THE CHINESE UNIVERSITY OF HONG KONG, SHENZHEN

DDA 2020

MACHINE LEARNING

Assignment2 Report

Name: Xiang Fei

Student ID: 120090414

Contents

1. Written Questions

1.1 Question 1

1.2 Question 2

1.3 Question 3

1.4 Question 4

1.5 Question 5

1.6 Question 6

2. Programming Question

2.1 Question restatement

2.3 One vs rest strategy (Explanation and implementation)

2.4 Results (including the derivation of the optimization problem)

1、Written Questions

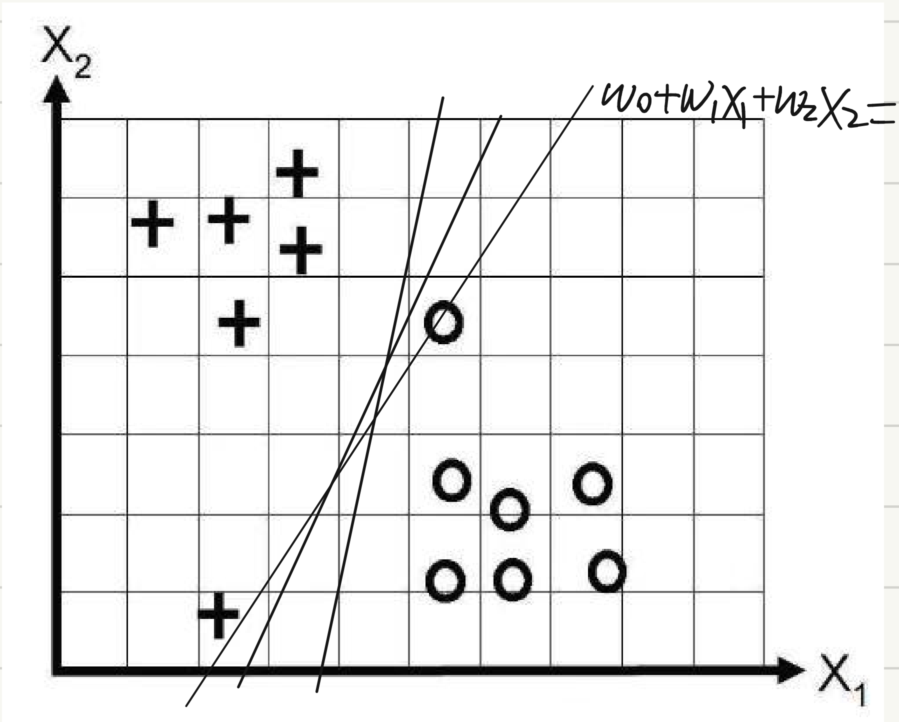
1.1 Question 1

a、 the decision boundary is given by:

$$w_0 + w_1 x_1 + w_2 x_2 = 0$$

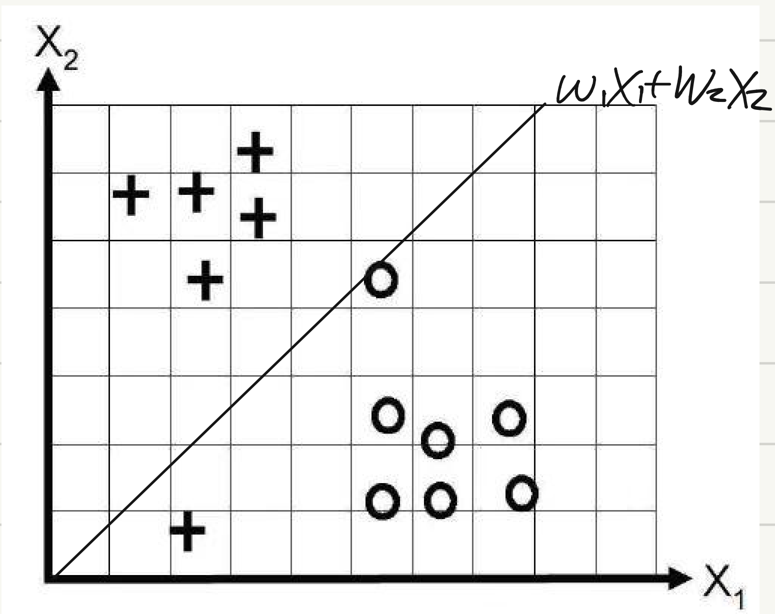
it is a line depends on $w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$

