(EC-2 Makeup)

Q.1 Consider the following dataset with 4 records. [4+2 = 6 Marks]

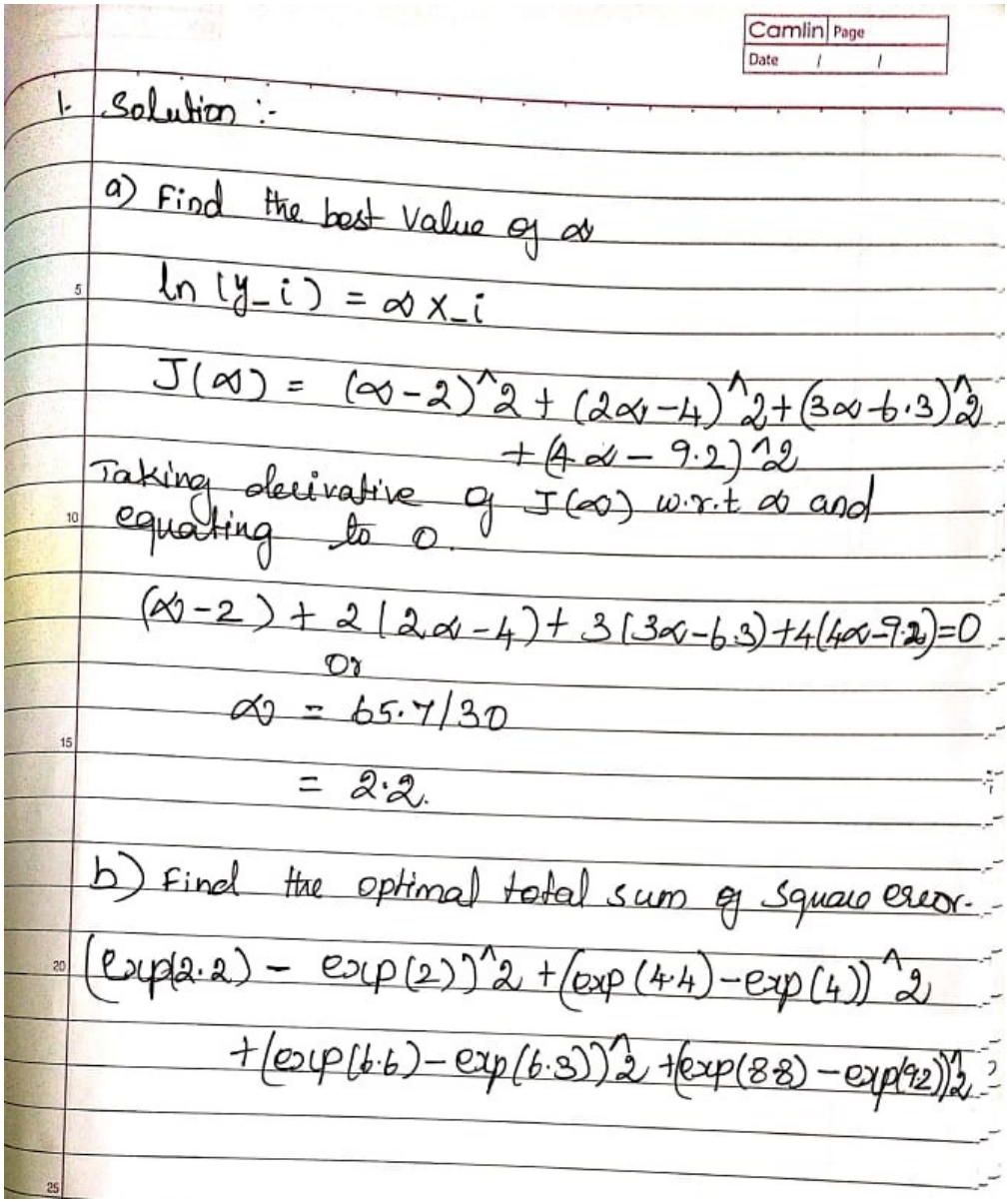
Input X	Output Y
1	exp(2)
2	exp(4)
3	exp(6.3)
4	exp(9.2)

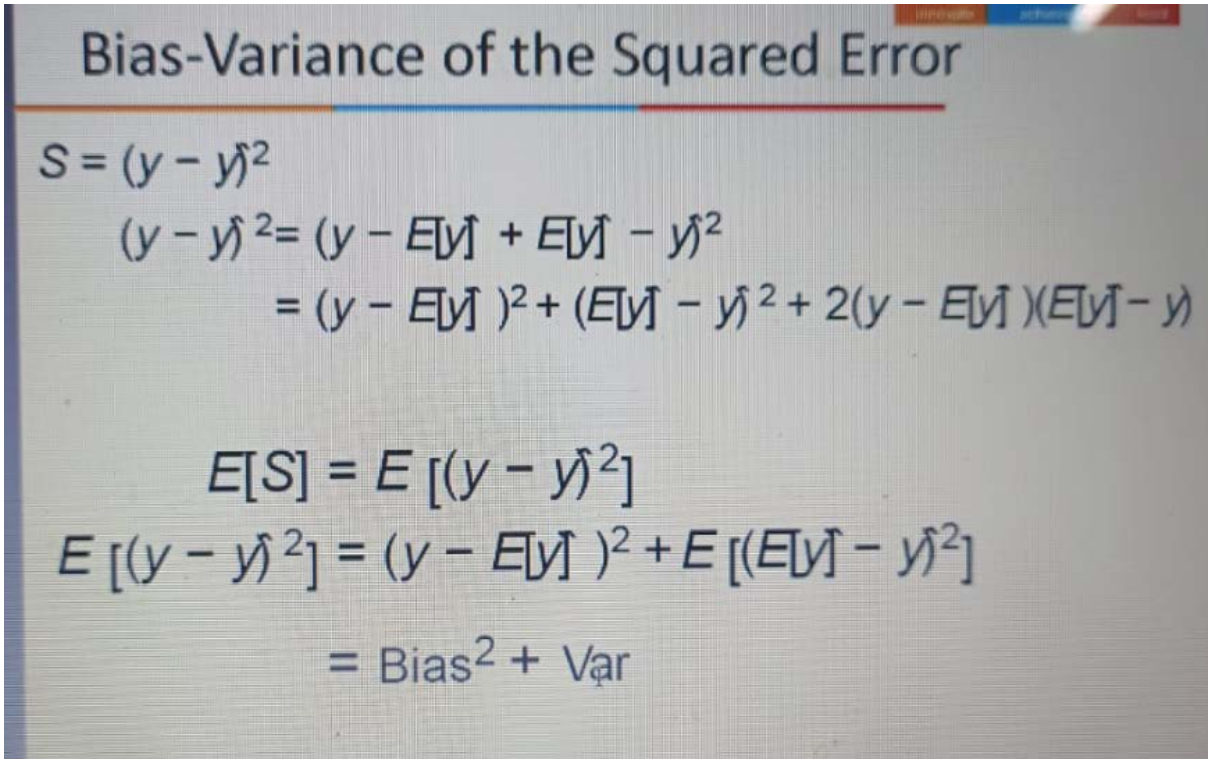
Assume output $y=e^{(\alpha * x)}$. Using linear regression,

(a) Find the best value of α.

(b) Find the optimal total sum of square error.

Solution:

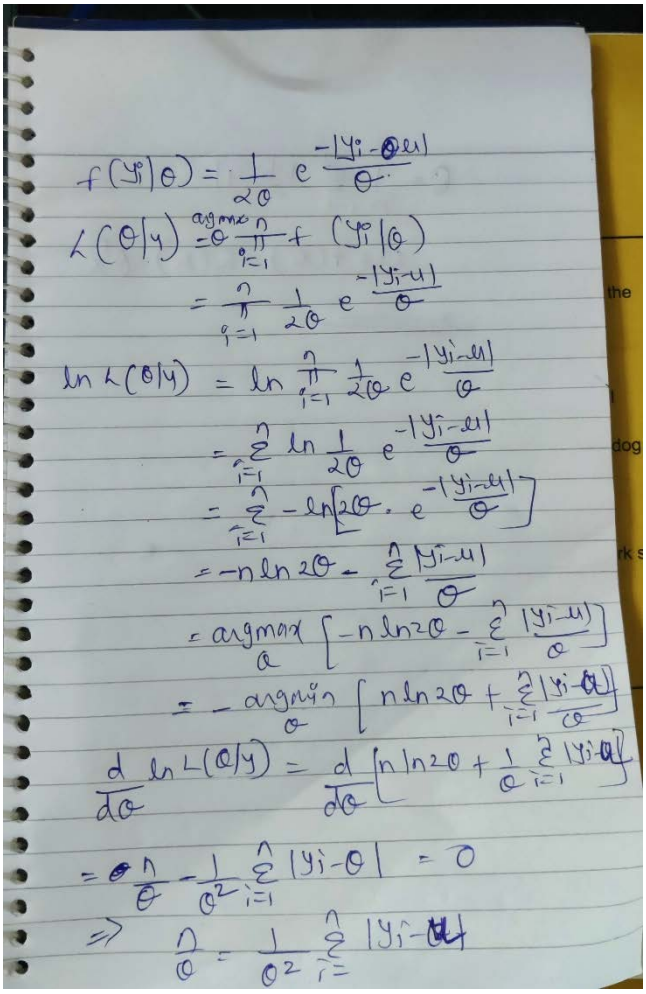




Q.2 Consider inputs x_i which are real valued attributes and the outputs y_i which are real valued of the form $y_i = f(x_i) + e_i$, where $f(x_i)$ is the true function and e_i is a random variable representing laplacian noise with PDF given by

