Machine learning – Assignment#2
Aditya Gautam(agautam1@andrew.cmu.edu)

Assignment #2 (Machine learning 601-B)

-ADITYA GAUTAM
(agautam1@andrew.cmu.edu)

Problem1) Independent events and Bayes theorem

a) let A & B be two events.
probability that event B occurs given A has already occured = P(A/B)
$P(A \cap B) = P(A)\, P(B/A)$ [probability of occurrence of A given B]
Similarly,
given event A has occured, probability of occurance of event A
would be $P(B)\, P(A/B) = P(A \cap B)$
From, above two points, we can say that
$$P(A \cap B) = P(A)\, P\left(\frac{B}{A}\right) = P(B)\, P\left(\frac{A}{B}\right)$$
So,
$$P(A/B)\, P(B) = P(A)\, P(B/A)$$
$$\boxed{P(A/B) = \frac{P(A)\, P(B/A)}{P(B)}}$$

b) Complete Sample Space $\Omega \rightarrow A_1, A_2, A_3 \ldots A_n$
Probability that event B occurs given A_1 is the space
$$= P(A_1 \text{ occurs})\, P\left(\frac{B \text{ occurs given } A_1}{\text{ett}}\right)$$
$$= P(A_1)\, P(B/A_1)$$
Likewise for any A_i $\quad P(B \cap A_i) = P(A_i)\, P(B/A_i)$
$$P(B) = P(B \cap \underset{\text{complete space}}{S}) = P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_i) \quad \left[\text{as } \bigcup_{i=1}^{n} A_i = \Omega(S)\right]$$
Applying ① in ②, we get
$$P(B) = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + \ldots + P(A_n)P(B/A_n)$$
So,
$$\boxed{P(B) = \sum_{i=1}^{n} P(A_i)\, P(B/A_i)}$$

c) $P(A_i)>0$ & $P(B)>0$.

Using the Results from part a) & b), we can say the following about A_i and B,

$$P(B)P(A_i|B) = P(B/A_i)\,P(A_i) \quad \text{--- (1)}$$

i.e probability that B has occured given A_i is same as that of probability of B given A_i has occured.

Also, from part b),

$$P(B) = \sum_i P(B/A_i)P(A_i) \quad \text{where } \bigcup_i A_i = \Omega \quad \text{--- (2)}$$

So, we can say that

$$P(A_i/B) = \frac{P(B/A_i)\,P(A_i)}{P(B)} \quad \text{— from (1)}$$

$$\Rightarrow \boxed{P(A_i/B) = \frac{P(B/A_i)\,P(A_i)}{\sum_k P(B/A_k)P(A_k)}} \quad \text{— applying (2)}$$

d) a) $P(A,B,C) = P(A\,|\,B,C)\,P(B|C)\,P(C)$

$\boxed{\text{True}}$ by directly applying chain rule / Bayes theorm.

b) $P(A,B) = P(A/B)\,P(B/A)$

$\boxed{\text{False}}$

Applying chain rule $P(A,B) = P(A/B)\,P(B)$

and $P(B) = P(B/A)$ only when A & B are independent events

c) $P(A,B,C) = P(B|A,C)\,P(C,A)$

$\boxed{\text{True}}$

Applying chain rule,

$$P(A,B,C) = P(B/A,C)\,P(A,C)$$

$$P(A,C) = P(C,A) = P(A\cap C) \rightarrow \text{proved in part (a)}$$

4) $P(A,B,C) = P(B/A,C)\, P(C,A)\, P(C) \rightarrow \boxed{\text{False}}$

Applying chain Rule of L.H.S

$P(A,B,C) = P(B/_{A,C})\, P(A,C) = P(B/_{A,C})\, P(C,A)$ → proved in part (a)

So, R.H.S has additional term $P(C)$.

