HW5

1.

因為 data 是 1 維，所以 SVM 的形式是$\mathrm{w} * \mathrm{x} + \mathrm{b}$。因為分界線在$\mathrm{x_m}$跟$\mathrm{x_{m+1}}$之間，所以根據 SVM 定義可知這條分界線的 margin=min(線與$\mathrm{x_m}$的距離,線與$\mathrm{x_{m+1}}$的距離)<=($\mathrm{x_{m+1}} - \mathrm{x_m})/2$。

2.

根據題意，Lagrange Function 是$\mathcal{L}(\mathrm{b}, \mathrm{w}, \alpha) = \frac{1}{2}\mathrm{w}^{\mathrm{T}} w + \sum_{n \in y_n = 1} \alpha_n (1 - y_n(w^T x_n + b)) + \sum_{n \in y_n = -1} \alpha_n \big(\rho - y_n(w^T x_n + b)\big)$ 。

下面跟講義一樣，為了避免 violate constraint，所以定所有α都必須>=0。$\min_{b,w}(\max_{all\ \ \alpha \geq 0} \mathcal{L}(\mathrm{b}, \mathrm{w}, \alpha)) \geq \max_{all\ \ \alpha \geq 0}(\min_{b,w}(\mathcal{L}(\mathrm{b}, \mathrm{w}, \alpha)))$，所以展開就是$\max_{all\ \ \alpha \geq 0}(\min_{b,w}(\frac{1}{2}\mathrm{w}^{\mathrm{T}} w + \sum_{n \in y_n = 1} \alpha_n (1 - y_n(w^T x_n + b)) + \sum_{n \in y_n = -1} \alpha_n \big(\rho - y_n(w^T x_n + b)\big)))$

接著對上式做微分。對 b 微分後，得到$\sum_{n=1}^{N} \alpha_n\, y_n = 0$。所以上式變成$\max_{all\ \ \alpha \geq 0}(\min_{b,w}(\frac{1}{2}\mathrm{w}^{\mathrm{T}} w + \sum_{n \in y_n = 1} \alpha_n (1 - y_n(w^T x_n)) + \sum_{n \in y_n = -1} \alpha_n \big(\rho - y_n(w^T x_n)\big)))$。然後對每個$\mathrm{w_i}$微分，可得$\mathrm{w_i} = \sum_{n=1}^{N} \alpha_n\, x_n y_n$，所以上式變成$\max_{all\ \ \alpha \geq 0}(\frac{-1}{2}\|\sum_{n=1}^{N} \alpha_n\, x_n y_n\|^2 + \sum_{n \in y_n = 1} \alpha_n + \sum_{n \in y_n = -1} \alpha_n \rho)$。

稍微整理一下，就變成$\max_{all\ \ \alpha \geq 0}(\frac{-1}{2}\|\sum_{n=1}^{N} \alpha_n\, x_n y_n\|^2 + \sum_{n \in y_n = 1} \alpha_n + \sum_{n \in y_n = -1} \alpha_n \rho)$。限制有$\sum_{n=1}^{N} \alpha_n\, y_n = 0$、$\mathrm{w_i} =$

$\sum_{n=1}^{N} \alpha_n x_n y_n$。

3.

因w_1, b_1是正常情況的解。所以對於任意$\mathrm{x_i}, y_i$有 4 種情況:
$\mathrm{w^T} x_i + b = 1 \ \text{ if } \ \mathrm{y_n} = 1$、$\mathrm{w^T} x_i + b > 1 \ \text{ if } \ \mathrm{y_n} = 1$、$\mathrm{w^T} x_i + b = -1 \ \text{ if } \ \mathrm{y_n} = -1$、$\mathrm{w^T} x_i + b < -1 \ \text{ if } \ \mathrm{y_n} = -1$。這裡構造出一個解:$\mathrm{w}_{1126} = \frac{1127}{2} w_1$，$\mathrm{b}_{1126} = \frac{1127}{2} b_1 - \frac{1125}{2}$。接下來證明這是對的。

