CHUKA UNIVERSITY

UNIVERSITY EXAMINATIONS

FIRST YEAR EXAMINATION FOR THE AWARD OF DOCTOR OF PHILOSOPHY IN STATISTICS

MATH 944: STATISTICAL COMPUTING AND DATA-BASED MANAGEMENT

STREAMS: PhD (STATS) Y1S2 **TIME: 3 HOURS**

DAY/DATE: THURSDAY 08/04/2021 **2.30 P.M. – 5.30 P. M.**

INSTRUCTIONS: Answer Question ONE and ANY other TWO Questions

QUESTION ONE (30 MARKS)

a) Distinguished between a data mart, and data ware house as used in data mining? (4 marks)

b) Discuss the key properties of data mining and how it is useful in statistics. (6 marks)

c) Distinguish role of decision support system in data analysis? (4 marks)

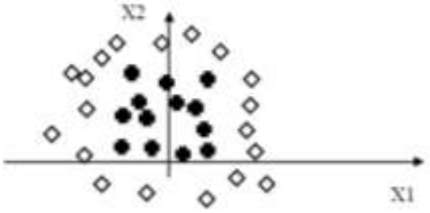
d) Discuss how will you know which data mining technique to choose for your classification problem? (6 marks)

e) There are two categories of supervised learning, classification and regression. When will you employ classification over regression? (4 marks)

f) ANN is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs. Discuss the basic structure and working of ANN. (6 marks)

QUESTION TWO (15 MARKS)

a) Suppose that we want to build a neural network that classifies two dimensional data (i.e., X = [x1, x2]) into two classes: diamonds and crosses. We have a set of training data that is plotted as follows:



Draw a network that can solve this classification problem. Justify your choice of the number of nodes and the architecture. Draw the decision boundary that your network can find on the diagram. (5 marks)

b) Prof Mark wants to use an artificial neural network (ANN) to automatically determine the species of bird's finches of the same family in images using the following measurements: Beak length, beak height, eye diameter, head length, and body length. He has a database of a few hundred labeled images of individuals of these species on which to train his ANN as shown.

i. Explain whether Prof Mark ANN be hidden units or not? (2 marks)

ii. Should the ANN use feedforward connections, recurrent connections, both, or neither? Explain your choice. (3 marks)

iii. Explain the learning mechanism(s) used by ANN. (5 marks)

QUESTION THREE (15 MARKS)

a) You have been hired by agricultural research institute to help them create an AI based system for Mushroom classification as either poisonous or not. You have the following data to consider.

Example	Is_Heavy	Is_Smelly	Is_Spotted	Is_Smooth	Is_Poisonous
A	NO	NO	NO	NO	NO
B	NO	NO	YES	NO	NO
C	YES	YES	NO	YES	NO
D	YES	NO	NO	YES	YES
E	NO	YES	YES	NO	YES
F	NO	NO	YES	YES	YES
G	NO	NO	NO	YES	YES
H	YES	YES	NO	NO	YES
U	YES	YES	YES	YES	?
V	NO	YES	NO	YES	?
W	YES	YES	NO	NO	?

You know whether or not mushrooms A through H are poisonous, but you do not know about U through W.

i. What is the entropy of Is_Poisonous? (2 marks)

ii. Which attribute should you choose as the root of a decision tree? (3 marks)

iii. What is the information gain of the attribute you chose in the previous question? (2 marks)

iv. Build a decision tree to classify mushrooms as poisonous or not. (5 marks)

v. Classify mushrooms U,V, and W using this decision tree as Poisonous or not. (3 marks)

QUESTION FOUR (15 MARKS)

a) You are given a data set on cancer detection. You've build a classification model and achieved an accuracy of 96%. Why shouldn't you be happy with your model performance? What can you do about it? (6 marks)

b) What are the functions of supervised learning and unsupervised learning? Discuss their applications in Modern Businesses. (4 marks)

c) You came to know that your model is suffering from low bias and high variance. What Is bias and variance in a machine learning model? What Is the Trade-off between bias and variance? What should you do when your model is suffering from low bias and high variance? Which algorithm should you use to tackle it? Why? (5 marks)

QUESTION FIVE (15 MARKS)

a) Explain the term ‘overfitting’ as studied in machine learning? Why does overfitting occur? How can overfitting be controlled. (4 marks)

b) Differentiate between K-means clustering algorithm and K-Nearest Neighbor. (6 marks)

c) What is ‘training set’ and ‘test set’ in a machine learning model? How much data will you allocate for your training, validation, and test sets? (5 marks)
