Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second Semester 2019-2020
M.Tech (Data Science and Engineering)
Mid-Semester Test (EC-2 Regular)

Course No.	: DSECL ZG565
Course Title	: MACHINE LEARNING
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 90 minutes
Date of Exam	: December 29, 2019 (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.

Answer All the Questions (only on the pages mentioned against questions. if you need more pages, continue remaining answers from page 20 onwards)

Question 1. [Marks 2+1+2=5]   [**to be answered only on pages 3-5**]

a) What are the steps in designing a machine learning system (2 marks)
Solution
Choose the training experience (data)
Choose exactly what is to be learned – i.e. the target function
Choose how to represent the target function
Choose a learning algorithm to infer the target function from the experience

b)A survey was conducted of 200 families to observe the relationship between average annual income per year and whether the family will buy car or not. Consider the following table:

	Income below Rs 10 lakhs	Income >= Rs 10 lakhs	Total
Buyer	38	42	80
Non-Buyer	82	38	120
Total	120	80	200

Solution:
a) What is the probability that a randomly selected family is a buyer of the Car? (1 mark)
  • 80/200 =0.40.
b) What is the probability that a randomly selected family is both a buyer of car and belonging to income of Rs. 10 lakhs and above? (1 mark)
  • 42/200 =0.21.
c) A family selected at random is found to be belonging to income of Rs 10 lakhs and above. What is the probability that this family is buyer of car? (1 mark)
  • 42/80 =0.525. Note this is a case of conditional probability of buyer given income is Rs. 10 lakhs and above.

Question 2. [Marks =5]

Consider there are two bags A and B, where A contains 5 white balls and 7 blue balls whereas B contains 2 white and 12 blue balls. We pick bag A, 50% of the time. After an experiment, a white ball is selected. What is the probability that the ball is drawn from bag B?

Let given the white ball is drawn, the probability that it belongs to bag B, is =p(B|White)

$$p(B|White) = \frac{p(White|B).p(B)}{p(White|B).p(B) + p(White|A).p(A)}$$

$$p(A) = p(B) = \frac{1}{2}; p(White|B) = \frac{2}{14}; p(White|A) = \frac{5}{12}$$

$$p(B|White) = \frac{0.143 \, X \, 0.5}{[0.143 + 0.417]X \, 0.5} = \frac{0.143}{0.56} = 0.2553$$

Question 3. [Marks =5]

Given the following labelled training data,
Flat 20% Cashback on Oyo Room bookings done via Paytm. (SPAM)
Lets Talk Fashion! Get flat 40% Cashback on Backpacks (SPAM)
Opportunity with Product firm for Fullstack (HAM)
Javascript Developer, Full Stack Developer in Bangalore (HAM)
Use Naive Bayes Classifier with laplace smoothing to identify classification of the sentence "Scan Paytm QR Code to Pay & Win 100% Cashback"
Solution:
Number of unique words across spam and ham - 28
Number of words in Spam – 18 , Number of words in ham - 13

word	P(word/spam)	P(word/ham)
Scan	(0+1)/(18+28)	(0+1)/(13+28)
PayTM	(1+1)/(18+28)	(0+1)/(13+28)
QRCode	(0+1)/(18+28)	(0+1)/(13+28)
to	(0+1)/(18+28)	(0+1)/(13+28)
Pay	(0+1)/(18+28)	(0+1)/(13+28)
&	(0+1)/(18+28)	(0+1)/(13+28)
Win	(0+1)/(18+28)	(0+1)/(13+28)
100%	(0+1)/(18+28)	(0+1)/(13+28)
Cashback	(2+1)/(18+28)	(0+1)/(13+28)

P (Spam) = 2/4 = 0.5, P(Ham) = 2/4 = 0.5
P (Scan Paytm QRCode to Pay & Win 100% Cashback / Spam)
=0.5*1/46*2/46*1/46*1/46*1/46*1/46*1/46*1/46*3/46 =3*10^-15
P (Scan Paytm QRCode to Pay & Win 100% Cashback / Ham) =0.5*(1/41)^9 = 2*10^-15
Since P (sentence/spam)>P(sentence/ham), the given sentence belongs to spam category.

