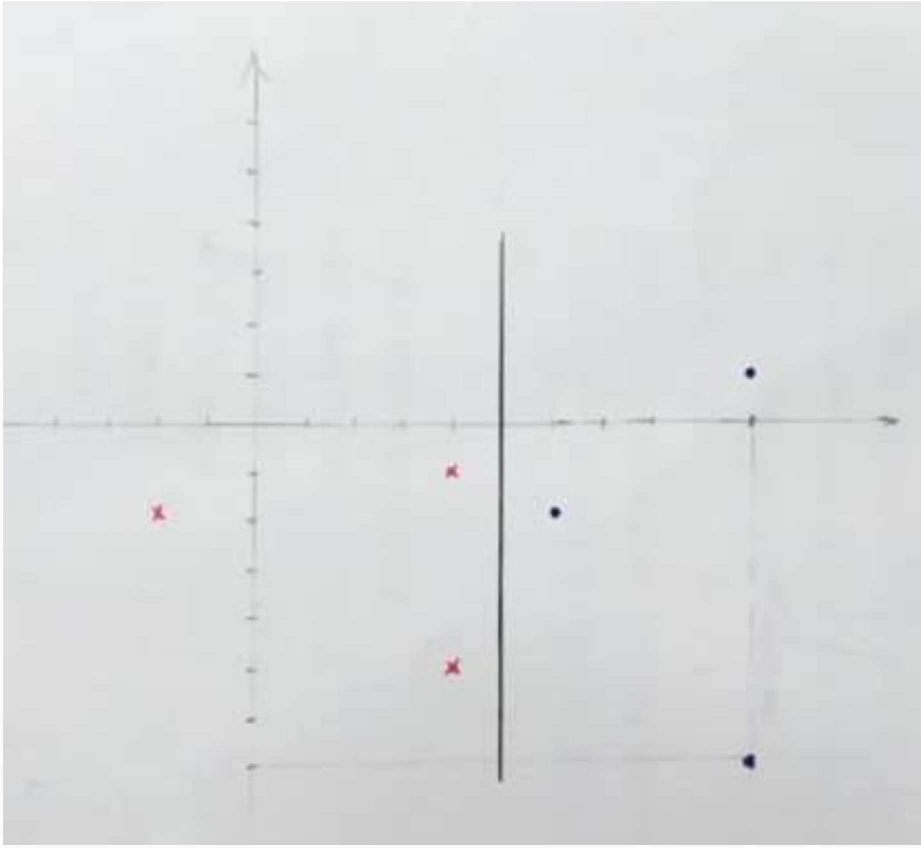
1.

這七點經過轉換後變成: (表成(z_1, z_2, y))

(-2, -2, -1), (4, -5, -1), (4, -1, -1), (6, -2, 1), (10, -7, 1), (10, 1, 1), (10, 1, 1)



易知，轉換後空間中之 optimal separating hyperplane 為$z_1 = 5$(或表示成$\phi_1(x) = 5$)。即為圖中之粗線

2.

直接使用 scikit-learn 來做 SVM。(version == 0.20.0)，程式如下：

```python
from sklearn.svm import SVC
import numpy as np

data = np.array([[1, 0, -1], [0, 1, -1], [0, -1, -1], [-1, 0, 1], [0, 2, 1], [0, -2, 1], [-2, 0, 1]])
X = data[:, 0:2]
Y = data[:, 2]
model = SVC(kernel = 'poly', degree = 2, gamma = 1, coef0 = 1, C=1e10)

k = model.fit(X, Y)
print(k)

print("support vector:", model.support_, )
print("dual coefficient:", model.dual_coef_)
```

Result：

```
Benson@Benson-Tsai MINGW64 ~/Desktop/ML_Technique/hw1 (master)
$ python 2.py
C:\Users\Benson\AppData\Local\Programs\Python\Python36\lib\site-packages\sklearn\externals\joblib\externals\cloudpickle\cloudpickle.py
:47: DeprecationWarning: the imp module is deprecated in favour of importlib; see the module's documentation for alternative uses
  import imp
SVC(C=10000000000.0, cache_size=200, class_weight=None, coef0=1,
  decision_function_shape='ovr', degree=2, gamma=1, kernel='poly',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False)
support vector: [1 2 3 4 5]
dual coefficient: [[-0.59647182 -0.81065085  0.8887034   0.20566488  0.31275439]]
```

