Home Work 1

Machine Learning Techniques

R04323050

經濟碩三　陳伯駒

1.

$(\phi_1(\mathbf{x}_1), \phi_2(\mathbf{x}_1)) = (-2, 0) \Longrightarrow$ ” $\times$ ”

$(\phi_1(\mathbf{x}_2), \phi_2(\mathbf{x}_2)) = (4, 3) \Longrightarrow$ ” $\times$ ”

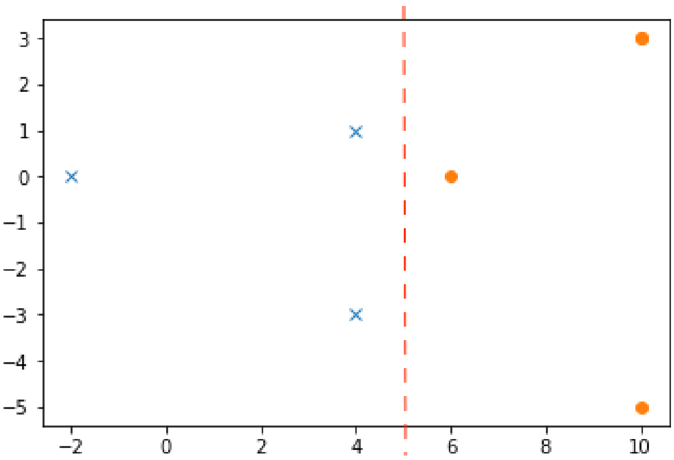
$(\phi_1(\mathbf{x}_3), \phi_2(\mathbf{x}_3)) = (4, 1) \Longrightarrow$ ” $\times$ ”

$(\phi_1(\mathbf{x}_4), \phi_2(\mathbf{x}_4)) = (6, 0) \Longrightarrow$ ” $\circ$ ”

$(\phi_1(\mathbf{x}_5), \phi_2(\mathbf{x}_5)) = (10, -5) \Longrightarrow$ ”$\circ$”

$(\phi_1(\mathbf{x}_6), \phi_2(\mathbf{x}_6)) = (10, 3) \Longrightarrow$ ” $\circ$ ”

$(\phi_1(\mathbf{x}_7), \phi_2(\mathbf{x}_7)) = (10, 3) \Longrightarrow$ ” $\circ$ ”



Pictorially, the optimal separating hyperplane is the equation: $z_1 = 5$。

2.

By implementing the sklearn.svm package in python:

$\boldsymbol{\alpha} = (-0.21970141, -0.28015714, 0.33323258, 0.06819373, 0.09843225)$

$\mathbf{b} = -1.66633495$。Support Vector indices: $[1, 2, 3, 4, 5] = \{\mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$。

3.

The optimal separating hyperplane:

$$
\begin{aligned}
g_{svm}(\mathbf{x}) &= \sum_{\text{SV indices n}} sign(\alpha_n y_n K(\mathbf{x}_n, \mathbf{x}) + \mathbf{b}) \\
&= -0.219 \cdot (-1) \cdot (1 + 2x_2)^2 - 0.28 \cdot (-1) \cdot (1 - 2x_2)^2 + 0.333 \cdot 1 \cdot (1 - 2x_2)^2 \\
&+ 0.068 \cdot 1 \cdot (1 + 4x_2)^2 + 0.098 \cdot 1 \cdot (1 - 4x_2)^2 - 1.66 \\
&= 0.219 \cdot (1 + 2x_2)^2 + 0.28 \cdot (1 - 2x_2)^2 + 0.333 \cdot (1 - 2x)^2 + 0.068 \cdot (1 + 4x_2)^2 \\
&+ 0.098 \cdot (1 - 4x_2)^2 - 1.66
\end{aligned}
$$

4.

$K(\mathbf{x}, \mathbf{x}')$ 所對應到的 z-space 為：$(1, 2x_1, 2x_2, 2x_1^2, 2x_2^2)$。
明顯地與第一題對應到的 $z-$space：$(2x_2^2 - 4x_1 + 2, x_1^2 - 2x_2 - 1)$ 不同，因此推導出的 hyperplane 顯然會不一樣。

5.