$$f(y_i/\theta) = \frac{1}{2\theta} * e^{\frac{-|y_i - \mu|}{\theta}}$$

Implementing a linear regression model of the form $h(x_i) = \sum_{i=0}^{n} \theta_i x_i$, and $\mu = h(x_i)$, find the maximum likelihood estimator of θ. Comment on the loss function. [4+1 Marks]



$$f(y_i|\theta) = \frac{1}{2\theta} e^{\frac{-|y_i-\theta|}{\theta}}$$

$$L(\theta|y) = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} f(y_i|\theta)$$

$$= \prod_{i=1}^{n} \frac{1}{2\theta} e^{\frac{-|y_i-u|}{\theta}}$$

$$\ln L(\theta|y) = \ln \prod_{i=1}^{n} \frac{1}{2\theta} e^{\frac{-|y_i-u|}{\theta}}$$

$$= \sum_{i=1}^{n} \ln \frac{1}{2\theta} e^{\frac{-|y_i-u|}{\theta}}$$

$$= \sum_{i=1}^{n} -\ln[2\theta \cdot e^{\frac{-|y_i-u|}{\theta}}]$$

$$= -n \ln 2\theta - \sum_{i=1}^{n} \frac{|y_i-u|}{\theta}$$

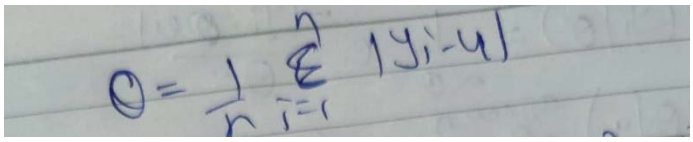
$$= \underset{\theta}{\operatorname{argmax}} \left[-n \ln 2\theta - \sum_{i=1}^{n} \frac{|y_i-u|}{\theta}\right]$$

$$= -\underset{\theta}{\operatorname{argmin}} \left[n \ln 2\theta + \sum_{i=1}^{n} \frac{|y_i-\theta|}{\theta}\right]$$

$$\frac{d}{d\theta} \ln L(\theta|y) = \frac{d}{d\theta}\left[n \ln 2\theta + \frac{1}{\theta} \sum_{i=1}^{n} |y_i-\theta|\right]$$

$$= \frac{n}{\theta} - \frac{1}{\theta^2} \sum_{i=1}^{n} |y_i-\theta| = 0$$

$$\Rightarrow \frac{n}{\theta} = \frac{1}{\theta^2} \sum_{i=1}^{n} |y_i-\theta|$$



$$\theta = \frac{1}{n} \sum_{i=1}^{n} |y_i-u|$$

Comment on Loss function: Instead of MSE, MAE is the maximum likelihood hypothesis. So MAE is appropriate for the loss function.

Q.3 Consider a result prediction system where student's efforts are encoded as percent of time a student has spent studying out of total available time.

- The input X is having just one feature representing the student's efforts having only four discrete values (25%, 50%, 75%, and 100%)
- The output Y is having 3 classes (First class, Second class, Fail)
- The priors for each class are: P(Y = First Class) = 0.5, P(Y = Second class) = 0.3, and P(Y = Fail) = 0.2.
- Based on the past data, the estimated the class-conditional probability P(X| Y) are shown in the following table.

Student's efforts	p(x|y=fail)	p(x|y=second class)	p(x|y=first class)
25	0.7	0.4	0.1
50	0.2	0.3	0.1

75	0.1	0.2	0.3
100	0	0.1	0.7

Consider a following loss function $\iota(\hat{y}, y)$ *where* $\hat{y}$ = *predicted class label* and y is true class label:

$$\iota(\hat{y}, y) = \begin{cases} 0 & \hat{y} = y \\ 1 & \hat{y} = Fail \text{ and } \hat{y} \neq y \\ 2 & \hat{y} = \text{Second class and } \hat{y} \neq y \\ 4 & \hat{y} = First \text{ class and } \hat{y} \neq y \end{cases}$$

Consider modified Naïve Bayes hypothesis function:

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}}\, l(y, \hat{y}) P(Y = y_k) \Pi_i P(X_i | Y = y_k)$$

Use this modified hypothesis function to classify each of the examples in the given table. [5 Marks]

Solution:

	p(y|x)		
	Fail	second class	first class
25	0.35	0.12	0.02
50	0.1	0.09	0.02
75	0.05	0.06	0.06
100	0	0.03	0.14

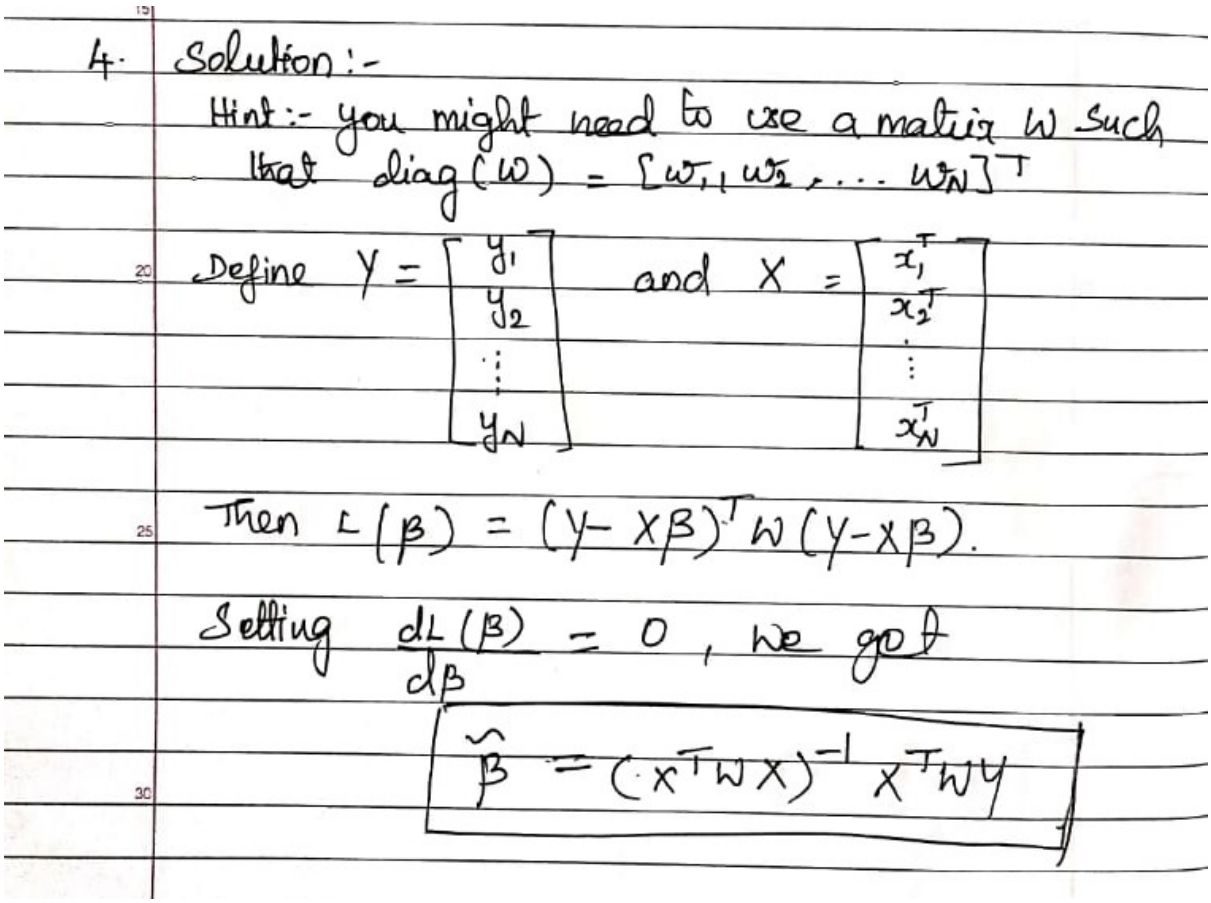
	L(y-hat,y) *p(y/x)	L(y-hat,y) *p(y/x)	L(y-hat,y) *p(y/x)		
	Fail	second class	first class	Highest value	
25	0.35	0.24	0.08	0.35	fail
50	0.1	0.18	0.08	0.18	second class
75	0.05	0.12	0.24	0.24	first class
100	0	0.06	0.56	0.56	first class

Q.4 If we modify the loss function of the linear regression model as follows:

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} w^{(i)} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2$$

Where $w^{(i)}$ is the weight assigned to each training example. Derive the equation to find the value of θ with this modified loss function. Suppose, we estimate the value of $w^{(i)}$ inversely proportional to the variance of the residuals, comment in **no more than 20 words** when you prefer to use this kind of modified loss function. [3+2=5 Marks]

Solution:



Comment: Robust against outliers. Outliers will have higher variance of the residuals resulting into lower weight.

