● Hard-Margin Support Vector Machine

1. “hard-margin” decision stump

首先可以知道一定是切在x_M跟x_{M+1}之間，接著就是要讓邊界最大化，那麼分界點就是取中間值$\frac{x_M+x_{M+1}}{2}$，所以最大邊界長度就是：

$$\frac{x_M+x_{M+1}}{2}-x_M=\frac{x_{M+1}-x_M}{2}$$

2. dual problem for the uneven margin SVM

Primal 問題：

$$\begin{array}{cc} \min\limits_{\mathbf{w},b} & \frac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{suject to} & \begin{array}{l} (\mathbf{w}^T\mathbf{x}_n+b)\geq 1 \ \text{ for } \ y_n=+1 \\ -(\mathbf{w}^T\mathbf{x}_n+b)\geq \rho \ \text{ for } \ y_n=-1 \end{array} \end{array}$$

首先列出 Lagrange function：

$$\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})=\frac{1}{2}\mathbf{w}^T\mathbf{w}+\sum_{y_n=+1}\alpha_n\big(1-(\mathbf{w}^T\mathbf{x}_n+b)\big)+\sum_{y_n=-1}\alpha_n\big(\rho+(\mathbf{w}^T\mathbf{x}_n+b)\big)$$

其中所有的α_n都大於等於 0。定義：

$$\text{SVM}\equiv\min_{\mathbf{w},b}\left(\max_{\text{all }\alpha_n\geq 0}\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})\right)$$

如果資料是線性可分的，那麼：

$$\min_{\mathbf{w},b}\left(\max_{\text{all }\alpha_n\geq 0}\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})\right)=\min_{\mathbf{w},b}\frac{1}{2}\mathbf{w}^T\mathbf{w}$$

因為線性可分確保了下面兩個式子：

$$0\geq 1-(\mathbf{w}^T\mathbf{x}_n+b) \ \text{ for } \ y_n=+1$$
$$0\geq \rho+(\mathbf{w}^T\mathbf{x}_n+b) \ \text{ for } \ y_n=-1$$

加上我們限制所有的α_n都大於等於 0，取 max 後會讓Σ裡面的東西變成 0。然後可以跟 Dual SVM 講義第 9 頁的說明一樣，得到：

$$\min_{\mathbf{w},b}\left(\max_{\text{all }\alpha_n\geq 0}\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})\right)\geq\max_{\text{all }\alpha_n\geq 0}\left(\min_{\mathbf{w},b}\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})\right)$$

接著因為我們確保了三個 constraint qualification：線性的限制(linear constraints)、convex primal 跟資料是線性可分(feasible primal)，所以上面的大於小於可以是等於：

$$\min_{\mathbf{w},b}\left(\max_{\text{all }\alpha_n\geq 0}\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})\right)=\max_{\text{all }\alpha_n\geq 0}\left(\min_{\mathbf{w},b}\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})\right)$$

接著來求取 min 部分，inner problem 的 optimal；先對 b 偏微分求極值：

$$\frac{\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})}{\partial b}=\sum_{y_n=+1}-\alpha_n+\sum_{y_n=-1}\alpha_n=-\sum\alpha_n y_n=0$$

然後將這個 inner optimal 代入 Lagrange function 不會影響最佳性質：

$$\max_{\text{all } \alpha_n \geq 0 \ \sum \alpha_n y_n = 0} \left(\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{y_n=+1} \alpha_n (1 - \mathbf{w}^T\mathbf{x}_n) + \sum_{y_n=-1} \alpha_n (\rho + \mathbf{w}^T\mathbf{x}_n) \right)$$

接著對 w 偏微分求極值：

$$\frac{\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})}{\partial w_i} = w_i + \sum_{y_n=+1} -\alpha_n \mathrm{x}_{n,i} + \sum_{y_n=-1} \alpha_n \mathrm{x}_{n,i} = w_i - \sum \alpha_n y_n \mathrm{x}_{n,i} = 0$$

一樣可以將這個 inner optimal 代入 Lagrange function 不會影響最佳性質：

$$\mathbf{w} = \sum \alpha_n y_n \mathbf{x}_n$$

$$\max_{\text{all } \alpha_n \geq 0 \ \sum \alpha_n y_n = 0 \ \mathbf{w} = \sum \alpha_n y_n \mathbf{x}_n} \left(\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{y_n=+1} \alpha_n + \sum_{y_n=-1} \rho\alpha_n - \mathbf{w}^T\mathbf{w} \right)$$

$$= \max_{\text{all } \alpha_n \geq 0 \ \sum \alpha_n y_n = 0 \ \mathbf{w} = \sum \alpha_n y_n \mathbf{x}_n} \left(\min_{\mathbf{w},b} -\frac{1}{2}\left\|\sum \alpha_n y_n \mathbf{x}_n\right\|^2 + \sum_{y_n=+1} \alpha_n + \sum_{y_n=-1} \rho\alpha_n \right)$$

