1. In OvO, each classifier is trained on data from 2 classes.

For K-class classification, there are $\binom{k}{2} = \frac{k(k-1)}{2}$ binary classifier

Each binary classifier is trained on $\frac{N}{k}$

Given for binary classification, we need CPU time aN^3.

$\Rightarrow$ For K-class classification, each binary classification has data size $2\left(\frac{N}{k}\right)$ $\Rightarrow$ CPU time $= a\left(2\left(\frac{N}{k}\right)\right)^3$

$\therefore$ Total CPU time $= \frac{k(k-1)}{2} \cdot a\left(2\left(\frac{N}{k}\right)\right)^3$

$= 4aN^3 \frac{k(k-1)}{k^3}$

$= \frac{4aN^3(k-1)}{k^2}$

#

2. $\{x_n, y_n\}$, $g(x) = \tilde{w}^T \Phi_Q(x)$, $z_n = \Phi_Q(x_n)$

run linear regression on $\{(z_n, y_n)\}_{n=1}^N$ find $\tilde{w}$, prove there's some Q s.t. $E_{in}(g) = 0$

$\because$ x_n is 1-D data $\Rightarrow$ after Φ_Q, we can simply get

$z_n = (1, x_n, x_n^2, x_n^3, \ldots x_n^Q)$

($Q+1$ dimension data)

For linear regression by squared error,

$$E_{in}(g) = \frac{1}{N}\sum_{n=1}^{N}(\tilde{w}^T z_n - y_n)^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}(z_n^T \tilde{w} - y_n)^2$$

$$= \frac{1}{N}\left\| \begin{matrix} z_1^T\tilde{w} - y_1 \\ z_2^T\tilde{w} - y_2 \\ \vdots \\ z_N^T\tilde{w} - y_N \end{matrix} \right\|^2$$

$$= \frac{1}{N}\left\| \underbrace{\begin{bmatrix} - & z_1^T & - \\ - & z_2^T & - \\ & \vdots & \\ - & z_N^T & - \end{bmatrix}}_{(N\times(Q+1))} \underbrace{\tilde{w}}_{(Q+1)\times 1} - \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{N\times 1} \right\|^2$$

$$= \frac{1}{N}\| Zw - y \|^2$$

$$= \frac{1}{N}\left(\tilde{w}^T Z^T Z \tilde{w} - 2\tilde{w}^T Z^T y + y^T y\right)$$

To find min,

$$\nabla E_{in}(\tilde{w}_{Lin}) = 0 \Rightarrow \nabla E_{in}(w) = \frac{1}{N}(2Z^T Z w - 2Z^T y) = 0$$

$$\therefore \text{ for } W_{lin} = (Z^T Z)^{-1} Z^T y$$

We know the dimension of Z is $(N\times(Q+1)) \Rightarrow$ We pick `Q=N-1` which can form a $N\times N$ matrix. (Vandermonde Matrix)

To detect if $(Z^T Z)$ is invertible, we check $\det(Z^T Z)$

$\because \det(Z^T Z) = \det(Z^T)\det(Z)$

$$= (\det(Z))^2 \quad (\because \det(Z^T) = \det(Z))$$

Vandermonde matrix

$$\stackrel{\downarrow}{=} \left(\begin{vmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{N-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{N-1} \\ & & \vdots & & \\ 1 & x_N & x_N^2 & \cdots & x_N^{N-1} \end{vmatrix} \right)^2$$

$$= \left(\prod_{1 \le n < m \le N} (x_m - x_n) \right)^2$$

$\because$ ~~Given that~~ Assume all $\{x_n\}_{n=1}^N$ are different

$\therefore \det(Z^T Z) = \left(\prod_{1 \le n < m \le N} (x_m - x_n) \right)^2 \neq 0 \quad \therefore (Z^T Z)^{-1}$ exists

$\therefore$ We can find unique $W_{Lin} = (Z^T Z)^{-1} Z^T y$ to form

$g(x) = \tilde{W}_{Lin}^T \Phi_Q(x_n)$ such that $E_{in}(g) = \frac{1}{N} \sum_{n=1}^{N} (W_{Lin}^T Z_n - y_n) = 0$

as we pick "$Q = N-1$" perfectly #

3. x is uniformly sampled from $[-1, 1]$

$y = x + \varepsilon$, ε is independently sampled from Gaussian distribution with mean "0" and Variance "1"

$\because (\Phi(x))_n = z_n = [[x = x_n]]$

$\because$ all x_n are different in training data set, z will have exact one "1" and rest will be "0"

$\therefore$ By squared error,

$$E_{in}(g) = \frac{1}{N}\sum_{n=1}^{N}(\tilde{w}^T Z_n - y_n)^2$$

$\therefore$ For $E_{in}(g)$,

Since the transform creates a one-hot encoded vector for each training sample, the linear regression model will perfectly predict output y for each training sample x.

