**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

No. of Pages = 2

# No. of Questions = 6

Course Title : MACHINE LEARNING

Nature of Exam : Closed Book

Weightage : 30%

Duration : 90 minutes

Date of Exam : December 29, 2019 (FN)

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+3=5] **[to be answered only on pages 3-5**]

a) What are the steps in designing a machine learning system (2 marks)

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 |

1. What is the probability that randomly selected family is a buyer? (1 marks)
2. What is the probability that a randomly selected family is both buyer of the car and has income of Rs 10 lakh and above? (1 mark)
3. A family selected at random belongs to the category of income greater than Rs 10 lakhs. What is the probability that they will buy a car? (1 marks)

**Question 2.** [Marks =5] **[to be answered only on pages 6-7**]

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? (5 marks)

**Question 3.** [Marks=5] **[to be answered only on pages 8-9]**

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”

**Question 4.** [Marks 2+3=5] **[to be answered only on pages 10-11**]

1. Explain the cost/error function used in logistic regression (2 marks)
2. Compare Probabilistic generative model and probabilistic discriminative models with examples. (3 marks)

**Question 5.** [Marks 3+2=5] **[to be answered only on pages 12-14**]

1. Plot cost function J (w) for linear regression y=w1x for the training data pair <0, 0>,

<0.5, 0.5>, <1, 1>, <1.5, 1.6> (3 marks)

1. Distinguish Bias and variance in the machine learning domain and discuss how model complexity is affected by these two. (2 marks)

**Question 6.** [Marks 2+1+2= 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 | - |

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