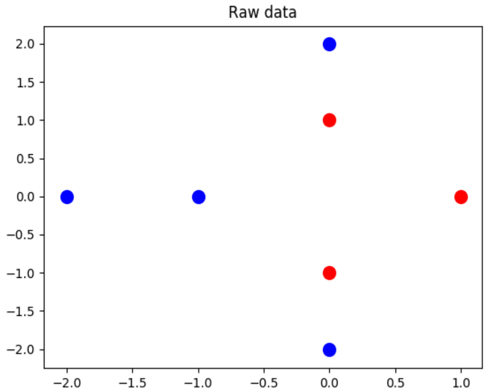
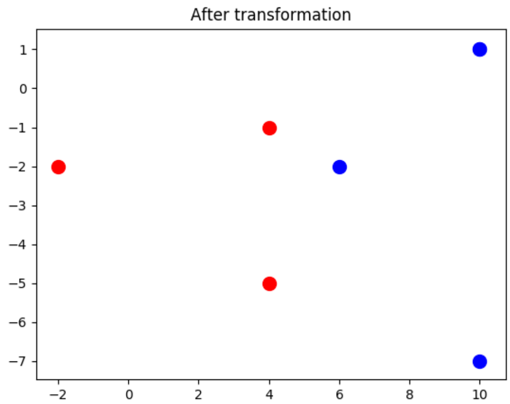
Machine Learning HW1
B04901069 電機四 林志皓

Question 1.

The data before transformation :
(1, 0), (0, 1), (0, -1), (-1, 0), (0, 2), (0, -2), (-2, 0)
The data after transformation:
(-2, -2), (4, -5), (4, -1), (6, -2), (10, -7), (10, 1), (10, 1)

illustration: (+1 denoted by blue dots, -1 denoted by red dots)





by looking the picture at right, we can pick the vertical line between (4, -1) & (6, -2) as the hyperplane in the Z space , whose function is x - 5 = 0. This plane separate the data correctly and with the largest margin. Also, I use python package sklearn to verify,

```
svm = SVC(C=100, kernel= 'linear')
svm.fit(x_transform, y_label)
support_id = svm.support_
support_vectors = svm.support_vectors_
coef = svm.dual_coef_
```

The support vectors found by the program are:
(4, -5) , (4, -1), (6, -2)
which are on the boundary of the SVM

then compute the w and b :

```
for i in range(support_num):
    w += coef[0][i] * support_vectors[i]
for i in range(support_num):
    b += (1 / y_label[support_id[i]]) - (w @ support_vectors[i])
b /= support_num
```

we got w = (0.9995266, 0.0000946)   b = -5.0000315
is almost the same as the result as above.

so the hyperplane in Z space is : x - 5 = 0

Question 2

I use python package sklearn to solve this problem, too.

```
svm = SVC(C=100, kernel= 'poly', degree= 2, gamma= 1, coef0= 1)
svm.fit(x_raw, y_label)

support_id = svm.support_
support_vecotors = svm.support_vectors_
coef = svm.dual_coef_
```

by the program, I found the alpha coefficients as below:
[0, 0.59647182, 0.81065085, 0.8887034, 0.20566488, 0.31275349, 0]
and the support vectors are the data with non-zero alpha values:
[0, 1], [0, -1], [-1, 0], [0, 2], [0, -2]

Question 3

$[\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7] =$
$[0, 0.59647182, 0.81065085, 0.8887034, 0.20566488, 0.31275439, 0]$

kernel function $K(x, x') = (1 + x^T x')^2$

$g_{svm}(x) = \text{sign}\left(\sum \alpha_n y_n K(x_n, x) + b\right)$

to compute b value. pick one support vector (0, 1)

$b = y_s - \sum \alpha_n K(x_n, x_s) = -1.666$

nonliear curve in X space : (let coordinate (x_1, x_2) in X space)

$$\sum \alpha_n y_n K(x_n, x) + b$$
$$= \alpha_2 y_2 (1 + x_2)^2 + \alpha_3 y_3 (1 - x_2)^2 + \alpha_4 y_4 (1 - x_1)^2 + \alpha_5 y_5 (1 + 2x_2)^2 + \alpha_6 y_6 (1 - 2x_2)^2 + b$$
$$= (\alpha_2 y_2 + \alpha_3 y_3 + 4\alpha_5 y_5 + 4\alpha_6 y_6) x_2^2 + 2(\alpha_2 y_2 - \alpha_3 y_3 + 2\alpha_5 y_5 - 2\alpha_6 y_6) x_2$$
$$+ \alpha_4 y_4 x_1^2 - 2\alpha_4 y_4 x_1 + b + (\alpha_2 y_2 + \alpha_3 y_3 + \alpha_4 y_4 + \alpha_5 y_5 + \alpha_6 y_6)$$
$$= 0.667 x_2^2 + 0.889 x_1^2 - 1.777 x_1 - 1.666 = 0.$$

