M.Sc. T.E. 2nd Semester 17 May 2024

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION SUMMER SEMESTER, 2022-2023

DURATION: 3 HOURS FULL MARKS: 150

CSE 6293: Data Warehousing and Mining

Programmable calculators are not allowed. Do not write anything on the question paper.

Answer **all 6 (six)** questions. Figures in the right margin indicate full marks of questions.

---

1. a) How is a data warehouse different from a database? How are they similar? 7

   b) Outliers are often discarded as noise. However, one persons garbage could be anothers treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Using fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable. 8

   c) It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation. Suppose you have the two-dimensional data set shown in Table 1. Given a new data point, $\mathbf{x} = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity. 10

Table 1: Data for Question 1.c

	A_1	A_2
x_1	1.5	1.7
x_2	2	1.9
x_3	1.6	1.8
x_4	1.2	1.5
x_5	1.5	1.0

2. a) Suppose that the data for analysis includes the attribute *age*. The *age* values for the data tuples are (in increasing order) 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, where the bin depth is 3. Illustrate your steps. 6

   ii. How would you determine outliers in the data? 6

   iii. What other methods are there for data smoothing? 3

   b) Suppose a data warehouse for a university consists of the following four dimensions: *student*, *course*, *semester*, and *instructor*, and two measures *count* and *avg_grade*. When 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. 5

   ii. Starting with the base cuboid [*student*, *course*, *semester*, *instructor*], what specific OLAP operations (e.g., roll-up from *semester* to *year*) should one perform to list the average grade of *Data Mining* courses for each University student. 5

3. a) What are the different architectures of OLAP servers? Discuss the motivation behind OLAP mining (OLAM). 2 + 3

b) What are the various types of data sampling method? What are the benefits of applying dimensionality reduction to a dataset? 2 + 3

c) Suppose you have 15 candidate itemsets of length 3: 1, 4, 5, 1, 2, 4, 4, 5, 7, 1, 2, 5, 4, 5, 8, 1, 5, 9, 1, 3, 6, 2, 3, 4, 5, 6, 7, 3, 4, 5, 3, 5, 6, 3, 5, 7, 6, 8, 9, 3, 6, 7, 3, 6, 8. Construct a hash tree for the above candidate itemsets of length 3. Assume the tree uses a hash function $H(p) = p \ mod \ 3$. 15

4. a) Consider the data set shown in Table 2.

Table 2: Data for Question 4.a

Customer ID	Transaction ID	Item Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

i. Compute the support for itemsets $\{e\}$, $\{b, d\}$, and $\{b, d, e\}$ by treating each transaction ID as a market basket. 5

ii. Use the results in part (i) to compute the confidence for the association rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$. 5

iii. Compute the support for itemsets $\{e\}$, $\{b, d\}$, and $\{b, d, e\}$ by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.) 5

iv. Use the results in part (iii) to compute the confidence for the association rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$. 5

b) What is the anti-monotone property of the Apriori algorithm? 5

5. a) Attribute 2 shown in Table 3 is continuous. Find the best split position for attribute 2 using the Gini index. 12

Table 3: Data for Question 5.a

Instance	Attribute 1	Attribute 2	Target Class
1	T	1.0	C1
2	T	6.0	C1
3	F	5.0	C2
4	F	4.0	C1
5	T	7.0	C2
6	T	3.0	C2
7	F	8.0	C2
8	F	7.0	C1
9	T	5.0	C2

b) What are the associated problems of choosing large and small values of k when using the K-Nearest Neighbors (KNN) algorithm? 8

c) What are the advantages of a rule-based classifier? 5

6. a) How does the K-means clustering algorithm work? 8

b) Mention the data preprocessing techniques used in text mining. 7

c) With an example explain the Bag of Words (BoW) technique for text mining. 10