LAST MOMENT TUITIONS

DATA WAREHOUSE AND MINING

MAY 2019

Computer Engineering (Semester 6)

Total marks: 80
Total time: 3 Hours

INSTRUCTIONS
(1) Question 1 is compulsory.
*(2) Attempt any **three** from the remaining questions.*
(3) Draw neat diagrams wherever necessary.

1.a. What are spatial data structures? Outline their importance in GIS. (5 marks)

1.b. What is Metadata? Why do we need metadata when search engines like Google seem so effective? (5 marks)

1.c. In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe various methods for handling this problem. (5 marks)

1.d. With respect to web mining, is it possible to detect visual objects using meta-objects? (5 marks)

2.a. Suppose that a data warehouse for *DB-University* consists of the four dimensions *student, course, semester,* and *instructor*, and two measures count and *avg-grade*. At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg-grade measure stores the actual course grade of the student. At higher conceptual levels, avg-grade stores the average grade for the given combination.

i. Draw a snowflake schema diagram for the data warehouse.

ii. Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should you perform in order to list the average grade of CS courses for each DB-University student. (10 marks)

2.b. What is the relationship between data warehousing and data replication? Which form of replication (synchronous or asynchronous) is better suited for data warehousing? Why? Explain with appropriate example. (10 marks)

3.a. The following table consists of training data from an employee database. The data have been generalized. For example, "31: : : 35" for *age* represents the age range of 31 to 35. For a given row entry, *count* represents the number of data tuples having the values for *department, status, age*, and *salary* given in that row.

department	status	age	salary	count
sales	senior	31…35	46k…50k	30
sales	junior	26…30	26k…30k	40
sales	junior	31…35	31k…35k	40
systems	junior	21…25	46k…50k	20
systems	senior	31…35	66k…70k	5
systems	junior	26…30	46k…50k	3
systems	senior	41…45	66k…70k	3
marketing	senior	36…40	46k…50k	10
marketing	junior	31…35	41k…45k	4
secretary	senior	46…50	36k…40k	4
secretary	junior	26…30	26k…30k	6

Let *status* be the class label attribute.

i. How would you modify the basic decision tree algorithm to take into consideration the *count* of each generalized data tuple (i.e., of each row entry)?

ii. Use your algorithm to construct a decision tree from the given data. (10 marks)

3.b. Why is *tree pruning* useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning? Given a decision tree, you have the option of (i) *converting* the decision tree to rules and then pruning the resulting rules, or (ii) *pruning* the decision tree and then converting the pruned tree to rules. What advantage does (i) have over (ii)? (10 marks)

4.a. Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are:
A1 (2, 10), A2 (2, 5), A3 (8, 4), B1 (5, 8), B2 (7,5), B3 (6, 4), C1 (1, 2), C2 (4, 9).
The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the *k-means* algorithm to show only (i) The three cluster centers after the first round of execution
(ii) The final three clusters. (10 marks)

4.b. Briefly outline with example, how to compute the dissimilarity between objects described by the following:

i. Nominal attributes

ii. Asymmetric binary attributes (10 marks)

5.a. Frequent pattern mining algorithms considers only distinct items in a transaction. However, multiple occurrences of an item in the same shopping basket, such as four cakes and three jugs of milk, can be important in transactional data analysis. How can one mine frequent itemsets efficiently considering multiple occurrences of items? Generate Frequent Pattern Tree for the following transaction with 30% minimum support: (10 marks)

Transaction ID	Items
T1	E,A,D,B
T2	D,A,C,E,B
T3	C,A,B,E
T4	B,A,D
T5	D
T6	D,B
T7	A,D,E
T8	B,C

5.b. Differentiate between simple linkage, average linkage and complete linkage algorithms. Use complete linkage algorithm to find the clusters from the following dataset. (10 marks)

X	4	8	15	24	24
Y	4	4	8	4	12

6.a. *Data quality* can be assessed in terms of several issues, including accuracy, completeness, and consistency. For each of the above three issues, discuss how data quality assessment can depend on the *intended use* of the data, giving examples. Propose two other dimensions of data quality. (10 marks)

6.b. Present an example where data mining is crucial to the success of a business. What *data mining functionalities* does this business need (e.g., think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis? (10 marks)