MLtech hw3 資工二 B05902109 柯上優

1.

With $\mathrm{u}_{-} = 1 - \mathrm{u}_{+}$, we can get

$$1 - \mathrm{u}_{+}^2 - u_{-}^2 = 1 - \mathrm{u}_{+}^2 - (1 - u_{+})^2 = -2\mathrm{u}_{+}^2 + 2\mathrm{u}_{+} = -2\left(\mathrm{u}_{+} - \frac{1}{2}\right)^2 + \frac{1}{2}$$

When $\mathrm{u}_{+} \in [0, 1]$, we have maximum value 0.5 when $\mathrm{u}_{+} = 0.5$.

2.

In the problem 1, we know the normalize Gini index is

$$\frac{1 - \mathrm{u}_{+}^2 - u_{-}^2}{0.5} = -4\mathrm{u}_{+}^2 + 4\mathrm{u}_{+}$$

So we compare,

[a] the maximum value is 0.5, so normalize function is

$$2\min(\mathrm{u}_{+}, \mathrm{u}_{-})$$

[b] first we calculate the maximum

$$\mathrm{u}_{+}\left(1 - \left(\mathrm{u}_{+} - (1 - \mathrm{u}_{+})\right)\right)^2 + (1 - \mathrm{u}_{+})(-1 - (\mathrm{u}_{+} - (1 - \mathrm{u}_{+})))^2 = -4\mathrm{u}_{+}^2 + 4\mathrm{u}_{+}$$

$$= -4\left(\mathrm{u}_{+} - \frac{1}{2}\right)^2 + 1$$

The maximum value is 1, and the normalize function is

$$-4\mathrm{u}_{+}^2 + 4\mathrm{u}_{+}$$

[c]

By wolframe alpha, we know maximum value is log 2, and the normalize function is

$$\frac{-\mathrm{u}_{+}\ln(\mathrm{u}_{+}) - (1 - \mathrm{u}_{+})\ln(1 - \mathrm{u}_{+})}{\ln(2)}$$

[d]

The maximum value is 1, and the normalize function is

$$1 - |2\mathrm{u}_{+} - 1|$$

So the answer is [b].

3.

Same procedure with lecture 10 page 8,

$$\lim_{N \to inf}\left(1 - \frac{1}{N}\right)^{pN} = \left(\lim_{N \to inf}\left(1 - \frac{1}{N}\right)^N\right)^p = \left(\frac{1}{e}\right)^p = e^{-p}$$

4.

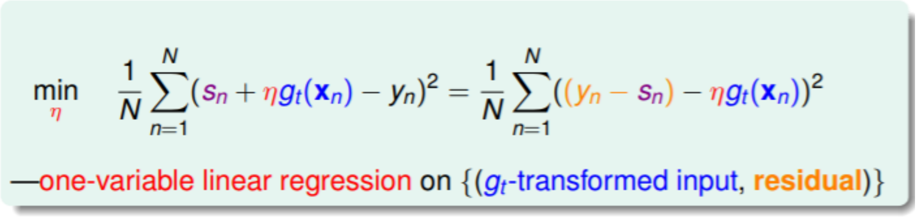
已知所有的錯誤(可能重疊) $\sum_{k=1}^{K} e_k$，在最糟的情況下，一個資料的錯誤需要

$\frac{\mathrm{K}+1}{2}$ 個樹判斷錯誤，因此，在最糟的狀況下，排除掉重複錯誤的總錯誤率為

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

5.

由 Lecture211, P.17



$$\min_{\eta} \quad \frac{1}{N}\sum_{n=1}^{N}(s_n + \eta g_t(\mathbf{x}_n) - y_n)^2 = \frac{1}{N}\sum_{n=1}^{N}((y_n - s_n) - \eta g_t(\mathbf{x}_n))^2$$

—one-variable linear regression on $\{(g_t\text{-transformed input, residual})\}$

$$\min_{\eta}\frac{1}{N}\sum_{n=1}^{N}\left((y_n - s_n) - \eta g_t(x_n)\right)^2 = \min_{\eta}\frac{1}{N}\sum_{n=1}^{N}\left((y_n) - 2*\eta\right)^2$$

$$\nabla\mathrm{E} = \frac{\partial}{\partial\eta}\frac{1}{N}\sum_{n=1}^{N}\left((y_n - s_n) - \eta g_t(x_n)\right)^2 = \frac{1}{N}\sum_{n=1}^{N}(-4y_n - 8*\eta) = 0$$

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

所求

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

6.

