大家好，这篇是有关Learning from data第八章习题的详解，这一章主要介绍了支持向量机。

我的github地址：
https://github.com/Doraemonzzz

个人主页：
http://doraemonzzz.com/

参考资料:
https://blog.csdn.net/a1015553840/article/details/51085129
http://www.vynguyen.net/category/study/machine-learning/page/6/
http://book.caltech.edu/bookforum/index.php
http://beader.me/mlnotebook/

# Chapter8 Support Vector Machines

## Part 1: Exercise

### Exercise 8.1 (Page 3)

Assume $\mathcal{D}$ contains two data points $(x_+, +1)$ and $(x_-, -1)$. Show that:

(a) No hyperplane can tolerate noise radius greater than $\frac{1}{2}||x_+ - x_-||$.

(b) There is a hyperplane that tolerates a noise radius $\frac{1}{2}||x_+ - x_-||$.

假设平面为$w^Tx + b = 0$，那么任意一点$x'$到该平面的距离为$\frac{|w^Tx' + b|}{||w||}$，接下来考虑这两个问题。

(a)记$x_+$到该平面的距离为$d_+$，$x_-$到该平面的距离为$d_-$

$$d_+ + d_- = \frac{|w^Tx_+ + b|}{||w||} + \frac{|w^Tx_- + b|}{||w||}$$

因为$x_+$标记为$1$，所以$w^Tx_+ + b > 0$，同理$w^Tx_- + b < 0$，所以距离之和为

$$\begin{aligned} d_+ + d_- &= \frac{|w^Tx_+ + b|}{||w||} + \frac{|w^Tx_- + b|}{||w||} \\ &= \frac{w^Tx_+ + b}{||w||} + \frac{-w^Tx_- - b}{||w||} \\ &= \frac{w^T(x_+ - x_-)}{||w||} \end{aligned}$$

由柯西不等式可知

$$d_+ + d_- = \frac{w^T(x_+ - x_-)}{||w||} \leq \frac{||w^T||||x_+ - x_-||}{||w||} = ||x_+ - x_-||$$

因为$\text{radius}=\min\{d_+, d_-\}$，所以

$$2\text{radius} \leq d_+ + d_- \leq ||x_+ - x_-||$$
$$\text{radius} \leq \frac{1}{2}||x_+ - x_-||$$

(b)从几何意义我们可知，只要取$x_+, x_-$的“中垂线”即可，类似二维情形，该平面为

$$(x_+ - x_-)^T(x - \frac{1}{2}(x_+ + x_-)) = 0$$

$x_+$到该平面的距离为

$$\begin{aligned} d_+ &= \frac{|(x_+ - x_-)^T(x_+ - \frac{1}{2}(x_+ + x_-))|}{||x_+ - x_-||} \\ &= \frac{1}{2}\frac{|(x_+ - x_-)^T(x_+ - x_-)|}{||x_+ - x_-||} \\ &= \frac{1}{2}||x_+ - x_-|| \end{aligned}$$

$x_-$到该平面的距离为

$$\begin{aligned} d_- &= \frac{|(x_+ - x_-)^T(x_- - \frac{1}{2}(x_+ + x_-))|}{||x_+ - x_-||} \\ &= \frac{1}{2}\frac{|(x_+ - x_-)^T(x_+ - x_-)|}{||x_+ - x_-||} \\ &= \frac{1}{2}||x_+ - x_-|| \end{aligned}$$

所以存在一个平面满足条件。

## Exercise 8.2 (Page 5)

Consider the data below and a 'hyperplane' $(b, w)$ that separates the data.

$$X = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \end{bmatrix}, y = \begin{bmatrix} -1 \\ -1 \\ +1 \end{bmatrix}, w = \begin{bmatrix} 1.2 \\ -3.2 \end{bmatrix}, b = -0.5$$

(a) Compute $\rho = \min\limits_{n=1,\ldots,N} y_n(w^T x_n + b)$.

(b) Compute the weights $\frac{1}{\rho}(b, w)$ and show that they satisfy (8.2).

(c) Plot both hyperplanes to show that they are the same separator

(a)计算即可

$$
\begin{aligned}
y_1(w^T x_1 + b) &= (-1) \times (0 - 0.5) = 0.5 \\
y_2(w^T x_2 + b) &= (-1) \times (2.4 - 6.4 - 0.5) = 4.5 \\
y_3(w^T x_3 + b) &= (1) \times (2.4 - 0.5) = 1.9 \\
\rho &= 0.5
\end{aligned}
$$

(b)(8.2)为

$$
\min_{n=1,\dots,N} y_n(w^T x_n + b) = 1
$$

计算$\frac{1}{\rho}(b, w)$可得

$$
\frac{w}{\rho} = \frac{\begin{bmatrix} 1.2 \\ -3.2 \end{bmatrix}}{0.5} = \begin{bmatrix} 2.4 \\ -6.4 \end{bmatrix}, \frac{b}{\rho} = -1
$$

重新计算可得

$$
\begin{aligned}
\frac{y_1(w^T x_1 + b)}{\rho} &= 1 \\
\frac{y_2(w^T x_2 + b)}{\rho} &= 9 \\
\frac{y_3(w^T x_3 + b)}{\rho} &= 3.8
\end{aligned}
$$

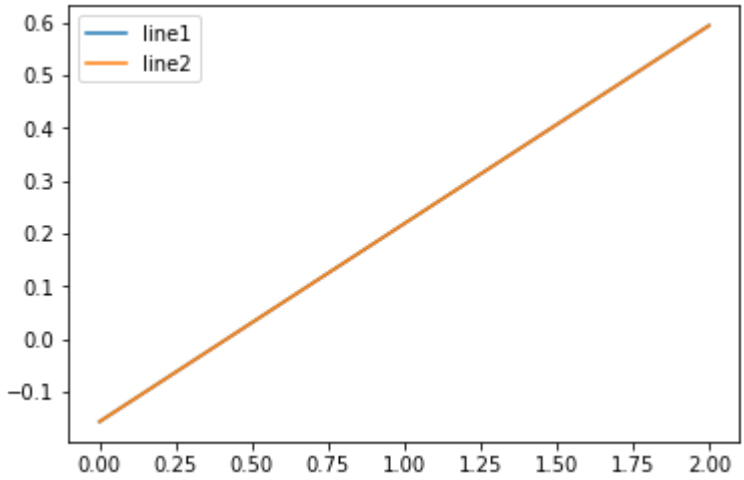
所以条件满足。

(c)

作图可得

```
# -*- coding: utf-8 -*-
"""
Created on Fri Mar 22 12:04:36 2019

@author: qinzhen
"""

import matplotlib.pyplot as plt
import numpy as np
w = np.array([1.2, -3.2])
b = -0.5
w1 = w / 0.5
b1 = b / 0.5

x = np.array([0, 2])
y = - (w[0] * x + b) / w[1]
y1 = -(w1[0] * x + b1) / w1[1]

plt.plot(x, y, label='line1')
plt.plot(x, y1, label='line2')
```

```
plt.legend()
plt.show()
```




## Exercise 8.3 (Page 8)

For separable data that contain both positive and negative examples, and a separating hyperplane $h$, define the positive-side margin $\rho_+(h)$ to be the distance between $h$ and the nearest data point of class $+1$. Similarly, define the negative-side margin $\rho_-(h)$ to be the distance between $h$ and the nearest data point of class $-1$. Argue that if $h$ is the optimal hyperplane, then $\rho_+(h) = \rho_-(h)$. That is, the thickness of the cushion on either side of the optimal $h$ is equal

设$\rho_+(h)$对应的点为$X_+ = \{x_{1+}, \dots, x_{n+}\}$，$\rho_-(h)$对应的点为$X_- = \{x_{1-}, \dots, x_{m-}\}$，由8.1可知

$$\rho_+(h) \le \frac{1}{2}||x_{i+} - x_{j-}||$$
$$\rho_-(h) \le \frac{1}{2}||x_{i+} - x_{j-}||$$
$$x_{i+} \in X_+, x_{j-} \in X_-$$

记$d = \min_{x_{i+} \in X_+, x_{j-} \in X_-} \{\frac{1}{2}||x_{i+} - x_{j-}||\}$，且当$x_{i+} = x_+^*, x_{i-} = x_-^*$时，$\frac{1}{2}||x_{i+} - x_{j-}||$取最小值，结合Exercise 8.1可知

$$\rho_+(h) \le d$$
$$\rho_-(h) \le d$$
$$\text{当超平面为} (x_+^* - x_-^*)^T(x - \frac{1}{2}(x_+^* + x_-^*)) = 0 \text{时，等号同时成立}$$

我们知道最优超平面为使得$\rho_+(h), \rho_-(h)$最小值最大的超平面，由上述结论知当超平面为

$$(x_+^* - x_-^*)^T(x - \frac{1}{2}(x_+^* + x_-^*)) = 0$$

时，$\rho_+(h), \rho_-(h)$同时取最大值且

$$\rho_+(h) = \rho_-(h) = d$$

所以结论成立。

## Exercise 8.4 (Page 10)

Let $Y$ be an $N \times N$ diagonal matrix with diagonal entries $Y_{nn} = y_n$ (a matrix version of the target vector y). Let $X$ be the data matrix augmented with a column of 1s. Show that $A = YX$.

不难验证我们有

$$Y = \text{diag}\{y_1, \dots, y_N\}$$

$$X = \begin{bmatrix} 1 & x_1^T \\ \dots & \dots \\ 1 & x_N^T \end{bmatrix}$$

$$YX = \begin{bmatrix} y_1 & y_1 x_1^T \\ \dots & \dots \\ y_N & y_N x_N^T \end{bmatrix}$$

回顾课本第10页$A$的定义可知

$$A = YX$$

## Exercise 8.5 (Page 11)

Show that the matrix $Q$ described in the linear hard-margin SVM algorithm above is positive semi-definite (that is $u^T Q u \geq 0$ for any $u$).

The result means that the QP-problem is convex. Convexity is useful be cause this makes it 'easy' to find an optimal solution. In fact, standard QP-solvers can solve our convex QP-problem in $O((N+d)^3)$.

回顾课本第10页定义

$$Q = \begin{bmatrix} 0 & 0_d^T \\ 0_d^T & I_d \end{bmatrix}$$

令$x = (x_0, x_1, \dots, x_d)$，那么

$$x^T Q x = \sum_{i=1}^{d} x_i^2 \geq 0$$

所以$Q$半正定，结论成立。

## Exercise 8.6 (Page 12)

Construct a toy data set with $N = 20$ using the method in Example 8.4.

(a) Run the SVM algorithm to obtain the maximum margin separator $(b, w)$ SVM and compute its $E_{\text{out}}$ and margin.

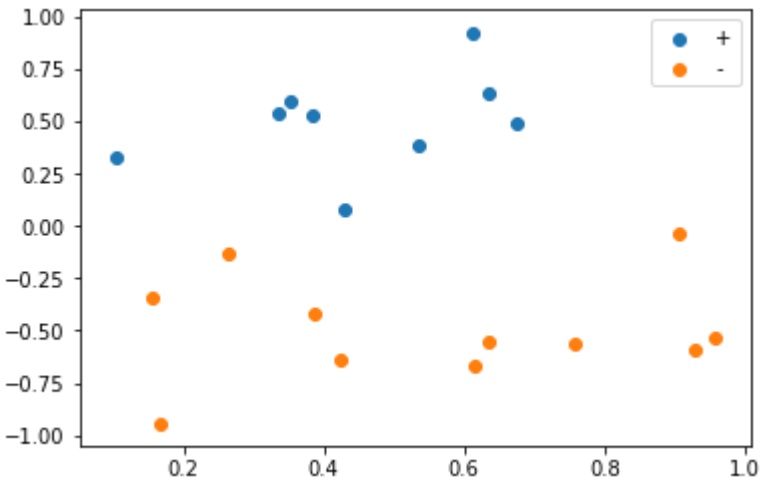
(b) Construct an ordering of the data points that results in a hyperplane with bad $E_{\text{out}}$ when PLA is run on it. [Hint: Identify a positive and negative data point for which the perpendicular bisector separating these two points is a bad separator. Where should these two points be in the ordering? How many iterations will PLA take?]

(c) Create a plot of your two separators arising from SVM and PLA.

回顾Example 8.4，$x_1 \in [0,1], x_2 \in [-1,1], f(x) = \text{sign}(x_2)$，下面做实验即可。

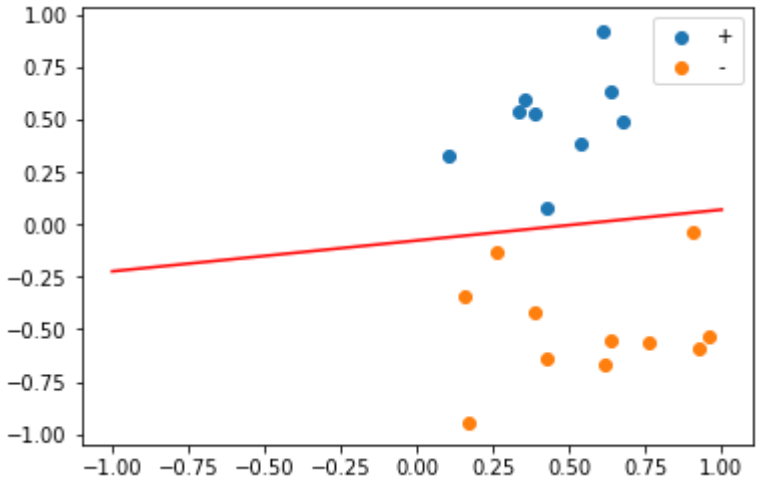
(a)首先作图

```
# -*- coding: utf-8 -*-
"""
Created on Fri Mar 22 12:07:03 2019

@author: qinzhen
"""

import numpy as np
import matplotlib.pyplot as plt
from sklearn import svm
from sklearn.linear_model import Perceptron

####(a)
####作图
N = 20
x1 = np.random.uniform(0, 1, N)
x2 = np.random.uniform(-1, 1, N)
x = np.c_[x1, x2]
y = np.sign(x2)

plt.scatter(x1[y>0], x2[y>0], label="+")
plt.scatter(x1[y<0], x2[y<0], label="-")
plt.legend()
plt.show()
```



接着使用svm算法，这里好像没有看到hard margin，所以取惩罚系数C为很大的一个数来代替。

```
####训练数据
clf = svm.SVC(kernel ='linear', C=1e10)
clf.fit(X, y)

#获得超平面
w = clf.coef_[0]
b = clf.intercept_[0]

#作图
m = np.array([-1, 1])
n = - (b + w[0] * m) / w[1]
plt.scatter(x1[y>0], x2[y>0], label="+")
plt.scatter(x1[y<0], x2[y<0], label="-")
plt.plot(m, n, 'r')
plt.legend()
plt.show()
```




计算margin，以及$E_{\text{out}}$，注意这里我$E_{\text{out}}$都是采用模拟的方法求得，没有采用积分的方法。

```
margin = 1 / np.sqrt(np.sum(clf.coef_ ** 2))
margin
```

```
margin = 0.09612367161991067
```

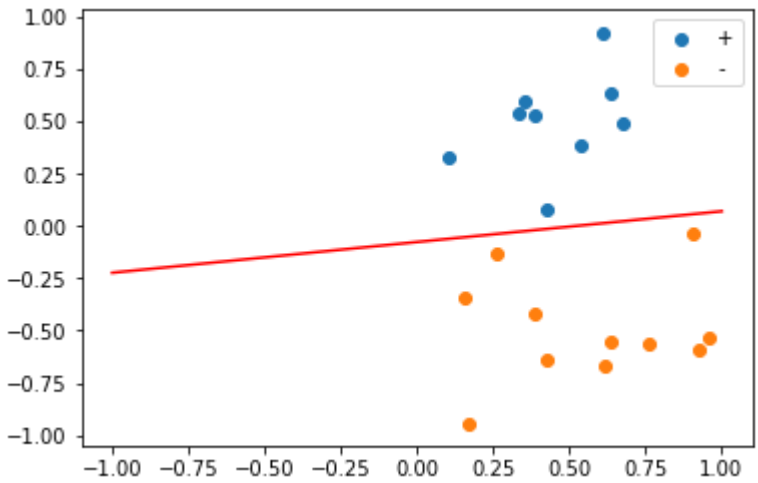
```
#计算Eout
def Eout(w, b, N=1000):
    x1 = np.random.uniform(0, 1, N)
    x2 = np.random.uniform(-1, 1, N)
    X = np.c_[x1, x2]
    y = np.sign(x2)
    y1 = np.sign(X.dot(w) + b)
    return np.mean(y != y1)

e = Eout(w, b)
print(e)
```

```
0.0183
```

