MAASAI MARA UNIVERSITY

2021/2022 ACADEMIC YEAR

FOURTH YEAR FIRST SEMESTER

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

SCHOOL OF PURE, APPLIED, AND HEALTH

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

COURSE CODE: COM 4213

COURSE TITLE: MACHINE LEARNING

DATE: 6TH APRIL 2022 TIME:2:30PM-4:30PM

INSTRUCTIONS TO CANDIDATES:

ANSWER ALL QUESTIONS IN SECTION A AND ANY 2 QUESTIONS IN

SECTION B

SECTION A: COMPULSORY [30 MARKS]

QUESTION ONE

i. Define the term "*machine learning*" **[2 Marks]**

ii. State the main **Four** machine learning techniques **[2 Marks]**

iii. Discuss any **Four** applications of machine learning **[8 Marks]**

iv. Consider the following table 1 below for SMS text classification.

[1] Draw word occurrence table after applying the feature extraction process **[2 Marks]**

[2] **DERIVE** Naïve Bayes probabilistic algorithm. **[4 Marks]**

[3] Demonstrate using the word occurrence table you derive in [1] whether the incoming unknown (Ham/Spam) SMS using the Naïve Bayes classification. **[6 Marks]**

v. Consider table 2 below.

[1] Define the term Information theory. **[2 Marks]**

[2] Derive the entropy (D) formula from the load data to be used

[3] Using the derived entropy calculate using loan data the best attribute (root node) to be used in decision tree. **[2 Marks]**

[4] Derive the information gain (G) to select the best attribute to partition the data **[2Marks]**

QUESTION TWO [20 MARKS]

i. Distinguish between ***Supervised Learning*** and ***Unsupervised learning*** as used in machine learning. **[4 Marks]**

ii. Discuss the following algorithms of unsupervised learning:

[1] K-Means Clustering **[4 Marks]**

[2] Hidden Markov model **[4 Marks]**

iii. Define the following terms in relation to Logistic Regression

a. Posterior Distribution **[2 Marks]**

b. Multiclass Classification **[2 Marks]**

iv. Derive the k-Means algorithms **[4 Marks]**

SECTION B: ANSWER ANY TWO QUESTION [**40 MARKS**]

QUESTION THREE [20 MARKS]

i. Define overfitting as used in classification**.** **[2 Marks]**

ii. Consider the text below:

"As the home to UVA's recognized undergraduate and graduate degree programs in systems engineering. In the UVA Department of Systems and Information Engineering, our students are exposed to a wide range of range"

[1] GENERATE Bag-of-Words (BoW) model from the sentence, and **[2 Marks]**

[2] Derive the Bag-of-Feature (BoF) to generate vector space. **[2 Marks]**

iii. Text classification system contains four different levels of scope that can be applied, Discuss. **[4 Marks]**

iv. Discuss any **TWO** techniques of feature extractions. **[4 Marks]**

v. **Discuss** the Principal Component Analysis (PCA) model. **[2 marks]**

QUESTION FOUR [20 MARKS]

i. Discuss Support Vector Machine (SVM). **[4 Marks]**

ii. Assume that we want to estimate the distribution of weights of a population. Sample data from a population might look as follows: X = {57, 88, 54, 84, 83, 59, 56, 43, 70, 63, 90, 98, 102, 97, 106, 99, 103, 112}.

   a. Transform X into a realistic estimate of the density p(x). Starting with a 'density estimate' with only discrete terms. **[6 Marks]**

iii. Consider a binary classification task, where we are given a training set {(x1,y1),...,(xm,ym)} with xi ∈ H and yi ∈ {±1}.
find a linear decision boundary parameterized by (w,b) such that hw,xii+b ≥ 0 **[10 Marks]**

QUESTION FIVE [20 MARKS]

v. Consider the following of a spam e-mail Example

x1: The quick brown fox jumped over the lazy dog.

x2: The dog hunts a fox. The dog hunts a fox.

   a. Provide the Vector space representation of strings **[5 Marks]**

   b. Derive Naive Bayes classifier **[5 Marks]**

   c. Derive k-Nearest Neighbor Classifier **[5 Marks]**

   d. Derive perceptron classifier **[5 Marks]**

TABLE I: VECTOR TABLE

SMS No.	Type	Word attributes		
		Good	Very	bad
1	Ham	1	0	0
2	Ham	1	1	0
3	Spam	0	0	1
4	Spam	0	1	1
5	Spam	0	2	2

Table 1: vector table

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

Table 2: loan application data table

/////END/////