

RS



(Knowledge for Development)  
**KIBABII UNIVERSITY**

**UNIVERSITY EXAMINATIONS  
2021 /2022 ACADEMIC YEAR**

**END OF SEMESTER EXAMINATIONS  
YEAR FOUR SEMESTER TWO EXAMINATIONS  
FOR THE DEGREE OF BACHELOR OF SCIENCE  
COMPUTER SCIENCE**

**COURSE CODE : CSC 423  
COURSE TITLE : MACHINE LEARNING**

**DATE: 30 /08/2022**

**TIME: 9:00 A.M – 11:00 A.M**

---

**INSTRUCTIONS TO CANDIDATES**

**ANSWER QUESTIONS ONE AND ANY OTHER TWO.**

**QUESTION ONE (COMPULSORY) [30 MARKS]**

- a)
- i. Describe the Machine Learning and Draw appropriate diagram [4 Marks]
  - ii. Suppose that the Naïve Bayes algorithm is used to classify an instance described by a vector of  $n$  attributes  $(A_1; A_2; \dots; A_n)$  into one of two categories,  $C_1$  and  $C_2$ . Given an unclassified instance described by the vector  $(x_1; x_2; \dots; x_n)$ , the Naïve Bayes algorithm computes a score for each category and returns the category with the highest score. The Naïve Bayes learning algorithm analyses its training data to obtain a set of conditional probabilities for computing the scores. Explain what probabilities the learning algorithm obtains from the training data, and how it computes the two scores for a new instance. [6 Marks]
  - iii. Differentiate between supervised learning and reinforcement learning [4 Marks]
  - iv. Describe the importance of data processing [2 Marks]
- b)
- i. State and Explain the **THREE** common filter methods used in feature selection [6 Marks]
  - ii. Describe Model Validity [4 Marks]
  - iii. Describe hyperparameters of a model [2 Marks]
  - iv. Describe the disadvantages of grid search [2 Marks]

**QUESTION TWO [20 MARKS]**

- a) Differentiate between validation set and Test set [2 Marks]
  - b) Describe the Classification of Ensemble Learning [4 Marks]
- c) A Suppose we train a model to predict whether an email is Spam or Not Spam. After training the model, we apply it to a test set of 500 new emails (also labeled) and the model produces the following contingency table.

		True Class	
		Spam	Not Spam
Predicted Class	Spam	70	30
	Not Spam	70	330

i. Compute the precision of this model with respect to the Spam class.

[4Marks]

ii. Compute the recall of this model with respect to the Spam class

[4Marks]

d) Suppose we have two users (Emily and Simon) with the following preferences. Emily hates seeing spam emails in her inbox! However, she doesn't mind periodically checking the "Junk" directory for genuine emails incorrectly marked as spam. Simon doesn't even know where the "Junk" directory is. He would much prefer to see spam emails in his inbox than to miss genuine emails without knowing! With reasons, explain which user is more likely to be satisfied with this classifier?

[6 Marks]

### QUESTION THREE [20 MARKS]

a) Determine if the following statements are TRUE/ FALSE

[6Marks]

- i. Artificial intelligence is a new technical science that studies and develops theories, methods and application systems for simulating, extending and extending human intelligence. It is one of the core research areas of machine learning.
- ii. The word recognition in the speech recognition service refers to the synchronous recognition of short speech. Upload the entire audio at once, and the recognition result will be returned in the response
- iii. Self-encoder is an unsupervised learning algorithm
- iv. Loss function and model function are the same thing.
- v. The commonly used functions for mathematical operations in Python are basically in the math module and the cmath module.
- vi. The Python dictionary is identified by "{}", and the internal data consists of the key and its corresponding value.

- b) Write a code in Python to show how splitting is achieved **[4 Marks]**
- c) State and Explain the Common kernel functions **[4 Marks]**
- d) Describe Unsupervised learning and Give Examples of Unsupervised learning **[6 Marks]**

