Home Work 3

Machine Learning Techniques

R04323050

經濟碩三　陳伯駒

1.

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

$$f.o.c. \quad \frac{\partial Gini}{\partial \mu_+} = -4\mu_+ + 2 = 0 \Rightarrow \mu_+^* = \frac{1}{2}$$

$$s.o.c. \quad \frac{\partial^2 Gini}{\partial \mu_+^2} = -4 < 0$$

因此，$\mu_+ = \frac{1}{2} = \mu_-$ 時、有最大值 $Gini = \dfrac{1}{2}$。

2.

By definition, the normalized Gini index in problem 1 would be:

$$\begin{aligned}\frac{1-\mu_+^2-\mu_-^2}{\frac{1}{2}} &= 2-2\mu_+^2-2\mu_-^2 = 2-2\mu_+^2-2(1-\mu)\\ &= 2-2\mu_+^2-2+4\mu_+-2\mu_+^2 = -4\mu_+^2+4\mu_+\\ &= 4\mu_+\cdot(1-\mu_+)\end{aligned}$$

(a). Normalized: $2\cdot min\left\{\mu_+,\mu_-\right\} = 4\mu_+(1-\mu_+)$(✗)

(b). 原式 $= \mu_+(2-\mu_+)^2+(1-\mu_+) = 4(1-\mu_+)(1-\mu_++\mu_+) = 4\mu_+(1-\mu_+)$ (✓)

(c). 原式 $= -\mu_+\cdot ln(\mu_+)-(1-\mu_+)\cdot ln(\mu_+)$(✗)

(d). 原式 $= 1-|\mu_+(1-\mu_+)| = 1-|2\mu_++1|$(✗)

3.

在 N 個樣本中、bootstrap 出 $p \cdot N$ 個、而每一個樣本沒有被取到的機率皆為 $1-\frac{1}{N}$。
$\therefore$ 對某一個樣本而言、bootstrap $N' = p \cdot N$ 次、都沒有被取到的機率為：

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

$\therefore$ Totally and approximately $N \cdot e^{-p}$ of the examples will not be sampled at all.

4.

在一個含有 K 個二元分類樹 $\{g_k\}_{k=1}^{K}$ 的隨機森林 G，若有一個點被隨機森林最終歸類錯誤、則至少平均有 $\frac{K+1}{2}$ 個分類樹將其分類錯誤，而每顆分類樹的錯誤次數皆為 e_k。因此總地而言、有 $\sum_{k=1}^{K} e_k$ 個錯誤。
所以在最極端的情況，存在 $\frac{\sum_{k=1}^{K} e_k}{\frac{1}{2}} = \frac{2}{K+1} \cdot \sum_{k=1}^{K} e_k$ 個錯誤點。$\therefore E_{out}(G) \leq \frac{2}{K+1} \cdot \sum_{k=1}^{K} e_k$

5.

已知 $g_1(x) = 2$，根據第 11 講投影片、p.17，optimal $\alpha_1 \to \eta$:

$$\min_{\eta} \frac{1}{N}\sum_{n=1}^{N}\left[(y_n - s_n^{(0)} - \eta g_1(\mathbf{x}_n))\right]^2 = \frac{1}{N}\sum_{n=1}^{N}(y_n - 2\eta)^2$$

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

$$\alpha_1 = \eta = \frac{1}{2N}\sum_{n=1}^{N} y_n \quad \therefore s_n = \alpha_1 \cdot g_1(\mathbf{x}_n) = \frac{1}{N}\sum_{n=1}^{N} y_n$$

6.

