Assignment2 Report

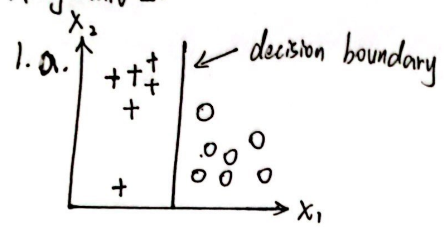
Author: 余永琦 Student Number: ID 120090761

Contents

Part1: Written Questions

Assignment 2.

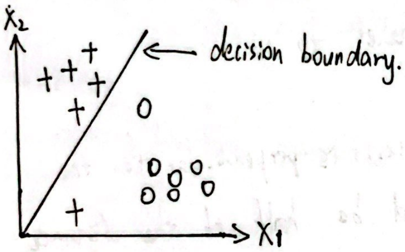
1. a.



The decision boundary is not unique, the line that seperate two classes can be a decision boundary.

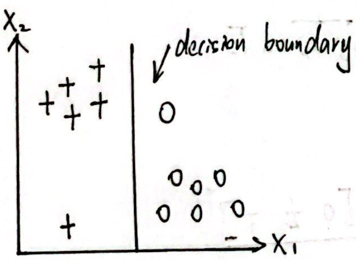
0 classification errors ~~are~~ is ~~maked~~ made by this boundary.

b.

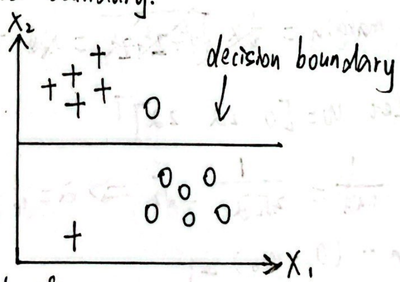


1 classification error is made by this boundary.

c.



d.



0 ~~cla~~ classification error is made by this boundary.

2 classification errors are made by this boundary.

2.

a. $\phi(x_1) = [1, 0, 0]^T$, $\phi(x_2) = [1, 2, 2]^T$

A vector from $\phi(x_1)$ to $\phi(x_2)$ can be represent by: $[0, 2, 2]^T$

Since w is perpendicular to the decision boundary, so it should parallel to the vector from $\phi(x_1)$ to $\phi(x_2)$.

$\therefore$ The vector $[0\ 2\ 2]^T$ is parallel to w.

b. Because the vector from $\phi(x_1)$ to $\phi(x_2)$ is perpendicular to the decision boundary, the margin should be half of the distance between $\phi(x_1)$ and $\phi(x_2)$.

$\therefore$ margin $= \cancel{\frac{1}{2} \times \sqrt{0^2+2^2+2^2}} = \frac{1}{2} \times \sqrt{0^2+2^2+2^2} = \sqrt{2}$

c. Let $w = [0\ \ 2a\ \ 2a]^T$.

$$\frac{1}{\|w\|} = \frac{1}{2\sqrt{2}a} = \sqrt{2} \Rightarrow a = \frac{1}{4}, \quad \underline{w = [0\ \ \tfrac{1}{2}\ \ \tfrac{1}{2}]^T}$$

d. $$\begin{cases} -(0 + w_0) = 1 \\ 2 + w_0 = 1 \end{cases} \Longrightarrow w_0 = -1$$

e. $f(x) = \frac{\sqrt{2}}{2} x + \frac{1}{2} x^2 - 1$

3. Yes. Since the dataset is linearly separable, we can always find a hyperplane that separate the data.

That is, the problem $\min \frac{1}{2}\|w\|^2$ s.t. $1-y_i(w^T x_i + b) \leq 0, \forall i$ is feasible and has a optimal solution w^*. Now consider $\min \frac{1}{2}\|w\|^2 + C\sum_i^m \xi_i$ s.t. $1-\xi_i - y_i(w^T x_i + b) \leq 0$, $\xi_i \geq 0, \forall i$,

consider its objective $\frac{1}{2}\|w\|^2 + C\sum_i^m \xi_i$, since $C>0$, then $\cancel{\frac{1}{2}\|w\|^2+}$ $\frac{1}{2}\|w\|^2 + C\sum_i^m \xi_i \geq \frac{1}{2}\|w\|^2$, and we're doing minimize here, so it guarantee that $\xi_i = 0, \forall i$.