Let α_n be the Largrange multuplier for constraint $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq \rho_n - \xi_n$ and β_n is for the constraint $\xi_n \geq 0$, then the primal problem will be:

$$
\begin{aligned}
\min_{\mathbf{b},\mathbf{w},\boldsymbol{\xi}} \max_{\alpha_n > 0, \beta_n > 0} \mathcal{L}((b, \mathbf{w}, \xi), \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N} \mu_n \xi_n + \sum_{n=1}^{N} \alpha_n(\rho_n - \xi_n - y_n(\mathbf{w}^T\mathbf{x}_n + b)) \\
+ \sum_{n=1}^{N} \beta_n \cdot (-\xi_n)
\end{aligned}
$$

First, we simplify β_n by taking the derivative of ξ_n:

$$
\frac{\partial \mathcal{L}}{\partial \xi_n} : C \cdot \mu_n - \alpha_n - \beta_n = 0 \Longrightarrow \begin{cases} \text{implicit constraint:} & \beta_n = C \cdot \mu_n - \alpha_n \\ \text{explicit constraint:} & 0 \leq \alpha_n \leq C \cdot \mu_n \end{cases}
$$

then we can rewrite the problem as:

$$\min_{\mathbf{b},\mathbf{w},\boldsymbol{\xi}} \max_{\substack{0\le\alpha_n\le C\cdot\mu_n\\ \beta_n=C\cdot\mu_n-\alpha_n}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\mu_n\xi_n + \sum_{n=1}^{N}\alpha_n(\rho_n - \xi_n - y_n(\mathbf{w}^T\mathbf{x}_n + b)) + \sum_{n=1}^{N}(C\cdot\mu_n - \alpha_n)\cdot(-\xi_n)$$

$$\therefore\ \mathcal{L}((b,\mathbf{w},\xi),\boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\mu_n\xi_n + \sum_{n=1}^{N}\alpha_n(\rho_n - \xi_n - y_n(\mathbf{w}^T\mathbf{x}_n + b)) + \sum_{n=1}^{N}(C\cdot\mu_n - \alpha_n)\cdot(-\xi_n)$$

6.

By strong duality, the solution would be same as:

$$\max_{\substack{0\le\alpha_n\le C\cdot\mu_n\\ \beta_n=C\cdot\mu_n-\alpha_n}} \min_{\mathbf{b},\mathbf{w},\boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\mu_n\xi_n + \sum_{n=1}^{N}\alpha_n(\rho_n - \xi_n - y_n(\mathbf{w}^T\mathbf{x}_n + b)) + \sum_{n=1}^{N}(C\cdot\mu_n - \alpha_n)\cdot(-\xi_n)$$

Now we simplify the ξ_n:

$$\max_{\substack{0\le\alpha_n\le C\cdot\mu_n\\ \beta_n=C\cdot\mu_n-\alpha_n}} \min_{\mathbf{b},\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\mu_n\xi_n + \sum_{n=1}^{N}\alpha_n(\rho_n - y_n(\mathbf{w}^T\mathbf{x}_n + b))) \equiv \mathcal{L}((b,\mathbf{w},\xi),\boldsymbol{\alpha})$$

which is the inner problem same as hard-margin SVM:

$$\frac{\partial\mathcal{L}}{\partial b} = 0 \Rightarrow \text{no loss of optimality if solving with constraint : } \sum_{n=1}^{N}\alpha_n y_n = 0.$$

$$\frac{\partial\mathcal{L}}{\partial w_i} = 0 \Rightarrow \text{no loss of optimality if solving with constraint : } \mathbf{w} = \sum_{n=1}^{N}\alpha_n y_n \mathbf{z}_n.$$

Hence, by the KKT conditions and Complementary Slackness, the dual problem will be:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \mathbf{z}_n^T\mathbf{z}_m - \sum_{n=1}^{N}\rho_n\alpha_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} \alpha_n y_n = 0$$

$$0 \le \alpha_n \le C \cdot \mu_n, \text{ for } n = 1, 2, \dots N$$

$$\text{implicity} \quad \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n$$

$$\beta_n = C \cdot \mu_n - \alpha_n, \text{ for } n = 1, 2, \dots N$$

7.

