B04901069 電機四 林志皓

**Problem 1**

① want max Gini impurity $1-\sum_{k=1}^{K}\mu_k^2$.

$\Rightarrow$ want min $\sum_{k=1}^{K}\mu_k^2$

$$\sum_{k=1}^{K}\mu_k^2 = \sum_{k=1}^{K}\mu_k^2 - \frac{2}{K}\left(\sum_{k=1}^{K}\mu_k\right) + \frac{2}{K} \qquad \because \sum_{k=1}^{K}\mu_k = 1$$

$$= \sum_{k=1}^{K}\left(\mu_k^2 - \frac{2}{K}\mu_k + \frac{1}{K^2}\right) + \frac{1}{K}$$

$$= \sum_{k=1}^{K}\left(\mu_k - \frac{1}{K}\right)^2 + \frac{1}{K}$$

let $x_k = \mu_k - \frac{1}{K}$ $\forall k=1\ldots K$

$$\sum_{k=1}^{K}x_k = \sum_{k=1}^{K}\left(\mu_k - \frac{1}{K}\right) = 1-1 = 0 \quad \text{— ①}$$

$\sum_{k=1}^{K}\left(\mu_k-\frac{1}{K}\right)^2 = \sum_{k=1}^{K}x_k^2 = 0$ when $x_k = 0$ $\forall k=1\ldots K$.

satisfy ① and reach minimum ($x_k^2 \geq 0$ $\forall k=1\ldots K$).

$$\therefore \sum_{k=1}^{K}\mu_k^2 = \sum_{k=1}^{K}x_k^2 + \frac{1}{K} \geq 0 + \frac{1}{K} = \frac{1}{K}$$

(minimum is reached when $\mu_k = \frac{1}{K}$ $\forall k=1\ldots K$)

$\therefore$ Gini impurity $= 1-\sum_{k=1}^{K}\mu_k^2 \leq 1-\frac{1}{K}$ (maximum)

#

**Problem 2**

② $\mu_+ + \mu_- = 1 \qquad \mu_- = 1 - \mu_+$

$$\mu_+\left(1-(\mu_+-\mu_-)\right)^2 + \mu_-\left(-1-(\mu_+-\mu_-)\right)^2$$

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

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

$$= \mu_+(1-\mu_+)\left[4(1-\mu_+) + 4\mu_+\right]$$

$$= 4\mu_+(1-\mu_+) \quad — \ ①$$

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

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

① = 2 * ② ⇒ ① is a scaled version of ② (Gini impurity)

**Problem 3**

③

Each time we select a sample from examples each example has probability of $\frac{1}{N}$ of chosen.

$\Rightarrow$ has probability of $1-\frac{1}{N}$ of "not chosen".

$\Rightarrow$ $pN$ rounds of selection. $\Rightarrow$ $\left(1-\frac{1}{N}\right)^{pN}$

$$\left(1-\frac{1}{N}\right)^{pN} = \left(\frac{N-1}{N}\right)^{pN} = \left(\frac{N}{N-1}\right)^{-pN} = \left(1+\frac{1}{N-1}\right)^{-pN}$$

$$\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^{pN} = \lim_{N\to\infty}\left[\left(1+\frac{1}{N-1}\right)^{N-1}\cdot\left(1+\frac{1}{N-1}\right)\right]^{-p}$$

$$= (e\cdot 1)^{-p} = e^{-p}$$

After bootstrapping, each example has probability of $e^{-p}$ of "not chosen"

$\Rightarrow$ Total $e^{-p}\cdot N$ of examples will not be sampled at all

**Problem 4**

④

∵ In Random Forest Algorithm. all Decision Trees votes "uniformly" for final result

∴ for each wrong example in $G$, at least need $\frac{K+1}{2}$ Decision Tree that also do wrong on this example, (outperform the correct ones in voting)

Let $G$ do wrong on $W_G$ examples. $g_k$ do wrong on $W_{g_k}$

$\Rightarrow$ at least need $\frac{K+1}{2} W_G$ wrong examples for all DT.