可以看出 support vector 為x_2, x_3, x_4, x_5, x_6，而 optimal α為：

[0, 0.59647182, 0.81065085, 0.8887034, 0.20566488, 0.31275439, 0]

*註：因為 sklearn 的α_n為我們上課定義的$\alpha_n * y_n$，故上述之答案為已修正後，符合我們定義之α。

3.

由 kernel function 我們可以知道其轉換式$\phi(x) = \{1, x_1, x_2, x_1^2, x_2^2, 2x_1x_2\}$。

(註：將x_1x_2放在最後面試為了使 coding 更方便)

在 z 空間中，其$w = \sum \alpha_n y_n z_n$。

接著使用以下程式：

```python
import numpy as np

data = np.array([[1, 0, -1], [0, 1, -1], [0, -1, -1], [-1, 0, 1], [0, 2, 1], [0, -2, 1], [-2, 0, 1]])
X = data[:, 0:2]
Y = data[:, 2].reshape(-1, 1)

z = np.hstack((np.ones(X.shape[0]).reshape(-1, 1), np.sqrt(2)*X, X ** 2, (2*X[:, 0]*X[:, 1]).reshape(-1, 1)))
alpha = np.array([0, 0.59647182, 0.81065085, 0.8887034, 0.20566488, 0.31275439, 0]).reshape(-1, 1)

w = np.sum(alpha * Y * z, axis = 0)
print("w = ", w)
b = Y[2] - w.dot(z[2, :])
print("b = ", b)
```

可以計算出：

w = [-1.11022302e-16, -1.25681640, 1.41421358e-08, 8.88703400e-01, 6.66554410e-01, 0]

b =-1.66655439

故可知nonliner curve為：(僅取至小數第4位)

$$-1.6666 - 1.2568x_1 + 0.8887x_1^2 + 0.6666x_2^2 = 0$$

4.

第一題之$\phi_1(x) = 5$轉換回原本的維度後，其方程式為：

$$2x_2^2 - 4x_1 + 3 = 0$$

這和上一題之結果是不一樣的。

而實際上這兩者也不應該一樣，因為他們轉換到的維度並不相同，一個轉換到 2 維，另一個轉換到 5 維。再轉回以原本維度來表示，本來就不應該相同。

5.

首先，先算出$\tilde{\phi}(x)$長度之平方：

$$\|\tilde{\phi}(x)\|^2 = \sum_{k=0} \frac{2^k}{k!} x^{2k}$$

而又因為e^{2x^2}的泰勒展開式為：

$$e^{2x^2} = \sum_{k=0} \frac{2^k}{k!} x^{2k}$$

由此可知：$\left|\widetilde{\phi}(x)\right|^2 = e^{2x^2} = \frac{1}{e^{-2x^2}}$

故我們可以知道$1 = \left\|\widetilde{\phi}(x)\right\|^2 * e^{-2x^2} = (\left\|\widetilde{\phi}(x)\right\| * e^{-x^2})^2$ ，即可得：

$$\left\|\widetilde{\phi}(x)\right\| * e^{-x^2} = \pm 1$$

又因為$|\phi(x)|$為$\phi(x)$之大小，必> 0。e^{-x^2}為指數函數，無論x直為何，恆正。
故兩者乘積必為正數，故可得$\left\|\widetilde{\phi}(x)\right\| * e^{-x^2} = 1$。即為：

$$e^{-x^2} = \frac{1}{\left\|\widetilde{\phi}(x)\right\|}$$

6.

