1.

Gini impurity: $1-\sum_{k=1}^{K} \mu_k^2$

max Gini impurity $\Rightarrow$ min $\sum_{k=1}^{K} \mu_k^2$

設一向量 $v=\sum_{k=1}^{K} \mu_k i_k$ 其中i_k為第k維度的單位向量

$u=\sum_{k=1}^{K} i_k$

$u \cdot v = |u||v|\cos\theta \le |u||v| \Rightarrow \sum_{k=1}^{K} \mu_k \cdot (\overset{=1}{i_k})^2 = \sum_{k=1}^{K} \mu_k \le \sqrt{\left(\sum_{k=1}^{K} \mu_k^2 i_k^2\right)} \cdot \sqrt{\left(\sum_{k=1}^{K} i_k^2\right)}$

$\sum_{k=1}^{K} \mu_k \le \sqrt{\sum_{k=1}^{K} \mu_k^2 \cdot K} \Rightarrow 1 \le \sum_{k=1}^{K} \mu_k^2 \cdot K \Rightarrow \frac{1}{K} \le \sum_{k=1}^{K} \mu_k^2$

Gini impurity $= 1-\sum_{k=1}^{K} \mu_k^2 \ge 1-\frac{1}{K} = \frac{K-1}{K}$ ※

2.

$$\mu_+(1-(\mu_+-\mu_-))^2+\mu_-(-1-(\mu_+-\mu_-))^2 = \mu_+(2-2\mu_+)^2+(1-\mu_+)(-2\mu_+)^2$$
$$= 4\mu_+(1-\mu_+)[(1-\mu_+)+\mu_+]$$
$$= 4\mu_+(1-\mu_+) = 4\mu_+-4\mu_+^2$$

$1-\mu_+^2-\mu_-^2 = 1-\mu_+^2-(1-\mu_+)^2 = 1-\mu_+^2-(\mu_+^2-2\mu_++1) = 2\mu_+-2\mu_+^2$

$2\times$ Gini impurity $= 4\mu_+-4\mu_+^2 =$ SRE when using binary classification. ※

3.

一個數據沒被選到的機率：$(1-\frac{1}{N})^{N'} = (1-\frac{1}{N})^{pN} = \left[(1-\frac{1}{N})^{N}\right]^{p}$

for N large, $\left[(1-\frac{1}{N})^{N}\right]^{p} \approx e^{-p}$

共有約 $N\times e^{-p}$個數據不會被選到。

4.

$G = \text{Uniform}(\{g_t\})$，對一點 x 與 y，若分類錯誤，$\sum_{k=1}^{K}[g_k(x) \neq y] \geq \frac{K+1}{2}$，要有超過一半 g_k 分類錯誤

設有 N 個點，全部的 tree 共分類錯 $\sum_{k=1}^{K} e_k \cdot N$ 次 $\Rightarrow$ G 最多錯 $N \cdot \sum_{k=1}^{K} e_k / \left(\frac{K+1}{2}\right)$ 個點

$$E_{out}(G) = \frac{\sum_{n=1}^{N}[G(x_n) \neq y_n]}{N} \leq N \cdot \sum_{k=1}^{K} e_k / \left(\frac{K+1}{2}\right) \cdot N = \frac{2}{K+1}\sum_{k=1}^{K} e_k \quad \#$$

5.

$$\alpha_1 = \min_{\eta} \frac{1}{N}\sum_{n=1}^{N}((y_n - s_n) - \eta g_1(x_n))^2$$

$$\xrightarrow[\text{對}\eta\text{做偏微}]{} -\frac{2}{N}\sum_{n=1}^{N}((y_n - s_n) - \eta g_1(x_n)) \cdot g_1(x_n)$$

$$= -\frac{2}{N} \cdot \sum_{n=1}^{N}((y_n - 0) - \eta \cdot 11.26) \cdot 11.26$$

$$= -\frac{22.52}{N}\sum_{n=1}^{N}(y_n - 11.26\eta) = -\frac{22.52}{N}\left(-11.26N\eta + \sum_{n=1}^{N} y_n\right) = 0$$

$$\alpha_1 = \frac{1}{11.26N}\sum_{n=1}^{N} y_n$$

6.