$\therefore E_{in}(g) = 0$ #

For $E_{out}(g)$,

For any new sample x', $\Phi(x')$ will be a vector of zeros, ($\because x'$ is different from all training samples x_n)

$\therefore g(x') = \tilde{w}^T \Phi(x') = 0$ (for all x' not in the training set.)

By $y' = x' + \varepsilon$, $\because g(x') = 0 \Rightarrow$ The squared error for new sample is

$$(y' - g(x'))^2 = (x' + \varepsilon)^2$$

$$\therefore E_{out}(g) = \mathbb{E}((x'+\varepsilon)^2)$$

$$= \mathbb{E}(x'^2) + \mathbb{E}(2x'\varepsilon) + \mathbb{E}(\varepsilon^2)$$

$\mathbb{E}(x'^2)$ is uniform distribution over $[-1,1] \Rightarrow \mathbb{E}(x'^2) = \frac{1}{12}(1-(-1))^2 = \frac{1}{3}$

$\mathbb{E}(x'\varepsilon) = \mathbb{E}(x')\mathbb{E}(\varepsilon) = 0$ ($\because$ mean of ε is 0)

$\mathbb{E}(\varepsilon^2) = 1$ (variance of ε is 1) $\quad \therefore E_{out}(g) = 1 + \frac{1}{3} = \frac{4}{3}$ #

4. $$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_N \\ | & | & & | \end{bmatrix}^T = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_N & - \end{bmatrix}_{N\times(d+1)}$$

$$X_h = \begin{bmatrix} | & | & & | & | & | & & | \\ x_1 & x_2 & \cdots & x_N & \tilde{x}_1 & \tilde{x}_2 & \cdots & \tilde{x}_N \\ | & | & & | & | & | & & | \end{bmatrix}^T = \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_N & - \\ - & \tilde{x}_1 & - \\ & \vdots & \\ - & \tilde{x}_N & - \end{bmatrix}_{(2N)\times(d+1)}$$

For

$$X_h^T X_h = \left(\begin{bmatrix} | & | & & | & | & | & & | \\ x_1 & x_2 & \cdots & x_N & \tilde{x}_1 & \tilde{x}_2 & \cdots & \tilde{x}_N \\ | & | & & | & | & | & & | \end{bmatrix} \begin{bmatrix} - & x_1 & - \\ - & x_2 & - \\ & \vdots & \\ - & x_N & - \\ - & \tilde{x}_1 & - \\ - & \tilde{x}_2 & - \\ & \vdots & \\ - & \tilde{x}_N & - \end{bmatrix} \right)_{(d+1)\times(d+1)}$$

$$= \begin{bmatrix} | \\ x_1 \\ | \end{bmatrix} \begin{bmatrix} - & x_1 & - \end{bmatrix} + \begin{bmatrix} | \\ x_2 \\ | \end{bmatrix} \begin{bmatrix} - & x_2 & - \end{bmatrix} + \cdots + \begin{bmatrix} | \\ \tilde{x}_N \\ | \end{bmatrix} \begin{bmatrix} - & \tilde{x}_N & - \end{bmatrix}$$

$$= X^T X + \begin{bmatrix} | \\ \tilde{x}_1 \\ | \end{bmatrix} \begin{bmatrix} - & \tilde{x}_1 & - \end{bmatrix} + \begin{bmatrix} | \\ \tilde{x}_2 \\ | \end{bmatrix} \begin{bmatrix} - & \tilde{x}_2 & - \end{bmatrix} + \cdots \begin{bmatrix} | \\ \tilde{x}_N \\ | \end{bmatrix} \begin{bmatrix} - & \tilde{x}_N & - \end{bmatrix}$$

$$= \underbrace{X^T X}_{①} + \underbrace{\tilde{X}^T \tilde{X}}_{②}$$

For ① $\Rightarrow E(X^T X) = X^T X$

For ② $\Rightarrow E(\tilde{X}^T \tilde{X}) = E((X+\varepsilon)^T (X+\varepsilon))$

