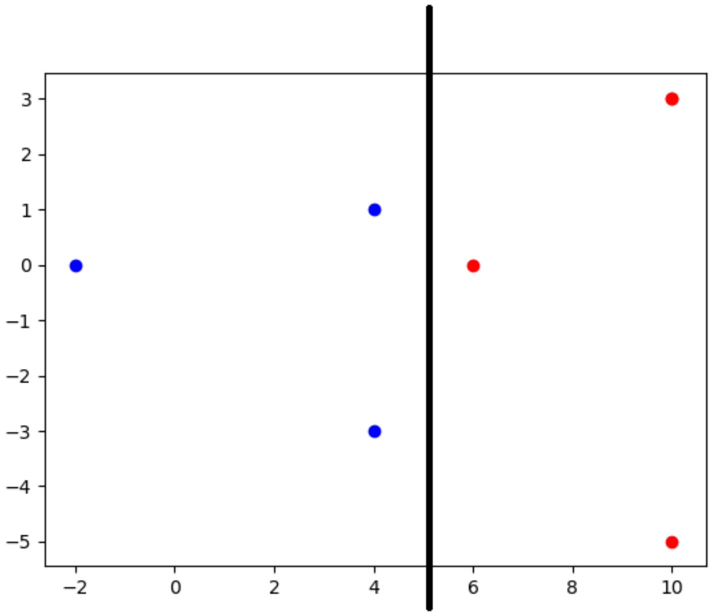
Machine Learning Techniques HW1

B05902109 資工二 柯上優

1.



經過轉換，數值如下，圖案如上，

Z1 = (-2 , 0), y = -1

Z2 = (4 , -3) , y = -1

Z3 = (4 , 1) , y = -1

Z4 = (6 , 0) , y = 1

Z5 = (10 , -5) , y =1

Z6 = (10 , 3) , y =1

Z7 = (10 , 3) , y = 1

直線 z1=5 可以完全分離，x2 x3 x4 為 support vector，皆距離為 1

2.

Alpha1 = 0.0

Alpha2 = 0.21970141

Alpha3 = 0.28015714

Alpha4 = 0.33323258

Alpha5 = 0.06819373

Alpha6 = 0.09843225

Alpha7 = 0.0

Support vector: x2, x3, x4, x5, x6

Library: sklearn.svm

reference: http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

3.

根據投影片公式，可以先展開得到

$$-0.2197(1+2x_2)^2 - 0.2802(1-2x_2)^2 + 0.3332(1-2x_1)^2 + 0.0681(1+4x_2)^2 + 0.0984(1-4x_2)^2 - 1.666 = 0$$

劃簡後得到

$$1.3328{x_1}^2 - 1.3328x_1 + 0.6716{x_2}^2 + 0.0068x_2 - 1.6644 = 0$$

方便計算，我將等式乘 3，忽略 x2 項

$$4{x_1}^2 - 4x_1 + 2{x_2}^2 = 5$$

4.

The two curve should be different, because the two curve is learned from different Z spaces.

5.

$$\frac{1}{2}w^Tw + C\sum_{n=1}^{N}\mu_n\varepsilon_n + \sum_{n=1}^{N}\alpha_n(\rho_n - \varepsilon_n - y_n(w^Tx+b)) + \sum_{n=1}^{N}-\beta_n\varepsilon_n$$

6.

By

$$\frac{\partial L}{\partial \varepsilon_n} = 0 = C\mu_n - \alpha_n - \beta_n$$

We can simplify into

$$\mathrm{L}(\beta, w, \varepsilon, \alpha) = \frac{1}{2}w^Tw + \sum_{n=1}^{N}\alpha_n(\rho_n - y_n(w^Tx_n+b))$$

Similar with Lecture 2 and 4, By

$$\frac{\partial L}{\partial b} = 0 = \sum_{n=1}^{N}\alpha_n y_n \quad \text{, for } \ \mathrm{n} = 1,2,\dots,\mathrm{N}$$

$$\frac{\partial L}{\partial w} = 0 \ \ , w = \sum_{n=1}^{N}\alpha_n y_n x_n \quad \text{, for } \ \mathrm{n} = 1,2,\dots,\mathrm{N}$$

we can get

$$\min \ \frac{1}{2}\left\|\sum_{n=1}^{N}\alpha_n y_n x_n\right\|^2 - \sum_{n=1}^{N}\alpha_n\rho_n$$

$$\text{subject to } \sum_{n=1}^{N} \alpha_n y_n = 0; \quad 0 \le \alpha_n \le C\mu_n \text{ , for } n = 1,2,\ldots,N$$

$$\text{implicitly } w = \sum_{n=1}^{N} \alpha_n y_n x_n \text{ , } \beta_n = C\mu_n - \alpha_n \text{ , for } n = 1,2,\ldots,N$$

7.

