SMAI - Assignment 3

Question 1

Results

```
mean_e =

    0.1199


var_e =

   1.6543e-04

>> arr'

ans =

    0.1030
    0.1240
    0.1090
    0.1370
    0.1070
    0.1350
    0.1190
    0.1110
    0.1170
    0.1370
```

Therefore, the classification based on Naive Baye's algorithm on the Bank Marketing dataset gave good result with a mean accuracy of around 88%-90% .
The tie cases were handled based on the prior probability of

class, i.e. in case of a tie the prior probabilities were used to decide the class assigned to the input.

%% CODE

```
clc;clear;

load('data');
l = length(data);
len= 1000;

for k=1:10
train = round(rand(len,1)*l);
test = round(rand(len,1)*l);

jobf=zeros(12,2);
maritialf=zeros(4,2);
educationf=zeros(8,2);
defaultf=zeros(3,2);
housingf=zeros(3,2);
loanf=zeros(3,2);

yes= 0;
no=0;
for i=1:len
    temp = data(train(i),:);
    job=cellstr(temp.job);
    maritial=cellstr(temp.maritail);
    educatin=cellstr(temp.educatin);
    default=cellstr(temp.default);
    housing=cellstr(temp.housing);
    loan=cellstr(temp.loan);
    contact=cellstr(temp.contact);
    dec = cellstr(temp.VarName21);
    if(strcmp(dec,'"yes"'))
        yes=yes+1;
    else
        no=no+1;
    end

    if(strcmp(job,'"housemaid"'))
```

```
    if(strcmp(dec,'"yes"'))
        jobf(4,1)=jobf(4,1)+1;
    else
        jobf(4,2)=jobf(4,2)+1;
    end
elseif(strcmp(job,'"admin."'))
    if(strcmp(dec,'"yes"'))
        jobf(1,1)=jobf(1,1)+1;
    else
        jobf(1,2)=jobf(1,2)+1;
    end
elseif(strcmp(job,'"blue-collar"'))
    if(strcmp(dec,'"yes"'))
        jobf(2,1)=jobf(2,1)+1;
    else
        jobf(2,2)=jobf(2,2)+1;
    end
elseif(strcmp(job,'"entrepreneur"'))
    if(strcmp(dec,'"yes"'))
        jobf(3,1)=jobf(3,1)+1;
    else
        jobf(3,2)=jobf(3,2)+1;
    end
elseif(strcmp(job,'"management"'))
    if(strcmp(dec,'"yes"'))
        jobf(5,1)=jobf(5,1)+1;
    else
        jobf(5,2)=jobf(5,2)+1;
    end
elseif(strcmp(job,'"retired"'))
    if(strcmp(dec,'"yes"'))
        jobf(6,1)=jobf(6,1)+1;
    else
        jobf(6,2)=jobf(6,2)+1;
    end
elseif(strcmp(job,'"self-employed"'))
    if(strcmp(dec,'"yes"'))
        jobf(7,1)=jobf(7,1)+1;
    else
```

```
        jobf(7,2)=jobf(7,2)+1;
    end
elseif(strcmp(job,'"services"'))
    if(strcmp(dec,'"yes"'))
        jobf(8,1)=jobf(8,1)+1;
    else
        jobf(8,2)=jobf(8,2)+1;
    end
elseif(strcmp(job,'"student"'))
    if(strcmp(dec,'"yes"'))
        jobf(9,1)=jobf(9,1)+1;
    else
        jobf(9,2)=jobf(9,2)+1;
    end
elseif(strcmp(job,'"technician"'))
    if(strcmp(dec,'"yes"'))
        jobf(10,1)=jobf(10,1)+1;
    else
        jobf(10,2)=jobf(10,2)+1;
    end
elseif(strcmp(job,'"unemployed"'))
    if(strcmp(dec,'"yes"'))
        jobf(11,1)=jobf(11,1)+1;
    else
        jobf(11,2)=jobf(11,2)+1;
    end
elseif(strcmp(job,'"unknown"'))
    if(strcmp(dec,'"yes"'))
        jobf(12,1)=jobf(12,1)+1;
    else
        jobf(12,2)=jobf(12,2)+1;
    end
end
%%%%%%% maritial

if(strcmp(maritial,'"divorced"'))
    if(strcmp(dec,'"yes"'))
        maritialf(1,1)=maritialf(1,1)+1;
    else
```

```
        maritialf(1,2)=maritialf(1,2)+1;
    end
elseif(strcmp(maritial,'"married"'))
    if(strcmp(dec,'"yes"'))
        maritialf(2,1)=maritialf(2,1)+1;
    else
        maritialf(2,2)=maritialf(2,2)+1;
    end
 elseif(strcmp(maritial,'"single"'))
    if(strcmp(dec,'"yes"'))
        maritialf(3,1)=maritialf(3,1)+1;
    else
        maritialf(3,2)=maritialf(3,2)+1;
    end
   elseif(strcmp(maritial,'"unknown"'))
    if(strcmp(dec,'"yes"'))
        maritialf(4,1)=maritialf(4,1)+1;
    else
        maritialf(4,2)=maritialf(4,2)+1;
    end
end
%%%%%%% education
if(strcmp(educatin,'"basic.4y"'))
    if(strcmp(dec,'"yes"'))
        educationf(1,1)=educationf(1,1)+1;