列出 primal-dual optimal 的四個 KKT 條件：

- primal feasible $\begin{array}{l}(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \ \text{ for } \ y_n = +1 \\ -(\mathbf{w}^T\mathbf{x}_n + b) \geq \rho \ \text{ for } \ y_n = -1\end{array}$
- dual feasible $\alpha_n \geq 0$
- dual-inner optimal $\sum \alpha_n y_n = 0, \ \mathbf{w} = \sum \alpha_n y_n \mathbf{x}_n$
- primal-inner optimal $\begin{array}{l}\alpha_n\left(1 - (\mathbf{w}^T\mathbf{x}_n + b)\right) = 0 \ \text{ for } \ y_n = +1 \\ \alpha_n\left(\rho + (\mathbf{w}^T\mathbf{x}_n + b)\right) = 0 \ \text{ for } \ y_n = -1\end{array}$

因此最後我們得到我們 SVM 的 dual problem：

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\left\|\sum \alpha_n y_n \mathbf{x}_n\right\|^2 + \sum_{y_n=+1} \alpha_n + \sum_{y_n=-1} \rho\alpha_n$$

$$\text{subject to } \sum \alpha_n y_n = 0$$
$$\alpha_n \geq 0$$

或著改成取 min：

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_n \sum_m \alpha_n \alpha_m y_n y_m \mathbf{x}_n \mathbf{x}_m - \sum_{y_n=+1} \alpha_n - \sum_{y_n=-1} \rho\alpha_n$$

$$\text{subject to } \sum_n \alpha_n y_n = 0$$
$$\alpha_n \geq 0$$

3. Converting without solving the QP problem

在此我先證明，even margin SVM 得到的超平面，其兩側必定都存在至少一個支持向量；這裡我使用反證法：

假設 SVM 最後求得的$\mathbf{w}$跟 b，有下列的情況：

$$\min_{\mathbf{x}_n}(\mathbf{w}^T\mathbf{x}_n + b) = 1 \text{ for } y_n = +1$$

$$\min_{\mathbf{x}_n}(\mathbf{w}^T\mathbf{x}_n + b) < -1 \text{ for } y_n = -1$$

也就是$y_n = -1$的資料中最小值並不在邊界上而在邊界外。

為了方便說明，我將$+1$跟-1的最小值的資料分別叫做$\mathbf{x}_{+1}$跟$\mathbf{x}_{-1}$，並且令$\mathbf{w}^T\mathbf{x}_{-1} + b$得到的值叫做 $\rho < -1$：

$$\mathbf{w}^T\mathbf{x}_{+1} + b = 1$$

$$\mathbf{w}^T\mathbf{x}_{-1} + b = \rho$$

這兩個式子也代表了夾住原本超平面的兩個邊界超平面：

$$\mathbf{w}^T\mathbf{x} + b = 1$$

$$\mathbf{w}^T\mathbf{x} + b = \rho$$

這時我們來去找，在相同的法向量$\mathbf{w}$還有b的情況下，哪個超平面可以和那兩個邊界超平面之間的距離是 1 比 1：

我們想要找的超平面：

$$\mathbf{w}^T\mathbf{x} + b = \alpha$$

算出離兩個邊界超平面的垂直距離：

$$1-\alpha : \alpha - \rho = 1 : 1$$

$$\Rightarrow \alpha - \rho = 1 - \alpha$$

$$\Rightarrow 2\alpha = 1 + \rho$$

$$\Rightarrow \alpha = \frac{1+\rho}{2}$$

這個超平面就是我們新的最佳超平面，但是要做一些調整：

$$\mathbf{w}^T\mathbf{x} + b = \alpha = \frac{1+\rho}{2}$$

$$\Rightarrow \mathbf{w}^T\mathbf{x} + b - \alpha = 0 = \mathbf{w}^T\mathbf{x} + b - \frac{1+\rho}{2}$$

我令這個新的截距部分叫做$b_{opt} = b - \alpha$：

$$\mathbf{w}^T\mathbf{x} + b_{opt} = 0$$

雖然此時這個新的超平面跟那兩個邊界超平面距離是 1 比 1，但$\mathbf{w}^T\mathbf{x} + b$的部分不是真的等於 1：

這是$+1$資料方向的邊界超平面：$\mathbf{w}^T\mathbf{x} + b = 1$

$$\Rightarrow \mathbf{w}^T\mathbf{x} + b - \frac{1+\rho}{2} = 1 - \frac{1+\rho}{2}$$

