



MACHAKOS UNIVERSITY

University Examinations for 2018/2019

SCHOOL OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTING AND INFORMATION TECHNOLOGY

FOURTH YEAR SECOND SEMESTER EXAMINATION FOR

BACHELOR OF SCIENCE (INFORMATION TECHNOLOGY)

SIT 409 DATA MINING AND KNOWLEDGE DISCOVERY

DATE: 2/5/2019

TIME: 8.30-10.30 AM

INSTRUCTIONS

Answer **question ONE** and any other **TWO** questions.

QUESTION ONE (30 MARKS)

- a) Define the following terms as used in data mining.
 - i. Knowledge discovery (2 marks)
 - ii. Data warehouse (2 marks)
 - iii. Data cube (2 marks)
- b) Describe the steps and their purposes in knowledge discovery from databases. (10 marks)
- c) It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms in a class or a concepts. By providing examples, explain three of these **class/concept descriptions**. (10 marks)
- d) Differentiate OLAP and OLTP as used in data mining. (4 marks)

QUESTION TWO (20 MARKS)

Data pre-processing and conditioning is one of the key factors that determine whether a data mining project will be a success. For each of the following topics, describe the affect this issue can have on our data mining session and what techniques can we use to counter this problem.

- a) Noisy data (6 marks)
- b) Missing data (6 marks)
- c) As a system analyst of Hightech software Technologies, You note that many tuples in the company's data warehouse have no recorded values for several attributes for customer *income*. Explain four probable techniques that you could use to fill in these missing data. (8 marks)

QUESTION THREE (20 MARKS)

- a) Compare two definitions (views) of prediction tasks. (6 marks)
- b) Compare the pros and cons of decision tree and neural network classification methods. (8 marks)
- c) Briefly discuss the major difference between classification and clustering. List one real application for each of them. (6 marks)

QUESTION FOUR (20 MARKS)

- a) For each data mining technique identified below, describe the technique, identify which problems it is best suited for, identify which problems it has difficulties with, and describe any issues or limitations of the technique.
 - i. Decision trees (7 marks)
 - ii. Association rules (7 marks)
- b) Given five frequent itemsets: {beach, ocean}, {beach, sunshine} , {holiday, sunshine} {ocean sunshine} and {beach, ocean sunshine}. Generate association rules from the frequent itemsets. Calculate the confidence of each rule and identify all the strong association rules.
 Min_confidence=70%

$$\text{Confidence}(X \rightarrow Y) = P(X \mid Y) = P(X \cup Y) / P(Y)$$
 (6 marks)

QUESTION FIVE (20 MARKS)

- a) What is sampling (1 mark)
- b) Explain any three types of sampling (3 marks)
- c) You have trained a classifier to recognize your abstracts describing medical conditions for patients. The probability of the abstract describing a true medical condition is $P(\text{abstract})=0.79$. The abstract mentions two medical conditions in each different sentence. The first sentence describes diabetes with a probability is $P(\text{sentence}) = 0.98$ and the second condition is stroke with a probability of $P(\text{sentence}) = 0.67$ (8 marks)
- d) Given the following pairs of credit ranking and fraud outcome in a data set, give the formulation for entropy reduction and fill in the weights and probabilities in the formulation for splitting the records by Ranking= L. (8 marks)

	1	2	3	4	5	6	7	8	9	10
Fraud	No	No	No	Yes	No	No	No	Yes	No	Yes
Ranking	L	L	M	H	H	L	L	H	M	M