    else
        educationf(1,2)=educationf(1,2)+1;
    end
elseif(strcmp(educatin,'"basic.6y"'))
    if(strcmp(dec,'"yes"'))
        educationf(2,1)=educationf(2,1)+1;
    else
        educationf(2,2)=educationf(2,2)+1;
    end
elseif(strcmp(educatin,'"basic.9y"'))
    if(strcmp(dec,'"yes"'))
        educationf(3,1)=educationf(3,1)+1;
    else
        educationf(3,2)=educationf(3,2)+1;
```

```
    end
  elseif(strcmp(educatin,'"high.school"'))
    if(strcmp(dec,'"yes"'))
        educationf(4,1)=educationf(4,1)+1;
    else
        educationf(4,2)=educationf(4,2)+1;
    end
  elseif(strcmp(educatin,'"illiterate"'))
    if(strcmp(dec,'"yes"'))
        educationf(5,1)=educationf(5,1)+1;
    else
        educationf(5,2)=educationf(5,2)+1;
    end
  elseif(strcmp(educatin,'"professional.course"'))
    if(strcmp(dec,'"yes"'))
        educationf(6,1)=educationf(6,1)+1;
    else
        educationf(6,2)=educationf(6,2)+1;
    end
   elseif(strcmp(educatin,'"university.degree"'))
    if(strcmp(dec,'"yes"'))
        educationf(7,1)=educationf(7,1)+1;
    else
        educationf(7,2)=educationf(7,2)+1;
    end
 elseif(strcmp(educatin,'"unknown"'))
    if(strcmp(dec,'"yes"'))
        educationf(8,1)=educationf(8,1)+1;
    else
        educationf(8,2)=educationf(8,2)+1;
    end
end
%%%%%%%%     default

if(strcmp(default,'"yes"'))
    if(strcmp(dec,'"yes"'))
        defaultf(1,1)=defaultf(1,1)+1;
    else
        defaultf(1,2)=defaultf(1,2)+1;
```

```
    end
elseif(strcmp(default,'"no"'))
    if(strcmp(dec,'"yes"'))
        defaultf(2,1)=defaultf(2,1)+1;
    else
        defaultf(2,2)=defaultf(2,2)+1;
    end
elseif(strcmp(default,'"unknown"'))
    if(strcmp(dec,'"yes"'))
        defaultf(3,1)=defaultf(3,1)+1;
    else
        defaultf(3,2)=defaultf(3,2)+1;
    end
end

%%%%%%% housing
 if(strcmp(housing,'"yes"'))
    if(strcmp(dec,'"yes"'))
        housingf(1,1)=housingf(1,1)+1;
    else
        housingf(1,2)=housingf(1,2)+1;
    end
 elseif(strcmp(housing,'"no"'))
    if(strcmp(dec,'"yes"'))
        housingf(2,1)=housingf(2,1)+1;
    else
        housingf(2,2)=housingf(2,2)+1;
    end
 elseif(strcmp(housing,'"unknown"'))
    if(strcmp(dec,'"yes"'))
        housingf(3,1)=housingf(3,1)+1;
    else
        housingf(3,2)=housingf(3,2)+1;
    end
 end

 %%%%%%%% loan
  if(strcmp(loan,'"yes"'))
    if(strcmp(dec,'"yes"'))
```

```
            loanf(1,1)=loanf(1,1)+1;
        else
            loanf(1,2)=loanf(1,2)+1;
        end
    elseif(strcmp(loan,'"no"'))
        if(strcmp(dec,'"yes"'))
            loanf(2,1)=loanf(2,1)+1;
        else
            loanf(2,2)=loanf(2,2)+1;
        end
    elseif(strcmp(loan,'"unknown"'))
        if(strcmp(dec,'"yes"'))
            loanf(3,1)=loanf(3,1)+1;
        else
            loanf(3,2)=loanf(3,2)+1;
        end
     end
     %%%%%


end


     jobf(:,1)=jobf(:,1)./sum(jobf(:,1));
     jobf(:,2)=jobf(:,2)./sum(jobf(:,2));

      maritialf(:,1)= maritialf(:,1)/sum( maritialf(:,1));
      maritialf(:,2)= maritialf(:,2)/sum( maritialf(:,2));

     educationf(:,1)=educationf(:,1)/sum(educationf(:,1));
     educationf(:,2)=educationf(:,2)/sum(educationf(:,2));

     defaultf(:,1)=defaultf(:,1)/sum(defaultf(:,1));
     defaultf(:,2)=defaultf(:,2)/sum(defaultf(:,2));

     housingf(:,1)=housingf(:,1)/sum(housingf(:,1));
     housingf(:,2)=housingf(:,2)/sum(housingf(:,2));

     loanf(:,1)=loanf(:,1)/sum(loanf(:,1));
```

```
      loanf(:,2)=loanf(:,2)/sum(loanf(:,2));
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%% test



error=0;
 for i=1:len
    temp = data(test(i),:);
    job=cellstr(temp.job);
    maritial=cellstr(temp.maritail);
    educatin=cellstr(temp.educatin);
    default=cellstr(temp.default);
    housing=cellstr(temp.housing);
