الجامعةالسعوديةالالكترونية
Saudi Electronic University

Ministry of Higher Education
Saudi Electronic University
College of Computing and Informatics

Mid-term Examination Cover Sheet

Fall Semester: 1436 / 2015

Course Title:	**Datamining and Data Warehousing**	**Course Code:**	**IT446**
Exam Duration:	**60 Minutes**	**Number of Pages:** (including cover page)	**8**

Fill all the information below so that your answer sheet is not misplaced

Student Name: ______________________ **Student ID:** ______________

Class Day & Time ______________________ **CRN:** ______________

Instructor Name: ______________________ **Exam Date:** ______________

Exam Guidelines

- **Mobile phones are not permitted.**

Marking Scheme

Questions	Score
Q1 (10 Marks)	/10
Q2 (2.5 Marks)	/2.5
Q3 (2.5 Marks)	/2.5
Q4 (2.5 Marks)	/2.5
Q5 (2.5 Marks)	/2.5
Q6 (2.5 Marks)	/2.5
Q7 (2.5 Marks)	/2.5
Total (100)	

Section 1: Multiple Choice Questions (10 Marks)

Every question carries 1 mark

Q1. Choose the appropriate answer:

1. This is a Data Mining task:

   A. Computing the total sales of a company.
   B. Sorting a student database based on student identification numbers.
   C. Predicting the future stock price of a company using historical records.
   D. Dividing the customers of a company according to their gender.

2. Data mining, as a research discipline, does not draw ideas from this research discipline:

   A. Statistics
   B. Linguistics
   C. Artificial Intelligence
   D. Information Theory

3. The set of odd integers from n=5 to n=41 is which of these types of variables:

   A. Categorical
   B. Interval
   C. Independent
   D. Ordinal

4. Data preprocessing does not include which of these tasks:

   A. Data classification
   B. Data integration
   C. Data reduction
   D. Data cleaning

5. Which of the following is not a factor in data quality?

   A. Accuracy
   B. Completeness
   C. Relevance
   D. Timeliness

6. Attribute Subset Selection is a

    A. Way to reduce the dimensionality of the data
    B. Step in applying the wavelet transformation
    C. Step in applying principal component analysis
    D. Way of performing histogram analysis

7. A data warehouse is a ---------------- collection of data in support of management's decision-making process

    A. subject-oriented
    B. integrated
    C. time-variant
    D. subject-oriented, integrated, time-variant, and nonvolatile

8. Data Warehouse consists of ----------------- .

    A. relational databases
    B. flat files
    C. on-line transaction records
    D. All of the above are correct

9. OLTP is the acronym of -------------------.

    A. On line transaction processing
    B. Only large transfer process
    C. Open line transmission process
    D. Object linking & transmission process

10. One way to interact OLAP-styled analysis with data mining

    A. Using cube space to define data space for mining
    B. Using OLAP queries to generate features and targets for mining, e.g., multi-feature cube
    C. Using data-mining models as building blocks in a multi-step mining process, e.g., prediction cube
    D. All of the above

Section 2: True/False Questions (2.5 Marks)

Q2. Please specify whether the following statements are true or false.

Statement	True/False
1) A data cube can be usefully viewed as a lattice of cuboids.	**True**
2) Cluster analysis is a means for discovering and grouping together (clustering) sets of observations that are closely related	**True**
3) Binary attributes are special case of ordinal attributes.	**False**
4) Info Cube is a 3-D visualization technique where hierarchical information is displayed as nested semi-transparent cubes.	**True**
5) The term “noise” has a technical meaning in data mining referring to the distortion of data from their true value and/or the addition of spurious objects.	**True**

Section 3: Fill in the Blanks (2.5 Marks)

Q3. Fill in the blanks with appropriate answers from the table below

apex cuboid	**Iceberg**	**drilling down**	**ordered**	**Apriori Pruning**

1. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid.

2. Computing *only* the cuboid cells whose measure satisfies a certain minimum support, then the resulting materialized data cube is called Iceberg data cube.

3. If there is any item set which is infrequent, its superset should not be generated/tested. This is called Apriori Pruning Principle.

4. Ordered data can be sequentially ordered in terms of spatial (positional)- or time-based attributes.

5. If we analyze monthly data in terms of the days of each month we are drilling down.

Section 4: Subjective Questions (10 Marks)

All questions carry equal marks

Q4. In the context of Data Preprocessing, what are the three methods to handle missing data values? Briefly explain every method (roughly in two to three lines).

1. Ignore the tuple

Skip the tuple and don’t include it in the processed data. It is usually done when class label is missing.

2. Fill in the missing value manually

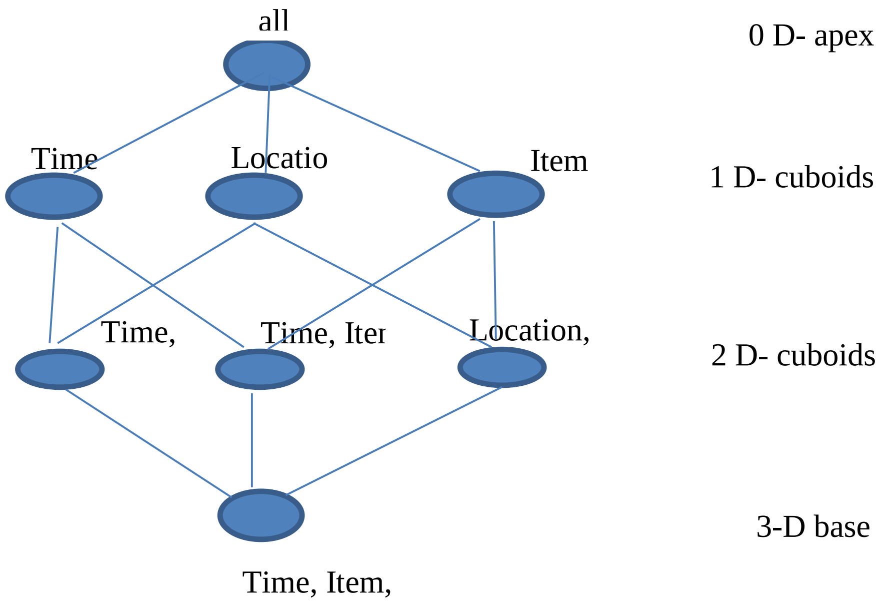
Instead of ignoring the entire tuple, fill in the missing values with appropriate ones. This is a tedious manually fill in the values and becomes infeasible when there are hundreds and thousands of tuples and several of them are having missing data.

3. Fill in it automatically

When manual filling becomes infeasible, automatic filling with a global constant : e.g., “unknown”, or the attribute mean for the samples belonging to the same class. Another alternative is to use the most probable value: inference-based such as Bayesian formula or decision tree.

Q5. Draw the lattice of cuboids (i.e. all cuboid levels from apex cuboid to base cuboid) for the following 3-D sales cube:

- Dimension 1: Time
- Dimension 2: Location
- Dimension 3: Item



Q5. How can data visualization help in decision-making?

Answer:

Data visualization helps the analyst gain intuition about the data being observed. Visualization applications frequently assist the analyst in selecting display formats, viewer perspective and data representation schemas that faster deep intuitive understanding thus facilitating decision-making.

Q6. What is data reduction? Why it is used? Write names of 3 data reduction strategies.

Data reduction:

It is the process to obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results

Why data reduction?

A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set. Using data reduction strategies, we reduce the volume of data which enables us to complete complex data analysis tasks in reasonable time.

Data reduction strategies

1. Dimensionality reduction,
2. Numerosity reduction,
3. Data compression.

G☺☺D LUCK