定義$\phi(x) = \left[\frac{x_1}{|x|}, \frac{x_2}{|x|}, \dots, \frac{x_n}{|x|}\right]^T$，kernel function 為$k(x, x') = \cos(x, x') = \phi(x)^T\phi(x')$。

Kernel matrix 如下：$(k_{ij} = k(x_i, x_j))$

$$K = \begin{bmatrix} \phi(x_1)^T\phi(x_1) & \phi(x_1)^T\phi(x_2) & \dots & \phi(x_1)^T\phi(x_2) \\ \phi(x_2)^T\phi(x_1) & \phi(x_2)^T\phi(x_2) & \dots & \phi(x_2)^T\phi(x_2) \\ \dots & \dots & & \dots \\ \phi(x_n)^T\phi(x_1) & \phi(x_n)^T\phi(x_2) & \dots & \phi(x_n)^T\phi(x_2) \end{bmatrix}$$

易可看出：

$$K = [\phi(x_1), \phi(x_2), \dots, \phi(x_n)]^T[\phi(x_1), \phi(x_2), \dots, \phi(x_n)] = Z^TZ$$

由此可知 K 為半正定矩陣。

7.

先列出題目所需之條件：

$$\min_{R\in\mathbb{R}, c\in\mathbb{R}^d} R^2, \text{s.t.} \|z_n - c\|^2 \le R^2 \ \text{ for } n = 1, 2, \dots, N$$

由上課所提之方法(with Lagrange multiplier)，可知要表示成以下形式：

$$L(R, c, \lambda) = "objective" + \sum_{n=1}^{N} \lambda_n * "constraint"$$

故$L(R, c, \lambda)$即為：

$$L(R, c, \lambda) = R^2 + \sum_{n=1}^{N} \lambda_n(\|z_n - c\|^2 - R^2)$$

8.

Dual problem 為：

$$\max_{\lambda_n \ge 0} \min_{R\in\mathbb{R}, c\in\mathbb{R}^d} L(R, c, \lambda) = \max_{\lambda_n \ge 0} \min_{R\in\mathbb{R}, c\in\mathbb{R}^d} R^2 + \sum_{n=1}^{N} \lambda_n(\|z_n - c\|^2 - R^2)$$

在 optimal 時：

(1) 對R偏微：

$$\frac{\partial \mathrm{L}(\mathrm{R},\mathrm{c},\lambda)}{\partial R}=\frac{\partial(\mathrm{R}^2-\sum_{n=1}^{N}\lambda_n\mathrm{R}^2)}{\partial R}=2R\left(1-\sum_{n=1}^{N}\lambda_\mathrm{n}\right)=0$$

$$\sum_{n=1}^{N}\lambda_\mathrm{n}=1 \text{ or } \mathrm{R}=0$$

(2) 對c偏微：

$$\frac{\partial \mathrm{L}(\mathrm{R},\mathrm{c},\lambda)}{\partial c}=\frac{\partial\sum_{n=1}^{N}\lambda_n\|\mathrm{z_n}-\mathrm{c}\|^2}{\partial c}=\sum_{n=1}^{N}\lambda_n(2\mathrm{c}-2\mathrm{z_n})=0$$

$$\sum_{n=1}^{N}\lambda_n(\mathrm{c}-\mathrm{z_n})=0$$

可得以下 KKT conditions:

(1) primal feasible: $|\mathrm{z_n}-\mathrm{c}|^2\le\mathrm{R}^2$

(2) dual feasible: $\lambda_n\ge0$

(3) optimal (partial over R): $\sum_{n=1}^{N}\lambda_\mathrm{n}=1\ or\ R=0$

(4) optimal (partial over c): $\sum_{n=1}^{N}\lambda_n(\mathrm{c}-\mathrm{z_n})=0$

(5) primal-inner optimal: $\lambda_n(\|\mathrm{z_n}-\mathrm{c}\|^2-\mathrm{R}^2)=0$

而由第四項(optimal of partial over c)，可推得以下：

$$\sum_{\mathrm{n}=1}^{\mathrm{N}}\lambda_\mathrm{n}(\mathrm{c}-\mathrm{z_n})=\mathrm{c}\sum_{\mathrm{n}=1}^{\mathrm{N}}\lambda_\mathrm{n}-\sum_{n=1}^{N}\lambda_n\mathrm{z_n}=0$$

$$\mathrm{c}\sum_{\mathrm{n}=1}^{\mathrm{N}}\lambda_\mathrm{n}=\sum_{n=1}^{N}\lambda_n\mathrm{z_n}$$

