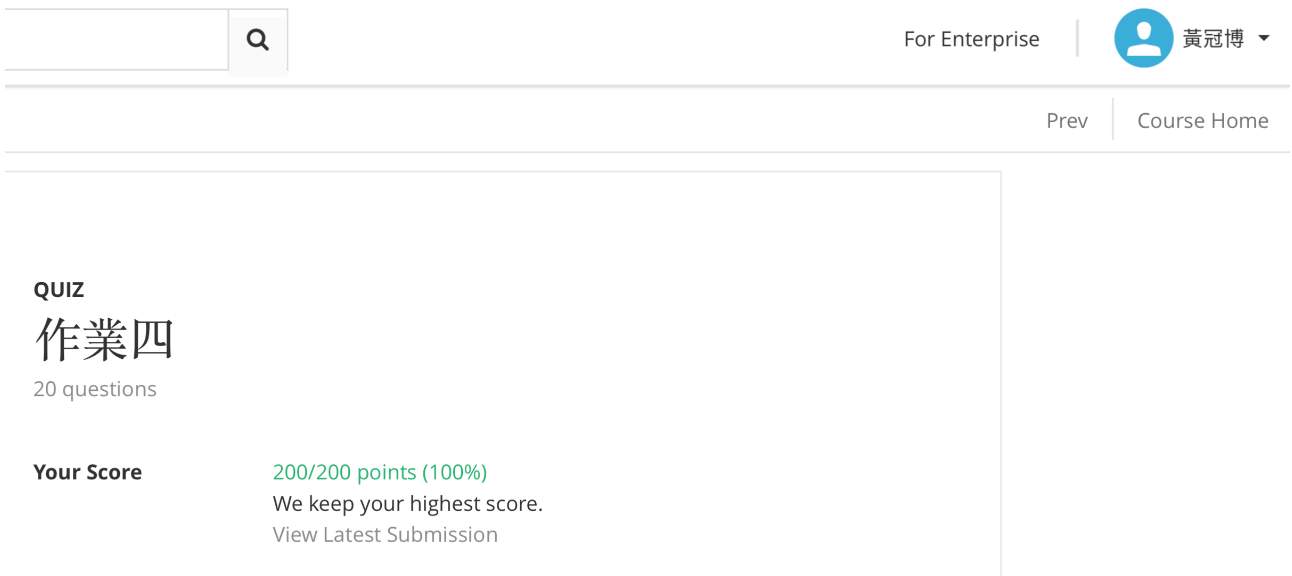
Machine Learning Foundations
Homework 4

b05902121 黃冠博

January 18, 2018

1.



2. Assume that $\mathbf{w}$ is $(w_0, w_1, \cdots, w_n)$:

$$\begin{aligned} E_{aug}(w) &= E_{in}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w^T w} \\ &= E_{in}(\mathbf{w}) + \frac{\lambda}{N}(w_0^2 + w_1^2 + \cdots + w_n^2) \end{aligned}$$

Now we have to calculate $\nabla E_{aug}(w)$:

$$\nabla E_{aug}(\mathbf{w}) = \left(\frac{\partial E_{aug}(\mathbf{w})}{\partial w_0}, \frac{\partial E_{aug}(\mathbf{w})}{\partial w_1}, \cdots, \frac{\partial E_{aug}(\mathbf{w})}{\partial w_n}\right)$$

while

$$\frac{\partial E_{aug}(\mathbf{w})}{\partial w_i} = \frac{\partial E_{in}(\mathbf{w})}{\partial w_i} + \frac{2\lambda}{N}w_i$$

Hence,

$$\begin{aligned} \nabla E_{aug}(\mathbf{w}) &= \nabla E_{in}(\mathbf{w}) + \frac{2\lambda}{N}(w_0, w_1, \cdots, w_n) \\ &= \nabla E_{in}(\mathbf{w}) + \frac{2\lambda}{N}\mathbf{w} \end{aligned}$$

