The Center for Advanced Computer Studies

University of Louisiana at Lafayette

CMPS 566

Term Test

Date: March 23, 2015 Instructor: Dr. V. Raghavan

Time: 3:30 – 4:45pm Total Marks: 75

PART A (20 Marks)

NOTE: There are **five** parts. Answer **any 4.**

Q1. Classification vs. Prediction

Q2. Pattern Novelty

Q3. Fact Constellation Schema

Q4. Set-grouping Hierarchy

Q5. Smoothing (for noise handling) by bin boundaries

PART B (55 Marks)

NOTE: There are three questions. Answer all questions.

Q6. Suppose the following *age* values are extracted from the tuples of a dataset and are listed 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

a) Use *smoothing by bin means* to smooth the data, assuming we desire bin depth of (approximately) 3. Adopt an equi-depth approach. [5 marks]

b) How might you determine *outliers* in the data? [2 marks]

c) Use *min-max* normalization to transform the value 35 for *age* onto the range [0.0, 1.0]. Also, apply normalization by *decimal scaling* to obtain the transformed value for the person having *age* value = 35. Which normalization method would you prefer for this data? Give reasons why. [3] marks]

Q7. Suppose that a data warehouse for U.S. Traffic Ticket System consists of the following four dimensions: *driver, violation, location,* and *date,* and two measures count and amount (fine amount). For *location,* the concept hierarchy involves "street < city < state" and for *date,* the hierarchy consists of "date < month < quarter < year." Design the hierarchies (each having exactly two levels) for the *driver* and *violation* dimensions by yourself.

a) Draw a snowflake schema diagram for the data warehouse. [8 marks]

b) Starting with the base cuboid *[driver, violation, location,* and *date],* what specific OLAP operations (e.g., roll-up from street to city) should one perform in order to list the total amount of fines for the year of 2010 in Lafayette Louisiana. [8 marks]

c) How many cuboids will this cube contain (including the base and apex cuboids)? [3 marks]

d) Suppose there are four cuboids (including the Base cuboid) materialized: [6 marks]
cuboid 1: {year, street}
cuboid 2: { year, city)
cuboid 3: { state } where year =2010
Which of the above cuboids would you select for the query in part b)? Explain your reasons.

Q8.

ID	Power	Mileage
T1	high	med
T2	high	med
T4	med	med
T9	med	med
T11	med	med
T14	high	med
T3	high	low
T7	high	low
T8	med	low

a) Write a DMQL query to find the *comparisons* of cars according to 'Power.' The class of 'med power' is to be compared to the class of 'high power.' [2 marks]

b) Provide a comparison of naïve predictions, with regard to Power class under two situations: i) Mileage attribute is not assumed known, and ii) Mileage attribute is known. State the prediction accuracy associated with the predictions that you make. [4 marks]

c) Provide a bi-directional quantitative rule for predicting the target class Mileage = ‘low’. The Power attribute should be used to predict the target class. [3 marks]

d) Using the above table as the context for generating formal concepts, briefly explain and illustrate the following notions [4 marks]:

i. Non-feasible set of objects (tuples)

ii. A formal concept

iii. Non-comparable concepts

e) Briefly discuss the computational complexity associated with the computation of the concept lattice, including the factors the complexity depends on. [7 marks]