山东大学 计算机科学与技术 学院

机器学习与模式识别 课程实验报告

学号：201600301304	姓名：贾乘兴	班级： 人工智能 16
实验题目：支持向量机		
实验学时：2 小时	实验日期： 2018/11/23	
实验目的：在给定的数据集下，实现支持向量机算法，并进行测试		
硬件环境： 16 GB 内存		
软件环境：mac os，matlab 2017b		

实验步骤与内容：

一．支持向量机算法

1. 对于一个二分类任务，我们分类的思路可以是，寻找一个分类的直线、平面或者超平面，使正负两类到分隔面的最小距离最大化，而我们选取的直线可以放置于两个类别的最近的点中间，另这个距离为 m，则

$$m = \frac{\left|w^T x + b\right|}{||w||}$$

对左边归一化，得到

$$\frac{\left|w^T x + b\right|}{m||w||} = 1$$

对系数整理可得

$$\left|w_0^T x + b_0\right| = 1, w_0 = \frac{w}{m||m||}, b_0 = \frac{b}{m||w||}$$

我们将类别分为正类与负类两类，y=1 与 y=-1，可知所有的点满足如下条件

$$y_i \bullet \left(w^T x_i + b\right) = 1, i = 1, ...m$$

而我们的问题是最大化 1/w，将其转为倒数并平方，得到优化的目标如下

$$\min . \frac{1}{2}||w||^2$$

$$s.t. \ \left(w^T x_i + b\right) y_i \geq 1, i = 1, ..., m$$

对于该问题，我们可以由拉格朗日乘数法转化为对偶问题

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \left\langle x_i, x_j \right\rangle$$

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

该方法可以用二次规划的方法求解，也可以通过 SMO 算法进行求解，固定其他数值不

变，每次只优化其中两个。本次实验采用的是二次规划，同时，我们加入软间隔，对该问题训练集的出错加入一定的容忍，加入惩罚因子 C，则该问题可写为

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

$$s.t.\ \ \left(w^T x_i + b\right)y_i \geq 1-\xi_i, i=1,...,m$$

最终得到的对偶问题的二次规划形式如下

$$\max_{\alpha} \quad \frac{1}{2}\alpha^T H\alpha + f^T\alpha$$

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

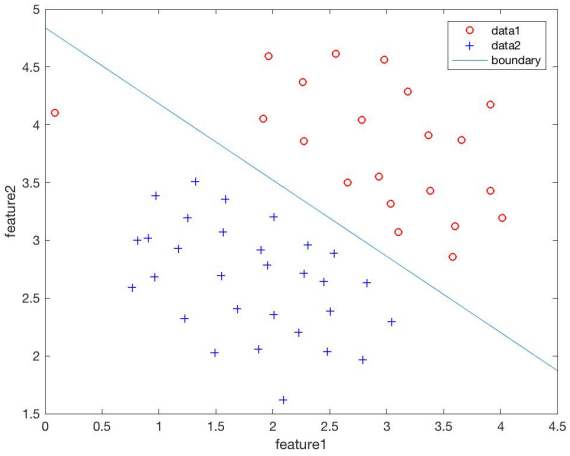
（注意拉格朗日问题求的是最大，而该问题在二次规划下求的是最小

二．实验内容与结果

1. 二维数据的分类

首先导入数据，将数据可视化，然后进行二次规划求解

在设置惩罚因子 c=1 下分类如下



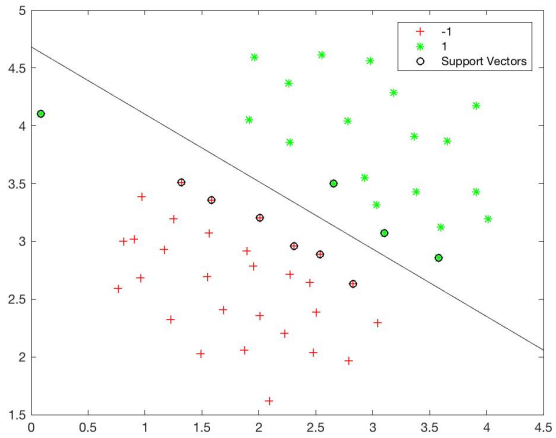
代码如下

```
pos=(y==1);
neg=(y==-1);
figure;
plot(x(pos,1),x(pos,2),'ro');
hold on
plot(x(neg,1),x(neg,2),'b+');
m=length(y);
C=1;
f=-ones(1,m);
H=(y*y').*(x*x');
lb=zeros(1,m);
ub=C*ones(1,m);
Aeq=y';
beq=0;
alpha=quadprog(H,f,[],[],Aeq,beq,lb,ub);
w=sum((alpha.*y).*x);

t=x*w';
max=-1000;
min=1000;
for i=1:m
    if(y(i)==-1&&t(i)>max)
       max=t(i);
    end
end
for i=1:m
    if(y(i)==1&&t(i)<min&&t(i)>max)
       min=t(i);
    end
end
b=-(max+min)/2;
```

```
% w1*x1+w2*x2+b=0
x1=[0:0.1:4.5];
x2=-((w(1)*x1)+b)./w(2);
hold on
plot(x1,x2,'-');
xlabel('feature1');
ylabel('feature2');
legend('data1','data2','boundary');
```

