Machine Learning Assignment2

Xiaoxu Gao | 4504348

highsmallxu@gmail.com

Question I:

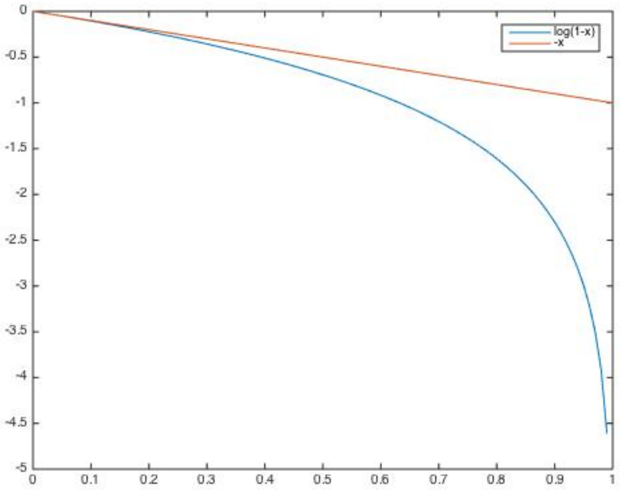
To prove $e^{-x} \geq (1-x)$ in this case, we can use MATLAB and plot function to see the result. The following plot shows the curves of $y_1 = log(-x)$ and $y_2 = -x$. We can see that $log(-x)$ is always above $-x$ from [0,1). Therefore, we can get:

$$log(1-x) + x \geq 0$$

, and then get the final formula:

$$e^{-x} \geq (1-x)$$

, since it is also true when $x = 1$, the final domain for this function is $[0, 1]$.



Question a

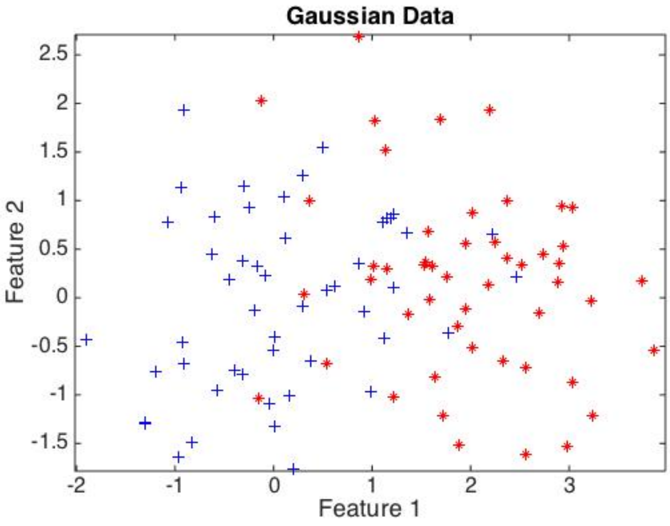
Implement a weak learner: the decision stump.

```
load data.mat
x = X; % 2 features
y = Y; % 200 samples
w = ones(size(x,1), 1); % Give each object a weight w=1
stump = stump(x,y)

--------------------------------------
function [stump]=stump(x,y)
d = size(X,2) % number of features
stump = cell(d,1);
werr  = zeros(d,1);
for i = 1:d
    stump{i} = stump_onedim(x(:,i),y); % go through each feature
    stump{i}.ind = i;
    werr(i)  = stump{i}.werr;
end
[min_werr,ind] = min(werr);
stump = stump{ind(1)};  % return the most optimal stump

---------------------------------------
function [stump] = stump_onedim(x,y)
[sorted,I] = sort(x);
Ir = I(end:-1:1);
score_left = cumsum(y(I));
score_right= cumsum(y(Ir));
score = -score_left(1:end-1) + score_right(end-1:-1:1); % score the boundary
Idec  = find(sorted(1:end-1)<sorted(2:end)); % find distinguishable points
if(length(Idec)>0)
    [maxscore,ind] = max(abs(score(Idec)));
    ind = Idec(ind(1));
    stump.werr = 0.5 - 0.5*maxscore; % weighted error
    stump.x0 = (sorted(ind)+sorted(ind+1))/2; % threshold
    stump.s  = sign(score(ind));
else
    stump.werr = 0.5;
    stump.x0 = 0;
    stump.s  = 1;
end
```

Question b

Dataset is generated by **gendats** from Protools. **Dataset:**
X : each row is a 2-d feature vector
Y : each row is the label (+1/-1), 200 samples in total

```
[a,labels] = gendats;
scatterd(a);
```



Results:

```
werr:0.1500
thres:1.2211
sign(y):-1
ind:1
```

Question c

Dataset: optdigitsubset.mat
X : each row is a 64-d feature vecotr
Y : each row is the label(+1/-1), 50 samples in each class
Training Result - first 50 rows:

```
werr:0.0100
thres = 24
sign(y) = 1
ind = 37
```

Test Result

```
result=[];
for i = 1:1025
    if tt(i,1)>24
        key = 1;
    else
        key = -1;
    end
    result = [result;key];
end
ErrorM  = sum(yt ~= result) / length(yt);

sum(yt ~= result) = 18
ErrorM = 0.0176
```

Training Result - random 50 rows

```
MATLAB Code:
load optdigitsubset.mat
x1 = ex1([1:554],:);
x2 = ex1([555:1125],:);
idx1 = randperm(size(x1,1),50);
idx2 = randperm(size(x2,1),50);

r1 = x1(idx1,:);
r2 = x2(idx2,:);
r  = [r1;r2];
y  = [ones(50,1).*(-1);ones(50,1)];
w = ones(size(r,1), 1);
stump = build_stump(r,y,w);

x1(idx1,:)=[];
x2(idx2,:)=[];
t  = [x1;x2];
thresh = stump.x0;
ind = stump.ind;
tt = t(:,ind);
yt = [ones(504,1).*(-1);ones(521,1)];
result=[];
for i = 1:1025
    if tt(i,1)>thresh
        key = 1;
    else
        key = -1;
    end
    result = [result;key];
end
```

```
sum(yt~=result)
ErrorM  = sum(yt ~= result) / length(yt);

Final Result:
When randomly choosing 50 rows from each class, the result doesn't change a
lot.
ErrorM is from 0.0170~0.0300
```

Question d&e

AdaBoost

```
function [output, thresholds, signs,hypothesis, weights,error] =
adaboostc(a, w, T)
w = ones(size(x,1), 1);
T = 100;
for m=1:M
    p = w ./ sum(w);
    [stump,h] = build_stump(x,y,p);
    error = stump.werr;
    beta = error/(1-error);
    beta_1 = ones(size(w)) * beta;
    w = w.*(beta.^(ones(100,1)-abs(h-y)));
end
L(:,i) = log(1/beta) * h;
R(:,i) = log(1./beta_1) / 2;

lhs = sum(L,2);
rhs = sum(R,2);

h = (lhs >= rhs);
Error = sum(h~=y)/length(y)
```

Question f

Dataset: gendatb

Use 50 samples. 25 of them are training examples, the others are for the test.

Result:

Objects near the boundary have lower weights, while objects far from the boundary have higher weights.

Question g

Dataset: optdigitsubset

Result:

```
ErrorM :0.0107
```

The result doesn't change a lot when we change the number of iterations.

```
It shows that for class 1, the weight of 16th,37th image is quite high. For
class 2, the weight of 29th image is quite hight.
```