National Institute of Technology Calicut

Department of Computer Science and Engineering

S6 - B. Tech. (CSE)/S4-MCA

Machine Leaning (CS4044D)

End Semester Exam-Online (Winter 2019-20)

Total Time: 2 Hrs Max Marks: 20

Notes:

- All questions are compulsory.
- Total time is **Two Hours** which includes downloading the question paper, writing answers, taking snapshots of answer sheets, making a single PDF for all answers, and uploading PDF**.**
- Write your name and roll number in your answer sheets.
- If any plagiarism is found in answer scripts, strict action will be taken against respective group of students by concerned institute committee.

1. You decided to purchase a personal computer (C) online from Flipkart. You take decision based on K-Nearest neighbour classifier. Training data set is available in the following table in which each data point (each computer) has two features; first, grade of computer based on user review (P1) and second, price of the computer (P2). What will be the decision if one computer has P1=5 and P2=2. Follow step-by step procedure to decide. Assume that number of nearest neighbour, K=04. **[05 Marks]**

Computer	P1	P2	Decision
C1	6	7	Yes
C2	8	5	Yes
C3	4	6	No
C4	10	9	No
C5	9	7	Yes

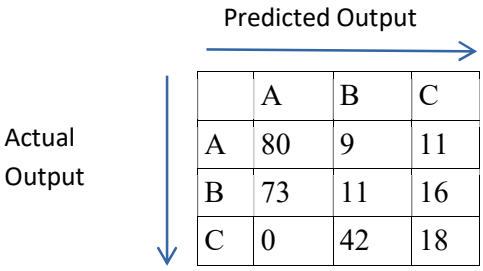
2. (a) You train a logistic regression algorithm, and find that it has unacceptably high error on the test set. However, your hypothesis performs well (has low error) on the training set. What are possible steps you should take to resolve mentioned issues? **[2.5 Marks]**

(b) If above scenario acts in reverse manner then what are possible steps you should take in this case? Write answers precisely, clearly, and point wise**. [2.5 Marks]**

3. "The Indian Rebellion of 1857 began as a mutiny of sepoys of the East India Company's army on 10 May 1857, in the town of Meerut, and soon escalated into other mutinies and civilian rebellions largely in the upper Gangetic plain and central India, with the major hostilities confined to present-day Uttar Pradesh, Bihar, northern Madhya Pradesh, and the Delhi region." Write the output for each text preprocessing task as mentioned below on the above sentence. **[4 marks]**

i. Sentence Tokenization
ii. Word Tokenization
iii. Stopwords removal
iv. Pos tagging
v. Chunking

4. Calculate **Accuracy** and **F1 score** for the given confusion matrix: **[6 marks]**



	A	B	C
A	80	9	11
B	73	11	16
C	0	42	18