$$\Rightarrow \mathbf{w}^T\mathbf{x} + b_{opt} = \frac{1-\rho}{2}$$

這是$+1$資料方向的邊界超平面：$\mathbf{w}^T\mathbf{x}+b=\rho$

$$\Rightarrow \mathbf{w}^T\mathbf{x}+b-\frac{1+\rho}{2}=\rho-\frac{1+\rho}{2}$$

$$\Rightarrow \mathbf{w}^T\mathbf{x}+b_{opt}=\frac{\rho-1}{2}$$

可以觀察到他們確實是 1 比 1。所以接著要對$\mathbf{w}$進行調整，同除以$\frac{1-\rho}{2}$：

$$\frac{2}{1-\rho}(\mathbf{w}^T\mathbf{x}+b_{opt})=1 \text{ 這是}+1\text{資料方向的邊界超平面}$$

$$\frac{2}{1-\rho}(\mathbf{w}^T\mathbf{x}+b_{opt})=-1 \text{ 這是}-1\text{資料方向的邊界超平面}$$

$$\frac{2}{1-\rho}(\mathbf{w}^T\mathbf{x}+b_{opt})=0 \text{ 這是新的最佳超平面}$$

而這個新得到的最佳法向量(權重)：

$$\frac{2}{1-\rho}\mathbf{w}$$

因為ρ是小於-1的負數，因此可以知道$\frac{2}{1-\rho}<1$，所以可以知道這個新的最佳法向量比原本的更好，造成矛盾。

因此可以知道 SVM 求得的最佳超平面，其兩側必定有支持向量存在。

上面的推論過程，剛好可以拿來計算此題提到的轉換。

原本$(b_1^*,\mathbf{w}_1^*)$是從 even margin 得到的最佳超平面，如果想要得到$(b_{1126}^*,\mathbf{w}_{1126}^*)$，方法就跟上面一樣，找到一個新的超平面，他跟原本的兩個邊界超平面之間的距離變成 1 比 1126，所以上面推論過程中，比例的部分就會變成(為了更 general 我將 1126 換成了β)：

$$1-\alpha：\alpha-(-1)=1：\beta$$
$$\Rightarrow \alpha+1=\beta-\beta\alpha$$
$$\Rightarrow \alpha(1+\beta)=\beta-1$$
$$\Rightarrow \alpha=\frac{\beta-1}{\beta+1}$$

$b_{opt}=b_1^*-\alpha=b_1^*-\frac{\beta-1}{\beta+1}$；接著我們可以確認和兩個邊界超平面的距離的確是 1 比 β：

$$1-\alpha=1-\frac{\beta-1}{\beta+1}=\frac{2}{\beta+1}$$

$$\alpha-(-1)=\frac{\beta-1}{\beta+1}+1=\frac{2\beta}{\beta+1}$$

並且可知需要同除以的係數就是$\frac{2}{\beta+1}$，恰好對應上面推導時的$\frac{1-\rho}{2}$：

$$\frac{\beta+1}{2}(\mathbf{w}_1^{*T}\mathbf{x}+b_{opt})=0$$

所以可以知道$\mathbf{w}_\beta^*=\frac{\beta+1}{2}\mathbf{w}_1^*$，$b_\beta^*=\frac{\beta+1}{2}(b_1^*-\frac{\beta-1}{\beta+1})=\frac{\beta+1}{2}b_1^*-\frac{\beta-1}{2}$。

將 1126 代入即可：$\mathbf{w}_{1126}^*=\frac{1127}{2}\mathbf{w}_1^*$，$b_{1126}^*=\frac{1127}{2}b_1^*-\frac{1125}{2}$。

4. optimal solution of the uneven margin

我認為$\boldsymbol{\alpha}_1^*$並不會是任何$\boldsymbol{\alpha}_\rho^*$的最佳解，原因在於 dual-inner optimal：

$$\sum \alpha_n y_n = 0 \ , \mathbf{w}=\sum \alpha_n y_n \mathbf{x}_n$$

從第三題的推論可以知道$\mathbf{w}_\beta^*=\frac{\beta+1}{2}\mathbf{w}_1^*$，只差了常數倍，所以可以知道

$\sum \boldsymbol{\alpha}_{1,\mathrm{n}}^* y_n$跟$\sum \boldsymbol{\alpha}_{\rho,\mathrm{n}}^* y_n$應該會差常數倍，而不是等於的關係，換句話說就是$\boldsymbol{\alpha}_1^*$跟$\boldsymbol{\alpha}_\rho^*$應該是差常數倍，而不是等於。(這裡我的$\beta$就是$\rho$)

● Operation of Kernels

5. multiplication of valid kernels is still a valid kernel

這裡我令：

$$\phi_1(\mathbf{x})=(z_1,z_2,\dots,z_n,\dots)$$
$$\phi_2(\mathbf{x})=(k_1,k_2,\dots,k_m,\dots)$$