5) $P(A,B) = P(A)P(B) \rightarrow \boxed{\text{False}}$

Chain Rule & Bayes Theorem ⇒ $P(A,B) = P(A)P(B/A) = P(B)P(A/B)$

This will be true only when A & B are independent event.

c) $X = \begin{cases} -1 & \text{if A occurs} \\ 0 & \text{otherwise} \end{cases}$

$P(A)$ → probability of A occurs.

$E[X] = P(A) \times -1 + (1-P(A)) \times 0$

$\Rightarrow E[X] = -P(A)$

$\Rightarrow \boxed{E[X] + P(A) = 0}$

when $X=1$, if event A occurs.

so $X = \begin{cases} 1 & \text{if A occurs} \\ 0 & \text{otherwise} \end{cases}$

$E[X] = P(A) \times 1 + (1-P(A)) \times 0 = P(A)$

$\Rightarrow \boxed{E[X] = P(A)}$ so, old result will not hold.

Problem 2: Maximum Likelihood Estimation

a) $P(X_i = k) = (1-\theta)^k \theta$

$L(\hat{\theta}) = P(x_1, \ldots x_n / \hat{\theta})$ x_i are i.i.d

as x_i are i.i.d $\Rightarrow P(x_1, x_2/\theta) = P(x_1/\theta) P(x_2/\theta)$

$\Rightarrow L(\hat{\theta}) = P(\frac{x_1}{\theta}) P(x_2/\theta) P(x_3/\theta) \ldots P(x_n/\theta)$

$L(\hat{\theta}) = \prod_{i=1}^{N} P(\frac{x_i}{\theta})$

$$l(\hat{\theta}) = \log L(\hat{\theta})$$

$$= \log \prod_{i=1}^{N} P\left(\frac{x_i}{\theta}\right) = \sum_{i=1}^{N} \log P(x_i/\theta)$$

$$= \log\left(P(x_1/\theta)\right) + \log\left(P(x_2/\theta)\right) + \log\left(P\left(\frac{x_3}{\theta}\right)\right) + \dots + \log P(x_i/\theta)$$

$$= \log((1-\theta)^{x_1}\theta) + \log((1-\theta)^{x_2}\theta) \dots + \log(1-\theta)^{x_n}\theta$$

$$= [x_1 \log(1-\theta) + \log\theta] + [x_2 \log(1-\theta) + \log\theta] + \dots$$

$$\dots + [x_n \log(1-\theta) + \log\theta]$$

$$= x_1 \log(1-\theta) + x_2 \log(1-\theta) + \dots + x_n \log(1-\theta)$$

$$+ \log\theta + \log\theta \dots + \log\theta$$

$$= \log(1-\theta)[x_1 + x_2 \dots + x_n] + n\log\theta$$

$$\boxed{l(\hat{\theta}) = \log(1-\theta)\sum_{i=1}^{N} x_i + N\log\theta}$$

This doesn't depend on the order of Random variable.

b) $x = (1, 0, 3, 5, 18, 14, 15, 7, 13, 9, 0, 17, 4, 24, 3)$

code, plots and figure submitted to autolab.

code → logMLE.m (file name).

log maximumlikelihood is maximum at $\hat{\theta}$ .80

c) fet say that

$$\frac{d\,l(\hat{\theta})}{\hat{\theta}} = 0$$

$$\Rightarrow \frac{d\left(\sum x_i \log(1-\hat{\theta}) + N\log\hat{\theta}\right)}{d\hat{\theta}} = 0$$

$$\Rightarrow \frac{\sum x_i}{(1-\theta)}(-1) + \frac{N}{\theta} = 0$$

$$\Rightarrow \frac{N}{\theta} = \frac{\sum x_i}{1-\theta} \Rightarrow N - N\theta = (\sum x_i)\theta$$

