**Matlab code for Problem 4**

X\_train = csvread('X\_train.csv');

y\_train = csvread('label\_train.csv');

y\_test = csvread('label\_test.csv');

X\_test = csvread('X\_test.csv');

setNum = size(X\_test,1);

setSize = size(X\_train,2);

N1 = length(find(y\_train));

N0 = length(find(~y\_train));

p\_pre1 = zeros(setNum,1);

sumX1 = sum(X\_train.\*repmat(y\_train,1,setSize),1);

sumX0 = sum(X\_train.\*repmat(1-y\_train,1,setSize),1);

log\_cX = sum((sumX0+1)\*(log(N0+1)-log(N0+2))) - sum((sumX1+1)\*(log(N1+1)-log(N1+2)));

log\_cN = log(N1+2)-log(N0+2);

for k = 1:setNum

log\_factor1 = 0;

log\_factor0 = 0;

for i = 1: 54

if X\_test(k,i) ~= 0

log\_factor1 = log\_factor1 + sum(log(sumX1(i)+1:sumX1(i)+X\_test(k,i)));

log\_factor0 = log\_factor0 + sum(log(sumX0(i)+1:sumX0(i)+X\_test(k,i)));

end

end

log\_fx = (sum(X\_test(k,:))-1)\*log\_cN + log\_factor0 - log\_factor1;

p0\_div\_p1 = exp(log\_fx + log\_cX + log(N0+1)-log(N1+1));

p\_pre1(k) = 1/(1+p0\_div\_p1);

end

y = (p\_pre1 > 0.5);

r = (y == y\_test);

s\_s = length(find(y.\*r));

s\_n = length(find(y.\*~r));

n\_n = length(find(~y.\*r));

n\_s = length(find(~y.\*~r));

cNames = {'classified spam', 'classified non\_spam'};

rNames = {'spam', 'non-spam'};

data = [s\_s n\_s;n\_s n\_n];

classified\_spam = [s\_s;s\_n];

classified\_non\_spam = [n\_s;n\_n];

table(classified\_spam,classified\_non\_spam,'RowName',rNames)

m = find(~r);

figure;

for i = 1:3

p\_pre1(m((i)))

plot(X\_train(m(i),:),'-\*');

hold on;

end

plot((sumX1+1)/(N1+1))

hold on;

plot((sumX0+1)/(N0+1))

legend('sample1','sample2','sample3','E1','E0');

title('misclassified')

[~,I] = sort(abs(p\_pre1-0.5));

figure;

for i = 1:3

p\_pre1(I((i)))

plot(X\_train(I(i),:),'-\*');

hold on;

end

plot((sumX1+1)/(N1+1))

hold on;

plot((sumX0+1)/(N0+1))

legend('sample1','sample2','sample3','E1','E0');

title('three most ambiguous')