If $\rho_n = 0.25$ and $\mu_n = 1$ for all n. The dual problem will be:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - \sum_{n=1}^{N} 0.25 \alpha_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} \alpha_n y_n = 0$$

$$0 \le \alpha_n \le C, \text{ for } n = 1, 2, \dots N$$

Let $\boldsymbol{\alpha}'^*$ be the solution for P_1'; $\boldsymbol{\alpha}^*$ be the solution for P_1. We know the dual problem for P_1 is:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - \sum_{n=1}^{N} \alpha_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} \alpha_n y_n = 0$$

$$0 \le \alpha_n \le C, \text{ for } n = 1, 2, \dots N$$

Hence, the optimal $\boldsymbol{\alpha}^* \times 0.25 = \boldsymbol{\alpha}'^*$ By implicity: $\mathbf{w}'^* = 0.25 \times \mathbf{w}^* \Rightarrow \mathbf{w}^* = 4\mathbf{w}^*$

Now we can solve for b'^* by complementary slackness:

$$\begin{aligned} b'^* &= y_s \rho_s - y_s \xi_s - \mathbf{w}'^T \mathbf{z}_s \quad, s \text{ denotes as the support vector} \\ &= 0.25 y_s - y_s \xi_s - 0.25 \mathbf{w}^T \mathbf{z}_s \\ &= 0.25 b^* - 0.75 y_s \xi_s \quad, \text{where } b^* \text{ is the solution for } P_1 \end{aligned}$$

$$\therefore \quad b^* = 4b'^* + 3y_s \xi_s$$

8.

In the class and slides4 p.10, we know the only difference between hard-margin and soft-margin SVM in dual problem is adding the upper bound C on α_n in soft-margin SVM.

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \mathbf{z}_n^T\mathbf{z}_m - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to} \quad \sum_{n=1}^{N}\alpha_n y_n = 0$$

$$0 \le \alpha_n \le C, \text{ for } n = 1, 2, \dots N$$

Let $\boldsymbol{\alpha}^*$ be the solution in hard-margin. If we set $C \ge \max\limits_{1\le n\le N} \alpha_n$, then the solution is also optimal in soft-margin problem intuitively.

9.

(a). Let $K_1 = \begin{bmatrix} \frac{4}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{4}{5} \end{bmatrix}$, then $K_= \begin{bmatrix} 1.2 & 1.8 \\ 1.8 & 1.2 \end{bmatrix} \Rightarrow det(K) < 0$. If the matrix is positive-semidefinite, determinants of all upper left sub-matrices are non-negative. (✗)

(b). $K(\mathbf{x}, \mathbf{x}') = \mathbf{I}$, let $\mathbf{a}$ be any vector in $\mathcal{R}^n$, then $\mathbf{a}^T\mathbf{I}\mathbf{a} = \mathbf{a}^T\mathbf{a} \ge 0$ (✓)

(c). $K(\mathbf{x}, \mathbf{x}') = (2 - K_1(\mathbf{x}, \mathbf{x}'))^{-1}$. Let $\mathbf{a}$ be any vector in $\mathcal{R}^m$, then:

$$\begin{aligned}\mathbf{a}^T K\ \mathbf{a} &= \sum_{i,j=1}^{m} a_i\ \frac{1}{2 - K_1(x_i, x_j)}\ a_j = \sum_{i,j=1}^{m} a_i\ \frac{K_1(x_i, x_j)}{[2 - K_1(x_i, x_j)]\,(K_1(x_i, x_j))}\ a_j \\ &> \sum_{i,j=1}^{m} a_i\ K_1(x_i, x_j)\ a_j = \mathbf{a}^T K_1\ \mathbf{a} \ge 0 \quad (✓)\end{aligned}$$

(d). $K(\mathbf{x}, \mathbf{x}') = (2 - K_1(\mathbf{x}, \mathbf{x}'))^{-2}$. Let $\mathbf{a}$ be any vector in $\mathcal{R}^m$, then:

$$\begin{aligned}\mathbf{a}^T K\ \mathbf{a} &= \sum_{i,j=1}^{m} a_i\ \frac{1}{(2 - K_1(x_i, x_j)) \cdot (2 - K_1(x_i, x_j))}\ a_j \\ &= \sum_{i,j=1}^{m} a_i\ \frac{K_1(x_i, x_j)}{[2 - K_1(x_i, x_j)]^2\,(K_1(x_i, x_j))}\ a_j \\ &> \sum_{i,j=1}^{m} a_i\ K_1(x_i, x_j)\ a_j = \mathbf{a}^T K_1\ \mathbf{a} \ge 0 \quad (✓)\end{aligned}$$

10.

