Machine Learning Techniques Homework 4

1

The probability that an example is not sampled by one sampling is $1 - \frac{1}{N}$. Since pN examples are sampled out, the probability of an example is not sampled at all is

$$(1 - \frac{1}{N})^{pN} = ((1 - \frac{1}{N})^{N})^{p} = (e^{-1})^{p} = e^{-p}$$

Thus, approximately $e^{-p}N$ examples will not be sampled at all.

2

Minimal $E_{in} = 0$. Since to achieve no prediction error, only two of the three tree predicting correctly is required. Let the test data set be X, and

$$\begin{cases} X_1 = \{x \in X |\ g_1(x) \text{ is incorrect}\} \\ X_2 = \{x \in X |\ g_2(x) \text{ is incorrect}\} \\ X_3 = \{x \in X |\ g_3(x) \text{ is incorrect}\} \end{cases}$$

since $E_{out}(g_1) + E_{out}(g_2) + E_{out}(g_3) \leq 1$, it is possible that X_1, X_2, X_3 are pairwaise disjoint. So if X_1, X_2, X_3 are pairwise disjoint, for all $x \in X$, if $x \in X_k$, $x \notin X_i$ for $i \in \{1, 2, 3\}$
k. It implies that for any $x \in X$, at least two of g_1, g_2, g_3 will predict correctly. Therefore, $E_{in} = 0$ is attained.

Maximun $E_{out} = 0.375$, which is achieveable only if $X_1 \cup X_2 \cup X_3 = \phi$ to maximize

$$E_{out} = \frac{|(X_1 \cap X_2) \cup (X_2 \cap X_3) \cup (X_1 \cap X_3)|}{|X|}$$

And also since an error prediction on x requires two g that predict incorrectly, thus the constraint becomes

$$\begin{cases} X_3 \subset X_1 \cup X_2 \\ X_3 \cap (X_1 \cap X_2) = \phi \end{cases}$$

So the $E_{out}(G) = (|X_3| + |X_1 \cap X_2|)/|X|$ where $|X_3| = 0.35|X|$, and

$$\begin{aligned} &\text{maximize} \quad |X_1 \cap X_2| \\ &\text{subject to} \quad \begin{cases} X_3 \subset X_1 \cup X_2 \\ X_3 \cap (X_1 \cap X_2) = \phi \end{cases} \end{aligned}$$

is to

$$\begin{aligned} &\text{maximize} \quad |X_4| \\ &\text{subject to} \quad \begin{cases} X_4 \subset X_1, X_2 \\ X_3 \subset (X_1 \setminus X_4) \cup (X_2 \setminus X_4) \\ X_3 \cap X_4 = \phi \end{cases} \end{aligned}$$

that is

$$\begin{aligned} &\text{maximize} \quad |X_4| \\ &\text{subject to} \quad \begin{cases} |X_3| \leq (|X_1| - |X_4|) + (|X_2| - |X_4|) \end{cases} \end{aligned}$$

we get the maximum $|X_4| = 0.025$. Thus, maximum $E_{out}(G) = 0.35 + 0.025 = 0.375$.

3

Let X be the test example set, and

$$X_i = \{x \in X | g_i(x) \text{ is incorrect}\}$$

Let

$$W_p = \bigcap_{i \in p} X_i$$

where $p \in P$ and

$$P = \{p \subset \{1, 2, 3, \cdots, K\} || p| = \frac{K+1}{2}\}$$

So for $x \in W_p$, x is predicted incorrectly by $\frac{K+1}{2}$ hypothesises gs, and thus predicted incorrectly by G. Therefore,

$$\begin{aligned}
&E_{out}(G) \cdot |X| \\
=&|\bigcup_{p \in P} W_p| \\
=&\sum_{p \in P} |W_p| - \sum_{p_1, p_2 \in P; p_1 \neq p_2} |W_{p_1} \cap W_{p_2}| + \sum_{p_1, p_2, p_3 \in P; p_1 \neq p_2 \neq p_3 \neq p_1} |W_{p_1} \cap W_{p_2} \cup W_{p_3}| \cdots
\end{aligned}$$

Since if $\{W_p | p \in P\}$ are not pairwise disjoint,

$$- \sum_{p_1, p_2 \in P; p_1 \neq p_2} |W_{p_1} \cap W_{p_2}| + \sum_{p_1, p_2, p_3 \in P; p_1 \neq p_2 \neq p_3 \neq p_1} |W_{p_1} \cap W_{p_2} \cup W_{p_3}| \cdots < 0$$