the decision boundary is not unique



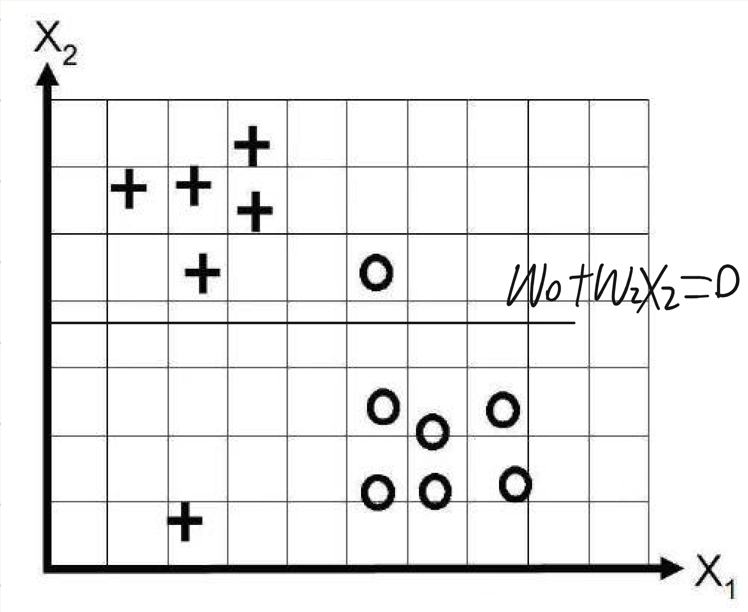
there is no classification error

b、 since w_0 all the way to 0, the decision is a line that passes the origin. $w_1 x_1 + w_2 x_2 = 0$



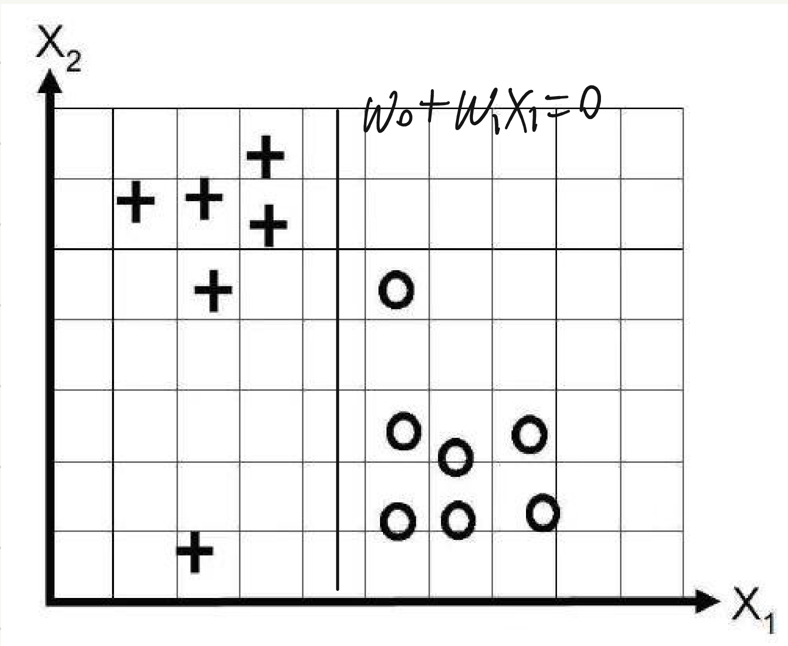
there is 1 classification error

c. w_1 all the way to 0, so the decision boundary is a line parallel to X_1 axis.



there are 2 classification errors.

d. W_2 all the way to 0, so the decision boundary is a line parallel to X_2 axis.



there is no classification error.

1.2 Question 2

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

$\phi(x_2)=[1,2,2]^T$

decision boundary has normal $[0,1,1]^T$ and passes the point $[1,1,1]^T$.

$\therefore$ a vector that is parallel to w is also perpendicular to the decision boundary, so the vector can be $[0,1,1]^T$

b. the margin is the distance between $[1,1,1]^T$ and $[1,2,2]^T$

$$\sqrt{(2-1)^2+(2-1)^2+(1-1)^2}=\sqrt{2}$$

c. $\frac{1}{\|w\|}=\sqrt{2} \Rightarrow \frac{1}{\|w\|}=\frac{\sqrt{2}}{2}$

$$w=k\cdot[0,1,1]^T$$

$$\therefore \sqrt{k^2+k^2}=\sqrt{2}k=\frac{\sqrt{2}}{2} \Rightarrow k=\frac{1}{2}$$

$$\therefore w=[0,\tfrac{1}{2},\tfrac{1}{2}]^T$$

d. $w^T\phi(x_1)=0$

$$\therefore -w_0=1 \Rightarrow w_0=-1$$

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

1.3 Question 3

No, the resulting decision boundary can't guaranteed to separate the classes, since the margin can be increased by considering the slack variable, which allows some points locate inside the margin and may appear on the wrong side.