$\therefore$ The resulting boundary guaranteed to separate the classes.

4.

(1) The dual problem is: $\max \sum_{i}^{m} \alpha_i - \frac{1}{2}\sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$

$$\text{s.t.} \quad \sum_{i}^{m} \alpha_i y_i = 0$$
$$\alpha_i \geq 0$$

So we can write it as: $\max \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\alpha_1^2 - \frac{1}{2}\alpha_2^2 - \frac{1}{2}\alpha_3^2 - \frac{1}{2}\alpha_4^2 - \alpha_1\alpha_3 - \alpha_2\alpha_4$

$$\text{s.t.} \quad -\alpha_1 - \alpha_2 + \alpha_3 \alpha_4 = 0$$
$$\alpha_i \geq 0$$

Then we have:

$$\left.\begin{aligned} \nabla_{\alpha_1} f &= 1 - \alpha_1 - \alpha_3 = 0 \\ \nabla_{\alpha_2} f &= 1 - \alpha_2 - \alpha_4 = 0 \\ \nabla_{\alpha_3} f &= 1 - \alpha_3 - \alpha_1 = 0 \\ \nabla_{\alpha_4} f &= 1 - \alpha_4 - \alpha_2 = 0 \end{aligned}\right. \Rightarrow \begin{cases} \alpha_1 = \frac{1}{2} \\ \alpha_2 = \frac{1}{2} \\ \alpha_3 = \frac{1}{2} \\ \alpha_4 = \frac{1}{2} \end{cases}$$

$W = \sum_{i}^{m} \alpha_i y_i x_i = (-1, -1)$, $b = \frac{1}{4}\sum_{i}^{m}(y_i - W^T x_i) = 0$

$\therefore$ The svm classifier is $f(x) = -x_1 - x_2$ for this data set.

(2) The data points with $\alpha > 0$ are the support vectors.

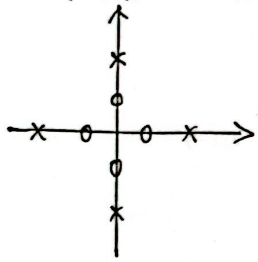
$\therefore$ The data points $(1\ 0)$, $(0\ 1)$, $(-1\ 0)$, $(0\ -1)$ are support vectors.

(3) $f(x) = -1 - 2 = -3 < 0$

$\therefore$ The label of $[1;2]$ is predicted as class -1.

5.

(1) Let first take a look on the graph:



We can see that the data set is not linearly separable.
And we can't find such a svm classifier without slack variable to make every points satisfy the constrain $1 - y_i(w^T x_i + b) \le 0$

So we can not find a svm classifier (with linear kernel, without slack variable) for the data set.

However if we use kernels, we may find one, as we do in the next problem. ($\phi(x) = [x_1^2; x_2^2]$).

(2) After the expanding, the new data is: class -1: $\begin{bmatrix} (1 & 0) \\ (0 & 1) \end{bmatrix}$ and class $+1$: $\begin{bmatrix} (4 & 0) \\ (0 & 4) \end{bmatrix}$.

$\therefore$ The dual problem is: $\max\ a_1 + a_2 + a_3 + a_4 - \frac{1}{2}a_1^2 - \frac{1}{2}a_2^2 - 8a_3^2 - 8a_4^2 + 4a_1a_3 + 4a_2a_4$.

s.t. $-a_1 - a_2 + a_3 + a_4 = 0$.

