COMSATS University Islamabad, Lahore Campus

Mid Term Examination – Semester Spring 2022

Course Title:	Introduction to Data Science			Course	CSC465	Credit	3
Course	Mr. Muhammad Shahid Bhatti			Program Name:	BS(CS)		
Semester:	6, 7th	Batch:		Section:		Date	13-May-2022
Time Allowed:	**90 Minutes**			**Maximum Marks:**			**40**
Student's Name:				Reg. No.			

Important Instructions / Guidelines:

- All parts are compulsory.

Question 1: (C-3) [Marks 05]

1. Suppose we have the continuous-valued attribute "Monthly Income" and a class label "Fashion Conscious" with values Yes and No. Training samples are given below: Monthly Income in thousands – (Fashion Conscious):

   35(Y) 150(N) 105(N) 3(N) 15(Y) 25(Y) 7(N) 75(Y) 101(N) 50(Y) 9(N)

   Convert the continuous-valued attribute into

   (a) a Boolean-valued attribute
   (b) a three-valued discrete attribute

Question 2: (C-3) [Marks 10]

We will use the dataset below to learn a decision tree which predicts if student pass data science (Yes or No) based on their previous GPA (High, Medium, or Low) and whether or not they studied.

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

For this problem, you can write your answers using $\log_2$, but it may be helpful to note that $\log_2 3 \approx 1.6$.

A. What is the entropy H(Passed)?

B. What is the entropy H(Passed | GPA)?

C. What is the entropy H(Passed | Studied)?

D. Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations.

Question 3: (C-3) [Marks 10]

Imagine you wish to recognize good and bad items produced by your company. You're able to measure three numeric properties of each widget: *P1*, *P2*, and *P3*. You randomly grab several items off of your shipping dock and extensively test whether or not they are good, obtaining the following results:

P1	*P2*	*P3*	*Result*
0.0	0.2	0.8	good
9.2	0.7	1.5	bad
4.9	0.1	2.9	good
2.7	5.3	6.2	bad

Scanned with CamScanner

2.4	0.0	3.7	good

What a *three-nearest neighbour (3-NN)* algorithm would classify the following new example.

P1 = 6.3 *P2 = 5.1* *P3 = 0.4*

Question 4: (C-1) [10 Marks]

Write short answers to the following questions.

A. What is Hadoop? Write the names of its core components and two to three lines of their description?

B. How does the Map-Reduce algorithm work to solve a big computational task?

C. What to consider when choosing large data file formats? Write the names of the two best formats. Why are these best?

csv
binary.

Question 5: (C-3) [5 Marks]

Logistic regression is a supervised learning algorithm that predicts a dependent categorical target variable. Describe the cost function of logistic regression and how it penalizes the inaccurate values.

Scanned with CamScanner