所以我們可以得到：

$$\mathrm{c}=\frac{\sum_{\mathrm{n}=1}^{\mathrm{N}}\lambda_\mathrm{n}\mathrm{z_n}}{\sum_{\mathrm{n}=1}^{\mathrm{N}}\lambda_\mathrm{n}}\ ,\text{if}\sum_{\mathrm{n}=1}^{\mathrm{N}}\lambda_\mathrm{n}\ne0$$

9.

因為$\mathrm{R}\ne0$，故可知上面的條件有$\sum_{n=1}^{N}\lambda_\mathrm{n}=1$，又可因此帶入c的條件中，可得 $\mathrm{c}=\sum_{\mathrm{n}=1}^{\mathrm{N}}\lambda_\mathrm{n}\mathrm{z_n}$。帶回原本之$\mathrm{L}(\mathrm{R},\mathrm{c},\lambda)$：

$$\mathrm{L}(\mathrm{R},\mathrm{c},\lambda)=\mathrm{R}^2+\sum_{n=1}^{N}\lambda_n(\|\mathrm{z_n}-\mathrm{c}\|^2-\mathrm{R}^2)$$

$$=\mathrm{R}^2\left(1-\sum_{n=1}^{N}\lambda_n\right)+\sum_{n=1}^{N}\lambda_n\|\mathrm{z_n}-\mathrm{c}\|^2=\sum_{n=1}^{N}\lambda_n\|\mathrm{z_n}-\mathrm{c}\|^2$$

$$=\sum_{n=1}^{N}\lambda_n(\mathrm{z_n^Tz_n}-2\mathrm{z_n^Tc}+\mathrm{c^Tc})$$

$$=\sum_{n=1}^{N}\lambda_n z_n^T z_n-2c^T\sum_{n=1}^{N}\lambda_n z_n+c^T c\sum_{n=1}^{N}\lambda_\mathrm{n}$$

$$= \sum_{n=1}^{N} \lambda_n z_n^T z_n - (\sum_{n=1}^{N} \lambda_n z_n)^T \sum_{n=1}^{N} \lambda_n z_n$$

故原本之 dual problem 即為：

$$\max_{\lambda_n \geq 0, \sum_{n=1}^{N} \lambda_n = 1} \sum_{n=1}^{N} \lambda_n z_n^T z_n - (\sum_{n=1}^{N} \lambda_n z_n)^T \sum_{n=1}^{N} \lambda_n z_n$$

10.

對於某個$i \in [1, N]\ s.t. \lambda_i > 0$，必可得以下：$\|z_i - c\|^2 - R^2 = 0$

可推導成：

$$R^2 = z_i^T z_i - 2z_i^T c + c^T c = z_i^T z_i - 2z_i^T \sum_{n=1}^{N} \lambda_n z_n + \left(\sum_{n=1}^{N} \lambda_n z_n\right)^T \sum_{n=1}^{N} \lambda_n z_n$$

$$= K(x_i, x_i) - 2\sum_{n=1}^{N} \lambda_n K(x_i, x_n) + \sum_{m=1}^{N} \sum_{n=1}^{N} \lambda_m \lambda_n K(x_m, x_n)$$

故 R 為：

$$R = \sqrt{K(x_i, x_i) - 2\sum_{n=1}^{N} \lambda_n K(x_i, x_n) + \sum_{m=1}^{N} \sum_{n=1}^{N} \lambda_m \lambda_n K(x_m, x_n)}$$

其中，$i \in [1, N]\ s.t. \lambda_i > 0$

11.