Q.5 Fit a logistic regression. Find the updated weights after the 3 iterations of modified Gradient Descent algorithm where gradient update happens after every training example using a learning rate of 0.5 and initial weights (W_0, W_1, W_2) = (1, 1, 1) for the following data with the logistic regression output given by

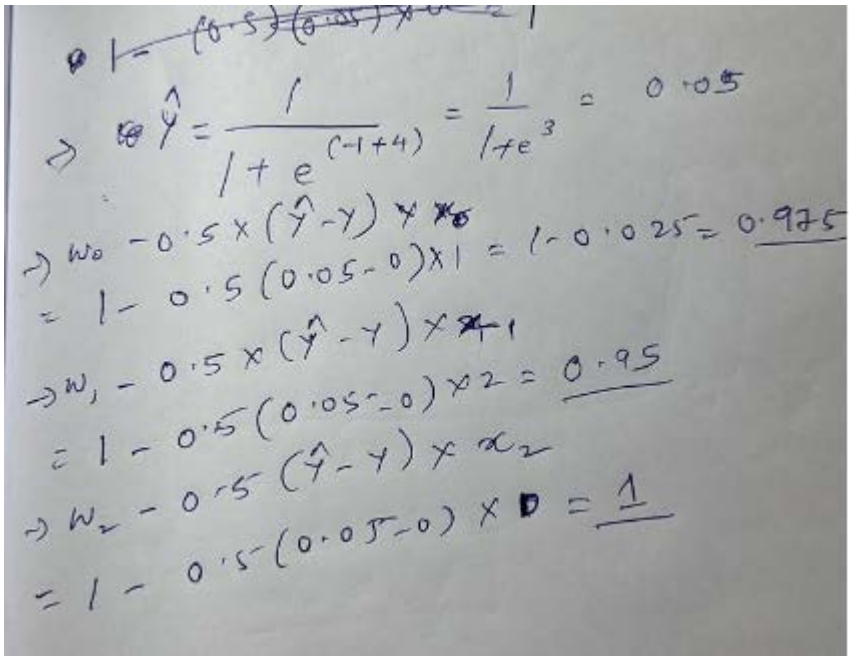
$$\frac{1}{1+e^{\left(-W_0+W_1X_1^2-W_2X_2\right)}}$$

Assume the results obtained after 3 iterations is the final weights. Using this construct the confusion matrix for given below training data. [4+2=6 Marks]

Input X1	Input X2	Output Label
2	0	0
0	2	0
0	-2	0
-2	0	0
0	1	1
0	-1	1

Solution:

				Wi-LR*[Y-Pred – Y]* Xi		
X1	X2	y	Y-Pred = h(X)	W0	w1	w2
2	0	0	0.05	0.975	0.95	1
0	2	0	0.95	0.5	0.95	0.05
0	-2	0	0.27	0.365	0.95	0.32
-2	0	0	0.05			



w0	w1	w2
0.365	0.95	0.32

X1	X2	y	Y-Pred = h(X)	Y
2	0	0	0.03	0
0	2	0	0.73	1
0	-2	0	0.43	0
-2	0	0	0.03	0
0	1	1	0.66	1
0	-1	1	0.51	1

Confusion Matrix		True Class	
		Y=0	Y=1
Predicted Class	Y=0	3	0
	Y=1	1	2

1.) Calculate y-pred. first.

2.) calculate w_0, w_1, w_2 acc.^{n} to formula given:

$$w_i - LR * [y_{pred} - y] * x_i$$

[consider $x_0 \rightarrow 1$] [LR = 0.5, given in ques]

[calculate w_0, w_1, w_2 for 3 iterations & consider the 3rd one to calculate y-pred again]

x_1	x_2	y_{output}	y_{Pred}	$w_i - LR * [y_{pred} - y] * x_i$		
				w_0	w_1	w_2
2	0	1	0.05	0.975	0.95	1
0	2	0	0.9532	0.5	0.95	0.05
0	-2	0	0.2695	0.365	0.95	0.32
-2	0	0	0.05			

For iteration (1) :- w_0 :- $w_0 - (0.5) \times (0.05 - 0) \times (x_0)$.

$= 1 - 0.5 \times 0.05 \times 1$

$\therefore w_0$ = ~~0.525~~ 0.975.

$w_1 = w_1 - 0.5 \times 0.05 \times 2$.

$= 1 - 0.05 = 0.95$

$w_2 = w_2 - 0 = 1$

For iteration (2) :- $w_0 = 0.975 - (0.5 \times 0.95 \times 1)$

(use iteration 1 w_0 value here)

$\therefore w_0 = 0.5$.

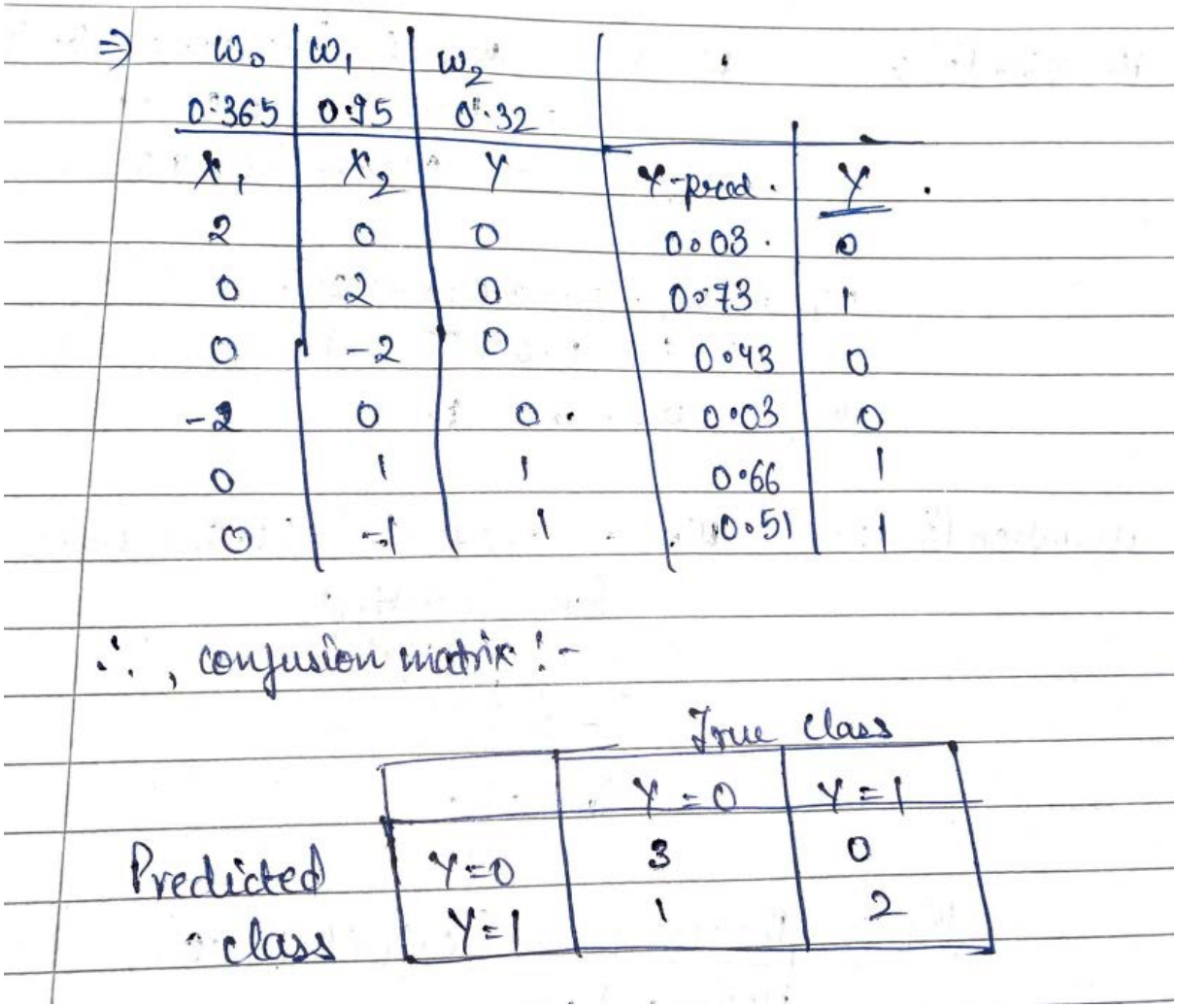
$w_1 = \underbrace{0.95}_{\text{iteration 1 value}} - (0.5 \times 0.95 \times 0)$

$= 0.95$

$w_2 = \underbrace{1}_{\text{iteration 1 value}} - (0.5 \times 0.95 \times 0.2) = 0.05$

⇒ Similarly, calculate for iteration-3. & those values are considered to further calculate Ypred.