1.4 Question 4

(1) prime problem : $\min_{w,b} \frac{1}{2}\|w\|^2$

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

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

the corresponding dual problem :

$$\max_\alpha \sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}^m \alpha_i\alpha_j y_i y_j X_i^T X_j$$

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

$\therefore \max_\alpha \sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}^m \alpha_i\alpha_j y_i y_j X_i^T X_j$

$= \max_\alpha \alpha_1+\alpha_2+\alpha_3+\alpha_4-\frac{1}{2}\alpha_1^2-\frac{1}{2}\alpha_2^2-\frac{1}{2}\alpha_3^2-\frac{1}{2}\alpha_4^2-\alpha_1\alpha_3-\alpha_2\alpha_4$

$= \max_\alpha g(\alpha)$

we have $-\alpha_1-\alpha_2+\alpha_3+\alpha_4=0$

$\Rightarrow \max_\alpha g(\alpha) = 2\alpha_1+2\alpha_2-\alpha_1^2-2\alpha_2^2-\alpha_3^2-2\alpha_1\alpha_2+2\alpha_2\alpha_3$

$\frac{\partial g}{\partial \alpha_1} = 2-2\alpha_1-2\alpha_2=0$

$\frac{\partial g}{\partial \alpha_2} = 2-4\alpha_2-2\alpha_1+2\alpha_3=0 \quad \Rightarrow \quad \alpha_1+\alpha_2=\alpha_3+\alpha_4=1,\ \alpha_2=\alpha_3$

$\frac{\partial g}{\partial \alpha_3} = -2\alpha_3+2\alpha_2=0$

$$w=-\alpha_1\begin{bmatrix}1\\0\end{bmatrix}-\alpha_2\begin{bmatrix}0\\1\end{bmatrix}+\alpha_3\begin{bmatrix}-1\\0\end{bmatrix}+\alpha_4\begin{bmatrix}0\\-1\end{bmatrix}=\begin{bmatrix}-\alpha_1-\alpha_3\\-\alpha_2-\alpha_4\end{bmatrix}=\begin{bmatrix}-\alpha_1-\alpha_2\\-\alpha_3-\alpha_4\end{bmatrix}=\begin{bmatrix}-1\\-1\end{bmatrix}$$

$$b=-1-(-\alpha_1-\alpha_3)=-1+\alpha_1+\alpha_3=-1+\alpha_1+\alpha_2=-1+1=0$$

$\therefore$ the svm : $w=\begin{bmatrix}-1\\-1\end{bmatrix},\ b=0$

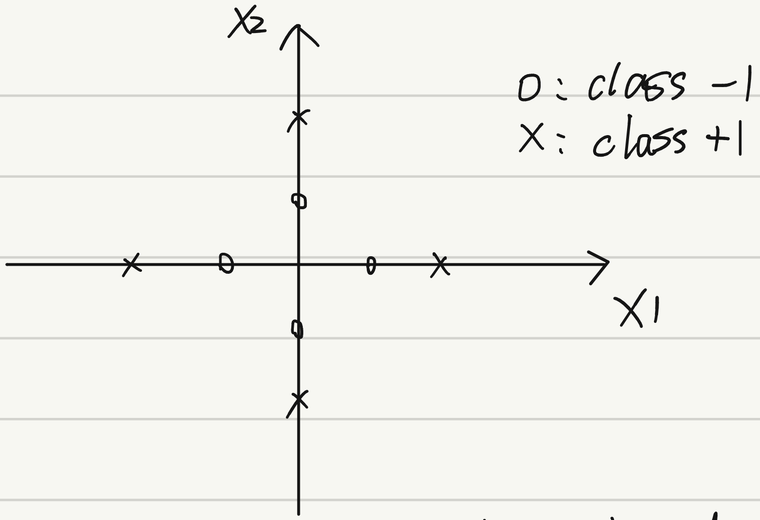
(2) since $\alpha_1,\alpha_2,\alpha_3,\alpha_4>0$, so the four given data points are all support vectors.

(3)
$$w^T x+b = \begin{bmatrix}-1 & -1\end{bmatrix}\begin{bmatrix}1\\2\end{bmatrix} = -3<0$$

$\therefore$ the predicted label of [1; 2] is -1

1.5 Question 5

(1)



Yes, we can use kernel to make the data points become separable.

(2) dual problem :

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

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

$$\phi(x_1) = [1;0] = \phi(x_3)$$

$$\phi(x_2) = [0;1] = \phi(x_4)$$

$$\phi(x_5) = [4;0] = \phi(x_7)$$

$$\phi(x_6) = [0;4] = \phi(x_8)$$

$$\max_{\alpha} g(\alpha)$$

$$= \max_{\alpha} \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\left(\alpha_1^2 + \alpha_2^2 + 16\alpha_3^2 + 16\alpha_4^2 - 8\alpha_1\alpha_3 - 8\alpha_2\alpha_4\right)$$

$$= \max_{\alpha} \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\alpha_1^2 - \frac{1}{2}\alpha_2^2 - 8\alpha_3^2 - 8\alpha_4^2 + 4\alpha_1\alpha_3 + 4\alpha_2\alpha_4$$