which is bounded : $\frac{K+1}{2} W_G \leq \sum_{k=1}^{K} W_{g_k}$

Assume there are all $N$ examples

$$\Rightarrow \frac{K+1}{2} \cdot \frac{W_G}{N} \leq \sum_{k=1}^{K} \frac{W_{g_k}}{N} \Rightarrow \frac{K+1}{2} E_{out}(G) \leq \sum_{k=1}^{K} E_{out}(g_k)$$

$$\Rightarrow E_{out}(G) \leq \frac{2}{K+1} \sum_{k=1}^{K} e_k \quad \#$$

**Problem 5**

(5)

$\alpha_1$ optimizes $\sum_{n=1}^{N} err(y_n, s_n)$

$$= \sum_{n=1}^{N} (y_n - 11.26\alpha)^2 = E$$

$$\frac{\partial E}{\partial \alpha} = \sum_{n=1}^{N} \frac{\partial (y_n - 11.26\alpha)^2}{\partial \alpha} = \sum_{n=1}^{N} 2(y_n - 11.26\alpha)\cdot(-11.26).$$

$\frac{\partial E}{\partial \alpha} = 0$ when $\sum_{n=1}^{N} (y_n - 11.26\alpha) = 0.$

$$\Rightarrow \sum_{n=1}^{N} y_n - \sum_{n=1}^{N} 11.26\alpha = \sum_{n=1}^{N} y_n - N\cdot 11.26\alpha = 0$$

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

**Problem 6**

⑥ Let $s_n^{(t-1)}$ be the $s_n$ before iteration $t$.

$s_n^{(t)}$ be the $s_n$ after iteration $t$.

Steepest $\eta \Rightarrow$ optimize $\sum_{n=1}^{N}\left(y_n - s_n^{(t-1)} - \alpha g_t(x_n)\right)^2$

$$\frac{\partial E}{\partial \alpha} = \sum_{n=1}^{N} 2\left(y_n - s_n^{(t-1)} - \alpha g_t(x_n)\right)\cdot\left(-g_t(x_n)\right)$$

$$\frac{\partial E}{\partial \alpha} = 0 \text{ when } \alpha = \frac{\sum_{n=1}^{N}\left(y_n - s_n^{(t-1)}\right) g_t(x_n)}{\sum_{n=1}^{N} g_t^2(x_n)}.$$

$$\therefore \sum_{n=1}^{N} s_n g_t(x_n) = \sum_{n=1}^{N}\left(s_n^{(t-1)} + \alpha g_t(x_n)\right) g_t(x_n)$$

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

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

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

**Problem 7**

7

Let $\{x_n, y_n\}_{n=1}^{N}$ be all $N$ examples and ground truths.

Polynomial Regression $\Rightarrow$ Optimize $E = \sum_{n=1}^{N} (y_n - w^T x_n)^2$.

let $g_1(x) = (w^*)^T x$ where $w^*$ leads to min $E$.

To get $\alpha_1 \Rightarrow$ Optimize $E' = \sum_{n=1}^{N} (y_n - \alpha g_1(x_n))^2$

assume $\alpha_1 \neq 1 \Rightarrow \sum_{n=1}^{N} (y_n - \alpha g_1(x_n))^2 < \sum_{n=1}^{N} (y_n - g_1(x_n))^2$

$\Rightarrow \sum_{n=1}^{N} (y_n - \alpha w^{*T} x_n)^2 \leq \sum_{n=1}^{N} (y_n - w^{*T} x_n)^2$

$\Rightarrow \alpha \cdot w^*$ leads to smaller $E$ than $w^*$.

$\Rightarrow$ contradiction. $\Rightarrow \alpha_1 = 1$ #

**Problem 8**

(8)

$x_0 = 1$ (constant).

Let $w_0 = d - \frac{1}{2}$. $w_i = 1$ $\forall i = 1, 2 \ldots d$

and if $x_1 = x_2 = \cdots = x_d = -1$.

