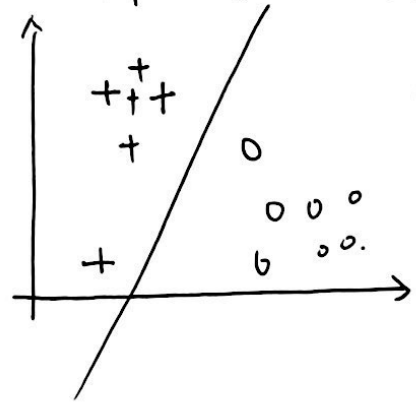
Zhaliang Yu
Assignment 2
12048077.

Exercise 8.7.

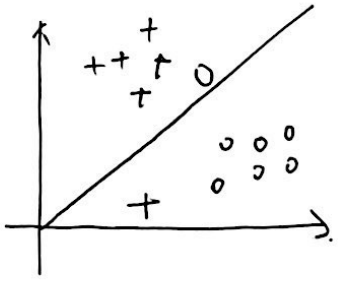
a. $P(y=1|x,w) = \delta(w_0+w_1x_1+w_2x_2)$ $J(w) = -l(w, D_{train})$.



① boundary is not unique

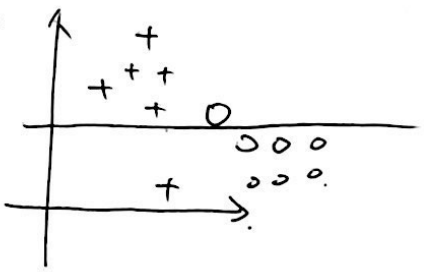
② 0.

b. ~~to~~



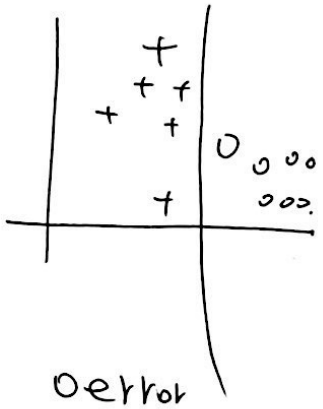
① 2 errors

c.



② 2 errors

D.



0 error

Exercise §14.1

max margin classifier:

$\min \|w\|^2$ S.t.

$y_1(w^T\phi(x_1)+w_0) \geq 1$

$y_2(w^T\phi(x_2)+w_0) \geq 1$

a. $\phi(x_1) = \begin{bmatrix}1\\0\\0\end{bmatrix}$ $\phi(x_2) = \begin{bmatrix}1\\2\\2\end{bmatrix}$

as w is perpendicular to the decision boundary:

$w // \begin{bmatrix}0\\1\\1\end{bmatrix}$

b. $d_m = \frac{\sqrt{0^2+1^2+1^2}}{2} = \sqrt{2}$.

c. $d_m = \frac{1}{\|w\|} \Rightarrow \|w\| = \frac{1}{\sqrt{2}}$ $w = (0, \frac{1}{2}, \frac{1}{2})$.

d.

$$\begin{cases} -1\left([0,\frac{1}{2},\frac{1}{2}]\begin{bmatrix}1\\0\\0\end{bmatrix} + w_0\right) \geq 1 \\ 1\left([0,\frac{1}{2},\frac{1}{2}]\begin{bmatrix}1\\2\\2\end{bmatrix} + w_0\right) \geq 1 \end{cases} \Rightarrow w_0 = -1$$

e. $f(x) = -1 + [0, \frac{1}{2}, \frac{1}{2}]\begin{bmatrix}1\\\sqrt{5}x\\x^2\end{bmatrix}$

Exercise 14.6 Linear Separability

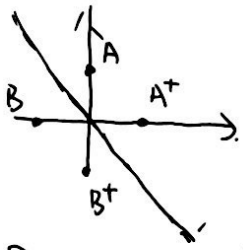
Yes, Consider 2-d Linear Separable data

So the separating hyperplane guarantee to be exist.

So we can use SVM to correctly find the max-imum margin by SVM

4. class -1: $\begin{bmatrix}(1 & 0)\\(0 & 1)\end{bmatrix}$ class 1: $\begin{bmatrix}(-1 & 0)\\(0 & -1)\end{bmatrix}$

(1) The dataset is linear separable



$A, A^+ \in$ class -1

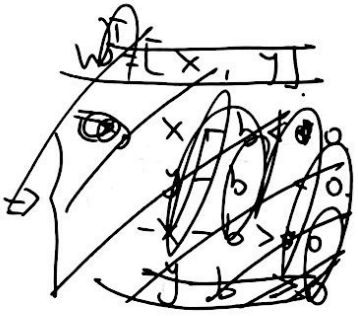
$B, B^+ \in$ class 1.