上課時導出以下 hard margin 和 soft margin 的 dual problem：

Hard margin:

$$\max_{\alpha_n \geq 0}(\min_{b,w} \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n (1 - y_n(w^T z_n + b)))$$

Soft margin:

$$\max_{C \geq \alpha_n \geq 0}(\min_{b,w} \frac{1}{2} w^T w + \sum_{n=1}^{N} \alpha_n (1 - y_n(w^T z_n + b)))$$

而若所有在 hard margin 中的α^*滿足：$C \geq \max_{1 \leq n \leq N} \alpha_n^*$則其亦符合 soft margin 中之

$C \geq \alpha_n \geq 0$之條件。故α^*亦為 soft-SVM 之 optimal solution。

用反證法可以證得：

若存在一組α'為 soft-SVM 的 optimal solution，表示$\alpha = \alpha'$時比$\alpha = \alpha^*$好。而因為只要符合 soft-SVM，必會符合 hard-SVM。故在 hard-SVM 中，$\alpha = \alpha'$時，其解亦會比$\alpha = \alpha^*$好。但又因$\alpha = \alpha^*$為 optimal solution，矛盾。

故可得之α^*亦為 soft-SVM 之 optimal solution。

12.
先列出原本的 dual problem 的 QP problem：

$$\min_{\alpha} \frac{1}{2}\alpha^{T}Q\alpha + p_{QP}^{T}\alpha$$

subject to $A\alpha \geq C_{QP}$, 其要滿足：$\sum y_n \alpha_n = 0\,, C \geq \alpha_n \geq 0\ for\ n = 1, 2, \ldots, N$

其中，(Q, p, A, C_{QP})如下：

$$Q = \begin{bmatrix} y_1y_1K(x_1, x_1) & y_1y_2K(x_1, x_2) & \ldots & y_1y_NK(x_1, x_N) \\ y_2y_1K(x_2, x_1) & y_2y_2K(x_2, x_2) & & y_2y_NK(x_2, x_N) \\ \ldots & & & \ldots \\ y_Ny_1K(x_N, x_1) & y_Ny_2K(x_N, x_2) & \ldots & y_Ny_NK(x_N, x_N) \end{bmatrix}$$

$$p_{QP} = -I_n, A = \begin{bmatrix} y^T \\ -y^T \\ I_N \\ -I_N \end{bmatrix}, C_{QP} = \begin{bmatrix} 0 \\ 0 \\ 0_N \\ -CI_N \end{bmatrix}$$

若 K(x, x')使用新的$\tilde{K}(x, x') = pK(x, x')$，C 亦使用新的$\tilde{C} = \frac{C}{p}$：

$$Q' = \begin{bmatrix} y_1y_1pK(x_1, x_1) & y_1y_2pK(x_1, x_2) & \ldots & y_1y_NpK(x_1, x_N) \\ y_2y_1pK(x_2, x_1) & y_2y_2pK(x_2, x_2) & & y_2y_NpK(x_2, x_N) \\ \ldots & & & \ldots \\ y_Ny_1pK(x_N, x_1) & y_Ny_2pK(x_N, x_2) & \ldots & y_Ny_NpK(x_N, x_N) \end{bmatrix}, C'_{QP} = \begin{bmatrix} 0 \\ 0 \\ 0_N \\ -\frac{C}{p}I_N \end{bmatrix}$$

可得$Q' = pQ, C'_{QP} = \frac{C_{QP}}{p}$，而$p_{QP}, A$不變。

故使用新的$\tilde{K}$, $\tilde{C}$，其 QP problem 為：(為求式子整潔，直接將p_{QP}用$-I_N$表示)

$$\min_{\alpha} \frac{1}{2}p\alpha^{T}Q\alpha - I_N\alpha\,, \text{subject to } pA\alpha \geq C_{QP}$$

而因 p 為常數，故$\min_{\alpha} \frac{1}{2}p\alpha^{T}Q\alpha - I_N\alpha$和$\min_{\alpha} p(\frac{1}{2}p\alpha^{T}Q\alpha - I_N\alpha)$解出之α會相同

故上式等價於：

$$\min_{\alpha} \frac{1}{2}p^2\alpha^{T}Q\alpha - pI_N\alpha\,, \text{subject to } pA\alpha \geq C_{QP}$$

$$\min_{\alpha} \frac{1}{2}(p\alpha)^{T}Q(p\alpha) - I_N(p\alpha)\,, \text{subject to } A(p\alpha) \geq C_{QP}$$