$E_{out}(G) \cdot |X|$ is upper bounded

$$E_{out}(G) \cdot |X| = |\bigcup_{p \in P} W_p| \leq \sum_{p \in P} |W_p|$$

And we know that when $\{W_p | p \in P\}$ are pairwise disjoint

$$\begin{aligned}
\sum_{p \in P} |W_p| = |\bigcup_{p \in P} W_p| \\
&= |\{x \in W_p \text{ for some } p\}| \\
&= |\{x | x \text{ in } \frac{K+1}{2} \text{ sets } (X_i, \forall i \in p) \text{ that contain } x\}| \\
&= \frac{2}{K+1} \sum_{k=1}^{K} |X_k| \\
&= \frac{2}{K+1} \sum_{k=1}^{K} |X| e_k
\end{aligned}$$

Therefore,

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

4

To solve

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

Let

$$\frac{\partial}{\partial\eta}\sum_{n=1}^{N}(y_n - 2\eta)^2 = \sum_{n=1}^{N} -4(y_n - 2\eta) = 0$$

Get

$$\alpha_1 = \eta = \frac{\sum_{n=1}^{N} y_n}{2N}$$

Thus

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

5

Let s'_n be the value of s_n before updated. Then the optimal η is the root of the derivative

$$\frac{\mathrm{d}}{\mathrm{d}\eta}\frac{1}{N}\sum_{n=1}^{N}((y_n - s'_n) - \eta g_t(x_n))^2 = \sum_{n=1}^{N} 2g_t(x_n)(y_n - s'_n - \eta g_t(x_n)) = 0$$

So we get

$$\alpha_t = \eta = \frac{\sum_{n=1}^{N} g_t(x_n)(y_n - s'_n)}{\sum_{n=1}^{N} g_t^2(x_n)}$$

Then

$$\begin{aligned}
&\sum_{n=1}^{N} s_n g_t(x_n) \\
=&\sum_{n=1}^{N}(s'_n + \alpha_t g_t(x_n))g_t(x_n) \\
=&\sum_{n=1}^{N} s'_n g_t(x_n) + \alpha_t \sum_{n=1}^{N} g_t^2(x_n) \\
=&\sum_{n=1}^{N} s'_n g_t(x_n) + \frac{\sum_{n=1}^{N} g_t(x_n)(y_n - s'_n)}{\sum_{n=1}^{N} g_t^2(x_n)} \sum_{n=1}^{N} g_t^2(x_n) \\
=&\sum_{n=1}^{N} g_t(x_n) y_n
\end{aligned}$$

6

A general polynomial regression problem is to find an optiomal w that minimizes

$$\|y - g(z)\|^2 = \|y - w^T z\|^2$$

where

$$z_n = (1, x_n, x_n^2, x_n^3, \cdots, x_n^k)$$

for a k-degree polynomial gregression. (x_n^k is elementwise power of x_n.)
So the optimal w satisfies

$$\frac{\partial}{\partial w}\|y - w^T z\|^2 = 2z^T(y - w^T z) = 0$$

where $w^T z = g(z)$ and since the first row of z^T is an all 1 vector, therefore we have

$$1^T(y - g(z)) = 1^T y - 1^T g(z) = 0$$

That is

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

So to find minimal η, let

$$\frac{\partial}{\partial \eta}\sum_{n=1}^{N}(y_n - \eta g(x_n))^2 = \sum_{n=1}^{N} -2g(x_n)(y_n - \eta g(x_n)) = 0$$

Get

$$\eta = \frac{\sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} g(x_n)} = 1$$

7

Let the optimal w in g_1, g_2 be w_1, w_2 respectively. Then finding the optimal g_2 is to find the optimal w_2 that minimize

$$\|y - s - \eta g_t(x)\|^2 = \|y - w_1^T z - \eta w_2^T z\|^2 = \|y - (w_1^T + \eta w_2^T)z\|^2$$

where z is defined as in problem 6. But we know that w_1 is the optimal solution that minimizes

$$\|y - w_1^T z\|^2$$

Thus $w_2 = 0$, and therefore $g_2(x) = 0$

8

$$w_i = \begin{cases} 1 & i \neq 0 \\ d-1 & i = 0 \end{cases}$$

then g_A is equivalent to OR since if $x_1 = x_2 = x_3 = \cdots = x_d = -1$,

$$g_A(x_1, x_2, x_3, \cdots, x_d) = \text{sign}(\sum_{i=1}^{d} w_i x_i + w_0) = \text{sign}(-d + d - 1) = -1 = \text{OR}(x_1, x_2, x_3, \cdots, x_d)$$