$a_i \geq 0,\ \forall i$.

$\nabla_{a_1} f = 1 - a_1 + 4a_3 = 0$

$\nabla_{a_2} f = 1 - a_2 + 4a_4 = 0$

$\nabla_{a_3} f = 1 - 16a_3 + 4a_1 = 0$

$\nabla_{a_4} f = 1 - 16a_4 + 4a_2 = 0$

$\Rightarrow$ Not feasible, use SMO to solve it.

First assume that $a_1 = a_2 = a_3 = a_4 = \frac{1}{4}$.

① Choose pair a_1, a_3. $W(a_1, a_3) = a_1 + a_3 - \frac{1}{2}a_1^2 - 8a_3^2 + 4a_1a_3 + \frac{1}{2} - \frac{1}{32} - \frac{1}{2} + \frac{1}{4}$

And $a_1 = a_3 \Rightarrow W(a_1) = 2a_1 - \frac{9}{2}a_1^2 + \text{constant}$.

$\nabla W = 2 - 9a_1 = 0 \Rightarrow a_1 = a_3 = \frac{2}{9}$

② Choose pair a_2, a_4, we can also update that $a_2 = a_4 = \frac{2}{9}$

③ Now choose pair a_1, a_2.

$W(a_1, a_2) = a_1 + a_2 - \frac{1}{2}a_1^2 - \frac{1}{2}a_2^2 + \frac{16}{9}a_1 + \frac{16}{9}a_2 + \text{constant}$

$= \frac{25}{9}a_1 + \frac{25}{9}a_2 - \frac{1}{2}a_1^2 - \frac{1}{2}a_2^2 + \text{constant}$. and $a_2 = \frac{4}{9} - a_1$

$\therefore W(a_1) = - \frac{1}{2}a_1^2 - \frac{1}{2}(\frac{4}{9} - a_1)^2 + \text{constant}$

$\nabla W = -a_1 + (\frac{4}{9} - a_1) = 0 \Rightarrow a_1 = a_2 = \frac{2}{9} \Rightarrow a_1, a_2$ doesn't change

④ Choose a_3, a_4:

$W(a_3, a_4) = a_3 + a_4 - 8a_3^2 - 8a_4^2 + \frac{16}{9}a_3 + \frac{16}{9}a_4 + \text{constant}$ and $a_4 = \frac{8}{9} - a_3$

$\nabla W = 0 \Rightarrow a_3 = a_4 = \frac{2}{9}$

④ Choose α_1, α_4.

$W(\alpha_1, \alpha_4) = \alpha_1 + \alpha_4 - \frac{1}{2}\alpha_1^2 - 8\alpha_4^2 + \frac{8}{9}\alpha_1 + \frac{8}{9}\alpha_4 + \text{constant}$ and $\alpha_1 = \alpha_4$

$\Rightarrow W(\alpha_1) = \frac{25}{9}\alpha_1 - \frac{17}{2}\alpha_1^2 \quad \nabla W = 0 \Rightarrow \alpha_1 = \alpha_4 = \frac{2}{9}$. α_1, α_4 not change

⑤ Choose α_2, α_3. Also get $\alpha_2 = \alpha_3 = \frac{2}{9}$, α_2, α_3 not change.

So now the sequence converge, and we get $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \frac{2}{9}$,

$w = \sum_{i=1}^{4} \alpha_i y_i \phi(x_i) = (\frac{2}{3}, \frac{2}{3})^T$. $b = \frac{1}{4} \cdot \sum_{i=1}^{4} (y_i - w^T\phi(x_i)) = \frac{1}{4} \times (-\frac{20}{3}) = -\frac{5}{3}$.

$\therefore$ The decision boundary is $w^T\phi(x) + b = 0 \Rightarrow \frac{2}{3}x_1^2 + \frac{2}{3}x_2^2 - \frac{5}{3} = 0$.