(為避免搞混，將使用原本之 K、C 所得出之α以α_{normal}表示。)
而從以上推導可看出，$\alpha_{normal} = p\alpha$
再來看兩者所產出之g(x)：
(1) 原本之 K、C：

$$g(x) = \text{sign}\left(\sum \alpha_{normal_n} y_n K(x_n, x) + b_{normal}\right)$$

$$= \text{sign}\left(\sum \alpha_{normal_n} y_n \, \mathrm{K}(\mathrm{x_n}, \mathrm{x}) + \left(\mathrm{y_s} - \sum \alpha_{normal_n} y_n \, \mathrm{K}(\mathrm{x_n}, \mathrm{x})\right)\right)$$

(2) 使用新的$\widetilde{K}$、 $\tilde{C}$：

$$\mathrm{g}'(\mathrm{x}) = \text{sign}\left(\sum \alpha_n y_n \, \widetilde{K}(\mathrm{x_n}, \mathrm{x}) + \mathrm{b}\right)$$

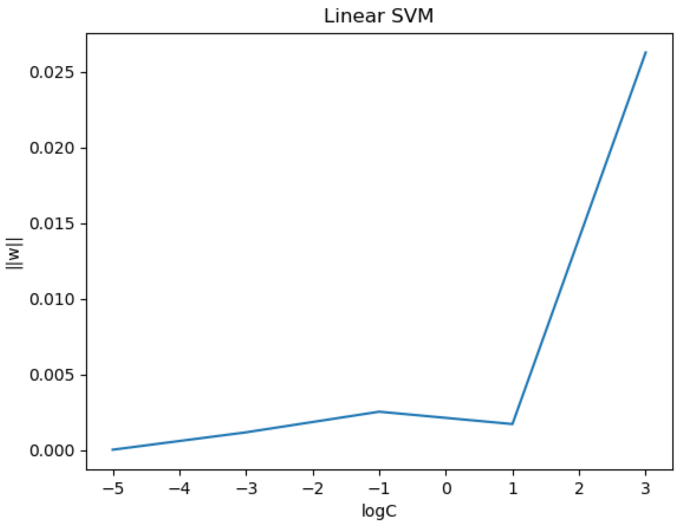
$$= \text{sign}\left(\sum \alpha_n y_n \, \widetilde{K}(\mathrm{x_n}, \mathrm{x}) + \left(\mathrm{y_s} - \sum \alpha_n y_n \, \widetilde{K}(\mathrm{x_n}, \mathrm{x})\right)\right)$$

$$= \text{sign}\left(\sum \alpha_n y_n \, \mathrm{pK}(\mathrm{x_n}, \mathrm{x}) + \left(\mathrm{y_s} - \sum \alpha_n y_n \, \mathrm{pK}(\mathrm{x_n}, \mathrm{x})\right)\right)$$

又由上面導出$\alpha_{\mathrm{normal}} = p\alpha$，故可知 g(x)=g’(x)

13.

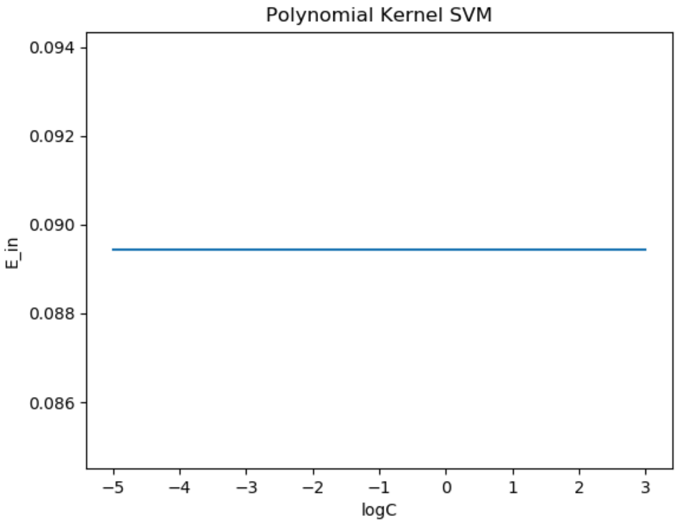
註：因原本即有安裝 scikit-learn 的套件，故以下幾小題均使用此套件完成。