$$\underset{\eta}{min}\frac{1}{N}\sum_{n=1}^{N}\left[(y_n - s_n - \eta \cdot g_t(\mathbf{x}_n))\right]^2$$

$$\frac{\partial E}{\partial \eta} = 0 : \frac{1}{N}\sum_{n=1}^{N} 2 \cdot \left[y_n - s_n^{(t-1)} - \eta \cdot g_t(\mathbf{x}_n)\right]^2 \cdot (-g_t(\mathbf{x}_n)) = 0$$

$$\Rightarrow \alpha_t = \eta = \frac{\sum_{n=1}^{N} g_t(\mathbf{x}_n) \cdot (y_n - s_n^{(t-1)})}{\sum_{n=1}^{N} g_t^2(\mathbf{x}_n)}$$

$$\Rightarrow \alpha_t \cdot \sum_{n=1}^{N} g_t^2(\mathbf{x}_n) + \sum_{n=1}^{N} g_t(\mathbf{x}_n) \cdot s_n^{(t-1)} = \sum_{n=1}^{N} g_t(\mathbf{x}_n) y_n$$

$$\therefore \sum_{n=1}^{N} s_n^{(t)} = \sum_{n=1}^{N}(s_n^{(t-1)} + \alpha_t \cdot g_t(\mathbf{x}_n)) \cdot g_t(\mathbf{x}_n) = \sum_{n=1}^{N} g_t(\mathbf{x}_n) \cdot y_n$$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w_1} = 2(w_1 x_1 + w_0 - 2x_1 + x_1^2) \cdot x_1 + 2(w_1 x_2 + w_0 - 2x2 + x_2^2) \cdot x_2 = 0 \\ \frac{\partial \mathcal{L}}{\partial w_2} = 2(w_1 x_1 + w_0 - 2x_1 + x_1^2) + 2(w_1 x_2 + w_0 - 2x2 + x_2^2) = 0 \end{cases}$$

7.

By the slides in lecture 11, p.17: $\underset{\eta}{min}\sum_{n=1}^{N}(y_n - s_n - \eta \cdot g_t(\mathbf{x}_n))^2$ could be viewed as the one variable linear regression on $\{(g_t\text{-transformed inputs, residuals})\}$。

In the linear regression problem, we want to :

$$\begin{aligned} min \quad \sum_{n=1}^{N}\{residual\}^2 &= \sum_{n=1}^{N}(y_n - s_n)^2 \\ &= \sum_{n=1}^{N}(y_n - s_n - \eta g_t(\mathbf{x}_n))^2 \quad \text{if} \quad g_t(\mathbf{x}_n) = 0 \end{aligned}$$

$\therefore$ If we impose gradient boosting with optimal $g_t(\mathbf{x}_n) = 0$, which means η does not matter, gradient boosting is not appropriate for linear regression! The intuition is that although both GB and linear regression are attempting to solving the following problem:

$$\widehat{\beta} = \underset{\beta}{argmin}(y - \mathbf{x}\beta)^T(y - \mathbf{x}\beta)$$

Linear regression just observe that we can solve it directly by finding the solution to the linear equation:

$$\mathbf{x}^T\mathbf{x}\beta = \mathbf{x}^T y$$

This automatically gives us the best possible value of β out of all possibilities.

However, in GB, whether our weal classifier is a one variable or multi-variable regression, gives us a sequences of coefficients $\beta_1, \beta_2, \cdots$. The final model prediction will be the weighted form as the full linear regression:

$$\mathbf{x}\beta_1 + \mathbf{x}\beta_2 + \mathbf{x}\beta_3 + \cdots + \mathbf{x}\beta_n = \mathbf{x}(\beta_1 + \beta_2 + \beta_3 + \cdots + \beta_n)$$

Each of these steps is chosen to further decrease the sum of squared errors, but we could find the minimum possible sum of squares within this functional form by just performing a full regression model to begin with.

8.

OR：有一個對就對、全錯才算錯。

$\therefore$ Let $w_0 = d - 1$, $w_1 = w_2 = \cdots = w_d = 1$, then:

$$\begin{cases} x_1, x_2, \cdots, x_d \text{ 均為 } -1 \text{ 時，} g_A(\mathbf{x}) = sign(-1) = -1 \\ x_1, x_2, \cdots, x_d \text{ 至少有一個為 } -1 \text{ 時，} \sum_{i=1}^{d} w_i x_i \geq d - 1 + 1 - (d-1) \times 1 = 1 \Rightarrow g_A(\mathbf{x}) = +1 \end{cases}$$

9.

