ii. Explain star schema. What are the problems with star schema design? When snow flake schema is useful? **6**

OR iii. Explain architecture of Data Ware house with labelled diagram. **6**

Q.6 Attempt any two:

i. Discuss major issues in Data Mining. **5**

ii. Discuss social impact of Data Mining with relevant example. **5**

iii. Write short note: **5**

(a) Spatial Data Mining

(b) Web Mining

******

Total No. of Questions: 6 *Total No. of Printed Pages:4*

Enrollment No......................................

Faculty of Engineering

End Sem (Odd) Examination Dec-2017

CA5CO15 Data Warehousing and Mining

Programme: MCA Branch/Specialisation: Computer Application

Duration: 3 Hrs. **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d.

Q.1 i. Which of the following is not a data mining functionality? **1**

(a) Characterization and Discrimination.

(b) Classification and regression.

(c) Selection and interpretation.

(d) Clustering and Analysis.

ii. The various aspects of data mining methodologies is/are ______ **1**

I. Mining various and new kinds of knowledge.

II. Mining knowledge in multidimensional space.

III. Pattern evaluation and pattern or constraint-guided mining.

IV. Handling uncertainty, noise, or incompleteness of data.

(a) I, II and IV only (b) II, III and IV only

(c) I, II and III only (d) All I, II, III and IV

iii. ________ is data about data **1**

(a) Mini data (b) Meta data (c) Micro data (d) Multi data

iv. Data cleaning is **1**

(a) Large collection of data mostly stored in a computer system

(b) The removal of noise errors and incorrect input from a database

(c) The systematic description of the syntactic structure of a specific database. It describes the structure of the attributes the tables and foreign key relationships

(d) None of the above

v. Classification is **1**

(a) A subdivision of a set of examples into a number of classes

(b) A measure of the accuracy, of the classification of a concept that is given by a certain theory

(c) The task of assigning a classification to a set of examples

(d) None of the above

P.T.O.

vi. A Cluster is **1**

(a) Group of similar objects that differ significantly from other objects

(b) Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm

(c) Symbolic representation of facts or ideas from which information can potentially be extracted

(d) None of the above

vii. Data is stored, retrieved and updated in **1**

(a) OLTP (b) OLAP (c) SMTP (d) FTP

viii. Star schema is composed of _______ fact table **1**

(a) One (b) Two (c) Three (d) Four

ix. K-means is an example of **1**

(a) Classification (b) Association

(c) Clustering (d) Prediction

x. PageRank is a metric for ________documents based on their quality **1**

(a) Ranking hypertext (b) Ranking document structure

(c) Ranking web content (d) None of these

Q.2 i. Define data mining. **3**

ii. Describe the steps in the process of knowledge discovery in databases with diagram. **7**

OR iii. Draw and explain architecture of a typical data mining system. **7**

Q.3 i. What do you mean by Data Pre-processing? **2**

ii. A database has four transactions. Let Min. Support = 60% and min. Conf = 80% : **8**

T ID	**Date**	**Item Bought**
T 100	15/10/17	{K,A,D,E}
T 200	15/10/17	{D,A,C,E,B}
T 300	19/10/17	{C,A,B,E}
T 400	20/10/17	{B,A,D}

Find all frequent item sets using A priori.

OR iii. Suppose that the data for analysis include the attribute age. The age values for the data tuples are (in increasing order): **8**

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) Use smoothing by bin means to smooth the above data, using a bin depth of 3.

(b) How might you determine outliers in the data?

Q.4 i. Define Classification and Prediction. **2**

ii. With the help of decision tree find means of predicting which company profiles will lead to a increase or decrease in profits based on the following data: **8**

Age	**Competition**	**Type**	**Profit**
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

Profit is class attribute.

OR iii. Given the car theft data. Attributes are Car No., Color , Type , Origin, and the class label **Stolen** can be either yes or no. **8**

Car No.	**Color**	**Type**	**Origin**	**Stolen**
A1	Blue	Racing	Domestic	Yes
A2	Blue	Racing	Domestic	No
A3	Blue	Racing	Domestic	Yes
A4	Yellow	Racing	Domestic	No
A5	Yellow	Racing	Important	Yes
A6	Yellow	SUV	Important	No
A7	Yellow	SUV	Important	Yes
A8	Yellow	SUV	Domestic	No
A9	Blue	SUV	Important	No
A10	Blue	Racing	Important	Yes

Apply the Bayesian Classification on above data.

Q.5 i. Differentiate between OLTP and OLAP. **4**

CA5CO15 Data Warehousing and Mining

Marking scheme

Q.1 i. Which of the following is not a data mining functionality? **1**

(a) Characterization and Discrimination

(b) Classification and regression

(c) Selection and interpretation

(d) Clustering and Analysis

Ans: (c) Selection and interpretation

ii. The various aspects of data mining methodologies is/are ______ **1**

i) Mining various and new kinds of knowledge

ii) Mining knowledge in multidimensional space

iii) Pattern evaluation and pattern or constraint-guided mining.

iv) Handling uncertainty, noise, or incompleteness of data

(a) i, ii and iv only (b) ii, iii and iv only

(c) i, ii and iii only (d) All i, ii, iii and iv

Ans: (d) All i, ii, iii and iv

iii. ________ is data about data **1**

(a) Mini data (b) Meta data (c) Micro data (d) Multi data

Ans: (b) Meta data

iv. Data cleaning is **1**

(a) Large collection of data mostly stored in a computer system

(b) The removal of noise errors and incorrect input from a database

(c) The systematic description of the syntactic structure of a specific database. It describes the structure of the attributes the tables and foreign key relationships.

