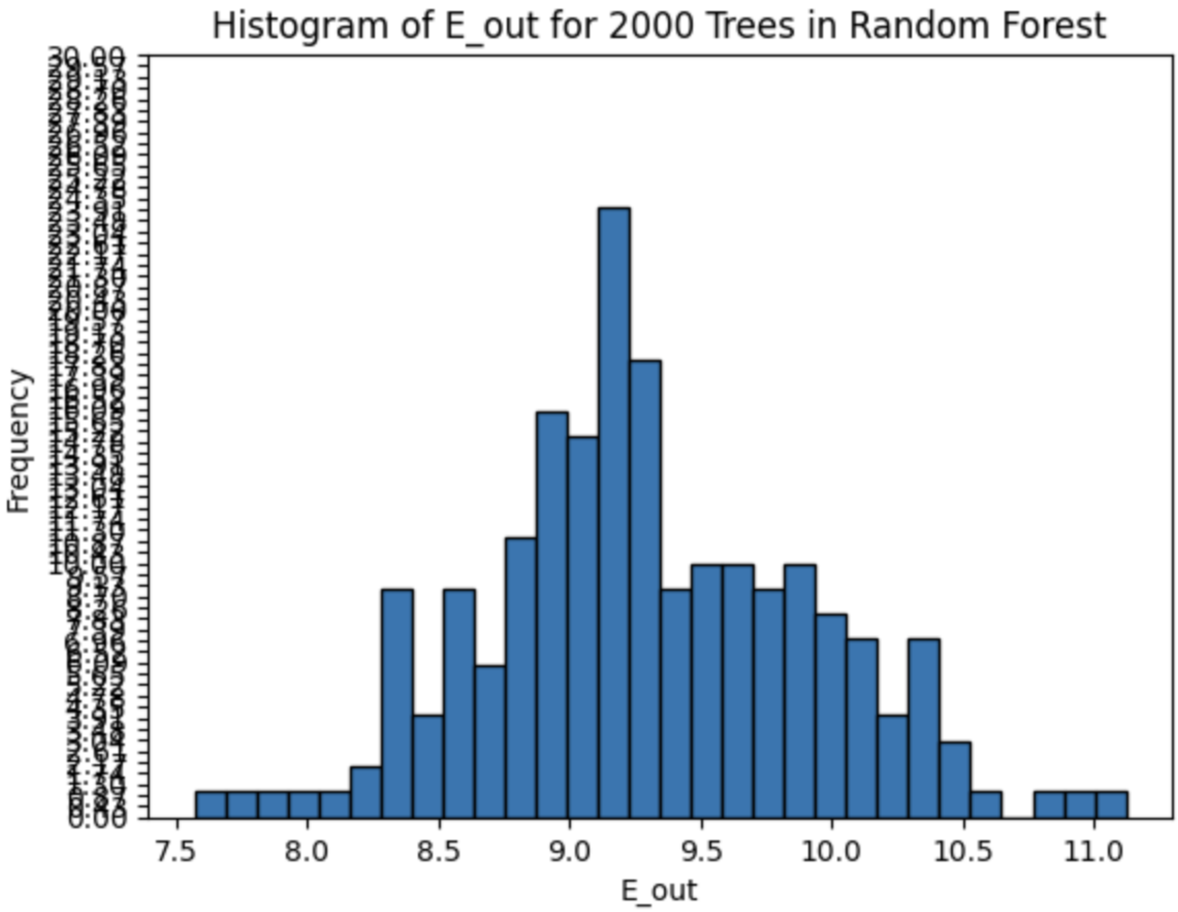
HTML hw6 solution