隨著 C 增加，$\|w\|$ 大致走向亦是增加的，這也代表其 SVM 的寬度($\frac{1}{\|w\|}$)變小。

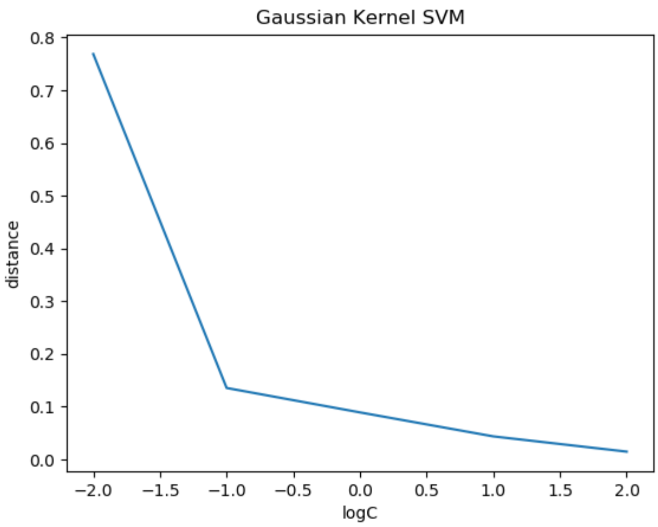
雖然有點誤差，但大致和理論相近。理論上，C 越大時，SVM 的寬度本來就會變窄，即表示$\|w\|$會遞增。

14.



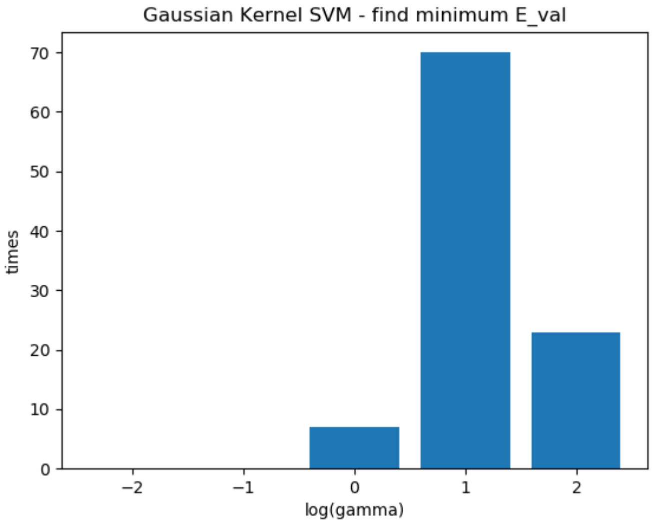
這題的結果維調整 C 並沒有使 E_in 有改變，我亦有將$\log_{10} C$調至-10, 10，但 E_in 仍然為水平線。我認為那些 error 應該是源自於雜訊，舉例而言可能有兩筆數據為(4, 2, 1)和(2, 2, 1)，故不論怎麼調整$\log_{10} C$，E_in 也不會再更好。且可能原本在 z 空間中，hard-margin 時邊界就足夠寬，沒有 overfit。故將 C 條很小也難以使其 E_in 變大。

15.



隨著 C 越大，在 Z 空間中其 SVM 的寬度會越來越窄。這和理論相符，因為 C 越大，代表其希望 violation 越少越好，故 SVM 的寬度本來就會隨 C 增加而變窄。

16.



此圖可看出若γ過小，其在 validation set 上表現並不好。而多數時候，在$\log_{10}\gamma = 1$時，其表現會最好。除此之外，因為γ過大時，會造成 overfitting，在 validation set 上的表現亦會較差，故結果亦可看出多數時候，在$\log_{10}\gamma = 2$時，表現較$\log_{10}\gamma = 1$差，代表$\log_{10}\gamma$設成 2 已經有點大了。

17.