(coefficient of x_2 is almost 0)   nonlinear curve

Question 4

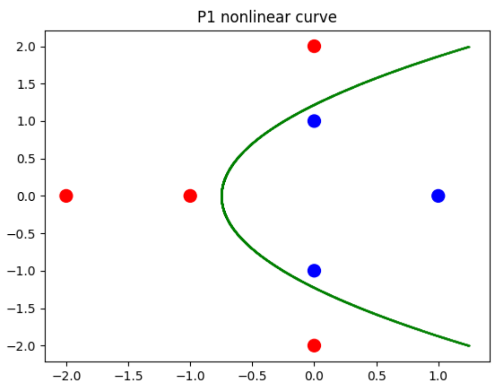
In Question 1, the nonlinear curve in X space is :

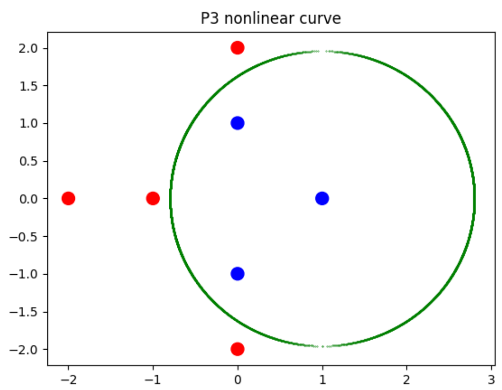
$$2X_2^2 - 4X_1 + 2 = 5$$

and in Question 3, the nonlinear curve in X space is :

$$0.667X_2^2 + 0.889(X_1 - 1)^2 = 2.555$$

Apparently, they are different. below are the illustration:





They are not the same because in the two questions we use different kernels to do nonlinear transformation, therefore the margin definition are different.

in Question 1 , the kernel is :

$$K(X, X') = (2X_2^2 - 4X_1 + 2) * (2X'^2_2 - 4X'_1 + 2) + (X_1^2 - 2X_2 - 3) * (X'^2_1 - 2X'_2 - 3)$$

and in Question3 the kernel is:

$$K(X, X') = (1 + X^T X')^2$$

Question 5

To prove $e^{-x^2} = \frac{1}{\|\Phi(x)\|}$ that is, prove $e^{x^2} = \|\Phi(x)\|$

consider the Taylor expansion of $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$

let $k = 2x^2$. $e^{2x^2} = e^k = \sum_{n=0}^{\infty} \frac{k^n}{n!} = \sum_{n=0}^{\infty} \frac{(2x^2)^n}{n!} = \sum_{n=0}^{\infty} \frac{2^n}{n!} x^{2n}$

$\|\Phi(x)\| = \left(1 + \frac{2}{1!}x^2 + \frac{2^2}{2!}x^4 + \cdots\right)^{\frac{1}{2}} = \left(\sum_{n=0}^{\infty} \frac{2^n}{n!} x^{2n}\right)^{\frac{1}{2}} = \left(e^{2x^2}\right)^{\frac{1}{2}} = e^{x^2}$

$\therefore e^{-x^2} = \frac{1}{\|\Phi(x)\|}$ #

Question 6

$\cos(x.x') = \frac{x \cdot x'}{\|x\| \cdot \|x'\|} = \left(\frac{x}{\|x\|}\right) \cdot \left(\frac{x'}{\|x'\|}\right) = \phi(x) \cdot \phi(x')$

and the transformation $\phi(x) = \frac{x}{\|x\|}$ which is the normalization of the original vector

the transformation is found $\Rightarrow$ $\cos(x.x')$ is a valid kernel

Question 7

$$L(R,c,\lambda) = R^2 + \lambda_1(\|z_1 - c\|^2 - R^2) + \lambda_2(\|z_2 - c\| - R^2) + \cdots + \lambda_n(\|z_n - c\|^2 - R^2)$$