所以可以知道：

$$K(\mathbf{x},\mathbf{x}')=K_1(\mathbf{x},\mathbf{x}')K_2(\mathbf{x},\mathbf{x}')$$
$$=\phi_1(\mathbf{x})^T\phi_1(\mathbf{x}')\cdot\phi_2(\mathbf{x})^T\phi_2(\mathbf{x}')$$

$$=\left(z_1z_1'+\cdots+z_nz_n'+\cdots\right)\cdot\left(k_1k_1'+\cdots+k_mk_m'+\cdots\right)$$
$$=z_1z_1'k_1k_1'+\cdots+z_nz_n'k_1k_1'+\cdots+z_1z_1'k_mk_m'+\cdots+z_nz_n'k_mk_m'+\cdots$$
$$=z_1k_1z_1'k_1'+\cdots+z_nk_1z_n'k_1'+\cdots+z_1k_mz_1'k_m'+\cdots+z_nk_mz_n'k_m'+\cdots$$

$$=(z_1k_1,\dots,z_nk_1,\dots,z_1k_m,\dots,z_nk_m,\dots)^T\left(z_1'k_1',\dots,z_n'k_1',\dots,z_1'k_m',\dots,z_n'k_m',\dots\right)$$

所以可知新的轉換就是：

$$\phi(\mathbf{x})=(z_1k_1,\dots,z_nk_1,\dots,z_1k_m,\dots,z_nk_m,\dots)$$

6. distances in the Z space

距離公式(平方)是：

$$\|\phi(\mathbf{x})-\phi(\mathbf{x}')\|^2=\phi(\mathbf{x})^T\phi(\mathbf{x})+\phi(\mathbf{x}')^T\phi(\mathbf{x}')-2\phi(\mathbf{x})^T\phi(\mathbf{x}')$$

使用我們的傳統藝能 kernel trick：

$$\Rightarrow (1+\mathbf{x}^{\boldsymbol{T}}\mathbf{x})^2+\left(1+\mathbf{x}'^{\boldsymbol{T}}\mathbf{x}'\right)^2-2(1+\mathbf{x}^{\boldsymbol{T}}\mathbf{x}')^2$$

$$=(1+1)^2+(1+1)^2-2\big(1+\|\mathbf{x}\|\|\mathbf{x}'\|cos(\theta)\big)^2$$

$$=8-2\big(1+cos(\theta)\big)^2$$

對θ微分求極值：

$$-4\big(1+cos(\theta)\big)\big(-sin(\theta)\big)=0$$

$$sin(\theta)\big(1+cos(\theta)\big)=0$$

所以極值發生在$\theta=0°$跟$\theta=180°$的時候，代回去原本的函數可得：

$$8-2\big(1+cos(0°)\big)^2=8-2(1+1)^2=0$$

$$8-2\big(1+cos(180°)\big)^2=8-2(1+-1)^2=8$$

別忘記最後還要開個根號：

最大距離為$2\sqrt{2}$

最小距離為 0

7. Gaussian kernel

$$\tilde{\phi}(x)=\left(1,\sqrt{\frac{2}{1!}}x,\sqrt{\frac{2^2}{2!}}x^2,\sqrt{\frac{2^3}{3!}}x^3,\dots\right)$$

$$\left\|\tilde{\phi}(x)\right\|^2=1+\frac{2}{1!}x^2+\frac{2^2}{2!}x^4+\frac{2^3}{3!}x^6+\cdots=e^{2x^2}$$

$$\Rightarrow \|\tilde{\phi}(x)\|=e^{x^2}$$

$$\Rightarrow \frac{e^{-x^2}}{\|\tilde{\phi}(x)\|}=\frac{e^{-x^2}}{e^{x^2}}=1$$

8. cosine

$$cos(\mathbf{x},\mathbf{x}')=\frac{\mathbf{x}^{\boldsymbol{T}}\mathbf{x}'}{\|\mathbf{x}\|\|\mathbf{x}'\|}$$