(b)题目的意思是数据按照什么顺序排列，使得PLA的$E_{\text{out}}$很大，根据提示，推测样本的顺序为一个正样本后面跟一个负样本，实验这里略过。（这题强调的是PLA的结果和样本的顺序有关，而SVM和样本顺序无关）。

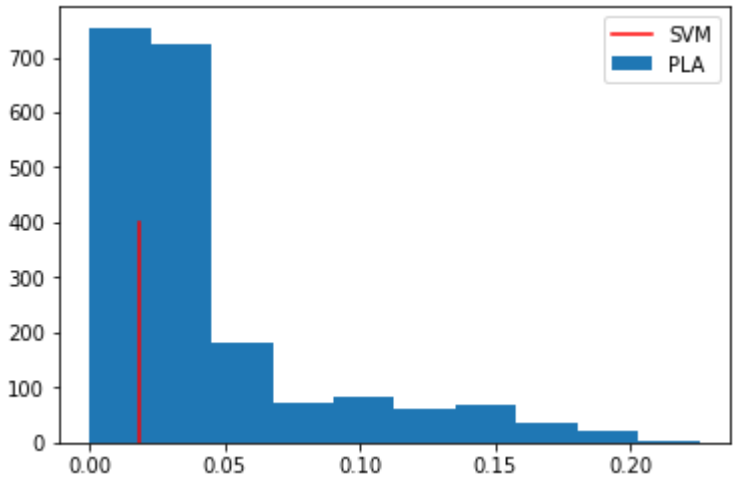
(c)打乱数据，多次采用PLA，然后计算$E_{\text{out}}$，作图

```
####(c)
clf = Perceptron()
clf.fit(X, y)

w1 = clf.coef_[0]
b1 = clf.intercept_[0]

#作图
m1 = np.array([-1, 1])
n1 = - (b + w[0] * m1) / w[1]
plt.scatter(x1[y>0], x2[y>0], label="+")
plt.scatter(x1[y<0], x2[y<0], label="-")
plt.plot(m1, n1, 'r')
plt.legend()
plt.show()
```




做多次实验，作出$E_{\text{out}}$的直方图

```
#多次实验，做直方图
result = np.array([])
for i in range(2000):
    np.random.shuffle(X)
    y = np.sign(X[:,1])
    clf.fit(X, y)
    w1 = clf.coef_[0]
    b1 = clf.intercept_[0]
    result = np.append(result,Eout(w1,b1))

plt.hist(result, label="PLA")
plt.plot([e] * 400, range(400), 'r', label="SVM")
plt.legend()
plt.show()
```




## Exercise 8.7 (Page 15)

Assume that the data is restricted to lie in a unit sphere.

(a) Show that $d_{\text{vc}}(\rho)$ is non-increasing in $\rho$.

(b) In 2 dimensions, show that $d_{\text{vc}}(\rho) < 3$ for $\rho > \frac{\sqrt{3}}{2}$. [Hint: Show that for any 3 points in the unit disc, there must be two that are within distance $\sqrt{3}$ of each other. Use this fact to construct a dichotomy that cannot be implemented by any ρ-thick separator.]

(a)如果$\rho < \rho^{'}$，由几何意义可知

$$\mathcal{H}(\rho^{'}) \subset \mathcal{H}(\rho)$$

从而

$$d_{\text{vc}}(\rho^{'}) < d_{\text{vc}}(\rho)$$

所以结论成立。

(b)如果能证明单位圆内任意三个点至少有两个点的距离小于等于$\sqrt{3}$，那么当$\rho > \frac{\sqrt{3}}{2}$时，$d_{\text{vc}}(\rho) < 3$，接下来证明单位圆内任意三个点至少有两个点的距离大于等于$\sqrt{3}$。

假设单位圆圆心为$O$，圆内三个点为$A, B, C$，那么三角形$ABC$内至少有个一个角小于等于$\frac{\pi}{3}$，不妨设为$\angle A$，所以$\angle BOC$小于等于$\frac{2\pi}{3}$，$\cos(\angle BOC) \geq -\frac{1}{2}$。在三角形BOC内利用余弦定理计算$BC$，注意$OB \leq 1, OC \leq 1$

$$BC = \sqrt{OB^2 + OC^2 - 2OB \times OC \cos(\angle BOC)} \leq \sqrt{OB^2 + OC^2 + OB \times OC} \leq \sqrt{1^2 + 1^2 + 1 \times 1} = \sqrt{3}$$

从而结论成立。

## Exercise 8.8 (Page 18)

(a) Evaluate the bound in (8.8) for the data in Figure 8.5.

(b) If one of the four support vectors in a gray box are removed, does the classifier change?

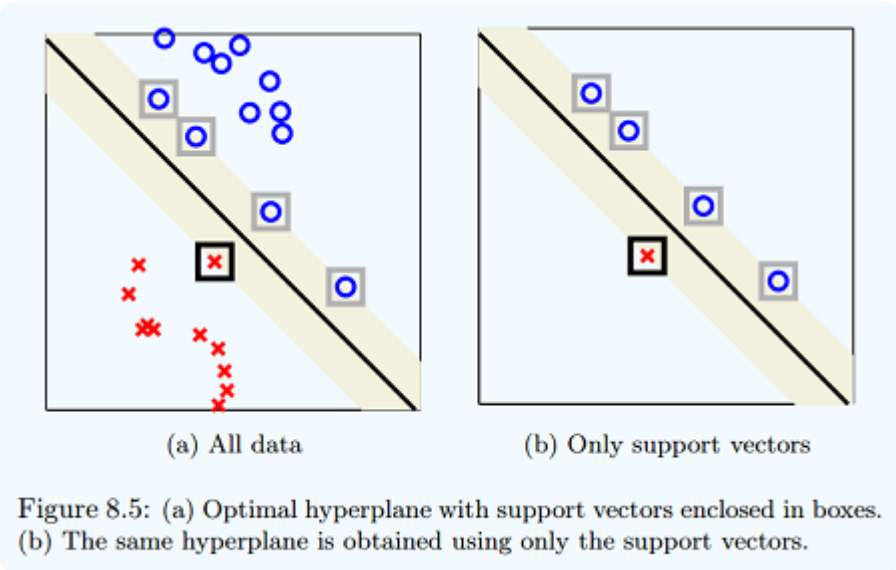
(c) Use your answer in (b) to improve your bound in (a).

The support vectors in gray boxes are non-essential and the support vector in the black box is essential. One can improve the bound in (8.8) to use only essential support vectors. The number of support vectors is unbounded, but the number of essential support vectors is at most $d + 1$ (usually much less).

首先回顾公式(8.8)

$$E_{\text{cv}}(\text{SVM}) = \frac{1}{N} \sum_{n=1}^{N} e_n \leq \frac{\#\text{ support vectors}}{N}$$

接着看图片8.5



(a) All data　　(b) Only support vectors

Figure 8.5: (a) Optimal hyperplane with support vectors enclosed in boxes. (b) The same hyperplane is obtained using only the support vectors.

(a)图中$\#\text{ support vectors} = 5, N = 23$，所以

$$E_{\text{cv}}(\text{SVM}) \leq \frac{5}{23}$$

(b)如果在灰色box中的某一个点被删除，是不会影响这个分类器的，因为margin没有变，且直线的方向向量没有变。

(c)如果按照b的操作，那么$\#\text{ support vectors} = 4, N = 22$，所以

$$E_{\text{cv}}(\text{SVM}) \leq \frac{4}{22}$$

## Exercise 8.9 (Page 23)

Let $u_0$ be optimal for (8.10), and let $u_1$ be optimal for (8.11).

(a) Show that $\max_{\alpha \geq 0} \alpha(c - a^T u_0) = 0$. [Hint: $c - a^T u_0 \leq 0$.]

(b) Show that $u_1$ is feasible for (8.10). To show this, suppose to the contrary that $c - a^T u_1 > 0$. Show that the objective in (8.11) is infinite, whereas $u_0$ attains a finite objective of $\frac{1}{2} u_0^T Q u_0 + p^T u_0$, which contradicts the optimality of $u_1$.

(c) Show that $\frac{1}{2} u_1^T Q u_1 + p^T u_1 = \frac{1}{2} u_0^T Q u_0 + p^T u_0$, and hence that $u_1$ is optimal for (8.10) and $u_0$ is optimal for (8.11).

(d) Let $u^*$ be any optimal solution for (8.11) with $\max_{\alpha \geq 0} \alpha(c - a^T u^*)$ attained at $\alpha^*$. Show that

$$\alpha^* (c - a^T u^*) = 0$$

Either the constraint is exactly satisfied with $c - a^T u^* = 0$, or $\alpha^* = 0$.

首先回顾8.10与8.11

8.10

$$\begin{aligned} \underset{u \in \mathbb{R}^L}{\text{minimize}}: \ & \frac{1}{2} u^T Q u + p^T u \\ \text{subject to}: \ & a^T u \geq c \end{aligned}$$

8.11

$$\underset{u \in \mathbb{R}^L}{\text{minimize}}: \frac{1}{2} u^T Q u + p^T u + \max_{\alpha \geq 0} \alpha(c - a^T u)$$

此题是证明问题8.10与问题8.11的等价性

(a)因为

$$c - a^T u_0 \leq 0, \alpha \geq 0$$

所以

$$\begin{aligned} \alpha(c - a^T u_0) &\leq 0 \\ \max_{\alpha \geq 0} \alpha(c - a^T u_0) &= 0 \end{aligned}$$

(b)下面证明$u_1$也是8.10的解，首先证明$c - a^T u_1 \leq 0$，利用反证法：如果$c - a^T u_1 > 0$，那么

$$\alpha(c-a^Tu_1)\geq 0,\max_{\alpha\geq 0}\alpha(c-a^Tu_1)=+\infty$$

$$\frac{1}{2}u_1^TQu_1+p^Tu_1+\max_{\alpha\geq 0}\alpha(c-a^Tu_1)=+\infty$$

但是

$$\frac{1}{2}u_0^TQu_0+p^Tu_0+\max_{\alpha\geq 0}\alpha(c-a^Tu_0)=\frac{1}{2}u_0^TQu_0+p^Tu_0<+\infty$$

这就与$u_1$为8.11的最优解矛盾，所以$c-a^Tu_1\leq 0$，从而$u_1$也是8.10的可行解。

(c)证明两个方向的不等式即可。先证明$\frac{1}{2}u_1^TQu_1+p^Tu_1\geq\frac{1}{2}u_0^TQu_0+p^Tu_0$，注意由上题可知$c-a^Tu_1\leq 0$，所以

$$\max_{\alpha\geq 0}\alpha(c-a^Tu_1)=0$$

从而

$$\frac{1}{2}u_1^TQu_1+p^Tu_1+\max_{\alpha\geq 0}\alpha(c-a^Tu_1)=\frac{1}{2}u_1^TQu_1+p^Tu_1\geq\frac{1}{2}u_0^TQu_0+p^Tu_0$$

最后一个不等式是由于$u_0$为8.10的解。

再证明$\frac{1}{2}u_1^TQu_1+p^Tu_1\leq\frac{1}{2}u_0^TQu_0+p^Tu_0$，因为$u_1$为8.11的解，所以

$$\frac{1}{2}u_0^TQu_0+p^Tu_0+\max_{\alpha\geq 0}\alpha(c-a^Tu_0)\geq\frac{1}{2}u_1^TQu_1+p^Tu_1+\max_{\alpha\geq 0}\alpha(c-a^Tu_1)$$

注意

$$\frac{1}{2}u_0^TQu_0+p^Tu_0+\max_{\alpha\geq 0}\alpha(c-a^Tu_0)=\frac{1}{2}u_0^TQu_0+p^Tu_0$$
$$\frac{1}{2}u_1^TQu_1+p^Tu_1+\max_{\alpha\geq 0}\alpha(c-a^Tu_1)=\frac{1}{2}u_1^TQu_1+p^Tu_1$$

所以

$$\frac{1}{2}u_0^TQu_0+p^Tu_0\geq\frac{1}{2}u_1^TQu_1+p^Tu_1$$

结合两个不等式可得

$$\frac{1}{2}u_1^TQu_1+p^Tu_1=\frac{1}{2}u_0^TQu_0+p^Tu_0$$

(d)由之前推导过程我们知道，如果$u^*$为8.11的解，必然有

$$\alpha^*(c-a^Tu^*)=0$$

所以结论成立。

## Exercise 8.10 (Page 24)

Do the algebra. Derive $(*)$ and plug it into $\mathcal{L}(u,\alpha)$ to obtain $\mathcal{L}(\alpha)$.

$u_1 = \frac{\alpha_1+\alpha_2}{2}, u_1 = \frac{2\alpha_1+\alpha_3}{2}$，将这两式带入$\mathcal{L}(u,\alpha)$

$$
\begin{aligned}
\mathcal{L}(u,\alpha) &= u_1^2 + u_2^2 + \alpha_1(2-u_1-2u_2) - \alpha_2 u_1 - \alpha_3 u_2 \\
&= (\frac{\alpha_1+\alpha_2}{2})^2 + (\frac{2\alpha_1+\alpha_3}{2})^2 + \alpha_1(2-\frac{\alpha_1+\alpha_2}{2} - 2\frac{2\alpha_1+\alpha_3}{2}) - \alpha_2(\frac{\alpha_1+\alpha_2}{2}) - \alpha_3(\frac{2\alpha_1+\alpha_3}{2}) \\
&= \frac{1}{4}\alpha_1^2 + \frac{1}{2}\alpha_1\alpha_2 + \frac{1}{4}\alpha_2^2 + \alpha_1^2 + \alpha_1\alpha_3 + \frac{1}{4}\alpha_3^2 + \alpha_1(2-\frac{5}{2}\alpha_1 - \frac{1}{2}\alpha_2 - \alpha_3) - \frac{1}{2}\alpha_1\alpha_2 - \frac{1}{2}\alpha_2^2 - \alpha_1\alpha_3 - \frac{1}{2}\alpha_3^2 \\
&= (1+\frac{1}{4}-\frac{5}{2})\alpha_1^2 + (\frac{1}{4}-\frac{1}{2})\alpha_2^2 + (\frac{1}{4}-\frac{1}{2})\alpha_3^2 + (\frac{1}{2}-\frac{1}{2}-\frac{1}{2})\alpha_1\alpha_2 + (1-1-1)\alpha_1\alpha_3 + 2\alpha_1 \\
&= -\frac{5}{4}\alpha_1^2 - \frac{1}{4}\alpha_2^2 - \frac{1}{4}\alpha_3^2 - \frac{1}{2}\alpha_1\alpha_2 - \alpha_1\alpha_3 + 2\alpha_1
\end{aligned}
$$

## Exercise 8.11 (Page 28)

(a) Show that the problem in (8.21) is a standard QP-problem:

$$
\begin{aligned}
\underset{\alpha\in\mathbb{R}^N}{\text{minimize}}: \quad & \frac{1}{2}\alpha^T Q_D \alpha - 1_N^T \alpha \\
\text{subject to}: \quad & A_D\alpha \geq 0_{N+2}
\end{aligned}
$$

where $Q_D$ and $A_D$ ($D$ for dual) are given by:

$$
Q_D = \begin{bmatrix} y_1y_1x_1^Tx_1 & \dots & y_1y_Nx_1^Tx_N \\ y_2y_1x_2^Tx_1 & \dots & y_2y_Nx_2^Tx_N \\ \dots & \dots & \dots \\ y_Ny_1x_N^Tx_1 & \dots & y_Ny_Nx_N^Tx_N \end{bmatrix} \text{ and } A_D = \begin{bmatrix} y^T \\ -y^T \\ I_{N\times N} \end{bmatrix}
$$

[Hint: Recall that an equality corresponds to two inequalities.]

(b) The matrix $Q_D$ of quadratic coefficients is $[Q_D]_{mn} = y_m y_n x_m^T x_n$. Show that $Q_D = X_s X_s^T$, where $X_s$ is the 'signed data matrix',

$$
X_s = \begin{bmatrix} -y_1x_1^T- \\ -y_2x_2^T- \\ \dots \\ -y_Nx_N^T- \end{bmatrix}
$$

Hence, show that $Q_D$ is positive semi-definite. This implies that the QP-problem is convex.