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

Question 8

(P) $\min_{R \in \mathbb{R}, c \in \mathbb{R}^d} \max_{\lambda_n \geq 0} L(R, c, \lambda)$. (D) $\max_{\lambda_n \geq 0} \min_{R \in \mathbb{R}, c \in \mathbb{R}^d} L(R, c, \lambda)$

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

at optimal solution, $\frac{\partial L}{\partial R} = 0$. $\frac{\partial L}{\partial c} = 0$.

$$\frac{\partial L(R, c, \lambda)}{\partial R} = 2R + \sum_{n=1}^{N} -2\lambda_n \cdot R = 2(1 - \sum_{n=1}^{N} \lambda_n) R = 0.$$

$$\frac{\partial L(R, c, \lambda)}{\partial c_k} = 0 + \sum_{n=1}^{N} \lambda_n [-2(z_{nk} - c_k)] = 2 \sum_{n=1}^{N} \lambda_n (c_k - z_{nk}) = 0.$$

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

KKT Optimality Conditions:

① Primal feasible: $\|z_n - c\|^2 \leq R^2$, $\forall n$

② Dual feasible: $\lambda_n \geq 0$, $\forall n$

③ Dual-inner Optimal: $(1 - \sum_{n=1}^{N} \lambda_n) R = 0$. $\sum_{n=1}^{N} \lambda_n (c - z_n) = 0$.

④ Primal-inner Optimal: $\lambda_n (\|z_n - c\|^2 - R^2) = 0$, $\forall n$.

by the Dual-inner optimality, $\sum_{n=1}^{N} \lambda_n (c - z_n) = 0$

$$\sum_{n=1}^{N} \lambda_n \cdot c - \lambda_n z_n = c\left(\sum_{n=1}^{N} \lambda_n\right) - \sum_{n=1}^{N} \lambda_n z_n = 0.$$

if $\sum_{n=1}^{N} \lambda_n \neq 0$. $\therefore c = \left(\sum_{n=1}^{N} \lambda_n z_n\right) / \left(\sum_{n=1}^{N} \lambda_n\right)$ #

Question 9

By $\left(1-\sum_{n=1}^{N}\lambda_n\right)R = 0$ and $R>0$. ∴ $\sum_{n=1}^{N}\lambda_n = 1$.

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

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

∴ $\sum_{n=1}^{N}\lambda_n z_n = c$.

$$\begin{aligned}
L(R,c,\lambda) &= \sum_{n=1}^{N}\lambda_n(z_n-c)^T(z_n-c) = \sum_{n=1}^{N}\lambda_n\left(z_n^Tz_n - 2c^Tz_n + c^Tc\right)\\
&= \sum_{n=1}^{N}\lambda_n z_n^Tz_n - 2\sum_{n=1}^{N}\lambda_n c^Tz_n + \sum_{n=1}^{N}\lambda_n c^Tc\\
&= \sum_{n=1}^{N}\lambda_n z_n^Tz_n - 2c^T\cdot\left(\sum_{n=1}^{N}\lambda_n z_n\right) + c^Tc\left(\sum_{n=1}^{N}\lambda_n\right)\\
&= \sum_{n=1}^{N}\lambda_n z_n^Tz_n - 2c^Tc + c^Tc = \sum_{n=1}^{N}\lambda_n z_n^Tz_n - c^Tc\\
&= \sum_{n=1}^{N}\lambda_n z_n^Tz_n - \left(\sum_{n=1}^{N}\lambda_n z_n^T\right)\left(\sum_{n=1}^{N}\lambda_n z_n\right)\\
&= \sum_{n=1}^{N}\lambda_n z_n^Tz_n - \sum_{m=1}^{N}\sum_{n=1}^{N}\lambda_m\lambda_n z_m^Tz_n
\end{aligned}$$

∴ $(P')\ \max_{\lambda_n\ge 0}\ \sum_{n=1}^{N}\lambda_n z_n^Tz_n - \sum_{m=1}^{N}\sum_{n=1}^{N}\lambda_m\lambda_n z_m^Tz_n$, subject to $\sum_{n=1}^{N}\lambda_n = 1$

Question 10

when solved D'