$$\text{s.t. } -\alpha_1 - \alpha_2 + \alpha_3 + \alpha_4 = 0, \text{ if } \alpha_1, \alpha_2, \alpha_3, \alpha_4 > 0$$

$$\therefore \max_{\alpha} 2\alpha_1 + 2\alpha_2 - \frac{17}{2}\alpha_1^2 - \frac{9}{2}\alpha_2^2 - 16\alpha_3^2 - 12\alpha_1\alpha_2 + 20\alpha_1\alpha_3 + 12\alpha_2\alpha_3$$

$$\triangleq \max_{\alpha} g(\alpha)$$

$$\frac{\partial g}{\partial \alpha_1} = 2 - 17\alpha_1 - 12\alpha_2 + 20\alpha_3 = 0, \quad \frac{\partial g}{\partial \alpha_2} = 2 - 9\alpha_2 - 12\alpha_1 + 12\alpha_3 = 0$$

$$\frac{\partial g}{\partial \alpha_3} = -32\alpha_3 + 20\alpha_1 + 12\alpha_2 = 0$$

$$w = \begin{bmatrix} -\alpha_1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -\alpha_2 \end{bmatrix} + \begin{bmatrix} 4\alpha_3 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 4\alpha_4 \end{bmatrix} = \begin{bmatrix} 4\alpha_3 - \alpha_1 \\ 4\alpha_4 - \alpha_2 \end{bmatrix}$$

$\therefore w = \begin{bmatrix} \frac{2}{3} & \frac{2}{3} \end{bmatrix}^T, \quad b = -1 - (-a_1 + 4a_3) = -\frac{5}{3}$

for $[1;2]$, $\phi(x) = [1;4]$

$$w^T\phi(x) + b = \frac{2}{3} + \frac{2}{3} \times 4 - \frac{5}{3} = \frac{5}{3} > 0$$

$\therefore$ the predicted label of $[1;2]$ is $+1$

1.6 Question 6

let x_i be a support vector and y_i is its label

$$r = \frac{y_i(w^T x_i + b)}{\|w\|}$$

since x_i is a support vector, we have:

$$y_i(w^T x_i + b) = 1$$

$$\therefore r = \frac{1}{\|w\|} \Rightarrow \frac{1}{r^2} = \|w\|^2$$

$$w = \sum_{n=1}^{N} a_n t_n x_n$$

$$\|w\|^2 = w^T w = w^T \sum_{n=1}^{N} a_n t_n x_n$$

$$= \sum_{n=1}^{N} a_n t_n w^T x_n$$

we have $\sum_{n=1}^{N} a_n t_n = 0$

$\therefore$ multiply a constant b: $\sum_{n=1}^{N} a_n t_n b = 0$

$$\therefore \|w\|^2 = \sum_{n=1}^{N} a_n t_n w^T x_n + \sum_{n=1}^{N} a_n t_n b$$

$$= \sum_{n=1}^{N} a_n t_n (w^T x_n + b)$$

for support vector, $a_n t_n (w^T x_n + b) = a_n$

otherwise: $a_n = 0 \Rightarrow a_n t_n (w^T x_n + b) = 0$

$\therefore \|w\|^2 = \sum_{n=1}^{N} a_n \quad \therefore \frac{1}{r^2} = \sum_{n=1}^{N} a_n$. Q.E.D.