回顾公式8.21

$$\underset{\alpha\in\mathbb{R}^N}{\text{minimize}}: \ \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N}y_n y_m \alpha_n \alpha_m x_n^T x_m - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to:} \ \sum_{n=1}^{N} y_n \alpha_n = 0, \alpha_n \geq 0 (n = 1, \dots, N)$$

(a)看到这个最小化的式子可以想到二次型，所以可以很自然的构造出$Q_D$，带入验证可得

$$\frac{1}{2}\alpha^T Q_D \alpha + 1_N^T \alpha = \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N}y_n y_m \alpha_n \alpha_m x_n^T x_m - \sum_{n=1}^{N}\alpha_n$$

接着看限制条件，将等式转化为两个不等式

$$\sum_{n=1}^{N} y_n \alpha_n = 0 \Leftrightarrow \sum_{n=1}^{N} y_n \alpha_n \geq 0, \sum_{n=1}^{N} y_n \alpha_n \leq 0 \Leftrightarrow y^T \alpha \geq 0, -y^T \alpha \geq 0$$

再看不等式限制条件

$$\alpha_n \geq 0 (n = 1, \dots, N) \Leftrightarrow I_{N\times N}\alpha \geq 0$$

从而$A_D \alpha \geq 0$等价于原来的约束条件。

(b)直接根据定义验证即可

$$\begin{aligned}
(X_s X_s^T)_{mn} &= X_s\text{第 }m\text{行 点积 }X_s^T\text{第 }n\text{列} \\
&= X_s\text{第 }m\text{行 点积 }X_s\text{第 }n\text{行} \\
&= (y_m x_m^T)^T (y_n x_n^T) \\
&= y_m y_n x_m x_n^T \\
&= y_m y_n x_m^T x_n
\end{aligned}$$

## Exercise 8.12 (Page 29)

If all the data is from one class, then $\alpha_n^* = 0$ for $n = 1, \dots, N$.

(a) What is $w^*$?

(b) What is $b^*$?

我们先来证明题目中的结论：

$$\text{如果所有数据都属于一类，那么}\alpha_n = 0, n = 1, \dots, N$$

因为所有数据都属于一类，$y_1 = y_2 = \ldots = y_N = t$，结合27页问题8.21的限制条件

$$\sum_{n=1}^{N} y_n \alpha_n = t(\sum_{n=1}^{N}\alpha_n) = 0, \alpha_n \geq 0 (n = 1, \dots, N)$$

可得

$$\sum_{n=1}^{N}\alpha_n = 0, \alpha_n \geq 0 (n = 1, \ldots, N)$$

所以$\alpha_n = 0, n = 1, \ldots, N$。

(a)因为$\alpha_n^* = 0, n = 1, \ldots, N$，那么

$$w^* = \sum_{n=1}^{N} y_n \alpha_n^* x_n = 0$$

(b)将$w^* = 0$带入限制条件$y_n(w^T x_n + b) \geq 1$

$$y_n(w^T x_n + b) = y_n b \geq 1$$

因为所有数据都属于一类，所以$y_1 = y_2 = \ldots = y_N = 1$或$y_1 = y_2 = \ldots = y_N = -1$。

如果$y_1 = y_2 = \ldots = y_N = 1$

$$y_n b = b \geq 1$$

如果$y_1 = y_2 = \ldots = y_N = -1$

$$y_n b = -b \geq 1$$
$$b \leq -1$$

## Exercise 8.13 (Page 31)

KKT complementary slackness gives that if $\alpha_n^* > 0$, then $(x_n, y_n)$ is on the boundary of the optimal fat-hyperplane and $y_n(w^{*T} x_n + b^*) = 1$. Show that the reverse is not true. Namely, it is possible that $\alpha_n^* = 0$ and yet $(x_n, y_n)$ is on the boundary satisfying $y_n(w^{*T} x_n + b^*) = 1$. [Hint: Consider a toy data set with two positive examples at (0, 0) and (1, 0), and one negative example at (0, 1).]

这题想要说明的是，即使$\alpha_n^* = 0$，也有可能$y_n(w^{*T} x_n + b^*) = 1$，说明我们找到的支持向量只是边界点的子集。

现在来看本题，考虑题目中给出的三个点$(0,0), (1,0), (0,1)$，其中$(0,0), (1,0)$标记为$+1$，$(0,1)$标记为$-1$。回忆之前的优化问题

$$\begin{aligned} \underset{\alpha \in \mathbb{R}^N}{\text{minimize}}: \quad & \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N} y_n y_m \alpha_n \alpha_m x_n^T x_m - \sum_{n=1}^{N}\alpha_n \\ \text{subject to:} \quad & \sum_{n=1}^{N} y_n \alpha_n = 0, \alpha_n \geq 0 (n = 1, \ldots, N) \end{aligned}$$

我们的条件为$N = 3, y_1 = y_2 = 1, y_3 = -1, x_1 = (0,0), x_2 = (1,0), x_3 = (0,1)$，从而

$$x_1^T x_1 = 0, x_1^T x_2 = 0, x_1^T x_3 = 0, x_2^T x_2 = 1, x_2^T x_3 = 0, x_3^T x_3 = 1$$

带入可得

$$\begin{aligned}\underset{\alpha\in\mathbb{R}^N}{\text{minimize}}:\ & \frac{1}{2}(\alpha_2^2+\alpha_3^2)-(\alpha_1+\alpha_2+\alpha_3)\\ \text{subject to:}\ & \alpha_1+\alpha_2-\alpha_3=0,\alpha_1\geq 0,\alpha_2\geq 0,\alpha_3\geq 0\end{aligned}$$

将$\alpha_3=\alpha_1+\alpha_2$带入可得

$$\begin{aligned}\frac{1}{2}(\alpha_2^2+\alpha_3^2)-(\alpha_1+\alpha_2+\alpha_3)&=\frac{1}{2}[\alpha_2^2+(\alpha_1+\alpha_2)^2]-2(\alpha_1+\alpha_2)\\ &=\frac{1}{2}(\alpha_1^2+2\alpha_1\alpha_2+2\alpha_2^2)-2\alpha_1-2\alpha_2\\ &=\alpha_2^2+(\alpha_1-2)\alpha_2+\frac{1}{2}\alpha_1^2-2\alpha_1\\ &=[\alpha_2+\frac{1}{2}(\alpha_1-2)]^2+\frac{1}{2}\alpha_1^2-2\alpha_1-\frac{1}{4}(\alpha_1-2)^2\\ &=[\alpha_2+\frac{1}{2}(\alpha_1-2)]^2+\frac{1}{4}(2\alpha_1^2-8\alpha_1-\alpha_1^2+4\alpha_1-4)\\ &=[\alpha_2+\frac{1}{2}(\alpha_1-2)]^2+\frac{1}{4}(\alpha_1^2-4\alpha_1-4)\\ &=[\alpha_2+\frac{1}{2}(\alpha_1-2)]^2+\frac{1}{4}(\alpha_1-2)^2-2\end{aligned}$$

所以当

$$\begin{gathered}\alpha_2+\frac{1}{2}(\alpha_1-2)=0,\alpha_1-2=0\\ \alpha_1=2,\alpha_2=0\end{gathered}$$

上式取最小值，此时

$$\alpha_3=\alpha_1+\alpha_2=2$$

结合$w,b$的公式

$$\begin{gathered}w^*=\sum_{n=1}^{N}y_n\alpha_n^*x_n\\ \text{对于}\alpha_s\neq 0,b^*=y_s-w^{*T}x_s\end{gathered}$$

带入可得

$$w=2x_1-2x_3=2(0,0)-2(0,1)=(0,-2)$$

取$s=1$，那么

$$b=y_1-w^Tx_1=1-(0,-2)^T(0,0)=1$$

注意$\alpha_2=0$，我们来看$x_2$的位置

$$w^Tx_2+b=(0,-2)^T(1,0)+1=1$$

所以$x_2$满足$w^Tx+b=1$，结论成立。

## Exercise 8.14 (Page 32)

Suppose that we removed a data point $(x_n, y_n)$ with $\alpha_n^* = 0$.

(a) Show that the previous optimal solution $\alpha^*$ remains feasible for the new dual problem (8.21) (after removing $\alpha_n^*$) .

(b) Show that if there is any other feasible solution for the new dual that has a lower objective value than $\alpha^*$, this would contradict the optimality of $\alpha^*$ for the original dual problem.

(c) Hence, show that $\alpha^*$ (minus $\alpha_n^*$) is optimal for the new dual.

(d) Hence, show that the optimal fat-hyperplane did not change.

(e) Prove the bound on $E_{\text{cv}}$ in (8.27).

首先回顾对偶问题

$$\begin{aligned}
&\underset{\alpha \in \mathbb{R}^N}{\text{minimize}}: \ \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N} y_n y_m \alpha_n \alpha_m x_n^T x_m - \sum_{n=1}^{N}\alpha_n \\
&\text{subject to:} \ \sum_{n=1}^{N} y_n \alpha_n = 0, \alpha_n \geq 0 (n = 1, \dots, N)
\end{aligned}$$

这题想要说明的是去除$\alpha_n = 0$的点并不影响该问题，不失一般性，不妨设$(x_N, y_N)$对应的$\alpha_N^* = 0$，那么删除$(x_N, y_N)$之后对偶问题变为

$$\begin{aligned}
&\underset{\alpha \in \mathbb{R}^{N-1}}{\text{minimize}}: \ \frac{1}{2}\sum_{m=1}^{N-1}\sum_{n=1}^{N-1} y_n y_m \alpha_n \alpha_m x_n^T x_m - \sum_{n=1}^{N-1}\alpha_n \\
&\text{subject to:} \ \sum_{n=1}^{N-1} y_n \alpha_n = 0, \alpha_n \geq 0 (n = 1, \dots, N-1)
\end{aligned}$$

接着定义以下符号

$$f(\alpha_1, \dots, \alpha_N) = \frac{1}{2}\sum_{m=1}^{N}\sum_{n=1}^{N} y_n y_m \alpha_n \alpha_m x_n^T x_m - \sum_{n=1}^{N}\alpha_n$$

所以对偶问题可以写成如下形式

$$\begin{aligned}
&\underset{\alpha \in \mathbb{R}^N}{\text{minimize}}: \ f(\alpha_1, \dots, \alpha_N) \\
&\text{subject to:} \ \sum_{n=1}^{N} y_n \alpha_n = 0, \alpha_n \geq 0 (n = 1, \dots, N)
\end{aligned}$$

注意到

$$\frac{1}{2}\sum_{m=1}^{N-1}\sum_{n=1}^{N-1} y_n y_m \alpha_n \alpha_m x_n^T x_m - \sum_{n=1}^{N-1}\alpha_n = f(\alpha_1, \dots \alpha_{N-1}, 0)$$

所以删除$(x_N, y_N)$之后对偶问题变为

$$\begin{aligned}
&\underset{\alpha\in\mathbb{R}^{N-1}}{\text{minimize}}: \ f(\alpha_1, \dots \alpha_{N-1}, 0) \\
&\text{subject to:} \ \sum_{n=1}^{N-1} y_n\alpha_n = 0, \alpha_n \geq 0 (n = 1, \dots, N-1)
\end{aligned}$$

(a)因为$f(\alpha_1, \dots, \alpha_N)$表示的范围比$f(\alpha_1, \dots \alpha_{N-1}, 0)$更大，所以

$$\underset{\alpha\in\mathbb{R}^{N}}{\text{minimize}}: f(\alpha_1, \dots, \alpha_N) \leq \underset{\alpha\in\mathbb{R}^{N-1}}{\text{minimize}}: f(\alpha_1, \dots \alpha_{N-1}, 0)$$

注意原问题的最优解为$\alpha^*$，且$\alpha_N^* = 0$，从而

$$\begin{aligned}
f(\alpha_1^*, \dots, \alpha_{N-1}^*, 0) &= f(\alpha_1^*, \dots, \alpha_{N-1}^*, \alpha_N^*) \\
&= \underset{\alpha\in\mathbb{R}^{N}}{\text{minimize}}: f(\alpha_1, \dots, \alpha_N) \\
&\leq \underset{\alpha\in\mathbb{R}^{N-1}}{\text{minimize}}: f(\alpha_1, \dots \alpha_{N-1}, 0)
\end{aligned}$$

显然

$$f(\alpha_1^*, \dots, \alpha_{N-1}^*, 0) \geq \underset{\alpha\in\mathbb{R}^{N-1}}{\text{minimize}}: f(\alpha_1, \dots \alpha_{N-1}, 0)$$

结合以上两点可知

$$f(\alpha_1^*, \dots, \alpha_{N-1}^*, 0) = \underset{\alpha\in\mathbb{R}^{N-1}}{\text{minimize}}: f(\alpha_1, \dots \alpha_{N-1}, 0)$$

从而$\alpha_1^*, \dots, \alpha_{N-1}^*$为删除$x_N, y_N$之后的对偶问题的最优解。

(b)如果存在$\alpha^{'} = (\alpha_1^{'}, \dots, \alpha_{N-1}^{'})$使得

$$f(\alpha_1^{'}, \dots, \alpha_{N-1}^{'}, 0) < f(\alpha_1^*, \dots, \alpha_{N-1}^*, 0) = f(\alpha_1^*, \dots, \alpha_{N-1}^*, \alpha_N^*) = \underset{\alpha\in\mathbb{R}^{N}}{\text{minimize}}: f(\alpha_1, \dots, \alpha_N)$$

这就与$\alpha^*$为原问题的最优解相矛盾。

(c)(a)说明原问题的解一定是新问题的解，(b)说明新问题解不会比原问题更优，从而$\alpha_1^*, \dots, \alpha_{N-1}^*$为删除$x_N, y_N$之后的对偶问题的最优解。

（补充，个人感觉应该从(a)的部分就足够得出这个结论了。）

(d)回顾公式

$$\begin{aligned}
w^* &= \sum_{n=1}^{N} y_n\alpha_n^* x_n \\
\text{对于}\alpha_s \neq 0, b^* &= y_s - w^{*T}x_s
\end{aligned}$$

所以删除$\alpha_n^* = 0$的点不会改变$w^*$，也不会改变$b^*$，所以超平面没有变，结论成立。

(e)回顾下公式

$$E_{\text{cv}} = \frac{1}{N}\sum_{n=1}^{N} e_n \leq \frac{\#\text{ number of } \alpha_n^* > 0}{N}$$

回顾之前结论可知，我们删除$\alpha_n^* = 0$的点不改变超平面，所以$E_{\text{cv}}$不会变，只有删除$\alpha_n^* > 0$的点才会改变超平面，所以

$$\sum_{n=1}^{N} e_n \leq \#\text{ number of } \alpha_n^* > 0$$

$$E_{\text{cv}} = \frac{1}{N}\sum_{n=1}^{N} e_n \leq \frac{\#\text{ number of } \alpha_n^* > 0}{N}$$

## Exercise 8.15 (Page 38)

Consider two finite-dimensional feature transforms $\Phi_1$ and $\Phi_2$ and their corresponding kernels $K_1$ and $K_2$.

(a) Define $\Phi(x) = (\Phi_1(x), \Phi_2(x))$. Express the corresponding kernel of $\Phi$ in terms of $K_1$ and $K_2$.

(b) Consider the matrix $\Phi_1(x)\Phi_2(x)^T$ and let $\Phi(x)$ be the vector representation of the matrix (say, by concatenating all the rows). Express the corresponding kernel of $\Phi$ in terms of $K_1$ and $K_2$.

(c) Hence, show that if $K_1$ and $K_2$ are kernels, then so are $K_1 + K_2$ and $K_1 K_2$.

The results above can be used to construct the general polynomial kernels and (when extended to the infinite-dimensional transforms) to construct the general Gaussian-RBF kernels.

直接根据定义验证即可

(a)

$\Phi(x)^T\Phi(x^{'}) = (\Phi_1(x), \Phi_2(x))^T(\Phi_1(x^{'}), \Phi_2(x^{'})) = \Phi_1(x)^T\Phi_1(x^{'}) + \Phi_2(x)^T\Phi_2(x^{'}) = K_1(x, x^{'}) + K_2(x, x^{'})$

所以$\Phi(x) = (\Phi_1(x), \Phi_2(x))$的kernel为$K_1 + K_2$

(b)$\Phi(x)$是$\Phi_1(x)\Phi_2^T(x)$每一行拼接而成的向量，设$\Phi_1(x), \Phi_2(x) \in \mathbb{R}^n$，给出以下记号

$$\Phi^i(x) = \Phi_1^i(x)\Phi_2^T(x) \in \mathbb{R}^{1\times n}$$
$$\Phi_1^i(x)\text{为}\Phi_1(x)\text{的第}i\text{个分量}$$

那么

$$\Phi(x) = \begin{bmatrix} \Phi^1(x) & \Phi^2(x) & \dots & \Phi^n(x) \end{bmatrix} \in \mathbb{R}^{1\times n^2}$$

接着计算$\Phi(x)\Phi^T(x')$，注意$\Phi^i(x)$为行向量

$$
\begin{aligned}
(\Phi(x)\Phi^T(x^{'})) &= \sum_{i=1}^{n}(\Phi^i(x))\Phi^i(x^{'})^T \\
&= \sum_{i=1}^{n}(\Phi_1^i(x)\Phi_2^T(x))(\Phi_1^i(x^{'})\Phi_2^T(x^{'}))^T \\
&= \sum_{i=1}^{n}\Phi_1^i(x)\Phi_1^i(x^{'})\Phi_2^T(x)\Phi_2(x^{'}) \\
&= \sum_{i=1}^{n}\Phi_1^i(x)\Phi_1^i(x^{'})K_2(x,x^{'}) \\
&= K_2(x,x^{'})\sum_{i=1}^{n}\Phi_1^i(x)\Phi_1^i(x^{'}) \\
&= K_2(x,x^{'})K_1(x,x^{'})
\end{aligned}
$$

所以$\Phi(x)$对应的kernel为$K_1(x,x^{'})K_2(x,x^{'})$

(c)由(a),(b)可以直接推出。

## Exercise 8.16 (Page 42)

Show that the optimization problem in (8.30) is a QP-problem.