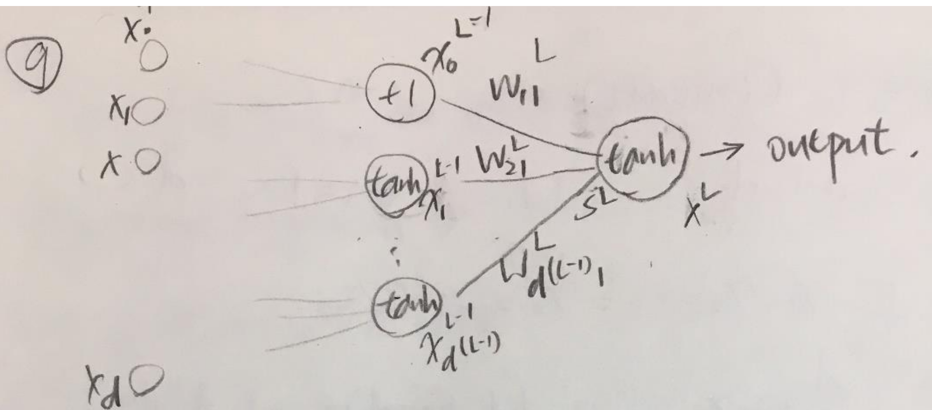
$\text{sign}\left(\sum_{i=0}^{d} w_i x_i\right) = \text{sign}\left(d - \frac{1}{2} - d\right) = -1$. (False)

if there's any $x_i = 1$. the $\text{sign}\left(\sum_{i=0}^{d} w_i x_i\right) = +1$ (True)

$\Rightarrow$ $OR(x_1, x_2 \cdots x_d)$ $\therefore$ $w_0 = d - \frac{1}{2}$. $w_i = 1$ $\forall i = 1, 2 \ldots d$

is one of the solution

**Problem 9**




$\because$ All weights $w_{ij}^{(l)}$ are $0$.

$\therefore$ $s_j^{(l)} = \sum_{i=0}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)} = 0$.

$\therefore$ $x_j^{(l)} = \tanh(s_j^{(l)}) = \tanh(0) = 0 \quad (l \neq 1)$

final output $x^L$ has gradient $\dfrac{\partial e}{\partial x^L}$

$$\frac{\partial e}{\partial w_{ij}^{(l)}} = \frac{\partial e}{\partial s_j^{(l)}} \cdot \frac{\partial s_j^{(l)}}{\partial w_{ij}^{(l)}} = \frac{\partial e}{\partial s_j^{(l)}} \cdot x_i^{(l-1)} = 0.$$

$$\frac{\partial e}{\partial x_i^{(l-1)}} = \frac{\partial e}{\partial s_j^{(l)}} \cdot \frac{\partial s_j^{(l)}}{\partial x_i^{(l-1)}} = \frac{\partial e}{\partial s_j^{(l)}} \cdot w_{ij}^{l} = 0.$$