2. Programming Question

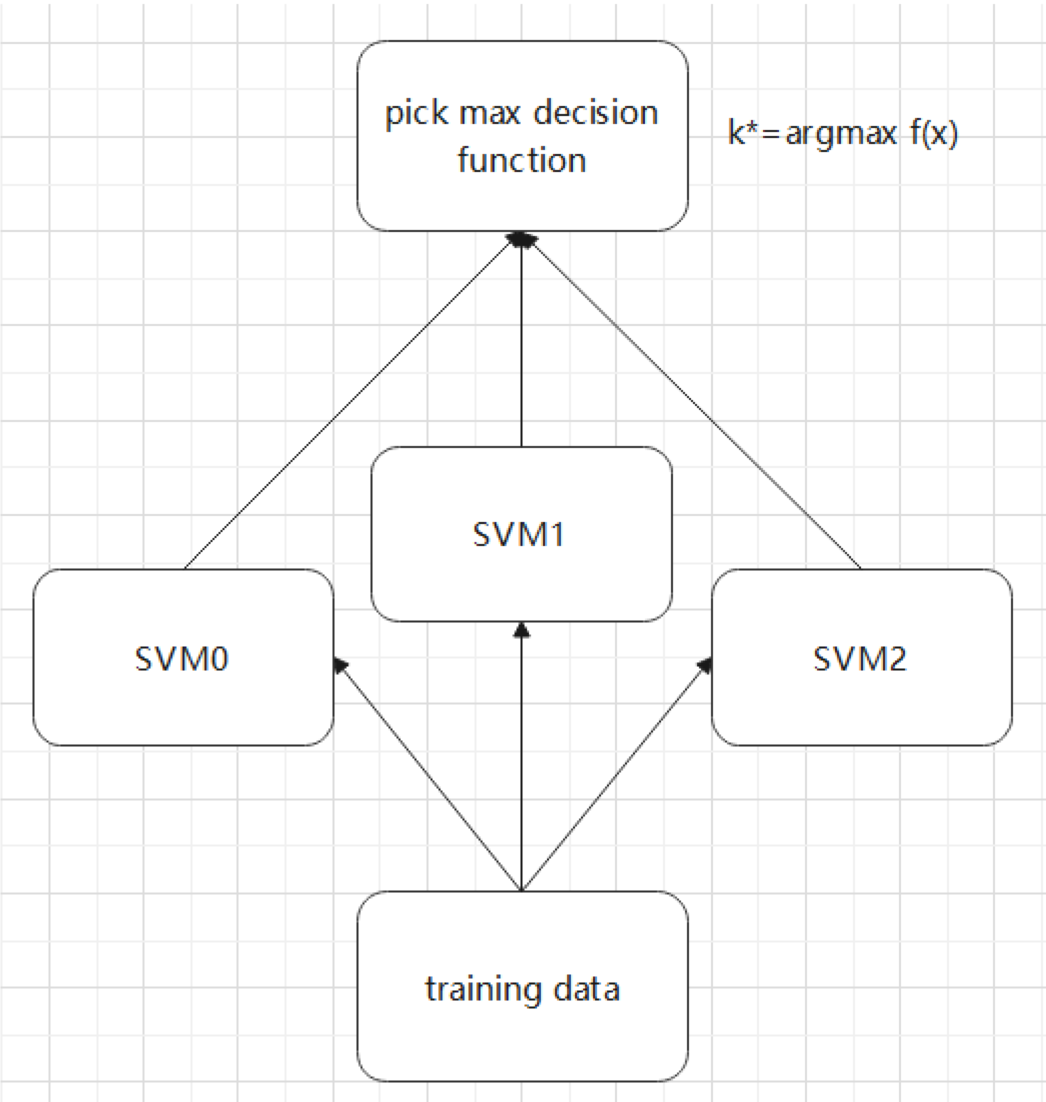
2.1 Question restatement

In this programming problem, we need to write a program that construct support vector machine models with different kernel functions and slack variable. We need to implement hard margin svm, soft margin svm, svm with Polynomial kernel(degree 2 and 3), svm with Radial Basis Function (RBF) kernel($\sigma = 1$) and svm with Sigmoidal kernel($\sigma = 1$). The datasets we use is one of most popular classification dataset in machine learning area: Iris dataset, including 120 training data and 30 testing data, respectively. The datasets contains 3 different classes(labels) of 50 instances each: setosa(label 0), versicolor(label 1) and virginica(label 2), which represents the type of iris plant. And for each instance, it has four different features: sepal width, sepal length, petal width and petal length.

2.3 One vs rest strategy (Explanation and implementation)

2.3.1 Explanation

one vs rest strategy is an approach used in Multi-class SVM. To classify multiple classes, we use this strategy to convert K binary classification to a K-class classification. The following graph shows the processing logic.



2.3.2 Implementation

We use sklearn.svm.SVC to construct support vector machine, since it provides attributes about support vectors while others don't provide these attributes. However, there is a problem that SVC only implement one vs one strategy. As a result, we need to implement one vs rest strategy manually. We can manually aggregating two classes into one and construct three binary classification svm using one vs one strategy. And then, we compare the decision function of each svm, and find the one which has the maximum decision function value. Finally, we can get the predicted label of given data. The python code is in the following (linear svm as example).

- construct three binary support vector machine and compute the corresponding parameters and support indices.

```
manual_Ytrain = [0]*len(self.Y_train)
clf1 = SVC(C=C,kernel='linear',decision_function_shape='ovo')
for i in range(len(self.Y_train)):
    if self.Y_train[i] == 0:
        manual_Ytrain[i] = 1
    else:
        manual_Ytrain[i] = -1
clf1.fit(self.X_train,manual_Ytrain)
w_setosa = clf1.coef_[0]
b_setosa = clf1.intercept_[0]
svi_setosa = clf1.support_

clf2 = SVC(C=C,kernel='linear',decision_function_shape='ovo')
for i in range(len(self.Y_train)):
    if self.Y_train[i] == 1:
        manual_Ytrain[i] = 1
    else:
        manual_Ytrain[i] = -1
clf2.fit(self.X_train,manual_Ytrain)
w_versicolor = clf2.coef_[0]
b_versicolor = clf2.intercept_[0]
svi_versicolor = clf2.support_

clf3 = SVC(C=C,kernel='linear',decision_function_shape='ovo')
for i in range(len(self.Y_train)):
    if self.Y_train[i] == 2:
        manual_Ytrain[i] = 1
    else:
        manual_Ytrain[i] = -1
clf3.fit(self.X_train,manual_Ytrain)
w_virginica = clf3.coef_[0]
b_virginica = clf3.intercept_[0]
svi_virginica = clf3.support_
```

