Machine Learning HW2

r07922100 資工碩一 楊力權

1.

$$\theta(s) = \frac{e^s}{e^s+1} = \frac{1}{e^{-s}+1} = 1 - \frac{1}{e^s+1} = 1 - \theta(-s)$$

$$F(A,B) = \frac{1}{N}\sum_{n=1}^{N} \ln(1+\exp(-y_n(A(w_{SVM}^T\phi(x_n)+b_{SVM})+B))) = \frac{1}{N}\sum_{n=1}^{N} \ln(1+\exp(-y_n(Az_n+B)))$$

$$= \frac{1}{N}\sum_{n=1}^{N} \ln(\frac{1}{\theta(y_n(Az_n+B))}) = \frac{-1}{N}\sum_{n=1}^{N} \ln(\theta(y_n(Az_n+B))) = \frac{-1}{N}\sum_{n=1}^{N} \ln(1-p_n)$$

$$\frac{\partial}{\partial A}F(A,B) = \frac{-1}{N}\sum(\frac{-1}{1-p_n})\frac{\partial p_n}{\partial A} = \frac{-1}{N}\sum(\frac{-1}{1-p_n})\frac{-(e^{y_n(Az_n+B)}y_nz_n)}{(1+e^{y_n(Az_n+B)})^2}$$

$$= \frac{1}{N}\sum y_nz_n(\frac{1}{1-p_n})(-1)\frac{e^{y_n(Az_n+B)}}{1+e^{y_n(Az_n+B)}}\frac{1}{1+e^{y_n(Az_n+B)}} = \frac{1}{N}\sum y_nz_n(\frac{1}{1-p_n})(-1)(1-p_n)(p_n)$$

$$= \frac{-1}{N}\sum_{n=1}^{N} y_nz_np_n$$

$$\frac{\partial}{\partial B}F(A,B) = \frac{-1}{N}\sum(\frac{-1}{1-p_n})\frac{\partial p_n}{\partial B} = \frac{-1}{N}\sum(\frac{-1}{1-p_n})\frac{-(e^{y_n(Az_n+B)}y_n)}{(1+e^{y_n(Az_n+B)})^2}$$

$$= \frac{-1}{N}\sum_{n=1}^{N} y_np_n$$

$$\therefore \nabla F(A,B) = (\frac{-1}{N}\sum_{n=1}^{N} y_nz_np_n\ ,\ \ \frac{-1}{N}\sum_{n=1}^{N} y_np_n)$$

2.

$$\frac{\partial}{\partial A}F(A,B) = \frac{-1}{N}\sum_{n=1}^{N} y_nz_np_n$$

$$\frac{\partial}{\partial B}F(A,B) = \frac{-1}{N}\sum_{n=1}^{N} y_np_n$$

Hessian Matrix要計算二次微分，已知$y_n^2 = 1$

$$\frac{\partial^2}{\partial^2 A}F(A,B) = \frac{-1}{N}\sum_{n=1}^{N} y_nz_n\frac{\partial}{\partial A}p_n = \frac{-1}{N}\sum_{n=1}^{N} y_nz_n(-1)(1-p_n)p_ny_nz_n = \frac{1}{N}\sum_{n=1}^{N} z_n^2(1-p_n)p_n$$

$$\frac{\partial^2}{\partial^2 B}F(A,B) = \frac{-1}{N}\sum_{n=1}^{N} y_n\frac{\partial}{\partial B}p_n = \frac{-1}{N}\sum_{n=1}^{N} y_n(-1)(1-p_n)p_ny_n = \frac{1}{N}\sum_{n=1}^{N} (1-p_n)p_n$$

$$\frac{\partial^2}{\partial A\partial B}F(A,B) = \frac{-1}{N}\sum_{n=1}^{N} y_nz_n\frac{\partial}{\partial B}p_n = \frac{-1}{N}\sum_{n=1}^{N} y_nz_n(-1)(1-p_n)p_ny_n = \frac{1}{N}\sum_{n=1}^{N} z_n(1-p_n)p_n$$

$$H(F)=\begin{bmatrix}\frac{\partial^2 F}{\partial^2 A} & \frac{\partial^2 F}{\partial A\partial B}\\ \frac{\partial^2 F}{\partial B\partial A} & \frac{\partial^2 F}{\partial^2 B}\end{bmatrix}=\begin{bmatrix}\frac{1}{N}\sum_{n=1}^{N} z_n^2(1-p_n)p_n & \frac{1}{N}\sum_{n=1}^{N} z_n(1-p_n)p_n\\ \frac{1}{N}\sum_{n=1}^{N} z_n(1-p_n)p_n & \frac{1}{N}\sum_{n=1}^{N}(1-p_n)p_n\end{bmatrix}$$

3.