The svm is $f(x) = w^T\phi(x) + b = \frac{2}{3}x_1^2 + \frac{2}{3}x_2^2 - \frac{5}{3}$.

$\because$ The point $[1;2]$ has the function value $f([1;2]) = \frac{1}{3} > 0$.

$\therefore$ We predict it as class: $+1$.

6. Let $f = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m)$.

At the optimal point, we'll have:

$$\nabla_{a_i} f = 1 - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n t_n t_m k(x_n, x_m) - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_m t_n t_m k(x_n, x_m).$$

$$= 1 - \sum_{n=1}^{N}\sum_{m=1}^{N} a_n t_n t_m k(x_n, x_m) = 0$$

Mutiple a_i, we get $a_i = a_i \sum_{n=1}^{N}\sum_{m=1}^{N} a_n t_n t_m k(x_n, x_m)$, for $\forall i = 1, \cdots, N$

Then, sum all the equations and we can get:

$$\sum_{n=1}^{N} a_n = (a_1 + a_2 + \cdots + a_n) \sum_{n=1}^{N}\sum_{m=1}^{N} a_n t_n t_m k(x_n, x_m)$$

$$\Rightarrow \sum_{n=1}^{N} a_n = \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m)$$

By the stationarity condition of the lagrange function, we have:

$$W^2 = \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m).$$

And we also know that the margin $\gamma = \frac{1}{\|W\|}$

$$\therefore \frac{1}{\gamma^2} = W^2 = \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m) = \sum_{n=1}^{N} a_n$$

$$\therefore \frac{1}{\gamma^2} = \sum_{n=1}^{N} a_n$$

Part2: Programming Questions

One vs Rest Strategy Implementation

I directly use the OneVsRestClassifier in sklearn package to implement the one vs rest strategy.

Question1

We are solving the linear SVM problem without slack in this question, and we can derive it as following:

Question 1

Primal problem: $\min \frac{1}{2}\|w\|^2$

s.t. $1 - y_i(w^T x_i + b) \le 0,\ \forall i$

Lagrange: $L(w, \alpha, b) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(w^T x_i + b))$

KKT: $\frac{dL}{d\alpha} = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{m} \alpha_i y_i x_i$

$\frac{dL}{db} = -\sum_{i=1}^{m} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$

$\therefore L = \frac{1}{2}\sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{m} \alpha_i - \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{m} \alpha_i y_i b$

$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$

$\therefore$ The dual problem is:

$\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$

s.t. $\sum_{i=1}^{m} \alpha_i y_i = 0,\ \alpha_i \ge 0,\ \forall i$

In this problem, x_i is the features and y_i is the class label, and α_i is the dual variable. which determine whether a point is a support vector. If $\alpha_i > 0$, it's a support vector. Otherwise, it's not.

The performance of this model on the datasets is: training_error = 0.041666666666666664 and testing_error = 0.0 .

We can know that all the three classes are linearly separable with SVM without slack and we can know it because for each class the SVM problem without slack is feasible which means we can find a linear hyperplane that separate this class with the other two. Hence all the three classes are linearly separable.

Question2

We are solving the linear SVM problem with slack in this question, and we can derive it as following:

Question 2

In the basic problem of question 1, we add slack variable to each point, and it makes the primal problem become:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t. } 1-\xi_i - y_i(w^T x_i + b) \le 0, \forall i$$

$$\xi_i \ge 0, \forall i$$

$$L(w,b,\alpha,\xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\left[\alpha_i(1-\xi_i - y_i(w^T x_i + b)) - \mu_i \xi_i\right].$$

$$\frac{dL}{dw} = 0 \Rightarrow w = \sum_{i=1}^{m}\alpha_i y_i x_i, \quad \frac{dL}{db} = 0 \Rightarrow \sum_{i=1}^{m}\alpha_i y_i = 0.$$

$$\frac{dL}{d\xi_i} = 0 \Rightarrow \alpha_i = C - \mu_i$$

$$\therefore L(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

The dual problem is: $\max \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j$

$\text{s.t. } \sum_{i=1}^{m}\alpha_i y_i = 0$

$0 \le \alpha_i \le C, \forall i$

When $\alpha_i = 0$, the point is outside the margin and it's not a support vector.

