Machine learning Homework 5

B01902032 江東峻

1. Answer: b
   the primal formulation is $\min_{b,\mathbf{w},\boldsymbol{\xi}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\cdot\sum_{n=1}^{N}\xi_n$ s.t. $y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1-\xi_n$ and $\xi_n \geq 0$ for all n
   variable = d + N + 1 (w長度為d , ξ 有N個, C有1個)
2. Answer: c
   依y分: {1,2,3} {4,5,6,7}

	z1	z2	y	z1+z2	z1-z2
x1	1	-2	-1	-1	3
x2	4	-5	-1	-1	9
x3	4	-1	-1	3	5
x4	5	-2	+1	3	7
x5	7	-7	+1	0	14
x6	7	1	+1	8	6
x7	7	-1	+1	6	8

   [a]: 分成{1,2,3,4,5} {6,7} 與y不符
   [b]: 分成{1} {2,3,4,5,6,7}與y不符
   [c]: 分成{1,2,3} {4,5,6,7} 與y相同
   [d]: 分成{1,2,3,4,5,6,7} {} 與y不符
3. Answer: b d
   a1~a7 = [0.0000 0.7037 0.7037 0.8889 0.2593 0.2593 0.0000]
   max = a4, min = a1 & a7
   sum = 2.8148
4. Answer: b
   直接代公式：
   [a]: 分成 {1,2,3,4,5,6,7} {} 與y不符
   [b]: 分成 {1,2,3} {4,5,6,7} 與y相同
   [c]: 分成 {1,2,3,4,5,6,7} {} 與y不符
   [d]: 分成 {1,2,4,5} {3,6,7} 與y不符
5. Answer: c
   兩個是不同的曲線，因為他們用到不同的Z空間，一個是用兩個給定的公式將x轉換成z，一個是用second-order polynomial transformation，是不同的轉換公式。
6. Answer: d
   先將constraint轉成 constraint' ≤ 0的形式，再乘上Lagrange multiplier 後加總，與原本的所求相加： => L(R, c, λ) = [d]
7. Answer: a c d
   KKT Optimality Conditions:
   primal feasible: $\|x_n-c\|^2 \leq R^2$ (a)
   dual feasible: $\lambda_n \geq 0$ (b)
   dual-inner optimal:
   (c) $\frac{\partial L(R,c,\lambda)}{\partial R} = 2R-2R\sum_{n=1}^{N}\lambda_n = R\left(1-\sum_{n=1}^{N}\lambda_n\right)=0$ (d) $\frac{\partial L(R,c,\lambda)}{\partial c} = \sum_{n=1}^{N}\lambda_n(-x_n+c) = 0$
   primal-inner optimal : $\lambda_n(\|x_n-c\|^2-R^2) = 0$ (e)
   [a]: by (a), 若 R≠0, 另一項必須為0
   [b]: by primal-inner optimal, $\lambda_n = 0$, 則另一項沒有限制
   [c]: by primal-inner optimal, 這項不為0, 則$\lambda_n = 0$
   [d]: by (b), 若λ_n 總和不為0, 移項後可得 c

8. Answer: e

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

=> 因為R > 0, $\sum_{n=1}^{N} \lambda_n = 1$代入消去開頭以及最後$R^2$ => $L(R,c,\lambda) = \sum_{n=1}^{N} \lambda_n |x_n - c|^2$

=> 用(d)將c換掉，又R > 0, $\sum_{n=1}^{N} \lambda_n = 1$得 $c = \sum_{m=1}^{N} \lambda_m x_m$ => $L(R,c,\lambda) = \sum_{n=1}^{N} \lambda_n \left| x_n - \sum_{m=1}^{N} \lambda_m x_m \right|^2$ = [a]

但這是max Objective(λ), 要換成min 的話需要做其他的動作，而在選項裡都沒有符合的答案 = [e]

9. Answer: e

將8的[a]的平方乘開:

=> $\sum_{n=1}^{N} \lambda_n \left(x_n^T x_n - 2 x_n^T \sum_{m=1}^{N} \lambda_m x_m + \left(\sum_{m=1}^{N} \lambda_m x_m \right)^2 \right)$

再展開，並將 $\sum_{n=1}^{N} \lambda_n = 1$代入，

=> $\sum_{n=1}^{N} \lambda_n x_n^T x_n - 2 \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m x_n^T x_m + \sum_{n=1}^{N} \lambda_n \left(\sum_{m=1}^{N} \lambda_m x_m \right)^2 = \sum_{n=1}^{N} \lambda_n x_n^T x_n - 2 \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m x_n^T x_m + \left(\sum_{m=1}^{N} \lambda_m x_m \right)^2$

最後一項由於沒有變數n，可以把其中一個m換成n，並且代入kernel形式

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

最後一項可以改成 $\sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K(x_n, x_m)$

跟中間項形式一樣 => 合併

=> $\sum_{n=1}^{N} \lambda_n K(x_n, x_n) - \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m K(x_n, x_m)$ = [c]

但這是max Objective(λ), 要換成min 的話需要做其他的動作，而在選項裡都沒有符合的答案 = [e]