#

**Problem 10**

10 Let $V_i = 1$, $V_k = 0 \ \forall k \neq i$

$$e = -\sum_{k=1}^{K} V_k \ln q_k = -\ln q_i$$

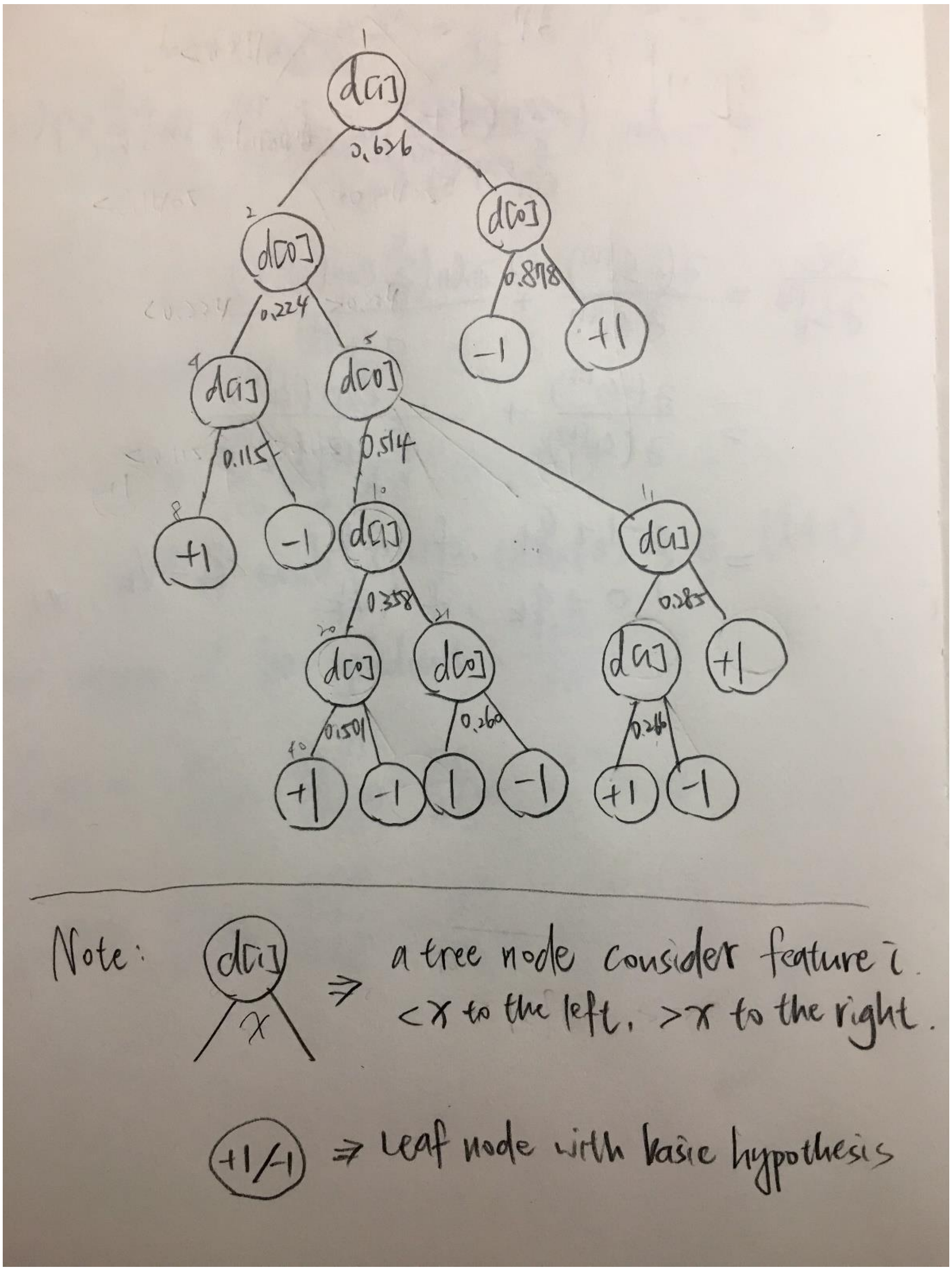
$$= -\ln\left(\frac{\exp(s_i^{(L)})}{\sum_{k=1}^{K}\exp(s_k^{(L)})}\right) = -s_i^{(L)} + \ln\left(\sum_{k=1}^{K}\exp(s_k^{(L)})\right)$$

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

$$= \frac{\partial(-s_i^{(L)})}{\partial (s_k^{(L)})} + \frac{\exp(s_k^{(L)})}{\sum_{k=1}^{K} \exp(s_k^{(L)})}$$

$$= \begin{cases} -1+q_k, & \text{if } i=k \\ 0+q_k, & \text{if } i\neq k\end{cases} = q_k - V_k \quad \#$$