- find the max decision function value of training data and test data to do the classification.

```
d1 = clf1.decision_function(self.X_train)
d2 = clf2.decision_function(self.X_train)
d3 = clf3.decision_function(self.X_train)
Y_train_pred = [0]*len(self.Y_train)
for i in range(len(self.Y_train)):
    f1 = d1[i]
    f2 = d2[i]
    f3 = d3[i]
    if f1>=f2 and f1>=f3:
        Y_train_pred[i] = 0
    elif f2>=f1 and f2>=f3:
        Y_train_pred[i] = 1
    else:
        Y_train_pred[i] = 2
d1_test = clf1.decision_function(self.X_test)
d2_test= clf2.decision_function(self.X_test)
d3_test = clf3.decision_function(self.X_test)
Y_test_pred = [0]*len(self.Y_test)
for i in range(len(self.Y_test)):
    f1_test = d1_test[i]
    f2_test = d2_test[i]
    f3_test = d3_test[i]
    if f1_test>=f2_test and f1_test>=f3_test:
        Y_test_pred[i] = 0
    elif f2_test>=f1_test and f2_test>=f3_test:
        Y_test_pred[i] = 1
    else:
        Y_test_pred[i] = 2
```

2.4 Results (including the derivation of the optimization problem)

2.4.1 Question 1: Standard SVM model

The primal problem:

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \ s.t. 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \leq 0, \forall i$$

The dual problem:

$$\max_{\alpha} \sum_i^m \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j \ s.t. \sum_i^m \alpha_i y_i = 0, \alpha_i \geq 0, \forall i$$

Since SVC doesn't support hard margin, or perfect separable data. Therefore, I simulate the standard SVM model by setting the penal parameter $C = 1e5$.

I compute the training error and testing error as the following:

```
train_error = 1 - sum(Y_train_pred==self.Y_train)/len(self.Y_train)
test_error = 1 - sum(Y_test_pred==self.Y_test)/len(self.Y_test)
```

The error:

$training\ error = 0.04166666666666663$

$testing\ error = 0.0$

Linear separable problem

Since the data provided is not necessarily linearly separable, therefore we need to find out which classes and the rest are not linearly separable. In fact, we can just check the training error of the three ovo model respectively. If the error is not 0, then the class and the rest are not linearly separable. The python code is like the following:

```
svm1_train_pred = clf1.predict(self.X_train)
svm1_train_error = 1 - sum(svm1_train_pred==manual_Ytrain)/len(manual_Ytrain)
print(svm1_train_error)
svm2_train_pred = clf2.predict(self.X_train)
svm2_train_error = 1 - sum(svm2_train_pred==manual_Ytrain)/len(manual_Ytrain)
print(svm2_train_error)
svm3_train_pred = clf3.predict(self.X_train)
svm3_train_error = 1 - sum(svm3_train_pred==manual_Ytrain)/len(manual_Ytrain)
print(svm3_train_error)
```

And we get the result:

```
0.0
0.21666666666666667
0.016666666666666672
```

It means the training error of model1 is 0, and the training error of model2 and model3 are bigger than 0. So we get the result:

Class1 and the rest are linearly separable.

Class2 and the rest are not linearly separable.

Class3 and the rest are not linearly separable.

2.4.2 Question 2: SVM with slack variables (linear kernel)

The primal problem:

$$\min_{w,b,\xi_i} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i^m \xi_i \ s.t. 1 - \xi_i - y_i(\boldsymbol{w^T x_i} + b) \leq 0, \xi_i \geq 0, \forall i$$

Lagrange function:

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i^m \xi_i + \sum_i^m [\alpha_i(1 - \xi_i - y_i(\boldsymbol{w^T x_i} + b)) - u_i\xi_i]$$

KKT condition:

Stationarity:

$$\frac{\partial L}{\partial \boldsymbol{w}}=0,\ \ \boldsymbol{w}=\sum_i^m \alpha_i y_i \boldsymbol{x_i}\ \frac{\partial L}{\partial b}=0,\ \ \sum_i^m a_i y_i=0\ \frac{\partial L}{\partial \xi_i}=0,\ \ \alpha_i=C-\mu_i,\forall i$$

Feasibility:

$$\alpha_i \geq 0, 1-\xi_i-y_i(\boldsymbol{w^T}\boldsymbol{x_i}+b)\leq 0, \xi_i\geq 0, \mu_i\geq 0, \forall i$$

Complementary slackness:

$$\alpha_i(1-\xi_i-y_i(\boldsymbol{w^T}\boldsymbol{x_i}+b))=0, \mu_i\xi_i=0, \forall i$$

Then, we can get the dual problem:

$$\max_{\alpha}\sum_i^m \alpha_i\alpha_j y_i y_j \boldsymbol{x_i^T}\boldsymbol{x_j}\ s.t.\sum_i^m \alpha_i y_i=0, \alpha_i\geq 0, \forall i$$

For each $C=0.1\times t, t=1,2,\ldots,10$, I fit my algorithm and get the error.