$W^T = [-1, -1]$

So we can simply use linear classifier:

$a_{(x)} = Sgn(w^T x - b)$

$\begin{cases} w^T x - b > 0 & \forall x \in \text{class } +1 \\ w^T x - b < 0 & \forall x \in \text{class } -1 \end{cases}$



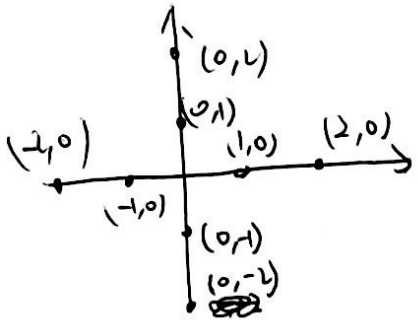
(2). Supporting vectors.

$C = \{(1\ 0)\ (0\ 1)\ (-1\ 0)\ (0,-1)\}$ are all supporting vectors.

$\forall x_i \in C \quad M_i(w, b) = 1.$

(3). $[1; 2]$ belongs to class -1.

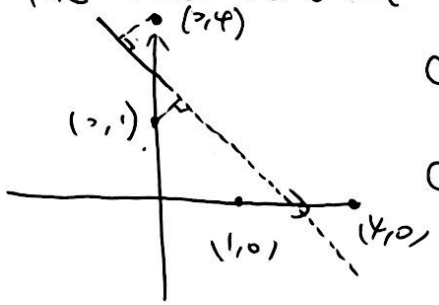
5. The dataset looks like this



(1) Linear classifier cannot classify the data

(2) by adding the feature $\phi(x) = [x_1^2 ; x_2^2]$

The new dataset becomes



$$\text{class } -1 : \begin{bmatrix} (1,0) \\ (0,1) \end{bmatrix}$$

$$\text{class } +1 : \begin{bmatrix} (4,0) \\ (0,4) \end{bmatrix}$$

$$\begin{cases} w^T x - b < 0 & \text{for } x \in \text{class } -1 \\ w^T x - b > 0 & \text{for } x \in \text{class } 1 \end{cases} \qquad w^T = [x, y]$$

$$\Rightarrow w^T = \left[\frac{2}{3}, \frac{2}{3}\right] \quad b = \frac{5}{3}$$

$$\left[\frac{2}{3}, \frac{2}{3}\right]\begin{bmatrix} 1 \\ 2 \end{bmatrix} - \frac{5}{3} = \frac{1}{3} > 0$$

So $[1;2]$ belongs to class 1

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

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

$$\text{s.t.} \sum_{n=1}^{N} a_n t_n = 0$$

$$a_n \geq 0 \quad \forall n = 1, 2 \cdots N$$

$$k(x_n, x_m) = x_n^T x_m$$

$$b = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \sum_{m=1}^{M} a_m t_m k(x_n, x_m)\right)$$

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

$$= \sum_{n=1}^{N} a_n t_n \cdot x_n^T \cdot x + \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \sum_{m=1}^{M} a_m t_m x_n^T t_m\right)$$

$$\Rightarrow \frac{1}{\gamma^2} = \frac{1}{\|w\|^2} = \sum_{n=1}^{N} a_n$$

SVM Programming Report

In this assignment we implemant various settings of support vector machine, including the regular linear SVM, SVM with slack variable, SVM with kernels.

SVM Formulation

The objective function of support vector machine is:

$$min_{w,b}\frac{1}{2}||w||^2$$
$$s.t.\, y_i(w^T x_i + b) \geq 1 \forall i$$

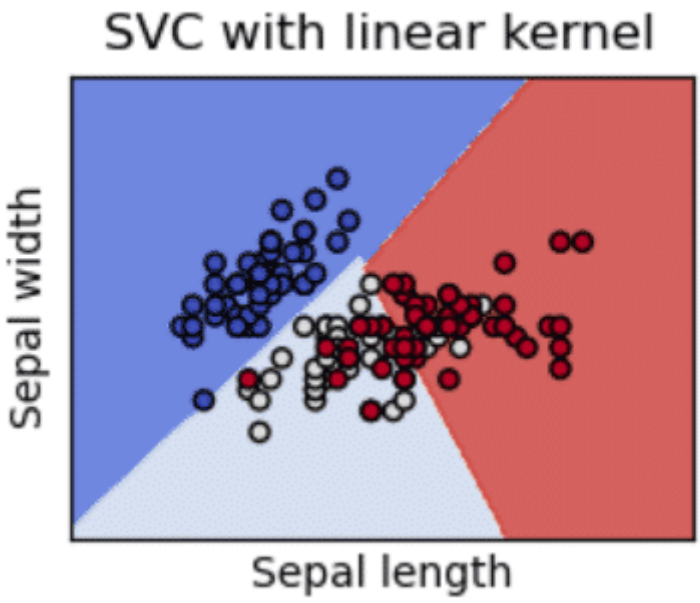
Regular SVM

So in our project we use the "w" to denote the weight, "b" to denote the bias, and x*{i} to denote training data, y*{i} to denote the training label.

