

NTUEE, WAN - CYUAN FAN 范萬泉

“no collaborators, with reference”

Homework 1

Problem 1 :

ML tech. HW2 Bo450205 范萬泉

Descent Methods for Probabilistic SVM

- $\min_{A, B} F(A, B) = \frac{1}{N} \sum_{n=1}^N \ln(1 + \exp(-y_n(A \cdot (w_{SVM}^T \phi(x_n) + b_{SVM}) + B)))$
 Let $Z_n = w_{SVM}^T \phi(x_n) + b_{SVM} \rightarrow p_n = \Theta(-y_n(AZ_n + B))$, $\Theta(s) = \frac{\exp(s)}{1 + \exp(s)}$

$$\Rightarrow \min_{A, B} F(A, B) = \frac{1}{N} \sum_{n=1}^N \ln(1 + \exp(-y_n(AZ_n + B)))$$

$$\frac{\partial F(A, B)}{\partial A} = \frac{1}{N} \sum_{n=1}^N \frac{-y_n Z_n \exp(-y_n(AZ_n + B))}{1 + \exp(-y_n(AZ_n + B))} = \frac{1}{N} \sum_{n=1}^N -y_n Z_n P_n$$

$$\frac{\partial F(A, B)}{\partial B} = \frac{1}{N} \sum_{n=1}^N \frac{-y_n \exp(-y_n(AZ_n + B))}{1 + \exp(-y_n(AZ_n + B))} = \frac{1}{N} \sum_{n=1}^N -y_n P_n$$

$$\nabla F(A, B) = \frac{1}{N} \sum_{n=1}^N P_n y_n (-Z_n \hat{e}_A + \hat{e}_B)$$

Problem 2 :

2. $H(F(A, B)) = \begin{bmatrix} \frac{\partial^2 F}{\partial A^2} & \frac{\partial^2 F}{\partial A \partial B} \\ \frac{\partial^2 F}{\partial B \partial A} & \frac{\partial^2 F}{\partial B^2} \end{bmatrix}$, Let $P_n = \theta(S_n)$, $\theta(S_n) = \frac{\exp(S_n)}{1 + \exp(S_n)}$
 $S_n = -y_n(Az_n + B)$

$$\frac{\partial P_n}{\partial A} = \frac{\partial P_n}{\partial S_n} \cdot \frac{\partial S_n}{\partial A} = \frac{\exp(S_n)}{(1 + \exp(S_n))^2} \cdot (-y_n z_n) = P_n(1 - P_n)(-y_n z_n)$$

$$\frac{\partial P_n}{\partial B} = \frac{\partial P_n}{\partial S_n} \cdot \frac{\partial S_n}{\partial B} = \frac{\exp(S_n)}{(1 + \exp(S_n))^2} \cdot (-y_n) = P_n(1 - P_n)(-y_n)$$

$$\frac{\partial^2 F}{\partial A^2} = \frac{\partial}{\partial A} \left(\frac{1}{N} \sum_{n=1}^N -y_n z_n P_n \right) = \frac{1}{N} \sum_{n=1}^N -y_n z_n \cdot \frac{\partial P_n}{\partial A} = \frac{1}{N} \sum_{n=1}^N P_n(1 - P_n)(y_n z_n)^2$$

$$\frac{\partial^2 F}{\partial B^2} = \frac{\partial}{\partial B} \left(\frac{1}{N} \sum_{n=1}^N -y_n P_n \right) = \frac{1}{N} \sum_{n=1}^N -y_n \frac{\partial P_n}{\partial B} = \frac{1}{N} \sum_{n=1}^N P_n(1 - P_n)(y_n)^2$$

$$\frac{\partial^2 F}{\partial A \partial B} = \frac{\partial}{\partial B} \left(\frac{1}{N} \sum_{n=1}^N -y_n z_n P_n \right) = \frac{1}{N} \sum_{n=1}^N -y_n z_n \frac{\partial P_n}{\partial B} = \frac{1}{N} \sum_{n=1}^N P_n(1 - P_n)y_n^2 z_n$$

$$\frac{\partial^2 F}{\partial B \partial A} = \frac{\partial}{\partial A} \left(\frac{1}{N} \sum_{n=1}^N -y_n P_n \right) = \frac{1}{N} \sum_{n=1}^N -y_n \frac{\partial P_n}{\partial A} = \frac{1}{N} \sum_{n=1}^N P_n(1 - P_n)y_n^2 z_n.$$

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

Problem 3 :

Extreme kernel & overfitting.