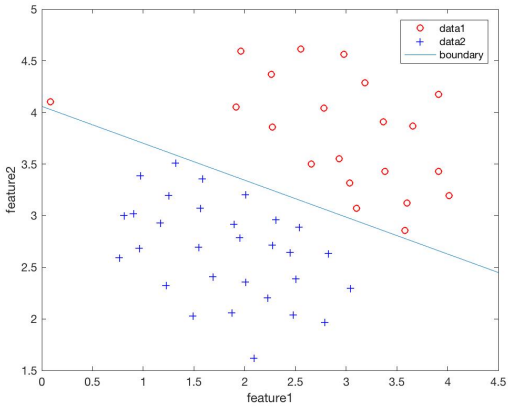
使用 matlab 自带的 svmtrain 函数得到图像如下



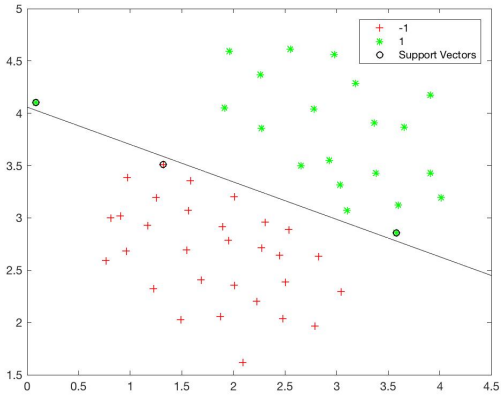
代码如下

```
% solution
figure;
svm=svmtrain(x,y,'boxconstraint',100,'Showplot',true);
```

令 c=100，得到的分类如下



使用 matlab 自带的函数如下



2. 文本分类

对文本进行分类，但需要注意的是，文本的向量长度不确定，这里取了长度为 3000，利用 test 分别对 50、100、400 得到的模型进行测试，并与 matlab 工具箱实现的方法进行对比，实现的 matlab 算法的正确率分别为 75%、88%、98%。Matlab 自带的工具箱的正确率分别为 74%、90%、96%，可见二次规划下实现的 svm 效果较好

代码如下

```
%%
clear,clc;
k=4000;
fidin=fopen('email_train-400.txt');
i=1;
apres=[];

while ~feof(fidin)
    tline = fgetl(fidin);
    apres{i} = tline;
    i=i+1;
end
m0=length(apres);
x=zeros(m0,k);
y=zeros(m0,1);
for i=1:m0
    a=char(apres(i));
    if a(1)=='1'
        a=['+',a];
    end
    l=length(a);
    x0=sscanf(a(4:l),'%d:%f');
    lx=length(x0);
    for j=2:2:lx
        if(x0(j)<=0)
            break
        end
        x(i,x0(j-1))=x0(j);
    end
    y0=sscanf(a(1:2), '%d');
    y(i,1)=y0;
end
svm=svmtrain(x,y);

% plot
m=length(y);
C=1;
f=-ones(1,m);
H=(y*y').*(x*x');
lb=zeros(1,m);
ub=C*ones(1,m);
Aeq=y';
beq=0;

alpha=quadprog(H,f,[],[],Aeq,beq,lb,ub);
w=sum((alpha.*y).*x);

max=-1000;
min=1000;
for i=1:m
    t=x(i,:)*w';
    if (y(i)==1&&t<min)
        min=t;
    elseif (y(i)==-1&&t>max)
        max=t;
    end
end
b=-(max+min)/2;

%
fidin=fopen('email_test.txt');
i=1;
apres=[];
while ~feof(fidin)
    tline = fgetl(fidin);
    apres{i} = tline;
    i=i+1;
end
m0=length(apres);
tx=zeros(m0,k);
ty=zeros(m0,1);
y=zeros(m0,1);
for i=1:m0
    a=char(apres(i));
    if a(1)=='1'
        a=['+',a];
    end
    l=length(a);
    x0=sscanf(a(4:l),'%d:%f');
    lx=length(x0);
    for j=2:2:lx
        if(x0(j)<=0)
            break
        end
        tx(i,x0(j-1))=x0(j);
    end
    y0=sscanf(a(1:2), '%d');
    y(i,1)=y0;
end
ty=tx*w'+b;
```

```
accuracy=sum(((ty>=0)*2-1)==y)/m0

ylabel=svmclassify(svm,tx,'Showplot',true);
accuracy0=sum(((ylabel>=0)*2-1)==y)/m0
```