So in regular SVM we set the slack variable to be a relatively very small number, and set the kernel to be linear.

```
clf = svm.SVC(C=1e5, kernel = 'linear')
```

then we can get the following output to notice that class 0 "Iris-setosa" is linear separable

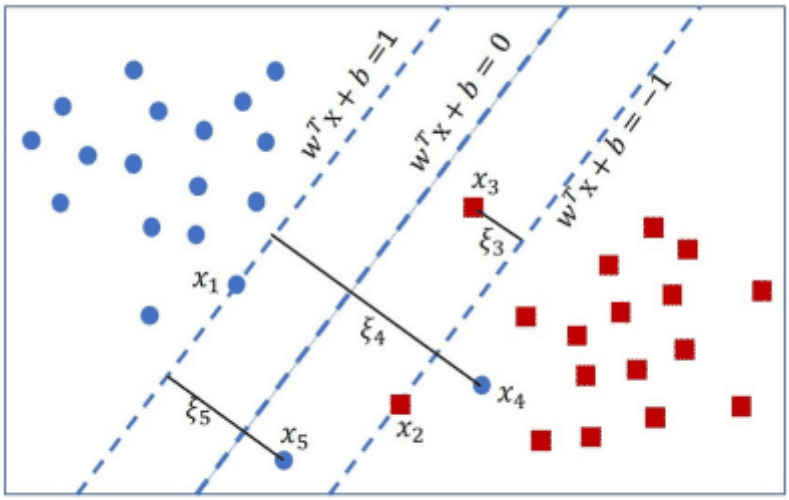


By training the regular SVM by the setting mentioned above we can get the following result:

```
training_error:0.016667
testing_error:0.000000
w_of_setosa:-0.046258538085542256, 0.5211827995531007, -1.0030446153124941,
-0.4641297849669344,
b_of_setosa: 1.452844
support_vector_indices_of_setosa: 13, 31, 34,
w_of_versicolor: -0.15228426789115657, 0.13536379368102813, -0.490693752093727,
-0.23688663894179923,
b_of_versicolor: 2.289340
support_vector_indices_of_versicolor: 50, 52, 57, 63, 78,
w_of_virginica: 4.264086882816628, 6.192174636991695, -8.641446592053398,
-12.560510518902447,
b_of_virginica: 19.189122
support_vector_indices_of_virginica: 97, 99, 103, 108,
```

SVM with Slack Formulation

A SVM with slack variable look like this, we allow little portion of data not in the decision region.



So the optimization problem formulation look like this:

- In this case, the SVM is formulated as follows

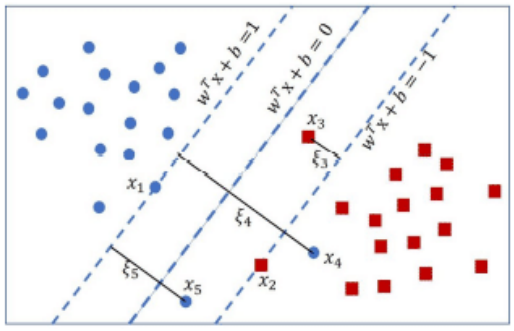
$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i^m \xi_i \tag{20}$$

$$s.t.\ 1-\xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 0, \xi_i \geq 0,\ \forall i$$

- Its Lagrange function is

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i^m \xi_i + \sum_i^m \left[\alpha_i\big(1-\xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\big) - \mu_i\xi_i\right],$$

and $\alpha_i, \mu_i \geq 0, \forall i$.



we can then derive its dual lagurangian and dual problem

- Lagrange function:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i^m \xi_i + \sum_i^m \left[\alpha_i\big(1-\xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\big) - \mu_i\xi_i\right],$$

- Replacing all KKT conditions into Lagrange function to eliminate primal variables, we have

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j \tag{26}$$

- Then, we obtain the following dual problem:

$$\max_{\boldsymbol{\alpha}} \sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j, \tag{27}$$

$$s.t.\ \sum_i^m \alpha_i y_i = 0,\ 0 \leq \alpha_i \leq C,\ \forall i \tag{28}$$

- Note that the only change in dual problem is the constraint $0 \leq \alpha_i \leq C$, which is $\alpha_i \geq 0$ in the dual problem of standard SVM.

