

Programme Code: TU082
Module Code: CMPU 4011

TECHNOLOGICAL UNIVERSITY DUBLIN

City Campus - Grangegorman

**TU082 – BSc. (Honours) in Information Systems and
Information Technology**

Year 4

SEMESTER 2 EXAMINATIONS 2024/25

CMPU 4011 Machine Learning for Predictive Analytics

Internal Examiner(s):

Dr. Bojan Božić
Dr. Paul Doyle

External Examiner(s):

Dr. Andrea Kealy

Instructions To Candidates:

Answer ALL questions.

Question 1 carries 40 marks and questions 2 and 3 carry 30 marks each.

Exam Duration: 2 hours

Special Instructions/Handouts: Available on the last page

1)

- a) **(5 marks)** Explain what overfitting is in machine learning. Provide an example to illustrate.
- b) **(5 marks)** What is the difference between restriction bias and preference bias? Give an example of each.
- c) **(5 marks)** Discuss what could go wrong if a machine learning model assumes feature independence when in reality the features are correlated. Give an example to support your answer.
- d) The table below is showing a model's predictions on whether an email is "spam" (true) or "ham" (false).

Actual / Predicted	Spam (True)	Ham (False)
Spam (True)	50	10
Ham (False)	5	35

Based on this table:

- i) **(6 marks)** Construct a **confusion matrix**.
- ii) **(4 marks)** Calculate **classification accuracy**.
- iii) **(15 marks)** Calculate **precision, recall and F1-score**. Explain what each metric means in this context.

2)

- a) The following table contains customer satisfaction data for two features: **Service Quality** and **Response Time**, labeled as "Satisfied" or "Unsatisfied."

Service Quality	Response Time	Satisfaction
5	10	Satisfied
3	20	Unsatisfied
4	15	Satisfied
2	25	Unsatisfied
5	12	Satisfied

Classify the following query using 3-Nearest Neighbour (3-NN) with Euclidean Distance.

Service Quality	Response Time
4	18

- i) **(8 marks)** Predict the **Satisfaction** label for the query using 3-NN. Show calculations and justification
 - ii) **(4 marks)** Normalize **Service Quality** and **Response Time** using **range normalization** (4 decimal places). Explain why normalization is important in k-NN.
 - iii) **(8 marks)** Repeat part (i) using the **normalized dataset**. Explain if and how the result changes.
- b) The next table shows decisions of a bank on whether to grant loans based on four features.

Credit Score	Income Level	Previous Loan	Age Group	Loan Granted
High	High	No	Adult	Yes
Medium	Medium	Yes	Senior	No
Low	Low	No	Young	No
High	Medium	Yes	Adult	Yes

- i) **(5 marks)** Calculate the **entropy** of the dataset. Explain what the value tells you.
- ii) **(5 marks)** Explain what **entropy** indicates about a dataset and provide 2 real-world examples of datasets with high and low entropy.

3)

- a) The following table contains a book purchase dataset.

Genre	Author known	Discounted	Purchased
Fiction	Yes	Yes	Yes
Non-fiction	No	Yes	No
Fiction	Yes	No	Yes
Poetry	No	No	No
Fiction	No	Yes	Yes

- i) **(18 marks)** Calculate Naïve Bayes probabilities for each feature given target values (4 decimals). Explain why Naïve Bayes is appropriate.
- b) **(10 marks)** Given a new book with:

Genre	Author Known	Discounted
Fiction	Yes	No

Calculate $P(\text{Purchased}=\text{yes})$ and $P(\text{Purchased}=\text{No})$.

- c) **(2 marks)** What would Naïve Bayes predict for this book? Explain how to interpret these probabilities.

Supplemental Material

Useful Formulae

The entropy of a dataset of examples with labels t_1, t_2, \dots, t_l

$$H(D) = - \sum_{i=1}^l (P(t_i) \times \log_2(P(t_i)))$$

Information Gain for descriptive feature d that splits D into partitions D_1, D_2, \dots, D_k .

$$IG(d, D) = H(D) - \sum_i^k \frac{|D_i|}{|D|} \times H(D_i)$$

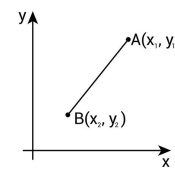
Table of base 2 log for different fractions

log ₂ (a/b)		a													
		1	2	3	4	5	6	7	8	9	10	11	12	13	14
b	1	0.00													
	2	-1.00	0.00												
	3	-1.58	-0.58	0.00											
	4	-2.00	-1.00	-0.42	0.00										
	5	-2.32	-1.32	-0.74	-0.32	0.00									
	6	-2.58	-1.58	-1.00	-0.58	-0.26	0.00								
	7	-2.81	-1.81	-1.22	-0.81	-0.49	-0.22	0.00							
	8	-3.00	-2.00	-1.42	-1.00	-0.68	-0.42	-0.19	0.00						
	9	-3.17	-2.17	-1.58	-1.17	-0.85	-0.58	-0.36	-0.17	0.00					
	10	-3.32	-2.32	-1.74	-1.32	-1.00	-0.74	-0.51	-0.32	-0.15	0.00				
	11	-3.46	-2.46	-1.87	-1.46	-1.14	-0.87	-0.65	-0.46	-0.29	-0.14	0.00			
	12	-3.58	-2.58	-2.00	-1.58	-1.26	-1.00	-0.78	-0.58	-0.42	-0.26	-0.13	0.00		
	13	-3.70	-2.70	-2.12	-1.70	-1.38	-1.12	-0.89	-0.70	-0.53	-0.38	-0.24	-0.12	0.00	
	14	-3.81	-2.81	-2.22	-1.81	-1.49	-1.22	-1.00	-0.81	-0.64	-0.49	-0.35	-0.22	-0.11	0.00

Euclidean distance:

Bayes Theorem:

Distance Formula


$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$