And if any $x_i = 1$, say $x_k = 1$

$$\begin{aligned}
g_A(x_1, x_2, x_3, \cdots, x_d) =& \text{sign}(\sum_{i=1}^{d} w_i x_i + w_0) \\
=& \text{sign}(\sum_{i=1; i \neq k}^{d} w_i x_i + w_k x_k + w_0) \\
\geq& \text{sign}(-(d-1) + 1 + d - 1) \\
=& +1 = \text{OR}(x_1, x_2, x_3, \cdots, x_d) \\
\therefore& g_A(x_1, x_2, x_3, \cdots, x_d) = \text{OR}(x_1, x_2, x_3, \cdots, x_d)
\end{aligned}$$

9

First layer is of 5 neuron:

$$g_1(x) = \text{sign}\left(\sum_{i=1}^{5} x_i + 4\right) \qquad g_2(x) = \text{sign}\left(\sum_{i=1}^{5} x_i + 2\right)$$

$$g_3(x) = \text{sign}\left(\sum_{i=1}^{5} x_i + 0\right) \qquad g_4(x) = \text{sign}\left(\sum_{i=1}^{5} x_i - 2\right)$$

$$g_5(x) = \text{sign}\left(\sum_{i=1}^{5} x_i - 4\right)$$

and the output layer is of neuron

$$g_6(x) = \text{sign}\left(-\sum_{i=1}^{5} (-1)^i x_i\right)$$

The neurons, g_1, g_2, g_3, g_4, g_5, are activated only when number of positive input is greater or equal to $1, 2, 3, 4, 5$ respectively (, wihch can be verify easily). So if there are k positive inputs, only neurons g_i for $i \leq k$ are avtivated. And the output layer ensure that it is only activated when only neuron $\{g_i | i \leq k, \}$, where k is odd, are activated (, which can also be verified easily). In conclusion, the output layer only output 1 when number of positive inputs is odd.

10

For $l \geq 2$, $0 \leq i \leq d^{(l-1)}$, $1 \leq j \leq d^{(l)}$:

$$\begin{aligned}\frac{\partial e_n}{\partial w_{ij}^{(l)}} =&\delta_j^{(l)} \cdot (x_i^{(l-1)})\\ =&\delta_j^{(l)} \cdot \tanh(\sum_k w_{ki}^{(l-1)} x_k^{(l-2)})\\ =&0\end{aligned}$$

since $w_{ij} = 0$.

11

For all j that satisfy $1 \leq j < d^{(1)}$,

$$s_j^{(1)} = \sum_{i=0}^{d^{(0)}} w_{ij}^{(1)} x_i^{(0)} = \sum_{i=0}^{d^{(0)}} x_i^{(0)} = \sum_{i=0}^{d^{(0)}} w_{i(j+1)}^{(1)} x_i^{(0)} = s_{j+1}^{(1)}$$

implies if $w_{j1}^{(2)} = w_{(j+1)1}^{(2)}$ (which holds at the start)

$$\delta_j^{(1)} = \delta_1^{(2)} w_{j1}^{(2)} \tanh'(s_j^{(1)}) = \delta_1^{(2)} w_{(j+1)1}^{(2)} \tanh'(s_{j+1}^{(1)}) = \delta_{j+1}^{(1)}$$

thus

$$\frac{\partial e_n}{\partial w_{ij}^{(1)}} = \delta_j^{(1)} \cdot x_i^{(0)} = \delta_{(j+1)}^{(1)} \cdot x_i^{(0)} = \frac{\partial e_n}{\partial w_{i(j+1)}^{(1)}}$$

That is, after one update of weights,

$$w_{ij}^{(1)} = w_{i(j+1)}^{(1)}$$

And thus

$$s_j^{(1)} = s_{j+1}^{(1)}$$

also holds after update, which implies

$$\frac{\partial e_n}{\partial w_{j1}^{(2)}} = \frac{\partial e_n}{\partial s_1^{(2)}} \cdot x_j^{(1)} = \frac{\partial e_n}{\partial s_1^{(2)}} \cdot \tanh(s_{(j)}^{(1)}) = \frac{\partial e_n}{\partial s_1^{(2)}} \cdot \tanh(s_{(j+1)}^{(1)}) = \frac{\partial e_n}{\partial s_1^{(2)}} \cdot x_{(j+1)}^{(1)} = \frac{\partial e_n}{\partial w_{(j+1)1}^{(2)}}$$

Thus after update of weights with gradient descent,

$$w_{(j)1}^{(2)} = w_{(j+1)1}^{(2)}$$

still holds. So previous assumption that $w_{j1}^{(2)} = w_{(j+1)1}^{(2)}$ holds after update. Therefore, by mathematical induction, equations above, including $w_{ij}^{(1)} = w_{i(j+1)}^{(1)}$, hold throughout the training process.

