DEDAN KIMATHI UNIVERSITY OF TECHNOLOGY
University Examinations 2015/2016

THIRD YEAR SEMESTER II EXAMINATION FOR THE DEGREE OF BACHELOR OF BUSINESS AND INFORMATION TECHNOLOGY

CBT 2301: DATA MINING

DATE:2ND SEPTEMBER 2015 **TIME: 11.00AM-1.00PM**

INSTRUCTIONS

- Answer **ALL** questions in section A and any **TWO** questions in section **B**
- All questions in section B carry equal marks

Section A: Answer ALL questions in this section

Question One (30 Marks)

a)

i) Explain the meaning of the term *"Data Mining"* [2 marks]

ii) List the steps of knowledge discovery in databases (KDD) and state the purpose of each one of them. [7 marks]

b)

i) Explain FOUR ways in which data mining can be applied in market analysis and management [4 marks]

ii) Distinguish between supervised and unsupervised classification [3 marks]

c) Use the following association analysis results for products P and Q to answer questions That follows

Product1	==>	Product2	Support (%)	Confidence (%)	Lift
P	==>	Q	2.18	26.33	1.49

i) State the meaning of support confidence and lift [3 marks

ii) Provide an interpretation of the results for each evaluation metrics in the table i.e. interpret the support, confidence and Lift results in the table. [4 marks]

iii) After evaluating these results the analysts advised the supermarket selling Items P and Q to invest on advertisement of both P and Q to boost sales. Dispute this opinion and provide your opinion. [3 marks]

d)

i) Explain the meaning of the terms "noise" and "Outlier" in a dataset [2 marks]

ii) State TWO different approaches to detect outliers in a dataset. [2 marks]

Question Two (20 Marks)

The following is an example of customer purchase transaction data set.

CID	TID	Date	Items Purchased
1	1	01/01/2001	10,20
1	2	01/02/2001	10,30,50,70
1	3	01/03/2001	10,20,30,40
2	4	01/03/2001	20,30
2	5	01/04/2001	20,40,70
3	6	01/04/2001	10,30,60,70
3	7	01/05/2001	10,50,70
4	8	01/05/2001	10,20,30
4	9	01/06/2001	20,40,60
5	10	01/11/2001	10,20,30,60

Note: CID = Customer ID and TID = Transactions ID

a) Calculate the support, confidence and lift of the association rule {10} → {50, 70}. Indicate if the items in the association rule are independent of each other or have negative or positive impacts on each other. [5 points]

b) The following is the list of large two item sets {10, 20} {10, 30} {20, 30} {20, 40}. Show the steps to apply the Apriori property to generate and prune the candidates for large three itemsets. Describe how the Apriori property is used in the steps. Give the final list of candidate large three item sets. [6 marks]

c) Does customer 1 support the sequence <{20} {50,70} {10}>? Justify your answer. [5 marks]
d) Calculate the support of <{10}, {30}>. [2 marks]
e) Based on the types of association rules, identify which type(s) of rules {10}→{50,70} is? [2 points]

Question Three (20 Marks)

a) Describe TWO different techniques to deal with missing values in a dataset. Explain when each of these techniques would be most appropriate. [4 marks]

b) Outline SIX possible recommendations For X=>Y Rule (Where X and Y are 2 separate Products and have high support, high confidence and high positive lift > 1) [6 marks]

c) Assume that the numerals in the following association rules and large sequences identify different music files that customers downloaded on the Internet in the same sessions or over multiple sessions. As a consultant to Amazon.com, make a recommendation to your client based on each of the following association rule. [10 marks]

i. 1 → 2 with low support, high confidence and lift = n where n is large.

ii. 1→ 2 with high support, high confidence and lift = 0

iii. 1→ 2 with high support, high confidence and lift = -n where n is large.

