5. **Bayesian spam filtering** (15 points)

In this problem you will apply the naïve Bayes classifier to the problem of spam detection, using a benchmark database assembled by researchers at Hewlett-Packard. Download the file `spambase.data` from Canvas and issue the following commands to load the data. In Matlab:

Test shows 282 errors (i.e. 10.84% misclassification rate). For sanity check, majority class are 0s (2788 samples). Misclassification rate for majority class is 0.

Code for quantization

```
function x_quant=quant(x,med)
x=x-repmat(med,[size(x,1) 1]);
for i=1:size(x,1)
    for j=1:size(x,2)
        if x(i,j)<0 || x(i,j)==0
            x_quant(i,j)=1;
        else
            x_quant(i,j)=2;
        end
    end
end
end
```

Code for classifier

```
function y_pred=fhat(x_test)

global pihat ghat_01 ghat_02 ghat_11 ghat_12

g1=pihat;
for j=1:numel(x_test)
    if x_test(j)==2
        g1=g1*ghat_12(j);
    else
        g1=g1*ghat_11(j);
    end
end

g0=1-pihat;

for j=1:numel(x_test)
    if x_test(j)==2
        g0=g0*ghat_02(j);
    else
        g0=g0*ghat_01(j);
    end
end
```

```
if g1>g0 || g1==g0
    y_pred=1;
else
    y_pred=0;
end
end
```

Code for determining majority class

```
close all; clear all; clc

z=dlmread('spambase.data',',');
rng(0);
rp=randperm(size(z,1));
z=z(rp,:);
x=z(:,1:end-1);
y=z(:,end);
med=median(x);
x=quant(x,med);

ind_nonzero=find(y);
numel(ind_nonzero)


ind_zero=find(y-1);
numel(ind_zero)
```

Code for generating results

```
close all; clear all; clc

global pihat ghat_01 ghat_02 ghat_11 ghat_12

z=dlmread('spambase.data',',');
rng(0);
rp=randperm(size(z,1));
z=z(rp,:);
x=z(:,1:end-1);
y=z(:,end);
med=median(x);
x=quant(x,med);

x_train=x(1:2000,:);
y_train=y(1:2000);
x_test=x(2001:end,:);
y_test=y(2001:end);

n=size(y_train,1);
```

```
n_1=sum(y_train);
pihat=n_1/n;

for j=1:size(x_train,2)
   sum=0;
   for i=1:size(x_train,1)
      if x_train(i,j)==2 && y_train(i)==1
         sum=sum+1;
      end
   end
   n_12(j)=sum;
   ghat_12(j)=n_12(j)/n_1;
   ghat_11(j)=1-ghat_12(j);
end

for j=1:size(x_train,2)
   sum=0;
   for i=1:size(x_train,1)
      if x_train(i,j)==2 && y_train(i)==0
         sum=sum+1;
      end
   end
   n_02(j)=sum;
   ghat_02(j)=n_02(j)/(n-n_1);
   ghat_01(j)=1-ghat_02(j);
end

for i=1:size(x_test,1)
   y_pred(i)=fhat(x_test(i,:));
end
y_pred=y_pred';
error=abs(y_pred-y_test);
errorSum=0;
for i=1:numel(error)
   errorSum=errorSum+error(i);
end

errorSum
misclas_rate=errorSum/numel(y_test)
```

Code for sanity check

```
close all; clear all; clc
global pihat ghat_01 ghat_02 ghat_11 ghat_12

z=dlmread('spambase.data',',');
rng(0);
```

```
rp=randperm(size(z,1));
z=z(rp,:);
x=z(:,1:end-1);
y=z(:,end);
med=median(x);
x=quant(x,med);

ind_zero=find(y-1);
x_train=x(ind_zero,:);
y_train=y(ind_zero)

x_test=x_train;
y_test=y_train;

n=size(y_train,1);
n_1=sum(y_train);
pihat=n_1/n;

for j=1:size(x_train,2)
   sum=0;
   for i=1:size(x_train,1)
      if x_train(i,j)==2 && y_train(i)==1
         sum=sum+1;
      end
   end
   n_12(j)=sum;
   ghat_12(j)=n_12(j)/n_1;
   ghat_11(j)=1-ghat_12(j);
end

for j=1:size(x_train,2)
   sum=0;
   for i=1:size(x_train,1)
      if x_train(i,j)==2 && y_train(i)==0
         sum=sum+1;
      end
   end
   n_02(j)=sum;
   ghat_02(j)=n_02(j)/(n-n_1);
   ghat_01(j)=1-ghat_02(j);
end

for i=1:size(x_test,1)
   y_pred(i)=fhat(x_test(i,:));
end
y_pred=y_pred';
error=abs(y_pred-y_test);
errorSum=0;
```

```
for i=1:numel(error)
   errorSum=errorSum+error(i);
end

misclas_rate=errorSum/numel(y_test)
```