    loan=cellstr(temp.loan);
    contact=cellstr(temp.contact);
    dec = cellstr(temp.VarName21);

    yes1 =1;
    no1=1;
    if(strcmp(job,'"housemaid"'))
        yes1=jobf(4,1)*yes1;
        no1=jobf(4,2)*no1;
    elseif(strcmp(job,'"admin."'))
        yes1=jobf(1,1)*yes1;
        no1=jobf(1,2)*no1;
    elseif(strcmp(job,'"blue-collar"'))
        yes1=jobf(2,1)*yes1;
        no1=jobf(2,2)*no1;
    elseif(strcmp(job,'"entrepreneur"'))
         yes1=jobf(3,1)*yes1;
        no1=jobf(3,2)*no1;
    elseif(strcmp(job,'"management"'))
        yes1=jobf(5,1)*yes1;
        no1=jobf(5,2)*no1;
    elseif(strcmp(job,'"retired"'))
         yes1=jobf(6,1)*yes1;
        no1=jobf(6,2)*no1;
    elseif(strcmp(job,'"self-employed"'))
```

```
   yes1=jobf(7,1)*yes1;
    no1=jobf(7,2)*no1;
elseif(strcmp(job,'"services"'))
    yes1=jobf(8,1)*yes1;
    no1=jobf(8,2)*no1;
elseif(strcmp(job,'"student"'))
    yes1=jobf(9,1)*yes1;
    no1=jobf(9,2)*no1;
elseif(strcmp(job,'"technician"'))
    yes1=jobf(10,1)*yes1;
    no1=jobf(10,2)*no1;
elseif(strcmp(job,'"unemployed"'))
    yes1=jobf(11,1)*yes1;
    no1=jobf(11,2)*no1;
elseif(strcmp(job,'"unknown"'))
   yes1=jobf(12,1)*yes1;
    no1=jobf(12,2)*no1;
end
%%%%%%% maritial

if(strcmp(maritial,'"divorced"'))
     yes1=maritialf(1,1)*yes1;
    no1=maritialf(1,2)*no1;
elseif(strcmp(maritial,'"married"'))
    yes1=maritialf(2,1)*yes1;
    no1=maritialf(2,2)*no1;
 elseif(strcmp(maritial,'"single"'))
   yes1=maritialf(3,1)*yes1;
    no1=maritialf(3,2)*no1;
   elseif(strcmp(maritial,'"unknown"'))
    yes1=maritialf(4,1)*yes1;
    no1=maritialf(4,2)*no1;
end
%%%%%%% education
if(strcmp(educatin,'"basic.4y"'))
   yes1=educationf(1,1)*yes1;
    no1=educationf(1,2)*no1;
elseif(strcmp(educatin,'"basic.6y"'))
     yes1=educationf(2,1)*yes1;
```

```
    no1=educationf(2,2)*no1;
 elseif(strcmp(educatin,'"basic.9y"'))
     yes1=educationf(3,1)*yes1;
    no1=educationf(3,2)*no1;
  elseif(strcmp(educatin,'"high.school"'))
    yes1=educationf(4,1)*yes1;
    no1=educationf(4,2)*no1;
  elseif(strcmp(educatin,'"illiterate"'))
     yes1=educationf(5,1)*yes1;
    no1=educationf(5,2)*no1;
  elseif(strcmp(educatin,'"professional.course"'))
     yes1=educationf(6,1)*yes1;
    no1=educationf(6,2)*no1;
   elseif(strcmp(educatin,'"university.degree"'))
     yes1=educationf(7,1)*yes1;
    no1=educationf(7,2)*no1;
 elseif(strcmp(educatin,'"unknown"'))
     yes1=educationf(8,1)*yes1;
    no1=educationf(8,2)*no1;
 end
%%%%%%%%     default

 if(strcmp(default,'"yes"'))
    yes1=defaultf(1,1)*yes1;
    no1=defaultf(1,2)*no1;
 elseif(strcmp(default,'"no"'))
    yes1=defaultf(2,1)*yes1;
    no1=defaultf(2,2)*no1;
 elseif(strcmp(default,'"unknown"'))
    yes1=defaultf(3,1)*yes1;
    no1=defaultf(3,2)*no1;
 end

 %%%%%%% housing
  if(strcmp(housing,'"yes"'))
    yes1=housingf(1,1)*yes1;
    no1=housingf(1,2)*no1;
  elseif(strcmp(housing,'"no"'))
    yes1=housingf(2,1)*yes1;
```

```
        no1=housingf(2,2)*no1;
    elseif(strcmp(housing,'"unknown"'))
        yes1=housingf(3,1)*yes1;
        no1=housingf(3,2)*no1;
    end

    %%%%%%%% loan
     if(strcmp(loan,'"yes"'))
        yes1=loanf(1,1)*yes1;
        no1=loanf(1,2)*no1;
    elseif(strcmp(loan,'"no"'))
      yes1=loanf(2,1)*yes1;
        no1=loanf(2,2)*no1;
    elseif(strcmp(loan,'"unknown"'))
        yes1=loanf(3,1)*yes1;
        no1=loanf(3,2)*no1;
     end

     yes1=yes1*(yes/len);
     no1=no1*(no/len);
        if(yes1 >= no1)
            if(strcmp(dec,'"yes"'))
            else
                error=error+1;
            end
        else
            if(strcmp(dec,'"yes"'))
                error=error+1;
            else
            end
        end

 end
 arr(k)=error/len;

end

mean_e=mean(arr)
var_e = var(arr)
```