the solution α_{i} has the following 3 cases:

1. $\alpha_{i} = 0$ the corresponding data are correctly classified and doesn't contribute to the classifier, locating outside of the margin
2. 0 < αi < C: in this case, μi > 0 due to αi = C − μi
   ; Since μiξi = 0,
   then we have ξi = 0. The corresponding data are correctly classified and contributes to the classifier, locating on the margin
3. αi = C: in this case, μi = 0; then we have ξi > 0. The corresponding data contributes to the classifier, locating inside the margin

If ξi ≤ 1, then the data is still correctly classified, not crossing decision boundary
If ξi > 1, then the data is incorrectly classified, crossing decision boundary

And we can derive the bias in the following way:

How to determine the bias parameter b?

- We define $\mathcal{M} = \{i|0 < \alpha_i < C\}$
- Since $0 < \alpha_i < C$, we have $\xi_i = 0$
- Then, for any support vector $\mathbf{x}_j, j \in \mathcal{M}$, we have

$$y_j(\mathbf{w}^\top \mathbf{x}_j + b) = 1, \ \forall j \in \mathcal{M} \tag{29}$$

$$\Rightarrow y_j(\sum_i^m \alpha_i y_i \mathbf{x}_i^\top \mathbf{x}_j + b) = 1, \ \forall j \in \mathcal{M} \tag{30}$$

- Utilizing $y_j \cdot y_j = 1$, we have

$$\sum_i^m \alpha_i y_i \mathbf{x}_i^\top \mathbf{x}_j + b = y_j, \ \forall j \in \mathcal{M} \tag{31}$$

$$\Rightarrow b = \frac{1}{|\mathcal{M}|} \sum_{j \in \mathcal{M}} (y_j - \sum_i^m \alpha_i y_i \mathbf{x}_i^\top \mathbf{x}_j) \tag{32}$$

- Note that using the average of all support vectors, rather than one single support vector, could make the solution of b more numerically stable.

so in this project we tested varios slack varible from 0 to 1 step size 0.1

we get the following result:
we select slack variable 0.1 as example

```
training_error:0.016667
testing_error:0.033333
w_of_setosa:-0.08461491940959265, 0.4452659076269341, -0.8404228999803351,
-0.39549567986979894,
b_of_setosa: 1.567953
support_vector_indices_of_setosa: 13, 14, 31, 34,
w_of_versicolor: -0.15228426789115657, 0.13536379368102813, -0.490693752093727,
-0.23688663894179923,
b_of_versicolor: 2.289340
support_vector_indices_of_versicolor: 43, 46, 48, 50, 52, 53, 56, 57, 58, 63, 64,
65, 66, 67, 71, 73, 78,
w_of_virginica: 0.12864428210513168, 0.42231488821646934, -1.5263467916532356,
-1.346810447700804,
b_of_virginica: 7.685070
support_vector_indices_of_virginica: 80, 81, 83, 89, 91, 93, 96, 97, 103, 104,
108, 111, 112, 116, 117, 119,
```

SVM with non-linear kernel

So in this part we set the slack variable to be 1, then we try to use kernels including: 2nd-order polynomial kernel, 3rd-order polynomial, rbf, sigmoid.

We use non-linear kernel because SVM can only handle linearly separable data, so we can add non-linearity to handle higher order polynomial data.

So we can formulate the SVM with kernel in the following manner:

- We firstly define the following kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j) \tag{38}$$

- Utilizing this kernel to replacing $\mathbf{x}_i^\top \mathbf{x}_j$, we have the following dual problem

$$\max_{\boldsymbol{\alpha}} \sum_i^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j), \tag{39}$$

$$s.t. \sum_i^m \alpha_i y_i = 0, \ \alpha_i \geq 0, \ \forall i \tag{40}$$

- The solution of b becomes

$$b = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \left(y_j - \sum_i^m \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}_j) \right) \tag{41}$$

- The prediction of new data $\mathbf{x}$ becomes

$$\mathbf{w}^\top \mathbf{x} + b = \sum_i^m \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b \tag{42}$$

- Since $\boldsymbol{\alpha}$ is sparse, the above classifier is also called sparse kernel classifier.

in the end we get the following result, take sigmoid as example:

```
training_error:0.958333
testing_error:0.966667
b_of_setosa: 0.328565
support_vector_indices_of_setosa: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,
14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
34, 35, 36, 37, 38, 39,
b_of_versicolor: 0.214481
support_vector_indices_of_versicolor: 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50,
51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70,
71, 72, 73, 74, 75, 76, 77, 78, 79,
b_of_virginica: 0.014337
support_vector_indices_of_virginica: 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90,
91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108,
109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119,
```