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

For $E(X^TX + \varepsilon^TX + X^T\varepsilon + \varepsilon^T\varepsilon)$

$$= X^TX + 0 + 0 + E(\varepsilon^T\varepsilon) \quad \left(\because E(\varepsilon^TX) = E(X^T\varepsilon) = 0\right)$$

For $E(\varepsilon^T\varepsilon)$,

$\because$ ε has components uniformly distributed in $[-\delta, \delta]$, its variance $Var(\varepsilon_i) = \frac{1}{12}(\delta-(-\delta))^2 = \frac{\delta^2}{3}$ for each i

$$\therefore E(\varepsilon^T\varepsilon) = \frac{\delta^2}{3} I_{(d+1)\times(d+1)}$$

In conclusion,

$$\therefore E(X_h^TX_h) = E(X^TX) + E(\tilde{X}^T\tilde{X})$$

$$= X^TX + \left(X^TX + \frac{\delta^2}{3} I\right)$$

$$= 2X^TX + \frac{\delta^2}{3} I \quad \#$$

(I is identity matrix with size $(d+1)\times(d+1)$)

5.

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

$$\nabla(E_{aug}(w)) = \nabla(E_{in}(w)) + \frac{\partial}{\partial w}\left(\frac{\lambda}{N} w^Tw\right)$$

$$= \nabla E_{in}(w) + \frac{2\lambda}{N} w$$

By gradient descent algorithm

$$w_{t+1} \leftarrow w_t - \eta \nabla E_{aug}(w_t)$$

$$\Rightarrow w_{t+1} \leftarrow w_t - \eta\left(\frac{2\lambda}{N} w_t + \nabla E_{in}(w_t)\right)$$

$$\Rightarrow w_{t+1} \leftarrow \left(1 - \frac{2\lambda\eta}{N}\right) w_t - \eta \nabla E_{in}(w_t)$$

$$\Rightarrow w_{t+1} \leftarrow \left(1 - \frac{2\lambda\eta}{N}\right)\left(w_t - \frac{\eta N}{N - 2\lambda\eta} \nabla E_{in}(w_t)\right)$$

$\therefore \alpha = 1 - \frac{2\lambda\eta}{N}$, $\beta = \frac{\eta N}{N - 2\lambda\eta}$

#

6. Find w^* by take gradient to $\min_{w \in R} \frac{1}{N}\sum_{n=1}^{N}(wx_n - y_n)^2 + \frac{\lambda}{N}w^2$

$$E_{aug}(w) = \frac{1}{N}\sum_{n=1}^{N}(w^2x_n^2 - 2wx_ny_n + y_n^2) + \frac{\lambda}{N}w^2$$

Let $\nabla E_{aug}(w^*) = 0$, (Given X is 1-D data)

$$\nabla E_{aug}(w^*) = \frac{1}{N}\sum_{n=1}^{N}(2x_n^2w^* - 2x_ny_n) + \frac{2\lambda}{N}w^* = 0$$

$$\Rightarrow \frac{2}{N}\sum_{n=1}^{N}(x_n^2w^* - x_ny_n) + \frac{2\lambda}{N}w^* = 0$$

$$\Rightarrow w^*\left(\frac{2}{N}\sum_{n=1}^{N}x_n^2 + \frac{2\lambda}{N}\right) - \frac{2}{N}\sum_{n=1}^{N}x_ny_n = 0$$

$$\Rightarrow w^* = \frac{\sum_{n=1}^{N}y_nx_n}{\sum_{n=1}^{N}x_n^2 + \lambda}$$

We check $(w^*)^2 = C \Rightarrow w^* = \pm\sqrt{C}$ (for every $\lambda > 0$)

① if $w^* = \sqrt{C}$,

$$\frac{\sum_{n=1}^{N}y_nx_n}{\sum_{n=1}^{N}x_n^2 + \lambda} = \sqrt{C} \Rightarrow \sqrt{C}\left(\sum_{n=1}^{N}x_n^2 + \lambda\right) = \sum_{n=1}^{N}y_nx_n$$

$$\Rightarrow \lambda = \frac{\sum_{n=1}^{N}y_nx_n}{\sqrt{C}} - \sum_{n=1}^{N}x_n^2$$