$$N = \theta(N + \sum x_i)$$

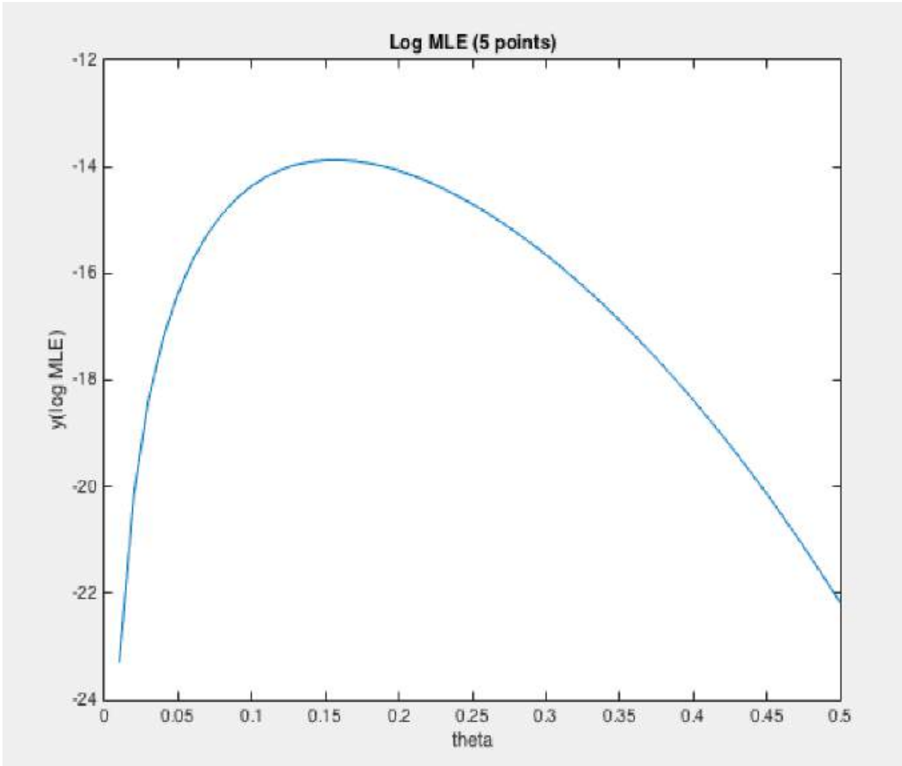
$$\boxed{\theta = \frac{N}{N + \sum x_i}}$$

Ques 2) Part b):

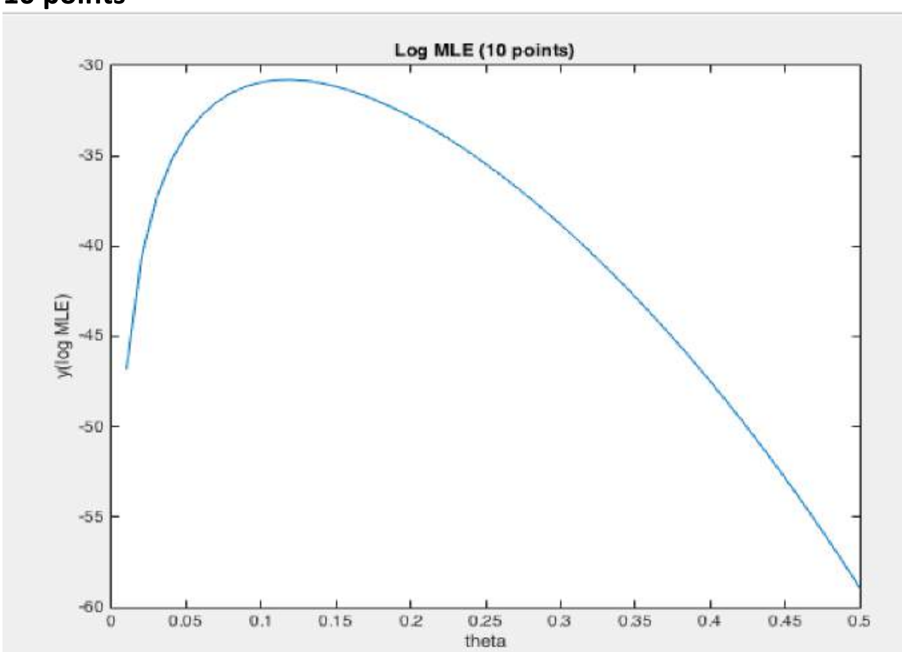
Matlab Code

```
function [ ] = logMLE()

theta = 0.01:.01:0.5;

x = [1 0 3 5 18 14 5 7 13 9 0 17 4 24 3];

y5= sum(x(1:5))*log(1-theta) + 5*log(theta);
y10 =  sum(x(1:10))*log(1-theta) + 10*log(theta);
y15 = sum(x(1:15))*log(1-theta) + 15*log(theta);

figure
plot(theta,y5)
xlabel('theta');
ylabel('y(log MLE)');
title('Log MLE (5 points)');

figure
plot(theta,y10)
xlabel('theta');
ylabel('y(log MLE)');
title('Log MLE (10 points)');

figure
plot(theta,y15)
xlabel('theta');
ylabel('y(log MLE)');
title('Log MLE (15 points)');

end
```