3. Gaussian kernel : $K(X, X') = \exp(-\gamma \|X - X'\|^2)$

If $\gamma \rightarrow \infty$, $K(X, X') = I$ ($\because X = X'$ 時, 对角线為 I)

Soft-SVM

$$\min_{\alpha} \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m y_n y_m (Z_n^T Z_m) - \sum_{n=1}^N \alpha_n$$

$\left(\begin{array}{l} \text{微} \\ \text{分} \end{array} \right) \quad \frac{\partial}{\partial \alpha} \left(\begin{array}{l} \min_{\alpha} \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m y_n y_m (Z_n^T Z_m) - \sum_{n=1}^N \alpha_n \\ \text{微} \end{array} \right) = 0$

$\Rightarrow \sum_{n=1}^N (\alpha_n - 1) = 0$ 則 $\min_{\alpha} \sum_{n=1}^N \alpha_n = 1 \rightarrow \alpha = 1$ vector.

$y_1 x_1 + y_2 x_2 + \dots + y_N x_N = 1$

Problem 4 :

Blending.

4. $f(x) = x - x^2$, i/p: uniform $[0, 1]$
 $(x_1, x_1 - x_1^2), (x_2, x_2 - x_2^2)$
 $h(x) = w_1 x + w_0$ 利用多項式法比較係數

$$\begin{aligned} &= \frac{(x_1 - x_1^2) - (x_2 - x_2^2)}{x_1 - x_2} x - \frac{(x_1 - x_1^2) - (x_2 - x_2^2)}{x_1 - x_2} x_1 + x_1 - x_1^2 \\ &= (1 - (x_1 + x_2)) x - (1 - (x_1 + x_2)) x_1 + x_1 - x_1^2 \\ &= (1 - x_1 - x_2) x + x_1 x_2. \end{aligned}$$

$$\begin{aligned} \bar{g}(x) &= ((-E[x_1] - E[x_2])x + E[x_1 x_2]) \\ &= (1 - \frac{1}{2} - \frac{1}{2})x + \frac{1}{4} = \frac{1}{4} \end{aligned}$$

Problem 5 :

Boosting.

5. $\min_w E_{in}^u(w) = \frac{1}{N} \sum_{n=1}^N \mu_n (y_n - w^T x_n)^2$

usual E_{in} of linear regression: $\min_w E_{in}(w) = \frac{1}{N} \sum_{n=1}^N (y_n - w^T x_n)^2$

將 $E_{in}^u(w)$ 中的 μ_n 放入平方中。

$\min_w E_{in}^u(w) = \frac{1}{N} \sum_{n=1}^N (\sqrt{\mu_n} y_n - w^T \sqrt{\mu_n} x_n)^2$

令 $\{(\tilde{x}_n, \tilde{y}_n)\}_{n=1}^N = \{(\sqrt{\mu_n} x_n, \sqrt{\mu_n} y_n)\}_{n=1}^N$

$$(\tilde{x}_n, \tilde{y}_n) = (\sqrt{\mu_n} x_n, \sqrt{\mu_n} y_n)$$

$\min_w E_{in}(w) = \frac{1}{N} \sum_{n=1}^N (\tilde{y}_n - w^T \tilde{x}_n)^2$

$$= \frac{1}{N} \sum_{n=1}^N (\sqrt{\mu_n} y_n - w^T \sqrt{\mu_n} x_n)^2$$

$$= \frac{1}{N} \sum_{n=1}^N \mu_n (y_n - w^T x_n)^2 = \min_w E_{in}^u(w)$$

Problem 6 :