Question 4. [Marks 5]
a) Explain the cost/error function used in logistic regression   (2 marks)
b) Compare Probabilistic generative model and probabilistic discriminative models with examples.
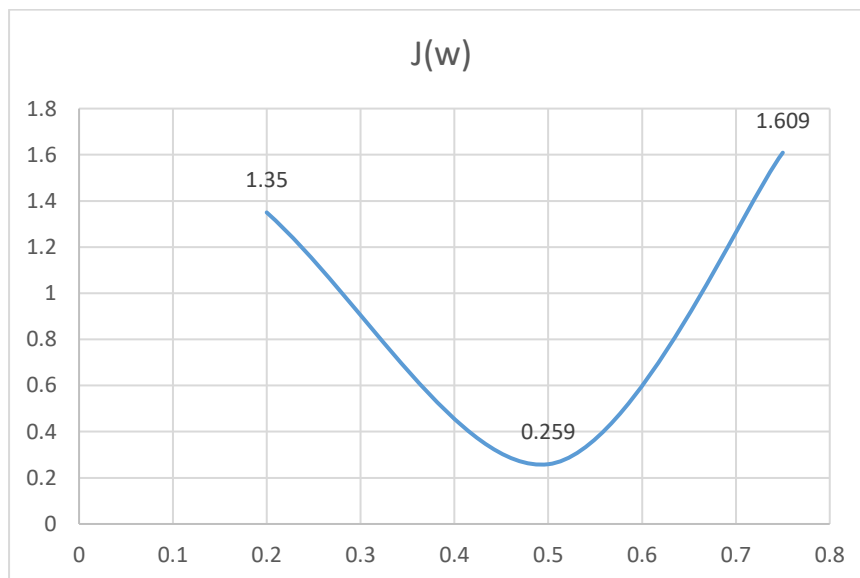   (3 marks)

Question 5. [Marks 3+2=5]
a) Plot cost function J (w) for linear regression $y=w_1x$ for the training data pair <0, 0>,
    <0.5, 0.5>, <1, 1>, <1.5, 1.6> for different values of w   (3 marks)

$$J(w) = \sum_i (wx_i - y_i)^2$$

$$= 0^2 + (0.5w - 0.5)^2 + (w - 1)^2 + (1.5w - 1.6)^2$$

w1	J(w)
0.2	1.35
0.5	0.259
0.75	1.609

b) Distinguish Bias and variance in the machine learning domain and discuss how model
    complexity is affected by these two.   (2 marks)
        Solution: Bias is error resulting from assumptions in the model and it could be seen as the
        difference between the expected value of the estimate and the expected value of the actual
        function. Variance is the change in the estimate of the target function for changes in data.
        Use the bias-variance trade-off with respect to model complexity to interpret their effect.

Question 6. [Marks = 5]   [to be answered only on pages 15-16]

Provide answers based on the following set of training examples

Instance	a1	a2	Classification
1	T	T	+
2	T	T	+
3	T	F	-
4	F	F	+
5	F	T	-
6	F	T	-

(a) What is the entropy of this collection of training examples with respect to the target function classification   (2 marks)
(b) What is the information gain of a2 relative to these training examples (1 marks)
(c) Why do we prefer shorter /smaller trees while learning decision tree? Does ID-3 guarantee shorter tree? (2 marks)

Solution
Entropy (S) = -P(+)logP(+)- P(-)logP(-)
= -3/6log 3/6-3/6log3/6 = 1

Gain (s, a2) = Entropy (S) - ∑|Sv|/|S| * Entropy (Sv)
= 1- 4/6 Entropy (S a2=T)-2/6 Entropy (S a2=F)
= 1- 4/6 - 2/6 =0

We prefer shorter trees to avoid overfitting
ID3 does not guarantee shorter tree. This is one of the short coming and techniques like pruning help us in getting the right fit.