In class, we have discussed that

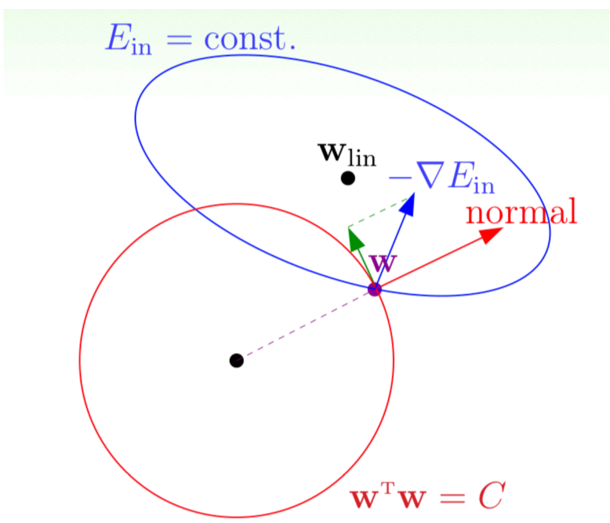
$$\mathbf{w_{t+1}} \leftarrow \mathbf{w_t} - \eta\mathbf{v}$$

where $\mathbf{v}$ is the direction we want to update.
So

$$\begin{aligned} &\mathbf{w_{t+1}} \leftarrow \mathbf{w_t} - \eta(\nabla E_{aug}(\mathbf{w_t})) \\ \Rightarrow\ &\mathbf{w_{t+1}} \leftarrow \mathbf{w_t} - \eta(\nabla E_{in}(\mathbf{w_t}) + \frac{2\lambda}{N}\mathbf{w_t}) \\ \Rightarrow\ &\mathbf{w_{t+1}} \leftarrow (1 - \frac{2\eta\lambda}{N}\mathbf{w_t}) - \eta\nabla E_{in}(\mathbf{w_t}) \end{aligned}$$

3.

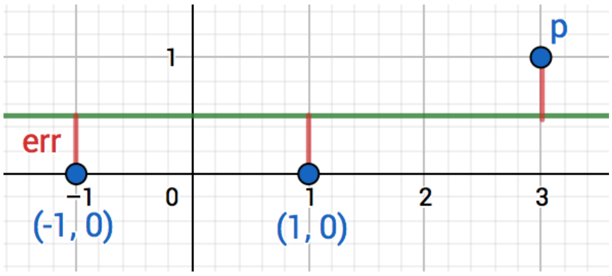
During class, we have discussed that: smaller λ decrease $\Leftrightarrow$ C increase $\Leftrightarrow$ larger $\mathbf{w}$.
However, when the circle touches $\mathbf{w_{lin}}$, $\mathbf{w_{reg}}(\lambda)$ will not keep decreasing with C. Hence, we know that $\|\mathbf{w_{reg}}(\lambda)\| \leq \|\mathbf{w_{lin}}\|$ when $\lambda > 0$.



4.

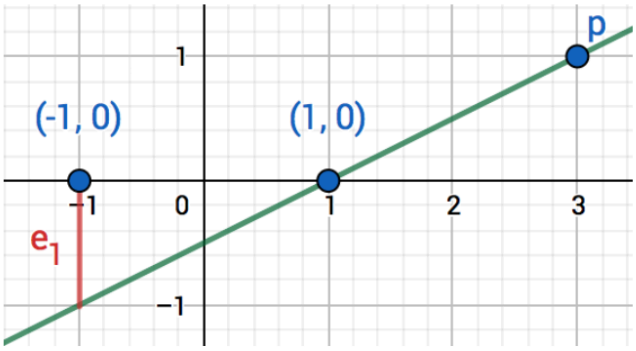
In class we've discussed that:

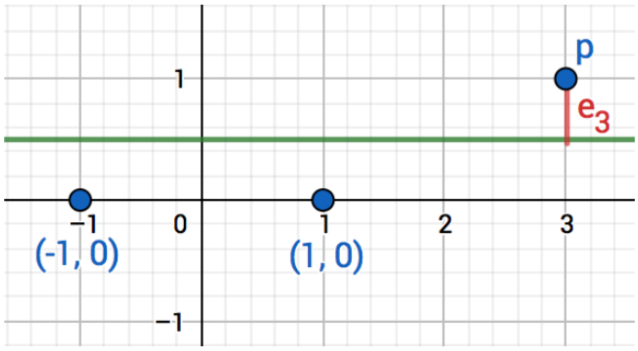
$$E_{loocv}(constant) = \frac{1}{3}(e_1 + e_2 + e_3)$$