6. positive : 78% negative : 22%

$$\epsilon_t = 22\% \quad , \quad \Delta_t = \sqrt{1 - \epsilon_t} = \sqrt{0.78} = 0.88$$

$$U_+^{(2)} = U_+^{(1)} / \Delta_t \quad , \quad U_-^{(2)} = U_-^{(1)} \cdot \Delta_t$$

$$U_+^{(2)} / U_-^{(2)} = \frac{\Delta_t}{1 - \Delta_t} \quad \text{want } g_1 \text{ best for } \bar{\epsilon}_{in} \quad U_n^{(1)} = \frac{1}{N}$$

$$= \frac{1}{0.88^2} = \frac{0.22}{0.78} = 0.28205\%$$

Problem 7 :

kernel for Decision stumps.

7. $d=2, M=5 \quad i \in \{1, 2\}, s \in \{-1, 1\}$

$$g_{s,i,\theta}(x) = s \cdot \text{sign}(x_i - \theta) \quad i/p \in [-5, 5] \text{ (z)}$$

input vector : $\{-5, -4, -3, -2, \dots, 0, 1, \dots, 4.5\}$

input space : $\{1, 2\}$

有 $2 + 10 \times 2 \times 2 = 42$ 種 $g_{s,i,\theta}(x)$

中間日数
 $\theta < -M, \theta > M$
只有2種全+/-

Problem 8 :

8.

$$\begin{aligned}
 K_{ds}(x, x') &= \phi_{ds}(x)^T \phi_{ds}(x') = \left(\sum_{t=1}^{|G|} g_t(x) \cdot g_t(x') \right) \\
 &= \sum_{t=1}^{|G|} \cancel{g_t} \cdot \text{sign}(x_{it} - \theta_t) \cdot \cancel{g_t} \cdot \text{sign}(x'_{it} - \theta_t) = 2 \sum_{t=1}^{|G|} \text{sign}(x_{it} - \theta_t) \cdot \text{sign}(x'_{it} - \theta_t)
 \end{aligned}$$

① case 1: $\max(x_{it}, x'_{it}) > \theta_t, \theta_t < \min(x_{it}, x'_{it})$ ② case 2: $\min(x_{it}, x'_{it}) < \theta_t \leq \max(x_{it}, x'_{it})$

$\text{sign}(x_{it} - \theta_t) = \text{sign}(x'_{it} - \theta_t)$ 則 $\text{sign}(x_{it} - \theta_t) \cdot \text{sign}(x'_{it} - \theta_t) = 1$
 \Rightarrow 相乘即為 1.

此 θ_t 組合在 $|G|$ 中有 $|G| - 2 \sum_{i=1}^d |x_i - x'_i|$ 此 θ_t 組合有 $2 \sum_{i=1}^d |x_i - x'_i|$ 種.

$$\begin{aligned}
 \Rightarrow |G| - 2 \sum_{i=1}^d |x_i - x'_i| &\quad \Rightarrow -2 \sum_{i=1}^d |x_i - x'_i| \\
 |G| - 4 \sum_{i=1}^d |x_i - x'_i| &
 \end{aligned}$$

Problem 9 ,10 ,11 ,12 :

```
Problem 9,10:  
lambda = 0.050000 ,Ein = 0.3175 ,Eout = 0.3600  
lambda = 0.500000 ,Ein = 0.3175 ,Eout = 0.3600  
lambda = 5.000000 ,Ein = 0.3200 ,Eout = 0.3600  
lambda = 50.000000 ,Ein = 0.3150 ,Eout = 0.4000  
lambda = 500.000000 ,Ein = 0.3300 ,Eout = 0.3700  
Problem 11,12:  
lambda = 0.050000 ,Ein = 0.3200 ,Eout = 0.3700  
lambda = 0.500000 ,Ein = 0.3200 ,Eout = 0.3700  
lambda = 5.000000 ,Ein = 0.3150 ,Eout = 0.3900  
lambda = 50.000000 ,Ein = 0.3150 ,Eout = 0.3900  
lambda = 500.000000 ,Ein = 0.3225 ,Eout = 0.3700
```