(a) Show that the optimization variable is $u = \begin{bmatrix} b \\ w \\ \xi \end{bmatrix}$, where $\xi = \begin{bmatrix} \xi_1 \\ \dots \\ \xi_N \end{bmatrix}$.

(b) Show that $u^* \leftarrow QP(Q,p,A,c)$, where

$$
Q = \begin{bmatrix} 0 & 0_d^T & 0_N^T \\ 0_d & I_d & 0_{d\times N} \\ 0_N & 0_{N\times d} & 0_{N\times N} \end{bmatrix}, p = \begin{bmatrix} 0_{d+1} \\ C1_N \end{bmatrix}, A = \begin{bmatrix} YX & I_N \\ 0_{N\times(d+1)} & I_N \end{bmatrix}, c = \begin{bmatrix} 1_N \\ 0_N \end{bmatrix}
$$

and $YX$ is the signed data matrix from Exercise 8.4.

(c) How do you recover $b^*$, $w^*$ and $\xi^*$ from $u^*$?

(d) How do you determine which data points violate the margin, which data points are on the edge of the margin and which data points are correctly separated and outside the margin?

首先回顾8.30

$$
\begin{aligned}
&\underset{w,b,\xi}{\text{minimize}} : \frac{1}{2}w^Tw + C\sum_{n=1}^{N}\xi_n \\
&\text{subject to: } y_n(w^Tx_n + b) \geq 1 - \xi_n \\
&\qquad\qquad\quad \xi_n \geq 0 \ (n = 1,2,\dots,N)
\end{aligned}
$$

Q-P问题的标准形式为

$$\underset{u\in R^L}{\text{minimize}}:\frac{1}{2}u^TQu+p^Tu$$
$$\text{subject to}:Au\geq c,$$

接下来验证结论。

(a)查看标准形式之后可以发现$u$为变量，8.30中的变量为$w,b,\xi_1,\dots,\xi_N$

所以可以记$u=\begin{bmatrix}b\\w\\\xi\end{bmatrix}$，其中$\xi=\begin{bmatrix}\xi_1\\\dots\\\xi_N\end{bmatrix}$

(b)分别验证即可，首先验证目标函数

$$\frac{1}{2}u^TQu=\frac{1}{2}\begin{bmatrix}b\\w\\\xi\end{bmatrix}^T\begin{bmatrix}0&0_d^T&0_N^T\\0_d&I_d&0_{d\times N}\\0_N&0_{N\times d}&0_{N\times N}\end{bmatrix}\begin{bmatrix}b\\w\\\xi\end{bmatrix}=\frac{1}{2}w^Tw$$
$$p^Tu=\begin{bmatrix}0_{d+1}\\C1_N\end{bmatrix}^T\begin{bmatrix}b\\w\\\xi\end{bmatrix}=C1_N^T\xi=C\sum_{n=1}^N\xi_n$$

接着验证限制条件

$$YX=\begin{bmatrix}y_1&y_1x_1^T\\\dots&\dots\\y_N&y_Nx_N^T\end{bmatrix},\begin{bmatrix}YX&I_N\end{bmatrix}\begin{bmatrix}b\\w\\\xi\end{bmatrix}=\begin{bmatrix}y_1b+y_1x_1^Tw+\xi_1\\\dots\\y_Nb+y_Nx_N^Tw+\xi_N\end{bmatrix}$$
$$\begin{bmatrix}0_{N\times(d+1)}&I_N\end{bmatrix}\begin{bmatrix}b\\w\\\xi\end{bmatrix}=\begin{bmatrix}\xi_1\\\dots\\\xi_N\end{bmatrix}$$

所以$Au\geq c$可以化为

$$Au=\begin{bmatrix}YX&I_N\\0_{N\times(d+1)}&I_N\end{bmatrix}\begin{bmatrix}b\\w\\\xi\end{bmatrix}=\begin{bmatrix}y_1b+y_1x_1^Tw+\xi_1\\\dots\\y_Nb+y_Nx_N^Tw+\xi_N\\\xi_1\\\dots\\\xi_N\end{bmatrix}\geq c=\begin{bmatrix}1_N\\0_N\end{bmatrix}$$

可以看出这就是题目中的条件。

(c)如果计算出了$u^*$，那么$b^*=u_0^*,w^*=(u_1^*,\dots,u_d^*),\xi^*=(u_{d+1}^*,\dots,u_{N+d}^*)$

(d)先看$\xi_i$，$\xi_i>0$表示误分，所以分类正确的点一定有$\xi_i=0$。对于$\xi_i=0$的点，如果$y_i(w^Tx_i+b)=1$，那么该点在边界上，否则该点不在边界上。

## Exercise 8.17 (Page 45)

Show that $E_{\text{svm}}(b,w)$ is an upper bound on the $E_{\text{in}}(b,w)$, where $E_{\text{in}}$ is the classification 0/1 error.

回顾两种误差的定义

$$E_{\text{svm}}(b,w)=\frac{1}{N}\sum_{n=1}^{N}\max(1-y_n(w^Tx_n+b),0)$$

$$E_{\text{in}}(b,w)=\frac{1}{N}\sum_{n=1}^{N}[\![\text{sign}(w^Tx_n+b)\neq y_n]\!]$$

所以只要比较

$$e_1=\max(1-y(w^Tx+b),0),e_2=[\![\text{sign}(w^Tx+b)\neq y]\!]$$

由于$y\in\{+1,-1\}$，对$e_2$进行变形

$$\begin{aligned}e_2&=[\![\text{sign}(w^Tx+b)\neq y]\!]\\&=[\![y\ \text{sign}(w^Tx+b)\neq 1]\!]\\&=[\![\ \text{sign}[y(w^Tx+b)]\neq 1]\!]\end{aligned}$$

设$t=y(w^Tx+b)$，所以$e_1,e_2$可以化为

$$e_1=\max(1-t,0)=\begin{cases}1-t & \text{如果 } t\leq 1\\0 & \text{如果 } t>1\end{cases}$$

$$e_2=[\![\ \text{sign}(t)\neq 1]\!]=\begin{cases}1 & \text{如果 } t<0\\0 & \text{如果 } t\geq 0\end{cases}$$

当$t>1$时，$e_1=e_2=0$，当$0\leq t\leq 1$时，$e_1\geq 0=e_2$，当$t<0$时，$e_1>1=e_2$，所以

$$e_1\geq e_2$$

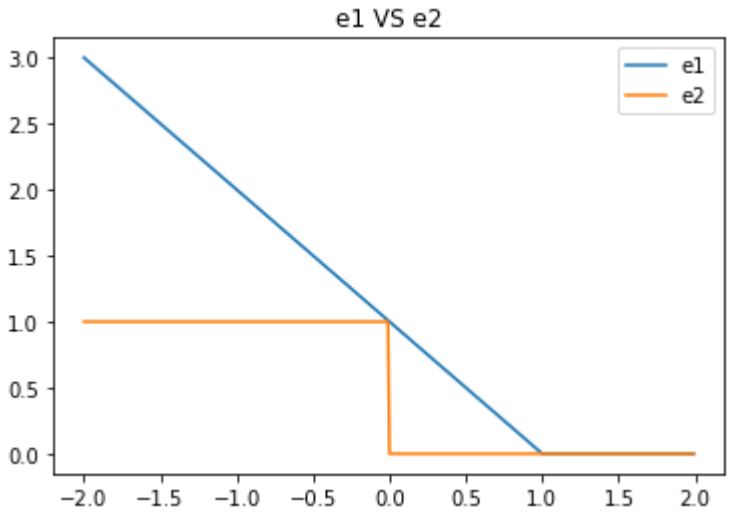
从而

$$E_{\text{in}}(b,w)\leq E_{\text{svm}}(b,w)$$

最后再从图像角度看一下这个结论。

```
# -*- coding: utf-8 -*-
"""
Created on Sun Mar 24 12:08:34 2019

@author: qinzhen
"""

import matplotlib.pyplot as plt
import numpy as np

def e1(t):
    return max(1 - t, 0)
```

```
def e2(t):
    if t >= 0:
        return 0
    else:
        return 1

x = np.arange(-2, 2, 0.01)
y1 = [e1(i) for i in x]
y2 = [e2(i) for i in x]

plt.plot(x, y1, label='e1')
plt.plot(x, y2, label='e2')
plt.legend()
plt.title('e1 VS e2')
plt.show()
```




# Part 2:Problems

## Problem 8.1 (Page 46)

Consider a data set with two data points $x_{\pm} \in \mathbb{R}^d$ having class $\pm 1$ respectively. Manually solve (8.4) by explicitly minimizing $||w||^2$ subject to the two separation constraints.

Compute the optimal (maximum margin) hyperplane $(b^*, w^*)$ and its margin. Compare with your solution to Exercise 8.1.

(8.4)在第7页，为原始的优化问题，来看下这里的条件

$$(w^T x_+ + b) \geq 1, -(w^T x_- + b) \geq 1$$

两式相加可得

$$w^T (x_+ - x_-) \geq 2$$

由柯西不等式可知

$$2 \leq w^T(x_+ - x_-) \leq ||w||||x_+ - x_-||,$$
$$\text{当且仅当}(w^T x_+ + b) = 1, -(w^T x_- + b) = 1\text{时等号成立}$$

所以

$$||w|| \geq \frac{2}{||x_+ - x_-||}$$

因此最优解为

$$||w^*|| = \frac{2}{||x_+ - x_-||}$$

其中$w^*$满足

$$(w^{*T} x_+ + b) = 1, -(w^{*T} x_- + b) = 1, w^{*T}(x_+ - x_-) = 2$$

接下来来求解$(b^*, w^*)$

$$w^{*T}(x_+ - x_-) = ||w^*||||x_+ - x_-|| \cos(\theta) = \frac{2}{||x_+ - x_-||}||x_+ - x_-|| \cos(\theta) = 2\cos(\theta) = 2$$
$$\cos(\theta) = 1$$

从而$w^*$与$x_+ - x_-$同向，可以设$w^* = k(x_+ - x_-), k > 0$，两边取模

$$||w^*|| = \frac{2}{||x_+ - x_-||} = k||x_+ - x_-||$$
$$k = \frac{2}{||x_+ - x_-||^2}$$
$$w^* = 2\frac{x_+ - x_-}{||x_+ - x_-||^2}$$

带入$(w^{*T} x_+ + b^*) = 1$可得

$$2(\frac{x_+ - x_-}{||x_+ - x_-||^2})^T x_+ + b^* = 1$$

$$b^* = 1 - 2\frac{x_+^T x_+ - x_-^T x_+}{||x_+ - x_-||^2} = 1 - \frac{2x_+^T x_+ - 2x_-^T x_+}{x_+^T x_+ + x_-^T x_- - 2x_+^T x_-} = \frac{x_-^T x_- - x_+^T x_+}{x_+^T x_+ + x_-^T x_- - 2x_+^T x_-} = \frac{||x_-||^2 - ||x_+||^2}{||x_+ - x_-||^2}$$

所以

$$(b^*, w^*) = (\frac{||x_-||^2 - ||x_+||^2}{||x_+ - x_-||^2}, 2\frac{x_+ - x_-}{||x_+ - x_-||^2})$$
$$\text{margin为}\frac{1}{||w^*||} = \frac{||x_+ - x_-||}{2}$$

最后验证Exercise 8.1的结论：$x_+, x_-$中至少有一点到超平面的距离为$\frac{||x_+ - x_-||}{2}$，由等号成立的条件可知

$$(w^{*T}x_+ + b) = 1, -(w^{*T}x_- + b) = 1$$

所以

$$d_+ = \frac{|w^{*T}x_+ + b^*|}{||w^*||} = \frac{1}{||w^*||} = \frac{||x_+ - x_-||}{2}, d_- = \frac{|w^{*T}x_- + b^*|}{||w^*||} = \frac{1}{||w^*||} = \frac{||x_+ - x_-||}{2}$$

（这题做的有些麻烦，直观结论就是最优超平面为中垂面。）

## Problem 8.2 (Page 46)

Consider a data set with three data points in $\mathbb{R}^2$:

$$X = \begin{bmatrix} 0 & 0 \\ 0 & -1 \\ -2 & 0 \end{bmatrix}, y = \begin{bmatrix} -1 \\ -1 \\ +1 \end{bmatrix}$$

Manually solve (8.4) to get the optimal hyperplane $(b^*, w^*)$ and its margin.

来看下此时的限制条件

$$-b \geq 1, -(-w_2 + b) \geq 1, (-2w_1 + b) \geq 1$$

第一个不等式和第三个不等式相加可得

$$-2w_1 \geq 2, w_1 \leq -1$$

对第二个式子进行变形

$$w_2 \geq 1 + b$$

所以

$$\frac{1}{2}w^T w = \frac{1}{2}(w_1^2 + w_2^2) \geq \frac{1}{2}(1 + 0)$$

当且仅当$w_1 = -1, w_2 = 0$时等号成立

将这两个条件带入第三个不等式可得

$$b + 2 \geq 1, b \geq -1$$

结合第一个不等式我们知道

$$b = -1$$

从而

$$(b^*, w^*) = (-1, (-1, 0))$$
$$\text{margin} = \frac{1}{||w^*||} = 1$$

## Problem 8.3 (Page 46)

Manually solve the dual optimization from Example 8.8 to obtain the same $\alpha^*$ that was obtained in the text using a QP-solver. Use the following steps.

(a) Show that the dual optimization problem is to minimize

$$\mathcal{L}(\alpha) = 4\alpha_2^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 - 4\alpha_2\alpha_3 - 6\alpha_2\alpha_4 + 6\alpha_3\alpha_4 - \alpha_1 - \alpha_2 - \alpha_3 - \alpha_4$$

subject to the constraints.

$$\begin{aligned}\alpha_1 + \alpha_2 &= \alpha_3 + \alpha_4\\ \alpha_1, \alpha_2, \alpha_3, \alpha_4 &\geq 0.\end{aligned}$$

(b) Use the equality constraint to replace $\alpha_1$ in $\mathcal{L}(\alpha)$ to get

$$\mathcal{L}(\alpha) = 4\alpha_2^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 - 4\alpha_2\alpha_3 - 6\alpha_2\alpha_4 + 6\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4$$

(c) Fix $\alpha_3, \alpha_4 \geq 0$ and minimize $\mathcal{L}(\alpha)$ in (b) with respect to $\alpha_2$ to show that

$$\alpha_2 = \frac{\alpha_3}{2} + \frac{3\alpha_4}{4} \text{ and } \alpha_1 = \alpha_3 + \alpha_4 - \alpha_2 = \frac{\alpha_3}{2} + \frac{\alpha_4}{4}$$

Are these valid solutions for $\alpha_1, \alpha_2$?

(d) Use the expressions in (c) to reduce the problem to minimizing

$$\mathcal{L}(\alpha) = \alpha_3^2 + \frac{9}{4}\alpha_4^2 + 3\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4$$

subject to $\alpha_3, \alpha_4 \geq 0$. Show that the minimum is attained when $\alpha_3 = 1$ and $\alpha_4 = 0$. What are $\alpha_1, \alpha_2$? It's a relief to have QP-solvers for solving such problems in the general case!