12

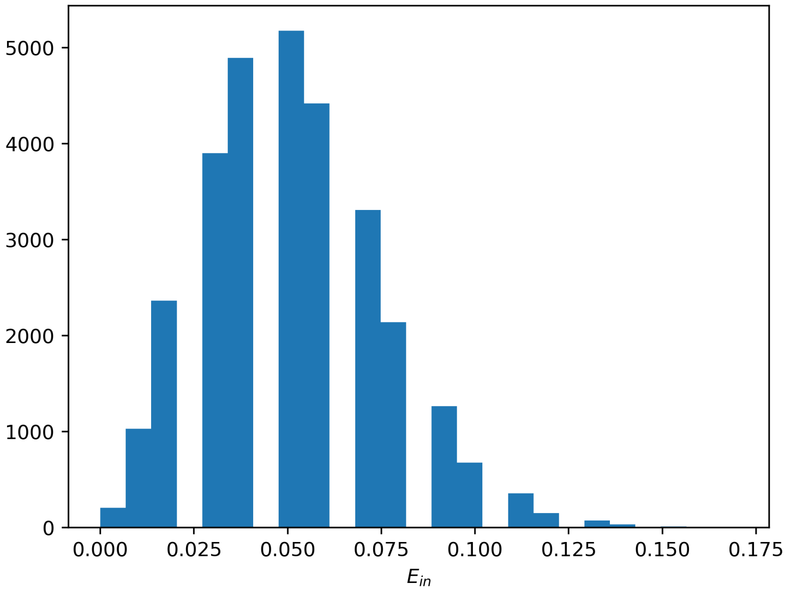

Figure 1: E_{in} over 30000 trees

13

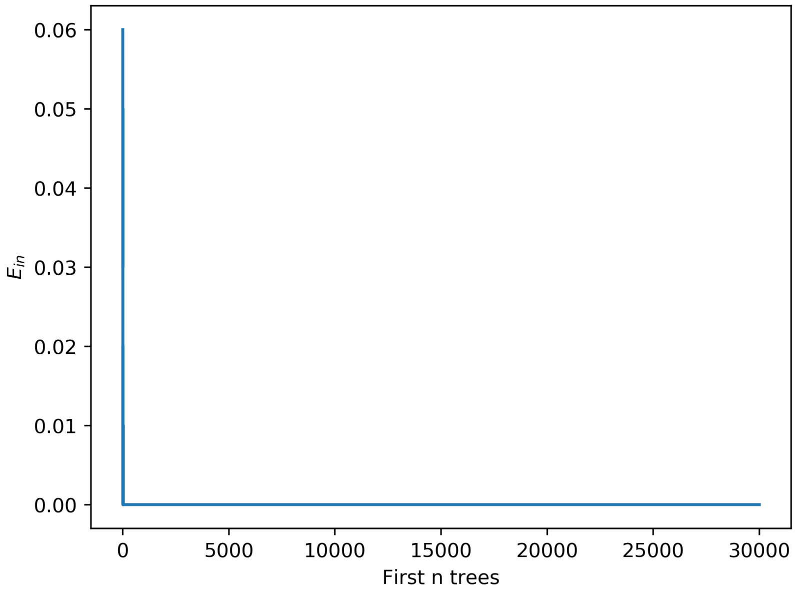

Figure 2: E_{in} over 30000 trees

14

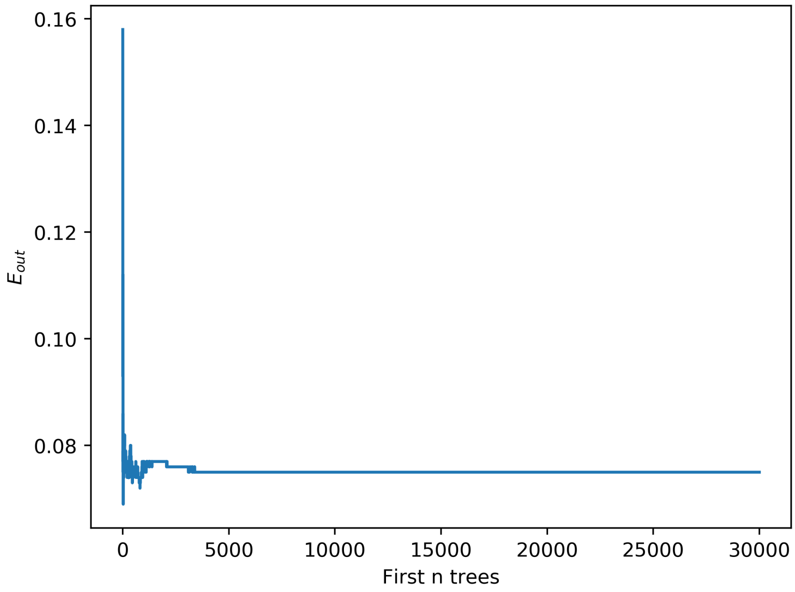

Figure 3: E_{out} over 30000 trees

In figure 2, E_{in} goes to 0 in first few trees, and keeps 0 until the end. In contrast, in figure 3, E_{out} oscilate in first few thousands of trees, and keeps about 0.08 until the end, which is higher than that of E_{in}.