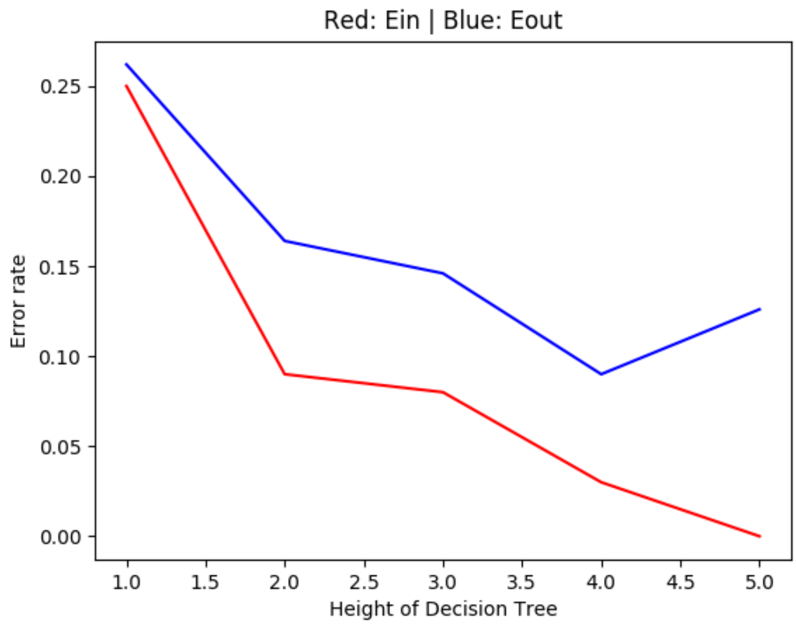
**Problem 11**




**Problem 12**

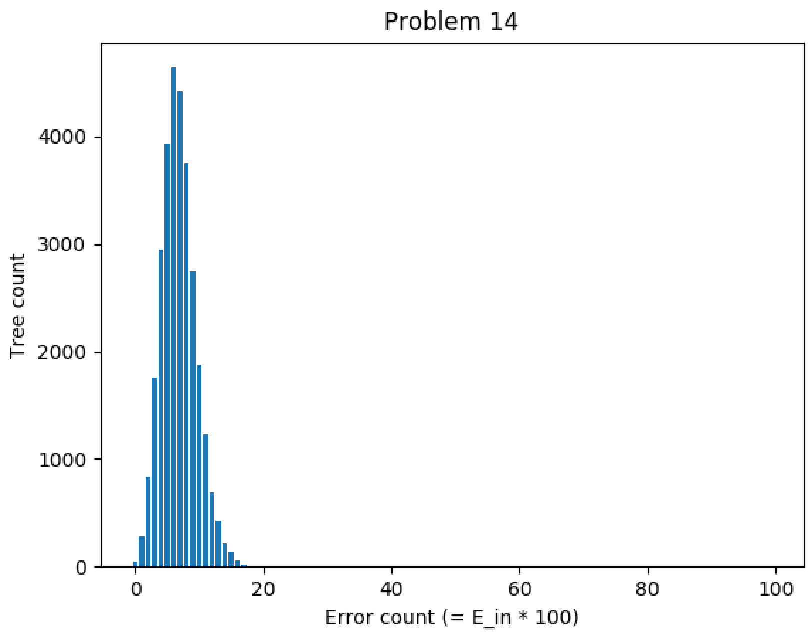
E in = 0.0, E out = 0.126

## Problem 13

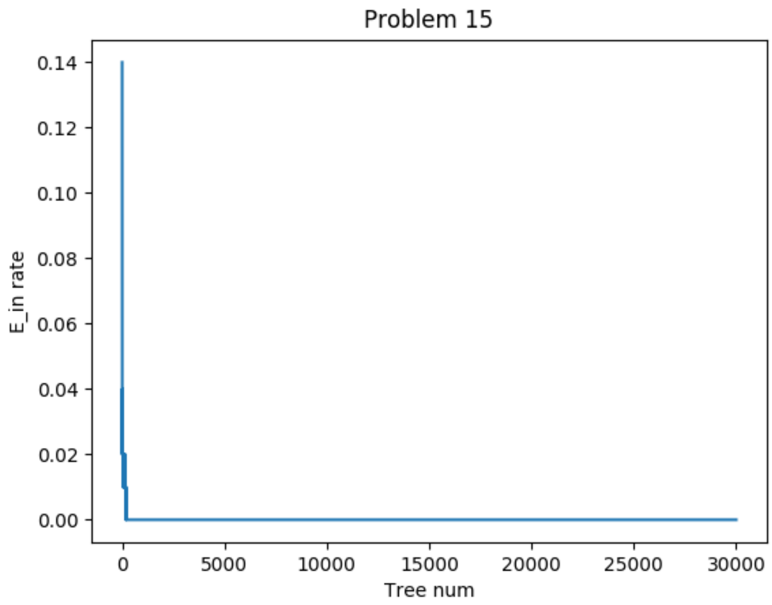



According to this graph, with the height increasing, the E_in dedreases to 0 gradually, and the E_out also decreases, it shows that the heigher the decision tree is, it's more powerful.