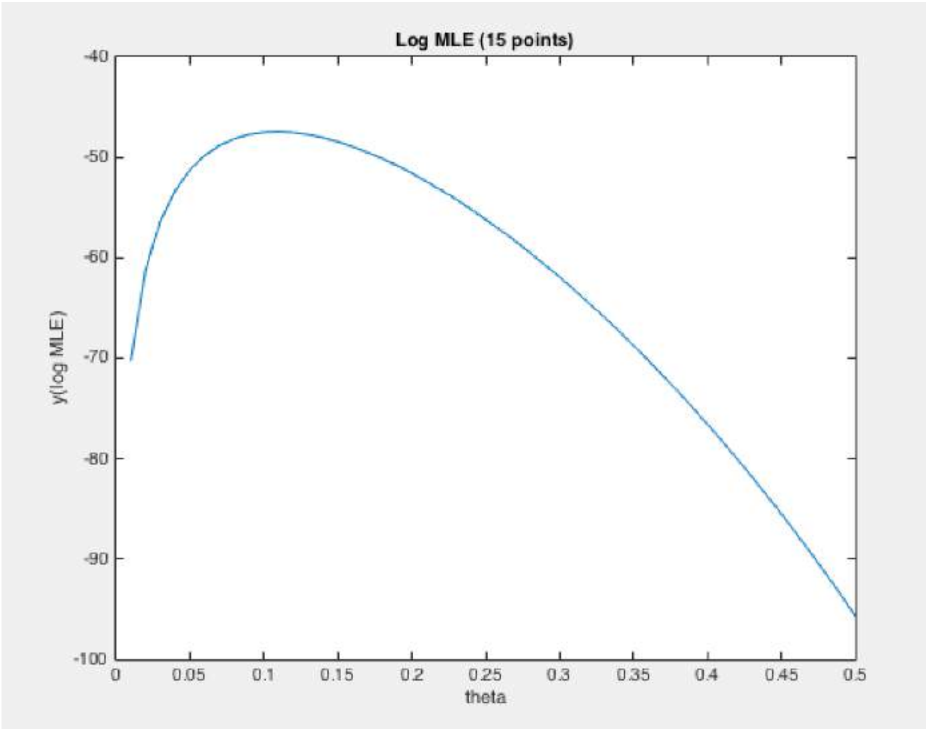
Graphs : (2b)
5 Points



10 points



15 points


Log MLE (15 points)
-40
-50
-60
-70
-80
-90
-100
y(log MLE)
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
theta

Yes, the closed form expression matches with the plots as the maximum of plot $(\hat{\theta})$ is same as one derived from the previous equation.

Like for $N=5$ $\sum x_i = 27$.

$$\hat{\theta} = \frac{5}{5+27} = \frac{5}{32} = \boxed{0.15}$$

We are getting maximum log likelihood around the same point in the matlab plot so it is consistent.