⇒ Then take round-off of Ypred & consider it as the output prediction value Y for confusion matrix



⇒

w_0	w_1	w_2
0.365	0.95	0.32

x_1	x_2	y	y-pred.	y
2	0	0	0.08	0
0	2	0	0.73	1
0	−2	0	0.43	0
−2	0	0	0.03	0
0	1	1	0.66	1
0	−1	1	0.51	1

∴, confusion matrix :-

		True class $y=0$	True class $y=1$
Predicted class	$y=0$	3	0
	$y=1$	1	2

Q.6 Consider the following set of training examples:

Instance	Classification	A_1	A_2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

What is the information gain of A_2 relative to these training examples? Provide the equation for calculating the information gain as well as intermediate results. [3 Marks]

Camlin Page
Date / /

6. Solution :-

Entropy $E(S) = E([3+, 3-])$

$= -(3/6)\log_2(3/6) - (3/6)\log_2(3/6)$

Entropy $E(S) = 1$

Gain $(S, a_2) = E(S) - (4/6)E(T) - (2/6)E(F)$

$= 1 - (4/6) - (2/6) \approx 0$

$E(T) = E([2+, 2-]) = 1$

$E(P) = E([1+, 1-]) = 1$

Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
MTech. Software Engineering at DSE (FC04, FA04_1-2021) Cluster

Second Semester 2021-2022
Mid-Semester Test
(EC-2 Regular)

Course No.	: DSECLZG565
Course Title	: Machine Learning
Nature of Exam	: Open Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	: 10-07-2022(FN)

No. of Pages = 2
No. of Questions = 6

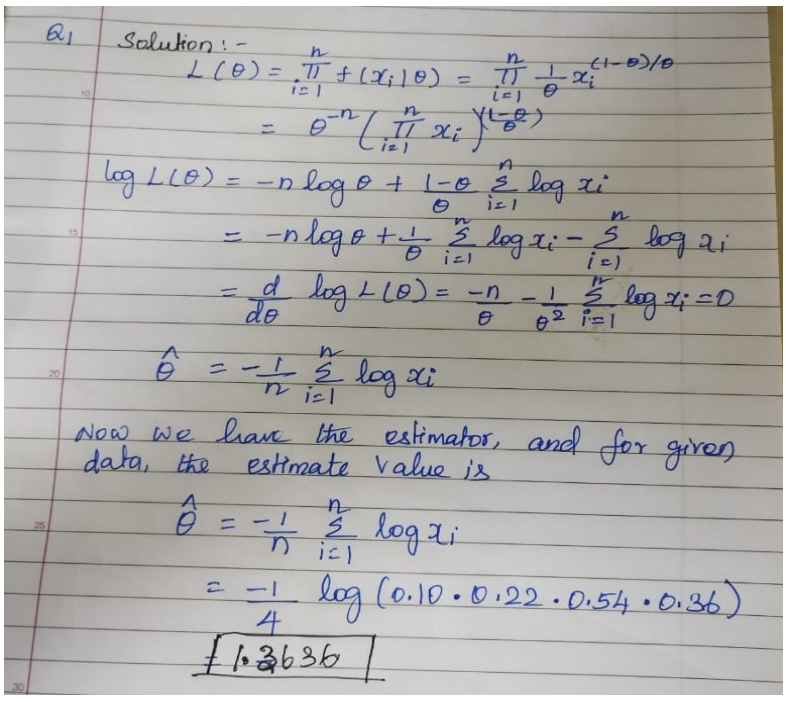
Note:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q.1 Let $T_1, T_2, \ldots T_n$ *be* a random sample of a population describing the website loading time on a mobile browser with probability density function given as:

$$f(t/\theta) = \frac{1}{\theta} t^{\frac{(1-\theta)}{\theta}} \quad where \;\; 0 < t < 1 \; and \; 0 < \theta < \infty$$

Find the maximum likelihood estimator of θ. What is the estimate of θ, if the website loading time from four samples are $t_1 = 0.10$, $t_2 = 0.22$, $t_3 = 0.54$, $t_4 = 0.36$. [5 Marks]

Solution:



Marking Scheme: Derivation of θ =3 marks (step wise marks)

θ Computation = 2 marls (wrong value = 0 marks)

Q.2 As a part of efforts to improve students' performance in the exams, you have been given the data showing number of study hours spent by students, their gender and their final results as pass or fail. Using this sample dataset, apply Naïve Bayes classification technique, to classify the test case {No of study hours = 3.5, Gender="male"} either as "Pass", or "Fail". [5 Marks]

No of study hours	Gender	Final result
4.5	Male	Pass
7	Female	Pass
2	Male	Fail
4	Female	Fail
2.5	Male	Fail
3	Female	Fail
8.3	Male	Fail
8	Female	Pass
9	Male	Pass

Solution:

1. Prior: [1M]

p(y=Pass)	p(y=Fail)
0.444444	0.555556

2. No of study hours –X1: continuous variable, applying class conditional PDF [1M]
3. $$p(X_i = x|Y = y_k) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} \; e^{-\frac{1}{2}(\frac{x-\mu_{ik}}{\sigma_{ik}})^2}$$

	Variance	mean
Pass class	2.945	7.2
Fail class	4.64	3.9

4. X1=3.5, X2="male" [3M]

$$\hat{Y} \leftarrow \underset{y_k}{\operatorname{argmax}} P(Y = y_k)\Pi_i P(X_i|Y = y_k)$$

p(X1/ y=Pass)	0.105614
p(X1/ y=Fail)	0.184564

p(X2/ y=Pass)	0.5
p(X2/ y=Fail)	0.6

P(y=Pass/X)	0.02346969
P(y=Fail/X)	0.061521395

Class : Fail

Q.3 The 2-input AND gate is implemented using logistic regression classifier with gradient descent optimization algorithm. The model parameters at time *t* are given by θ_0=0, θ_1=0, and θ_2=0. Given binary input (x1,x2), [2+3 = 5 Marks]

a) What will be value of the loss function at t? [2M]

Solution:

Cross entropy loss:

$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

x1	x2	Target y	Actual Output-yhat	y.ln(yhat)+(1-y)ln(1-yhat)
0	0	0	0.5	0*ln0.5+(1-0)*ln(1-0.5)
0	1	0	0.5	0*ln0.5+(1-0)*ln(1-0.5)
1	0	0	0.5	0*ln0.5+(1-0)*ln(1-0.5)
1	1	1	0.5	1*ln(0.5)

total loss=	0.693147181

Q.3 The 2-input AND gate is **implemented using logistic regression classifier with gradient descent optimization algorithm**. The model parameters at time *t* are given by $\theta_0=0$, $\theta_1=0$, and $\theta_2=0$. Given binary input (x1,x2), [2+3 = 5 Marks]

a) What will be value of the loss function at t? [2M]

Solution:

Cross entropy loss:

$$= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log (1-h_\theta(x^{(i)}))]$$

x1	x2	Target y	Actual Output-yhat	y.ln(yhat)+(1-y)ln(1-yhat)
0	0	0	0.5	0*ln0.5+(1-0)*ln(1-0.5)
0	1	0	0.5	0*ln0.5+(1-0)*ln(1-0.5)
1	0	0	0.5	0*ln0.5+(1-0)*ln(1-0.5)
1	1	1	0.5	1*ln(0.5)

total loss=	0.693147181

b) What will be the values of θ_0, θ_1 and θ_2 at (t+1) with learning rate $\alpha=1$ and L2 regularization constant $\lambda=1$? [3M]

Solution:
Cost function

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

Apply gradient descent update rule

y-hat	y	yhat-y	x0	(yhat-y)x0	w0-new
0.5	0	0.5	1	0.5	-0.25
0.5	0	0.5	1	0.5	
0.5	0	0.5	1	0.5	
0.5	1	-0.5	1	-0.5	

y-hat	y	yhat-y	x1	(yhat-y)x1	regularized w1-new
0.5	0	0.5	0	0	0
0.5	0	0.5	0	0	
0.5	0	0.5	1	0.5	
0.5	1	-0.5	1	-0.5	

y-hat	y	yhat-y	x2	(yhat-y)x1	regularized w2-new
0.5	0	0.5	0	0	0
0.5	0	0.5	1	0.5	
0.5	0	0.5	0	0	
0.5	1	-0.5	1	-0.5	

b) What will be the values of θ_0, θ_1 and θ_2 at (t+1) with **learning rate** α=1 and **regularization constant** λ=1? [3M]

L1 - Lasso
L2 - Ridge

Solution:
Cost function

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\ \log(h_\theta(x^{(i)})) + (1-y^{(i)})\ \log(1-h_\theta(x^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

R

$\frac{\lambda}{2m}\sum|\theta_j|$ ✓

$\theta_0 = \theta_0 - \frac{\alpha}{m}\sum[y-\hat{y}]\cdot x_0$

Apply gradient descent update rule

y-hat ✓	y ✓	yhat-y ✓	x0 ✓	(yhat-y)x0	w0-new
0.5	0	0.5	1	0.5	-0.25
0.5	0	0.5	1	0.5	
0.5	0	0.5	1	0.5	
0.5	1	-0.5	1	-0.5	

$\theta_0 = 0 - \frac{1}{4}\times 1$

$\theta_0 = -0.25$

y-hat	y	yhat-y	x1	(yhat-y)x1	regularized w1-new
0.5	0	0.5	0	0	0
0.5	0	0.5	0	0	
0.5	0	0.5	1	0.5	
0.5	1	-0.5	1	-0.5	

$\theta_1 = 0 - \frac{1}{4}\times 0 = 0$

y-hat	y	yhat-y	x2	(yhat-y)x1	regularized w2-new
0.5	0	0.5	0	0	0
0.5	0	0.5	1	0.5	
0.5	0	0.5	0	0	
0.5	1	-0.5	1	-0.5	

$\theta_2 = 0 - \frac{1}{4}\times 0 = 0$

Q.4 We claim that there exists a value for α in the following data : (1.0, 4.0), (2.0, 9.0) , (3.0, α) such that the line $y = 2 + 3x$ is the best least-square fit for the data. Is this claim true? If the claim is true, find the value of α. Otherwise, explain why the claim is false. Give **detailed** mathematical justification for your answer. [5 Marks]