第一種情況，if $\mathrm{y_n} = 1$，$\mathrm{w^T} x_i + b = 1 \rightarrow \mathrm{w}_{1126}^{\mathrm{T}} x_i + b_{1126} = \frac{1127}{2} w_1^T x_i + \frac{1127}{2} b_1 - \frac{1125}{2} = 1$。第二種情況，if $\mathrm{y_n} = 1$，$\mathrm{w^T} x_i + b > 1 \rightarrow \mathrm{w}_{1126}^{\mathrm{T}} x_i + b_{1126} = \frac{1127}{2} w_1^T x_i + \frac{1127}{2} b_1 - \frac{1125}{2} > 1$。第三種情況，if $\mathrm{y_n} = -1$，$\mathrm{w^T} x_i + b = -1 \rightarrow \mathrm{w}_{1126}^{\mathrm{T}} x_i + b_{1126} = \frac{1127}{2} w_1^T x_i + \frac{1127}{2} b_1 - \frac{1125}{2} = -1126$。第四種情況，if $\mathrm{y_n} = -1$，$\mathrm{w^T} x_i + b < -1 \rightarrow \mathrm{w}_{1126}^{\mathrm{T}} x_i + b_{1126} = \frac{1127}{2} w_1^T x_i + \frac{1127}{2} b_1 - \frac{1125}{2} < -1126$。

以上證明了w_1, b_1的四種情況可以對應到$\mathrm{w}_{1126}, b_{1126}$，而且驗證結果符合$\mathrm{w}_{1126}, b_{1126}$的定義，所以答案是$\mathrm{w}_{1126} = \frac{1127}{2} w_1$，$\mathrm{b}_{1126} = \frac{1127}{2} b_1 - \frac{1125}{2}$。

4.

不是。因為根據前面的結果，$\mathrm{w_i} = \sum_{n=1}^{N} \alpha_n \mathrm{y_n} x_n$，如果$\rho =$ 1126 跟 1時的α^*是一樣的，那麼$\mathrm{w_1}$跟$\mathrm{w_{1126}}$的結果也應該是一樣的。但上面的答案是$\mathrm{w_{1126}} = \frac{1127}{2} w_1$，所以$\alpha_1^*$跟$\alpha_{1126}^*$並不一樣。

5.

$\mathrm{K_1(x, x')} = \Phi_1(x)^T \Phi_1(x'),\ \mathrm{K_2(x, x')} = \Phi_2(x)^T \Phi_2(x')$

令向量$\Phi_1(x) = [\Phi^1(x_1), \Phi^1(x_2), \dots, \Phi^1(x_M)]^\mathrm{T}$，向量$\Phi_2(x) = [\Phi^2(x_1), \Phi^2(x_2), \dots, \Phi^2(x_K)]^\mathrm{T}$

→$\mathrm{K_1(x, x')K_2(x, x')} = \Phi_1(x)^T \Phi_1(x') \Phi_2(x)^T \Phi_2(x') = \sum_{i=1}^{M} \sum_{j=1}^{K} \Phi^1(x_i)\Phi^1(x_i')\Phi^2(x_j)\Phi^2(x_j') = \sum_{i=1}^{M} \sum_{j=1}^{K} \left[\Phi^1(x_i)\Phi^2(x_j)\right]\left[\Phi^1(x_i')\Phi^2(x_j')\right]$。令$\Phi_3(x)$是個向量，$\Phi_3(x)_{ij} = \Phi^1(x_i)\Phi^2(x_j)$，把所有 i 跟 j 組合的$\Phi_3(x)_{ij}$排成一個向量就是$\Phi_3(x)$，因此$\mathrm{K_1(x, x')K_2(x, x')} = \Phi_3(x)^T \Phi_3(x)$。所以$\mathrm{K_1(x, x')K_2(x, x')}$是 valid kernel。

6.

這個 k 對應的γ是 1，所以在 Z space，零次項的係數是 1，一次項的係數是$\sqrt{2}$，二次項的係數是 1。考慮說 x 跟 x’都是單位向量，也就是說向量中只有一個數字是 1，其他都是 0。這時有 2 種情況，x=x’ or x!=x’。當 x=x’時，在 Z space 中對應的向量也是一樣的，所以距離是 0。因為距離是根號值必為正，所以最小值是 0。當 x!=x’