9. Lambda = 50 的時候有最小值 $Ein(g) = 0.3150$ 。

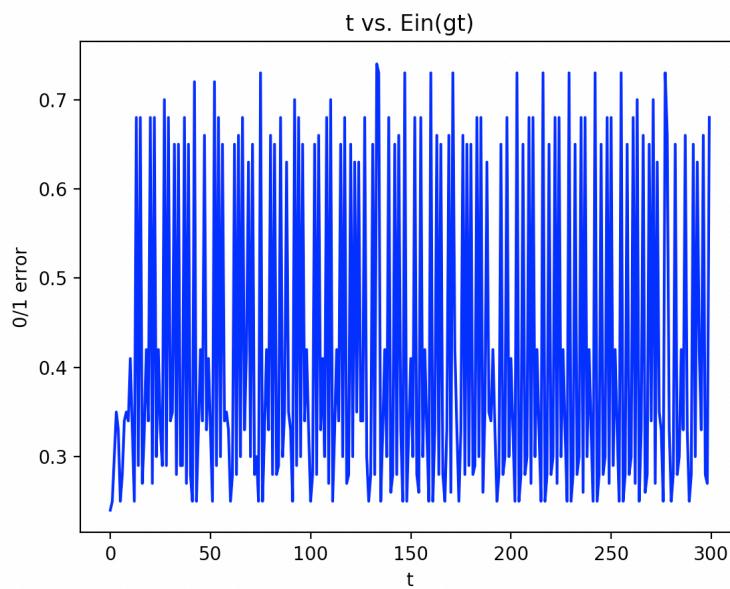
10. Lambda = 0.05 , 0.5 , 5 三個數字的時候有最小值 $Eout(g) = 0.3600$ 。

11. Lambda = 5 , 50 的時候有最小值 $Ein(G) = 0.3150$ 。

經過bagging後，對於 λ 在5,50者，確實有使得 Ein 下降的情況，但基本上都不是下降很多，表示多個g結合後計算的分隔線，實際上可能和原本在第9題一個g形成的linear 差不多。

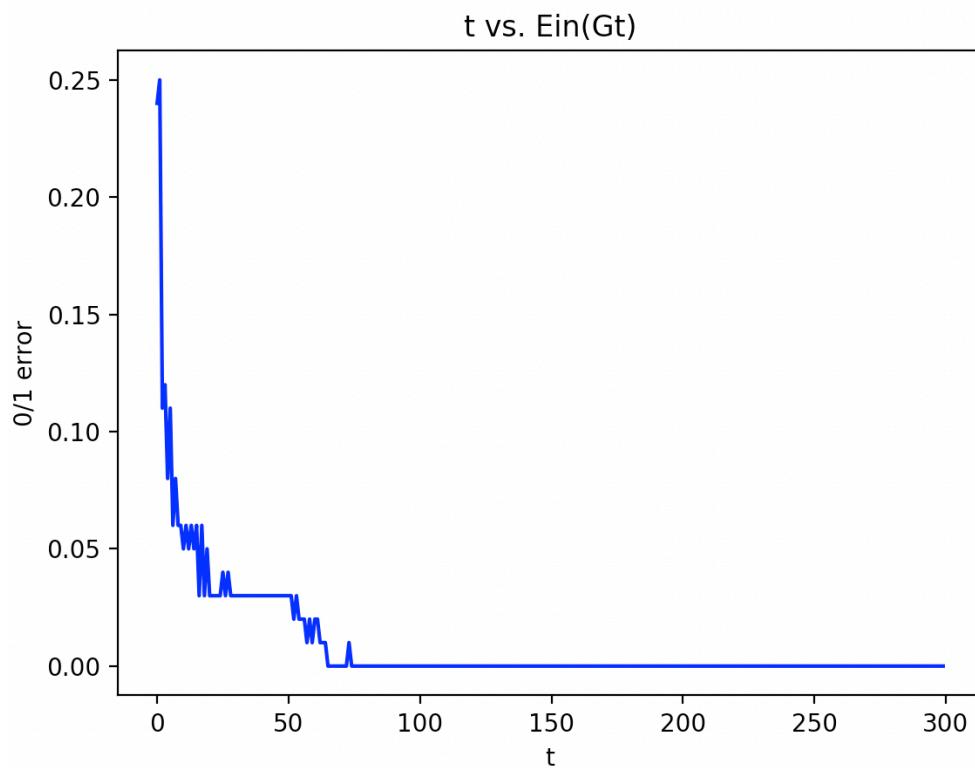
12. Lambda = 0.05 , 0.5 , 500 三個數字的時候有最小值 $Eout(g) = 0.3700$ 。