② if $w^* = -\sqrt{C}$

$$\frac{\sum_{n=1}^{N}y_nx_n}{\sum_{n=1}^{N}x_n^2 + \lambda} = -\sqrt{C} \Rightarrow -\sqrt{C}\left(\sum_{n=1}^{N}x_n^2 + \lambda\right) = \sum_{n=1}^{N}y_nx_n$$

$\therefore (\alpha, \beta) = \left(\sum_{n=1}^{N}y_nx_n, -\sum_{n=1}^{N}x_n^2\right)$ (when $w^* = \sqrt{C}$)

#

$$\Rightarrow \lambda = \frac{-\sum_{n=1}^{N} y_n x_n}{\sqrt{c}} + \sum_{n=1}^{N} x_n^2$$

$$\therefore (\alpha, \beta) = \left(-\sum_{n=1}^{N} y_n x_n , \sum_{n=1}^{N} x_n^2\right)$$

(when $w^* = -\sqrt{c}$)

#

7. $$\min_{\tilde{w} \in \mathbb{R}^{d+1}} \frac{1}{N}\sum_{n=1}^{N}\left(\tilde{w}^T \Phi(x_n) - y_n\right)^2 + \frac{\lambda}{N}\|\tilde{w}\|_1$$

$\because$ Γ is diagonal matrix which will scaling i-th term in 'x'

$$\frac{1}{N}\sum_{n=1}^{N}\left(\tilde{w}^T \Gamma x_n - y_n\right)^2 + \frac{\lambda}{N}\|\tilde{w}\|_1$$

$\longrightarrow$ which means $(|w_1| + |w_2| + \cdots + |w_d|)$

Comparing to regularized linear regression on original data

$$\frac{1}{N}\sum_{n=1}^{N}\left(w^T x_n - y_n\right)^2 + \frac{\lambda}{N}\Omega(w)$$

We observe that $w^T = \tilde{w}^T \Gamma$, Suppose Γ is a diagnal matrix which can be written as $\Gamma = \mathrm{diag}(v_1, v_2, \ldots, v_i, \ldots, v_{d+1})$

$$\therefore \Omega(w) = \sum_{i=1}^{d+1}\left|\frac{w_i}{v_i}\right|$$

#

Futhermore, the relation between w and $\tilde{w}$, we have

$$w^T = \tilde{w}^T \Gamma \Rightarrow (w^T)^T = (\tilde{w}^T \Gamma)^T$$

$$\therefore w = \Gamma^T \tilde{w}$$

#

8. Divided into "2" cases,

Case1. When positive example is left out

$\because$ Now, we have N-1 positive and N negative examples

$\therefore$ For $A_{minority}$, we always predict the positive because we have N negative examples (majority), which is more than N-1 positive example (minority) (Given that $A_{minority}$ always predict minority.)

$$E_1 = \frac{N}{2N-1}$$

Case2. when negative example is left out

$\because$ Now, we have N-1 negative, N positive examples,

Similarly, $A_{minority}$ will always predict negative

$$\therefore E_2 = \frac{N}{2N-1} \quad (\because \underset{\text{(minority for negative)}}{N-1} < \underset{\text{(majority for positive)}}{N})$$

$$\therefore E_{loocv}(A_{minority}) = \frac{1}{2N}\Big(\underbrace{N \cdot \frac{N}{2N-1}}_{case1} + \underbrace{N\frac{N}{2N-1}}_{case2}\Big)$$

$$= \frac{N}{2N-1}$$

#

9.

Consider expectation of bionomial distribution

$$\therefore E = \frac{1}{32}(n \cdot p)\Big|_{(n=5,\, p=\frac{1}{2})}$$

$$= \frac{1}{32} \cdot 5 \cdot \frac{1}{2} = \frac{5}{64}$$

#

13.