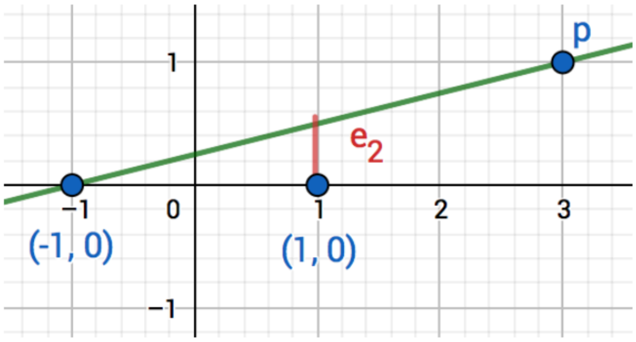


From the figure above, we can calculate that:

$$e_1 = e_2 = (\frac{1}{2})^2 = \frac{1}{4}$$







In class we've discussed that:

$$E_{loocv}(linear) = \frac{1}{3}(e_1 + e_2 + e_3)$$

By similar triangles, we can calculate e_1 and e_2:

$$\overline{AB} : 2 = 1 : \rho - 1 \Rightarrow \overline{AB} = \frac{2}{\rho - 1}$$

$$2 : \overline{CD} = \rho + 1 : 1 \Rightarrow \overline{CD} = \frac{2}{\rho + 1}$$

What we want is to let:

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

In the two cases e_3 are the same. Hence,

$$\begin{aligned}
&\frac{1}{4}+\frac{1}{4}+e_3=(\frac{2}{\rho-1})^2+(\frac{2}{\rho+1})^2+e_3\\
\Rightarrow\quad &(\frac{2}{\rho-1})^2+(\frac{2}{\rho+1})^2=\frac{1}{2}\\
\Rightarrow\quad &\frac{8\rho^2+8}{(\rho^2-1)^2}=\frac{1}{2}\\
\Rightarrow\quad &\rho^4-18\rho^2-15=0\\
\Rightarrow\quad &\rho=\sqrt{9+4\sqrt{6}}
\end{aligned}$$

5.

To solve this problem, we have to calculate the $\mathbf{w}$ for

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

by setting its gradient to zero.
However, we have to rewrite it into another form as shown below.

$$\begin{aligned}
&min_{\mathbf{w}}\frac{1}{N+K}\left(\|\mathbf{Xw}-\mathbf{y}\|^2+\|\tilde{\mathbf{X}}\mathbf{w}-\tilde{\mathbf{y}}\|^2\right)\\
\Rightarrow\quad &min_{\mathbf{w}}\frac{1}{N+K}\left((\mathbf{Xw}-\mathbf{y})^T(\mathbf{Xw}-\mathbf{y})+\left(\tilde{\mathbf{X}}\mathbf{w}-\tilde{\mathbf{y}}\right)^T\left(\tilde{\mathbf{X}}\mathbf{w}-\tilde{\mathbf{y}}\right)\right)\\
\Rightarrow\quad &min_{\mathbf{w}}\frac{1}{N+K}\left((\mathbf{w^TX^TXw}-2\mathbf{w^TX^Ty}+\mathbf{y^Ty})+\left(\mathbf{w^T}\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{X}}\mathbf{w}-2\mathbf{w^T}\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{y}}+\tilde{\mathbf{y}}^\mathbf{T}\tilde{\mathbf{y}}\right)\right)\\
\Rightarrow\quad &min_{\mathbf{w}}\frac{1}{N+K}\left(\mathbf{w^T}(\mathbf{X^TX}+\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{X}})\mathbf{w}-2\mathbf{w^T}(\mathbf{X^Ty}+\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{y}})+(\mathbf{y^Ty}+\tilde{\mathbf{y}}^\mathbf{T}\tilde{\mathbf{y}})\right)
\end{aligned}$$

According to lecture 9, slide P8,

$$\mathbf{w_{lin}}=A^{-1}b$$

with

$$\begin{aligned}
A&=\mathbf{X^TX}+\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{X}}\\
b&=\mathbf{X^Ty}+\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{y}}
\end{aligned}$$

the optimal $\mathbf{w}$ to the optimization problem is:

$$(\mathbf{X^TX}+\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{X}})^{-1}(\mathbf{X^Ty}+\tilde{\mathbf{X}}^\mathbf{T}\tilde{\mathbf{y}}) \tag{1}$$

6.