經過bagging，我們發現 $Eout(G)$ 反而變大，相對於前者0.36上升0.01，表示多個結合的G形成之model，反而有點overfitting在我們的training data上。

Problem 13 :

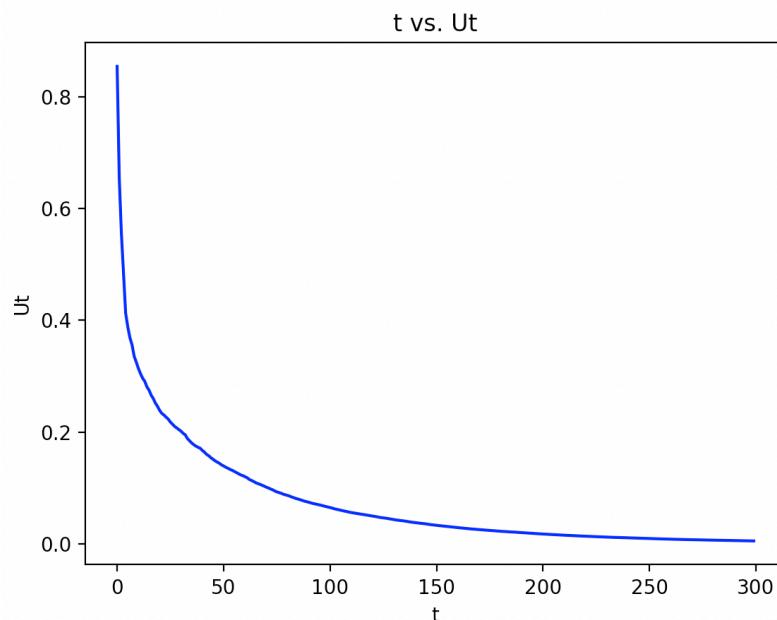
$$E_{in}(g_T) = 0.68$$

我們發現 E_{in} 值基本上時好時壞，並沒有一個穩定的遞減/增方向，因為每一次 $g(t)$ 決定都是找所有可能切法中最好的，不會有累積前面所有 $g(t)$ 的效果。

Problem 14 :