By the slides in class 3, we know:

$$g_{svm}(\mathbf{x}) = sign(\sum_{\text{SV indices n}} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}) + b)$$

$$= c$$

s denotes as the support vector.

Now we are using $\tilde{K} = p \cdot K(\mathbf{x}, \mathbf{x}')$. To make the result same, we then let $\alpha'_n = \frac{\alpha_n}{p}$ and $\tilde{C} = \frac{C}{p}$, we can solve the dual problem based on $\tilde{K}, \alpha'_n$ and $\tilde{C}$:

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha'_n \alpha'_m y_n y_m \tilde{K} - \sum_{n=1}^{N} \alpha'_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} \alpha'_n y_n = 0$$

$$0 \leq \alpha'_n \leq \tilde{C}, \text{ for } n = 1, 2, \dots N$$

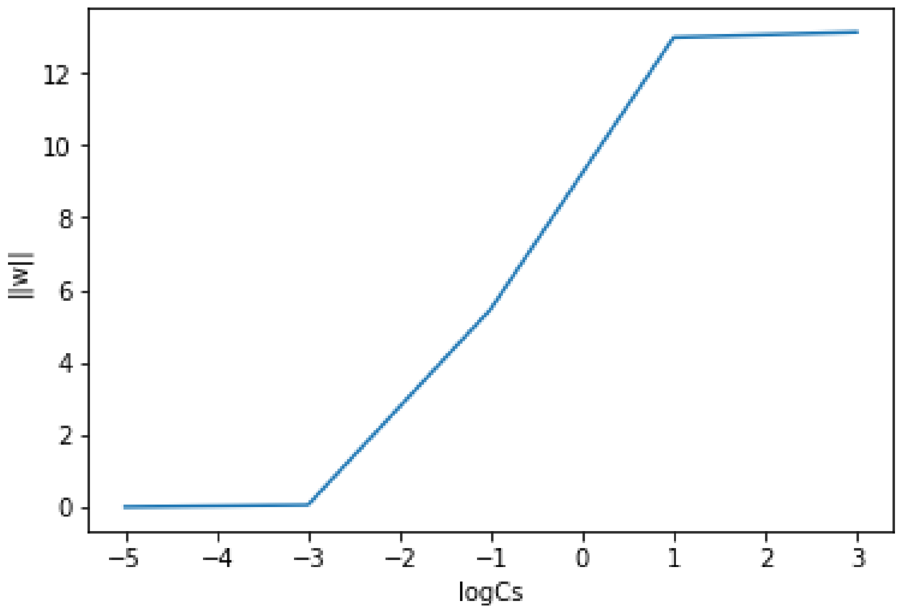
the we know the optimal separating hyperplane will be:

$$\begin{aligned} g_{svm}(\mathbf{x}) &= sign(\sum_{\text{SV}} \alpha'_n y_n \tilde{K}(\mathbf{x}_n, \mathbf{x}) + b) \\ &= sign(\sum_{\text{SV}} \alpha'_n y_n \tilde{K}(\mathbf{x}_n, \mathbf{x}) + y_s - \sum_{\text{SV}} \alpha'_n y_n \tilde{K}(\mathbf{x}_n, \mathbf{x}_s)) \\ &= sign(\sum_{\text{SV}} \frac{\alpha_n}{p} y_n p \cdot K(\mathbf{x}_n, \mathbf{x}) + y_s - \sum_{\text{SV}} \frac{\alpha_n}{p} y_n p \cdot K(\mathbf{x}_n, \mathbf{x}_s)) \\ &= sign(\sum_{\text{SV}} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}) + y_s - \sum_{\text{SV}} \alpha_n y_n K(\mathbf{x}_n, \mathbf{x}_s)) \end{aligned}$$

, which is equivalent to the solution of original problem.