15

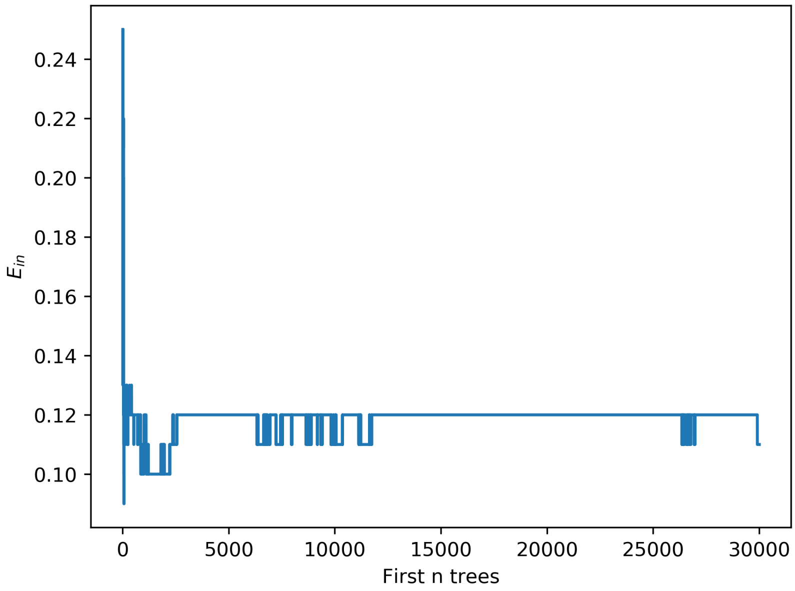

Figure 4: E_{in} over 30000 stumps

16

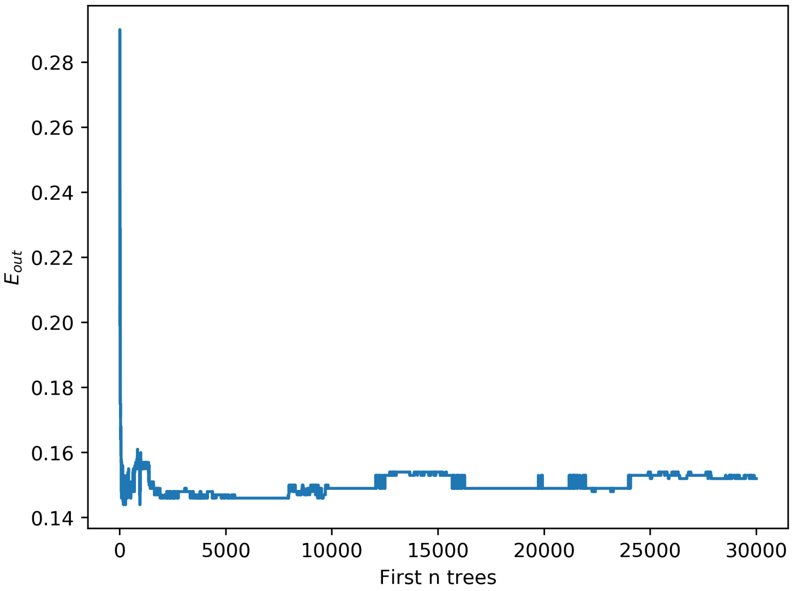

Figure 5: E_{out} over 30000 stumps

Both E_{in} and E_{out} go down drastically in first few stumps. But E_{in} goes down and up in first few thousands of trees, while E_{out} goes up and down for the same interval. And for the following trees, E_{in} is more flat than E_{out}, also E_{in} converges at lower value than E_{out}.