根據上題：$\alpha_t = \min_{\eta} \frac{1}{N}\sum_{n=1}^{N}((y_n - s_n) - \eta g_t(x_n))^2$

$$\xrightarrow[\text{對}\eta\text{做偏微}]{} -\frac{2}{N}\sum_{n=1}^{N}((y_n - s_n) - \eta g_t(x_n)) \cdot g_t(x_n) = 0$$

$$\Rightarrow \sum_{n=1}^{N}(y_n - s_n) \cdot g_t(x_n) = \eta \sum_{n=1}^{N} g_t^2(x_n) \Rightarrow \alpha_t = \sum_{n=1}^{N} g_t(x_n) \cdot (y_n - s_n) \Big/ \sum_{n=1}^{N} g_t^2(x_n)$$

$$\alpha_t \cdot \sum_{n=1}^{N} g_t^2(x_n) = \sum_{n=1}^{N} g_t(x_n) \cdot y_n - \sum_{n=1}^{N} g_t(x_n) \cdot s_n \Rightarrow \sum_{n=1}^{N} g_t(x_n) s_n = \sum_{n=1}^{N} g_t(x_n) \cdot y_n - \alpha_t \cdot \sum_{n=1}^{N} g_t^2(x_n)$$

但此時 s_n 還未更新，新的 s_n 為 $s_n + \alpha_t g_t(x_n)$

$$\sum_{n=1}^{N} s_n g_t(x_n) = \sum_{n=1}^{N}(s_n + \alpha_t g_t(x_n)) \cdot g_t(x_n) = \sum_{n=1}^{N} s_n \cdot g_t(x_n) + \alpha_t \sum_{n=1}^{N} g_t^2(x_n)$$

$$= \sum_{n=1}^{N} g_t(x_n) y_n - \cancel{\alpha_t \sum_{n=1}^{N} g_t^2(x_n)} + \cancel{\alpha_t \sum_{n=1}^{N} g_t^2(x_n)} = \sum_{n=1}^{N} g_t(x_n) \cdot y_n \quad \#$$

7. squared error polynomial regression: $\beta = (\lambda I + K)^{-1} y$.

$\because$ the regression is without regularization $\Rightarrow \lambda = 0 \Rightarrow \beta = K^{-1} y \Rightarrow K\beta = y$

根據 problem 5. $\alpha_1 = \min_{\eta} \frac{1}{N} \sum_{n=1}^{N} ((y_n - \underset{0}{\underset{\shortparallel}{s_n}}) - \eta g_1(x_n))^2$

$\Longrightarrow$ 對η偏微 $-\frac{2}{N} \sum_{n=1}^{N} (y_n - \eta g_1(x_n)) \cdot g_1(x_n)$

對於所有 $x_n, n \in N$, $g_1(x_n) = \beta^T z_n = \beta \cdot K_n \Rightarrow$ 已知 $K\beta = y \Rightarrow y_n = K_n \cdot \beta = \beta \cdot K_n$

$\Rightarrow$ 對於所有 x_n, $g_1(x_n) = y_n$

$-\frac{2}{N} \sum_{n=1}^{N} (y_n - \alpha_1 \cdot g_1(x_n)) \cdot g_1(x_n) = -\frac{1}{N} \sum_{n=1}^{N} (y_n - \alpha_1 y_n) \cdot y_n = 0 \Rightarrow \alpha_1 = 1$ #

8.

$$OR(x_1, x_2 \cdots x_d) = \begin{cases} FALSE & \text{iff } x_1 = x_2 = \cdots x_d = FALSE \\ TRUE & \text{else.} \end{cases}$$

set $w_0 = d - 0.5$ $\quad w_1 = w_2 = \cdots w_d = 1$

如此一來 $\sum_{i=0}^{d} w_i x_i < 0$ when $x_1 = x_2 = \cdots x_d = -1$

$\sum_{i=0}^{d} w_i x_i > 0$ for other cases.

9.

initial $s_j^{(l)} = \sum_{i=0}^{d^{(l-1)}} w_{ij}^{(l)} x_j^{(l-1)} = 0$ for $l = 1, 2 \cdots L$.

$\delta_1^{(L)} = -2(y_n - s_1^{(L)}) = -2 \cdot y_n \neq 0$

$\delta_j^{(l)} = \frac{\partial e_n}{\partial s_j^{(l)}} = \sum_k (\delta_k^{(l+1)})(\underset{0}{\underset{\shortparallel}{w_{jk}^{(l+1)}}})(\tanh'(s_j^{(l)})) = 0$ for $l = 1 \cdots L-1$. #

10、

$\frac{\partial e}{\partial s_k^{(L)}}$ 有2种情形：① $y=k'\neq k \Rightarrow v_k=0$ ② $y=k \Rightarrow v_k=1$