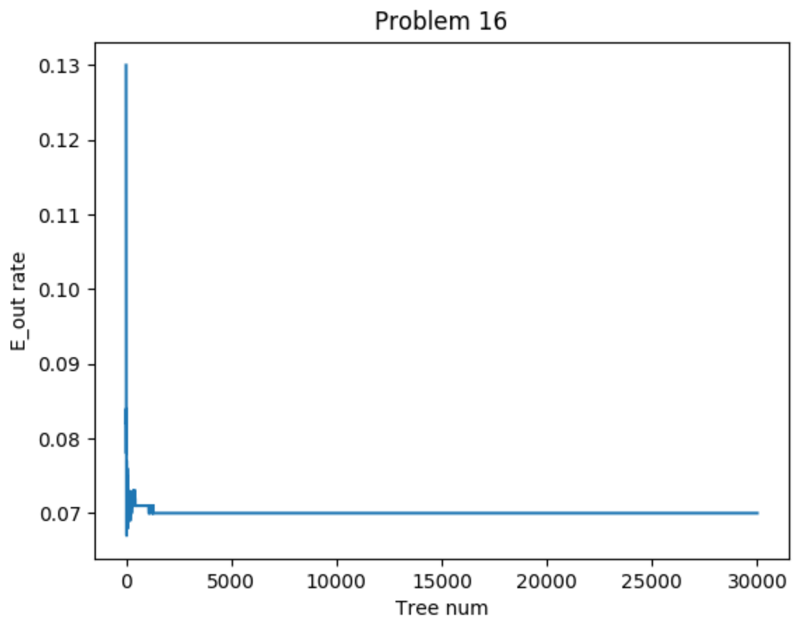
## Problem 14



## Problem 15



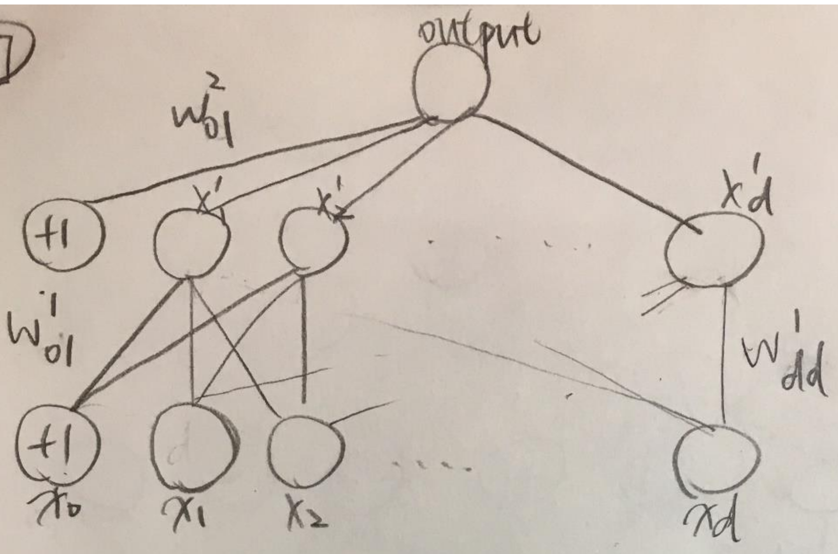

## Problem 16




Compared with problem 15, the E_in and E_out both decreases while the size of random forest becomes bigger. With 30000 trees, E_in =0, E_out = 0.07, and it's not overfitting, I think it's bagging that makes random forest more stable than single decision tree.

**Problem 17**

17




$XOR(x_1, x_2 \cdots x_d) \equiv$ There are odd number of +1 among $x_1 \sim x_d$

$w^1_{0i} = d-2i+1$ $\quad w^1_{jk} = 1$, $j \neq 0$

$\Rightarrow$ $x'_i = +1$ iff there are $n$ +1 among $x_1 \sim x_d$. $n \geq i$

$= -1$ otherwise.

$w^2_{i1} = (-1)^{i+1}$, $\quad w^2_{01} = -0.5$.

$\Rightarrow$ output $= +1$ iff $x'_i = 1 \ \forall i \leq k$ $\ x'_i = -1 \ \forall i > k$.

k is an odd number

output $= -1$ iff $x'_i = 1 \ \forall i \leq k$. $x'_i = -1 \ \forall i > k$

k is an even number.

$\Rightarrow$ Implemented $XOR(x_1, x_2 \cdots x_d)$

**Problem 18**