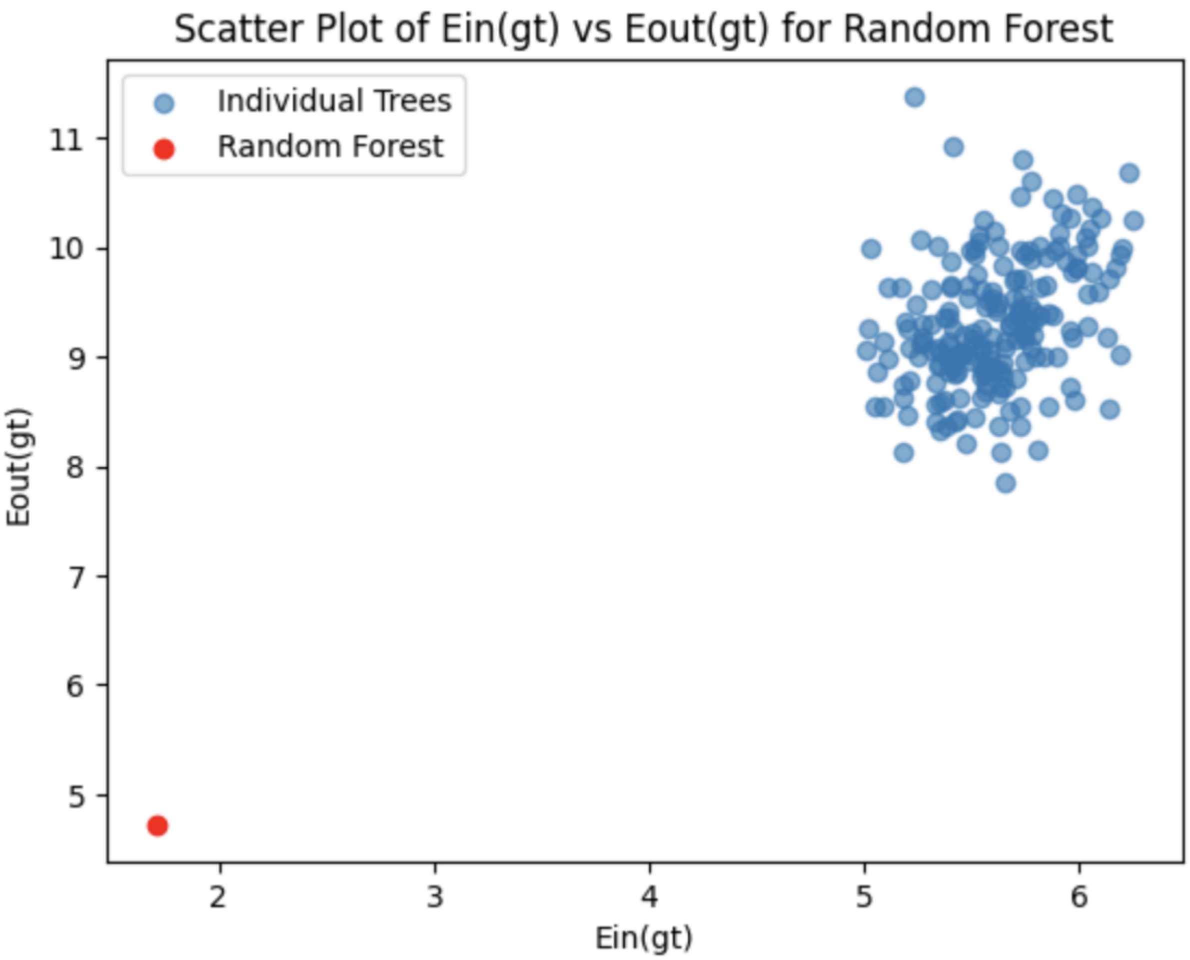
姓名: 謝銘倫, 系級:電機三, 學號:B10502166

9.
E_out = 8.736

10.