For the best η, we can get it by derive

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

$$\nabla\mathrm{E} = \frac{\partial}{\partial\eta}\frac{1}{N}\sum_{n=1}^{N}\left((y_n - s_n) - \eta g_t(x_n)\right)^2 = \frac{1}{N}\sum_{n=1}^{N} -2(y_n - s_n)g_t(x_n) + 2\eta g_t^2(x_n)$$

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

又 $\eta = \alpha$

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

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

所求，由前一次 $t-1$ 得到的 α_t 更新，恰好等於上面湊出的部分

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

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

7.

Solution 1>

We all know that Gradient Boosting need a weak learning model, and this is a good example. In the class, teacher told us that gradient is like walking in the montains. By updata S_n every iterations, we can gradually go to the point. But regression is too powerful that by solving inverse matrix it immediately go to the point. So with the same regression function, we will stay at the same point. Starting from the second iteration, the optimal g_2(x) will always return 0 because gradient is already 0, can not reinforce anymore.

Solution 2>

Thanks to B05902127 劉俊緯

For regression, $\mathrm{w}_{\mathrm{lin}}^1 = X^\dagger Y$, and $\mathrm{g}_1(x) = xX^\dagger Y$

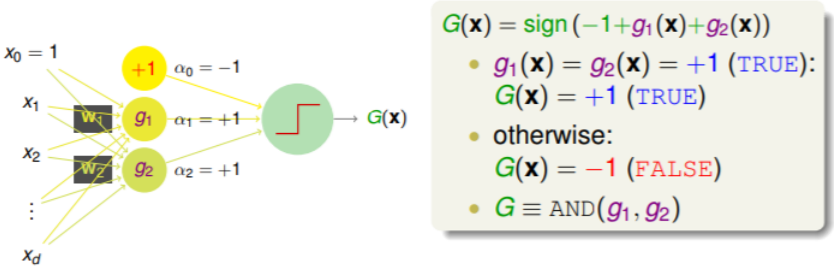
And we can derive $\mathrm{w}_{\mathrm{lin}}^2$ by

$$\mathrm{w}_{\mathrm{lin}}^2 = X^\dagger(Y - S_n) = X^\dagger(Y - XX^\dagger Y) = X^\dagger Y - X^\dagger X X^\dagger Y = X^\dagger Y - X^\dagger Y = 0$$

And thus $\mathrm{g}_2(x) = xO = 0$

8.

Similar with Lecture 212 P.3



$G(\mathbf{x}) = \mathrm{sign}\left(-1 + g_1(\mathbf{x}) + g_2(\mathbf{x})\right)$

- $g_1(\mathbf{x}) = g_2(\mathbf{x}) = +1$ (TRUE): $G(\mathbf{x}) = +1$ (TRUE)
- otherwise: $G(\mathbf{x}) = -1$ (FALSE)
- $G \equiv \mathrm{AND}(g_1, g_2)$

Only with all are False, the $\mathrm{OR}(x_1, x_2, \dots, x_d)$ will output False

My answer is

$$\{\mathrm{d}-1, 1, 1, \dots, 1\}, \text{where I let } \mathrm{sign}(0) = +1$$

Prove:

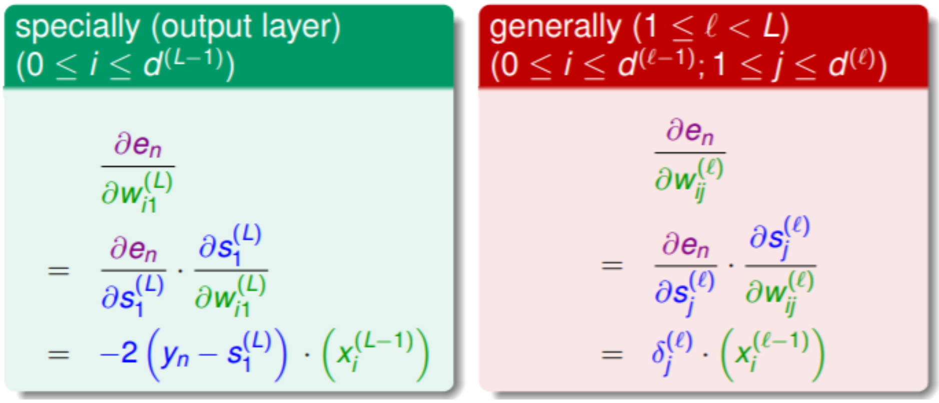
When all are True(+1), $\mathrm{sign}\left(\sum_{i=0}^{d} w_i x_i\right) = sign(2d-1) = 1$

When α True, β False, $\beta \neq \mathrm{d}$, $\mathrm{sign}\left(\sum_{i=0}^{d} w_i x_i\right) = sign(d - 1 + \alpha - \beta) = 1$

When all are False(-1), $\mathrm{sign}\left(\sum_{i=0}^{d} w_i x_i\right) = sign(-1) = -1$

9.