iv. <{1, 2}, {3}> with high support

v. <{1, 2, 3}, {4}> with high support

Question Four (20 Marks)

a) Distinguish between descriptive and predictive modeling as applied in classification analysis [3 marks]

b) With aid of graphical illustration, explain the general approach for building a classification model [4 marks]

c) An Internet marketer is interesting in segmenting Internet with a clustering tool using the input attributes – top ten search key words used, top 10 URLs, recent 10 online purchases (vendor, product, qty, amt), Internet usage level, heaviest access hour, and heaviest access day of a week. Answer the following questions:

i) Can we find users with different income level? Why or why not. [2 ½ Marks]

ii) Can we expect to find clusters differentiated based on Internet usage level? Why or why not. [2 ½ Marks]

d) Consider the following confusion matrix for a binary classification matrix

		Predicted class	
		Class = 1	Class = 0
Actual Class	Class = 1	F_{11}	F_{10}
	Class = 0	F_{01}	F_{00}

i) Explain the meaning of the entries $F_{11,}$ $F_{10,}$ $F_{01,}$ and $F_{00,}$ in the matrix [2 marks]

ii) From the confusion matrix, derive the expressions for accuracy and the error rate performance metrics that may be used to evaluate the performance of a classification model [2 marks]

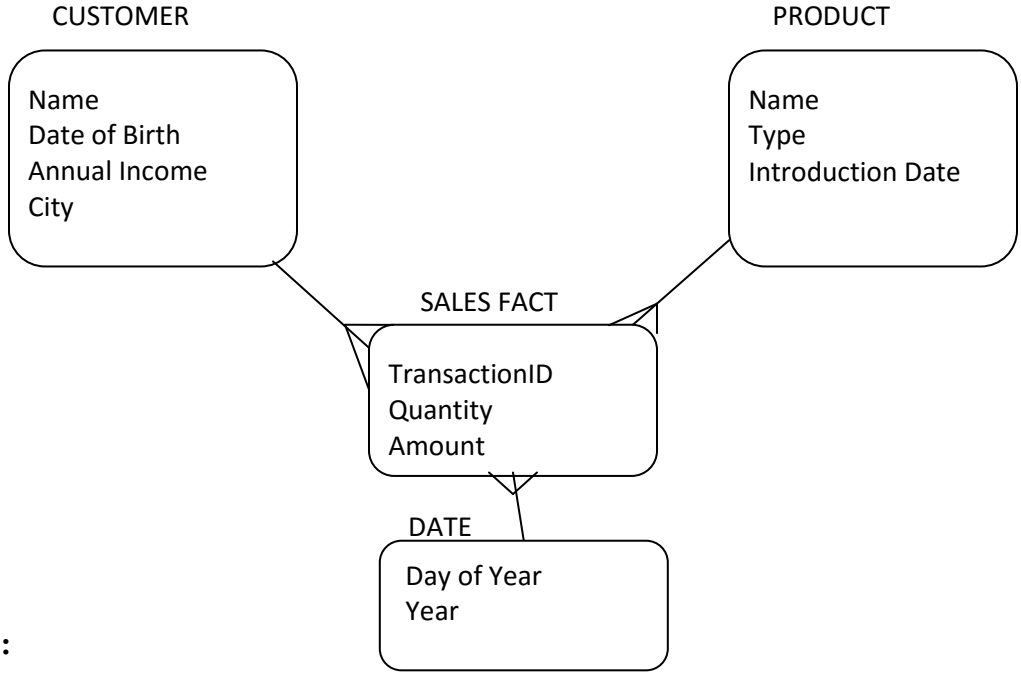
e)

i) In a decision tree classifier, the measures developed for determining the best way to split the records are based on the degree of impurity of the child nodes. Given that node N1 has class distribution (0.2, 0.8) and N2 has class distribution (0.4, 0.6). Which node has the lowest impurity and why [2 ½ marks]

ii) State three metrics for assessing the degree of impurity of a node [1 ½ marks]

Question Five (20 Marks)

The following is the star schema for Sales Department of your company.



Note:

1. TransactionID is used as the primary key in the fact table because there might be more than one transaction for each customer and product in a given day.

2. The Introduction Date for a product is the date when it is first introduced into the market.

a) The clustering task was selected to identify *customer segmentation.* Suggest the attributes including derived attributes to be used in the clustering task and justify your answer. [4 Marks]

b) Recommend a standardization or normalization method for the attributes in a distance function. [3 Marks]

c) You are asked to recommend a classification/predication task to be performed on the above data set.

i. Specify the input and class label attributes you choose for this classification/prediction task. Give an example of business decision(s) that can benefit from the classification/prediction results using the input and class label attributes of your choice. [4 marks]

ii. Define and give an example of noise using the data set above. [3 marks]

iii. Assume that you will use a decision tree classifier. Specify and compare the different tree pruning approaches. [3 marks]

iv. Suppose you are using a neural network instead of a decision tree. List at least three possible parameters you want to tune to improve its performance during the training period. [3 marks]