3.

- Matlab code

```
%Training
filename_Y = 'DENSE.Y.TRAIN.Y';
filename_X = 'DENSE.TRAIN.X';
Y = importdata(filename_Y);
X = importdata(filename_X);
count = zeros(2,size(X,2));
count(1,:) = sum(X(Y == 1,:));
count(2,:) = sum(X(Y == -1,:));
p_w_y = zeros(size(count,2),size(count,1));
p_w_y(:,1) = (1 + count(1,:))/(size(count,2) + sum(count(1,:)));
p_w_y(:,2) = (1 + count(2,:))/(size(count,2) + sum(count(2,:)));
py = [sum(Y == 1)/size(Y,1), sum(Y == -1)/size(Y,1)];
%Testing
filename_Ytest = 'DENSE.TEST.Y';
filename_Xtest = 'DENSE.TEST.X';
Y_test = importdata(filename_Ytest);
X_test = importdata(filename_Xtest);
Test_result = zeros(size(Y_test));
prob = zeros(size(Y_test,1),2);
prob(:,1) = log10(py(1)) + X_test*log10(p_w_y(:,1));
prob(:,2) = log10(py(2)) + X_test*log10(p_w_y(:,2));
Test_result = 1*(prob(:,1) > prob(:,2));
Test_result(Test_result == 0) = -1;
error = sum(Test_result ~= Y_test)/size(Y_test,1);
token_ratio = log10(p_w_y(:,1))- log10(p_w_y(:,2));
[n,I] = sort(token_ratio,'descend');
token = importdata('TOKENS_LIST');
Indicator = token(I(1:5));
disp(Indicator);
```

(a) Error rate

Size of training set	50	100	200	400	800	1400
Err rate	0.0388	0.0263	0.0263	0.0188	0.0175	00163

(b) Five tokens : 'httpaddr', 'spam', 'unsubscrib', 'ebai', ' valet'.

4.

- Matlab code

```
1   load mnist_data.mat;
2   K = [1 5 9 13];
3   test_sample_index = randsample(size(test,1),100);
4   class_l2 = zeros(100,4);
5   class_l1 = zeros(100,4);
6   for i = 1:100
7       index = test_sample_index(i);
8       diff = train(:,2:end) - repmat(test(index,2:end),size(train,1),1);
9       diff_l2 = sqrt(sum(diff.^2,2)); % L2-norm
10      diff_l1 = sum(abs(diff),2); % L1-norm
11      [B2 I2] = sort(diff_l2);
12      [B1 I1] = sort(diff_l1);
13      for j = 1:size(K,2)
14          class_l2(i,j) = mode(train(I2(1:K(j)),1));
15          class_l1(i,j) = mode(train(I1(1:K(j)),1));
16      end
17  end
18  accuracy_rate_l2 = zeros(1,4);
19  accuracy_rate_l1 = zeros(1,4);
20  for n = 1:4
21      accuracy_rate_l2(n) = sum(class_l2(:,n) == test(test_sample_index,1))/100;
22      accuracy_rate_l1(n) = sum(class_11(:,n) == test(test_sample_index,1))/100;
23  end
```

(c) For K = 1, 5, 9, 13, accuracy rate = 0.95, 0.95, 0.94, 0.94. The K value with best performance are 1,9. However, the result varies while choosing different test samples.

(d) For K = 1, 5, 9, 13, accuracy rate = 0.93, 0.95, 0.94, 0.94. The K value with best performance are 9. However, the result varies when we choose different test samples.

In this problem, I'll choose L2-norm because it always outperforms the L1-norm in terms of accuracy rate.