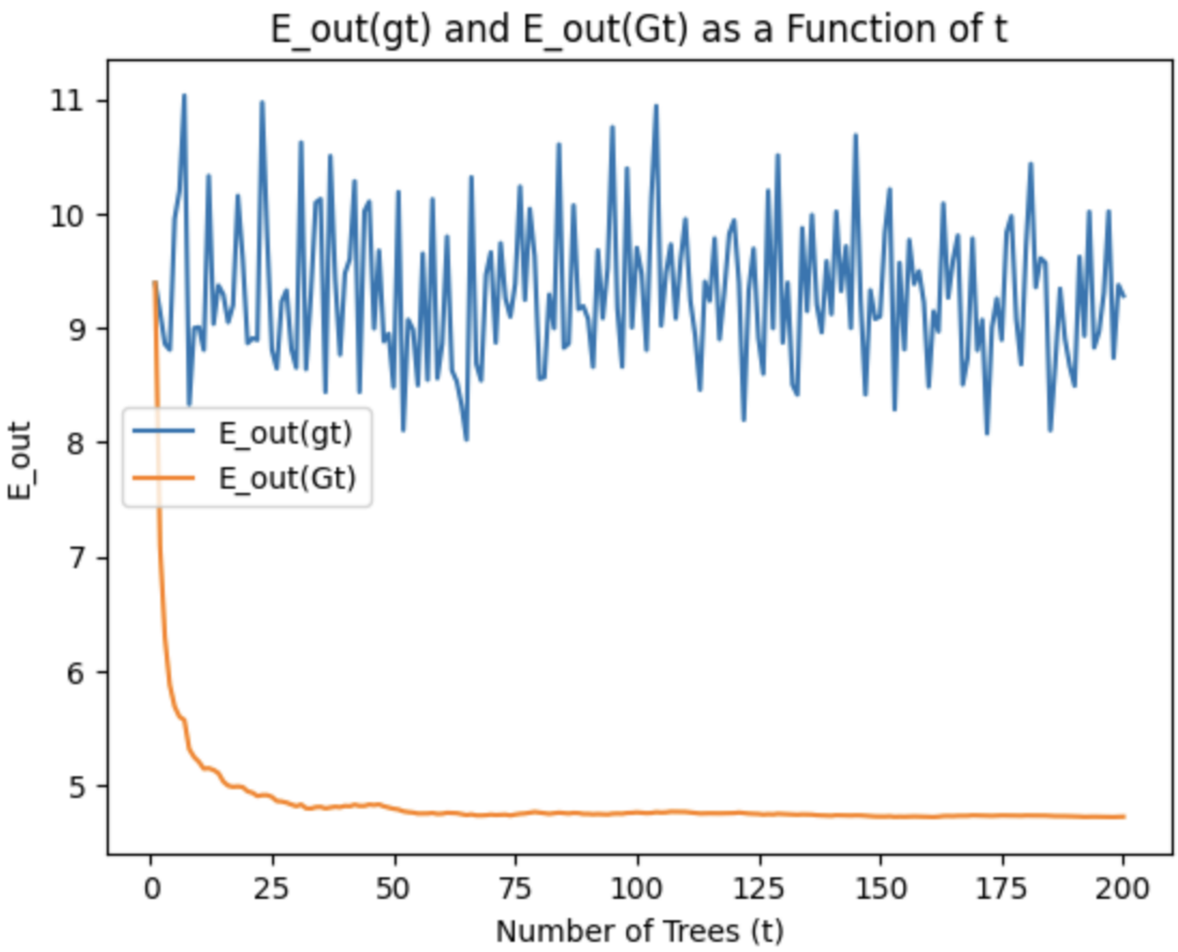


11.



Findings:
the individual trees are clustered at near positions, but the E_in(G), E_out(G) is very small obviously. (as the number of tree grows, error will be very small)

12.



Findings:
as the number of tree increases, the E_out(g_t) is oscillating at about error of 10, but the E_out(G) is decreasing obvious as the number of tree grows.

due to the lack of time, i just train 200 times.

HTML hw6 B10502166 謝銘倫 電機三

1.

$g_{s,i,\theta}(x) = s \cdot \mathrm{sign}(x_i - \theta)$

find $K_{ds}(x, x') = (\phi_{ds}(x))^T(\phi_{ds}(x'))$

$$g_{s,i,\theta}(x) \cdot g_{s,i,\theta}(x') = s_t\, \mathrm{sign}(x_i - \theta) \cdot s_t\, \mathrm{sign}(x_i' - \theta)$$
$$= \mathrm{sign}(x_i - \theta) \cdot \mathrm{sign}(x_i' - \theta)$$

$$K_{ds}(x, x') = (\phi_{ds}(x))^T(\phi_{ds}(x'))$$
$$= \sum_{t=1}^{|g|} \mathrm{sign}(x_t - \theta_t) \cdot \mathrm{sign}(x_t' - \theta_t)$$

Consider $\theta_t \in [\min(x_t, x_t'), \max(x_t, x_t'))$, the product of $\mathrm{sign}(x_t - \theta_t) \cdot \mathrm{sign}(x_t' - \theta_t) < 0$, in other cases, the product of $\mathrm{sign}(x_t - \theta_t)\mathrm{sign}(x_t' - \theta_t) > 0$

We have $|x_t - x_t'|$ numbers in interval $[\min(x_t, x_t'), \max(x_t, x_t'))$ such that $\mathrm{sign}(x_t - \theta_t)\mathrm{sign}(x_t' - \theta_t) = -1$,

We have $2\sum_{j=1}^{d}|x_j - x_j'| = 2\|x - x'\|_1$ numbers of such case.