In the lecture 212 P.14 and P.15



specially (output layer) $(0 \le i \le d^{(L-1)})$

$$\frac{\partial e_n}{\partial w_{i1}^{(L)}} = \frac{\partial e_n}{\partial s_1^{(L)}} \cdot \frac{\partial s_1^{(L)}}{\partial w_{i1}^{(L)}} = -2\left(y_n - s_1^{(L)}\right) \cdot \left(x_i^{(L-1)}\right)$$

generally $(1 \le \ell < L)$ $(0 \le i \le d^{(\ell-1)}; 1 \le j \le d^{(\ell)})$

$$\frac{\partial e_n}{\partial w_{ij}^{(\ell)}} = \frac{\partial e_n}{\partial s_j^{(\ell)}} \cdot \frac{\partial s_j^{(\ell)}}{\partial w_{ij}^{(\ell)}} = \delta_j^{(\ell)} \cdot \left(x_i^{(\ell-1)}\right)$$

$$\frac{\partial \mathrm{e}}{\partial \mathrm{w}_{\mathrm{ij}}^{l}} = \delta_j^l x_i^{l-1} = x_i^{l-1} \sum_{k=1}^{d^{l+1}} \delta_k^{l+1}\ w_{jk}^{l+1}\ tanh'(s_j^l)$$

可知，當 $\mathrm{w}_{\mathrm{ij}}^{l}$ 皆為 0，可知對於以下條件的 gradient component 皆為 0

$$\frac{\partial \mathrm{e}}{\partial \mathrm{w}_{\mathrm{ij}}^{l}}\ ,(1 \le \mathrm{l} < \mathrm{L}); (0 \le \mathrm{i} \le \mathrm{d}^{\mathrm{l}-1}; 1 \le \mathrm{j} \le \mathrm{d}^{\mathrm{l}})$$

10.

$$\frac{\partial \mathrm{e}}{\partial \mathrm{s_k}} = \frac{\partial - \sum_{n=1}^{K} v_n \ln q_n}{\partial \mathrm{s_k}} = \frac{\partial - \sum_{n=1}^{K} v_n \ln q_n}{\partial \mathrm{s_k}} = \frac{-\partial \sum_{n=1}^{K} v_n \ln q_n}{\partial q_n} * \frac{\partial q_n}{\partial \mathrm{s_k}}$$

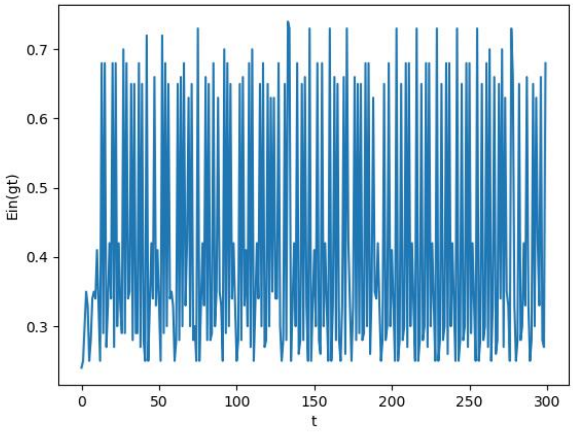
$$= \frac{-\sum_{n=1}^{K} v_n}{q_n} * \frac{\partial q_n}{\partial \mathrm{s_k}}$$

Only $n = k$ will make $v_n = 1$, which will stay. So we only discuss this condiction

$$\frac{\partial q_n}{\partial \mathrm{s_k}} = \frac{\partial \dfrac{\exp(s_k)}{\sum_{n=1}^{K} \exp(s_n)}}{\partial \mathrm{s_k}} = \frac{\exp(s_k) * \sum_{n=1}^{K} \exp(s_n) - \exp(s_k) * \exp(s_k)}{\left(\sum_{n=1}^{K} \exp(s_n)\right)^2} = q_k - q_k^2$$

$$\frac{\partial \mathrm{e}}{\partial \mathrm{s_k}} = \frac{-v_k}{q_k} * q_k - q_k^2 = q_k - v_k$$

11.