所以我們可以構造出一個轉換：

$$\phi_{cos}(\mathbf{x})=\frac{\mathbf{x}}{\|\mathbf{x}\|}$$

這樣一來：

$$cos(\mathbf{x},\mathbf{x}')=\phi_{cos}(\mathbf{x})^T\phi_{cos}(\mathbf{x}')$$

● Experiments with Soft-Margin Support Vector Machine

9. smallest number of support vectors

   根據 libsvm 的執行結果：

   # (0.1,2):860 # (0.1,3):789 # (0.1,4):740

   # (1,2):783 # (1,3):721 # (1,4):666

   # (10,2):712 # (10,3):659 # (10,4):629

   可以發現(10,4)具有最少的支持向量。

10. lowes E_{out}

   我依序傳入[0.01,0.1,1,10,100]，根據 libsvm 的執行結果：

   # Accuracy = 95.4% (1908/2000) (classification)

   # 0.01:95.39999999999999

   # Accuracy = 98.8% (1976/2000) (classification)

   # 0.1:98.8

   # Accuracy = 99.5% (1990/2000) (classification)

   # 1:99.5

   # Accuracy = 99.4% (1988/2000) (classification)

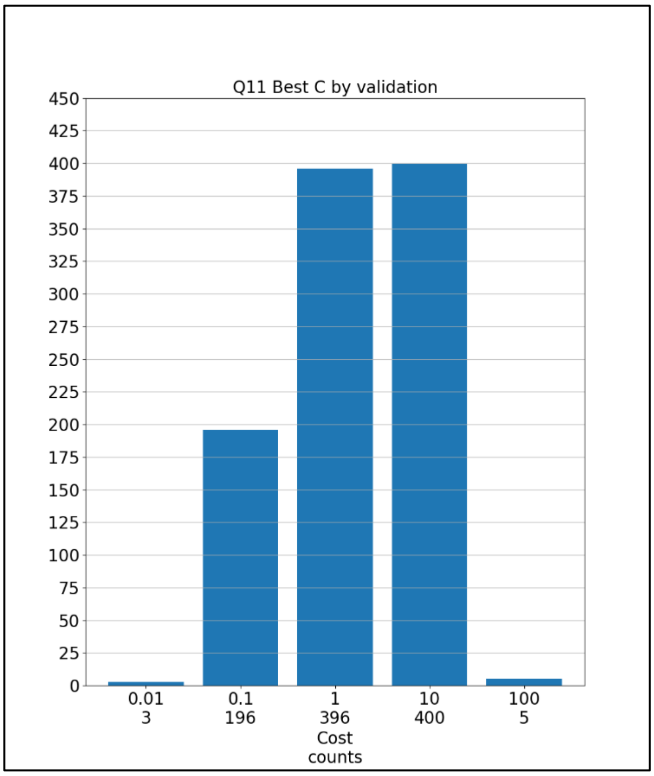
   # 10:99.4

   # Accuracy = 99.45% (1989/2000) (classification)

   # 100:99.45

   E_{out}最低的是$C = 1$。

11. times of each C is selected



可以看到在 1000 次實驗中，E_{val}最低大多落在$C = 1$跟$C = 10$。
$C = 100$會這麼低的原因是對錯誤的容忍太低了，導致超平面變得太複雜，發生 overfitting。$C = 0.01$會這麼低則是因為對錯誤的容忍度太高了，大多都是錯的，而$C = 0.1$則是逐漸提高準確率，並在$C = 10$的時候達到最高……。

正當我想這樣敘述的時候，仔細回想第十題的正確率：

\# Accuracy = 95.4% (1908/2000) (classification)
\# 0.01:95.39999999999999
\# Accuracy = 98.8% (1976/2000) (classification)
\# 0.1:98.8
\# Accuracy = 99.5% (1990/2000) (classification)
\# 1:99.5
\# Accuracy = 99.4% (1988/2000) (classification)
\# 10:99.4
\# Accuracy = 99.45% (1989/2000) (classification)
\# 100:99.45