Same is true with other two plots having 10 and 15 points.

for 10 points

$$\hat{\theta} = \frac{10}{10+75} = \frac{10}{85} = \boxed{0.11}$$

for 15 points

$$\hat{\theta} = \frac{15}{15+123} = \frac{15}{138} = \boxed{0.10}$$

All above values are consistent with matlab plots submitted.

d) $l(\hat{\theta}) = \log(1-\theta) \sum_{i=1}^{N} x_i + N \log \theta$.

as $\theta < 1 \Rightarrow \log \theta$ & $\log(1-\theta)$ both will be negative values as no. of sample increases, N increases or $N \log \theta$ will becomes more negative.

also $\sum_{i=1}^{N} x_i < \sum_{i=1}^{N+1} x_i$, and $\log(1-\theta)$ is negative as $0<(1-\theta)<1$ & base of log is greater than zero.

So, if number of sample increases, log likelihood function became more negative.

Problem 3) Implementing Naive Bayes

a) $\hat{Y} = \underset{y}{\text{argmax}}\ P(Y=y|X)$

If x_i's are conditionally independent give a class Y.

$\Rightarrow P(X_1, X_2/Y) = P(X_1/Y)\,P(X_2/Y)$

likewise for N features.

$P(X_1, X_2, X_3, \ldots X_n/Y=y) = P(X_1/y)\,P(X_2/y)\,P(X_3/y) \ldots P(X_n/y)$ — ①

$\Rightarrow P(X_1, X_2 \ldots X_n/Y) = \prod_{i=1}^{n} P(X_i/Y)$ — ②

$P(Y=y/X)\,P(X) = P(X/Y)\,P(Y)$

$\Rightarrow P(Y=y/X) = \dfrac{P(X_n/Y)\,P(Y=y)}{P(X)}$ → normalization.

$\Rightarrow \hat{Y} = \text{argmax}\ P(X/Y)\,P(Y=y)$

$\hat{Y} = \text{argmax}\ P(X_1, X_2 \ldots X_n/Y)\,P(Y=y)$

$\boxed{\hat{Y} = \prod_{i=1}^{n} P(X_i/Y=y)\,P(Y=y)}$ — from ②.

b) No. of parameter with naive Bayes assumption of conditionally independence would be $\boxed{2n+1}$, where n is the number of features.

If we don't assumption Naive Bayes assumption, then total number of parameters would by $\boxed{(2^n - 1)\times 2}$ as we would have 2 possibilities for every x_i and Y i.e 0 or 1, thus $(2^n - 1)\times 2$.

Gain is in the case of Naive Bayes i.e conditional independence.

c), d), e) → functions submitted on autolab.

f) Training error = ·0034483 (2 mismatch out of 580)
Test error = 0·027586 (2 Mismatch out of 145)

On a new collection of set, we should expect test Error as the best representative.
Since Naive Bayes is based on the training data, it will give comparatively good results for training data set when tested as ~~test~~ data matches with training data.
However, when we try different type of data (test data) we can expect ~~high~~ error rate since the distribution of test might be different from training data.
So, Naive Bayes attempt to minimise the training error.