10. Answer: a

by 第8題的(e)式，若$\lambda_i > 0$, 則 $\|x_n - c\|^2 = R^2$

代入第9題的形式，可得 $R^2 = \left| x_i - \sum_{m=1}^{N} \lambda_m x_m \right|^2$

展開平方式： $R^2 = x_i^T x_i - 2 \sum_{m=1}^{N} \lambda_m x_i^T x_m + \left(\sum_{m=1}^{N} \lambda_m x_m \right)^2 = x_i^T x_i - 2 \sum_{m=1}^{N} \lambda_m x_i^T x_m + \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m x_n^T x_m$

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

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

11. Answer: a

令$w^{\sim}$ = (w, tξ), t = constant, $x_n^{\sim}$ = (x_n, v_1, v_2, … , v_N)

用 (1) $\min_{b,w^{\sim}} \frac{1}{2} w^{\sim T} w^{\sim}$ *subject to* $y_n\left(w^{\sim T} x_n^{\sim} + b\right) \geq 1$

表示 (2) $\min_{b,w,\xi} \frac{1}{2} w^T w + C\sum_{n=1}^{N} \xi_n^2$ *subject to* $y_n(w^T x_n + b) \geq 1-\xi_n$

將(1) 展開得到： $\min_{b,w,\xi} \frac{1}{2} w^T w + \frac{1}{2} t^2 \sum_{n=1}^{N} \xi_n^2$ *subject to* $y_n\left(w_n^T x_n + t\sum_{m=1}^{N} \xi_m v_m + b\right) \geq 1$

移項得到(3) $\min_{b,w,\xi} \frac{1}{2} w^T w + \frac{1}{2} t^2 \sum_{n=1}^{N} \xi_n^2$ *subject to* $y_n\left(w_n^T x_n + b\right) \geq 1 - y_n t\sum_{m=1}^{N} \xi_m v_m$

其中令 $\frac{1}{2} t^2 = C, t = \sqrt{2C}$ $v_m = \frac{1}{\sqrt{2C}}[m=n]$

則(3)可以換成 $\min_{b,w,\xi} \frac{1}{2} w^T w + C\sum_{n=1}^{N} \xi_n^2$ *subject to* $y_n\left(w_n^T x_n + b\right) \geq 1 - y_n\xi_n$

若有ξ = (ξ_1,…, ξ_i ,…, ξ_N) 是最佳解，那 ξ' = ($-\xi_1$,…, $-\xi_i$,…, $-\xi_N$) = -ξ 也會是最佳解。
而不等式可以選較寬鬆的範圍。
所以 ξ 的正負不影響取最小值的結果，那可以把y_n拿掉，得到 (2)式
而$x_n^{\sim}$ 為[a]的形式。

12. Answer: a c

令A為N*N的矩陣，$A_{ij} = \Phi_1(x_i)^T\Phi_1(x_j)$ 可表示K_1
B為N*N的矩陣，$B_{ij} = \Phi_2(x_i)^T\Phi_2(x_j)$ 可表示K_2

[a] : K = K_1+K_2 , 則 可以做矩陣C = N*N 表示K: $C_{ij} = \Phi_1(x_i)^T\Phi_1(x_j)+\Phi_2(x_i)^T\Phi_2(x_j)=A_{ij}+B_{ij}$
可知C = A+B。 由於K_1,K_2為valid kernel，有symmetric 和 PSD性質。
$C_{ij} = A_{ij}+B_{ij} = A_{ji}+B_{ji} = C_{ji}$ (A,B有symmetric性質)
A,B有PSD性質： $(X^TAX)_{ij} \geq 0$, $(X^TBX)_{ij} \geq 0$
$X^TCX = X^T(A+B)X = X^T(AX+BX) = X^TAX+X^TBX$ (矩陣的分配律)
則 $(X^TCX)_{ij} = (X^TAX)_{ij} + (X^TBX)_{ij} \geq 0$, C有PSD性質 => K is valid

[b]: K = K_1 - K_2 , 則 可以做矩陣C = N*N 表示K: $C_{ij} = \Phi_1(x_i)^T\Phi_1(x_j)-\Phi_2(x_i)^T\Phi_2(x_j)=A_{ij} - B_{ij}$
可知C = A-B。令A = I, B = 2I (I, 2I 都是PSD 且symmetric), C = -I。
則trace(C) < 0 不是PSD => K 不是valid kernel。

[c]:
PSD:
令$f_i(x_n)$為$\Phi_1(x_n)$轉換的第i個係數，令$g_j(x_m)$為$\Phi_2(x_m)$轉換的第j個係數。
則$K_1(x,x')K_2(x,x') = (\sum_{i=1}^{\infty} f_i(x)f_i(x'))(\sum_{j=1}^{\infty} g_j(x)g_j(x')) = \sum_{i,j} f_i(x)g_j(x)f_i(x')g_j(x')$
則定義Φ_3的feature為 $h_{i,j}(x) = f_i(x)g_j(x)$, for 每一對i,j。
=> $K(x,x') = K_1(x,x')K_2(x,x') = \Phi_3(x)^T\Phi_3(x')$
symmetric:
$K(x',x) = K_1(x',x)K_2(x',x) = K_1(x,x')K_2(x,x') = K(x,x')$ (K_1,K_2有對稱性)
=> K(x,x') 為新的kernel 。