$$e_n = [y_n - NNet(\mathbf{x}_n)]^2 = (y_n - s_1^{(L)})^2 = \left[y_n - tanh(\mathbf{x}_1^{(L)})\right]^2 = \left[y_n - tanh\sum_{i=1}^{d^{(L-1)}} w_{il}^{(L)} x_i^{(L-1)}\right]^2$$

$$\frac{\partial e_n}{\partial w_{i1}^{(L)}} = -2(y_n - tanh(\mathbf{x}_1^{(L)})) \cdot tanh'(\mathbf{x}_1^L) \cdot (x_i^{L-1})$$

$$\frac{\partial e_n}{\partial w_{ij}^{(l)}} = \frac{\partial e_n}{\partial s_j^{(l)}} \cdot \frac{\partial s_j^{(l)}}{\partial w_{ij}^{(l)}} = \delta_j^{(l)} \cdot x_i^{(l-1)}, 1 \leq l < L$$

$$\delta_j^{(j)} = \sum_K \delta_K^{(l+1)} \cdot (w_{jK}^{(l+1)}) \cdot (tanh'(s_j^{(l)})), 1 \leq l < L$$

$$w_{ij}^{(l)} = 0, \text{由向前傳遞規則知：} x_i^{(l)} = 0 (l \geq 1, i > 0)$$

$$\begin{cases} 1 \leq l < L\text{時：} \delta_j^{(l)} = 0, \text{則梯度} \dfrac{\partial e_n}{\partial w_{ij}^{(j)}} = 0 \\ l = L\text{時：} \begin{cases} \text{若} i > 0\text{，則梯度} \dfrac{\partial e_n}{\partial w_{ij}^{(l)}} = 0 \\ \text{若} i = 0\text{，則注意} x_1^{(l)} = 0 \ \dfrac{\partial e_n}{\partial w_{ij}^{(l)}} = -2 y_n x_0^{(l-1)} \end{cases} \end{cases}$$

綜合以上：$y_n \neq 0$ 且 $x_0^{l-1} \Leftrightarrow \frac{\partial e_n}{\partial w_{01}^{(L)}} \neq 0$

10.

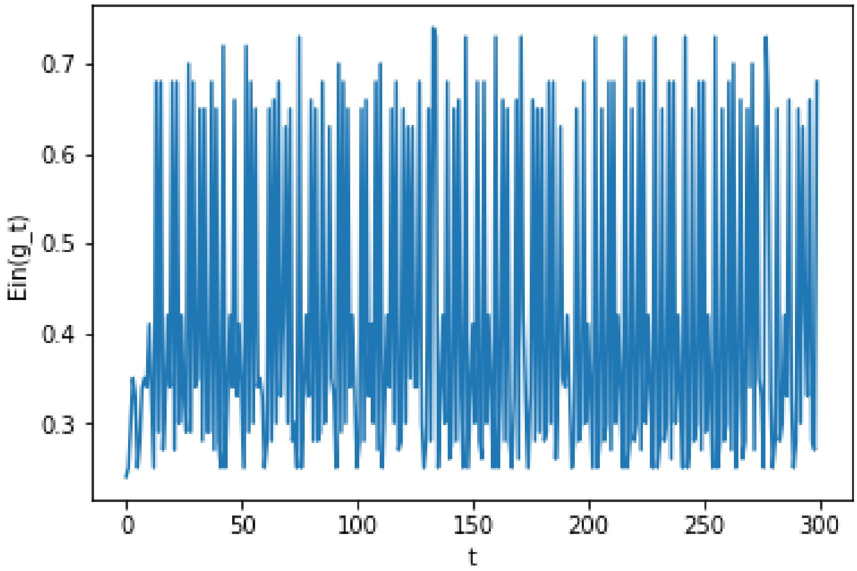
Denotes that $q_k = \dfrac{e^{s_k}}{\sum_{k=1}^{K} e^{s_k}}$, $E = -\sum_{k=1}^{K} v_k \, ln(q_k)$ and $\dfrac{\partial E}{\partial q_k} = \dfrac{-v_k}{q_k}$
Let index i denotes the class we want among all possible class K, then:

$$\frac{\partial q_k}{\partial s_i} = \begin{cases} \frac{e^{s_k}}{\sum_{k=1}^{K} e^{s_k}} - (\frac{e^{s_k}}{\sum_{k=1}^{K} e^{s_k}})^2 \ , \ i = k \\ -\dfrac{e^{s_k} \cdot s^{s_i}}{(\sum_{k=1}^{K} e^{s_k})^2} \ , \ i \neq k \end{cases} = \begin{cases} q_k \cdot (1 - q_k) \ , \ i = k \\ -q_i \cdot q_k \ , \ i \neq k \end{cases}$$

$$\therefore \frac{\partial E}{\partial s_k} = \sum_{k=1}^{K} \frac{\partial E}{\partial q_i} \cdot \frac{\partial q_i}{\partial s_k} = \frac{\partial E}{\partial q_k} \cdot \frac{\partial q_k}{\partial s_k} - \sum_{k \neq i} \frac{\partial E}{\partial q_i} \cdot \frac{\partial q_i}{s_k}$$
$$= v_k \cdot (1 - q_k) + \sum k \neq i v_i q_k = -v_k + q_k \cdot \sum_i v_i = q_k - v_k$$

11.

$E_{in}(g_1) = 0.24$, $\alpha_1 = 0.576$

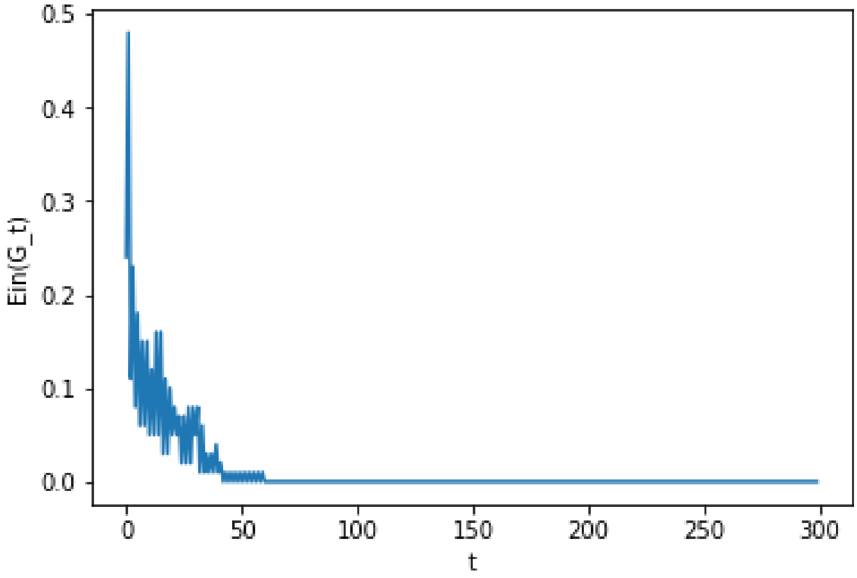


12.

由圖中可看出，$E_{in}(g_t)$ 隨著 t 並無某一特定規律。直觀上，因為 g_t 只是針對上一輪的錯誤樣本較敏感，而對整體 E_{in} 的下降並無特化。

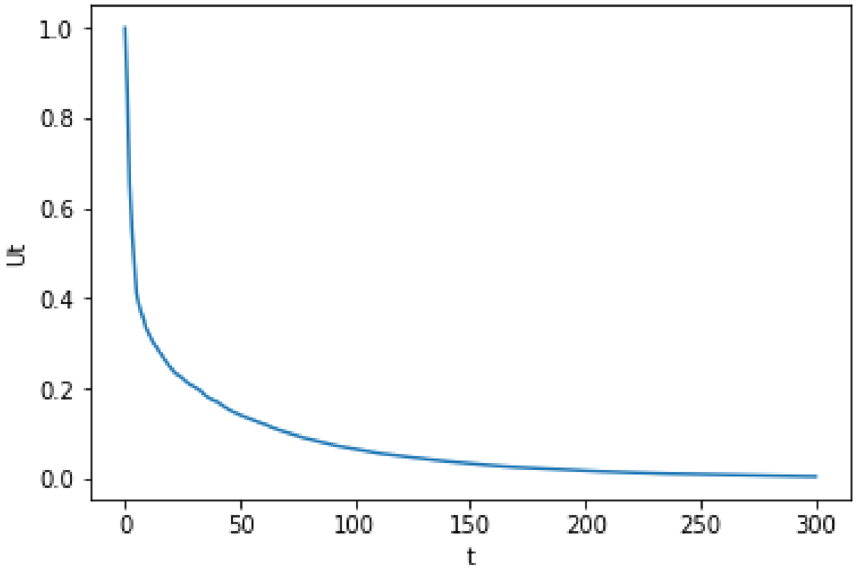
13.