The Primal-inner Optimality in KKT condition must be satisfied

$$\lambda_n (\|z_n - c\|^2 - R^2) = 0, \ \forall \lambda_n.$$

for some i with $\lambda_n > 0$. $\|z_i - c\|^2 - R^2 = 0$

$$R = \sqrt{\|z_i - c\|^2} = \sqrt{(z_i - c)^T (z_i - c)}$$

$$= \sqrt{z_i^T z_i - 2c^T z_i + c^T c}$$

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

$$= \sqrt{z_i^T z_i - 2\sum_{n=1}^{N} \lambda_n z_n^T z_i + \sum_{m=1}^{N}\sum_{n=1}^{N} \lambda_m \lambda_n z_m^T z_n}$$

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

#

Question 11

Let $f(\alpha) = \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m z_n^T z_m - \sum_{n=1}^{N}\alpha_n$

Hard-margin SVM dual:

$\min_{\alpha} f(\alpha)$, subject to $\sum_{n=1}^{N} y_n\alpha_n = 0$, $\alpha_n \geq 0$, $n = 1, 2, 3 \ldots n$

Soft margin SVM dual:

$\min_{\alpha} f(\alpha)$, subject to $\sum_{n=1}^{N} y_n\alpha_n = 0$ $C \geq \alpha_n \geq 0$ $n = 1, 2 \cdots n$.

α^* is the optimal solution for Hard-margin SVM., $C \geq \max_{1\leq n\leq N}\alpha_n^*$.

Let α' is the optimal solution for Soft-margin SVM, and $\alpha' \neq \alpha^*$.

$\because \alpha^*$ satisfy the constrain of soft-margin SVM.

$\therefore \alpha^*$ is one of the solution of soft-margin SVM.

and $f(\alpha') < f(\alpha^*)$ $\because \alpha'$ is the optimal.

α' satisfy the constrain, so α' is one solution of Hard-margin SVM

and $f(\alpha') < f(\alpha^*)$ $\therefore \alpha^*$ is not the optimal $\longrightarrow \times$

$\therefore \alpha^*$ is also the optimal solution for soft margin SVM.

Question 12

Consider Soft-margin SVM dual:

$$\min_{\alpha} \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m z_n^T z_m - \sum_{n=1}^{N} \alpha_n$$

$$\equiv \min_{\alpha} \frac{1}{2} \alpha^T Q \alpha + q^T \alpha, \quad Q_{n,m} = y_n y_m K(x_n, x_m) \quad q = -1_N$$

subject to $\sum_{n=1}^{N} y_n \alpha_n = 0$, $C \geq \alpha_n \geq 0$. $n = 1, 2, \dots N$.

Let $f(\alpha) = \frac{1}{2} \alpha^T Q \alpha + q^T \alpha$, with kernel K,

$\tilde{f}(\alpha) = \frac{1}{2} \alpha^T \tilde{Q} \alpha + q^T \alpha$ with kernel $\tilde{K}(x, x') = pK(x, x')$

Original soft-margin SVM dual: $\min_{\alpha} f(\alpha)$, $0 \leq \alpha_n \leq C$, $\sum_{n=1}^{N} y_n \alpha_n = 0$

New soft-margin SVM dual: $\min_{\alpha} \tilde{f}(\alpha)$, $0 \leq \alpha_n \leq C/p$, $\sum_{n=1}^{N} y_n \alpha_n = 0$

Let α^* is the optimal solution for original problem. $0 \leq \alpha_n^* \leq C$, $n = 1 \dots N$.

$\because$ $0 \leq \frac{\alpha_n^*}{p} \leq \frac{C}{p}$, $n = 1 \dots N$ $\therefore$ $\frac{\alpha_n^*}{p}$ is one solution for new SVM

Let α' is the optimal solution of new SVM and $\alpha' \neq \alpha^*$

$\therefore$ $\tilde{f}(\alpha') < \tilde{f}(\frac{\alpha^*}{p}) = \frac{1}{2} (\frac{\alpha^*}{p})^T \tilde{Q} (\frac{\alpha^*}{p}) + q^T (\frac{\alpha^*}{p}) = \frac{1}{p} \left[\frac{1}{2} \alpha^{*T} Q \alpha^* + q^T \alpha^* \right] = \frac{1}{p} f(\alpha^*)$

$\because$ $0 \leq \alpha_n' \leq \frac{C}{p}$ $\therefore$ $0 \leq p\alpha_n' \leq C$ $\therefore$ $p\alpha_n'$ is one of the solution of original SVM.

