


National University of Computer and Emerging Sciences, Lahore Campus

	Course Name:	NLP	Course Code:	CS4063
	Degree Program:	BS-CS	Semester:	Spring 2023
	Exam Duration:	60 Minutes	Total Marks:	80
	Paper Date:	10-04-2023	Weight	10
	Sections:	ALL	No of Page(s):	2
	Exam Type:	Midterm II		

Student : Name: _____ Roll No. _____ Section: _____

Instruction/Notes: Attempt all questions. *Do not use pencil or red ink to answer the questions.* Attempt all questions. Programmable calculators are not allowed.
Solve Q3 on the question paper and others on the answer sheet.

Q1: In a population of 100, a medical test identifies 55 sick individuals. Out of these, 33 are actually sick. Of the remaining 45 individuals, 37 are also sick. Please fill in the following blanks: (mention formulas where required) (14)

- a. True positives _____ 33
- b. False positives _____ 22
- c. True negatives _____ 8
- d. False negatives _____ 37
- e. Accuracy _____ $= (tn+tp)/(tn+tp+fn+fp) = (33+8)/100 = 0.41$
- f. Precision _____ $= tp/tp+fp = 33/55 = 0.6$
- g. Recall _____ $= tp/(tp+fn) = 33/(33+37) = 33/70 = 0.47$

Q2: You are given the output confusion matrix for a 3-class email classification task. Compute **micro-averaged** and **macro-averaged**, **precision** and **recall**. Show your work for partial credit, mention the formulas too. (5x5)

		Gold Labels		
		urgent	normal	spam
System Output	urgent	9	12	2
	normal	5	45	40
	spam	23	33	300

- i. Micro-averaged precision = 0.754797
- ii. Macro-averaged precision = 0.578
- iii. Micro-averaged recall = 0.754797
- iv. Macro-averaged recall = 0.540145
- v. Which one of the two measures (micro vs macro) reflected the true picture in this case and why?
MACRO

Q3. Suppose you are working on a speech recognition system that is designed to recognize three different phonemes: "aah", "eeh", and "ooh". You have a training dataset of recorded speech samples that have been labeled with the corresponding phoneme for each sample. However, the transitions between phonemes in the speech samples are not well defined, and there is some overlap between the phonemes in some of the samples. To model the transitions between phonemes and the probabilities of observing each phoneme in the dataset, you want to use the Baum-Welch algorithm to estimate the A and B matrices for a hidden Markov model (HMM).

Here is some additional information about the problem:

The training dataset consists of 100 speech samples, with an average length of 50 phonemes per sample. The observations are discrete, with each phoneme represented by a vector of 10 features. You have already determined the number of hidden states (N) in the HMM, which is 5.

For this problem, how would you initialize the A and B matrices for the Baum-Welch algorithm? What factors would you consider when choosing the initial values for the matrices? (10)

A: 5x5 (random, all ones, specify the underlying structure of the data and then decide, uniform distribution)

B: 5x10x50 (random, row sum should be 1 for each state)

Therefore, the B matrix needs to have 5 rows (one for each hidden state) and 10 columns (one for each feature of the phoneme representation, and 50 phonemes per sample).