For the line $y = a + bx$, the MSE turns out to be

$$(4.0 - (a+b))^2 + (9.0 - (a+2b))^2 + (\alpha - (a+3b))^2$$

$$\frac{\partial MSE}{\partial a} = 0$$

$$\Rightarrow 2(4.0 - a - b)(-1) + 2(9.0 - a - 2b)(-1) + 2(\alpha - a - 3b)(-1) = 0$$

$$\Rightarrow 4.0 - a - b + 9.0 - a - 2b + \alpha - a - 3b = 0$$

$$\alpha + 13 - 3a - 6b = 0$$

$$\Rightarrow \alpha = 3a + 6b - 13 \quad - (1)$$

$$\frac{\partial MSE}{\partial b} = 0$$

From ① and ②, the parameters a and b for the best fit must solve

$$\frac{6a + 14b - 22}{3} = 3a + 6b - 13$$

$$\Rightarrow 6a + 14b - 22 = 9a + 18b - 39$$

$$3a + 4b = 17$$

Substituting $a = 2$ and $b = 3$, we see that the above requirement is not met $\Rightarrow$ the given claim is false

Marking Scheme: calculation of 1 and 2 – 3M
Equation of a and b = 1M
Final answer =1M

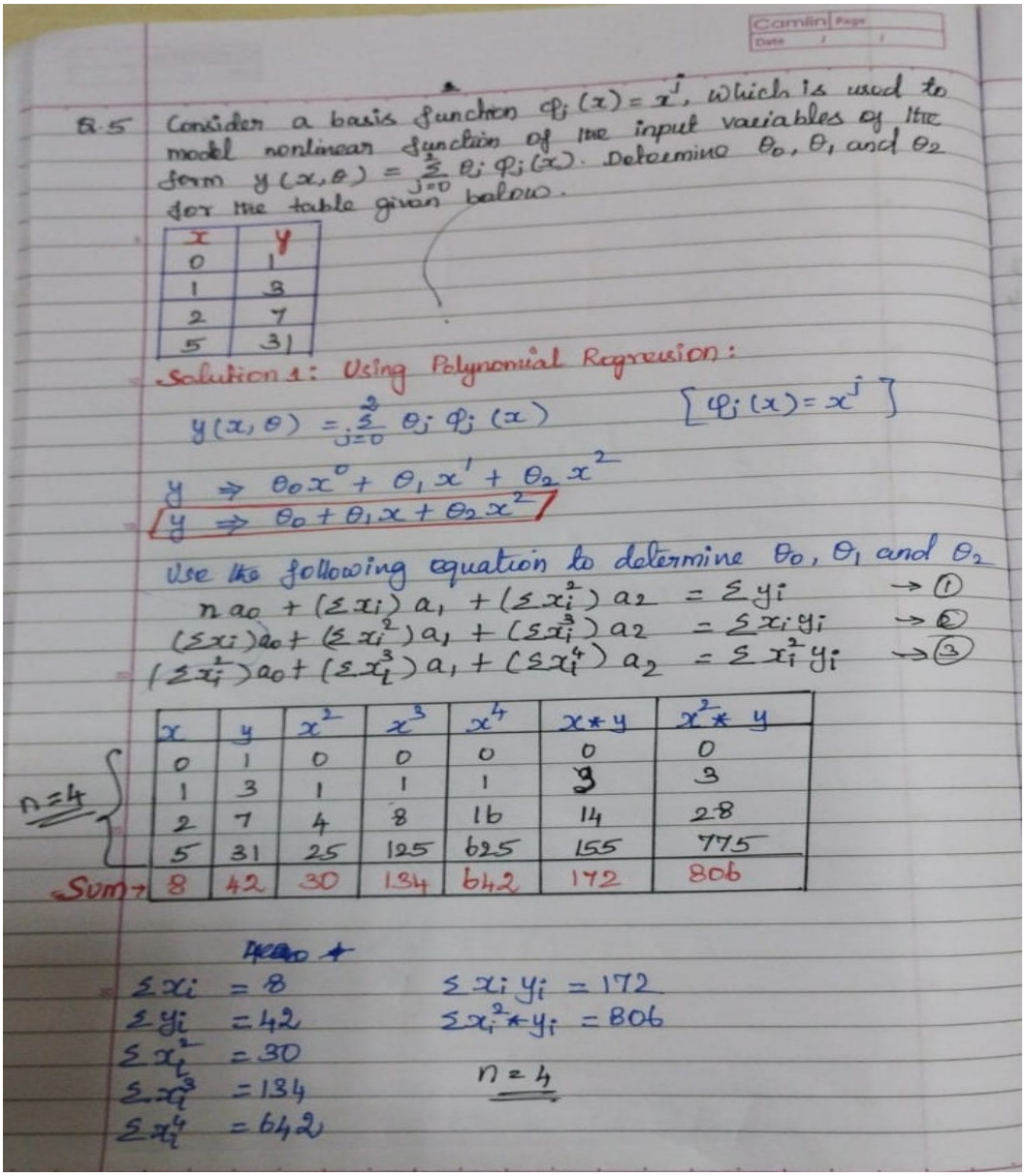
Q.5 Consider a basis function $\phi_j(x) = x^j$, which is used to model nonlinear function of the input variables of the form $y(x,\theta) = \sum_{j=0}^{2} \theta_j \phi_j(x)$. Determine θ_0, θ_1 and θ_2 for the table given below. [6 Marks]

x	y
0	1
1	3
2	7
5	31

Solution:

Polynomial Regression: $y = \theta_0 + \theta_1 x + \theta_2 x^2$ [2M]

Solution: Method 1 **[4M]**



Q.5 Consider a basis function $\varphi_j(x) = x^j$, which is used to model nonlinear function of the input variables of the form $y(x,\theta) = \sum_{j=0}^{2} \theta_j \varphi_j(x)$. Determine θ_0, θ_1 and θ_2 for the table given below.

x	y
0	1
1	3
2	7
5	31

Solution 1: Using Polynomial Regression:

$y(x,\theta) = \sum_{j=0}^{2} \theta_j \varphi_j(x)$ $\quad [\varphi_j(x) = x^j]$

$y \Rightarrow \theta_0 x^0 + \theta_1 x^1 + \theta_2 x^2$

$y \Rightarrow \theta_0 + \theta_1 x + \theta_2 x^2$

Use the following equation to determine θ_0, θ_1 and θ_2

$n a_0 + (\Sigma x_i) a_1 + (\Sigma x_i^2) a_2 = \Sigma y_i \rightarrow (1)$

$(\Sigma x_i) a_0 + (\Sigma x_i^2) a_1 + (\Sigma x_i^3) a_2 = \Sigma x_i y_i \rightarrow (2)$

$(\Sigma x_i^2) a_0 + (\Sigma x_i^3) a_1 + (\Sigma x_i^4) a_2 = \Sigma x_i^2 y_i \rightarrow (3)$

x	y	x^2	x^3	x^4	x*y	x^2*y
0	1	0	0	0	0	0
1	3	1	1	1	3	3
2	7	4	8	16	14	28
5	31	25	125	625	155	775
Sum →	8	42	30	134	642	172	806

n = 4

$\Sigma x_i = 8$

$\Sigma y_i = 42$

$\Sigma x_i^2 = 30$

$\Sigma x_i^3 = 134$

$\Sigma x_i^4 = 642$

$\Sigma x_i y_i = 172$

$\Sigma x_i^2 * y_i = 806$

$n = 4$

Camlin | Page
Date / /

Substitute the above values in all 3 equation

$$4a_0 + 8a_1 + 30a_2 = 42$$
$$8a_0 + 30a_1 + 134a_2 = 172$$
$$30a_0 + 134a_1 + 642a_2 = 806.$$

Coffient Matrix — Constant Matrix

$$\begin{bmatrix} 4 & 8 & 30 \\ 8 & 30 & 134 \\ 30 & 134 & 642 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 42 \\ 172 \\ 806 \end{bmatrix}$$