(d) None of the above

Ans: (b) The removal of noise errors and incorrect input from a database

v. Classification is **1**

(a) A subdivision of a set of examples into a number of classes

(b) A measure of the accuracy, of the classification of a concept that is given by a certain theory

(c) The task of assigning a classification to a set of examples

(d) None of the above

Ans: (a) A subdivision of a set of examples into a number of classes

vi. A Cluster is **1**

(a) Group of similar objects that differ significantly from other objects

(b) Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm

(c) Symbolic representation of facts or ideas from which information can potentially be extracted

(d) None of these

Ans: (a) Group of similar objects that differ significantly from other objects

vii. Data is stored, retrieved and updated in **1**

(a) OLTP (b) OLAP (c) SMTP (d) FTP

Ans: (a) OLTP

viii. Star schema is composed of _______ fact table **1**

(a) One (b) Two (c) Three (d) Four

Ans: (a) One

ix. k-means is an example of **1**

(a) Classification (b) Association

(c) Clustering (d) Prediction

Ans: (c) Clustering

x. PageRank is a metric for ________documents based on their quality **1**

(a) ranking hypertext (b) ranking document structure

(c) ranking web content (d) None of these

Ans: (c) ranking web content

Q.2 i. Define data mining. **3**

0.75 * 4 marks for each explained term of definition.

ii. Describe the steps in the process of knowledge discovery in databases with diagram. **7**

2 Marks for diagram.

5 marks for explanation of each step.

OR iii. Draw and explain architecture of a typical data mining system. **7**

3 Marks for diagram

4 Marks for explanation

Q.3 i. What do you mean by Data Pre-processing? **2**

2 marks for explanation

ii. A database has four transactions. Let Min. Support = 60% and min. Conf = 80% : **8**

T ID	**Date**	**Item Bought**
T 100	15/10/17	{K,A,D,E}
T 200	15/10/17	{D,A,C,E,B}
T 300	19/10/17	{C,A,B,E}

T 400	20/10/17	{B,A,D}

Find all frequent item sets using A priori.

5 marks for generating all the frequent n-item sets

3 marks for properly applying pruning

OR iii. Suppose that the data for analysis include the attribute age. The age values for the data tuples are (in increasing order): **8**

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

i) Use smoothing by bin means to smooth the above data, using a bin depth of 3.

ii) How might you determine outliers in the data?

5 marks for applying smoothing by bin method

3 marks for explaining methods of outlier detection.

Q.4 i. Define Classification and Prediction. **2**

1* 2 marks each for definition

ii. With the help of decision tree find means of predicting which company profiles will lead to a increase or decrease in profits based on the following data: **8**

Age	**Competition**	**Type**	**Profit**
Old	Yes	Software	Down
Old	No	Software	Down
Old	No	Hardware	Down
Mid	Yes	Software	Down
Mid	Yes	Hardware	Down
Mid	No	Hardware	Up
Mid	No	Software	Up
New	Yes	Software	Up
New	No	Hardware	Up
New	No	Software	Up

Profit is class attribute.

3 Marks for calculating information gain for 3 attributes

4 Marks for calculating 2nd level splitting attribute

1 Mark for drawing the tree

OR iii. Given the car theft data. Attributes are Car No., Color , Type , Origin, and the class label **Stolen** can be either yes or no. **8**

Car No.	**Color**	**Type**	**Origin**	**Stolen**
A1	Blue	Racing	Domestic	Yes
A2	Blue	Racing	Domestic	No
A3	Blue	Racing	Domestic	Yes
A4	Yellow	Racing	Domestic	No
A5	Yellow	Racing	Important	Yes
A6	Yellow	SUV	Important	No
A7	Yellow	SUV	Important	Yes
A8	Yellow	SUV	Domestic	No
A9	Blue	SUV	Important	No
A10	Blue	Racing	Important	Yes

Apply the Bayesian Classification on above data.

2 Marks for calculating probabilities of two classes

4 marks for calculating conditional probabilities

2 marks for calculating posterior probabilities

Q.5 i. Differentiate between OLTP and OLAP. **4**

1/2 * 8 marks for each difference

ii. Explain star schema. What are the problems with star schema design ? When snow flake schema is useful ? **6**

2 marks for explaining of star schema

2 marks for highlighting problems of star schema

2 marks for usefulness of snowflake schema

OR iii. Explain architecture of Data Ware house with labelled diagram. **6**

2 marks for diagram

4 marks for explanation

Q.6 Attempt any two:

i. Discuss major issues in Data Mining. **5**

1 * 5 mark for discussion of each issue

ii. Discuss social impact of Data Mining with relevant example. **5**

3 marks for outlining social impact

2 marks for example

iii. Write short note : **5**

a) Spatial Data Mining

b) Web Mining

2.5 * 2 marks for each topic

******