(a)根据28页的4个矩阵，可以计算出原问题化为

$$\begin{aligned}\text{minmize: } & \mathcal{L}(\alpha) = 4\alpha_2^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 - 4\alpha_2\alpha_3 - 6\alpha_2\alpha_4 + 6\alpha_3\alpha_4 - \alpha_1 - \alpha_2 - \alpha_3 - \alpha_4\\ \text{subject to: } & \alpha_1 + \alpha_2 = \alpha_3 + \alpha_4\\ & \alpha_1, \alpha_2, \alpha_3, \alpha_4 \geq 0.\end{aligned}$$

(b)将$\alpha_1 + \alpha_2 = \alpha_3 + \alpha_4$带入可得

$$\begin{aligned}\mathcal{L}(\alpha) &= 4\alpha_2^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 - 4\alpha_2\alpha_3 - 6\alpha_2\alpha_4 + 6\alpha_3\alpha_4 - \alpha_1 - \alpha_2 - \alpha_3 - \alpha_4\\ &= 4\alpha_2^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 - 4\alpha_2\alpha_3 - 6\alpha_2\alpha_4 + 6\alpha_3\alpha_4 - (\alpha_3 + \alpha_4) - \alpha_3 - \alpha_4\\ &= 4\alpha_2^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 - 4\alpha_2\alpha_3 - 6\alpha_2\alpha_4 + 6\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4\end{aligned}$$

(c)固定$\alpha_3, \alpha_4$，求$\mathcal{L}(\alpha)$的最小值，当成二次函数处理即可

$$\begin{aligned}\mathcal{L}(\alpha) &= 4\alpha_2^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 - 4\alpha_2\alpha_3 - 6\alpha_2\alpha_4 + 6\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4 \\ &= 4\alpha_2^2 - (4\alpha_3 + 6\alpha_4)\alpha_2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 + 6\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4 \\ &= 4(\alpha_2 - \frac{4\alpha_3 + 6\alpha_4}{8})^2 + 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 + 6\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4 - \frac{(2\alpha_3 + 3\alpha_4)^2}{4}\end{aligned}$$

利用二次函数的性质可知，当$\alpha_2 = \frac{4\alpha_3+6\alpha_4}{8} = \frac{\alpha_3}{2} + \frac{3\alpha_4}{4}$时，$\mathcal{L}(\alpha)$取最小值。

再来看下$\alpha_1$, $\alpha_1 = \alpha_3 + \alpha_4 - \alpha_2 = \frac{\alpha_3}{2} + \frac{\alpha_4}{4}$。由于$\alpha_3, \alpha_4 \geq 0$，所以

$$\alpha_2 = \frac{\alpha_3}{2} + \frac{3\alpha_4}{4} \geq 0, \alpha_1 = \frac{\alpha_3}{2} + \frac{\alpha_4}{4} \geq 0$$

从而这两个表达式是合理的解。

(d)将$\alpha_2 = \frac{4\alpha_3+6\alpha_4}{8}$带入，然后再进行配方

$$\begin{aligned}\mathcal{L}(\alpha) &= 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 + 6\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4 - \frac{(2\alpha_3 + 3\alpha_4)^2}{4} \\ &= 2\alpha_3^2 + \frac{9}{2}\alpha_4^2 + 6\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4 - \alpha_3^2 - 3\alpha_3\alpha_4 - \frac{9}{4}\alpha_4^2 \\ &= \alpha_3^2 + \frac{9}{4}\alpha_4^2 + 3\alpha_3\alpha_4 - 2\alpha_3 - 2\alpha_4 \\ &= \alpha_3^2 + (3\alpha_4 - 2)\alpha_3 + \frac{9}{4}\alpha_4^2 - 2\alpha_4 \\ &= (\alpha_3 + \frac{3\alpha_4 - 2}{2})^2 + \frac{9}{4}\alpha_4^2 - 2\alpha_4 - \frac{(3\alpha_4 - 2)^2}{4} \\ &= (\alpha_3 + \frac{3\alpha_4 - 2}{2})^2 + \alpha_4 - 1\end{aligned}$$

由于$\alpha_4 \geq 0$可得

$$(\alpha_3 + \frac{3\alpha_4 - 2}{2})^2 + \alpha_4 - 1 \geq -1$$

$$\text{当且仅当} \alpha_4 = 0, \alpha_3 + \frac{3\alpha_4 - 2}{2} = 0 \text{时等号成立}$$

此时

$$\alpha_4 = 0, \alpha_3 = 1, \alpha_2 = \frac{4\alpha_3 + 6\alpha_4}{8} = \frac{1}{2}, \alpha_1 = \alpha_3 + \alpha_4 - \alpha_2 = \frac{1}{2}$$

实际上一般的二次规划问题都可以按照这个方式求解。

## Problem 8.4 (Page 47)

Set up the dual problem for the toy data set in Exercise 8.2. Then, solve the dual problem and compute $\alpha^*$, the optimal Lagrange multipliers.

首先看下数据

$$X=\begin{bmatrix}0&0\\2&2\\2&0\end{bmatrix},y=\begin{bmatrix}-1\\-1\\+1\end{bmatrix},w=\begin{bmatrix}1.2\\-3.2\end{bmatrix},b=-0.5$$

所以

$$x_1^Tx_j=0,x_2^Tx_2=8,x_2^Tx_3=4,x_3^Tx_3=4$$
$$y_1=y_2=-1,y_3=1$$

所以最小化的式子为

$$f=\frac{1}{2}(8\alpha_2^2-4\alpha_2\alpha_3+4\alpha_3^2)-(\alpha_1+\alpha_2+\alpha_3)$$

限制条件为

$$-\alpha_1-\alpha_2+\alpha_3=0$$

将$\alpha_3=\alpha_1+\alpha_2$带入可得

$$\begin{aligned}f&=4\alpha_2^2-2\alpha_2\alpha_3+2\alpha_3^2-(\alpha_1+\alpha_2+\alpha_3)\\&=4\alpha_2^2-2\alpha_2(\alpha_1+\alpha_2)+2(\alpha_1+\alpha_2)^2-2(\alpha_1+\alpha_2)\\&=4\alpha_2^2-2\alpha_2\alpha_1-2\alpha_2^2+2\alpha_1^2+4\alpha_1\alpha_2+2\alpha_2^2-2\alpha_1-2\alpha_2\\&=4\alpha_2^2+2(\alpha_1-1)\alpha_2+2\alpha_1^2-2\alpha_1\\&=4[\alpha_2+\frac{1}{4}(\alpha_1-1)]^2+2\alpha_1^2-2\alpha_1-\frac{1}{4}(\alpha_1-1)^2\\&=4[\alpha_2+\frac{1}{4}(\alpha_1-1)]^2+\frac{1}{4}(8\alpha_1^2-8\alpha_1-\alpha_1^2+2\alpha_1-1)\\&=4[\alpha_2+\frac{1}{4}(\alpha_1-1)]^2+\frac{1}{4}(7\alpha_1^2-6\alpha_1-1)\end{aligned}$$

由二次函数性质可知，当

$$\alpha_1=\frac{3}{7},\alpha_2+\frac{1}{4}(\alpha_1-1)=0,\alpha_2=\frac{1}{7}$$

上式取最小值，此时

$$\alpha_3=\alpha_1+\alpha_2=\frac{4}{7}$$

可以计算超平面的$w,b$，带入可得

$$w=\sum_{n=1}^{N}\alpha_ny_nx_n=-\frac{3}{7}x_1-\frac{1}{7}x_2+\frac{4}{7}x_3=-\frac{3}{7}(0,0)-\frac{1}{7}(2,2)+\frac{4}{7}(2,0)=(\frac{6}{7},-\frac{2}{7})$$
$$b=y_2-w^Tx_2=-1-(\frac{6}{7},-\frac{2}{7})^T(2,2)=-1-\frac{12}{7}+\frac{4}{7}=-\frac{15}{7}$$

## Problem 8.5 (Page 47)

[Bias and Variance of the Optimal Hyperplane] In this problem, you are to investigate the bias and variance of the optimal hyperplane in a simple setting. The input is $(x_1, x_2) \in [-1, 1]^2$ and the target function is $f(x) = \text{sign}(x_2)$.

The hypothesis set $\mathcal{H}$ contains horizontal linear separators $h(x) = \text{sign}(x_2 - a)$, where $-1 \leq a \leq 1$. Consider two algorithms:

**Random**: Pick a random separator from $\mathcal{H}$.

**SVM**: Pick the maximum margin separator from $\mathcal{H}$.

(a) Generate $3$ data point uniformly in the upper half of the input-space and $3$ data points in the lower half, and obtain $g_{\text{Random}}$ and $g_{\text{SVM}}$.

(b) Create a plot of your data, and your two hypotheses.

(c) Repeat part (a) for a million data sets to obtain one million Random and SVM hypotheses.

(d) Give a histogram of the values of $a_{\text{Random}}$ resulting from the random algorithm and another histogram of $a_{\text{SVM}}$ resulting from the optimal separators. Compare the two histograms and explain the differences.

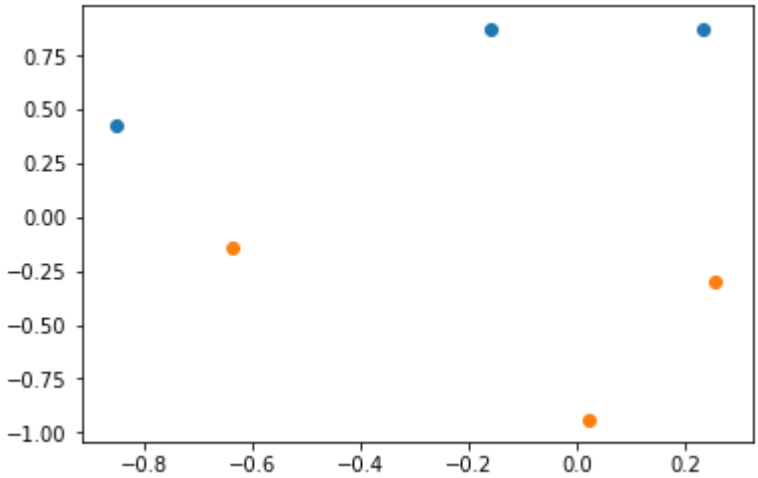
(e) Estimate the bias and var for the two algorithms. Explain your findings, in particular which algorithm is better for this toy problem.

8.5(a)(b)

构造产生数据集的函数，生成数据并作图

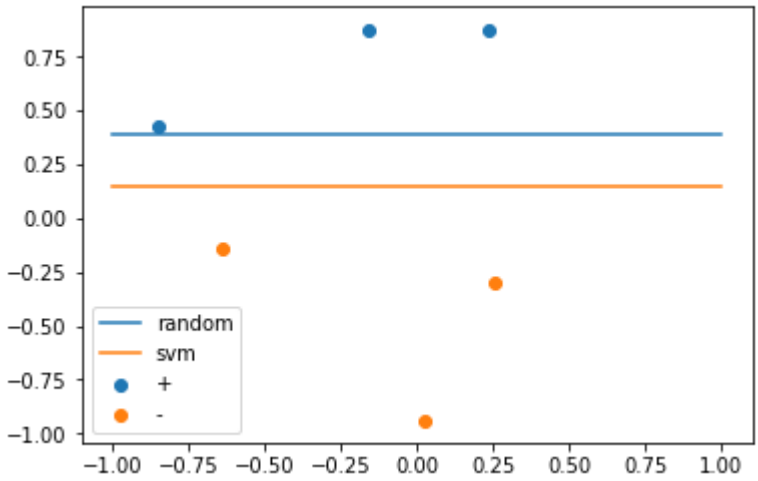
```
# -*- coding: utf-8 -*-
"""
Created on Sun Mar 24 12:38:01 2019

@author: qinzhen
"""

import numpy as np
import matplotlib.pyplot as plt

####a,b
def generate(n=3):
    """
    生成n个点
    """
    X1p = np.random.uniform(-1, 1, n)
    X2p = np.random.uniform(0, 1, n)
    X1n = np.random.uniform(-1, 1, n)
    X2n = np.random.uniform(-1, 0, n)

    return X1p, X2p, X1n, X2n

#生成数据
X1p, X2p, X1n, X2n = generate()
```

```
plt.scatter(X1p, X2p)
plt.scatter(X1n, X2n)
plt.show()
```




下面分别利用随机算法和svm产生判别函数，注意这里判别函数的的形式$h(x) = \text{sign}(x_2 - a)$的形式 决定了我们使用SVM时只要考虑纵坐标即可，只要找到+1类中纵坐标最小的点$(x_1, y_1)$，以及-1类中横坐标最大的点$(x_2, y_2)$，然后取$a = \frac{y_1+y_2}{2}$即可

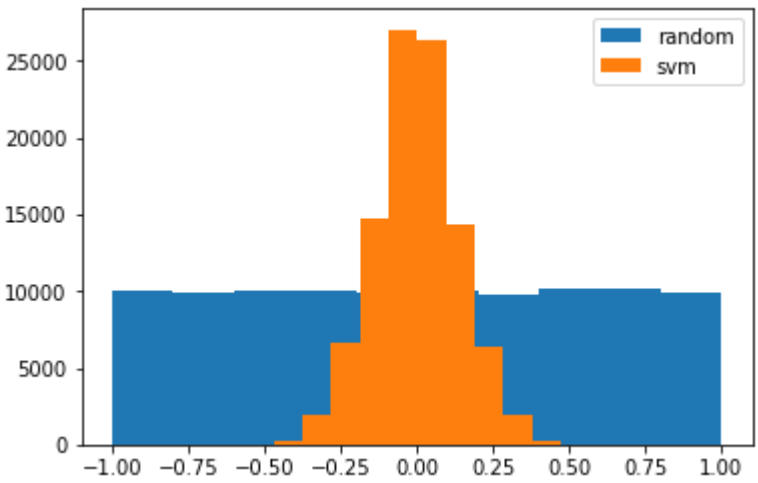
```
#产生结果
def mysvm(X2p, X2n):
    """找到+1类中纵坐标最小的点，-1类中纵坐标最大的点"""
    return (np.min(X2p) + np.max(X2n)) / 2

a_random = np.random.uniform(-1, 1)
a_svm = mysvm(X2p, X2n)

plt.scatter(X1p, X2p, label="+")
plt.scatter(X1n, X2n, label="-")
plt.plot([-1, 1], [a_random, a_random], label="random")
plt.plot([-1, 1], [a_svm, a_svm], label="svm")
plt.legend()
plt.show()

print("a_random = {}".format(a_random))
print("a_svm = {}".format(a_svm))
```




```
a_random = 0.3897481680806236
a_svm = 0.1437049653382313
```

(c)(d)题目中要求重复(a)100万次，我这里重复10万次

```
####c,d
N = 100000
A_random = []
A_svm = []
for i in range(N):
    X1p, X2p, X1n, X2n = generate()
    a_random = np.random.uniform(-1, 1)
    a_svm = mysvm(X2p, X2n)
    A_random.append(a_random)
    A_svm.append(a_svm)

#画直方图
plt.hist(A_random, label='random')
plt.hist(A_svm, label='svm')
plt.legend()
plt.show()
```



比较直方图可以发现，$a_{\text{Random}}$是均匀分布于$[-1,+1]$，而$a_{\text{SVM}}$集中在0附近，呈现钟形图。$a_{\text{Random}}$产生这样的直方图是因为本来就是在$[-1,+1]$上随机取的，所以每个值的呈现次数应该大体一致，而$a_{\text{SVM}}$是取最大间隔分类器，因为我们的数据一半$x$轴上方，令一半在$x$轴下方，且$y$的绝对值范围一致，所以总体来说在大部分$a_{\text{SVM}}$集中在0附近。

(e)首先我们需要计算

$$\overline{a}_{\text{Random}}=\frac{1}{K}\sum_{n=1}^{K}a_{\text{Random}},\overline{a}_{\text{SVM}}=\frac{1}{K}\sum_{n=1}^{K}a_{\text{SVM}}$$

这部分利用上题的数据计算即可。接着回忆方差，偏差公式

$$\text{bias}(x)=(\overline{g}(x)-f(x))^2,\text{bias}=\mathbb{E}[\text{bias}(x)]$$
$$\text{var}(x)=\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(x)-\overline{g}(x))^2],\text{var}=\mathbb{E}[\text{var}(x)]$$

这里计算积分比较麻烦，我使用

$$\frac{1}{N}\sum_{i=1}^{N}\text{bias}(x_i)$$

模拟bias

```
#计算a_random_mean,a_svm_mean
A_random = np.array(A_random)
A_svm = np.array(A_svm)
a_random_mean = np.mean(A_random)
a_svm_mean = np.mean(A_svm)

X_2 = np.random.uniform(-1,1,100000)
Y = np.sign(X_2)
Y_random_mean = np.sign(X_2 - a_random_mean)
Y_svm_mean = np.sign(X_2 - a_svm_mean)

bias_random = np.mean(Y != Y_random_mean)
bias_svm = np.mean(Y != Y_svm_mean)
print("bias_random =",bias_random)
print("bias_svm =",bias_svm)
```

```
bias_random = 0.00038
bias_svm = 1e-05
```

