Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
Second Semester 2018-2019
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	: August 11, 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 3+1=4] **[to be answered only on pages 3-5]**

Derive the equation and shape of decision surface for real-valued random variable **X**=<X_1, X_2, .., X_n> and boolean output Y for logistic regression. P(Y=1|**X**) is given by

$$P(Y=1|X=<X_1,...X_n>)=\frac{1}{1+exp(w_0+\sum_i w_i X_i)}$$

Question 2. [Marks 4+1=5] **[to be answered only on pages 6-7]**

Consider the hypothesis function $h(\mathbf{x}) = w_0 + w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_2^2$ with learnt $\mathbf{w} =< w_0, w_1, w_2, w_3, w_4 >=< 36, 0, 0, 4, 9 >$. What is the equation and shape of the decision boundary $g(x_1, x_2)$ for logistic regression given by $P(Y = 1|\mathbf{X}) = \frac{1}{1+e^{h(\mathbf{x})}}$

Question 3. [Marks 1+1+2.5+2.5+1=8] **[to be answered only on pages 8-10]**

First five documents in the following figure are used to train a Naive Bayes classifier. Calculate Prob(+), Prob(-1), Prob(+ | Test), Prob(- | Test) for the bag of words model. Which class does the Test document belong to?

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no originality

Question 4. [Marks 5] **[to be answered only on page 11]**

a) Consider a classification model with logistic regression and L2 regularization. Assuming that model is suffering from the problem of over-fitting, decreasing the value of regularization parameter helps in reduction of over-fitting. **True or False**

b) In the case of large feature space, Naïve Bayes algorithm outperforms logistic regression. **True or False**

c) Gaussian Naive Bayes classifier can have linear decision surface. **True** or **False**

d) Bagging is used in decision tree to reduce bias. **True** or **False**

e) What techniques can be used to reduce overfitting in Decision tree? i) **Pruning** ii) **Bagging** iii) **Feature Randomization.** Choose all that apply.

Question 5. [Marks 1+3=4] **[to be answered only on pages 12-14]**

a) A coin is tossed 250 times and lands heads 50 times. What is the maximum likelihood estimate for θ = probability of heads?

b) A 6-sided die is rolled 16 times resulting in 2 ones, 4 twos, 0 threes, 5 fours, 2 fives, 3 sixes. What is the maximum likelihood estimate for all values of θ_i where i is <1,2,3,4,5,6> for each side of the die?

Question 6. [Marks 1+1+2=4] **[to be answered only on pages 15-16]**

Draw the decision boundary (shape and position w.r.t. training points labelled as class A, B, and C in the figure below) for Decision Tree, Logistic Regression and Gaussian Naïve Bayes (different means and different variances for different classes) classifiers.