① $y=k'\neq k$：

$$e=-\sum_{k=1}^{K} v_k \ln q_k = -\ln q_{k'}$$

$$\frac{\partial e}{\partial s_k^{(L)}}=\frac{\partial}{\partial s_k^{(L)}}\cdot\left(-\ln\left(\frac{\exp(s_{k'}^{(L)})}{\sum_{t=1}^{K}(\exp(s_t^{(L)}))}\right)\right)=\frac{\partial}{\partial s_k^{(L)}}\cdot\left(-\ln(\exp(s_{k'}^{(L)}))+\ln\left(\sum_{t=1}^{K}(\exp(s_t^{(L)}))\right)\right)$$

$$=0+\frac{1}{\sum_{t=1}^{K}(\exp(s_t^{(L)}))}\cdot\frac{\partial}{\partial s_k^{(L)}}\left(\sum_{t=1}^{K}(\exp(s_t^{(L)}))\right)=\frac{\exp(s_k^{(L)})}{\sum_{t=1}^{K}(\exp(s_t^{(L)}))}=q_k-\overset{=0}{v_k}$$

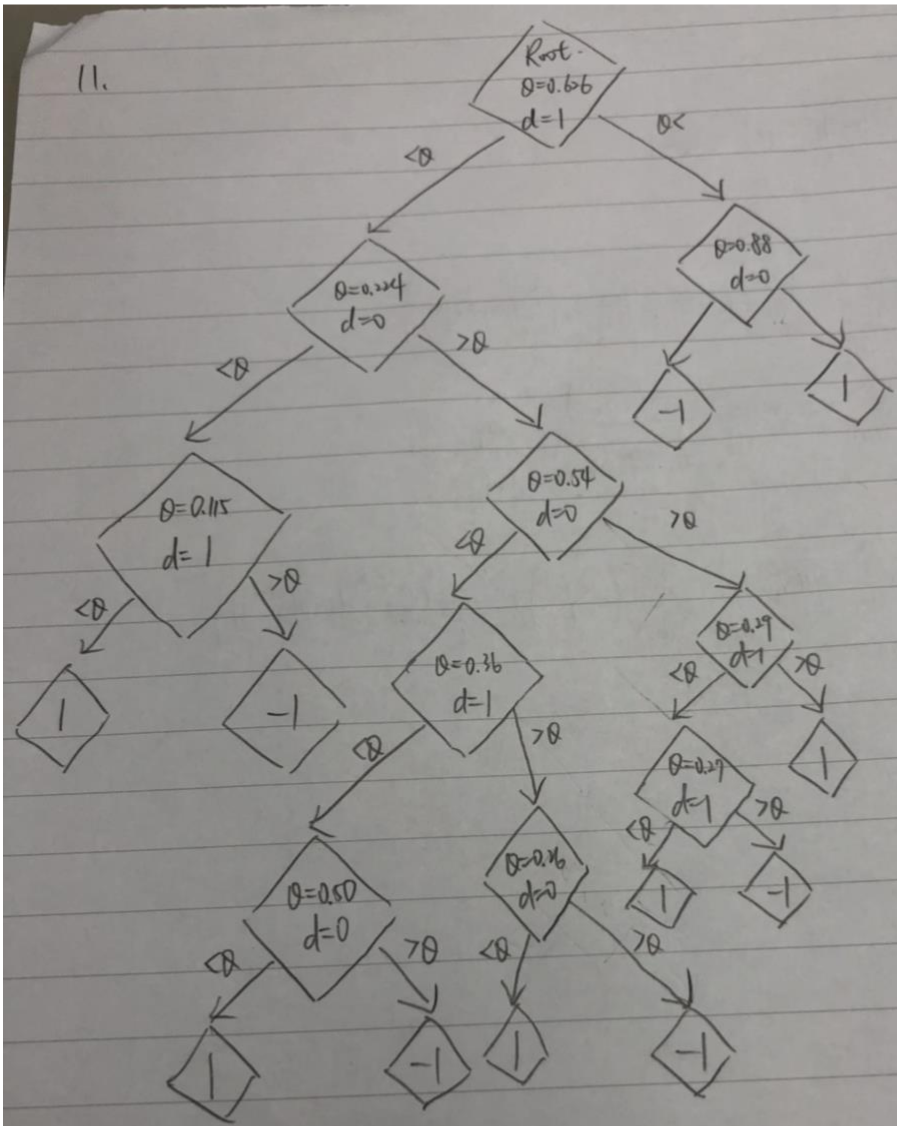
② $y=k$：

$$\frac{\partial e}{\partial s_k^{(L)}}=\frac{\partial}{\partial s_k^{(L)}}\left(-\ln(\exp(s_k^{(L)}))+\ln\left(\sum_{t=1}^{K}\exp(s_t^{(L)})\right)\right)=\frac{\partial}{\partial s_k^{(L)}}\cdot\left(-s_k^{(L)}\right)+\frac{\partial}{\partial s_k^{(L)}}\left(\ln\left(\sum_{t=1}^{K}(\exp(s_t^{(L)}))\right)\right)$$

$$=-1+\frac{\exp(s_k^{(L)})}{\sum_{t=1}^{K}\exp(s_t^{(L)})}=q_k-\overset{=1}{v_k}$$

$$\Rightarrow \frac{\partial e}{\partial s_k^{(L)}}=q_k-v_k \ \#$$

11.

