The Center for Advanced Computer Studies
University of Louisiana at Lafayette
CMPS 566
Term Test

Date: March 23, 2006 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 a group of 12 sales price records has been sorted as follows:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

a) Partition them into four bins, using a distance-based method (e.g. MaxDiff) to decide bin boundaries. [3 marks]

b) Use the **3-4-5 rule** to segment the above data, where you may go up to two levels. If in any particular bin there are two items or less, then it is not further segmented. [5 marks]

c) Compare the above two methods of segmentation, by providing two advantages and/or disadvantages. [2 marks]

d) Compare the result of part a) with the segmentation at the top level of part b), using the V-Optimal criterion and indicate which one is preferred. [5 marks]

Q7. Suppose that a data warehouse for a major sports league (e.g. NBA) consists of the four dimensions *date, spectator, location,* and *game,* and two measures, *count* and *charge,* where *charge* is the fare that a spectator pays when watching a game on a given date. Spectators may be of three categories: students, adults or seniors, with each category having its own charge rate. The spectator dimension has two levels (not counting *all*). For *location,* the concept hierarchy involves "stadium-name < city < state <

country-region" and for *date,* the hierarchy consists of "date < month < quarter < year." No hierarchy is defined for the *game* dimension.

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

b) Starting from the base cuboid, what specific OLAP operations should one perform in order to list the total charge paid by student spectators in stadiums located in the West-Coast region in 2004? [6 marks]

c) Suppose there are four cuboids materialized:
cuboid 1: {year, spectator-id, city}
cuboid 2: {year, spectator-category}
cuboid 3: {year, spectator-category, country-region}
cuboid 4: {spectator-id, country-region} where year =2004
Which of the above cuboids would you select for the query in part b)? Explain your reasons. [8 marks]

d) How many cuboids does this data warehouse contain, including base and apex cuboids? [3 marks]

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. [2 marks]