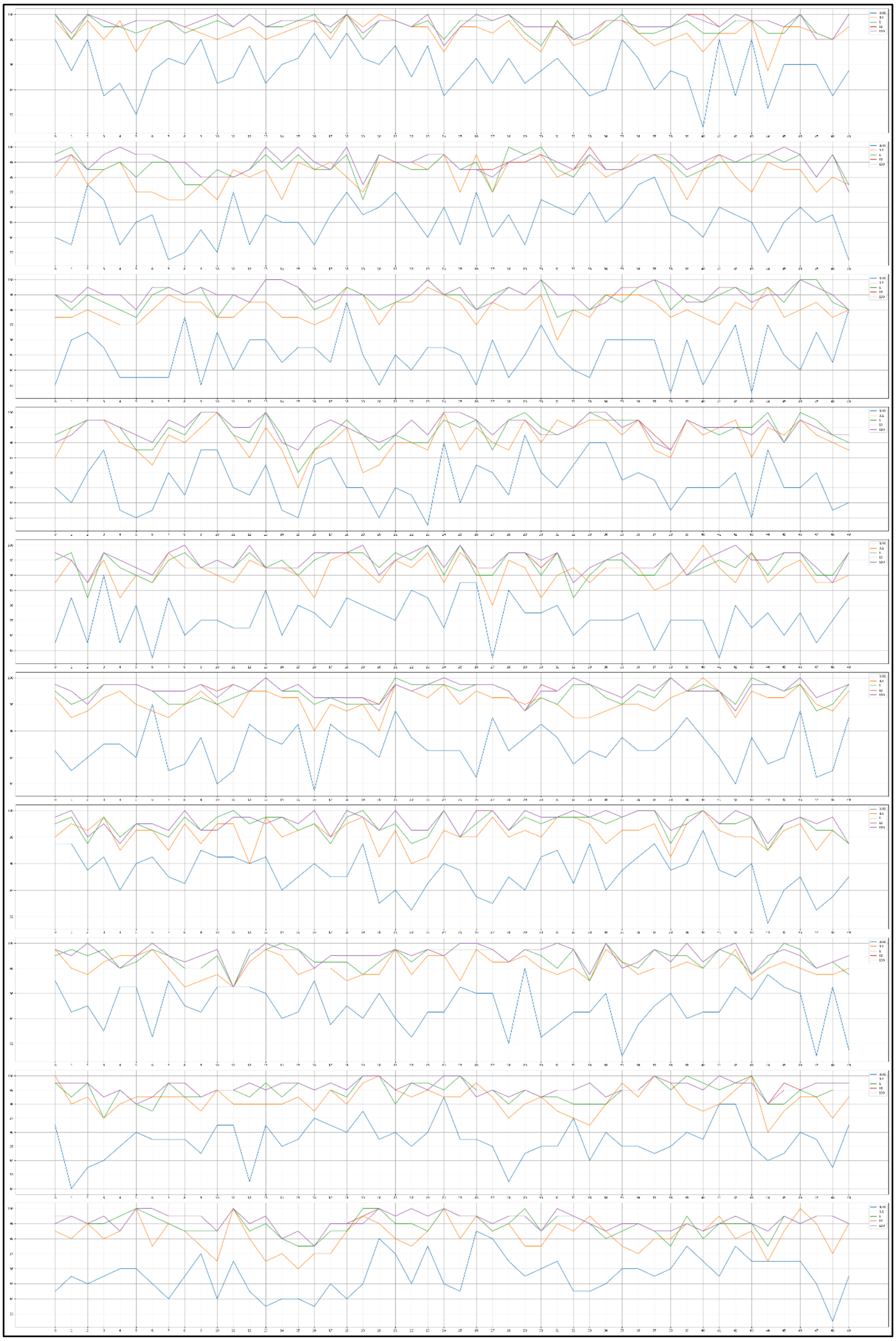
發現其實他們準確率都很高，尤其$C = 1$、10、100這三個非常接近，於是我就想有沒有可能是因為每次的模擬，$C = 100$ 跟 $C = 10$ 的正確率一樣，但是因為我們都是挑數值小的，所以導致$C = 100$被選的次數才這麼低，因此我將每次實驗得到的數據做出下面的圖表。(在下一頁)

可以看到確實如此，$C = 100$跟$C = 10$甚至跟$C = 1$，三者在模擬中的表現有很多時候都是一樣高的，尤其$C = 100$，他就只有 5 筆數據是準確率大於其他人，其他都是等於跟一點點的小於，導致他不容易被選到。

這時結合 11 題，E_{out}最低的是$C = 1$，我認為是因為模擬的過程中$C = 10$跟$C = 100$出現了「一點點」的 overfitting：由於我們都是從同個資料集合中隨機抽 200 個出來作為驗證資料，經過這 1000 次模擬下來，就好像是做了很多遍的 cross validation，而每次準確率都可以很高，代表模型在這樣手上有的資料的分布可以有良好的泛化能力，不論是怎樣的相同取樣方法，都不會發生 overfitting；不過如果測試資料有稍微不一樣的分佈，那麼預測可能就會稍微不準，也就是我認為 11 題發生的狀況。