(multiply by 2 because $s \in \{-1, +1\}$)

We denote that $\sum_{t=1}^{|g|} |\mathrm{sign}(x_{ti} - \theta_t)\mathrm{sign}(x_{ti}' - \theta_t)| = |g|$

Hence, we have # of $|g| - 2\|x - x'\|_1$ satisfy that

$$\text{sign}(x_{t_1} - \theta_t)\,\text{sign}(x'_{t_1} - \theta_t) = 1$$

$$\therefore K_{ds}(x, x') = \sum_{t=1}^{|g|} \text{sign}(x_{t_i} - \theta_t)\,\text{sign}(x'_{t_i} - \theta_t)$$

$$= \left(|g| - 2\|x - x'\|_1\right) + \left(-2\|x - x'\|_1\right)$$

$$= |g| - 4\|x - x'\|_1$$

$$= 2d(R - L) - 4\|x - x'\|_1 + 2$$

#

2.

3. $G(x) = \text{Sign}\left(\sum_{t=1}^{17} g_t(x)\right)$

Given $\text{avg}(E_{out}(g_t)) = \text{avg}\left(\mathcal{E}\left[(g-\bar{g})^2\right]\right) + E_{out}(G)$

worst case:

$\left\lceil \frac{17}{2} \right\rceil = 9 \Rightarrow$ since aggregated classifier G make its decision based on the majority vote of "17" classifier, more than half have to "agree" on the classification.

$\because$ In worst case, the smallest majority "9" classifiers are making incorrect predictions there are , 9, 10, 11, 12, 13, 14, 15, 16, 17 incorrect predictions are possible.

$\therefore \left(\frac{E_{out}(G)}{E}\right)_{max} = \frac{9}{17}$ #

4.

pick one example hold propability for $\frac{1}{N}$

$$\therefore P_{OOB} = \left(1 - \frac{1}{N}\right)^{\frac{3}{4}N}$$

$$= \left(\left(1 - \frac{1}{N}\right)^N\right)^{\frac{3}{4}}$$

$$= \left(\frac{1}{e}\right)^{\frac{3}{4}} = e^{-\frac{3}{4}}$$ #

5. total $u_n^{(1)}$ of 98% correct

$u_n^{(1)}$ of 2% incorrect

$\therefore$ correct $u_{n+}^{(2)} \leftarrow u_n^{(1)} \cdot 2$

incorrect $u_{n-}^{(2)} \leftarrow u_n^{(1)} \cdot 98$

$$\therefore \frac{u_+^{(2)}}{u_-^{(2)}} = \frac{\sum_{n: y_n > 0} u_n^{(2)}}{\sum_{n: y_n < 0} u_n^{(2)}} = \frac{2}{98}$$

#

6. $U_t = \sum_{n=1}^{N} u_n^{(t)}$, $0 < \varepsilon_t < \frac{1}{2} \Rightarrow \blacklozenge_t = \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} \geq 1$

correct: $u_{t+1} = u_t \cdot \sqrt{\frac{(1-\varepsilon_t)}{\varepsilon_t}}$

Incorrect: $u_{t+1} = u_t \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}$, Hence the incorrect example is multiplied by $\frac{\varepsilon_t}{1-\varepsilon_t}$

$$U_{t+1} = \sum_{n=1}^{N} u(t+1)_n = \sum_{n=1}^{N}{}_{\text{correct}} u(t)_n + \sum_{n=1}^{N} u(t)_n \beta_t$$

$$= (1-\varepsilon_t) U_t + \frac{\varepsilon_t}{1-\varepsilon_t} \varepsilon_t U_t$$

$$= U_t \left(1 - \varepsilon_t + \frac{\varepsilon_t^2}{1-\varepsilon_t}\right)$$

$$\Rightarrow = U_t \left(2\sqrt{\varepsilon_t(1-\varepsilon_t)}\right)$$

$\therefore$ Iterative over $\frac{U_2}{U_1}, \frac{U_3}{U_2} \ldots \frac{U_{t+1}}{U_t}$

Multiply by all ratios $\frac{U_{T+1}}{U_1} = \prod_{t=1}^{T} 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$