可以看到两者都是非常小的，再来看下var，这部分相对麻烦一点，我们根据(a)的方法生成一组数据，对这组数据求出$g^{(\mathcal{D})}(x)$，然后利用

$$\frac{1}{N}\sum_{i=1}^{N}(g^{(\mathcal{D})}(x_i)-\overline{g}(x_i))^2$$

计算$\mathcal{D}$固定的var。我们重复这个方法多次，取不同的$\mathcal{D}$，就可以模拟var

```
#计算var_random_mean,var_svm
var_random = np.array([])
var_svm = np.array([])
for i in range(1000):
    #生成数据
    X1p, X2p, X1n, X2n = generate()
    #计算随机选择以及svm算法对应的系数
    a_random = np.random.uniform(-1, 1)
    a_svm = mysvm(X2p, X2n)
    #生成用于模拟的数据
    X2 = np.random.uniform(-1, 1, 1000)
    #计算标签
    Y_random = np.sign(X2 - a_random)
    Y_svm = np.sign(X2 - a_svm)
    #计算平均值
    Y_random_mean = np.sign(X2 - a_random_mean)
    Y_svm_mean = np.sign(X2 - a_svm_mean)
    #计算样本方差
    var_random = np.append(var_random, np.mean(Y_random_mean != Y_random))
    var_svm = np.append(var_svm, np.mean(Y_svm_mean != Y_svm))

print("var_svm = {}".format(np.mean(var_svm)))
print("var_random = {}".format(np.mean(var_random)))
```

```
var_svm = 0.05110900000000001
var_random = 0.250094
```

可以看到，SVM的var很小，这也符合我们刚刚看到的直方图。

## Problem 8.6 (Page 47)

Show that $\sum_{n=1}^{N} ||x_n - \mu||^2$ is minimized at $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$

直接求梯度即可，记$f(\mu) = \sum_{n=1}^{N} ||x_n - \mu||^2$

$$
\begin{aligned}
\nabla f(\mu) &= \nabla \sum_{n=1}^{N} ||x_n - \mu||^2 \\
&= \nabla \sum_{n=1}^{N} (x_n - \mu)^T (x_n - \mu) \\
&= \nabla (\sum_{n=1}^{N} x_n^T x_n - 2x_n^T \mu + \mu^T \mu) \\
&= 2N\mu - 2\sum_{n=1}^{N} x_n
\end{aligned}
$$

显然$f(\mu)$有最小值，所以当

$$2N\mu - 2\sum_{n=1}^{N} x_n = 0$$

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n \text{时取最小值}$$

## Problem 8.7 (Page 47)

For any $x_1, \ldots, x_N$ with $||x_n|| \leq R$ and $N$ even, show that there exists a balanced dichotomy $y_1, \ldots, y_n$ that satisfies

$$\sum_{n=1}^{N} y_n = 0$$

$$||\sum_{n=1}^{N} y_n x_n|| \leq \frac{NR}{\sqrt{N-1}}$$

(This is the geometric lemma that is need to bound the VC-dimension of $\rho$-fat hyperplanes by $\lceil R^2/\rho^2 \rceil + 1$.) The following steps are a guide for the proof. Suppose you randomly select $N/2$ of the labels $y_1, \ldots, y_N$ to be $+1$, the others being $-1$. By construction, $\sum_{n=1}^{N} y_n = 0$.

(a) Show $||\sum_{n=1}^{N} y_n x_n||^2 = \sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m x_n^T x_m$.

(b) When $n = m$, what is $y_n y_m$? Show that $\mathbb{P}[y_n y_m = 1] = (\frac{N}{2} - 1)/(N-1)$

when $n \neq m$. Hence show that

$$\mathbb{E}[y_n y_m] = \begin{cases} 1 & m = n \\ -\frac{1}{N-1}, & m \neq n \end{cases}$$

(c) Show that

$$\mathbb{E}[||\sum_{n=1}^{N} y_n x_n||^2] = \frac{N}{N-1}\sum_{n=1}^{N} ||x_n - \overline{x}||^2$$

where the average vector $\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$. [Hint: Use linearity of expec tation in (a), and consider the cases $m = n$ and $m \neq n$ separately.]

(d) Show that $\sum_{n=1}^{N} ||x_n - \overline{x}||^2 \leq \sum_{n=1}^{N} ||x_n||^2 \leq NR^2$ [Hint: Problem 8.6.]

(e) Conclude that

$$\mathbb{E}[||\sum_{n=1}^{N} y_n x_n||^2] \leq \frac{N^2 R^2}{N-1}$$

and hence that

$$\mathbb{P}[||\sum_{n=1}^{N} y_n x_n|| \leq \frac{NR}{\sqrt{N-1}}] > 0$$

This means for some choice of $y_n, ||\sum_{n=1}^{N} y_n x_n|| \leq \frac{NR}{\sqrt{N-1}}$

This proof is called a probabilistic existence proof: if some random process can generate an object with positive probability, then that object must exist. Note that you prove existence of the required dichotomy without actually constructing it. In this case, the easiest way to construct a desired dichotomy is to randomly generate the balanced dichotomies until you have one that works.

(a)

$$\begin{aligned}
||\sum_{n=1}^{N} y_n x_n||^2 &= (\sum_{n=1}^{N} y_n x_n)^T \sum_{n=1}^{N} y_n x_n \\
&= (\sum_{n=1}^{N} y_n x_n^T) \sum_{m=1}^{N} y_m x_m \\
&= \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m x_n^T x_m
\end{aligned}$$

(b)因为$y_i \in \{1, -1\}$，所以当$m = n$时，$y_n y_m = 1$，$\mathbb{E}[y_n y_m] = 1$。当$m \neq n$时，计算$\mathbb{P}[y_n y_m = 1]$

$$\begin{aligned}
\mathbb{P}[y_n y_m = 1] &= \mathbb{P}[y_n = 1, y_m = 1] + \mathbb{P}[y_n = -1, y_m = -1] \\
&= 2\frac{N/2}{N}\frac{N/2-1}{N-1} \\
&= \frac{N/2-1}{N-1}
\end{aligned}$$

所以

$$\mathbb{P}[y_n y_m = -1] = 1 - \mathbb{P}[y_n y_m = 1] = \frac{N/2}{N-1}$$

$$\mathbb{E}[y_n y_m] = \frac{N/2-1}{N-1} - \frac{N/2}{N-1} = -\frac{1}{N-1}$$

$$\mathbb{E}[y_n y_m] = \begin{cases} 1 & m = n \\ -\frac{1}{N-1}, & m \neq n \end{cases}$$

(c)结合(a),(b)的结论一起即可

$$
\begin{aligned}
\mathbb{E}[||\sum_{n=1}^{N} y_n x_n||^2] &= \mathbb{E}[\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m x_n^T x_m] \\
&= \sum_{n=1}^{N}\sum_{m=1}^{N} \mathbb{E}[y_n y_m x_n^T x_m] \\
&= \sum_{n=1}^{N} x_n^T x_n - \frac{1}{N-1}\sum_{n=1}^{N}\sum_{m\neq n} x_n^T x_m \\
&= \frac{1}{N-1}[(N-1)\sum_{n=1}^{N} x_n^T x_n - \sum_{n=1}^{N}\sum_{m\neq n} x_n^T x_m] \\
&= \frac{1}{N-1}[N\sum_{n=1}^{N} x_n^T x_n - \sum_{n=1}^{N}\sum_{m=1}^{N} x_n^T x_m]
\end{aligned}
$$

注意$\overline{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$，所以

$$
\overline{x}^T\overline{x} = \frac{1}{N^2}(\sum_{n=1}^{N} x_n)^T \sum_{n=1}^{N} x_n = \frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N} x_n^T x_m
$$

接着处理等式右边

$$
\begin{aligned}
\frac{N}{N-1}\sum_{n=1}^{N}||x_n - \overline{x}||^2 &= \frac{N}{N-1}\sum_{n=1}^{N}(x_n-\overline{x})^T(x_n-\overline{x}) \\
&= \frac{N}{N-1}\sum_{n=1}^{N}(x_n^T x_n - 2x_n^T\overline{x} + \overline{x}^T\overline{x}) \\
&= \frac{N}{N-1}[\sum_{n=1}^{N} x_n^T x_n - 2(\sum_{n=1}^{N} x_n)^T\overline{x} + N\overline{x}^T\overline{x}] \\
&= \frac{N}{N-1}[\sum_{n=1}^{N} x_n^T x_n - 2N\overline{x}^T\overline{x} + N\overline{x}^T\overline{x}] \\
&= \frac{N}{N-1}[\sum_{n=1}^{N} x_n^T x_n - N\overline{x}^T\overline{x}] \\
&= \frac{1}{N-1}[N\sum_{n=1}^{N} x_n^T x_n - N^2\overline{x}^T\overline{x}] \\
&= \frac{1}{N-1}[N\sum_{n=1}^{N} x_n^T x_n - N^2(\frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N} x_n^T x_m)] \\
&= \frac{1}{N-1}[N\sum_{n=1}^{N} x_n^T x_n - \sum_{m=1}^{N}\sum_{n=1}^{N} x_n^T x_m]
\end{aligned}
$$

结合之前的论述可知，左边=右边，即

$$\mathbb{E}[||\sum_{n=1}^{N} y_n x_n||^2] = \frac{N}{N-1}\sum_{n=1}^{N}||x_n - \overline{x}||^2$$

(d)由8.6可知

$$\sum_{n=1}^{N}||x_n - \overline{x}||^2 \leq \sum_{n=1}^{N}||x_n||^2$$

注意$||x_n|| \leq R$，所以

$$\sum_{n=1}^{N}||x_n - \overline{x}||^2 \leq \sum_{n=1}^{N}||x_n||^2 \leq NR^2$$

(e)结合(c),(d)可知

$$\mathbb{E}[||\sum_{n=1}^{N} y_n x_n||^2] = \frac{N}{N-1}\sum_{n=1}^{N}||x_n - \overline{x}||^2 \leq \frac{N}{N-1}\sum_{n=1}^{N}||x_n||^2 \leq \frac{N^2R^2}{N-1}$$

由概率知识可知存在$y_1, \ldots, y_n$使得$\mathbb{P}[||\sum_{n=1}^{N} y_n x_n|| \leq \frac{NR}{\sqrt{N-1}}] > 0$，因为如果上式不成立，那么

$$\mathbb{P}[||\sum_{n=1}^{N} y_n x_n|| \leq \frac{NR}{\sqrt{N-1}}] = 0, \mathbb{P}[||\sum_{n=1}^{N} y_n x_n|| > \frac{NR}{\sqrt{N-1}}] = 1$$

$$\mathbb{E}[||\sum_{n=1}^{N} y_n x_n||^2] > \mathbb{E}[(\frac{NR}{\sqrt{N-1}})^2] \geq \frac{N^2R^2}{N-1}$$

这就与(d)矛盾。

## Problem 8.8 (Page 48)

We showed that if $N$ points in the ball of radius $R$ are shattered by hyperplanes with margin $\rho$, then $N \leq R^2/\rho^2 + 1$ when $N$ is even. Now consider $N$ odd, and $x_1, \ldots, x_N$ with $||x_n|| \leq R$ shattered by hyperplanes with margin $\rho$. Recall that $(w, b)$ implements $y_1, \ldots, y_N$ with margin $\rho$ if

$$\rho||w|| \leq y_n(w^T x_n + b), \text{for } n = 1, \ldots, N \tag{8.31}$$

Show that for $N = 2k + 1$ (odd), $N \leq R^2/\rho^2 + \frac{1}{N} + 1$ as follows: Consider random labelings $y_1, \ldots, y_N$ of the $N$ points in which $k$ of the labels are $+1$ and $k + 1$ are $-1$. Define $\ell_n = \frac{1}{k}$ if $y_n = +1$ and $\ell_n = \frac{1}{k+1}$ if $y_n = -1$.

(a) For any labeling with $k$ labels being $+1$, show, by summing (8.31) and using the Cauchy-Schwarz inequality, that

$$2\rho \leq ||\sum_{n=1}^{N} \ell_n y_n x_n||$$

(b) Show that there exists a labeling, with $k$ labels being $+1$, for which

$$||\sum_{n=1}^{N}\ell_n y_n x_n|| \leq \frac{2NR}{(N-1)\sqrt{N+1}}$$

(i) Show $||\sum_{n=1}^{N}\ell_n y_n x_n||^2 = \sum_{n=1}^{N}\sum_{m=1}^{N}\ell_n\ell_m y_n y_m x_n^T x_m$.

(ii) For $m=n$, show $\mathbb{E}[\ell_n\ell_m y_n y_m] = \frac{1}{k(k+1)}$

(iii) For $m\neq n$, show $\mathbb{E}[\ell_n\ell_m y_n y_m] = -\frac{1}{(N-1)k(k+1)}$

[Hint: $\mathbb{P}[\ell_n\ell_m y_n y_m = 1/k^2] = k(k-1)/N(N-1)$]

(iv) Show $\mathbb{E}[||\sum_{n=1}^{N} y_n x_n||^2] = \frac{N}{(N-1)k(k+1)}\sum_{n=1}^{N}||x_n-\overline{x}||^2$

(v) Use Problem 8.6 to conclude the proof as in Problem 8.7.

(c) Use (a) and (b) to show that $N \leq R^2/\rho^2 + \frac{1}{N} + 1$

(a)注意$N=2k+1$，现在假设$y_1,\ldots,y_k=1,y_{k+1},\ldots,y_{2k+1}=-1$

回顾(8.31)式

$$\rho||w|| \leq y_n(w^T x_n + b)$$

关于$y_i=1, y_j=-1$分别累加可得

$$k\rho||w|| = \sum_{n=1}^{k}\rho||w|| \leq \sum_{n=1}^{k} y_n(w^T x_n+b) = \sum_{n=1}^{k} y_n w^T x_n + kb$$

$$(k+1)\rho||w|| = \sum_{n=k+1}^{2k+1}\rho||w|| \leq \sum_{n=k+1}^{2k+1} y_n(w^T x_n+b) = \sum_{n=k+1}^{2k+1} y_n w^T x_n - (k+1)b$$

我们现在的任务是消去$b$，第一行的不等式乘以$k+1$加上第二行的不等式乘以$k$可得

$$\begin{aligned}2k(k+1)\rho||w|| &= (k+1)k\rho||w|| + k(k+1)\rho||w|| \\ &\leq (k+1)(\sum_{n=1}^{k} y_n w^T x_n + kb) + k(\sum_{n=k+1}^{2k+1} y_n w^T x_n - (k+1)b) \\ &= (k+1)\sum_{n=1}^{k} y_n w^T x_n + k\sum_{n=k+1}^{2k+1} y_n w^T x_n\end{aligned}$$

现在对$2(k+1)k\rho||w|| \leq (k+1)\sum_{n=1}^{k} y_n w^T x_n + k\sum_{n=k+1}^{2k+1} y_n w^T x_n$两边同除$(k+1)k$可得

$$
\begin{aligned}
2\rho||w|| &\leq \frac{1}{k}\sum_{n=1}^{k} y_n w^T x_n + \frac{1}{k+1}\sum_{n=k+1}^{2k+1} y_n w^T x_n \\
&= \sum_{n=1}^{k} \ell_n y_n w^T x_n + \sum_{n=k+1}^{2k+1} \ell_n y_n w^T x_n \\
&= \sum_{n=1}^{N} \ell_n y_n w^T x_n (\text{注意 } N = 2k+1) \\
&= w^T \sum_{n=1}^{N} \ell_n y_n x_n
\end{aligned}
$$

由Cauchy-Schwarz不等式可知

$$
w^T \sum_{n=1}^{N} \ell_n y_n x_n \leq ||w||.||\sum_{n=1}^{N} \ell_n y_n x_n||
$$

所以

$$
2\rho||w|| \leq ||w|||\sum_{n=1}^{N} \ell_n y_n x_n||
$$

$$
2\rho \leq ||\sum_{n=1}^{N} \ell_n y_n x_n||
$$

(b)

(i)直接展开计算即可

$$
\begin{aligned}
||\sum_{n=1}^{N} \ell_n y_n x_n||^2 &= (\sum_{n=1}^{N} \ell_n y_n x_n)^T (\sum_{n=1}^{N} \ell_n y_n x_n) \\
&= \sum_{n=1}^{N}\sum_{m=1}^{N} \ell_n \ell_m y_n y_m x_n^T x_m
\end{aligned}
$$

(ii)当$m = n$时，$y_n y_m = 1$，注意$N = 2k+1$，所以

$$
\begin{aligned}
\mathbb{E}[\ell_n \ell_m y_n y_m] &= \mathbb{E}[\ell_n \ell_n] \\
&= \frac{1}{k^2}\frac{k}{N} + \frac{1}{(k+1)^2}\frac{(k+1)}{N} \\
&= \frac{1}{kN} + \frac{1}{(k+1)N} \\
&= \frac{2k+1}{k(k+1)(2k+1)} \\
&= \frac{1}{k(k+1)}
\end{aligned}
$$