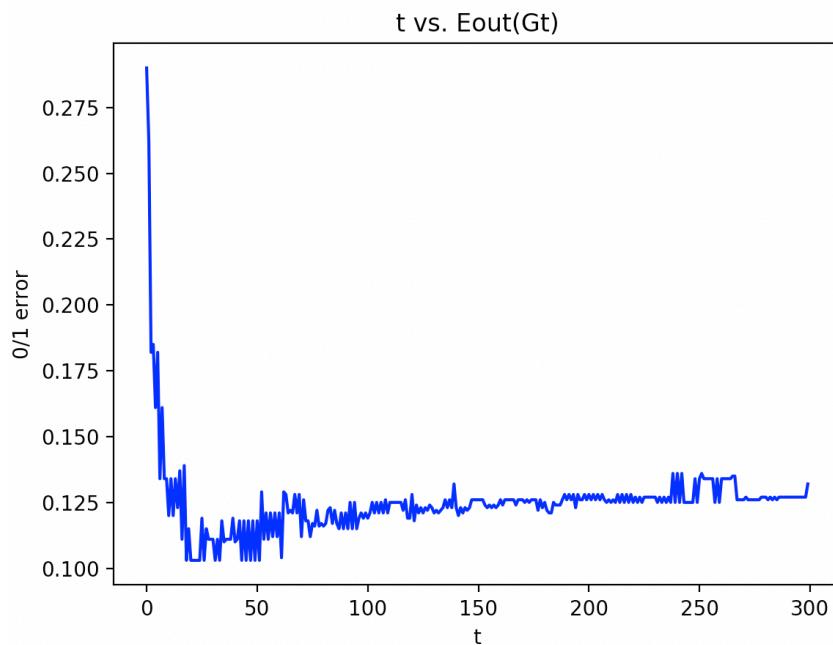
$$E_{in}(G_T) = 0.0$$

這題我們將前面的 $g(t)$ 累積計算求出對應的 $G(t)$ ，從實驗結果看出來， E_{in} 數值不斷下降，直到完全沒有錯誤。這是因為我們累計了前面所有的 $g(t)$ 因此就像講義上寫，會有更細的非線性切割來符合模型。

Problem 15 :

$$U_T : 0.005401486582613612$$

這裡的大U是將前面的u總和在每一個iteration記錄一次，我們知道此演算法會每次修正u值，如果正確的分類，我們把該data對應到的u值變小，因此合理我們每次 $g(t)$ 越來越多結合成 G_t 時，分類效果會越來越好，也就導致對的分類越來越多，u值也隨之不斷下降。

Problem 16:

$$E_{out}(G_T) = 0.132$$

這階段我們做testing，也看出來整體error數值不斷下降，大概在 $t=40$ 左右有最小值出現，在那之後又持續緩慢的上升，推測應該是model overfitting 在training data上頭，使得分類效果對於test data沒有那麼佳。

Problem 17 :

Power of Adaptive
Boosting.

$$\begin{aligned}
 U_{t+1} &= \sum_{n=1}^N U_n^{(t+1)} = \left\{ \begin{array}{l} \sum_{n=1}^N U_n^{(t)} \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}, y_n \neq g_t(x_n) \\ \sum_{n=1}^N U_n^{(t)} / \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}, y_n = g_t(x_n) \end{array} \right. \\
 &= \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \sum_{n=1}^N U_n^{(t)} [y_n \neq g_t(x_n)] + \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} \sum_{n=1}^N U_n^{(t)} [y_n = g_t(x_n)] \\
 &= \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \frac{\sum_{n=1}^N U_n^{(t)} [y_n \neq g_t(x_n)]}{\sum_{n=1}^N U_n^{(t)}} \cdot \sum_{n=1}^N U_n^{(t)} + \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} \frac{\sum_{n=1}^N U_n^{(t)} [y_n = g_t(x_n)]}{\sum_{n=1}^N U_n^{(t)}} \cdot \sum_{n=1}^N U_n^{(t)} \\
 &= \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \epsilon_t \left[\sum_{n=1}^N U_n^{(t)} \right] + \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} (1-\epsilon_t) \left[\sum_{n=1}^N U_n^{(t)} \right] \\
 &= 2\sqrt{\epsilon_t(1-\epsilon_t)} \sum_{n=1}^N U_n^{(t)} = 2\sqrt{\epsilon_t(1-\epsilon_t)} U_t
 \end{aligned}$$

$\because \epsilon_t \leq \epsilon < \frac{1}{2} \rightarrow \epsilon_t(1-\epsilon_t) = \epsilon_t - \epsilon_t^2 = -(\epsilon_t - \frac{1}{2})^2 + \frac{1}{4} \Rightarrow \epsilon_t = \frac{1}{2}$ 有 Max
 $\because \epsilon_t(1-\epsilon_t)$ 在 $\epsilon_t < \frac{1}{2}$ 区間遞減, $\therefore \epsilon_t - \epsilon_t^2 \leq \epsilon - \epsilon^2 = \epsilon(1-\epsilon)$
 $U_{t+1} = U_t \cdot 2\sqrt{\epsilon_t(1-\epsilon_t)} \leq U_t \cdot 2\sqrt{\epsilon(1-\epsilon)}$