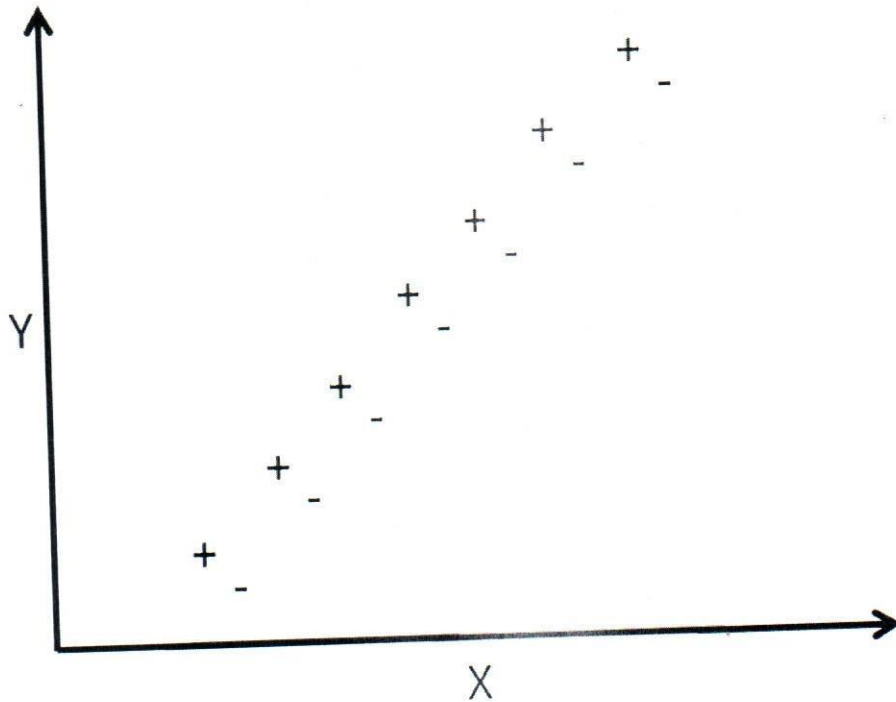
#### **QUESTION FOUR [20 MARKS]**

The goal in predictive analysis is to use training data to learn a model that can make predictions on new data. Answer the following questions a) and b).

- a) Suppose we increased the size of the training set. Would this likely improve or deteriorate the performance of the model on new data? Why? **[5 Marks]**
- b) Suppose we reduced the feature representation to include only the features with the highest mutual information with the target concept. Would this likely improve or deteriorate the performance of the model on new data? Why? **[5 Marks]**
- c) Describe Semi-Supervised learning **[4 Marks]**
- d) Describe dirty data **[3 Marks]**
- e) Describe the Procedure of a filter method **[3 Marks]**

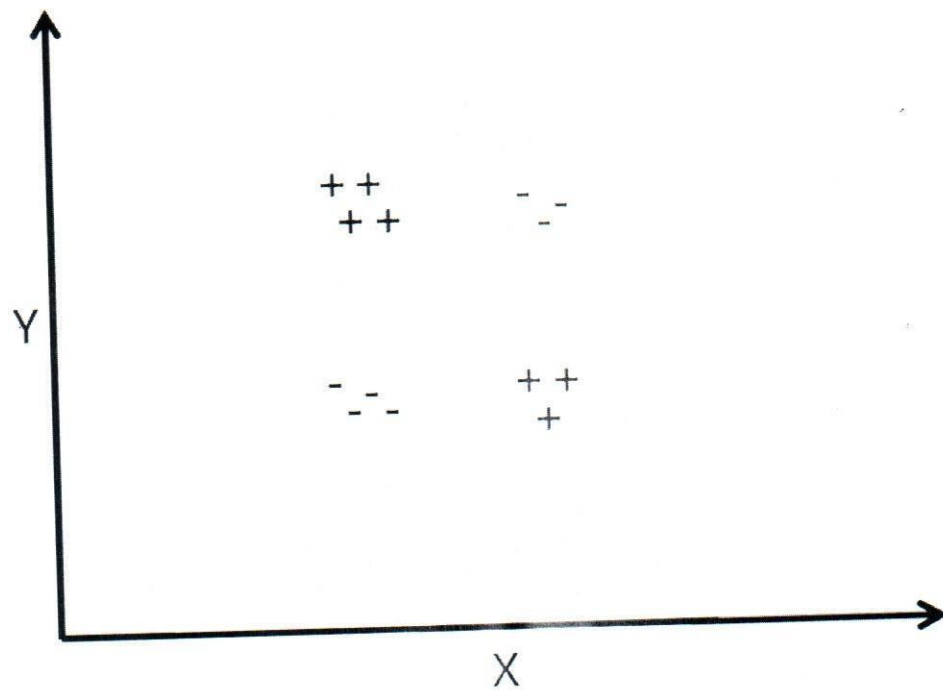
**QUESTION FIVE [20 MARKS]**

- (a) Suppose we have the following data, represented using two real-valued features (X and Y, as in Question 2) and suppose that our goal is to randomly split this data into a training set (90%) and a test set (10%) and to train and evaluate a model.



Which classifier do you think would have a higher chance of doing well in terms of accuracy: KNN (with  $K=1$ ) or Naïve Bayes? Explain with reasons. **[7.5 Marks]**

- b) Suppose we have the following data, represented using two real-valued features (X and Y, as in Question 2) and suppose that our goal is to randomly split this data into a training set (90%) and a test set (10%) and to train and evaluate a model.



Which classifier do you think would have a higher chance of doing well in terms of accuracy: KNN (with  $K=1$ ) or Naïve Bayes? Explain with reasons. **[7.5 Marks]**

c) Describe the Decision Tree Construction Process **[5 Marks]**