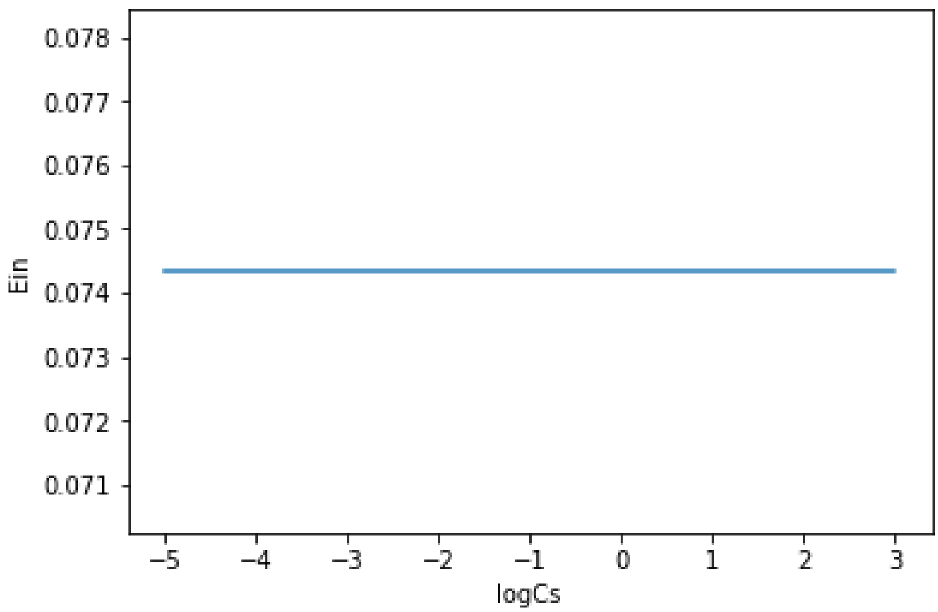
11.

隨著 C 的提升，$\mathbf{w}$ 的長度會越來越長。如下圖：



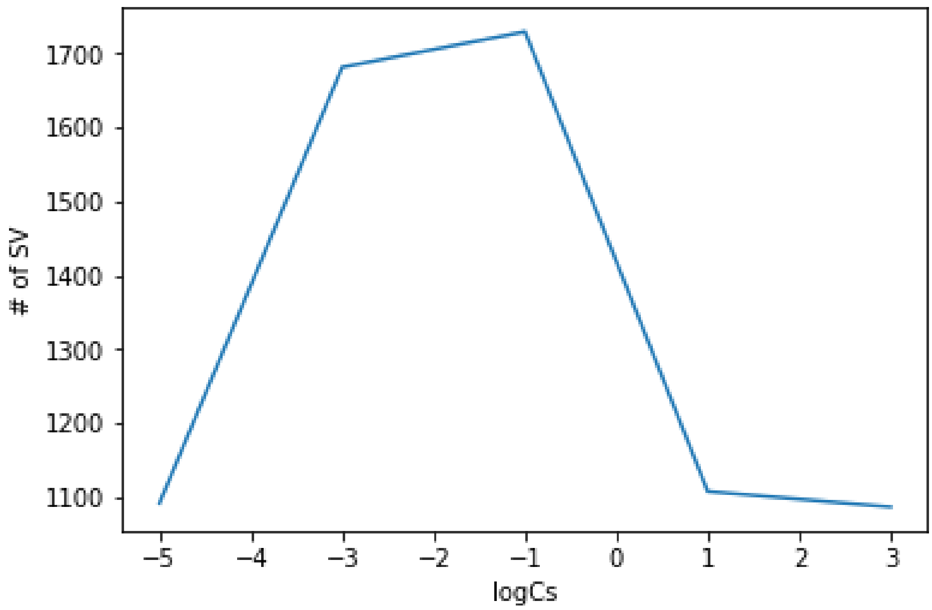
12.