時，令 x 和 x′中是大小是 1 的那個數值是$\mathrm{x_i}$跟$\mathrm{x'_j}$，考慮說在 Z space 中，只有$\mathrm{x_i}$跟$\mathrm{x'_j}$本身的次方不為 0，其他因為乘上項量中的其他項所以都是 0。次方為 2，所以只有零次項是 1，一次項$\sqrt{2}\mathrm{x_i}$跟$\sqrt{2}\mathrm{x'_j}$，二次項$\mathrm{x_i^2}$跟$\mathrm{x'_j}^2$。考慮說$\mathrm{x_i}$跟$\mathrm{x'_j}$的大小都是 1，所以$\mathrm{x_i^2} = \mathrm{x'_j}^2$，距離為 0，所以有距離的只有一次項$\sqrt{2}\mathrm{x_i}$跟$\sqrt{2}\mathrm{x'_j}$。如果 i!=j，則距離是 $\sqrt{\left(\sqrt{2}(\mathrm{x_i}-0)\right)^2+\left(\sqrt{2}(\mathrm{x'_j}-0)\right)^2} = 2$。當 i=j，則距離是 $|\sqrt{2}(\mathrm{x_i}-\mathrm{x'_j})|$，當 $\mathrm{x'_j}$跟$\mathrm{x_i}$同號時是 0，異號時是$2\sqrt{2}$。所以最大值是$2\sqrt{2}$，最小值是 0。

7.

題目所對應的 kernel 是$\mathrm{K(x,x')} = \exp(-(\mathrm{x}-\mathrm{x'})^2)$。當 x=x′時，$\mathrm{K(x,x)} = \exp(-(\mathrm{x}-\mathrm{x})^2) = 1 = \Phi(x)^T\Phi(x)$。也就是說$\left||\Phi(x)|\right| = 1$。也就是說，$\left||\exp(-\mathrm{x}^2) * \widetilde{\Phi}(x)|\right| = 1$，也就是說，$\frac{1}{\left||\widetilde{\Phi}(x)|\right|} = \exp(-\mathrm{x}^2)$。

8.

Cosine similarity 是$\mathrm{K(x,x')} = \frac{x^T x'}{\|x\|\|x'\|} = (\frac{x}{\|x\|})(\frac{x'}{\|x'\|})$，可以構造一個$\Phi(x) = (\frac{x}{\|x\|})$，使$\mathrm{K(x,x')} = \Phi(x)^T\Phi(x')$。所以 Cosine similarity 是 kernel。

9.

首先用 svm_read_problem 讀取資料，把 label 不是 4 的 label 改成-1，label 是 4 的改成 1。用 svm_train 去訓練 model，參數設定如下:-s=0 代表是有 C constraint 的 SVM，-t=1 代表是 1+rx^Tx 的 Q 次方這種形式。用-c 設定 C，-d 設定 Q。-g=1 設定 gamma=1，-r=1 設定常數是 1。接著用 get_nr_sv 得到 model 的 support vectors 的數量。

結果如下:

C:0.1, Q:2, num_sv=860
C:0.1, Q:3, num_sv=789
C:0.1, Q:4, num_sv=740
C:1, Q:2, num_sv=783
C:1, Q:3, num_sv=721
C:1, Q:4, num_sv=666
C:10, Q:2, num_sv=712
C:10, Q:3, num_sv=659
C:10, Q:4, num_sv=629

所以最小的 num_sv 是 629。

10.

參數設定如下:-s=0 代表是有 C constraint 的 SVM，-t=1 代表是 1+rx^Tx 的 Q 次方這種形式。用-c 設定 C，-d 設定 Q。-g=1 設定 gamma=1，-r=1 設定常數是 1。把 test data 輸入 svm_predict 去得到 predict。其中有個參數是 p_acc。根據 document，p_acc[0]是 accuracy，單位是百分比，所以 Eout 就是 100- p_acc[0]

結果如下:

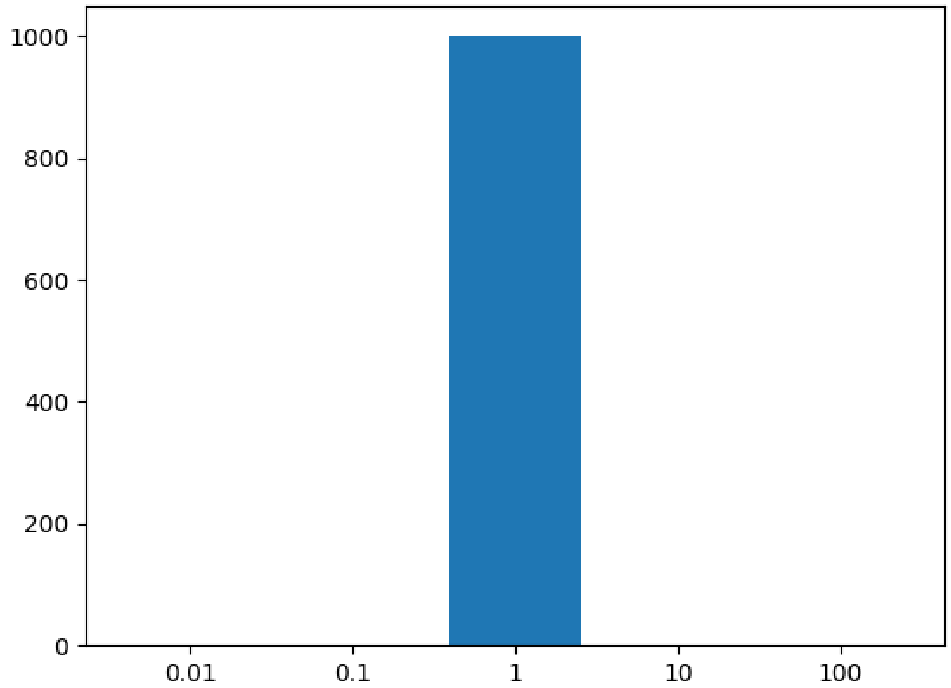
C=0.01→ Accuracy = 95.4% (1908/2000)
C=0.1→ Accuracy = 98.8% (1976/2000)
C=1→ Accuracy = 99.5% (1990/2000)
C=10→ Accuracy = 99.4% (1988/2000)
C=100→ Accuracy = 99.45% (1989/2000)

所以 Eout 最低的是 C=1

11.

參數設定一樣。重複一千次，每次設 seed 方便 reproduce。我是用 sklearn train_test_split 做分割。因為 train data 大小是 4435，所以切割比例設 200/4435。Random_state 每次設不同數字，方便 reproduce。

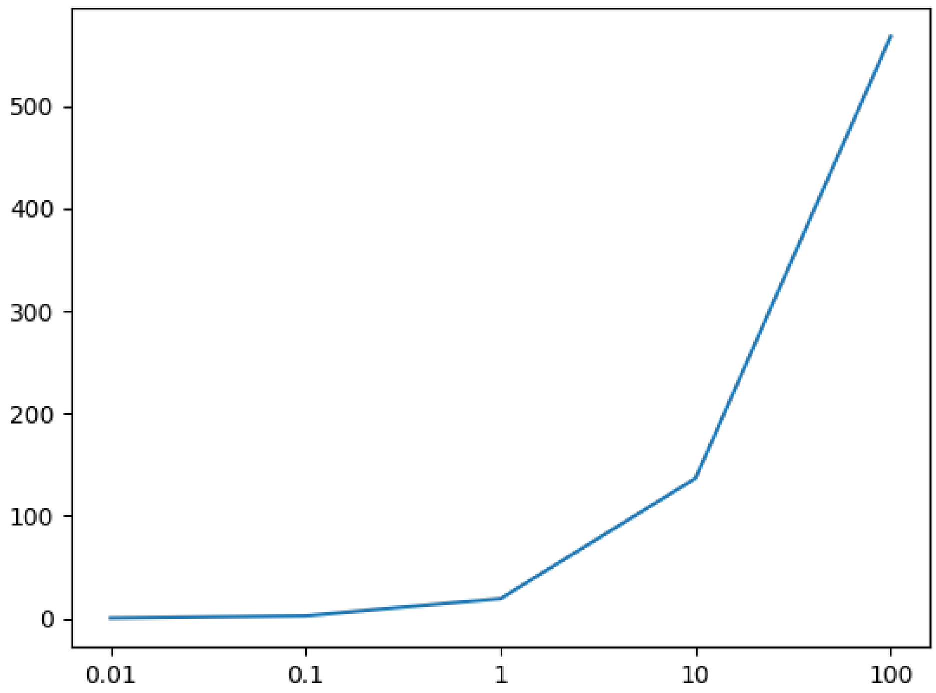
結果一千次都是 1。圖如下:



跟第 10 題結果一樣。可以說明 SVM 是個很穩定的方法。也就是說幾乎沒有 overfitting 的問題。

12.