In the (P1’)

$$\min_{w,b,\varepsilon} \frac{1}{2} w'^T w' + C \sum_{n=1}^{N} \varepsilon_n$$

$$\text{s.t.} \quad y_n\left(w'^T w'_n + b'\right) \ge 0.25 - \varepsilon_n \qquad \varepsilon_n \ge 0$$

目標是將 0.25 改成 1，才能與(P1)互相比較。

將上式中 min 乘 16 並進行分配，同時使 s.t.內有相同的限制，將會得到

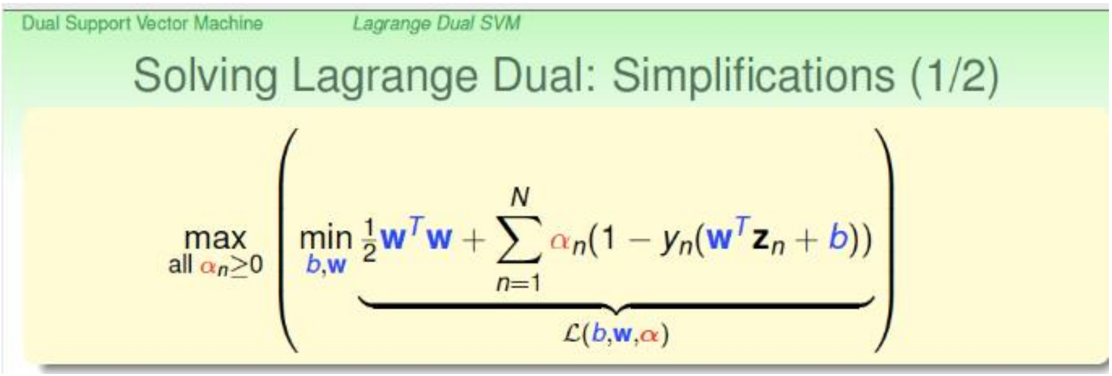
$$\min_{w,b,\varepsilon} \frac{1}{2} (4w')^T (4w') + (4C) \sum_{n=1}^{N} (4\varepsilon_n)$$

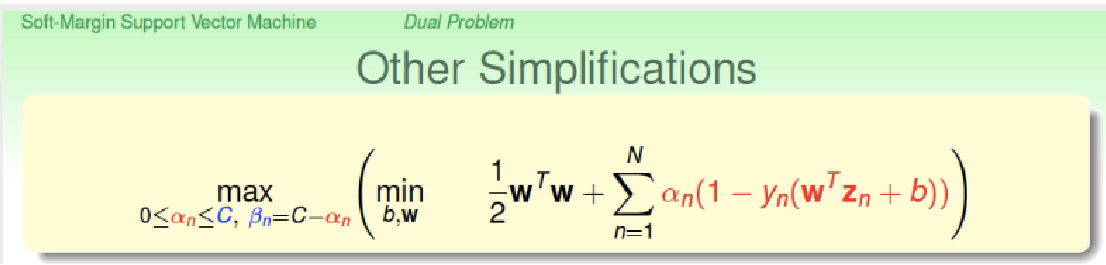
$$\text{s.t.} \quad y_n((4w')^T (4w')_n + 4b') \ge 1 - 4\varepsilon_n \qquad 4\varepsilon_n \ge 0$$

此時得到的$4w'$與 4b’即為(P1)的解，我選擇的 C 為(P1’)的四倍，4C。

8.

由上課投影片中的 hard-margin(上)和 soft-margin(下)





可知差異在於$\beta_n = C - \alpha_n \ge 0$，須滿足$0 \le \alpha_n \le C$。

題目要求，若 C 滿足$C \geq \max_{1 \le n < N} \alpha_n^*$，此要求恰好符合 soft-margin 的要求而不使 alpha 改變，故得證，在此要求下的 C，會使兩者有相同 alpha 的解。

9.
Ans = (b.)(c.)(d.)
(a.)

反例:令$\begin{bmatrix} a & b \\ b & a \end{bmatrix}$為合法 kernel，特徵值有 a+b, a-b 兩個，滿足 a+b>=0 且 a>=b。

新 kernel$\begin{bmatrix} 2-a & 2-b \\ 2-b & 2-a \end{bmatrix}$有特徵值 4-a-b, b-a，b-a 必定小於 0，非半正定。

(b.)