隨著 C 的提升，E_{in} 維持不變。如下圖：



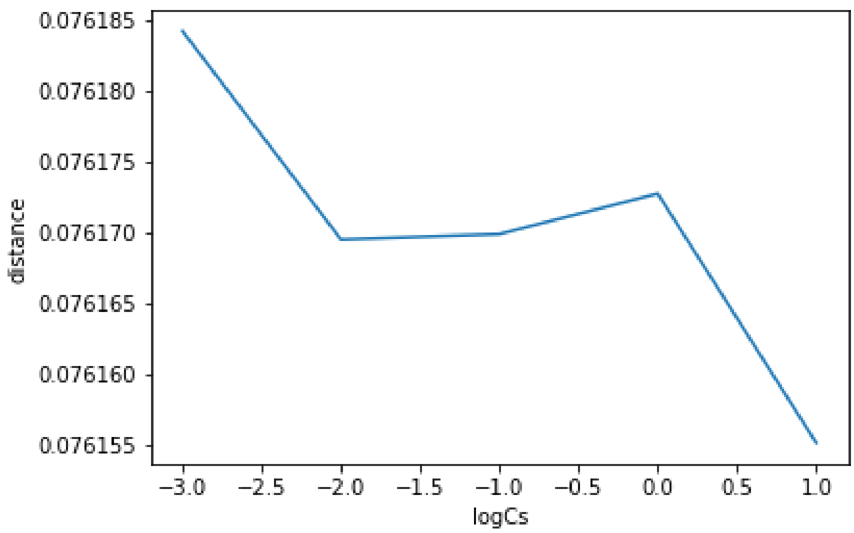
13.

在 $C = 0.1$ 時，有 optimal # of support vectors= 1729



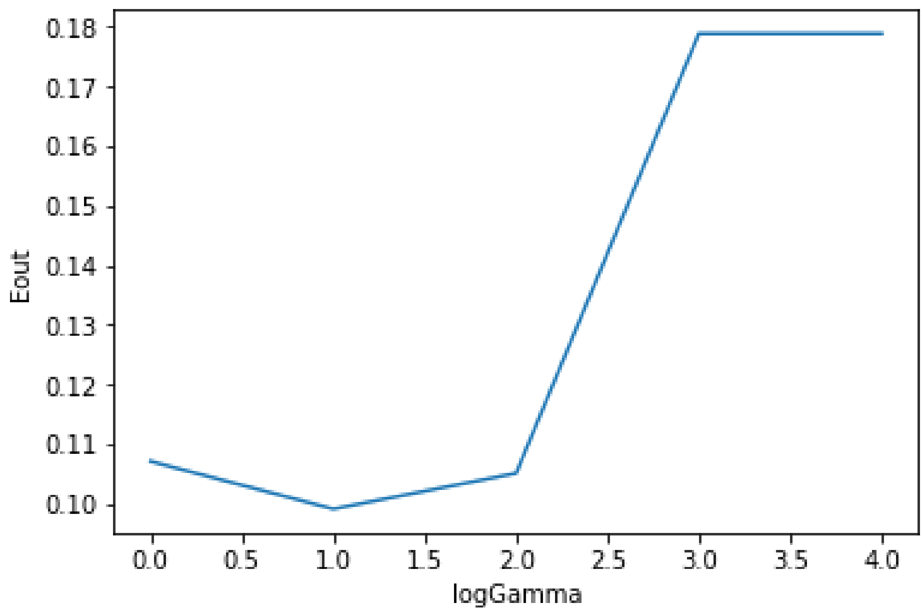
14.

If we choose any free support vector, and compute its distance to separating hyperplane. By the slides in class 1, we know in the primal hard-margin SVM, the distance would be $dist(z, b, \mathbf{w}) = \frac{1}{\|\mathbf{w}\|} \left|\mathbf{w}^T z + b\right|$. Hence, the distance is tend to decrease as C increases due to the increment of $\|\mathbf{w}\|$.



