UNIVERSITI MALAYA

EXAMINATION FOR THE DEGREE OF MASTER OF DATA SCIENCE

ACADEMIC SESSION 2024/2025 : SEMESTER I

WQD7006 : Machine Learning for Data Science

Jan 2025 Time : 2 hours

INSTRUCTIONS TO CANDIDATES :

*Answer **ALL** questions. (50 marks)*

(Kertas soalan ini mengandungi 3 soalan dalam 4 halaman yang bercetak)
(This question paper consists of 3 questions on 4 printed pages)

Online assessment/30 marks – 1.5 hours
Submission through Spectrum. Specific instructions to be given on Spectrum as well. Tentatively Week 12.

Question 1: 12 marks

1. (a) Liam participates in a cooking competition where he can prepare one of three dishes: pasta, sushi, or salad. The probabilities that he will win based on the dish he prepares are as follows:

   - If he makes pasta, there is a 40% chance he will win.
   - If he makes sushi, there is a 25% chance he will win.
   - If he makes salad, there is a 10% chance he will win.

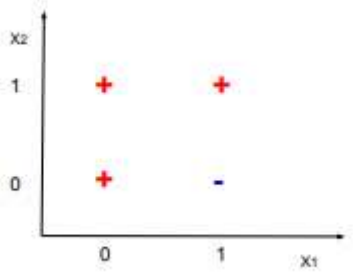
   The prior probabilities for each dish are assumed to be equal: P(Pasta)=P(Sushi)=P(Salad)=1/3

   i. Liam wins the competition one day. What is the probability that he made pasta? Use Bayes' theorem to solve the problem. (4 marks)

   ii. Now suppose Liam's friend knows that he usually makes sushi, never makes pasta, and 20% of the time makes salad. What is the probability that Liam made salad that day, given he won?

   (3 marks)

   (b) We are interested in predicting whether a person makes over 50K a year, and we model the two features with two boolean variables X1, X2 ∈ {0,1}, and label Y ∈ {0,1} where Y = 1 indicates a person makes over 50K. Figure below shows three positive samples ("+" for Y = 1) and one negative sample ("-" for Y = 0). Answer the following questions:



i. For the above scenario, which model would be better in predicting: Linear or Logistic Regression? Why? (2 marks)

ii. Is there any Logistic Regression classifier using X1 and X2 that can perfectly classify the examples in the figure above? Explain. (2 marks)

iii. If we change the label of point (0,1) from "+" to "-", will there be a perfect Logistic Regression classifier? (1 mark)

Question 2: 12 marks

A company is deciding whether to hire a candidate based on their qualifications (High, Medium, Low) and their interview performance (High, Low). The data collected from past candidates is as follows:

Qualifications	Interview Score	Hired
High	High	Yes
High	Low	Yes
Low	High	No
Low	Low	No
Medium	High	Yes
Medium	Low	No

(a) Calculate the entropy of the dataset. (2 marks)

(b) Calculate the entropy for Qualifications and Interview Score. (6 marks)

(c) Determine which feature has the highest information gain and is the best to split on (2 marks)

(d) Draw the decision tree based on your calculations. (2 marks)

Question 3: 6 marks

You are given the following distance matrix:

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

(a) Perform hierarchical clustering using single link technique. Show the distance matrix at each step and draw the final dendogram. (6 marks)

THE END