NATIONAL SCHOOL OF BUSINESS MANAGEMENT

BSc in Management Information Systems (Special) & BSc (Hons) Software Engineering

4th Year 1st Semester Examination

30 May 2021

CS405.3 - Data Warehousing and Data Mining

Instructions to Candidates

1) **Answer all questions.**
2) **Time allocated for the examination is three (03) hours and 30 minutes (Including downloading and uploading time)**
3) Download the paper, provide answers to the selected questions in a word document.
4) Please upload the document with answers (Answer Script) to the submission link before the submission link expires
5) Answer script should be uploaded in PDF Format
6) Under any circumstances E-mail submissions would not be taken into consideration for marking. Incomplete attempt would be counted as a MISSED ATTEMPT.
7) The Naming convention of the answer script – Module Code_Subject name_Index No

8) You must adhere to the online examination guidelines when submitting the answer script to N-Learn.

9) Your answers will be subjected to Turnitin similarity check, hence, direct copying and pasting from internet sources, friend's answers etc. will be penalized.

Question 1 -20 Marks

1. Briefly explain the difference between Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) in terms of **users**, **functions** and **storage size** (3 marks)
2. Briefly explain Snowflake schema and why Data Warehouse applications still prefer star schema over snowflake schema (5 marks)
3. What are the differences of independent and dependent data marts? (2marks)
4. By using an illustration explain Base cuboid and Apex cuboid (3 marks)
5. What are the differences between **information processing**, **analytical processing**, and **data mining** applications? (3 marks)
6. Briefly explain the tasks involved in data preprocessing (4 marks)

Question 2 -20 Marks

1. Briefly describe Market Basket Analysis technique (2 marks)
2. Briefly discuss why Knowledge Discovery from Data (KDD) is different to Data Mining. (2 marks)
3. What is the **frequent sequence** mining? (1 marks)
4. In order to optimize the network traffic of an organization, the management decided to collect some information about the web browsing of their employees during the office hours. They have collected data against their 6 main private IP addresses and 5 famous web sites. Below table summarizes the output. Records with “1” represents that web site being visited by that IP address and “0” represents that web site not being visited. By using apriori algorithm answer the questions accordingly.

   Minimum support count = 2

IP address	Google (G)	Bing (B)	YouTube (Y)	Facebook (F)	Wikipedia (W)
172.25.70.221	1	0	1	1	0
172.25.70.222	1	0	0	0	1
172.25.70.223	0	1	1	1	0
172.25.70.224	1	0	0	1	0
172.25.70.225	1	0	0	0	1
172.25.70.226	1	1	1	1	1

I. Find the candidate 1-Itemset for the dataset. (3 marks)

II. Find the candidate 2-itemset and frequent 2-item set for the dataset. (8 marks)

III. Mine the most frequently visited websites with respective support counts. (4 marks)

Question 3 -20 Marks

1. Following table extracted from a transaction processing database of a supermarket. Data is collected for 5 items which are Apple (A), Banana (B), Carrot (C), Dhal (D), and Eggs (E). By using FP-tree algorithm answer below questions.
   Minimum Support Count = 2.

TransactionId	Items
T1	E, D, A
T2	C, E
T3	C, D, A
T4	C, A, B
T5	C, E, D
T6	E, D, B
T7	D
T8	A, B
T9	D, E
T10	E, D, A
T11	C, E, D, A, B

I. Find the candidate 1-itemset for the dataset and the List denoted by "L" with ordered set of items in descending order. (4 marks)

II. Draw the FP-Tree for the given data set. (6 marks)

III. Find the Conditional pattern bases for each item in the dataset. (5 marks)

IV. Mine the frequent patterns of each item using FP-Tree for each. (5 marks)

Question 4 -20 Marks

1. What is the importance of cluster analysis in data mining? (2 marks)
2. Bellow table shows the result of a classifier which used to predict Covid19 patients. Horizontally (Rows) represents the actual classes and Vertically (Columns) represents the predictions of the classifier. By considering Accuracy, Sensitivity and Specificity measures evaluate the performance of this classifier. (3 marks)

Classes	Covid19 = Positive	Covid19 = Negative	Total
Covid19 = Positive	90	210	300
Covid19 = Negative	120	9580	9700
Total	210	9790	10000