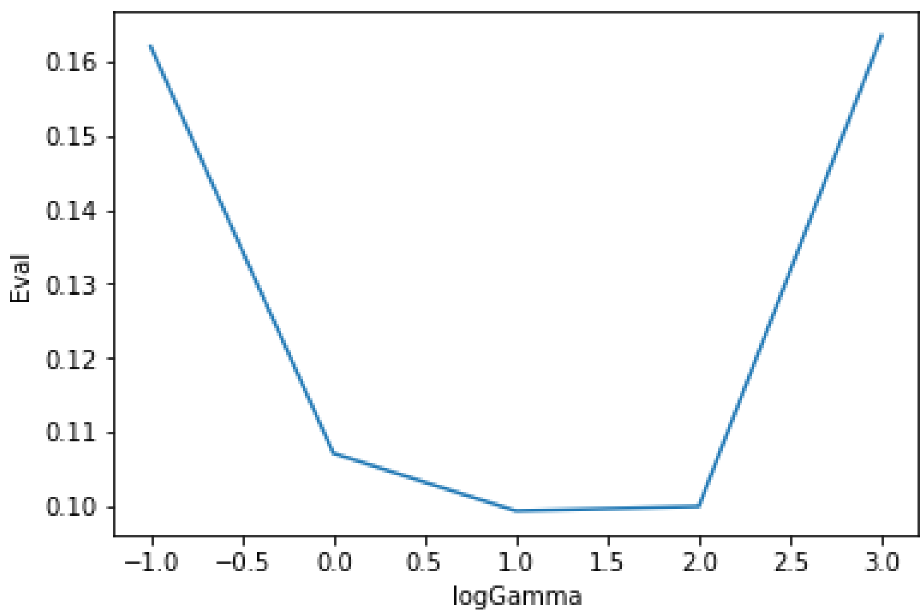
15.

隨著 γ 增加、E_{out} 先降後升，並在 $\gamma = 10$ 時達到最小值。



16.

隨著 γ 增加、E_{val} 先降後升，並在 $\gamma = 10$ 時達到最小值，與我們所期待的結果相符合 (透過 validation 選擇的 γ 也能使 E_{out} 極小)



17.

The optimal kernel SVM solution is: $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n$, for those constant feature component z_i=c ：

$$\sum_{n=1}^{N} \alpha_n y_n z_i = c \cdot \sum_{n=1}^{N} \alpha_n y_n = 0$$

直觀：Constant features will be capture in b^*, which is the intercept term. Unlike what we've learned in PLA, we do not stack up the intercept term.

18.

Let λ be the Lagrange multiplier for constraint $\mathbf{w}^T\mathbf{w} \le C$, then the Lagrange dual problem will be:

$$\min_{\mathbf{w},\lambda} \quad \frac{1}{N}\sum_{n=1}^{N}(y_n - \mathbf{w}^T\mathbf{x}_n)^2 + \lambda(\mathbf{w}^T\mathbf{w} - C), \text{ which is a convex problem by slides in class 2.}$$

f.o.c.

$$\begin{aligned}
&\frac{\partial \mathcal{L}}{\partial \mathbf{w}} : \frac{2}{N}\sum(y_n - \mathbf{w}^T\mathbf{x}_n)(-\mathbf{x}_n) + 2\lambda\mathbf{w} = 0 \Rightarrow \sum(y_n - \mathbf{w}^T\mathbf{x}_n)(-\mathbf{x}_n) = N\lambda\mathbf{w} - ① \\
&\frac{\partial \mathcal{L}}{\partial \lambda} : \mathbf{w}^T\mathbf{w} - C = 0 - ②
\end{aligned}$$

Transform ① condition to the matrix form:

$$\begin{aligned}
\mathbf{x}^T\mathbf{y} - \mathbf{x}^T\mathbf{X}\mathbf{w} = N\lambda\mathbf{w} \Rightarrow \mathbf{x}^T\mathbf{y} = \mathbf{x}^T\mathbf{x}\mathbf{w} + N\lambda\mathbf{w} = (\mathbf{x}^T\mathbf{x} + N\lambda\mathbf{I}_k)\mathbf{w} \\
\Rightarrow \mathbf{w}^* = (\mathbf{w}^T\mathbf{w} + N\lambda\mathbf{I}_k)^{-1}\mathbf{x}^T\mathbf{y}
\end{aligned}$$