(iii)当$m \neq n$时，我们来计算$\mathbb{P}[\ell_n \ell_m y_n y_m = 1/k^2], \mathbb{P}[\ell_n \ell_m y_n y_m = 1/(k+1)^2], \mathbb{P}[\ell_n \ell_m y_n y_m = -1/k(k+1)]$

$$\begin{aligned}
\mathbb{P}[\ell_n \ell_m y_n y_m = 1/k^2] &= \mathbb{P}[y_n = 1, y_m = 1, n \neq m] \\
&= \mathbb{P}[y_n = 1]\mathbb{P}[y_m = 1, n \neq m | y_n = 1] \\
&= \frac{k}{N}\frac{k-1}{N-1} \\
&= \frac{k(k-1)}{N(N-1)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}[\ell_n \ell_m y_n y_m = 1/(k+1)^2] &= \mathbb{P}[y_n = -1, y_m = -1, n \neq m] \\
&= \mathbb{P}[y_n = -1]\mathbb{P}[y_m = -1, n \neq m | y_n = 1] \\
&= \frac{k+1}{N}\frac{k}{N-1} \\
&= \frac{k(k+1)}{N(N-1)}
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}[\ell_n \ell_m y_n y_m = -1/k(k+1)] &= 1 - \mathbb{P}[\ell_n \ell_m y_n y_m = 1/k^2] - \mathbb{P}[\ell_n \ell_m y_n y_m = 1/(k+1)^2] \\
&= 1 - \frac{k(k-1)}{N(N-1)} - \frac{k(k+1)}{N(N-1)} \\
&= 1 - \frac{2k^2}{N(N-1)} (\text{注意} N = 2k+1, 2k = N-1) \\
&= 1 - \frac{k}{N} \\
&= \frac{k+1}{N}
\end{aligned}$$

所以

$$\begin{aligned}
\mathbb{E}[\ell_n \ell_m y_n y_m] &= \frac{1}{k^2}\mathbb{P}[\ell_n \ell_m y_n y_m = 1/k^2] + \frac{1}{(k+1)^2}\mathbb{P}[\ell_n \ell_m y_n y_m = 1/(k+1)^2] - \frac{1}{(k+1)k}\mathbb{P}[\ell_n \ell_m y_n y_m = -1/k(k+1)] \\
&= \frac{1}{k^2}\frac{k(k-1)}{N(N-1)} + \frac{1}{(k+1)^2}\frac{k(k+1)}{N(N-1)} - \frac{1}{(k+1)k}\frac{k+1}{N} \\
&= \frac{1}{N(N-1)}(\frac{k-1}{k} + \frac{k}{k+1} - \frac{N-1}{k}) (\text{注意} N-1 = 2k) \\
&= \frac{1}{N(N-1)}(1 - \frac{1}{k} + 1 - \frac{1}{k+1} - 2) \\
&= -\frac{1}{N(N-1)}\frac{2k+1}{k(k+1)} \\
&= -\frac{1}{(N-1)k(k+1)}
\end{aligned}$$

(iv)将(i)(ii)(iii)带入

$$\begin{aligned}\mathbb{E}[||\sum_{n=1}^{N}\ell_n y_n x_n||^2] &= \mathbb{E}[\sum_{n=1}^{N}\sum_{m=1}^{N}\ell_n\ell_m y_n y_m x_n^T x_m] \\ &= \frac{1}{k(k+1)}\sum_{n=1}^{N}x_n^T x_n - \frac{1}{(N-1)k(k+1)}\sum_{n=1}^{N}\sum_{m\neq n}x_m^T x_n \\ &= \frac{1}{k(k+1)(N-1)}[(N-1)\sum_{n=1}^{N}x_n^T x_n - \sum_{n=1}^{N}\sum_{m\neq n}x_m^T x_n] \\ &= \frac{1}{k(k+1)(N-1)}[N\sum_{n=1}^{N}x_n^T x_n - \sum_{n=1}^{N}\sum_{m=1}^{N}x_m^T x_n]\end{aligned}$$

同8.7(c)的证明过程可知

$$\frac{N}{(N-1)k(k+1)}\sum_{n=1}^{N}||x_n-\overline{x}||^2 = \frac{1}{k(k+1)(N-1)}[N\sum_{n=1}^{N}x_n^T x_n - \sum_{n=1}^{N}\sum_{m=1}^{N}x_n^T x_m]$$

所以左边等于右边，即

$$\mathbb{E}[||\sum_{n=1}^{N}\ell_n y_n x_n||^2] = \frac{N}{(N-1)k(k+1)}\sum_{n=1}^{N}||x_n-\overline{x}||^2$$

(v)利用8.6可知

$$\begin{aligned}\mathbb{E}[||\sum_{n=1}^{N}\ell_n y_n x_n||^2] &= \frac{N}{(N-1)k(k+1)}\sum_{n=1}^{N}||x_n-\overline{x}||^2 \\ &\leq \frac{N}{(N-1)k(k+1)}\sum_{n=1}^{N}||x_n||^2 \\ &\leq \frac{N^2R^2}{(N-1)k(k+1)}\end{aligned}$$

同8.7(e)的讨论我们知道，存在一种label，使得

$$||\sum_{n=1}^{N}\ell_n y_n x_n|| \leq \sqrt{\frac{N^2R^2}{(N-1)k(k+1)}} = \frac{NR}{\sqrt{k(k+1)(N-1)}}$$

因为$N=2k+1$，所以

$$k(k+1) = \frac{N-1}{2}\frac{N+1}{2} = \frac{(N-1)(N+1)}{4}$$

$$\sqrt{k(k+1)(N-1)} = \sqrt{\frac{(N-1)(N+1)(N-1)}{4}} = \frac{N-1}{2}\sqrt{N+1}$$

$$\frac{NR}{\sqrt{k(k+1)(N-1)}} = \frac{2NR}{(N-1)\sqrt{N+1}}$$

$$||\sum_{n=1}^{N}\ell_n y_n x_n|| \leq \frac{NR}{\sqrt{k(k+1)(N-1)}} = \frac{2NR}{(N-1)\sqrt{N+1}}$$

所以结论成立。

(c)结合(a)(b)可知

$$2\rho \leq ||\sum_{n=1}^{N}\ell_n y_n x_n|| \leq \frac{2NR}{(N-1)\sqrt{N+1}}$$

$$\frac{R}{\rho} \geq \frac{(N-1)\sqrt{N+1}}{N}$$

$$\frac{R^2}{\rho^2} \geq \frac{(N-1)^2(N+1)}{N^2} = \frac{N^3-N^2-N+1}{N^2} = N-1-\frac{1}{N}+\frac{1}{N^2} > N-1-\frac{1}{N}$$

$$\text{即}\, N \leq R^2/\rho^2 + \frac{1}{N} + 1$$

所以结论成立。

## Problem 8.9 (Page 49)

Prove that for the separable case, if you remove a data point that is not a support vector, then the maximum margin classifier does not change. You may use the following steps are a guide. Let $g$ be the maximum margin classifier for all the data, and $g^-$ the maximum margin classifier after removal of a data point that is not a support vector.

(a) Show that $g$ is a separator for $\mathcal{D}^-$ , the data minus the non-support vector.

(b) Show that the maximum margin classifier is unique.

(c) Show that if $g^-$ has larger margin than $g$ on $\mathcal{D}^-$, then it also has larger margin on $\mathcal{D}$, a contradiction. Hence, conclude that $g$ is the maximum margin separator for $\mathcal{D}^-$.

删除非支持向量相当于删除$\alpha_i = 0$的点，所以这题和Exercise 8.14一致。

## Problem 8.10 (Page 49)

An essential support vector is one whose removal from the data set changes the maximum margin separator. For the separable case, show that there are at most $d+1$ essential support vectors. Hence, show that for the separable case,

$$E_{\text{cv}} \leq \frac{d+1}{N}$$

回顾第9,10页，我们求解SVM最后转化为一个二次规划问题

$$\begin{aligned}&\underset{u\in\mathbb{R}^L}{\text{minimize}}: \frac{1}{2}u^TQu+p^Tu\\&\text{subject to: } Au\geq c\\&\qquad \text{其中} A\in\mathbb{R}^{N\times(d+1)}\end{aligned}$$

假设这个问题的解为$u^*=\begin{bmatrix}b^* & w^*\end{bmatrix}^T\in\mathbb{R}^{d+1}$，由求解的过程我们知道，$u^*$必然满足条件

$$A_iu^*=c_i\text{, }A_i\text{表示}A\text{的第}i\text{个行向量, }c_i\text{表示}c\text{的第}i\text{个分量}$$

将满足这个等式的行向量构成的子矩阵记为$A^*\in\mathbb{R}^{n\times(d+1)}$，对应的$c_i$构成的向量记为$c^*$，那么必然有

$$A^*u^*=c^*$$

设$\text{rank}(A^*)=r\leq d+1$，所以求解该方程组最多需要$r$个方程即可，即只需要$r$个$A_i$就可以确定$u^*$，删除其余的$A_j$不影响$u^*$的值，从而对应的$e_j=0$。所以

$$E_{\text{cv}}(\text{SVM})=\frac{1}{N}\sum_{n=1}^{N}e_n\leq\frac{r}{N}\leq\frac{d+1}{N}$$

## Problem 8.11 (Page 50)

Consider the version of the PLA that uses the misclassified data point $x_n$ with lowest index $n$ for the weight update. Assume the data is separable and given in some fixed but arbitrary order, and when you remove a data point, you do not alter this order. In this problem, prove that

$$E_{\text{cv}}(\text{PLA})\leq\frac{R^2}{N\rho^2}$$

where $\rho$ is the margin (half the width) of the maximum margin separating hyperplane that would (for example) be returned by the SVM. The following steps are a guide for the proof.

(a) Use the result in Problem 1.3 to show that the number of updates $T$ that the PLA makes is at most

$$T\leq\frac{R^2}{\rho^2}$$

(b) Argue that this means that PLA only 'visits' at most $\frac{R^2}{\rho^2}$ different points during the course of its iterations.

(c) Argue that after leaving out any point $(x_n,y_n)$ that is not 'visited', PLA will return the same classifier.

(d) What is the leave-one-out error $e_n$ for these points that were not visited? Hence, prove the desired bound on $E_{\text{cv}}(\text{PLA})$

(a) Problem 1.3的结论为

$$T \leq \frac{R^2||w^*||^2}{\rho^{*2}}$$

其中$R = \max_{1\leq n\leq N}||x_n||, w^*$为任意一个分离数据的权重，$\rho^* = \min_{1\leq n\leq N} y_n(w^{*T}x_n), T$为迭代次数所以，这里我们取$w$为$SVM$算法的解，那么此处$||w|| = 1, \rho^* = \rho$，带入可得

由条件可得此处$||w|| = 1, \rho^* = \rho$，带入上式得到

$$\frac{R^2||w^*||^2}{\rho^{*2}} = \frac{R^2}{\rho^2} \geq T$$

(b)根据(a)可知，最多更新$\frac{R^2}{\rho^2}$次，所以最多访问$\frac{R^2}{\rho^2}$个不同的点。

(c)由PLA的更新过程可知，最后的解$w^*$为访问过的点的线性组合，所以删除没访问过的点不会影响最终的解。

(d)由于删除没访问过的点不会影响最终的解，所以对应的$e_i = 0$，从而

$$E_{cv}(\text{PLA}) = \frac{\#\text{访问过的点}}{N} \leq \frac{R^2}{N\rho^2}$$

## Problem 8.12 (Page 50)

Show that optimal solution for soft-margin optimal hyperplane (solving optimization problem (8.30)) with $C \to \infty$ will be the same solution that was developed using linear programming in Problem 3.6(c).

先分别回顾下这两个问题。

8.30

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}w^T w + C\sum_{n=1}^{N}\xi_n \\ &\text{subject to } y_n(w^T x_n + b) \geq 1 - \xi_n \\ &\qquad\qquad \xi_n \geq 0 \end{aligned}$$

Problem 3.16(c)

$$\begin{aligned} \min_{w,\xi_n} \quad & \sum_{n=1}^{N}\xi_n \\ \text{subject to} \quad & y_n(w^T x_n) \geq 1 - \xi_n \\ & \xi_n \geq 0 \end{aligned}$$

注意这两个问题的$w$含义是不一样的，Problem 3.16(c)的$w = (w^*, b), x_n = (1, x_n^*)$，这样转化之后两个问题的条件就一致了。现在开始考虑题目中的问题，如果8.30的$C \to \infty$，那么如果我们要最小化$\frac{1}{2}w^T w + C\sum_{n=1}^{N}\xi_n$，必然要最小化$\sum_{n=1}^{N}\xi_n$，这就将问题转化为3.16(c)。

## Problem 8.13 (Page 50)

The data for Figure 8.6(b) are given below: Use the data on the left with the 2nd and 3rd order polynomial transforms $\Phi_2, \Phi_3$ and the pseudo-inverse algorithm for linear regression from Chapter 3 to get weights $\tilde{w}$ for your final final hypothesis in $\mathcal{Z}$-space. The final hypothesis in $\mathcal{X}$-space is:

$$g(x) = \text{sign}(\tilde{w}^T \Phi(x) + \tilde{b})$$

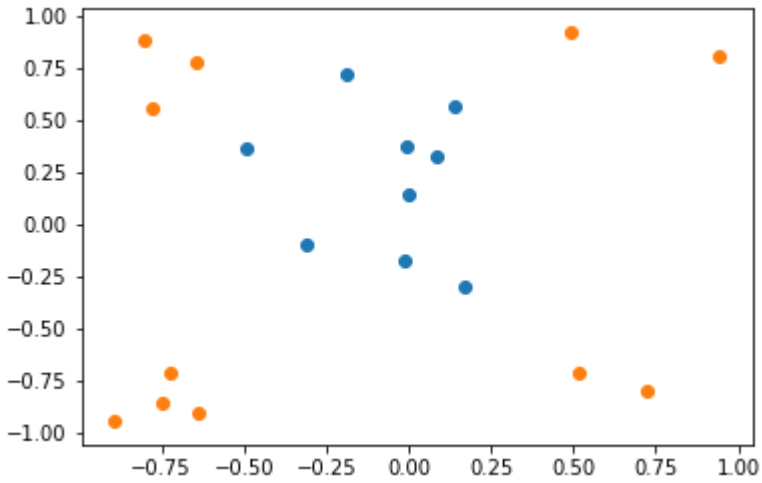
(a) Plot the classification regions for your final hypothesis in $\mathcal{X}$-space. Your results should look something like:

(b) Which of fits in part (a) appears to have overfitted?

(c) Use the pseudo-inverse algorithm with regularization parameter $\lambda = 1$ to address the overfitting you identified in part (d) Give a plot of the resulting classifier.