结论分析与体会： 通过本次实验，对 svm 有了更好的掌握，同时通过对 svm 的实现，代码能力也有一定的提升

附录：程序源代码

ex7.m

```
clear,clc;
fidin=fopen('twofeature.txt');
i=1;
apres=[];

while ~feof(fidin)
    tline = fgetl(fidin);
    apres{i} = tline;
    i=i+1;
end
m0=length(apres);
for i=1:m0
    a=char(apres(i));
    if a(1)=='1'
        a=['+',a];
    end
    l=length(a);
    x0=sscanf(a(4:l),'%d:%f');
    lx=length(x0);
    for j=2:2:lx
        if(x0(j)<=0)
            break
        end
        x(i,x0(j-1))=x0(j);
    end
    y0=sscanf(a(1:2), '%d');
    y(i,1)=y0;
end

% plot
pos=(y==1);
neg=(y==-1);
figure;
plot(x(pos,1),x(pos,2),'ro');
hold on
plot(x(neg,1),x(neg,2),'b+');
m=length(y);
C=1;
f=-ones(1,m);
H=(y*y').*(x*x');
lb=zeros(1,m);
ub=C*ones(1,m);
Aeq=y';
beq=0;
alpha=quadprog(H,f,[],[],Aeq,beq,lb,ub);
w=sum((alpha.*y).*x);
```

```
t=x*w';
max=-1000;
min=1000;
for i=1:m
    if(y(i)==-1&&t(i)>max)
        max=t(i);
    end
end
for i=1:m
    if(y(i)==1&&t(i)<min&&t(i)>max)
        min=t(i);
    end
end
b=-(max+min)/2;

% w1*x1+w2*x2+b=0
x1=[0:0.1:4.5];
x2=-((w(1)*x1)+b)./w(2);
hold on
plot(x1,x2,'-');
xlabel('feature1');
ylabel('feature2');
legend('data1','data2','boundary');

% solution
figure;
svm=svmtrain(x,y,'boxconstraint',100,'Showplot',true);

%%
clear,clc;
k=4000;
fidin=fopen('email_train-400.txt');
i=1;
apres=[];

while ~feof(fidin)
    tline = fgetl(fidin);
    apres{i} = tline;
    i=i+1;
end
m0=length(apres);
x=zeros(m0,k);
y=zeros(m0,1);
for i=1:m0
    a=char(apres(i));
    if a(1)=='1'
        a=['+',a];
    end
    l=length(a);
    x0=sscanf(a(4:l),'%d:%f');
    lx=length(x0);
    for j=2:2:lx
        if(x0(j)<=0)
            break
        end
        x(i,x0(j-1))=x0(j);
    end
    y0=sscanf(a(1:2), '%d');
    y(i,1)=y0;
end
svm=svmtrain(x,y);

% plot
m=length(y);
C=1;
f=-ones(1,m);
H=(y*y').*(x*x');
lb=zeros(1,m);
ub=C*ones(1,m);
Aeq=y';
beq=0;
```

```
alpha=quadprog(H,f,[],[],Aeq,beq,lb,ub);
w=sum((alpha.*y).*x);

max=-1000;
min=1000;
for i=1:m
    t=x(i,:)*w';
    if (y(i)==1&&t<min)
        min=t;
    elseif (y(i)==-1&&t>max)
        max=t;
    end
end
b=-(max+min)/2;

%
fidin=fopen('email_test.txt');
i=1;
apres=[];
while ~feof(fidin)
    tline = fgetl(fidin);
    apres{i} = tline;
    i=i+1;
end
m0=length(apres);
tx=zeros(m0,k);
ty=zeros(m0,1);
y=zeros(m0,1);
for i=1:m0
    a=char(apres(i));
    if a(1)=='1'
        a=['+',a];
    end
    l=length(a);
    x0=sscanf(a(4:l),'%d:%f');
    lx=length(x0);
    for j=2:2:lx
        if(x0(j)<=0)
            break
        end
        tx(i,x0(j-1))=x0(j);
    end
    y0=sscanf(a(1:2), '%d');
    y(i,1)=y0;
end
ty=tx*w'+b;
accuracy=sum(((ty>=0)*2-1)==y)/m0

ylabel=svmclassify(svm,tx,'Showplot',true);
accuracy0=sum(((ylabel>=0)*2-1)==y)/m0
```