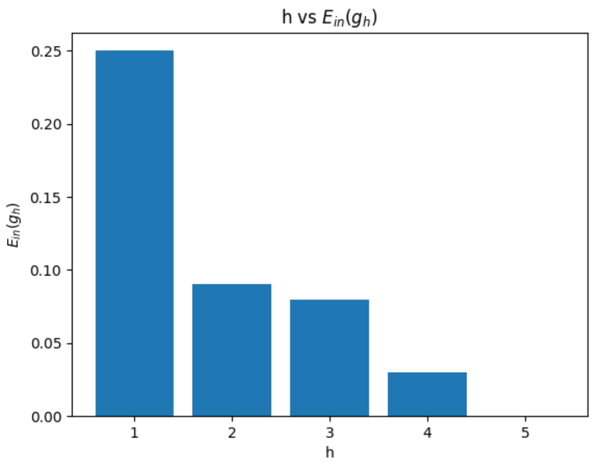


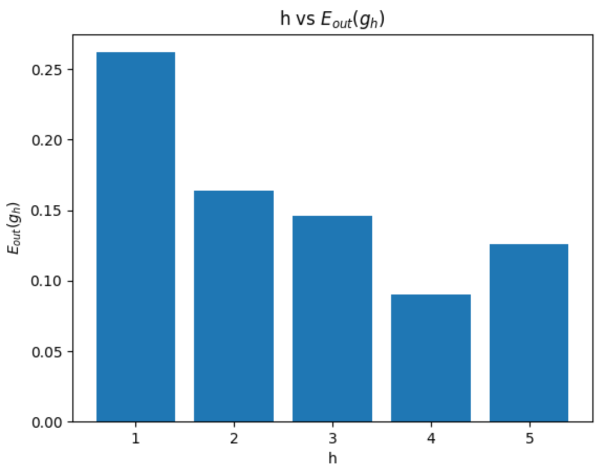
Root.
θ=0.66
d=1
<θ
θ<
θ=0.224
d=0
θ=0.88
d=0
<θ
>θ
-1
1
θ=0.115
d=1
θ=0.54
d=0
<θ
>θ
<θ
>θ
1
-1
θ=0.36
d=1
θ=0.29
d=1
<θ
>θ
<θ
>θ
θ=0.50
d=0
θ=0.26
d=0
θ=0.27
d=1
1
<θ
>θ
<θ
>θ
<θ
>θ
1
-1
1
-1
1
-1

12.

E_{in} = 0.0, E_{out} = 0.126

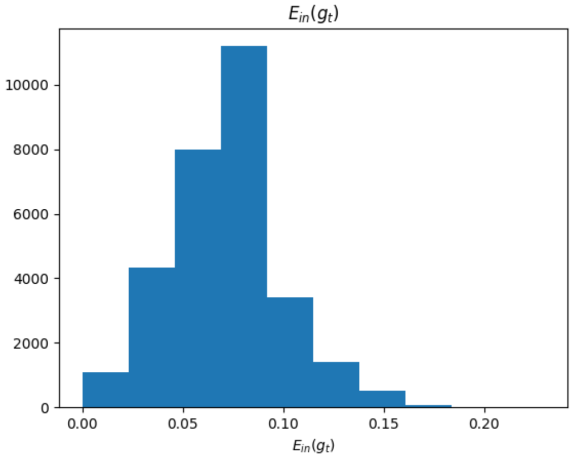
13.