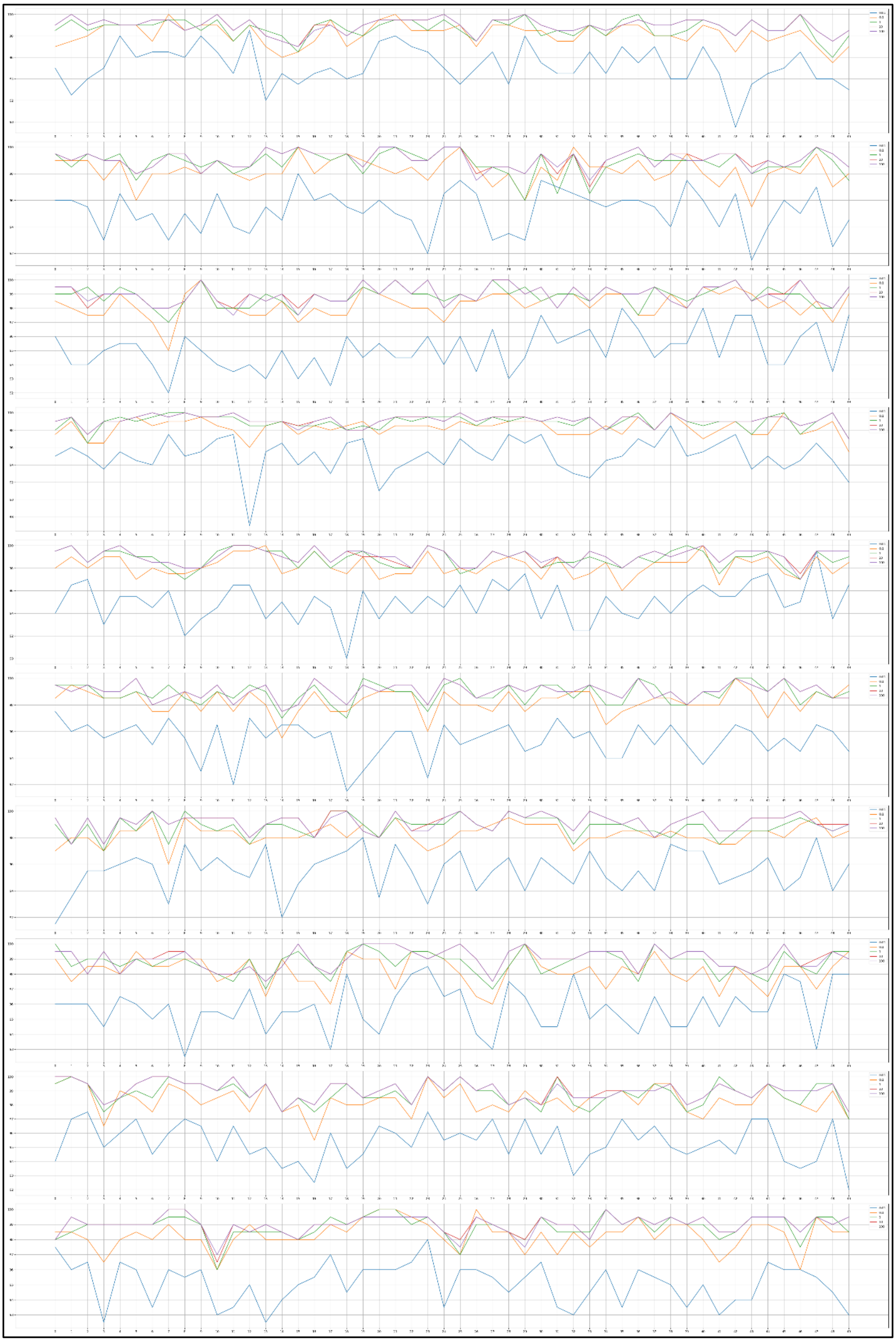
但是 overfitting 的狀況其實真的不嚴重，從那非常高的準確率就可以知道資料是有一定的分佈存在的。

這是前 500 筆數據



這是前 500 筆數據

藍色是 0.01，橘色是 0.1，綠色是 1，紅色是 10，紫色是 100。

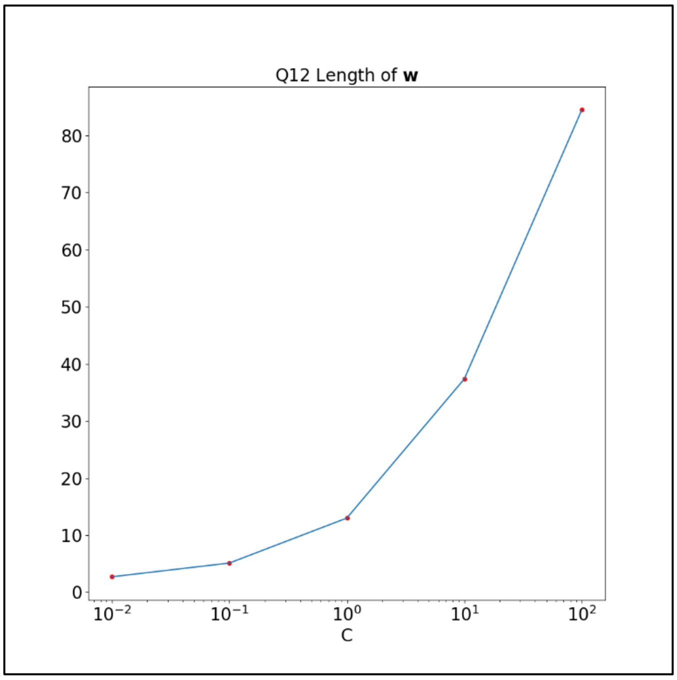


這是後 500 筆數據

藍色是 0.01，橘色是 0.1，綠色是 1，紅色是 10，紫色是 100。

可以明顯看到綠色紅色跟紫色非常頻繁的重疊在一起，那麼此時就會選小的。

12. $\|\mathbf{w}\|$



可以發現當 C 越大，$\|\mathbf{w}\|$就越大。C 越大代表我們對錯誤量的容忍程度越低，而$\|\mathbf{w}\|$越大代表 SVM 的 margin 越小。