Unconstrained linear regression solution is

$$W_{Lin} = (X^T X)^{-1} X^T y$$

By method of Dr. Regularize, subject to $\|w\|^2 \leq C$

$$W_C = \frac{W_{lin}}{\|W_{Lin}\|} \cdot \sqrt{C}$$

Assume $X^T X = \alpha I$,

$$W_{Lin} = (X^T X)^{-1} X^T y = \frac{1}{\alpha} X^T y$$

Now, we consider the regularized problem (ridge problem), which minimize $\|Xw - y\|^2 + \lambda \|w\|^2$. The solution is

$$W_{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

With $X^T X = \alpha I$, Wridge will be $(\alpha I + \lambda I)^{-1} X^T y$

$$= \frac{1}{\alpha + \lambda} X^T y$$

When $\|W_{lin}\|^2 > C \Rightarrow \|(\frac{1}{\alpha})^2 (X^T y)^T (X^T y)\| > C$,

By equating $\frac{W_{lin}}{\|W_{lin}\|}\sqrt{C} = W_{ridge} \Rightarrow \frac{\sqrt{C}}{\alpha \|W_{lin}\|} X^T y = \frac{1}{\alpha + \lambda} X^T y \Rightarrow \frac{\sqrt{C}}{\alpha \|W_{lin}\|} = \frac{1}{\alpha + \lambda}$

$\Rightarrow \lambda = \frac{\alpha \|W_{lin}\|}{\sqrt{C}} - \alpha \quad \therefore$ We still can find 'λ' to solve C-Constrained

linear regression if $X^TX = \alpha I$

For 'Only if' part

prove: if scaling W_{lin} is equivalent to solving the C-constrained problem, then $X^TX = \alpha I$

By equating $\frac{W_{lin}}{\|W_{lin}\|}\sqrt{C} = W_{ridge}$

$$\frac{W_{lin}}{\|W_{lin}\|} \cdot \sqrt{C} = (X^TX + \lambda I)^{-1} X^T y$$

Substituting $W_{lin} = (X^TX)^{-1}X^Ty$, we have

$$\frac{(X^TX)^{-1}X^Ty}{\|(X^TX)^{-1}X^Ty\|}\sqrt{C} = (X^TX+\lambda I)^{-1}X^Ty$$

$$\Rightarrow \left(\frac{X^TX}{\|(X^TX)^{-1}X^Ty\|\sqrt{C}}\right)^{-1} X^Ty = (X^TX+\lambda I)^{-1}X^Ty$$

$$\Rightarrow \left(\underbrace{\left(\frac{\sqrt{C}}{\|(X^TX)^{-1}X^Ty\|}\right)}_{(=\text{constant})} X^TX\right)^{-1} X^Ty = (X^TX+\lambda I)^{-1}X^Ty$$

If we want to equate above, we must let $X^TX = \alpha I$

$$\therefore \frac{\sqrt{C}}{\|(X^TX)^{-1}X^Ty\|}\alpha I = \alpha I + \lambda I \Rightarrow \frac{\alpha\sqrt{C}}{\|W_{lin}\|} = \alpha + \lambda$$

$$\Rightarrow \lambda = \frac{\alpha\sqrt{C}}{\|W_{lin}\|} - \alpha$$

$\therefore$ we can find such λ to equating $\frac{W_{lin}}{\|W_{lin}\|}\sqrt{C} = W_{ridge}$

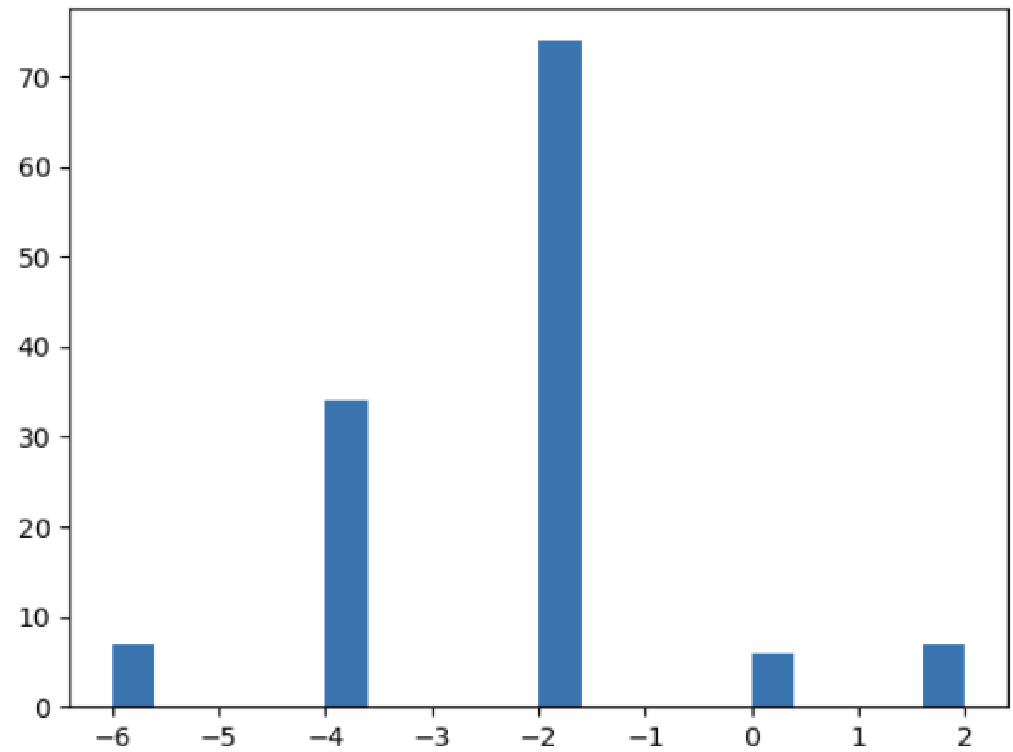
$\therefore$ W_{lin} is equivalent to solving the C-constrained only if $(X^TX) = \alpha I$ #