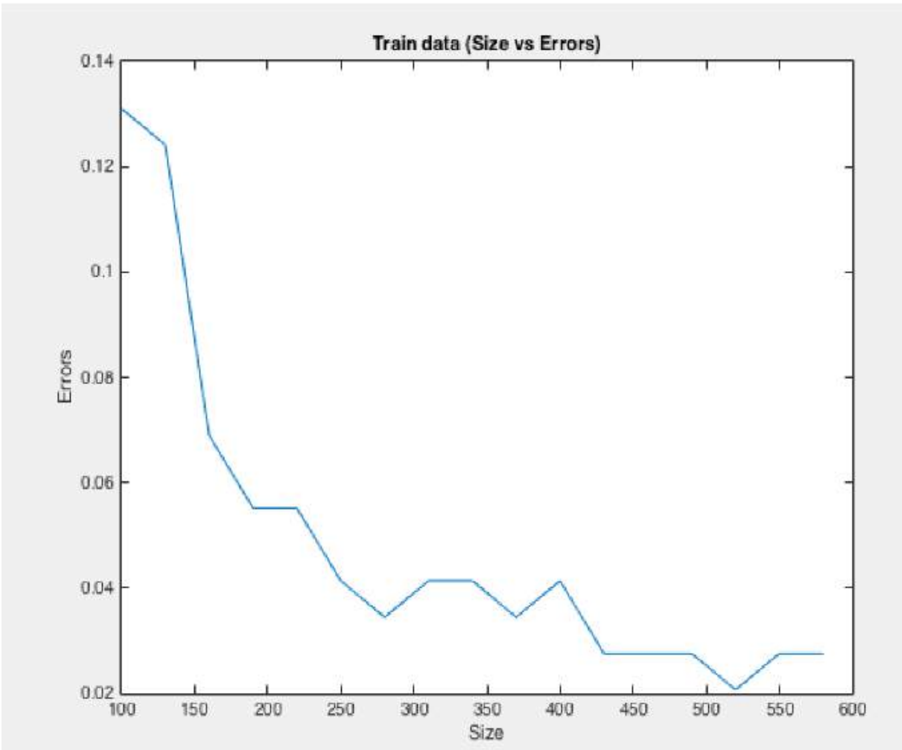
~~g)~~

g) Plots and Matlab code present in folder "Q3_g". Contains two subfolder i.e Test_Data_Error and Training_Data_Error.

Observation: In general we can say that as training data increase the classification error reduces or prediction gets better.
The error in the training data experiment is much lower than testing data for some no. of samples (documents).
However, there are some glitches in the error i.e sometime it goes up when I increased the samples by 30. This is observed both in training and test samples.
Some spikes with increase in sample set could be because additional document might not have a distribution as existing sample thus causing distortion.

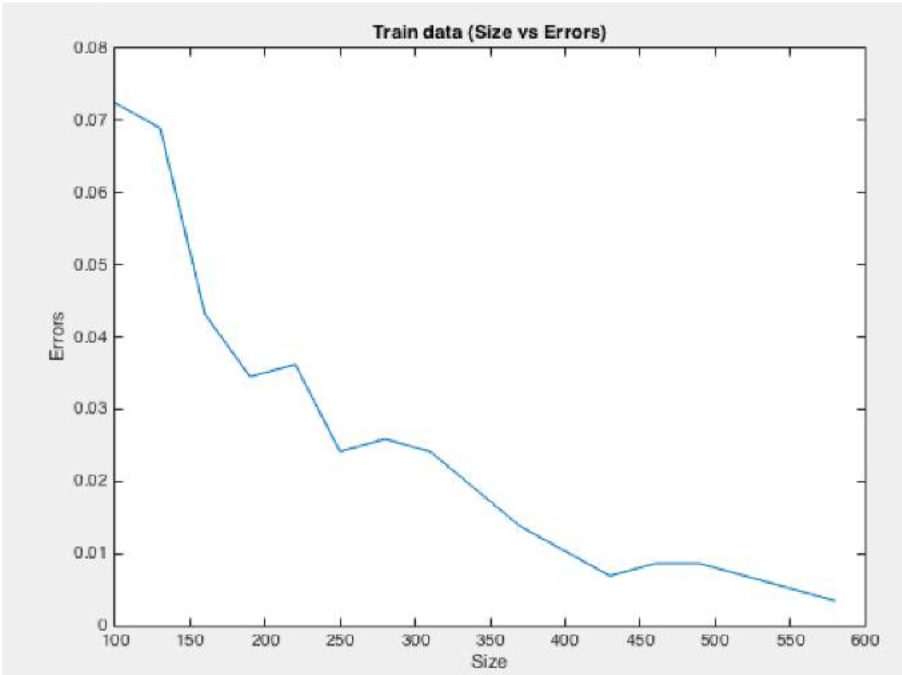
3 g)

Test data errors :



Train data (Size vs Errors)
Errors
Size
0.14
0.12
0.1
0.08
0.06
0.04
0.02
100
150
200
250
300
350
400
450
500
550
600

Training Data Error:



h) Top five words.