因為我們對錯誤量的容忍程度越低，會需要讓超平面更靠近犯錯的資料，所以才會導致 SVM 的 margin 變小，導致$\|\mathbf{w}\|$變大。

- Bonus

13. Dual of Dual

先列出原本的 hard margin SVM dual：

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{\boldsymbol{n}}\sum_{\boldsymbol{m}} \alpha_n\alpha_m y_n y_m \mathbf{x}_n\mathbf{x}_m - \sum_n \alpha_n$$

$$\text{subject to } \sum_n \alpha_n y_n = 0$$

$$\alpha_n \geq 0$$

首先建構出 Lagrange function：

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma_1, \gamma_2) =$$

$$\frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m \mathbf{x}_n\mathbf{x}_m - \sum_n \alpha_n + \sum_n \beta_n(-\alpha_n) + \gamma_1\left(\sum_n \alpha_n y_n\right) + \gamma_2\left(-\sum_n \alpha_n y_n\right)$$

然後經過跟講義一樣的步驟，可以得到 Lagrange dual：

$$\max_{\text{all } \beta_n, \gamma_1, \gamma_2 \geq 0}\left(\min_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma_1, \gamma_2)\right)$$

接著求 inner optimal，對α_i偏微分求極值：

$$\frac{\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma_1,\gamma_2)}{\partial\alpha_i}=\sum_n \alpha_n y_i y_n \mathbf{x}_i\mathbf{x}_n-1-\beta_i+\gamma_1 y_i+\gamma_2(-y_i)=0$$

$$\Rightarrow \beta_i=\sum_n \alpha_n y_i y_n \mathbf{x}_i\mathbf{x}_n-1+\gamma_1 y_i+\gamma_2(-y_i)$$

將 inner optimal 代回原 lagrange function 當中的$\sum_n \beta_n(-\alpha_n)$，可得：

$$\sum_n \beta_n(-\alpha_n)=\sum_n\left(\sum_m \alpha_m y_n y_m \mathbf{x}_n\mathbf{x}_m-1+\gamma_1 y_n+\gamma_2(-y_n)\right)(-\alpha_n)$$

$$=-\sum_n\sum_m \alpha_n\alpha_m y_n y_m \mathbf{x}_n\mathbf{x}_m+\sum_n \alpha_n-\sum_n \alpha_n\gamma_1 y_n+\sum_n \alpha_n\gamma_2 y_n$$

然後再代回原本 lagrange function 當中，會發現很多東西都削掉了：

$$\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\gamma_1,\gamma_2)=-\frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m \mathbf{x}_n\mathbf{x}_m$$

所以我們就得到了 dual 的 dual problem：

$$\max_{\boldsymbol{\beta},\gamma_1,\gamma_2}\ \min_{\boldsymbol{\alpha}}\ -\frac{1}{2}\sum_n\sum_m \alpha_n\alpha_m y_n y_m \mathbf{x}_n\mathbf{x}_m$$

$$\text{subject to }\ \beta_i=\sum_n \alpha_n y_i y_n \mathbf{x}_i\mathbf{x}_n-1+\gamma_1 y_i+\gamma_2(-y_i)$$

$$\text{all }\ \beta_n,\gamma_1,\gamma_2\geq 0$$

如果用 $\mathbf{w}=\sum\alpha_n y_n \mathbf{x}_n$ 進行替換可以得到

$$\max_{\boldsymbol{\beta},\gamma_1,\gamma_2\ \mathbf{w}=\sum\alpha_n y_n\mathbf{x}_n}\ -\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

$\boldsymbol{\beta},\gamma_1,\gamma_2$都不見了，所以可以不用寫上了。如果換成取 min，我們就會得到：

$$\min_{\mathbf{w}}\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

列出 KKT 條件：

- primal feasible $\sum_n \alpha_n y_n=0$ 跟 $\alpha_n\geq 0$
- dual feasible $\beta_n,\gamma_1,\gamma_2\geq 0$
- dual-inner optimal $\beta_i=\sum_n \alpha_n y_i y_n \mathbf{x}_i\mathbf{x}_n-1+\gamma_1 y_i+\gamma_2(-y_i)$
  $\sum_n \beta_n(-\alpha_n)=\mathbf{0}$
- primal-inner optimal $\gamma_1\left(\sum_n \alpha_n y_n\right)=\mathbf{0}$
  $\gamma_2\left(-\sum_n \alpha_n y_n\right)=\mathbf{0}$

可以發現形式就跟以前的 primal 一樣，所以只要補回 primal 的限制就會得到 SVM 的 primal problem：

$$y_n(\mathbf{w}^T\mathbf{x}_n+b)\geq 1$$

但是還有其背後的一堆 KKT 條件。