在上課推導時，我們知道$\mathrm{w} = \sum \alpha_n y_n z_n$。
而若選擇有常數項存在的 kernel function 時，轉換式子可表示成$\phi(\mathrm{x}) = \{1, \ldots\}$。
單就看w_0，可知$\mathrm{w}_0 = \sum \alpha_n y_n z_{n0}$。其中，$\mathrm{z}_{n0}$為第 n 個 hyperplane 中的向量的第 0 項(即常數項)。
而z_{n0}是常數項可知$\forall \mathrm{n} \epsilon [1, \mathrm{N}], \mathrm{z}_{n0} = 1$，故可推得：

$$\mathrm{w}_0 = \sum \alpha_n y_n z_{n0} = \sum \alpha_n y_n * 1 = \sum \alpha_n y_n$$

又因推導時可得：$\sum \alpha_n y_n = 0$
故結合以上，可知$\mathrm{w}_0 = 0$。

18.

dual 的 QP problem 為：

$$\min_{\alpha} \frac{1}{2} \alpha^{\mathrm{T}} \mathrm{Q} \alpha - \mathrm{I}_{\mathrm{N}} \alpha$$

subject to $\mathrm{A}\alpha \geq \mathrm{C}_{\mathrm{QP}}$, 其要滿足：$\sum y_n \alpha_n = 0\,, C \geq \alpha_n \geq 0\ for\ n = 1, 2, \ldots, \mathrm{N}$

若使用的是$\widetilde{K} = K(x, x') + q$，則只有 Q 會跟原本的不同。Q'為：

$$
\mathrm{Q}' = \begin{bmatrix} y_1y_1(K(x_1,x_1)+q) & y_1y_2(K(x_1,x_2)+q) & \dots & y_1y_N(K(x_1,x_N)+q) \\ y_2y_1(K(x_2,x_1)+q) & y_2y_2(K(x_2,x_2)+q) & & y_2y_N(K(x_2,x_N)+q) \\ \dots & & & \dots \\ y_Ny_1(K(x_N,x_1)+q) & y_Ny_2(K(x_N,x_2)+q) & \dots & y_Ny_N(K(x_N,x_N)+q) \end{bmatrix}
$$

$$
= \mathrm{Q} + \mathrm{q}\mathrm{Y}\mathrm{Y}^{\mathrm{T}}
$$

故其 dual problem 可表成：(subject to 的條件不變，故不再列出)

$$
\min_{\alpha} \frac{1}{2}\alpha^{\mathrm{T}}(\mathrm{Q}+\mathrm{qYY}^{\mathrm{T}})\alpha - \mathrm{I}_{\mathrm{N}}\alpha
$$

$$
= \min_{\alpha} \frac{1}{2}\alpha^{\mathrm{T}}\mathrm{Q}\alpha + \frac{1}{2}\mathrm{q}\alpha^{\mathrm{T}}\mathrm{YY}^{\mathrm{T}}\alpha - \mathrm{I}_{\mathrm{N}}\alpha
$$

$$
= \min_{\alpha} \frac{1}{2}\alpha^{\mathrm{T}}\mathrm{Q}\alpha + \frac{1}{2}\mathrm{q}(\alpha^{\mathrm{T}}\mathrm{Y})(\alpha^{\mathrm{T}}\mathrm{Y})^{\mathrm{T}} - \mathrm{I}_{\mathrm{N}}\alpha
$$

而在推導出此形式時，有個 KKT condition 為$\sum \alpha_n y_n = 0$，即表示$\alpha^{\mathrm{T}}\mathrm{Y} = 0$。
故上式會等同於：

$$
\min_{\alpha} \frac{1}{2}\alpha^{\mathrm{T}}\mathrm{Q}\alpha - \mathrm{I}_{\mathrm{N}}\alpha
$$

其和原本之 dual problem 一模一樣。
由此可知，即使 kernel function 用$\widetilde{K} = K(x,x') + q$，結果仍和用$K(x,x')$相同。