因為Gaussian Kernel : $K(x,x')=\exp(-\gamma\|x-x'\|^2)$

所以在條件$\gamma\to\infty$下，$\begin{cases}K(x,x')=1 \text{ if } x=x'\\ K(x,x')=0 \text{ if } x\neq x'\end{cases}$

因此SVM的目標函數從$\min_{\alpha}\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m K(x_n,x_m)-\sum_{n=1}^{N}\alpha_n$變為下式子

$\min_{\alpha}\frac{1}{2}\sum_{n=1}^{N}\alpha_n^2-\sum_{n=1}^{N}\alpha_n=\min_{\alpha}\frac{1}{2}\sum_{n=1}^{N}(\alpha_n^2-2\alpha_n+1)-\frac{N}{2}$，因此$\alpha_n=1$能得到最小值。

而在$[y_n=1]=[y_n=-1]$與C>1的條件下，$\alpha_n=1$滿足$\sum_{n=1}^{N}y_n\alpha_n=0$與$0\le\alpha_n\le C$

因此$\alpha_n=1$是此SVM的最佳解。

4.

求mean square error最佳解使用pseudo inverse : $A^{\dagger}=(A^TA)^{-1}A^T$

$$\begin{bmatrix}x_1 & 1\\ x_2 & 1\end{bmatrix}\begin{bmatrix}w_1\\ w_0\end{bmatrix}=\begin{bmatrix}x_1-x_1^2\\ x_2-x_2^2\end{bmatrix}\Rightarrow\begin{bmatrix}w_1\\ w_0\end{bmatrix}=\begin{bmatrix}x_1 & 1\\ x_2 & 1\end{bmatrix}^{\dagger}\begin{bmatrix}x_1-x_1^2\\ x_2-x_2^2\end{bmatrix}=\frac{1}{(x_1-x_2)^2}\begin{bmatrix}x_1-x_2 & x_2-x_1\\ x_2^2-x_1x_2 & x_1^2-x_1x_2\end{bmatrix}\begin{bmatrix}x_1-x_1^2\\ x_2-x_2^2\end{bmatrix}$$

$$=\frac{1}{(x_1-x_2)^2}\begin{bmatrix}(x_1-x_2)^2(1-x_1-x_2)\\ x_1x_2(x_1-x_2)^2\end{bmatrix}=\begin{bmatrix}1-x_1-x_2\\ x_1x_2\end{bmatrix}$$

$w_1=1-x_1-x_2,\ w_0=x_1x_2$

因為input x_1,x_2的probability是[0,1]上的uniform distribution

[0,1] uniform相加期望值 $=\frac{1}{2}$[0,2] uniform期望值 $=1$

[0,1] uniform相乘期望值 $=\lim_{n\to\infty}\frac{1}{n^2}(\frac{1}{n}+\cdots\frac{n}{n})^2=\lim_{n\to\infty}\frac{1}{4}\frac{(n+1)^2}{n^2}=\frac{1}{4}$

因此$w_1=1-(x_1+x_2)=1-1=0,\ w_0=x_1x_2=\frac{1}{4},\ \bar{g}(x)=0x+\frac{1}{4}$

5.

pseudo data $(\tilde{x}_n,\tilde{y}_n)=(\sqrt{u_n}x_n,\sqrt{u_n}y_n)$，Linear regression minimizes

$$\min_{w}E_{in}(w)=\frac{1}{N}\sum_{n=1}^{N}(\tilde{y}_n-w^T\tilde{x}_n)^2=\frac{1}{N}\sum_{n=1}^{N}(\sqrt{u_n}y_n-w^T\sqrt{u_n}x_n)^2=\frac{1}{N}\sum_{n=1}^{N}u_n(y_n-w^Tx_n)^2$$

使用此組pseudo data可與做Adaptive boosting有一樣的optimize效果。

6.

$$u_+^{(1)} = u_-^{(1)} = \frac{1}{N},\ \ \epsilon = 0.22,\ \ \blacklozenge_t = \sqrt{\frac{1-\epsilon}{\epsilon}} \quad \therefore u_+^{(2)} = \frac{u_+^{(1)}}{\blacklozenge_t} = \frac{\frac{1}{N}}{\blacklozenge_t}, \therefore u_-^{(2)} = u_-^{(1)} \times \blacklozenge_t = \frac{1}{N} \times \blacklozenge_t$$

$$\frac{u_+^{(2)}}{u_-^{(2)}} = \frac{1}{\blacklozenge_t^2} = \frac{\epsilon}{1-\epsilon} = 0.282$$

7.