新 kernel = $\begin{bmatrix} 1 & \dots & 1 \\ \dots & \dots. & \dots \\ 1 & \dots & 1 \end{bmatrix}$，$[x_1 \quad \dots \quad x_n]\begin{bmatrix} 1 & \dots & 1 \\ \dots & \dots. & \dots \\ 1 & \dots & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix} = (x_1 + \cdots + x_n)^2 \geq 0$

其他半正定性值皆符合。

(c.)

$$令 \text{ Kernel 滿足} x^T K x = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i k_{ij} x_j \geq 0\text{，現有新 kernel} = (2-K)^{-1}\text{，}$$

$$則 x^T (2-K)^{-1} x = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{x_i x_j}{2-k_{ij}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{x_i x_j k_{ij}}{(2-k_{ij})k_{ij}} \geq \sum_{i=1}^{N}\sum_{j=1}^{N} x_i k_{ij} x_j \geq 0,$$

$$\text{by } \ 1 \geq (2-\mathrm{x})\mathrm{x} > 0 \ \text{ where } \ 0 < \mathrm{x} < 2\text{。}$$

(d.)
Kernel has a property of production.。
Reference:
http://gpss.cc/gpss13/assets/Sheffield-GPSS2013-Durrande.pdf
Set

$$K_1(x,x') = a(x)^T a(x') \ , \qquad K_2(x,x') = b(x)^T b(x')$$
$$a(x) = [a_1(x), a_2(x), \dots\dots, a_N(x)]$$
$$b(x) = [b_1(x), b_2(x), \dots\dots, b_M(x)]$$

$$K_3(x,x') = K_1(x,x')K_2(x,x') = \left(\sum_{n=1}^{N} a_n(x)^T a_n(x')\right)\left(\sum_{m=1}^{M} b_m(x)^T b_m(x')\right)$$

$$= \sum_{n=1}^{N}\sum_{m=1}^{M} [a_n(x)b_m(x)][a_n(x')b_m(x')] = \sum_{n=1}^{N}\sum_{m=1}^{M} c_{nm}(x)c_{nm}(x')$$

$$= c(x)^T c(x'), \text{where c is a N} - \text{M dimention vector.}$$

接著，由(c.)知道$(2-K)^{-1}$合法，可知道$(2-K)^{-1}(2-K)^{-1} = (2-K)^{-2}$合法。

10.

Reference: B05902013 吳宗翰

由 dual problem 中

$$\min \ \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n\alpha_m y_n y_m K(x,x') - \sum_{n=1}^{N}\alpha_n$$

令新的問題中，用 alpha’表示，並使用$\widetilde{K}$。新的問題:

$$\min \ \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha'_n\alpha'_m y_n y_m * pK(x,x') - \sum_{n=1}^{N}\alpha'_n$$

現在將式子乘 p，

$$\min p\left(\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha'_n\alpha'_m y_n y_m pK(x,x') - \sum_{n=1}^{N}\alpha'_n\right)$$

$$= \min \ \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} (p\alpha'_n)(p\alpha'_m) y_n y_m K(x,x') - \sum_{n=1}^{N} p\alpha'_n$$