根據講義的推導，$\mathrm{w} = \sum a_n y_n z_n$ ，所以$|\mathrm{w}| = \sqrt{\sum_{sv}(a_n y_n z_n)^2} = \sqrt{\sum_{sv}(a_n z_n)^2} = \sqrt{\sum_{sv} a_n^2 z_n^T z_n} = \sqrt{\sum_{sv} a_n^2 K(x_n, x_n)} = \sqrt{\sum_{sv} a_n^2}$，其中$a_n$可用 get_sv_coef 獲得。結果如下:



C 是個 regularization 的參數，所以當 C 很小時，||w||的大小就變得十分重要，所以會很低。當 C 很大時，||w||的大小就變得不太重要，可以變很大。所以觀察上圖，||w||的大小隨著 C 變大，是符合理論預測的。

13.

Hard margin SVM dual 的式子是

$$\max_{all\ \ \alpha\geq 0, \sum_{n=1}^{N}\alpha_n y_n=0、\mathrm{w_i}=\sum_{n=1}^{N}\alpha_n z_n y_n} \left(\frac{-1}{2}\left\|\sum_{n=1}^{N}\alpha_n\, z_n y_n\right\|^2+\sum_{n=1}^{N}\alpha_n\right)。$$

改寫成 min:

$$\min_{all\ \ \alpha\geq 0, \sum_{n=1}^{N}\alpha_n y_n=0、\mathrm{w_i}=\sum_{n=1}^{N}\alpha_n z_n y_n} \left(\frac{1}{2}\left\|\sum_{n=1}^{N}\alpha_n\, z_n y_n\right\|^2-\sum_{n=1}^{N}\alpha_n\right)$$

把 constraint 寫成 Lagrange multipliers 放進式子就變成

$$\min_{a}\left(\max_{all\ \ \beta\leq 0}\left(\frac{1}{2}\left\|\sum_{n=1}^{N}\alpha_n\, z_n y_n\right\|^2-\sum_{n=1}^{N}\alpha_n+\sum_{n=1}^{N}\beta_{1n}\alpha_n+\gamma\sum_{n=1}^{N}\alpha_n y_n\right)\right)$$

跟講義一樣，先做個調換:

$$\max_{all\ \ \beta\leq 0}\left(\min_{a}\left(\frac{1}{2}\left\|\sum_{n=1}^{N}\alpha_n\, z_n y_n\right\|^2-\sum_{n=1}^{N}\alpha_n+\sum_{n=1}^{N}\beta_{1n}\alpha_n+\gamma\sum_{n=1}^{N}\alpha_n y_n\right)\right)$$

變數是α。對任意α_n微分會變成$z_n y_n(\sum_{n=1}^{N}\alpha_n\, z_n y_n)-1+\beta_{1n}+\gamma y_n=0=\mathrm{w}^{\mathrm{T}}z_n y_n-1+\beta_{1n}+\gamma y_n\to \mathrm{w}^{\mathrm{T}}z_n y_n+\beta_{1n}+\gamma y_n\geq \mathrm{w}^{\mathrm{T}}z_n y_n+\gamma y_n=y_n(\mathrm{w}^{\mathrm{T}}z_n+\gamma)\geq 1$。如果令$\gamma=\mathrm{b}$，則$y_n(\mathrm{w}^{\mathrm{T}}z_n+b)\geq 1$，跟原本的 constraint 一樣。所以改寫成

$$\min_{all\ \ \beta\leq 0}(\frac{1}{2}\mathrm{w}^{\mathrm{T}}w-\sum_{n=1}^{N}\alpha_n+\sum_{n=1}^{N}\beta_{1n}\alpha_n+\mathrm{b}\sum_{n=1}^{N}\alpha_n y_n)$$

$$=\max_{all\ \ \beta\leq 0}(\frac{1}{2}\mathrm{w}^{\mathrm{T}}w-\mathrm{w}^{\mathrm{T}}\sum_{n=1}^{N}\alpha_n\, z_n y_n+\sum_{n=1}^{N}\alpha_n-\sum_{n=1}^{N}\beta_{1n}\alpha_n$$

$$-\mathrm{b}\sum_{n=1}^{N}\alpha_n y_n)$$

$$=\max_{all\ \ \beta\leq 0}(\frac{1}{2}\mathrm{w}^{\mathrm{T}}w+\alpha_n\sum_{n=1}^{N}(1-\mathrm{w}^{\mathrm{T}}\, z_n y_n-\mathrm{b})-\sum_{n=1}^{N}\beta_{1n}\alpha_n)$$

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

$$-\sum_{n=1}^{N}\beta_{1n}\alpha_n)$$

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

可以發現就推回來了。所以其實 dual of dual 跟 primal 是一樣的。