%%%%%%%%%%%%%%%%%%%%%%%%%% END

Question 2

The decision rule to assign class in the Bayesian system is : choose ω_i if $P(\omega_i|x)$ if $P(\omega_i|x) > P(\omega_j|x)$
Therefore, applying Baye's formula on it we get:

$$P(\omega_i|\, x,\, D) \;=\; \frac{p(\,x \mid \omega_i,\, D)\, P(\omega_i \mid D)}{\sum_{j=1}^{c} p(\,x \mid \omega_j,\, D)\, P(\omega_j \mid D)}$$

We assume that the true values of the a priori probabilities are known and substitute

$$P(\omega_i \mid D) \;= P(\omega_i)$$

Also we assume that the desired p(x) has a parametric form θ and the function $p(x|\theta)$ is known. Therefore, we are changing our supervised learning problem into an unsupervised density estimation problem.
We use a set D of samples drawn independently according to the fixed but unknown probability distribution p(x) to determine p(x|D).
We do this by integrating the joint density p(x, θ|D) over θ :

$$p(x|D) \;=\; \int p(x, \theta|D)\, d\theta$$

$$p(x|D) \;=\; \int p(x|\theta) p(\theta|D)\, d\theta$$

Now if $p(\theta|D)$ peaks very sharply(~1) at some $\widehat{\theta}$ then $p(x|\theta) \sim 0$ for all θ *except* $\widehat{\theta}$ such that $p(x|D) \simeq p(x|\widehat{\theta})$ which can be modelled as Gaussian density.