$E_{in}(G_t) = 0$



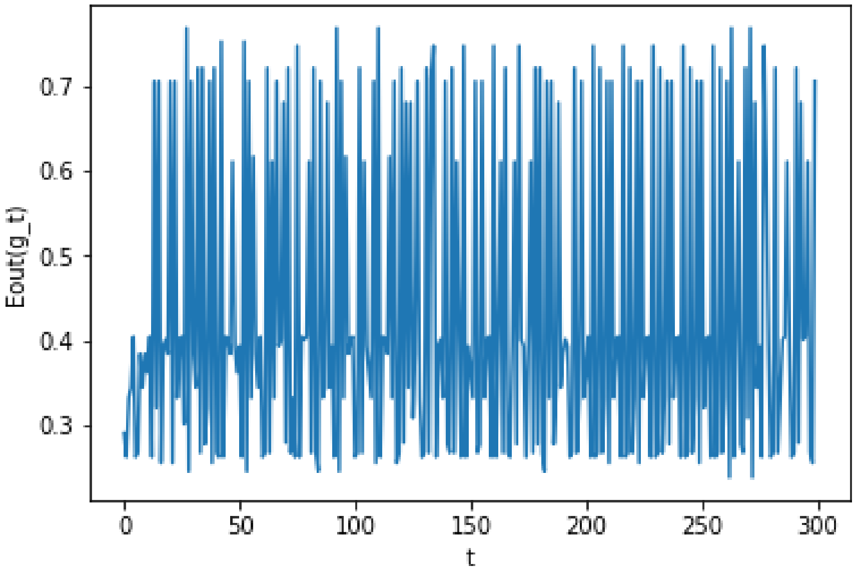
14.

$U_2 = 0.85416$, $U_T = 0.0054$



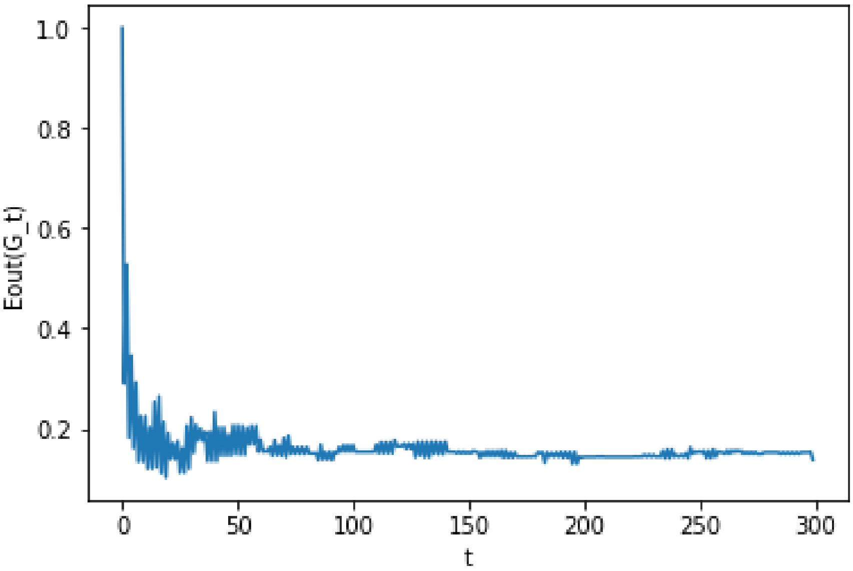
15.

$E_{out}(g_t) = 0.29$



16.