$$U_1 = \sum_{n=1}^N U_n^{(1)} = \sum_{n=1}^N \frac{1}{N} = 1$$

Problem 18 :

$$\begin{aligned}
 18. \quad E_{in}(G_T) &\leq U_{T+1} = U_T \cdot \sqrt{\epsilon_T(1-\epsilon_T)} \leq U_T \exp(-2(\frac{1}{2}-\epsilon)^2) \\
 &\leq U_{T-1} \cdot \sqrt{\epsilon_{T-1}(1-\epsilon_{T-1})} \cdot \exp(-2(\frac{1}{2}-\epsilon)^2) \leq U_{T-1} \exp(-2(\frac{1}{2}-\epsilon)^2) \\
 &\leq 1 \cdot \exp(-2(\frac{1}{2}-\epsilon)^2)^T \\
 &\because \exp(-2(\frac{1}{2}-\epsilon)^2) \quad , \quad \frac{1}{2}-\epsilon > 0 \quad , \quad 0 < (\frac{1}{2}-\epsilon)^2 < \frac{1}{2} \\
 &\quad 0 < 2(\frac{1}{2}-\epsilon)^2 < 1 \\
 &\therefore 0 < \exp(-2(\frac{1}{2}-\epsilon)^2) < 1 \\
 &\text{証法} \rightarrow \\
 &\text{若 } T = O(\log N) \text{ 則 } T \leq C \log N \\
 &\text{令 } T \text{ 為 } \frac{1}{-2(\frac{1}{2}-\epsilon)^2} \cdot \log(N) \in O(\log N) \\
 &E_{in}(G_T) \leq e^{-2(\frac{1}{2}-\epsilon)^2 \cdot \frac{1}{-2(\frac{1}{2}-\epsilon)^2} \cdot \log(N)} = \frac{1}{N} \\
 &\text{則當 } N \rightarrow \infty \text{ 時} \quad E_{in}(G_T) \leq \lim_{N \rightarrow \infty} \frac{1}{N} = 0 \\
 &\text{証法} \Rightarrow \\
 &E_{in}(G_T) \leq \exp(-2(\frac{1}{2}-\epsilon)^2)^T \quad \text{若在 } N \text{ 夠大時, } \exp(-2(\frac{1}{2}-\epsilon)^2) = 0 \\
 &\therefore \exp(-2(\frac{1}{2}-\epsilon)^2)^T = \lim_{N \rightarrow \infty} \frac{1}{N} = 0 \\
 &\stackrel{\log}{\Rightarrow} T \cdot \log(\exp(-2(\frac{1}{2}-\epsilon)^2)) = \lim_{N \rightarrow \infty} \log \frac{1}{N} \\
 &T = \lim_{N \rightarrow \infty} \frac{-\log N}{-2(\frac{1}{2}-\epsilon)^2} = O(\log N) \\
 &\Rightarrow N \text{ 夠大時, } E_{in}(G_T) = E_{in}(G_{O(\log N)}) \leq \frac{1}{N} \rightarrow 0
 \end{aligned}$$

My Code is available on my Github :

“https://github.com/davidhalladay/Computer-vision-project/tree/master/mini_02”

reference:

1. *SVM* , “<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>”
2. *Matplotlib.contour* , “https://matplotlib.org/api/_as_gen/matplotlib.pyplot.contour.html”
3. *SVM.math* , “<https://scikit-learn.org/stable/modules/svm.html#svm-kernels>”
4. *SVM parameters details* , “<http://www.stardustsky.net/index.php/post/53.html>”
5. *Train_test_split* , “https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html”
6. *Meyer’s theorem* , “https://en.wikipedia.org/wiki/Mercer%27s_theorem”
7. *Construct kernel func.* , “<http://mlweb.loria.fr/book/en/constructingkernels.html>”
8. *Standford SVM* , “<http://cs229.stanford.edu/notes/cs229-notes3.pdf>”
9. *Kernel RR* , “https://scikit-learn.org/stable/modules/generated/sklearn.kernel_ridge.KernelRidge.html#sklearn.kernel_ridge.KernelRidge”