Due to minimum problem

$$\mathrm{p} * \alpha' = \alpha$$

又由投影片可知，其中

$$0 \le p * \alpha'_n \le \mathrm{C}$$

$$0 \le {\alpha_n}' \le \frac{\mathrm{C}}{\mathrm{p}}$$

設$\tilde{C} = \frac{C}{p}$

由分類器公式

$$g_{svm} = sign(\sum_{SV} a_n y_n K(x_n, x') + b)$$

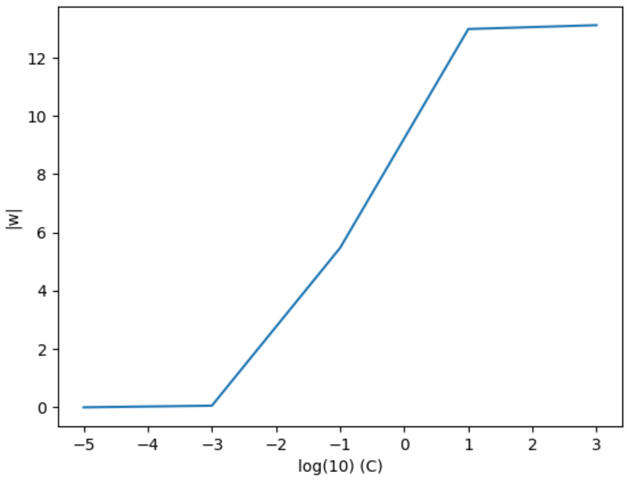
可知，新的分類器

$$\tilde{g}_{svm} = sign(\sum_{SV} \alpha'_n y_n * \tilde{K}(x_n, x') + b)$$

$$\tilde{g}_{svm} = sign(\sum_{SV} \frac{a_n}{p} y_n * pK(x_n, x') + b)$$

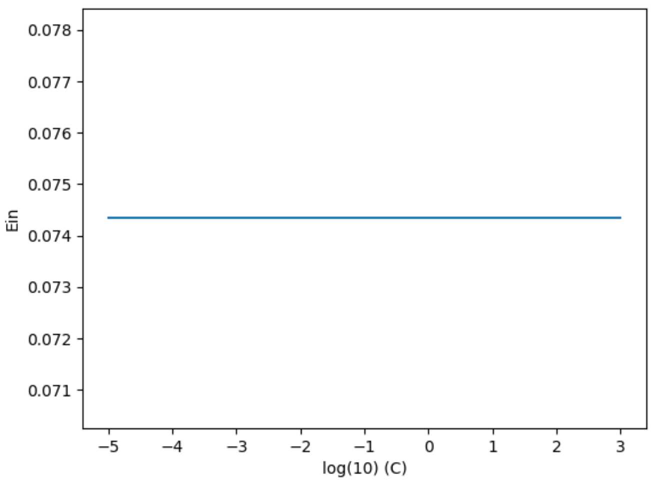
和原分類器一樣，且使用$\tilde{C} = \frac{C}{p}$

11.



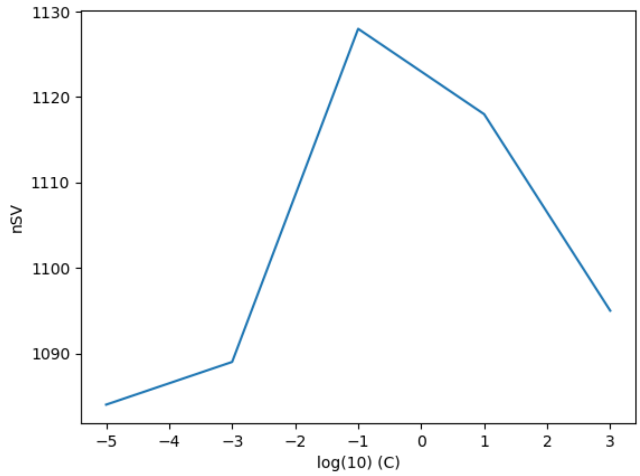
由數據可知 C 越大|w|有明顯的增加，但不會無限制增長而是會趨近。
我認為是容錯越大後，分類線變得更彎曲更容易 overfit 所致。

12.



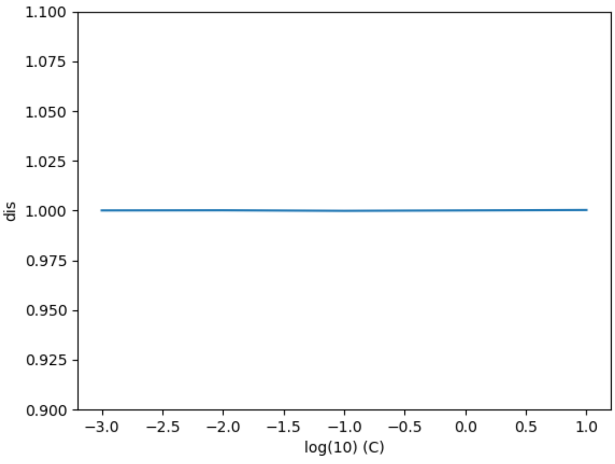
由圖可知，Ein 在不同 C 下都有著相同的值。
我認為是原始資料的分布方式，使不同分類器的樣子易有相似的結果。

13.