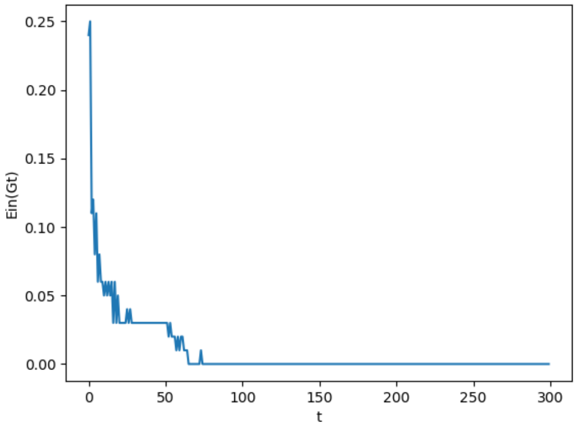


Ein(g1) = 0.240000, alpha_1 = 0.576340

12.

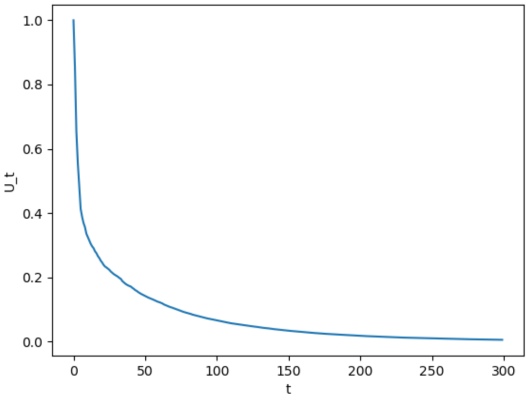
$\mathrm{E}_{\mathrm{in}}(g_t)$震盪，因為 AdaBoost 希望對於每個$g_t$在 u+1 時都表現很差，所以以新的權重讓其猜錯達到 $\frac{1}{2}$，因此圖會出現震盪。

13.



Ein(G) = 0.0

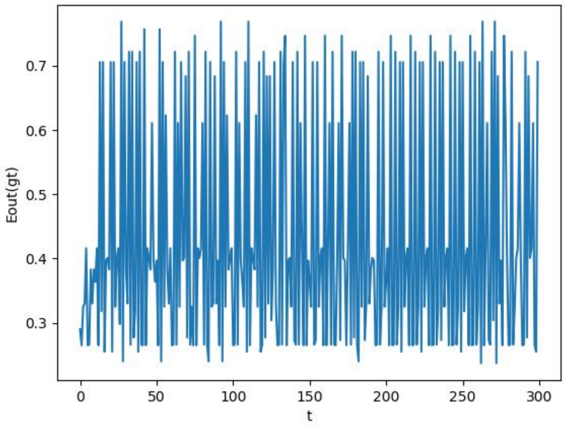
14.



U_2 = 0.854166

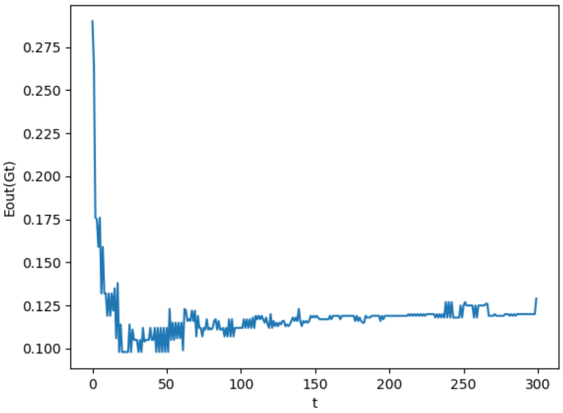
U_300 = 0.005465

15.



Eout(g1) = 0.29

16.