- $C=0.1$:$training\ error=0.125$; $testing\ error=0.23333333333333328$
- $C=0.2$:$training\ error=0.05833333333333335$; $testing\ error=0.16666666666666663$
- $C=0.3$:$training\ error=0.050000000000000044$; $testing\ error=0.1333333333333333$
- $C=0.4$:$training\ error=0.050000000000000044$; $testing\ error=0.09999999999999998$
- $C=0.5$:$training\ error=0.050000000000000044$; $testing\ error=0.09999999999999998$
- $C=0.6$:$training\ error=0.050000000000000044$; $testing\ error=0.09999999999999998$
- $C=0.7$:$training\ error=0.050000000000000044$; $testing\ error=0.09999999999999998$
- $C=0.8$:$training\ error=0.050000000000000044$; $testing\ error=0.09999999999999998$
- $C=0.9$:$training\ error=0.050000000000000044$; $testing\ error=0.06666666666666665$
- $C=1.0$:$training\ error=0.050000000000000044$; $testing\ error=0.06666666666666665$

2.4.3 Question 3: SVM with 2nd-order polynomial kernel and slack variables

The primal problem:

$$\min_{w,b,\xi_i}\frac{1}{2}||\boldsymbol{w}||^2+C\sum_i \xi_i\ s.t. 1-y_i(\boldsymbol{w^T}\boldsymbol{\phi(x_i)}+b)\leq 0, \forall i$$

Lagrange function:

$$L(\boldsymbol{w},b,\alpha)=\frac{1}{2}||\boldsymbol{w}||^2+C\sum_i^m \xi_i+\sum_i^m[\alpha_i(1-\xi_i-y_i(\boldsymbol{w^T}\boldsymbol{\phi(x_i)}+b))-u_i\xi_i]$$

KKT condition:

Stationarity:

$$\frac{\partial L}{\partial \boldsymbol{w}}=0,\ \ \boldsymbol{w}=\sum_i^m \alpha_i y_i \phi(\boldsymbol{x_i})\ \frac{\partial L}{\partial b}=0,\ \ \sum_i^m a_i y_i=0\ \frac{\partial L}{\partial \xi_i}=0,\ \ \alpha_i=C-\mu_i,\forall i$$

Feasibility:

$$\alpha_i \geq 0, 1-\xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i})+b) \leq 0, \xi_i \geq 0, \mu_i \geq 0, \forall i$$

Complementray slackness:

$$\alpha_i(1-\xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i})+b)) = 0, \mu_i\xi_i = 0, \forall i$$

Then, we can get the dual problem:

$$\max_{\alpha_i} \sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_i y_i y_i \phi(\boldsymbol{x_i})^T\phi(\boldsymbol{x_j})$$

Here, we use the kernel ($\gamma = 1$):

$$k(\boldsymbol{x}, \boldsymbol{x_i}) = (\boldsymbol{x^T}\boldsymbol{x_i})^2$$

Therefore, we can get the final dual problem

$$\max_{\alpha_i} \sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j (\boldsymbol{x^T}\boldsymbol{x_i})^2 \ s.t. \sum_i^m \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, \forall i$$

We set the panel parameter $C = 1$ and get the error:

$training\ error = 0.025000000000000022$

$testing\ error = 0.0$

2.4.4 Question 4: SVM with 3nd-order polynomial kernel and slack variables

The primal problem:

$$\min_{w,b,\xi_i} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i \ s.t. 1 - y_i(\boldsymbol{w^T}\boldsymbol{\phi(x_i)}+b) \leq 0, \forall i$$

Lagrange function:

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i^m \xi_i + \sum_i^m[\alpha_i(1-\xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i})+b)) - u_i\xi_i]$$

KKT condition:

Stationarity:

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0,\ \ \boldsymbol{w} = \sum_i^m \alpha_i y_i \phi(\boldsymbol{x_i})\ \frac{\partial L}{\partial b} = 0,\ \ \sum_i^m a_i y_i = 0\ \frac{\partial L}{\partial \xi_i} = 0,\ \ \alpha_i = C - \mu_i, \forall i$$

Feasibility:

$$\alpha_i \geq 0, 1-\xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i})+b) \leq 0, \xi_i \geq 0, \mu_i \geq 0, \forall i$$

Complementray slackness:

$$\alpha_i(1-\xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i})+b)) = 0, \mu_i\xi_i = 0, \forall i$$