$$f(p\alpha') = (p\alpha')^T Q (p\alpha') + q^T (p\alpha') = p(\alpha'^T \tilde{Q} \alpha' + q^T \alpha') = p\tilde{f}(\alpha')$$

$$< p \cdot \frac{1}{p} f(\alpha^*) = f(\alpha^*).$$

$\therefore$ $p\alpha'$ is more optimal than α^* on original SVM — $\times$

$\therefore$ the optimal solution for new SVM $= \frac{\alpha^*}{p}$

Let b^* be b of original SVM select $\alpha_s < C$ (free support vector)

$$b^* = y_s - \sum_{SV} \alpha_n^* y_n K(x_n, x_s)$$

Let b' be b of new SVM. $\because \frac{\alpha^*}{p}$ has the same support vectors as α^*

select $0 \le \frac{\alpha_s^*}{p} \le C$ (free)

$$b' = y_s - \sum_{n=1} \frac{\alpha_n^*}{p} y_n \tilde{K}(x_n, x_s)$$

$$= y_s - \sum_{SV} \frac{\alpha_n^*}{p} y_n p K(x_n, x_s) = y_s - \sum_{SV} \alpha_n^* y_n K(x_n, x_s) = b^*$$

For Original SVM.

$$g_{svm} = \text{sign}\left(\sum_{SV} \alpha_n^* y_n K(x_n, x) + b^* \right)$$

For New SVM.

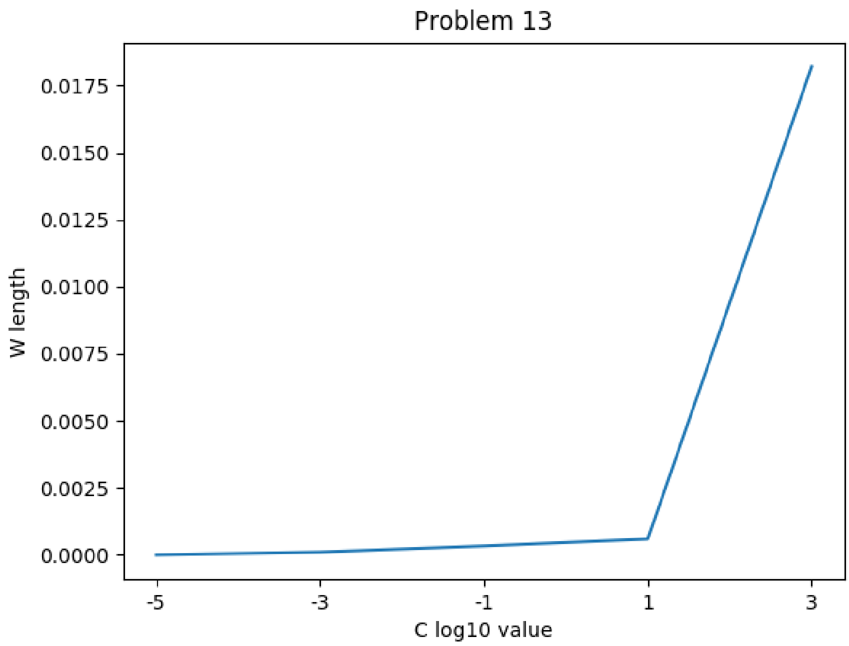
$$g_{svm} = \text{sign}\left(\sum_{SV} \frac{\alpha_n^*}{p} y_n \tilde{K}(x_n, x) + b^* \right)$$

$$= \text{sign}\left(\sum_{SV} \frac{\alpha_n^*}{p} y_n p K(x_n, x) + b^* \right)$$