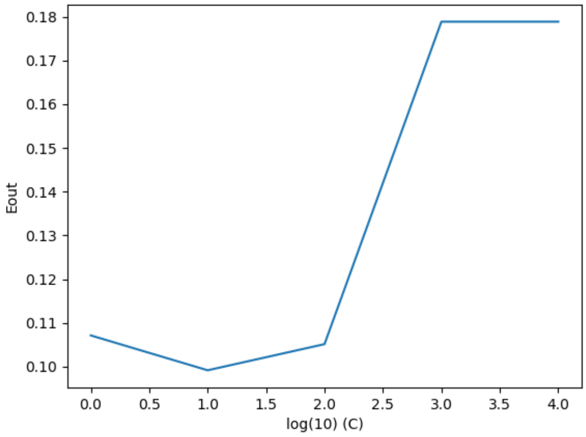
由數據可知，support vectors 的數量明顯成長到某個 C 後又開始驟減。

14.



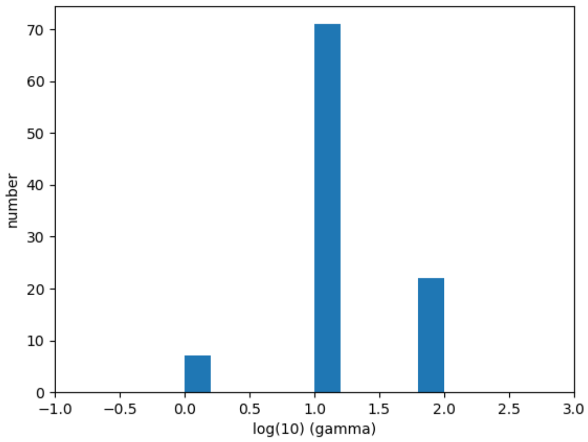
根據圖片可知，距離都大約在 1 附近。

15.



由圖片可知，Eout 減少後又增加。我認為可能發生 overfit。

16.



由數據可知，gamma 在 10 時，有五種狀況中的最好 E_val。

17.

Reference: B05902028 王元益

假設算出了根據原始問題

$$\min_{b,w} \frac{1}{2} w^T w$$

$$\text{s.t.} \quad y_n(w^T x_n + b) \geq 1 \quad , for \quad n = 1,2, \dots, N$$

得到 optimal solotion: w, b，其中 w 皆非 0

現在假設新的 w’ = w，但其中第 i 項 w_i’ = 0。

假設新的 b’ = b + w_i * z_i

在 SVM 上有

$$g'_{svm} = sign\left(\left(\sum_{j=1}^{N} w'_j x_j\right) + b'\right) = sign\left(\left(\sum_{j=1, j!=i}^{N} w_j x_j\right) + b + w_i * z_i\right)$$

$$= sign\left(\sum_{j=1}^{N} w_j x_j + b\right) = g_{svm}$$

雖然是完全相同的 SVM，但是在原始問題中，其中一項 $w'_i = 0$，會使出現更小的 minima，求出來的 w',b'有更好的解，矛盾，w,b 非 optimal。故不會有全非 0 的 w，prove。

18.

將 constrain 放入題目，會變成

$$\max_{\lambda \geq 0} \min_{w,C} \left(\frac{1}{N}\sum_{n=1}^{N}(y_n - w^T x_n)^2\right) + \lambda(w^T w - C)$$

分別對 w 與λ偏微分，會得到以下

$$\frac{\partial L}{\partial w} = \frac{1}{N}(-2XY + 2wX^T X) + 2\lambda w = 0$$

$$\mathrm{w} = \left(\frac{1}{N}X^T X - \lambda I\right)^{-1}\frac{1}{N}XY$$

原式等於

$$\max_{\lambda \geq 0}\left(-\sum_{n=1}^{N}\left(y_n - \left(\left(\frac{1}{N}X^T X - \lambda I\right)^{-1}\frac{1}{N}XY\right)^T x_n\right)^2\right)$$

$$+\lambda(\left(\left(\frac{1}{N}X^T X - \lambda I\right)^{-1}\frac{1}{N}XY\right)^T (\left(\frac{1}{N}X^T X - \lambda I\right)^{-1}\frac{1}{N}XY) - C)$$

最後得到

$$\min_{\lambda}\sum_{n=1}^{N}\left(y_n - (\left(\frac{1}{N}X^T X - \lambda I\right)^{-1}\frac{1}{N}XY)^T x_n\right)^2$$

$$-\lambda(\left(\left(\frac{1}{N}X^T X - \lambda I\right)^{-1}\frac{1}{N}XY\right)^T (\left(\frac{1}{N}X^T X - \lambda I\right)^{-1}\frac{1}{N}XY) - C)$$

$$\text{subject to} \quad \lambda \geq 0$$