$\Rightarrow$ Final Expression $\frac{U_{T+1}}{U_1} = \prod_{t=1}^{T} 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$

since $0 < \varepsilon_t < \frac{1}{2}$, each term $\sqrt{\varepsilon_t(1-\varepsilon_t)}$ is less than $\frac{1}{2}$
cause the product decrease rapidly.
∴ Adaboost will converge within $O(\log N)$ #

7. The update rule for gradient boosting:

$$s_n^{(t)} = s_n^{(t-1)} - \alpha_n g_t(x_n)$$

We aim to prove after the update $\sum_{n=1}^{N} (y_n - s_n^{(t)})\, g_n(x_n) = 0$

$$\sum_{n=1}^{N} (y_n - s_n^{(t)})(g_t(x_n))$$

$$= \sum_{n=1}^{N} (y_n - (s_n^{(t-1)} - \alpha_n g_t(x_n))(g_t(x_n))$$

$$= \sum_{n=1}^{N} (y_n - s_n^{(t-1)} + \alpha_n g_t(x_n))(g_t(x_n))$$

$$= \sum_{n=1}^{N} (y_n - s_n^{(t-1)})\, g_t(x_n) + \alpha_t \sum_{n=1}^{N} g_t(x_n)^2$$

We know:
α_t is chosen to be minimize the loss, that is

$$\alpha_t = \underset{\alpha}{\arg\min} \left(err\{ s_n^{(t-1)} - \alpha g_t(x_n) \} \right)$$

which means the given α_t, the change in loss due to the update of s_n will be the smallest possible

By gradient descendent

$$\frac{d}{d\alpha} \left(err\{ s_n^{(t-1)} - \alpha g_t(x_n) \} \right) \Big|_{\alpha=\alpha_t} = 0$$

By expanding,

$$\Rightarrow \sum_{n=1}^{N} \left(\frac{\partial Loss}{\partial s_n} \cdot \frac{d s_n}{d\alpha} \right) = 0 \Rightarrow \sum_{n=1}^{N} -g_t(x_n) g_t(x_n) = 0 \Rightarrow \sum_{n=1}^{N} g_t(x_n)^2 = 0$$

$\because \alpha_t \sum_{n=1}^{N} g_t(x_n)^2 = 0$, which is generated by gradient descendent

$$\Rightarrow \sum_{n=1}^{N} (y_n - s_n^{(t)}) g_t(x_n) = \sum_{n=1}^{N} (y_n - s_n^{(t-1)}) g_t(x_n) + \alpha \sum_{n=1}^{N} g_t(x_n)^2$$

as $t \to$ large enough, we have

$$\sum_{n=1}^{N} (y_n - s_n^{(t)}) g_t(x_n) = \sum_{n=1}^{N} (y_n - s_n^{(t-1)}) g_t(x_n)$$

due to iteration $s_n \leftarrow s_n^{(t-1)} - \alpha_t g_n(x_n)$

$$\because \alpha_t \sum_{n=1}^{N} g_t(x_n)^2 = 0$$

$$\therefore \sum_{n=1}^{N} (y_n - s_n^{(t)}) g_t(x_n) = \sum_{n=1}^{N} (y_n - s_n^{(t-1)}) g_t(x_n) = 0$$

\#

8 initial $w_{ij}^{(l)} = 0 \Rightarrow$ after forward propogation $s_j^{(l)} = 0 \ (l = 0, 1, \ldots, L)$

Give backward-propogation rule:

$$\frac{\partial e_n}{\partial w_{ij}^{(l)}} = \delta_j^{(l)} (x_i^{(l-1)})$$

$$\delta_j^{(L)} = -2(y_n - s_j^{(L)})$$

$$\delta_j^{(l)} = \sum_k (\delta_k^{(l+1)})(w_{jk}^{(l+1)}) \underbrace{\tanh'(s_j^{(l)})}_{0} \quad \text{for } (l = 0, 1, \ldots, L-1)$$

By ① $\Rightarrow \delta_j^{(l)} = 0$ (for $L = 0, 1, \cdots, L-1$)

$\because$ Initial $W_{ij}^{(l)} = 0 \Rightarrow$ By update rule

$$W_{ij}^{(l)} = W_{ij}^{(l)} - \eta \delta_j^{(l)} x_i^{(l-1)}$$

We know $W_{ij}^{(l)} = 0$ (for $L = 0, 1, \cdots L-1$)

#