Step 1 :- $R_1 \rightarrow R_1/4$

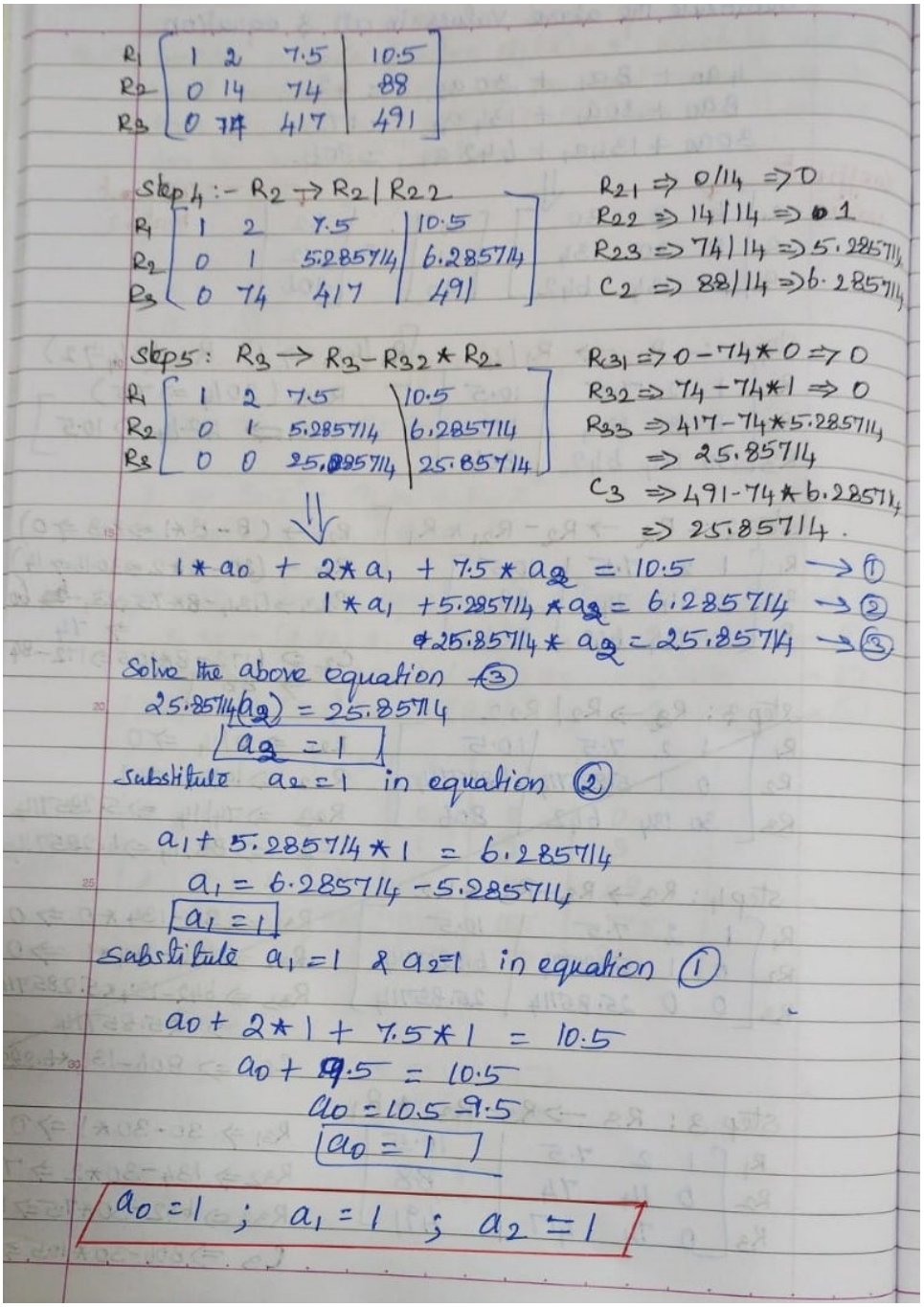
$$\begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \end{array} \left[\begin{array}{ccc|c} 1 & 2 & 7.5 & 10.5 \\ 8 & 30 & 134 & 172 \\ 30 & 134 & 642 & 806 \end{array}\right]$$

[R_{11} (4/4 ⇒ 1) R_{12} (8/4 ⇒ 2)
R_{13} (30/4 ⇒ 7.5)
C_1 ⇒ 42/4 ⇒ 10.5]

Step 2 :- $R_2 \rightarrow R_2 - R_{21} * R_1$

$$\begin{array}{c} \\ R_1 \\ R_2 \\ R_3 \end{array} \left[\begin{array}{ccc|c} 1 & 2 & 7.5 & 10.5 \\ 0 & 14 & 74 & 88 \\ 30 & 134 & 642 & 806 \end{array}\right]$$

[$R_{21} \rightarrow (8 - 8*1 \Rightarrow 8-8 \Rightarrow 0)$
$R_{22} \rightarrow (30 - 8*2 \Rightarrow 30-16 \Rightarrow 14)$
$R_{23} \rightarrow (134 - 8*7.5 \Rightarrow 134 - 60 \Rightarrow 74$
$C_2 \Rightarrow (172 - 8*10.5 \Rightarrow 172 - 84 \Rightarrow 88$]



Solution: Method 2

Using closed form solution: [4M]

$$\boldsymbol{\theta} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}$$

$\theta_0 = 1$, $\theta_1 = 1$ and $\theta_2 = 1$

Q.6 Consider the dataset of binary values in terms of attribute-value pairs where F is the value, and A,B, C are attributes. What is the entropy of the dataset? Fill in the columns for A and B, if it is known that A has maximum information gain and B has minimum information gain. Give mathematical justification for your answer. [4 Marks]

A	B	C	F
		0	0
		1	1
		0	1
		1	0
		0	1
		1	1
		0	1
		1	1

Solution:
For the entropy problem, the column F is the output attribute.

2 Marks:
Let column A = column F, so that the ith entries of the two column match each other. The information gain can be written as

$InformationGain(S, A) = Entropy(S) - \sum \frac{|S_A|}{|S|} Entropy(S_A)$ where the sum is over the attribute values of A.

Since the column entries of A match with those of F we see that the set $S_{A=0}$ is full of 0s and $S_{A=1}$ is full of 1s, so that $Entropy(S_{A=0})$ is 0 and so is $Entropy(S_{A=1})$. From the equation on Information Gain we can see that we get the maximum information gain possible in this case.

2 Marks:
The information gain with respect to column B can be written as $InformationGain(S, B) = Entropy(S) - \sum \frac{|S_B|}{|S|} Entropy(S_B)$.

For minimum information gain we see that if let the column B be the column of all 1s, then we have $S_{B=1} = S$ and $S_{B=0} = \phi$. Once again plugging this into the information gain equation shows that the information gain with respect to B is 0.

The arguments above work for maximum information gain when A is taken to be complement of F, and B is taken to be all zeroes rather than all 1s.

(1) It is known that a natural law obeys the quadratic relationship $y = ax^2$. What is the best linear curve that can be used to model this data if all of the data points are drawn uniformly at random in the interval (0,1)?

Let $y = px + q$ be the equation of the linear curve

We need to minimize $L = \int_0^1 (ax^2 - px - q)^2 dx$

$$L = \frac{a^2}{5} + \frac{p^2}{3} + q^2 - \frac{2ap}{4} - \frac{2aq}{3} + \frac{2pq}{2}$$

$$\frac{\partial L}{\partial p} = 0 \quad \frac{\partial L}{\partial q} = 0 \Rightarrow \frac{2p}{3} + q = \frac{a}{2}; \; 2q + p = \frac{2a}{3}$$

$$\Rightarrow \boxed{p = a, \; q = -\frac{a}{6}}$$

Marking Scheme

Setting up the integral → 1 marks

Getting this expression → 2 marks

Final Solution → 2 marks

(2) Consider the following dataset for text classification where three training instances are given with corresponding classifications into the '+' or –'- category:

Hindi India India	+
India Kannada Hindi	+
Chinese Hindi India	-

Showing all intermediate calculations, find the appropriate classification for the test instance: Chinese Kannada Chinese using the Naïve Bayes text classification algorithm.

First we observe $P(+) = \frac{2}{3}$ and $P(-) = \frac{1}{3}$

$docs_+$ = Hindi India India India Kannada Hindi

$docs_-$ = Chinese Hindi India

Vocabulary = {Hindi, India, Kannada, Chinese}

$P(\text{Hindi}/+) = \frac{2+1}{6+4} = \frac{3}{10}$ | $P(\frac{\text{Hindi}}{-}) = \frac{1+1}{3+4} = \frac{2}{7}$

$P(\text{India}/+) = \frac{3+1}{6+4} = \frac{4}{10}$ | $P(\frac{\text{India}}{-}) = \frac{1+1}{3+4} = \frac{2}{7}$

$P(\text{Kannada}/+) = \frac{1+1}{6+4} = \frac{2}{10}$ | $P(\frac{\text{Kannada}}{-}) = \frac{0+1}{3+4} = \frac{1}{7}$

$P(\text{Chinese}/+) = \frac{0+1}{6+4} = \frac{1}{10}$ | $P(\frac{\text{Chinese}}{-}) = \frac{1+1}{3+4} = \frac{2}{7}$

Decision $P(+)P(\frac{\text{Chinese}}{+})P(\frac{\text{Kannada}}{+})P(\frac{\text{Chinese}}{+})$

$P(-)P(\frac{\text{Chinese}}{-})P(\frac{\text{Kannada}}{-})P(\frac{\text{Chinese}}{-})$ vs

$\frac{2}{3} \times \frac{1}{10} \times \frac{2}{10} \times \frac{1}{10}$ vs $\frac{1}{3} \times (\frac{2}{7})^2 \times \frac{1}{7}$

0.00132 vs 0.0038

Marking Scheme:

Positive Conditional Probabilities → 1.5 mark
Negative Conditional Probs → 1.5 marks
Decision - 2 marks

Q.3. There are two varieties of cucumbers – C_1 and C_2 which have different distributions of length. The joint probability density function of the length of the cucumber and category 1 is denoted by $p(x, C_1)$, and is a uniform distribution over the range (10cm, 30cm). Similarly $p(x, C_2)$ is a uniform distribution over the range (20cm, 50cm). What is the error of classification we will make if we assert that all cucumbers of length less than 25cm are of Variety 1 and all cucumbers of length greater than 25cm are of Variety 2? [5 Marks]

The information in this question is incompletely specified

The functions given for $p(x, C_1)$ and $p(x, C_2)$ in the question are actually $p(x/C_1)$ and $p(x/C_2)$ respectively. To get $p(x, C_1)$ and $p(x, C_2)$ we need to multiply $p(x/C_1)$ and $p(x/C_2)$ by $p(C_1)$ and $p(C_2)$ respectively. $p(C_1)$ and $p(C_2)$ are not specified in the question, so p(mistake) should be calculated in terms of $p(C_1)$ and $p(C_2)$.