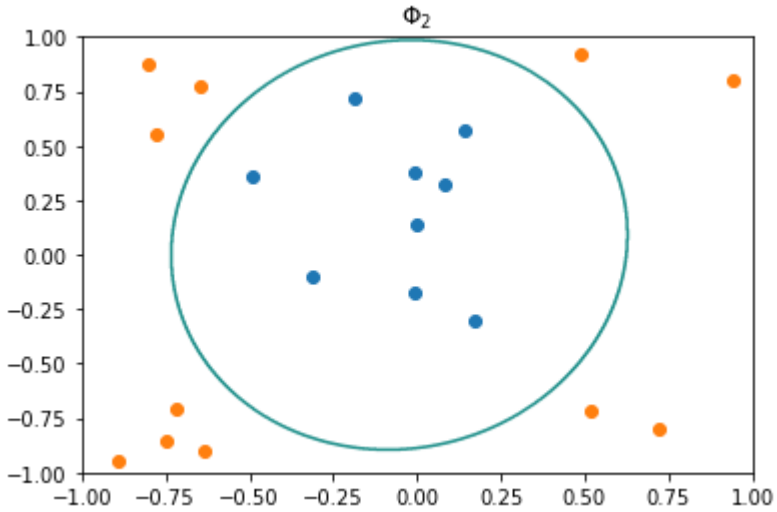
首先看下图像

```
# -*- coding: utf-8 -*-
"""
Created on Sun Mar 24 12:40:16 2019

@author: qinzhen
"""

import numpy as np
from sklearn import svm
from sklearn.preprocessing import PolynomialFeatures
import matplotlib.pyplot as plt
from numpy.linalg import inv

####(a)
#题目中的数据
X = np.array([(-0.494, 0.363), (-0.311, -0.101), (-0.0064, 0.374), (-0.0089, -0.173),
              (0.0014, 0.138), (-0.189, 0.718), (0.085, 0.32208), (0.171, -0.302), (0.142,
0.568),
              (0.491, 0.920), (-0.892, -0.946), (-0.721, -0.710), (0.519, -0.715),
              (-0.775, 0.551), (-0.646, 0.773), (-0.803, 0.878), (0.944, 0.801),
              (0.724, -0.795), (-0.748, -0.853), (-0.635, -0.905)])
#对应标签
y = np.array([1] * 9 + [-1] * 11)

#作图
x1p = X[y>0][:,0]
x2p = X[y>0][:,1]
x1n = X[y<0][:,0]
x2n = X[y<0][:,1]

plt.scatter(x1p, x2p)
plt.scatter(x1n, x2n)
plt.show()
```



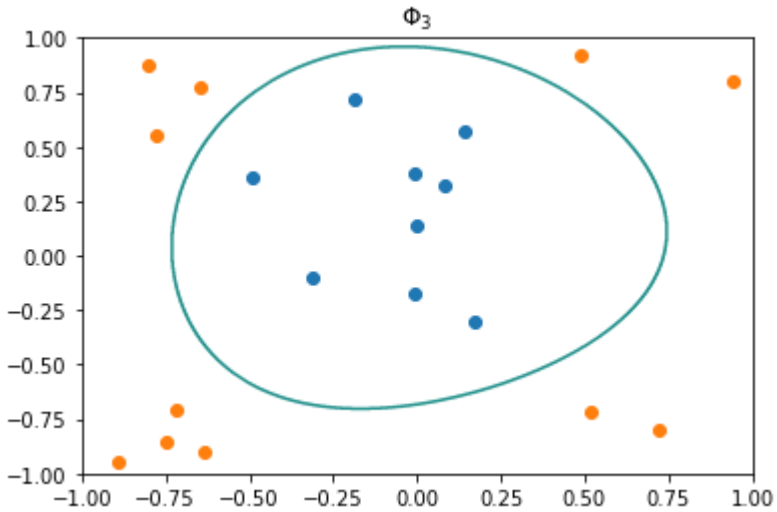

进行二次转换再使用SVM，然后作图

```
#二次转换
def contour(X1, X2, clf, poly):
    """
    计算每个点的标签
    """
    X = np.c_[X1.ravel(), X2.ravel()]
    X_poly = poly.fit_transform(X)
    label = clf.predict(X_poly)
    label = label.reshape(X1.shape)

    return label

poly2 = PolynomialFeatures(2)
X_poly2 = poly2.fit_transform(X)

clf = svm.SVC(kernel="linear", C=1e10)
clf.fit(X_poly2, y)

#点的数量
n = 1000
r = 1

#作点
a = np.linspace(-r, r, n)
b = np.linspace(-r, r, n)

#构造网格
A, B = np.meshgrid(a, b)
C = contour(A, B, clf, poly2)

#绘制等高线
plt.contour(A, B, C, 0)
plt.scatter(x1p, x2p)
plt.scatter(x1n, x2n)
plt.title('$\Phi_2$')
plt.show()
```




进行三次转换再使用SVM，然后作图

```
#三次转换
poly3 = PolynomialFeatures(3)
X_poly3 = poly3.fit_transform(X)

clf = svm.SVC(kernel="linear", C=1e10)
clf.fit(X_poly3, y)

#构造网格
A, B = np.meshgrid(a, b)
C = contour(A, B, clf, poly3)

#绘制等高线
plt.contour(A, B, C, 0)
plt.scatter(x1p, x2p)
plt.scatter(x1n, x2n)
plt.title('$\Phi_3$')
plt.show()
```



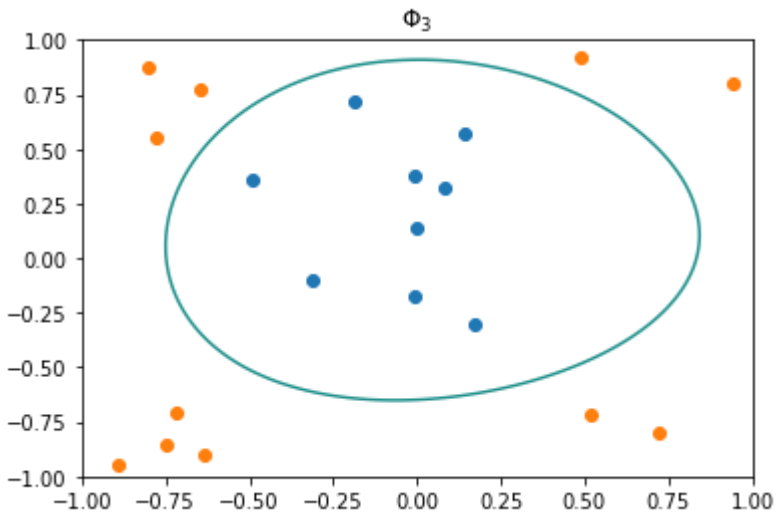

(b)这部分我和课本的图像不大一致，暂时没找到原因。

(c)直接利用

$$w = (X^T X + \lambda I)^{-1} X^T y$$

计算即可，这里转换成3次多项式

```
####(c)
Lambda = 1
n, d = X_poly3.shape
w = inv(X_poly3.T.dot(X_poly3) + Lambda * np.eye(d)).dot(X_poly3.T.dot(y))

label = poly3.fit_transform(np.c_[A.ravel(), B.ravel()]).dot(w)
C = label.reshape(A.shape)

plt.contour(A, B, C, 0)
plt.scatter(x1p, x2p)
plt.scatter(x1n, x2n)
plt.title('$\Phi_3$')
plt.show()
```




## Problem 8.14 (Page 51)

The kernel trick can be used with any model as long as fitting the data and the final hypothesis only require the computation of dot-products in the $\mathcal{Z}$-space. Suppose you have a kernel $K$, so

$$\Phi(x)^T \Phi(x') = K(x, x')$$

Let $Z$ be the data in the $\mathcal{Z}$-space. The pseudo-inverse algorithm for regularized regression computes optimal weights $\tilde{w}^*$ (in the $\mathcal{Z}$-space) that minimize

$$E_{\text{aug}}(\tilde{w}) = ||Z\tilde{w} - y||^2 + \lambda \tilde{w}^T \tilde{w}$$

The final hypothesis is $g(x) = \text{sign}(\tilde{w}^T \Phi(x))$. Using the representor theorem, the optimal solution can be written $\tilde{w}^* = \sum_{n=1}^{N} \beta_n^* z_n = Z^T \beta^*$

(a) Show that $\beta^*$ minimizes

$$E(\beta) = ||K\beta - y||^2 + \lambda\beta^T K\beta$$

where $K$ is the $N \times N$ Kernel-Gram matrix with entries $K_{ij} = K(x_i, x_j)$.

(b) Show that $K$ is symmetric.

(c) Show that the solution to the minimization problem in part (a) is:

$$\beta^* = (K + \lambda I)^{-1} y$$

Can $\beta^*$ be computed without ever 'visiting' the $\mathcal{Z}$-space?

(d) Show that the final hypothesis is

$$g(x) = \text{sign}\Big(\sum_{n=1}^{N} \beta_n^* K(x_n, x)\Big)$$

(a)将$\tilde{w} = \sum_{n=1}^{N} \beta_n z_n = Z^T \beta$带入可得

$$\begin{aligned} E_{\text{aug}}(\tilde{w}) = E(\beta) &= ||Z\tilde{w} - y||^2 + \lambda \tilde{w}^T \tilde{w} \\ &= ||ZZ^T\beta - y||^2 + \lambda (Z^T\beta)^T (Z^T\beta) \\ &= ||ZZ^T\beta - y||^2 + \lambda \beta^T ZZ^T \beta \\ &= ||K\beta - y||^2 + \lambda \beta^T K \beta \end{aligned}$$

所以$\beta^*$最小化

$$E(\beta) = ||K\beta - y||^2 + \lambda\beta^T K\beta$$

(b)记$Z = (\Phi(x_1), \ldots, \Phi(x_N))$，那么$K = ZZ^T$，所以

$$\begin{aligned} K^T &= (ZZ^T)^T = ZZ^T \\ xKx^T &= xZZ^T x^T = (xZ)(xZ)^T = ||xZ||^2 \geq 0 \end{aligned}$$

所以$K$是半正定对称矩阵。

(c)求梯度即可，注意$K$对称

$$\begin{aligned} \nabla_\beta E(\beta) &= \nabla(||K\beta - y||^2 + \lambda\beta^T K\beta) \\ &= \nabla(\beta^T K^T K\beta - 2y^T K\beta + y^T y + \lambda\beta^T K\beta) \\ &= 2K^T K\beta - 2K^T y + 2\lambda K\beta \\ &= 2KK\beta - 2Ky + 2\lambda K\beta \\ &= 2K[(K + \lambda I)\beta - y] \end{aligned}$$

如果

$$(K + \lambda I)\beta - y = 0, \beta = (K + \lambda I)^{-1} y$$

那么

$$\nabla_\beta E(\beta) = 0$$

所以最优解为

$$\beta^* = (K + \lambda I)^{-1} y$$

因为$K = ZZ^T, Z = (\Phi(x_1), \ldots, \Phi(x_N))$，所以不访问$\mathcal{Z}$也能计算出$\beta^*$

(d)将$\tilde{w}^* = \sum_{n=1}^{N} \beta_n^* z_n = Z^T \beta^*$代入

$$\begin{aligned} g(x) &= \text{sign}\Big(\tilde{w}^{*T} \Phi(x)\Big) \\ &= \text{sign}\Big((\sum_{n=1}^{N} \beta_n^* z_n)^T \Phi(x)\Big) \\ &= \text{sign}\Big(\sum_{n=1}^{N} \beta_n^* z_n^T \Phi(x)\Big) \\ &= \text{sign}\Big(\sum_{n=1}^{N} \beta_n^* \Phi(x_n)^T \Phi(x)\Big) \\ &= \text{sign}\Big(\sum_{n=1}^{N} \beta_n^* K(x_n, x)\Big) \end{aligned}$$

## Problem 8.15 (Page 52)

Structural Risk Minimization (SRM) is a useful framework for model selection that is related to Occam 's Razor. Define a structure - a nested sequence of hypothesis sets:

The SRM framework picks a hypothesis from each $\mathcal{H}_i$ by minimizmg $E_{\text{in}}$. That is, $g_i = \underset{h \in \mathcal{H}_i}{\text{argmin}}\, E_{\text{in}}(h)$ . Then, the framework selects the final hypothesis by minimizing $E_{\text{in}}$ and the model complexity penaltyn $\Omega$. That is, $g^* = \underset{i=1,2\ldots}{\text{argmin}}(\, E_{\text{in}}(g_i) + \Omega(\mathcal{H}_i))$. Note that $\Omega(\mathcal{H}_i)$ should be non decreasing in $i$ because of the nested structure.

(a) Show that the in sample error $E_{\text{in}}(g_i)$ is non increasing in $i$.

(b) Assume that the framework finds $g^* \in \mathcal{H}_i$ with probability $p_i$ . How does $p_i$ relate to the complexity of the target function?

(c) Argue that the $p_i$'s are unknown but $p_0 \leq p_1 \leq p_2 \leq \cdots \leq 1$.

(d) Suppose $g^* = g_i$.Show that

$$\mathbb{P}[|E_{\text{in}}(g_i) - E_{\text{out}}(g_i)| > \epsilon | g^* = g_i] \leq \frac{1}{p_i} 4 m_{\mathcal{H}_i}(2N) e^{-\frac{\epsilon^2 N}{8}}$$

Here, the conditioning is on selecting $g_i$ as the final hypothesis by SRM. [Hint: Use the Bayes theorem to decompose the probability and then apply the VC bound on one of the terms]

You may interpret this result as follows: if you use SRM and end up with $g_i$ , then the generalization bound is a factor $\frac{1}{p_i}$ worse than the bound you would have gotten had you simply started with $\mathcal{H}$.

同Chapter 5 Problem 5.2。

## Problem 8.16 (Page 52)

Which can be posed within the SRM framework: selection among different soft order constraints $\{\mathcal{H}_C\}_{C>0}$ or selecting among different regularization parameters $\{\mathcal{H}_\lambda\}_{\lambda>0}$ where the hypothesis set is fixed at $\mathcal{H}$ and the learning algorithm is augmented error minimization with different regularization parameters $\lambda$?

首先回顾soft order constraints：对于权重$w$，soft order constraints为

$$\sum_{q=0}^{Q} w_i^2 \leq C$$

由包含关系可得，这种情形可以使用SRM框架。

但是对带正则项的$\{\mathcal{H}_\lambda\}_{\lambda>0}$，并没有包含关系，所以无法使用SRM框架。

## Problem 8.17 (Page 53)

Suppose we use "SRM" to select among an arbitrary set of models $\mathcal{H}_1, \ldots, \mathcal{H}_M$ with $d_{\text{vc}}(\mathcal{H}_{m+1}) > d_{\text{vc}}(\mathcal{H}_m)$ (as opposed to a structure in which the additional condition $\mathcal{H}_m \subset \mathcal{H}_{m+1}$ holds). (a) Is it possible for $E_{\text{in}}(\mathcal{H}_m) < E_{\text{in}}(\mathcal{H}_{m+1})$? (b) Let $p_m$ be the probability that the process leads to a function $g_m \in \mathcal{H}_m$,with $\sum_m p_m = 1$. Give a bound for the generalization error in terms of $d_{\text{vc}}(\mathcal{H}_m)$.

(a)可能，因为VC维只是代表最大shatter的点的数量。

(b)注意到我们有

$$\mathbb{P}[|E_{\text{in}}(g_i) - E_{\text{out}}(g_i)| > \epsilon | g^* = g_i] \leq \frac{1}{p_i} 4 m_{\mathcal{H}_i}(2N) e^{-\frac{\epsilon^2 N}{8}}$$

以及

$$m_{\mathcal{H}_i}(2N) \leq (2N)^{d_{\text{vc}}(\mathcal{H}_m)} + 1$$

所以

$$\mathbb{P}[|E_{\text{in}}(g_i) - E_{\text{out}}(g_i)| > \epsilon | g^* = g_i] \leq \frac{1}{p_i} 4\big((2N)^{d_{\text{vc}}(\mathcal{H}_m)} + 1\big) e^{-\frac{\epsilon^2 N}{8}}$$

## Problem 8.18 (Page 53)

Suppose that we can order the hypotheses in a model, $\mathcal{H} = \{h_1, h_2, \ldots\}$. Assume that $d_{\text{vc}}(\mathcal{H})$ is infinite. Define the hypothesis subsets $\mathcal{H}_m = \{h_1, h_2, \ldots, h_m\}$. Suppose you implement a learning algorithm for $\mathcal{H}$ as follows: start with $h_1$; if $E_{\text{in}}(h_1) \leq \nu$, stop and output $h_1$; if not try $h_2$; and so on . . .

(a) Suppose that you output $h_m$, so you have effectively only searched the $m$ hypotheses in $\mathcal{H}_m$. Can you use the VC-bound: (with high probability) $E_{\text{out}}(h_m) \leq \nu + \sqrt{\frac{\ln(2m/\delta)}{2N}}$ ?If yes, why? If no, why not?

(b) Formulate this process within the SRM framework. *[Hint: the $\mathcal{H}_m$'s form a structure.]*

(c) Can you make *any* generalization conclusion (remember, $d_{\text{vc}}(\mathcal{H}) = \infty$)? If yes, what is the bound on the generalization error, and when do you expect good generalization?If no, why?

(a)不行，实际上这里的分析要使用SRM框架。

(b)首先这里的$\mathcal{H}_m$定义了SRM中结构。其次，不难发现，如果我们输出$h_i$，那么必然有

$$E_{\text{in}}(h_j) < E_{\text{in}}(h_i), \forall j < i$$

这说明

$$h_i = \underset{h \in \mathcal{H}_i}{\operatorname{argmin}} E_{\text{in}}(h)$$

接着，我们定义

$$\Omega(\mathcal{H}_i) = i$$

从而

$$h^* = \underset{i=1,2\ldots}{\operatorname{argmin}}( E_{\text{in}}(h_i) + \Omega(\mathcal{H}_i))$$

那么就会优先选择下标较小的$\mathcal{H}_i$对应的$h_i$，符合题意。

(c)利用SRM框架，我们有

$$\mathbb{P}[|E_{\text{in}}(h_i) - E_{\text{out}}(h_i)| > \epsilon | h^* = h_i] \leq \frac{1}{p_i} 4 m_{\mathcal{H}_i}(2N) e^{-\frac{\epsilon^2 N}{8}}$$