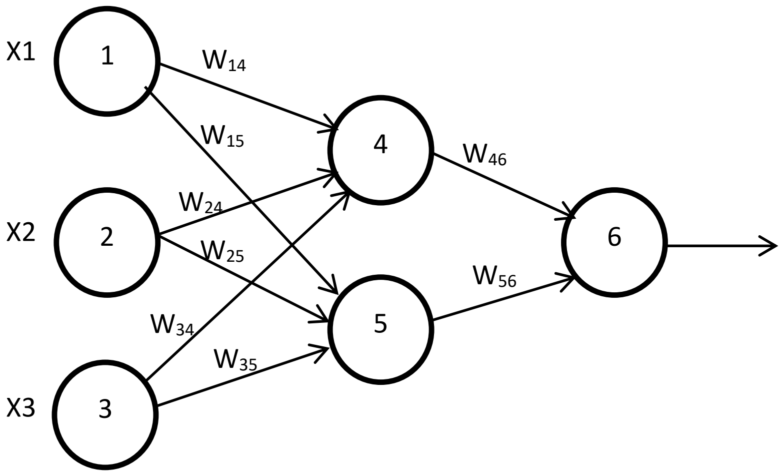
3. Consider the following database table which was extracted from a fraud detection system used by a bank. By using Naïve Bayes classification algorithm find out that below query would classifies as fraud = Yes or No. (15 marks)

   X= (Age = Middle, AssertDeclared = Yes, Income = Medium, EmploymentType = Self)

RecordID	Age	AssertDeclared	Income	EmploymentType	Fraud
T1	Senior	Yes	High	Permanent	No
T2	Middle	Yes	Medium	Permanent	No
T3	Middle	Yes	High	Permanent	No
T4	Middle	No	Medium	Self	Yes
T5	Youth	No	Low	Self	Yes
T6	Middle	Yes	High	Self	No
T7	Senior	No	High	Permanent	Yes
T8	Youth	Yes	Low	Contract	Yes
T9	Senior	Yes	Low	Contract	Yes
T10	Youth	No	High	Self	No
T11	Youth	Yes	Medium	Contract	No
T12	Middle	No	Low	Contract	No
T13	Senior	Yes	Medium	Permanent	No
T14	Middle	No	Low	Self	Yes
T15	Youth	No	High	Permanent	No

Question 5 -20 Marks

1. Below diagram shows a multilayer feed forward neural network with one hidden layer. X1, X2, X3 are the input tuples for the network and W_{14}, $W_{15,}$ …. Represents the respective weights of the links between nodes. Initial weights and biases (represented as Ɵ4, Ɵ5 and Ɵ6 for the bias of unit4, unit5 and unit6 respectively) are given in the table below the network diagram. By using back propagation algorithm answer the questions accordingly.



X1	X2	X3	W_{14}	W_{15}	W_{24}	W_{25}	W_{34}	W_{35}	W_{46}	W_{56}	Ɵ4	Ɵ5	Ɵ6
0	1	1	0.2	-0.5	-0.3	0.6	0.3	-0.2	-0.4	0.4	-0.3	0.5	0.2

I. If X1 = 0, X2 = 1 and X3 = 1 inputs to the network via unit1, unit2 and unit3 respectively compute the Net Input and Output of the each unit using below two equations. (6 marks)

$$I_j = \sum W_{ij} O_i + \theta_j$$

$$O_j = \frac{1}{1 + e^{-I_j}}$$

II. If Output of the unit 6 is given as 1 compute the error of output layer using below equation (2 marks)

$$Err_j = O_j(1 - O_j)(T_j - O_j)$$

III. By using below equation compute the errors of units 5 and 4. (4 marks)

$$Err_j = O_j\big(1 - O_j\big) \sum_K Err_K W_{jk}$$

IV. If learning rate (L) is given as 0.9 adjust the weights of W_{56} and W_{46} and bias of the unit6. (8 marks)

$$W_{ij} = W_{ij} + (L) Err_j O_i$$

$$\theta_j = \theta_j + (L) Err_j$$

…….END OF THE PAPER…….