We have

$$p(\text{mistake}) = \int_{R_1} p(x, C_2)\,dx + \int_{R_2} p(x, C_1)\,dx$$

$$= p(C_2) \int_{R_1} p(x/C_2)\,dx + p(C_1) \int_{R_2} p(x, C_1)\,dx$$

$$= p(C_2) \int_{20}^{25} p(x/C_2)\,dx + p(C_1) \int_{25}^{30} p(x, C_1)\,dx$$

$$= P(C_2)\frac{5}{30} + P(C_1)\frac{5}{20}$$

$$= P(C_2)\frac{1}{6} + P(C_1)\frac{1}{4}$$

Marking Scheme

Formula for $P(\text{mistake})$ = 3 marks

Final Calculation = 2 marks

(3) Consider the standard set of Gaussian Naïve Bayes assumptions used in the derivation of the logistic regression expression, but with one modification – the class conditional density for each class has unique values for both the mean and variance, rather than a common value for the variance, i.e $P(X_i/Y = y_k) = N(\mu_{ik}, \sigma_{ik})$. Find the expression for $P(Y = 1/X_1, X_2, \ldots X_n)$ in this case and find the decision boundary.

$$P(Y=1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$$

$$= \frac{1}{1 + \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$

$$= \frac{1}{1 + \exp(\ln \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)})}$$

$$= \frac{1}{1 + \exp(\ (\ln \frac{1-\pi}{\pi}) + \sum_i \ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)})}$$

Now $\ln \frac{P(X_i/Y=0)}{P(X_i/Y=1)} = -\frac{1}{2}\left(\frac{X_i - \mu_{i0}}{\sigma_{i0}}\right)^2 + \frac{1}{2}\left(\frac{X_i - \mu_{i1}}{\sigma_{i1}}\right)^2$

$$= \frac{1}{2}\frac{X_i^2}{\sigma_{i1}^2} - \frac{1}{2}\frac{X_i^2}{\sigma_{i0}^2} - \frac{\mu_{i1}X_i}{\sigma_{i1}} + \frac{\mu_{i0}X_i}{\sigma_{i0}}$$

$$+ \frac{\mu_{i0}^2}{\sigma_{i0}^2} - \frac{\mu_{i1}^2}{\sigma_{i1}^2}$$

This will give rise to an expression of the form $P(Y=1/X) = \frac{1}{1+\exp(\omega_0 + \sum \omega_i X_i + \alpha X_i^2)}$

where $\alpha = \frac{1}{2}\left(\frac{1}{\sigma_{i1}^2} - \frac{1}{\sigma_{i0}^2}\right)$

The decision boundary is quadratic instead of linear

Marking Scheme

Derivation for the new expression = 4 marks

Decision boundary quadratic → 1 mark

(4) Consider the following dataset

price	maintenance	capacity	Safety measures	Beneficial
lowpriced	cheap	5	yes	yes
lowpriced	average	5	yes	yes
lowpriced	cheap	5	yes	no
lowpriced	excessive	3	no	no
fair	average	5	no	no
fair	average	5	no	yes
fair	excessive	3	yes	no
fair	excessive	6	yes	yes
overpriced	average	5	yes	yes
overpriced	excessive	3	yes	no
overpriced	excessive	6	yes	no

Classify the new instance given: "price = fair, maintenance = cheap, capacity = 5, safety measures = yes". Use Laplace smoothing only when needed for an attribute.

We have $P(\text{Beneficial} = \text{Yes}) = \frac{5}{11}$

$P(\text{Beneficial} = \text{No}) = \frac{6}{11}$

$P(\text{Price} = \text{fair} \mid \text{Yes}) = \frac{2}{5}$ | $P(\text{Price} = \text{fair} \mid \text{no}) = \frac{2}{6}$

$P(\text{maint} = \text{cheap} \mid \text{Yes}) = \frac{1}{5}$ | $P(\text{maint} = \text{cheap} \mid \text{no}) = \frac{1}{6}$

$P(\text{capacity} = 5 \mid \text{Yes}) = \frac{4}{5}$ | $P(\text{capacity} = 5 \mid \text{no}) = \frac{2}{6}$

$P(\text{Safety} = \text{Yes} \mid \text{Yes}) = \frac{4}{5}$ | $P(\text{safety} = \text{Yes} \mid \text{no}) = \frac{4}{6}$

Compare $P(\text{Yes})\,P(\text{fair} \mid \text{yes})\,P(\text{cheap} \mid \text{yes})\,P(5 \mid \text{yes})\,P(\text{yes} \mid \text{yes})$

with $P(\text{No})\,P(\text{fair} \mid \text{No})\,P(\text{cheap} \mid \text{no})\,P(5 \mid \text{no})\,P(\text{yes} \mid \text{no})$

$\frac{5}{11} \times \frac{2}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{4}{5}$ (vs) $\frac{6}{11} \times \frac{2}{6} \times \frac{1}{6} \times \frac{2}{6} \times \frac{4}{6}$

0.023 vs 0.0067

The instance should be classified as Yes

Marking Scheme

$P(Yes)$, $P\left(\frac{x_i}{Yes}\right)$ → 2 Marks

$P(No)$, $P\left(\frac{x_i}{No}\right)$ → 2 Marks

final decision → 1 Mark

(5) Consider the dataset in terms of attribute-value pairs where F is the value, and A,B, C are attributes. What is the entropy of the dataset? Compute the information gain with respect to the attributes A, B and C. Which attribute would be used at the root of the decision tree constructed by the ID3 algorithm?

A	B	C	F
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	1
1	1	0	1
1	1	1	1

$$Entropy(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

$$= -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}$$

$$Gain(S, A) = Entropy(S) - \frac{|S_{A=0}|}{|S|} Entropy(S_{A=0})$$

$$- \frac{|S_{A=1}|}{|S|} Entropy(S_{A=1})$$

$$= \left(-\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}\right) - \frac{4}{8}\left(-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}\right)$$

$$- \frac{4}{8}\left(-\frac{4}{4}\log\frac{4}{4} - \frac{0}{4}\log\frac{0}{4}\right)$$

$$= 0.81 - 0.5 = \underline{0.31}$$

$$Gain(S, B) = Entropy(S) - \frac{|S_{B=0}|}{|S_B|} Entropy(S_{B=0})$$

$$\frac{|S_B=1|}{|S|} \text{Entropy}(S_B=1)$$

$$= \left(-\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}\right) - \frac{4}{8}\left(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}\right)$$

$$-\frac{4}{8}\left(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}\right) = 0.81 - 0.81 = 0$$

$$\text{Gain}(S,C) = \left(-\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}\right) - \frac{4}{8}\left(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}\right)$$

$$-\frac{4}{8}\left(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}\right) = 0.81 - 0.81 = 0$$

$$\text{Gain}(S,A) > \text{Gain}(S,B) = \text{Gain}(S,C)$$

$\Rightarrow$ A is the root attribute

Marking Scheme

Entropy of data set $\rightarrow$ 2 Marks

Gains of A, B, C $\rightarrow$ 2 Marks

Root attribute $\rightarrow$ 1 Mark

Format of Question paper for Mid-Semester Test

Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2020-2021

Mid-Semester Test
(EC-2 Regular/ EC-2 Make-up)

Course No.	:
Course Title	:
Nature of Exam	: Open Book
Weightage	: 30% or 35% (As per Course Handout)
Duration	: 2 Hours
Date of Exam	: 06/03/2021 or 19/03/2021 (FN/AN)

No. of Pages =
No. of Questions =

Note to Students:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Please Note:
A **maximum of 12 main questions** with Q.Nos. 1 to N**. Please do not include any objective type questions. Please do not include any reference material like data tables in the question paper, as all Mid-Semester Test are completely Open Book examinations.**

Q.1. We have five data points $x_1 = 1, x_2 = 3, x_3 = -1, x_4 = 4, x_5 = -3$ which are obtained from sampling a Gaussian distribution of zero mean. What is the Maximum Likelihood Estimate of the variance of the Gaussian distribution? Show all the steps in the calculation. [5 Marks]

We are sampling from a normal distribution $N(0, \sigma^2)$ where σ is ?

The likelihood is $\prod_{i=1}^{5} x_i / N(0, \sigma^2)$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2/\sigma^2}\right)\left(\frac{1}{\sigma\sqrt{2\pi}} \cdot \frac{1}{2}x_2^2/\sigma^2\right) \cdots \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{x_5^2}{\sigma^2}}\right)$$

$$L = \frac{1}{(2\pi)^{5/2}\sigma^5} e^{-\frac{1}{2}\left(\frac{x_1^2 + x_2^2 + \cdots x_5^2}{\sigma^2}\right)}$$

To find the value of σ that maximizes likelihood we have to take the derivative with respect to σ of L and equate it to

6.