$(10 \times 2) \times 2 + 2 = 42$ different decision stumps

先看一個維度，M是-5~5的整數，因此-5~5中的空隙 θ 共10個，s調整sign$(x_i - \theta)$左右正負乘上2得到20，本題d=2因此有兩個維度各有20種decision stumps，得20 x 2 = 40，最後加上全負與全正的stump共2種得42(不論哪個維度取全正全負都屬於g(x)=+1與g(x)=-1兩種stump)。

8.

$\phi_{ds}(x) = (g_1(x), g_2(x), \cdots, g_{|G|}(x))$

$$K_{ds}(x, x') = (\phi_{ds}(x))^T \phi_{ds}(x') = \sum_{t=1}^{|G|} g_t(x) g_t(x') = \sum_{t=1}^{|G|} s_t^2 \cdot \text{sign}(x_{t\_i} - \theta_t)\text{sign}(x'_{t\_i} - \theta_t)$$

若$(x_{t_i} - \theta_t)$與$(x'_{t_i} - \theta_t)$同號則sign$(x_{t_i} - \theta_t)$sign$(x'_{t_i} - \theta_t) = 1$，

若異號則sign$(x_{t_i} - \theta_t)$sign$(x'_{t_i} - \theta_t) = -1$。

因此$K_{ds}(x, x') =$ (同號數量) — (異號數量) $= |G| - 2 \times$ (異號數量)

異號表示$(x_{t_i} < \theta_t$且$x'_{t_i} > \theta_t)$或$(x_{t_i} > \theta_t$且$x'_{t_i} < \theta_t)$，因此對一個維度 i 而言x_i, x'_i 有 $|x_i - x'_i|$個$\theta \Rightarrow 2|x_i - x'_i|$個decision stump g能使g(x)g(x')=-1 (異號)

異號數量 $= \sum_{i=1}^{d} 2|x_i - x'_i| = 2\|x - x'\|_1$

給定(d,M)：$K_{ds}(x, x') = |G| - 2 \cdot 2\|x - x'\|_1 = 2Md + 2 - 4\|x - x'\|_1$

9.

$\lambda = 50$得最低$E_{in} = 0.315$

10.

$\lambda = 0.05,\ 0.5,\ 5$得最低$E_{out} = 0.36$

11.

$\lambda = 5$得最低$E_{in} = 0.315$

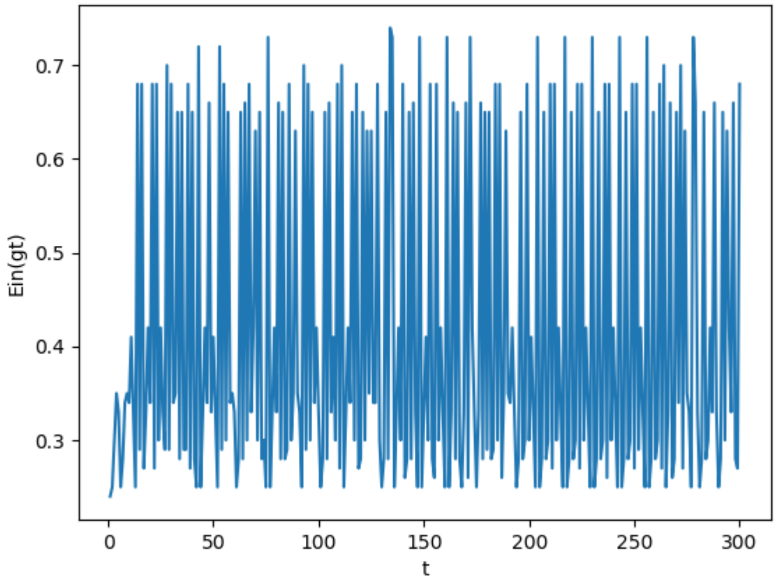
跟沒有做bagging效果一樣

12.

