30

(Knowledge for Development)

KIBABII UNIVERSITY

UNIVERSITY EXAMINATIONS
2020/2021 ACADEMIC YEAR

SPECIAL/SUPPLEMENTARY EXAMINATIONS
YEAR FOUR SEMESTER TWO EXAMINATIONS

FOR THE DEGREE OF BACHELOR OF SCIENCE COMPUTER SCIENCE

COURSE CODE : CSC 423

COURSE TITLE : MACHINE LEARNING

DATE: 13/01/2022 **TIME:** 02:00 P.M – 0:00 P.M

INSTRUCTIONS TO CANDIDATES

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

QUESTION ONE (COMPULSORY) [30 MARKS]

a)

i. Describe the Rational Understanding of Machine Learning Algorithms **[4 Marks]**

ii. Write a python code to show how Loading the Data Set, Viewing Data Attributes, and Visualizing the Data can be achieved **[6 Marks]**

iii. Differentiate between supervised learning and unsupervised learning **[4 Marks]**

iv. Distinguish between Training set and Testing Set **[2 Marks]**

b)

i. 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

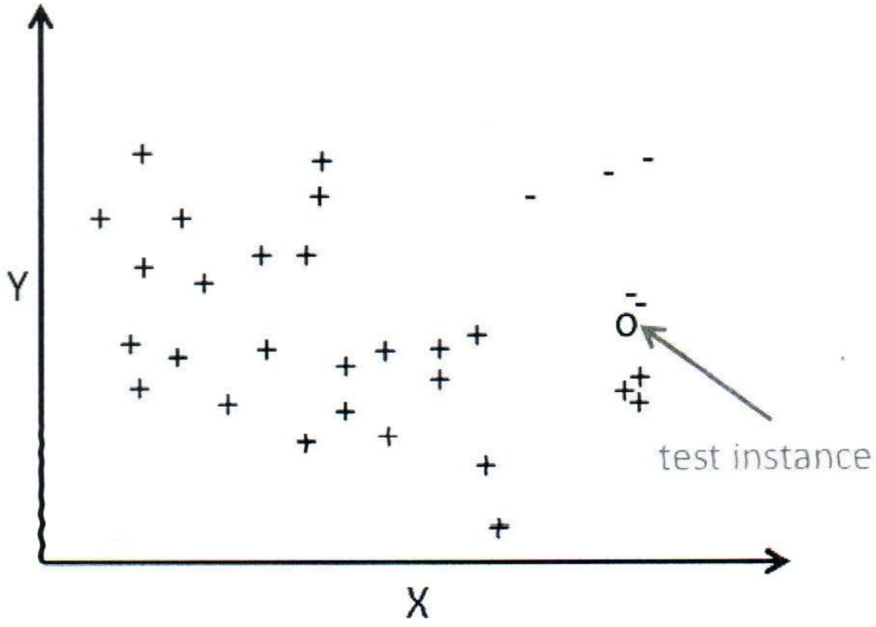
Compute the precision of this model with respect to the Spam class. **[4Marks]**

ii. Compute the recall of this model with respect to the Spam class **[4Marks]**

iii. 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! Which user is more likely to be satisfied with this classifier? Why? **[6 Marks]**

QUESTION TWO [20 MARKS]

A KNN classifier assigns a test instance the majority class associated with its K nearest training instances. Distance between instances is measured using Euclidean distance. Suppose we have the following training set of positive (+) and negative (-) instances and a single test instance (o). All instances are projected onto a vector space of two real-valued features (X and Y). Answer the following questions. Assume "unweighted" KNN (every nearest neighbor contributes equally to the final vote).



a) What would be the class assigned to this test instance for K=1 **[5 Marks]**

b) What would be the class assigned to this test instance for K=3 **[5 Marks]**

c) What would be the class assigned to this test instance for K=5 **[5 Marks]**

d) Setting K to a large value seems like a good idea. We get more votes! Given this particular training set, would you recommend setting K = 11? Why or why not? **[5 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]**