$E_{out}(G) = 0.138$



17.

$$U_{t+1} = \sum_{n=1}^{N} u_n^{(t+1)} = \sum_{n=1}^{N} u_n^{(t)} \cdot \blacklozenge_t \cdot \|y_n \neq g_t(\mathbf{x}_n)\| + \sum_{n=1}^{N} u_n^{(t)} / \blacklozenge_t \cdot \|y_n = g_t(\mathbf{x}_n)\|$$

$$(\text{where } \blacklozenge_t = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}})$$

$$= \epsilon_t \cdot \blacklozenge_t \cdot \sum_{n=1}^{N} u_n^{(t)} + (1-\epsilon_t)/\blacklozenge_t \cdot \sum_{n=1}^{N} u_n^{(t)}$$

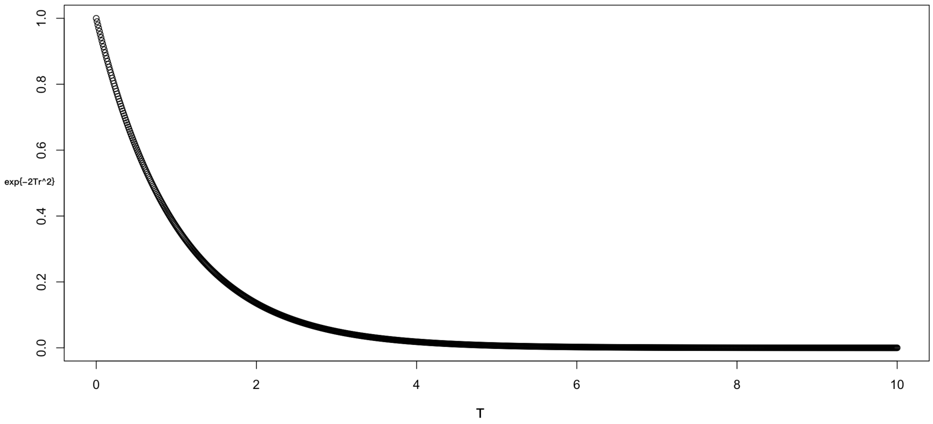
$$= U_t \cdot (\epsilon_t \cdot \blacklozenge_t + \frac{1-\epsilon_t}{\blacklozenge_t}) = 2\sqrt{\epsilon_t \cdot (1-\epsilon_t)} \cdot U_t \leq 2\sqrt{\epsilon \cdot (1-\epsilon)} \cdot U_t,$$

$$\forall\ \epsilon_t \leq \epsilon < \frac{1}{2}$$

18.

$$
\begin{aligned}
\widehat{E}_{ADA}^{T} &= \sum_{n=1}^{N} u_n^{(t)} \cdot \left[(1-\epsilon_t)e^{-n} + \epsilon_t e^{n}\right] \\
&= U_T \cdot \left[(1-\epsilon_t)e^{-n} + \epsilon_t e^{n}\right] \\
&\leq U_T \cdot 2\sqrt{\epsilon \cdot (1-\epsilon)} \cdot \left[(1-\epsilon_t)e^{-n} + \epsilon_t e^{n}\right] \\
&\leq U_T \cdot exp\left\{-2(\frac{1}{2}-\epsilon)^2\right\} \cdot \left[(1-\epsilon_t)e^{-n} + \epsilon_t e^{n}\right] \\
&= U_1 \cdot exp\left\{-2T(\frac{1}{2}-\epsilon)^2\right\} \cdot \left[(1-\epsilon_t)e^{-n} + \epsilon_t e^{n}\right] \\
&= exp\left\{-2T(\frac{1}{2}-\epsilon)^2\right\} \cdot \left[(1-\epsilon_t)e^{-n} + \epsilon_t e^{n}\right]
\end{aligned}
$$

令 $\gamma = \frac{1}{2} - \epsilon$，$\because \epsilon < \frac{1}{2} \quad \therefore \gamma > 0$，$exp\left\{-2T\gamma^2\right\}$ 之函數圖形如下：



According to above figure, we know the value of $exp\left\{-2T\gamma^2\right\}$ will diminish exponentially fast as T iterations. $\therefore$ after $O(log\ N)$ iterations, $E_{in}(G_T)$ goes to 0.