$\lambda = 0.5$得最低$E_{out} = 0.36$

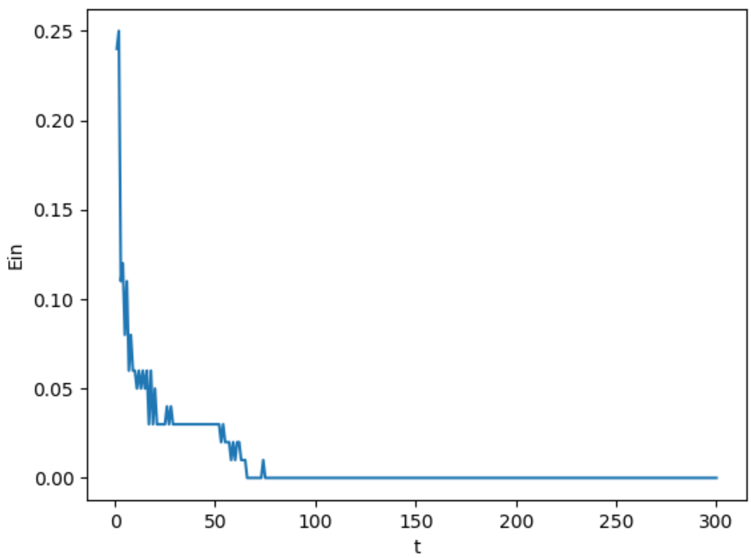
跟沒有做bagging效果一樣

13.



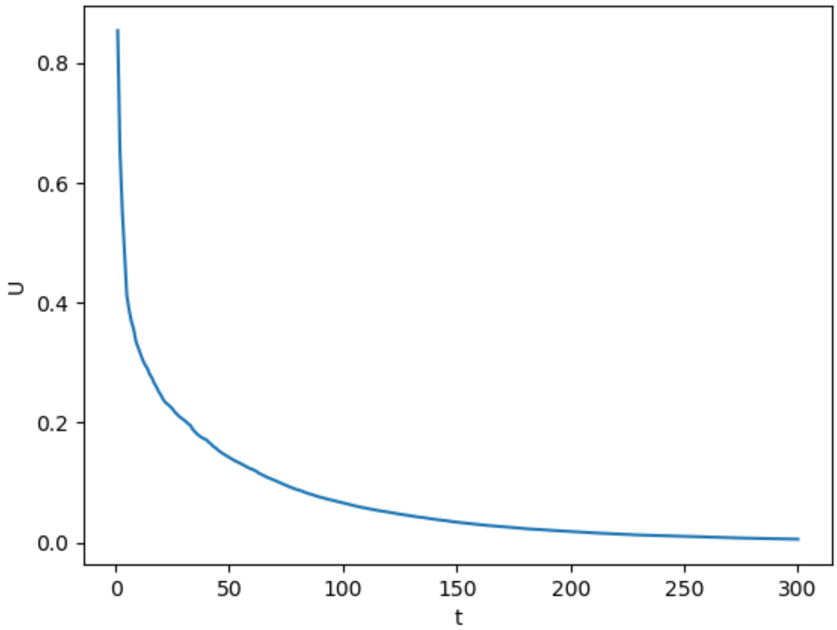
E_{in}在每個時間點都不小，從0.1~0.7都有，$E_{in}(g_T) = 0.68$，因為每個gt都是一個單一維度弱弱的stump，因此無法僅憑一己之力區分出多個維度表現的資料。

14.



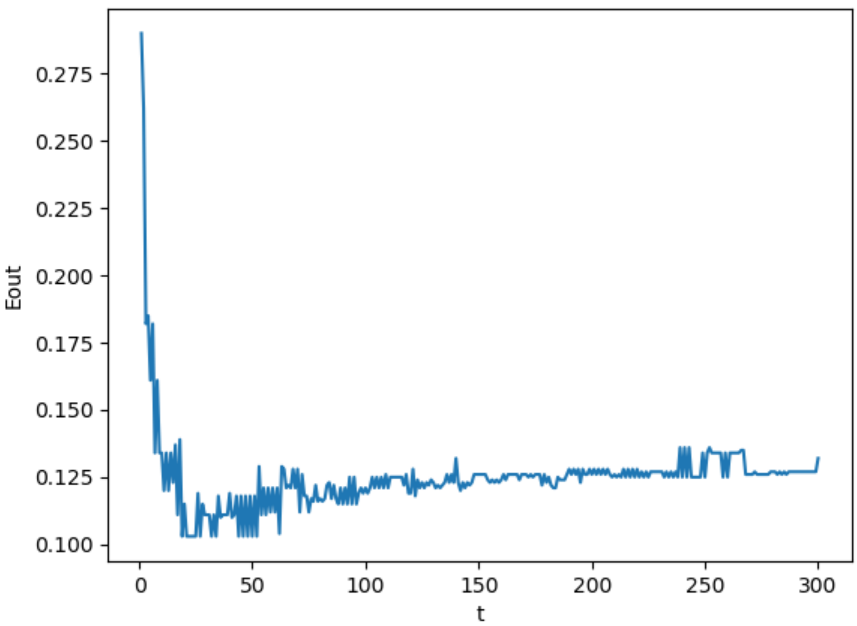
E_{in}隨t遞減，$E_{in}(G_T) = 0$，adaboost透過切分維度空間可以完美分開training set。

15.