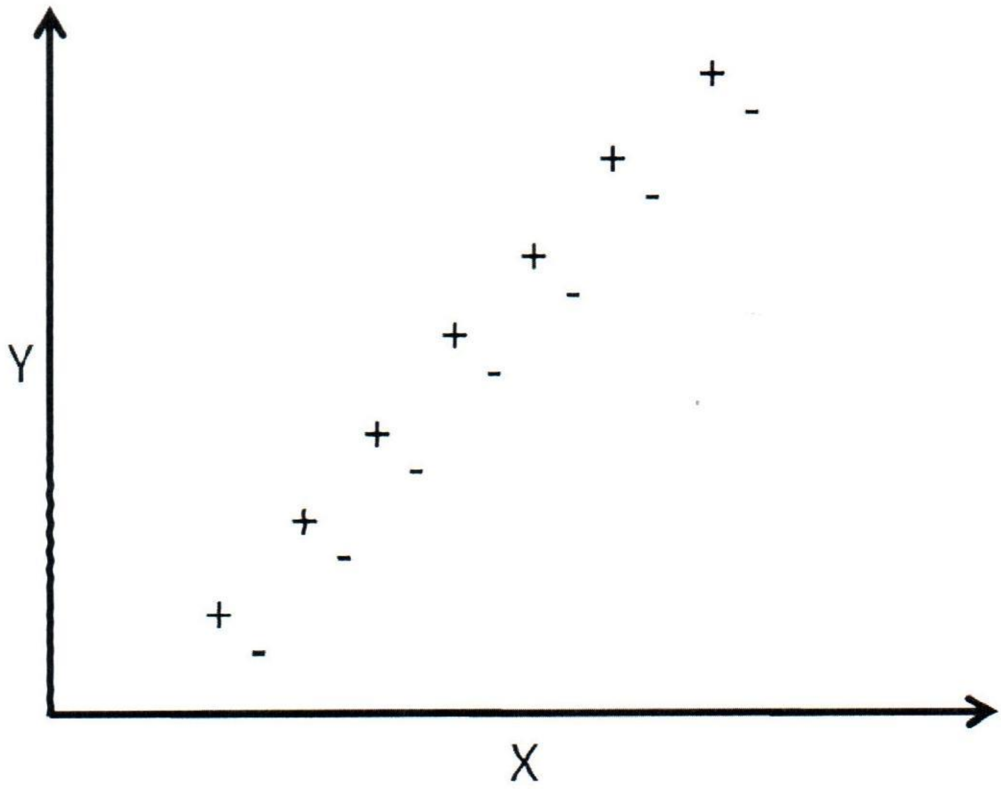
c) Describe Semi-Supervised learning **[4 Marks]**

d) Describe dirty data **[3 Marks]**

e) Describe the Procedure of a filter method **[3 Marks]**

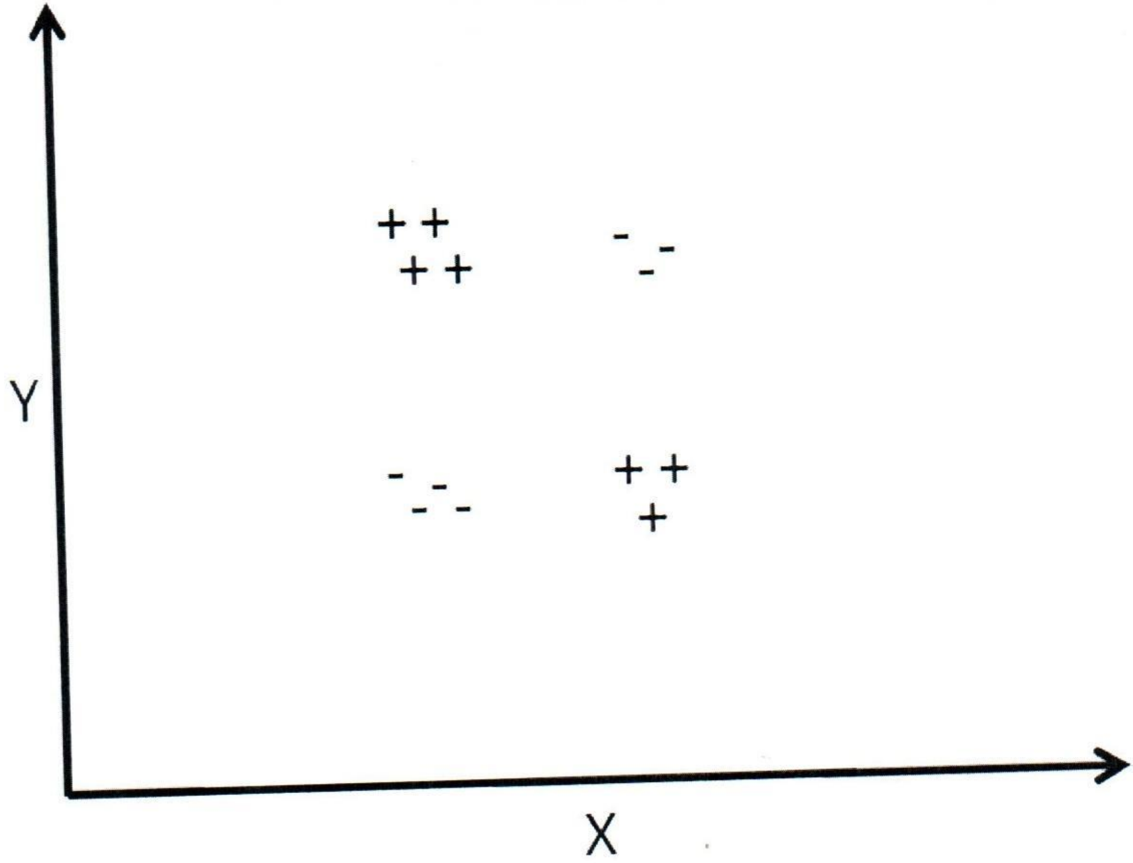
QUESTION FIVE [20 MARKS]

(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? Why? **[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? Why? [7.5 Marks]

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