<u>UNIVARIATE CASE</u>

$$p(x|\mu) \sim N(\mu, \sigma^2)$$
$$p(\mu) \sim N(\mu_0, \sigma_0{}^2)$$

For the univariate case only μ is unknown, therefore:

$$p(\mu|D) = \frac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)\, d\mu}$$

$$= \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu)$$

Now both these can be modelled as gaussian densities themselves- $p(x_k|\mu) \sim N(\mu, \sigma^2)$ and $p(\mu) \sim N(\mu_0, {\sigma_0}^2)$ to get:

$$p(\mu|D) = \alpha \prod_{k=1}^{n} \frac{1}{\sqrt{2\Pi\sigma^2}} exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2] \frac{1}{\sqrt{2\Pi{\sigma_0}^2}} exp[-\frac{1}{2}(\frac{\mu-\mu_0}{\sigma_0})^2]$$

If we write $p(\mu|D) \sim N(\mu_n, \sigma_n^2)$ then μ_n, σ_n can be found by equating it with the above equation, this yields:

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2} \quad , \text{ and}$$

$$\mu_n = (\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2})\overline{x}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

$$p(x|D) = \frac{1}{2\pi\sigma\,\sigma_n} exp[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}]\, f(\sigma, \sigma_n)$$

MULTIVARIATE CASE

Following similar steps we get :

$$p(\mu|D) = \alpha\, exp[-\frac{1}{2}(\mu-\mu_n)^t \Sigma_n^{-1}(\mu-\mu_n)$$

On equating it with gaussian multivariate density, we get:

$$\hat{\mu}_n = \frac{1}{n}\sum_{k=1}^{n} x_k$$

$$\mu_n = \Sigma_0(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\hat{\mu}_n + \frac{1}{n}\Sigma(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\mu_0$$

$$\Sigma_n = \Sigma_0(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\frac{1}{n}\Sigma$$

On performing the integral for $p(x|D)$, *we get* :

$$p(x|D) \sim N(\mu_n, \Sigma + \Sigma_n)$$

Question 3

Just the PCA is implemented and classification is done using the gaussian naive Bayes Classifier.

Psedo-code.

1)Preprocessing-data : make mean of data zero and divide it by variance.

2)Finding the covariance matrix of the data and finding the eigen values and vectors.

3)Now selecting the the largest 10,100,1000 eigenvalues to transform the data points onto new features space.

4)Once the data is reduced to lower dimension, then the gaussian bayesian classifier is used to classify.

```
%%%%%% CODE %%%%%%
clc;clear;
load('arcene_train.data');

k=1000;
m = mean(arcene_train);
vari = var(arcene_train);
[b,l]=size(arcene_train);

for i=1:b
    arcene_train_m(i,:)=(arcene_train(i,:)-m);
end

% Cov = cov(arcene_train); %% computed and saved..
% [V,D] = eigs(Cov);
load ('eigen')
v = zeros(length(V),k);

for i=1:k
    v(:,i)=V(:,i);
end
```

```
ai = zeros(b,k);
for i=1:b
    ai(i,:)= arcene_train_m(i,:)*v;
end

%%% ai has n x d' here d' is 5

%%% gaussian bayesian....


load('labels');
one = find(labels==1);
two = find(labels==-1);

aio = ai(one,:);
ait = ai(two,:);

one_p = size(aio,1)/(size(labels,1)); %%%  prior
two_p = size(ait,1)/(size(labels,1));


mean_aio = mean(aio);%%% p(x/1)
var_aio = var(aio);

mean_ait = mean(ait);%%%% p(x/2)
var_ait = var(ait);
%%%%%%%%%%%%%%%%%%   test  %%%%%%%%%%%%%%%%%%%%
load('arcene_valid.data')
load('avl')
arcene_test = arcene_valid;
k=1000;
m = mean(arcene_test);
vari = var(arcene_test);
error = 0;



[b,l]=size(arcene_test);

load('eigen')
v = zeros(length(V),k);
for i=1:b
    arcene_test(i,:)=(arcene_test(i,:));
```

```
end


for i=1:k
    v(:,i)=V(:,i);
end

aitr = zeros(b,k);

for i=1:b
    aitr(i,:)= arcene_test(i,:)*v;
    tempo = -1*sum( ((aitr(i,:)-mean_aio)./(var_aio)).^2 );
    tempo = tempo+log(one_p)-sum(log(var_aio));


    tempt = -1*sum( ((aitr(i,:)-mean_ait)./(var_ait)).^2 );
    tempt = tempt+log(two_p)-sum(log(var_ait));
    if(tempo > tempt)
        if(labels(i)==1)
        else
            error=error+1;
        end
    else
         if(labels(i)==-1)
        else
            error=error+1;
        end
    end
end
%%%%%%%%%%%%%%%%%%  END %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Results : Error is quite high 44 %
The error should decrease as the PCA (K) increases but a increase in K would increase the complexity.