U隨著時間變小，$U_T = 0.0054$。是因為adaboost使用$\blacklozenge_t = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$，會有以下關係式(17.會推導)：$U_{t+1} = U_t \cdot 2\sqrt{\epsilon_t(1-\epsilon_t)}$而$2\sqrt{\epsilon_t(1-\epsilon_t)}$為0~1的值，因此U只會遞減。

16.



一開始陡降之後趨於平緩，$E_{out}(G_T) = 0.132$，沒辦法跟E_{in}一樣變成0可能是因為太多次的Adaboost其實還是會overfit在training set上。

17.

Adaboost一開始令$u_n^{(1)} = \frac{1}{N}$ 所以 $U_1 = \sum_{n=1}^{N} u_n^{(1)} = N \times \frac{1}{N} = 1$

首要條件為adaboost的性質：

$$\alpha_t = \ln\sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \Rightarrow \exp(\alpha_t) = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$$

$$U_t = \sum_{n=1}^{N} u_n^{(t)}$$

$$\epsilon_t = \frac{\sum_{n=1}^{N} u_n^{(t)}[y_n \neq g_t(x)]}{\sum_{n=1}^{N} u_n^{(t)}}，1-\epsilon_t = \frac{\sum_{n=1}^{N} u_n^{(t)}[y_n = g_t(x)]}{\sum_{n=1}^{N} u_n^{(t)}}$$

$$\blacklozenge_t = \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}，[y_n \neq g_t(x)] : u_n^{(t+1)} = u_n^{(t)} \cdot \blacklozenge_t，[y_n = g_t(x)] : u_n^{(t+1)} = u_n^{(t)} / \blacklozenge_t$$

開始推導：

$$U_{t+1} = \sum_{n=1}^{N} u_n^{(t+1)} = \sum_{y_n = g_t(x_n)} u_n^{(t)} \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \sum_{y_n \neq g_t(x_n)} u_n^{(t)} \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$$

$$= \sum_{n=1}^{N} u_n^{(t)} \frac{\sum_{y_n = g_t(x_n)} u_n^{(t)} \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \sum_{y_n \neq g_t(x_n)} u_n^{(t)} \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}}{\sum_{n=1}^{N} u_n^{(t)}}$$

$$= U_t \cdot \frac{\sum_{y_n = g_t(x_n)} u_n^{(t)} \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \sum_{y_n \neq g_t(x_n)} u_n^{(t)} \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}}{\sum_{n=1}^{N} u_n^{(t)}} = U_t \cdot ((1-\epsilon_t)\sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \epsilon_t\sqrt{\frac{1-\epsilon_t}{\epsilon_t}})$$

$$= U_t \cdot (\sqrt{\epsilon_t(1-\epsilon_t)} + \sqrt{(1-\epsilon_t)\epsilon_t}) = U_t \cdot 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

$\frac{\partial}{\partial \epsilon}\sqrt{\epsilon(1-\epsilon)} = \frac{1-2\epsilon}{2\sqrt{\epsilon(1-\epsilon)}}$ 在 $\epsilon_t \leq \epsilon < \frac{1}{2}$ 的條件下恆正，表示$\sqrt{\epsilon(1-\epsilon)}$絕對遞增。

因此$U_{t+1} = U_t \cdot 2\sqrt{\epsilon_t(1-\epsilon_t)} \leq U_t \cdot 2\sqrt{\epsilon(1-\epsilon)}$

18.

$$E_{in}(G_T) \leq U_{T+1} \leq U_T \cdot 2\sqrt{\epsilon(1-\epsilon)} \leq U_1 \cdot (2\sqrt{\epsilon(1-\epsilon)})^T \leq e^{(-2(\frac{1}{2}-\epsilon)^2)T}$$

使$E_{in} = 0$必須讓E_{in}至少小於$\frac{1}{N} \Rightarrow E_{in}(G_T) \leq e^{(-2(\frac{1}{2}-\epsilon)^2)T} < \frac{1}{N}$

兩邊取ln $\Rightarrow 2T(\frac{1}{2}-\epsilon)^2 > \ln N \ \therefore T = O(\log N)$