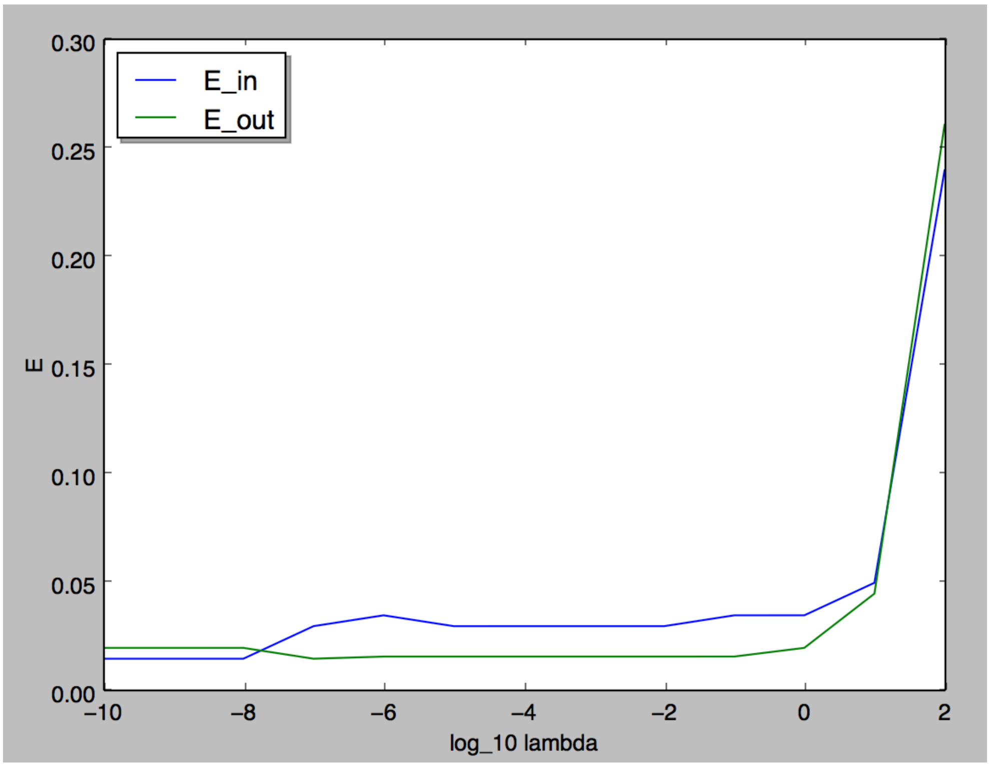
According to lecture 14, slide P10, the optimal solution is:

$$\mathbf{w_{reg}}\leftarrow(Z^TZ+\lambda I)^{-1}Z^Ty \tag{2}$$

Compare equation (1) and (2), we can derive

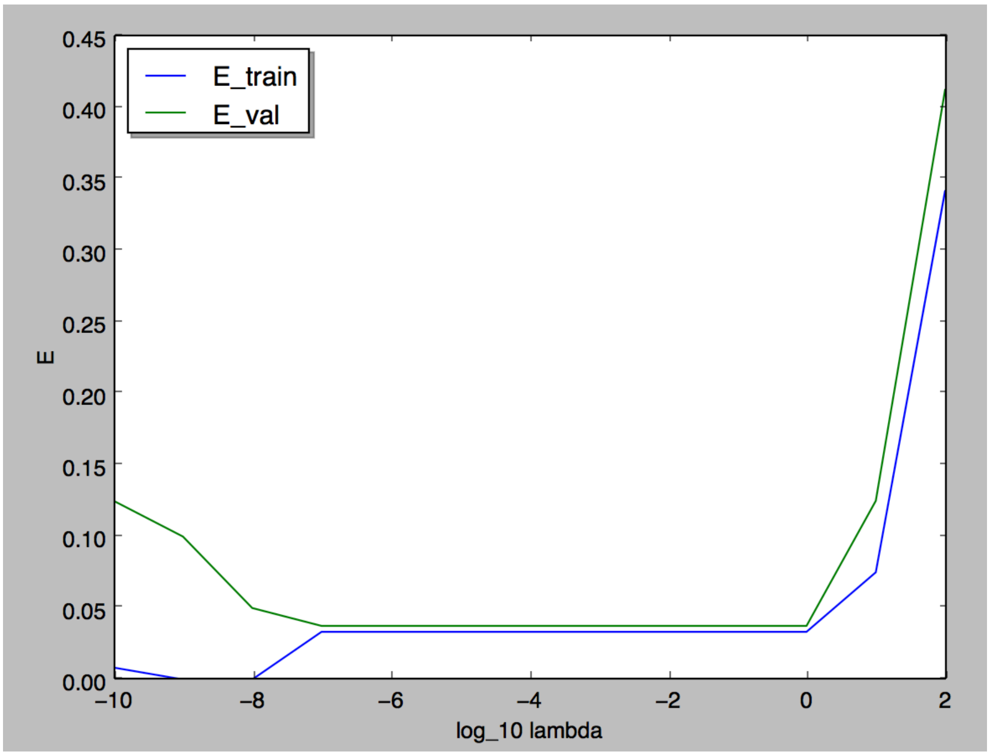
$$\begin{aligned}
\tilde{X}&=\sqrt{\lambda}\mathbf{I}\\
\tilde{\mathbf{y}}&=0
\end{aligned}$$

7.



When $log_{10}\lambda = -7 \sim 1$, $E_{in} > E_{out}$.
The algorithm performs good when λ is small.

8.



E_{val} is always larger than E_{train}.
However, the two are close when $log_{10}\lambda = -7 \sim 0$.
Therefore, the algorithm performs good when $log_{10}\lambda = -7 \sim 0$.

9.(a)

There are two cases for this problem.
Case 1: Pick a positive data as the "leave one out".
Now there are 1125 positive data and 1126 negative data.
$A_{majority}$ returns negative. However, if we use the positive data that we left out, the validation would be wrong.
$A_{minority}$ returns positive, if we use the positive data that we left out, the validation would be correct.

Case 2: Pick a negative data as the "leave one out".
Now there are 1126 positive data and 1125 negative data.
$A_{majority}$ returns positive. However, if we use the negative data that we left out, the validation would be wrong.
$A_{minority}$ returns negative, if we use the negative data that we left out, the validation would be correct.

From the result above, we see that $E_{loocv}(A_{majority}) = 1$ and $E_{loocv}(A_{minority}) = 0$.
Hence, we choose $A_{minority}$.

9.(b)

Consider the "leave one out" element to be y_i, $A_{average}$ returns:

$$\frac{1}{n-1}\left(\sum_{j\in\{1,\cdots,n\}\setminus\{i\}} y_j\right)$$

Now we calculate E_{val} with square error.

$$\begin{aligned}
E_{val} &= \left(y_i - \frac{1}{n-1}\left(\sum_{j\in\{1,\cdots,n\}\setminus\{i\}} y_j\right)\right)^2 \\
&= \left(y_i - \frac{n\mu - y_i}{n-1}\right)^2 \text{ where } \mu \text{ is the average of y} \\
&= \left(\frac{ny_i - y_i}{n-1} - \frac{n\mu - y_i}{n-1}\right)^2 \\
&= \frac{n^2}{(n-1)^2}(y_i-\mu)^2
\end{aligned}$$

Hence, we can calculate that:

$$\begin{aligned}
E_{loocv}(A_{average}) &= \frac{1}{n}\sum_{i=1}^{n}\frac{n^2}{(n-1)^2}(y_i-\mu)^2 \\
&= \frac{n}{(n-1)^2}\sum_{i=1}^{n}(y_i-\mu)^2
\end{aligned}$$

where

$$var = \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2$$

Therefore, we can conclude that $E_{loocv}(A_{average})$ is a scaled version of the variance of $\{y_n\}_{n=1}^{N}$.