Eout(G) = 0.132

17.

$$proof\ : \mathrm{U}_1 = 1$$

In the Lecture 208 P.17, we set $\mathrm{u}^1 = [\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}]$, so $U_1 = \sum_{n=1}^{N} u_n^1 = N * \frac{1}{N} = 1$

$$proof\ : \mathrm{U}_{t+1} = U_t * 2\sqrt{\epsilon_t(1-\epsilon_t)} \leq U_t * 2\sqrt{\epsilon(1-\epsilon)}$$

$$\mathrm{U}_{t+1} = \sum_{n=1}^{N} u_n^t * \begin{cases} e^{-\alpha_t} \ \ if\ \ g_t(x_n) = y_n \\ e^{\alpha_t} \ \ if\ \ g_t(x_n) \neq y_n \end{cases}$$

$$= \sum_{n:g_t(x_n)=y_n} u_n^t * e^{-\alpha_t} + \sum_{n:g_t(x_n)\neq y_n} u_n^t * e^{\alpha_t}$$

$$= e^{-\alpha_t} * \sum_{n:g_t(x_n)=y_n} u_n^t + e^{\alpha_t} * \sum_{n:g_t(x_n)\neq y_n} u_n^t$$

By $\epsilon_\mathrm{t} = \frac{\sum_{n=1}^{N} u_n^t [[g_t(x_n) \neq y_n]]}{\sum_{n=1}^{N} u_n^t}$

$$\frac{\mathrm{U}_{t+1}}{\mathrm{U}_t} = e^{-\alpha_t} * (1-\epsilon_\mathrm{t}) + e^{\alpha_t} * \epsilon_\mathrm{t}$$

By $\alpha_t = ln\sqrt{\frac{(1-\epsilon_\mathrm{t})}{\epsilon_\mathrm{t}}}$

$$= e^{-ln\sqrt{\frac{(1-\epsilon_\mathrm{t})}{\epsilon_\mathrm{t}}}} * (1-\epsilon_\mathrm{t}) + e^{ln\sqrt{\frac{(1-\epsilon_\mathrm{t})}{\epsilon_\mathrm{t}}}} * \epsilon_\mathrm{t}$$

$$= 2\sqrt{\epsilon_\mathrm{t}(1-\epsilon_\mathrm{t})}$$

So, we can get that

$$\mathrm{U}_{t+1} = U_t * 2\sqrt{\epsilon_\mathrm{t}(1-\epsilon_\mathrm{t})}$$

And for the satisfaction

$$\epsilon_\mathrm{t} \leq \epsilon < \frac{1}{2}$$

We can know that

$$U_t * 2\sqrt{\epsilon_\mathrm{t}(1-\epsilon_\mathrm{t})} \leq U_t * 2\sqrt{\epsilon(1-\epsilon)}$$

Finally,

$$\mathrm{U}_{t+1} = U_t * 2\sqrt{\epsilon_\mathrm{t}(1-\epsilon_\mathrm{t})} \leq U_t * 2\sqrt{\epsilon(1-\epsilon)}$$

Reference:

http://www.academia.edu/8148348/A_Simple_Proof_of_AdaBoost_Algorithm

18.

Thanks to B05902013 吳宗翰

We know that

$$\mathrm{E_{in}}(G_T) \le U_{T+1} = U_T * 2\sqrt{\mathrm{\epsilon_T}(1-\mathrm{\epsilon_T})} = U_{T-1} * 2^2\sqrt{\mathrm{\epsilon_T}(1-\mathrm{\epsilon_T})}\sqrt{\mathrm{\epsilon_{T-1}}(1-\mathrm{\epsilon_{T-1}})}$$

$$= \cdots = U_1 2^T \prod_{t=1}^{T} \sqrt{\mathrm{\epsilon_t}(1-\mathrm{\epsilon_t})} \le U_1 2^T \left(\sqrt{\mathrm{\epsilon}(1-\mathrm{\epsilon})}\right)^T$$

$$\le U_1 2^T \left(\frac{1}{2}\exp\left(-2\left(\frac{1}{2}-\epsilon\right)^2\right)\right)^T = U_1\left(\exp\left(-2\left(\frac{1}{2}-\epsilon\right)^2\right)\right)^T$$

$$= \exp\left(-2T\left(\frac{1}{2}-\epsilon\right)^2\right)$$

It is trivial that if $\mathrm{E_{in}}(G_T) < \frac{1}{N}$, $\mathrm{E_{in}}(G_T)$ is 0.

$$\exp\left(-2T\left(\frac{1}{2}-\epsilon\right)^2\right) < \frac{1}{N}$$

$$-2T\left(\frac{1}{2}-\epsilon\right)^2 < -lnN$$

$$T > \frac{lnN}{2\left(\frac{1}{2}-\epsilon\right)^2}$$

I can claim that in $T = O(\log N)$, $\mathrm{E_{in}}(G_T) = \frac{1}{N}$, but for $\mathrm{E_{in}}(G_T) = 0$, it need

more than O(logN)