A) $y=1$ (Economist) words $\Rightarrow$ 'the', 'to', 'of', 'in', 'a' ~~$P(x_w=1|y=y)$~~

Word indexs = 32, 36, 46, 55, 2

$y=2$ (onion)

Word indexes = 2, 24, 32, 36, 46 $\longrightarrow$ 'a', 'and', 'the', 'to', 'of'

B) Top five words for $\frac{P(x=1|y=1)}{P(x=1/y=2)}$ $\rightarrow$ economist / $\rightarrow$ onion $\frac{P(x_w=1|Y=y)}{P(x_w=1|Y\neq y)}$

Word Indexes = 827, 1450, 3550, 199, 1516

Top five words for $\frac{P(x=1/y=2)}{P(x=1/y=1)}$ $\searrow$ organis, recbon, favour, cras, ~~centr~~ labour

Word Indexes = 20846, 20899, 21003, 20849, 2175

4emlang, 5emlang, retrofit, realiz, coach

Words which has the largest ratio value for two given classes would have the highest contribution in classification or are most informative about class y.

These words are:

$\boxed{\text{4emlang, 5emlang, retrofit, realiz, coach.}}$

Code for 3(g) :

```
function [] = NB_Size_Error(XTest,yTest,XTrain,yTrain)

load('HW2Data.mat');

error = zeros([1 length(100:30:580)]);
count =1 ;

for m= 100:30:580

D = NB_XGivenY_Size(XTrain,yTrain,m);
p = sum(yTrain(1:m)==1)/length(yTrain);
fprintf('p = %d \n',p);
[yHat] = NB_Classify(D,p,XTest);

error(count)=sum((yTest~=yHat))/length(yTest);
fprintf('size = %d, error = %d \n',m,error(count));
count = count+1;

end

data_size = 100:30:580;

figure
plot(data_size,error)
xlabel('Size');
ylabel('Errors');
title('Train data (Size vs Errors)');

end
```

Code for 3(h):

```
function [] = NB_FreqWords(XTrain,yTrain)

load('HW2Data.mat');
[n, v] = size(XTrain);
D = zeros([2 v]);

word_count = zeros([2 v]);

[rows,col,value]= find(XTrain);

num_eco_docs = sum(yTrain==1);
num_oni_docs = sum(yTrain==2);
```

```matlab
length_spare_matrix = length(rows);

for i = 1:length_spare_matrix
    if yTrain(rows(i),1)==1
        word_count(1,col(i)) = word_count(1,col(i))+1;
    else
        word_count(2,col(i)) = word_count(2,col(i))+1;
    end
end

D(1,:) = (word_count(1,:) + .001)/(num_eco_docs + .901);
D(2,:) = (word_count(2,:) + .001)/(num_oni_docs + .901);

B = D(1,:);
for i=1:5
    [M,I] = max(B);
    fprintf(' Economist : val = %d, idx = %d \n',M,I);
    B(1,I)=0;
end

B = D(2,:);
for i=1:5
    [M,I] = max(B);
    fprintf(' Onion : val = %d, idx = %d \n',M,I);
    B(1,I)=0;
end

B = D(1,:) ./ D(2,:);
for i=1:5
    [M,I] = max(B);
    fprintf(' Eco/Onion : val = %d, idx = %d \n',M,I);
    B(1,I)=0;
end

B = D(2,:) ./ D(1,:);
for i=1:5
    [M,I] = max(B);
    fprintf(' Onion/Eco : val = %d, idx = %d \n',M,I);
    B(1,I)=0;
end

end
```