$$= \text{sign}\left(\sum_{SV} \alpha_n^* y_n K(x_n, x) + b^* \right)$$

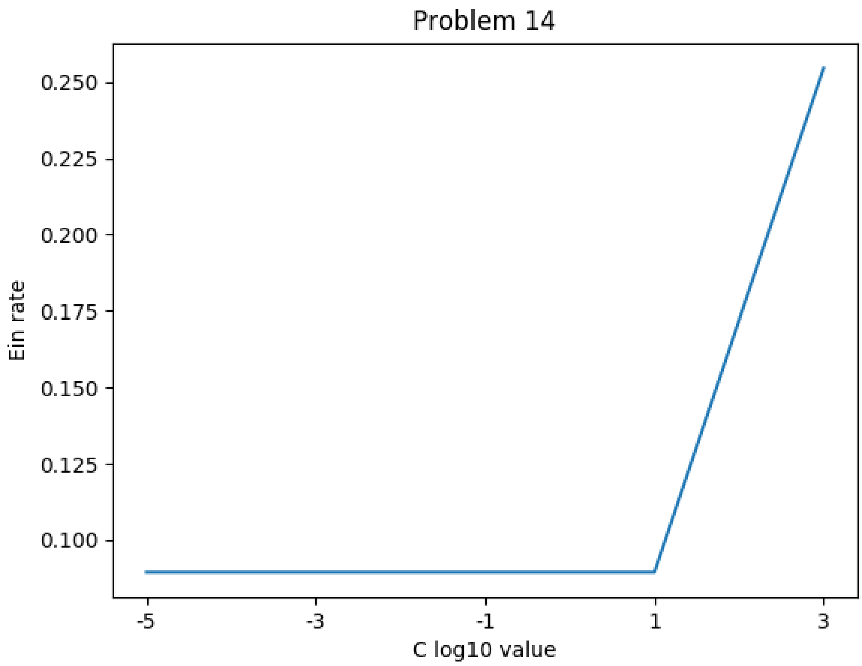
$\Rightarrow$ is the same as g_{svm} of original SVM #

Question 13



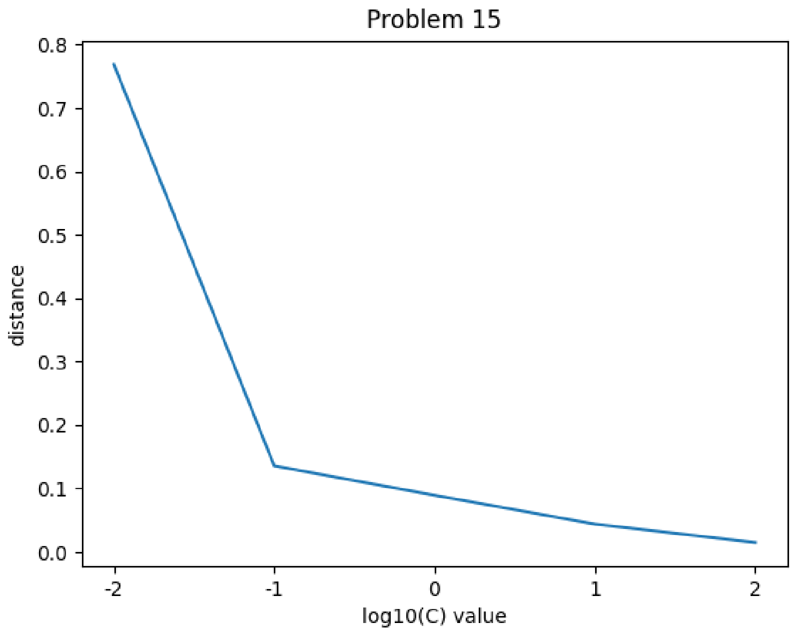
Length of W becomes bigger when cost value (C) increases. Larger C means less tolerance toward violations, so the model is less regularized. Therefore, the margin of the hyperplane would be more complicated, and the length of W increases.

Question 14



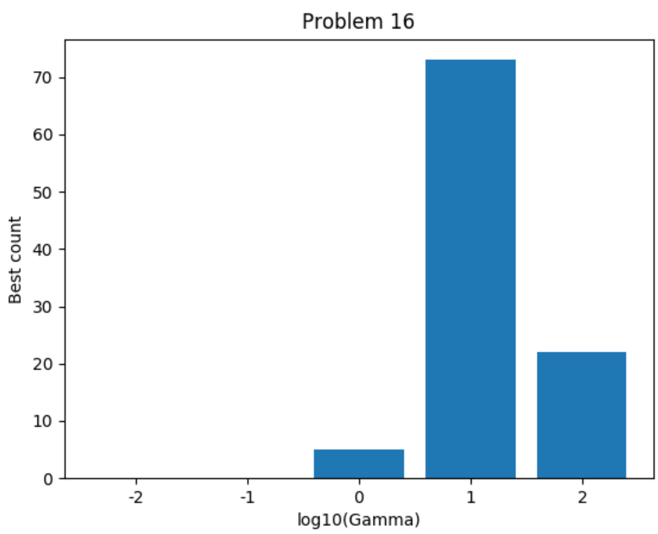
It's harder to solve the QP problem when the C gets bigger. The E_in should be smaller when the model becomes more powerful. When C gets larger, the model is less regularized, so the model is more powerful, E_in should be smaller as a result. I use "LIBSVM" in this problem, when the C is set to 1000, the optimization process reach its max number of iterations, and return a less optimized result, so the rightmost point in figure above is a little bit incorrect.