HTML hw4 solution

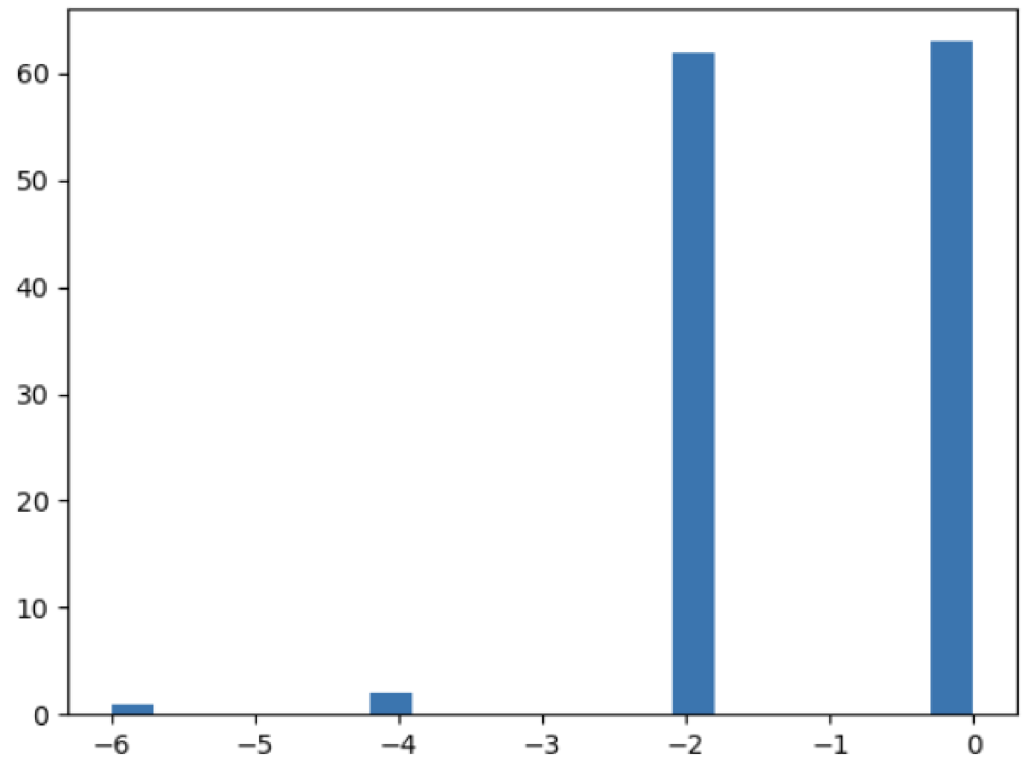
姓名: 謝銘倫, 系級:電機三, 學號:B10502166

10.
ans: $\log_{10}(\lambda*)$= -4

11.



12.



Findings compare with p11:

1. In problem 11, we have log10(λ∗) more on -2 and -4
2. In problem 11, we have log10(λ∗) more on 0 and -2
3. In problem 11, we use simple validation data, which is vulnerable to the only one training data set. if the log10(λ∗) generate from training data is not well enough, it will cause bad E_in validation data.

On the othre hand, the v-fold using different training data set with each fold, which is more safer to get log10(λ∗), which will more authentically show the real log10(λ∗) for data set.