It is more convenient to work with $\log L$

$$\frac{d}{d\sigma}(\log L) = 0$$

$$\Rightarrow \frac{d}{d\sigma}\left(\log \frac{1}{(2\pi)^{5/2}} + \log \frac{1}{\sigma^5} + -\frac{1}{2}\frac{(x_1^2 + x_2^2 + \cdots x_5^2)}{\sigma^2}\right)$$

$$= 0$$

$$\Rightarrow \frac{d}{d\sigma}\left(-5\log\sigma - \frac{x_1^2 + x_2^2 + \cdots x_5^2}{2\sigma^2}\right) = 0$$

$$\Rightarrow \frac{-5}{\sigma} + \frac{1}{\sigma^3}\left(x_1^2 + x_2^2 + \cdots x_5^2\right) = 0$$

$$\Rightarrow 5\sigma^2 = \left(x_1^2 + x_2^2 + \cdots x_5^2\right)$$

$$\Rightarrow \sigma^2 = \frac{x_1^2 + x_2^2 + \cdots x_5^2}{5}$$

$$\sigma^2 = \frac{1^2 + 3^2 + (-1)^2 + (-3)^2 + (4)^2}{5}$$

$$\sigma^2 = \frac{36}{5} = 7.2$$

Thus the maximum likelihood estimate of the variance of the Gaussian distribution is 7.2

Marking Scheme

2 Marks → Setting up max likelihood expression

3 Marks - differentiation and final calculation

Q.2. Consider a table with a single attribute "wind" and category "rain", where "wind" can take two attribute values – high and low, and "rain" has two classes – yes and no. There are 10 entries in the table, and it is known that 8 entries in the table have wind=high. It is also known that 8 entries in the table also have rain=yes. What is the highest and lowest possible information gain if we split the table on the attribute "wind"? [5 Marks]

For the highest information gain the table looks like this:

Wind	Rain
High	Yes
High	Yes
High	Yes
High	Yes
High	Yes
High	Yes
High	Yes
High	Yes
Low	No
Low	No

$$Gain(S, Wind) = Entropy(S) - \frac{|S_{w=H}|}{|S|} Entropy(S_{w=H}) - \frac{|S_{w=L}|}{|S|} Entropy(S_{w=L})$$

$$Entropy(S) = -p_+ \log p_+ - p_- \log p_- =$$

$$= -\frac{8}{10}\log\frac{8}{10} - \frac{2}{10}\log\frac{2}{10} = 0.721$$

$$Entropy(S_{W=H}) = -\frac{8}{8}\log\frac{8}{8} - \frac{0}{8}\log 0 = 0$$

$$Entropy(S_{W=L}) = \frac{0}{2}\log\frac{0}{2} - \frac{2}{2}\log 1 = 0$$

Highest Information gain : $0.721 - \frac{8}{10}\times 0 - \frac{2}{10}\times 0 = \underline{0.721}$

For the lowest information gain the table looks like this:

Wind	Rain
High	No
High	No
High	Yes
High	Yes
High	Yes
High	Yes
High	Yes
High	Yes
Low	Yes
Low	Yes

In this case

$$Gain(S, Wind) = Entropy(S) - \frac{|S_{W=H}|}{|S|} Entropy(S_{W=H}) - \frac{|S_{W=L}|}{|S|} Entropy(S_{W=L})$$

$$Entropy(S_{Wind=H}) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8} = 0.8075$$

$$Entropy(S_{Wind=L}) = -\frac{2}{2}\log\frac{2}{2} - \frac{0}{2}\log\frac{0}{2} = 0$$

$$Gain(S, Wind) = 0.721 - \frac{8}{10} \times 0.8075 - \frac{2}{10} \times 0$$

$$= 0.075$$

Marking Scheme

Highest Information Gain = 2.5 marks

Lowest Information Gain = 2.5 marks

Q.3. What is the best curve of the form $y = a + bx + cx^2$ in terms of minimizing square error that fits the following data of the form (x, y): (-1,0), (1,10), (2,24), (-2,4)? [5 Marks]

Computing the square loss function gives

$$L = (a - b + c - 0)^2 + (a + b + c - 10)^2 + (a + 2b + 4c - 24)^2 + (a - 2b + 4c - 4)^2$$

Setting $\frac{\partial L}{\partial a} = 0$, $\frac{\partial L}{\partial b} = 0$ and $\frac{\partial L}{\partial c} = 0$

We get

$$4a + 10c = 38$$

$$4a + 16b + 4c = 100$$

$$20a + 68c = 244$$

Final answer: $a = 2, b = 5, c = 3$

Marking Scheme

Setting up loss function = 2 marks

Final calculation = 3 marks

Q.4. There exists a training set consisting of 100 documents for text classification consisting of two types of document '+' and '-'. 75 of the 100 documents are '+' and the remaining are '-'. The total number of words including duplicates in the '+' documents is 150 and the total number of words including duplicates in the '-' documents is 100. The number of words in the vocabulary is 1000. What classification is given to a test text with 5 words consisting of words belonging to the vocabulary but which have not occurred in the training set at all using the Naïve Bayes algorithm for text classification?

[5 Marks]

We have $P(+) = \frac{75}{100} = \frac{3}{4}$

$P(-) = \frac{25}{100} = \frac{1}{4}$

Since the words ω_k in the test document do not occur in the training set, we have $P\left(\frac{\omega_k}{+}\right) = \frac{n_k + 1}{n + |V|} = \frac{0+1}{150+1000}$

$= \frac{1}{1150}$

$P\left(\frac{\omega_k}{-}\right) = \frac{n_k + 1}{n + |V|} = \frac{1}{100+1000} = \frac{1}{1100}$

$P(+) \prod_{k=1}^{5} P\left(\frac{\omega_k}{+}\right) = \frac{3}{4}\left(\frac{1}{1150}\right)^5$

$P(-) \prod_{k=1}^{5} P\left(\frac{\omega_k}{-}\right) = \frac{1}{4}\left(\frac{1}{1100}\right)^5$

$\frac{3}{4}\left(\frac{1}{1150}\right)^5 > \frac{1}{4}\left(\frac{1}{1104}\right)^5$

$\therefore$ we give a +ve classification to the new text.

Marking Scheme

3 marks → setting up $P\left(\frac{w_K}{+}\right)$ and $P\left(\frac{w_K}{-}\right)$

2 marks → final calculation

Q.5. One percent of women over 50 have breast cancer. Ninety percent of women who have breast cancer test positive on mammograms. Eight percent of women will have false positives. What is the probability that a woman has cancer if she has a positive mammogram result? [5 Marks]

Let C denote the event of cancer for women over 50.

$P(C) = 0.01$

$P(M/C) = 0.9$

$P(M/\neg C)$ = probability of showing a positive mammogram but given no cancer = false positive = 0.08

$P(C/M) = ?$

$$P(C/M) = \frac{P(M/C)P(C)}{P(M/\neg C)P(\neg C) + P(M/C)P(C)}$$

$$= \frac{(0.9)(0.01)}{(0.08)(0.99) + (0.9)(0.01)}$$

$$= \frac{0.009}{0.0882} = 0.102$$

Marking scheme:

2 Marks → calculating all the probabilities $P(C), P(\neg C), P(M/C)\ P(M/\neg C)$

3 Marks → Bayes equation

Q.6. Let X_1, X_2 be two real-valued features and Y be a Boolean-valued function of the given features such that the Gaussian Naïve-Bayes assumptions are satisfied. Suppose $P(Y/X_1, X_2) = \frac{1}{1+\exp(-0.1-0.2X_2-0.3X_3)}$. Assume that $P(X_1/Y=0) = N(1.0, \sigma_1)$ and $P(X_1/Y=1) = N(2.0, \sigma_1)$. Similarly $P(X_2/Y=0) = N(1.0, \sigma_2)$ and $P(X_2/Y=1) = N(2.0, \sigma_2)$ Calculate the standard deviations σ_1 and σ_2 and the probability $P(Y=1)$. [5 Marks]

From the formula for logistic regression we see that

$\frac{\mu_{i0} - \mu_{i1}}{\sigma_i}$ = coefficient of X_i in the formula $P(Y=1/X) = \frac{1}{1+\exp(\theta_0 + \Sigma\theta_i X_i)}$

$\therefore \frac{1.0 - 2.0}{\sigma_1^2} = -0.2 \Rightarrow \sigma_1^2 = 5$

$\frac{1.0 - 2.0}{\sigma_2^2} = -0.3 \Rightarrow \sigma_2^2 = \frac{10}{3}$

Also $\ln\left(\frac{1-\pi}{\pi}\right) + \frac{\mu_{11}^2 - \mu_{10}^2}{2\sigma_1^2} + \frac{\mu_{21}^2 - \mu_{20}^2}{2\sigma_2^2}$

= constant term where π = prior prob $= P(Y=1)$

$\ln\frac{1-\pi}{\pi} + \frac{4-1}{2\times 5} + \frac{4-1}{2\times\frac{10}{3}} = -0.1$

$\ln\frac{1-\pi}{\pi} = -0.1 - \frac{3.0}{10} - \frac{9}{20} = -0.85$

$\frac{1-\pi}{\pi} = e^{-0.85}$ $\quad \frac{1}{\pi} = 1 + e^{-0.85}$

$\Rightarrow \pi = \frac{1}{1+e^{-0.85}}$

$\frac{\mu_{i0} - \mu_{i1}}{\sigma_i}$ = Coefficient of x_i in the formula $P(Y=1/x_1, x_2) = \frac{1}{1+\exp(\theta_0 + \Sigma \theta_i x_i)}$

Marking Scheme

2 Marks → getting the formula right

3 Marks → final calculation