Question 15



As we can see in the figure above, the distance of any free vector to the hyperplane decreases when cost value C gets bigger. In other words, the width of the margin gets smaller, too.

Question 16



As we can see, Gamma = 10 is the best choice.

Question 17

(17) When solving the dual problem of SVM, we solve

$$\min_{\alpha} \left(\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m z_n^T z_m - \sum_{n=1}^{N}\alpha_n\right)$$ subject to $\sum_{n=1}^{N} y_n \alpha_n = 0$.

$0 \le \alpha_n \le C, \forall n$

$$W = \sum_{n=1}^{N} y_n \alpha_n z_n$$

let z_i be constant feature z_n and $z_i = c$

The the corresponding w_i to z_i

$$\sum_{n=1}^{N} y_n \alpha_n z_{n,i} = \sum_{n=1}^{N} y_n \alpha_n \cdot c = c\left(\sum_{n=1}^{N} y_n \alpha_n\right) = 0$$

$\therefore$ W_i correspond to some constant feature z_i is 0 #

Question 18

$$\min_{\alpha}\left(\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m k(x_n, x_m) - \sum_{n=1}^{N}\alpha_n\right) \text{ subject to } \sum_{n=1}^{N} y_n\alpha_n = 0 \quad \alpha_n \geq 0$$

When use another kernel $\tilde{k}$, s.t. $\tilde{k}(x,x') = k(x,x') + q$. $q \in \mathbb{R}$

the optimization problem become:

$$\min_{\alpha}\left(\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m \tilde{k}(x_n, x_m) - \sum_{n=1}^{N}\alpha_n\right) \text{ subject to } \sum_{n=1}^{N} y_n\alpha_n = 0 \quad \alpha_n \geq 0$$

$$\min_{\alpha}\left(\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m (k(x_n, x_m) + q) - \sum_{n=1}^{N}\alpha_n\right)$$

$$= \min\left(\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m k(x_n, x_m) + \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m q - \sum_{n=1}^{N}\alpha_n\right)$$

$$\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m q = q\left(\sum_{n=1}^{N} y_n \alpha_n\right)\left(\sum_{m=1}^{N} y_m \alpha_m\right) = q \cdot 0 \cdot 0 = 0$$

$\therefore$ The optimization problem of k and $\tilde{k}$ are the same!

$\Rightarrow$ same solution α

g_{svm} for k:

$b = y_s - \sum_{n=1}^{N} y_n \alpha_n k(x_n, x_s)$ for some free support vector x_s

$g_{svm}(x) = \text{sign}\left(\sum_{n=1}^{N} y_n \alpha_n k(x_n, x) + b\right)$

$\tilde{g}_{svm}$ for $\tilde{k}$

$$b^* = y_s - \sum_{n=1}^{N} y_n \alpha_n \tilde{k}(x_n, x_s) = y_s - \sum_{n=1}^{N} y_n \alpha_n (k(x_n, x_s) + q)$$

$$= y_s - \sum_{n=1}^{N} y_n \alpha_n k(x_n, x_s) - \underbrace{\sum_{n=1}^{N} y_n \alpha_n q}_{0} = b$$

$$\therefore \tilde{g}_{svm}(x) = \text{sign}\left(\sum_{n=1}^{N} y_n \alpha_n \tilde{k}(x_n, x) + b\right)$$

$$= \text{sign}\left(\sum_{n=1}^{N} y_n \alpha_n (k(x_n, x) + q) + b\right)$$

$$= \text{sign}\left(\sum_{n=1}^{N} y_n \alpha_n k(x_n, x) + \sum_{n=1}^{N} y_n \alpha_n q + b\right) = \text{sign}\left(\sum_{n=1}^{N} y_n \alpha_n k(x_n, x) + b\right)$$

is the same as g_{svm} #

NAN PAO