[d]: 令A = [1 1; 1 1], B = [100 1; 1 100] (皆為PSD 且symmetric)。
C = 2*2矩陣，表示K => C = [0.01 1; 1 0.01]
det(C) < 0 => K不是valid kernel。

13. Answer: b d

[a]: [a]可以拆成1 + $K_1(K_1-2)$ 。 取K_1 的表示矩陣 = I, 那K_1-2的表示矩陣為 [-1 -2; -2 -1]，$K_1(K_1-2)$的表示矩陣是[-1 0; 0 -1]，1 + $K_1(K_1-2)$的表示矩陣為 [0 1; 1 0] => det < 0 不是PSD， 則K不是valid kernel。

[b]: K_1是valid kernel，那K_1可拆成Z^TZ。構造一個Z' = √1126 * Z， 則 Z'^TZ = 1126 *Z^TZ = 1126 * K_1 = K。

[c]: 取K_1的表示矩陣 = 100I，則-K_1的表示矩陣為-100I，exp(-K_1)的表示矩陣為 [e^(-100) 1; 1 e^(-100)] => det <0 不是PSD， K不是valid kernel。

[d]: 因為 0 < $K_1(x,x')$ < 1 ， $K(x,x') = \frac{1}{1-K_1(x,x')} = \sum_{i=0}^{\infty} (K_1(x,x'))^i$

因為kernel在乘法及加法有封閉性 => K(x,x') 是valid kernel。

14. Answer: c

$$equivalent\ g_{SVM}\ classifer \Rightarrow w^Tz + b = \tilde{w}^Tz + \tilde{b}$$

$$\sum_{n=1}^{N} \alpha_n y_n K(x_n, x) + y_s - \sum_{n=1}^{N} \alpha_n y_n K(x_n, x_s) = \sum_{n=1}^{N} \tilde{\alpha_n} y_n \tilde{K}(x_n, x) + y_s - \sum_{n=1}^{N} \tilde{\alpha_n} y_n \tilde{K}(x_n, x_s)$$

消去 y_s 並把$\tilde{K}$ 用pK+q代入：

$$\sum_{n=1}^{N} \alpha_n y_n K(x_n, x) - \sum_{n=1}^{N} \alpha_n y_n K(x_n, x_s) = p\sum_{n=1}^{N} \tilde{\alpha_n} y_n K(x_n, x) + q\sum_{n=1}^{N} \tilde{\alpha_n} y_n - p\sum_{n=1}^{N} \tilde{\alpha_n} y_n K(x_n, x_s) - q\sum_{n=1}^{N} \tilde{\alpha_n} y_n$$

消去有q的項：

$$\sum_{n=1}^{N} \alpha_n y_n K(x_n, x) - \sum_{n=1}^{N} \alpha_n y_n K(x_n, x_s) = p\sum_{n=1}^{N} \tilde{\alpha_n} y_n K(x_n, x) - p\sum_{n=1}^{N} \tilde{\alpha_n} y_n K(x_n, x_s)$$

得 $\tilde{\alpha}_n = \alpha_n/p$ => 對於所有α_n , $0 < \alpha_n < C$ ，因為所有α_n縮小p倍 ，所以 bound也縮小p倍， $\tilde{C} = C/p$

15. Answer: b

w = model.SVs' * model.sv_coef;

16. Answer: e

train self and predict self

17. Answer: d

0.5*w*w-obj

18. Answer: a b e
19. Answer: b
20. Answer: b
21. Answer: NO

有可能是C設太小。若所有點都被包含在margin裡面，也是no free SV的狀況。

EX: x_1 = (-3, 1, O) , x_2 = (-1.5, -1, X) , x_3 = (1.5, 1, O) , x_4 = (3, -1, X) , C = 0.0001

結果是全部都在margin裡面，但這批資料是linear separable: {x_1,x_3}, {x_2,x_4}

22. Answer: YES

因為當SV的 $\xi_n > 1$ ，意思是只要是SV都是分錯的，而no free SV的情況，是margin包含所有的點，如果margin包含所有點(所有點都是SV)，且SV都是分錯的，那必定有個分法可以將所有SV分對 => 那一開始就應該選分對的狀況(矛盾)。

如果是沒有SV的狀況，那必定可以擴大margin => 還沒training完(矛盾)。

若所有SV的點是分錯的，那將label互換，使得SV的點是正確的(若是linear separable一定會有一種分法是這樣分，其他只是線的平移跟一點點旋轉在某個區間裡旋轉)，但SVM並沒有這樣分，所以在margin外至少有個有一點是label正確 的，才會使SV的點都是分錯的 => 此時若是linear separable的那條線，那就會發現至少有一點是錯誤的(矛盾)。

=> 不可能linear separable。