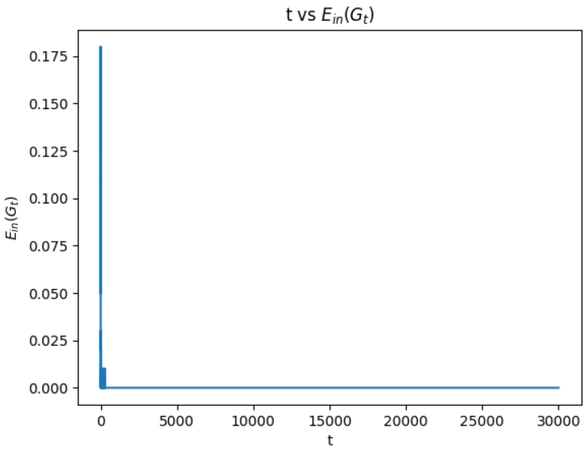


可以看出 $E_{in}(g_h)$隨 h 嚴格遞減，因為 tree 越深只會分越細，$E_{in}(g_h)$必然越來越小。而 $E_{out}(g_h)$雖然有遞減趨勢，但卻沒有嚴格遞減。可以看出整體來說 h 越大，$E_{in}(g_h)$和 $E_{out}(g_h)$都會越來越小，直到樹完全長完為止(此例 h 最高為 5)。

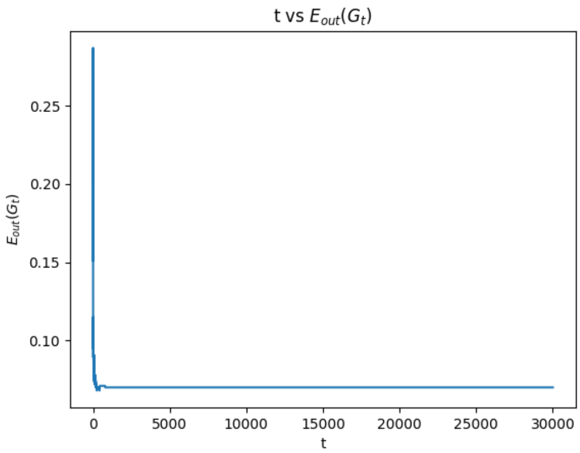
14.



15.



16.



15, 16 題的兩張圖非常相似，兩個都在 tree 數量小於 1000 時就趨於穩定。第一張圖的 $E_{in}(G_t)$最後穩定為 0，而 $E_{out}(G_t)$則在 0.05 左右。所以其實樹的數量並不需要太多，畢竟太多只會增加 model 的計算量，對降低 error rate 並沒有實質幫助。最好是能夠讓 tree 的 random 因素多一點增加 diversity 才會讓 forest 更有效。

Date:

17.

已知 $XOR(x_1, x_2 \cdots x_d)$ 即為計算當中為TRUE的邏輯元數量，奇數則 return TRUE，偶數則 return FALSE

對於 $x_1, x_2 \cdots x_d$，construct 一個 d-d-1 feed forward network such that

$$\begin{cases} W_{0k}^{(1)} = d+0.5-2k & \text{for } k \le d \\ W_{ij}^{(1)} = 1 & \text{for } i, j \le d,\ i \ne 0 \end{cases}$$

→ 如此一來，若 $x_i^{(1)}$ = TRUE，則代表 $x_1 \sim x_d$ 中至少有 i 個 1

$$\begin{cases} W_{01}^{(2)} = -0.5 \\ W_{i1}^{(2)} = -1^{i+1} & \text{for } i \le d,\ i \ne 0 \end{cases} \Rightarrow W_{i1}^{(2)} = (-0.5, 1, -1, 1, -1 \cdots)$$

當 $x_1 \cdots x_d$ 有k個TRUE，k為奇數：

$$f_d(x_1 \cdots x_d) = -0.5 + \overbrace{(\cancel{1+(-1)}+\cancel{1+(-1)}+\cancel{\cdots}+1)}^{\text{前k項}x_i^{(1)}\text{為}1} \times 1 + \overbrace{((-1)+1+(-1)+1+\cdots)}^{\text{後d-k項}x_i^{(1)}\text{為}-1} \times (-1)$$

$$= -0.5 + 1 + \{-1, 0\} \times -1 = 0.5 + \{1, 0\} > 0$$

※取決於d為奇或偶

$$XOR(x_1, x_2 \cdots x_d) = sign(f_d(x_1, \cdots x_d)) = 1$$

當 $x_1 \cdots x_d$ 有k個TRUE，k為偶數：

$$f_d(x_1 \cdots x_d) = -0.5 + (\cancel{1+(-1)}+\cancel{1+(-1)}+\cdots+\cancel{1+(-1)}) \times 1 + (1+(-1)+1+(-1)+\cdots) \times (-1)$$

$$= -0.5 + 0 + \{1, 0\} \times -1 = \{-1.5, -0.5\} < 0$$

※取決於d為奇或偶

$$XOR(x_1, x_2 \cdots x_d) = sign(f_d(x_1, \cdots x_d)) = -1$$

⇒ 此 d-d-1 feed-forward NN 即可 implement $XOR(x_1, x_2 \cdots x_d)$