$0 < \alpha_i < C$, the point lies on the margin and it's a support vector, $\xi_i = 0$

$\alpha_i = C$, the point is inside the margin and it's a support vector, and its $\xi_i > 0$.

ξ_i is the distance from the point to the margin and $\xi_i > 0$ only if $\alpha_i = C$.

The performance of this model on the datasets is:
C=0.1:training_error:0.125. testing_error:0.23333333333333334
C=0.2:training_error:0.058333333333333334
testing_error:0.16666666666666666
C=0.3:training_error:0.05. testing_error:0.13333333333333333
C=0.4:training_error:0.05. testing_error:0.1
C=0.5:training_error:0.05. testing_error:0.1
C=0.6:training_error:0.05. testing_error:0.1
C=0.7:training_error:0.05. testing_error:0.1
C=0.8:training_error:0.05. testing_error:0.1
C=0.9:training_error:0.05. testing_error:0.06666666666666667
C=1.0:training_error:0.05. testing_error:0.06666666666666667

Question3

We are solving the linear SVM problem with kernel functions and slack variables in this question, and we can derive it as following:

Question 3.

The only difference between question 2 and 3 is question 3 use a kernel function $k(x_i, x_j)$ to replace the term $x_i^T x_j$ in question 2.
So we can write a general problem for question 3:

Dual: $\max \sum_{i=1}^{m} \alpha_i - \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$

s.t. $\sum_{i=1}^{m} \alpha_i y_i = 0.$

$0 \le \alpha_i \le C, \forall i.$

3.(a) The second-order polynomial kernel: $k(x_i, x_j) = (x_i^T x_j)^2$

So we're solving the problem:

$\max \sum_{i=1}^{m} \alpha_i - \sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j)^2$

s.t. $\sum_{i=1}^{m} \alpha_i y_i = 0$

$0 \le \alpha_i \le C, \forall i$

3.(b) The third-order polynomial kernel: $k(x_i, x_j) = (x_i^T x_j)^3$

And ~~We~~ we're solving: $\max \sum_{i=1}^{m} \alpha_i - \sum_{i,j}^{m} \alpha_i, \alpha_j y_i y_j (x_i^T x_j)^3$

s.t. $\sum_{i=1}^{m} \alpha_i y_i = 0$

$0 \le \alpha_i \le C, \forall i$

3.(c) kernel: $k(x_i, x_j) = \exp(-\frac{1}{2}\|x_i - x_j\|^2)$, $\sigma = 1$

Problem: $\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j \exp(-\frac{1}{2}\|x_i - x_j\|^2)$

s.t. $\sum_{i=1}^{m} \alpha_i y_i = 0$

$0 \le \alpha_i \le C$

3.(d) kernel: $k(x_i, x_j) = \frac{1}{1+\exp(-x_i^T x_j)}$

$\sigma = 1, b = 0.$

Problem: $\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j}^{m} \alpha_i \alpha_j y_i y_j \exp\left(\frac{1}{1+\exp(-x_i^T x_j)}\right)$

s.t. $\sum_{i=1}^{m} \alpha_i y_i = 0$

$0 \le \alpha_i \le C, \forall i$

The performance of this model on the datasets is:
(a) 2nd-order polynomial kernel:
training_error:0.03333333333333333
testing_error:0.03333333333333333

(b) 3rd-order polynomial kernel:
training_error:0.025. testing_error:0.0

(c) Radial Basis Function kernel with σ = 1:
training_error:0.025. testing_error:0.03333333333333333

(d) Sigmoidal kernel with σ = 1:
training_error:0.6666666666666666
testing_error:0.6666666666666666