Then, we can get the dual problem:

$$\max_{\alpha_i}\sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_i y_i y_i \phi(\boldsymbol{x_i})^T\phi(\boldsymbol{x_j})$$

Here, we use the kernel ($\gamma = 1$):

$$k(\boldsymbol{x}, \boldsymbol{x_i}) = (\boldsymbol{x^T}\boldsymbol{x_i})^3$$

Therefore, we can get the final dual problem

$$\max_{\alpha_i}\sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j (\boldsymbol{x^T}\boldsymbol{x_i})^3 \ s.t. \sum_i^m \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, \forall i$$

We set the panel parameter $C = 1$ and get the error:

$training\ error = 0.008333333333333304$

$testing\ error = 0.0$

2.4.5 Question 5: SVM with Radial Basis Function kernel and slack variables

The primal problem:

$$\min_{w,b,\xi_i}\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i \ s.t. 1 - y_i(\boldsymbol{w^T}\boldsymbol{\phi(x_i)} + b) \leq 0, \forall i$$

Lagrange function:

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i^m \xi_i + \sum_i^m[\alpha_i(1 - \xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i}) + b)) - u_i\xi_i]$$

KKT condition:

Stationarity:

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0,\ \ \boldsymbol{w} = \sum_i^m \alpha_i y_i \phi(\boldsymbol{x_i})\ \frac{\partial L}{\partial b} = 0,\ \ \sum_i^m a_i y_i = 0\ \frac{\partial L}{\partial \xi_i} = 0,\ \ \alpha_i = C - \mu_i, \forall i$$

Feasibility:

$$\alpha_i \geq 0, 1 - \xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i}) + b) \leq 0, \xi_i \geq 0, \mu_i \geq 0, \forall i$$

Complementray slackness:

$$\alpha_i(1 - \xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i}) + b)) = 0, \mu_i\xi_i = 0, \forall i$$

Then, we can get the dual problem:

$$\max_{\alpha_i}\sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_i y_i y_i \phi(\boldsymbol{x_i})^T\phi(\boldsymbol{x_j})$$

Here, we use the kernel ($\gamma = \frac{1}{2}$):

$$k(\boldsymbol{x}, \boldsymbol{x_i}) = exp(-\frac{||\boldsymbol{x} - \boldsymbol{x_i}||^2}{2})$$

Therefore, we can get the final dual problem

$$\max_{\alpha_i}\sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j exp(-\frac{||\boldsymbol{x}-\boldsymbol{x_i}||^2}{2})\ s.t.\sum_i^m \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, \forall i$$

We set the panel parameter $C = 1$ and get the error:

$training\ error = 0.033333333333333326$

$testing\ error = 0.033333333333333326$

2.4.6 Question 5: SVM with Sigmoidal kernel and slack variables

The primal problem:

$$\min_{w,b,\xi_i}\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i\ s.t. 1 - y_i(\boldsymbol{w^T}\boldsymbol{\phi(x_i)} + b) \leq 0, \forall i$$

Lagrange function:

$$L(\boldsymbol{w}, b, \alpha) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i^m \xi_i + \sum_i^m[\alpha_i(1 - \xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i}) + b)) - u_i\xi_i]$$

KKT condition:

Stationarity:

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0,\ \ \boldsymbol{w} = \sum_i^m \alpha_i y_i \phi(\boldsymbol{x_i})\ \frac{\partial L}{\partial b} = 0,\ \ \sum_i^m a_i y_i = 0\ \frac{\partial L}{\partial \xi_i} = 0,\ \ \alpha_i = C - \mu_i, \forall i$$

Feasibility:

$$\alpha_i \geq 0, 1 - \xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i}) + b) \leq 0, \xi_i \geq 0, \mu_i \geq 0, \forall i$$

Complementray slackness:

$$\alpha_i(1 - \xi_i - y_i(\boldsymbol{w^T}\phi(\boldsymbol{x_i}) + b)) = 0, \mu_i\xi_i = 0, \forall i$$

Then, we can get the dual problem:

$$\max_{\alpha_i}\sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_i y_i y_i \phi(\boldsymbol{x_i})^T\phi(\boldsymbol{x_j})$$

Here, we use the kernel ($\gamma = \frac{1}{4}$ since $\boldsymbol{x}$ is 4 dimension, in SVC, we set it as 'auto'):

$$k(x, x_i) = \tanh(\frac{1}{4}\boldsymbol{x^T}\boldsymbol{x})$$

Therefore, we can get the final dual problem

$$\max_{\alpha_i}\sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \tanh(\frac{1}{4}\boldsymbol{x^T}\boldsymbol{x})\ s.t.\sum_i^m \alpha_i y_i = 0, 0 \leq \alpha_i \leq C, \forall i$$

We set the panel parameter $C = 1$ and get the error:

$training\ error = 0.825$

$testing\ error = 0.7666666666666666$