SYSC5703 – Assignment #2 – Fall 2013

Date out = Oct. 22nd, 2013 Date due = Nov. 28th, 2013
Marks: Q1 = 12 marks, Q2 = 3 marks, Q3 = 5 marks, Q4 = 4 marks, Q5 = 3 marks, Q6 = 13 marks, (Total marks = 40)

**** Make sure you work in group (same group for assignment #1 and the term paper).**

1. Consider the following functional dependencies over the attribute set ABCDEFGH:

   A → E
   AD → BE
   AC → E
   E → B
   BG → F
   BE → D
   BDH → E
   F → A
   D → H
   CD → A

   a. Find a minimal cover, and
   b. Decompose into lossless 3NF.
   c. After that, check if all the resulting relations are in BCNF. If you find a schema that is not, decompose it into a lossless BCNF. Explain all steps. (Exercise 6.20, page 248)

2. Find a projection of the following set of dependencies on the attribute AFE: A → BC E → HG C → FG G → A

3. For the attribute set ABCDEFG, let the MVDs be

   ABCD ⋈ DEFG
   ABCE ⋈ ABDFG
   ABD ⋈ CDEFG

   Find a lossless decomposition into 4NF. Is it unique? Why or why not?

4. Using the schema given below, define trigger that fire when a student grade average drops below certain threshold. (For simplicity, assume that there is a function, grade_avg(), which takes a student Id and returns the student average grade.)

   Student (*Id, Name, Country*)
   Course (*CrsCode, CrsName, Type, Instructor*)
   Results(*Id, CrsCode, Grade)*

5. It is possible to convert Datalog rules into equivalent relational algebra expressions. For each of the following Datalog rules, write an expression of relational algebra that defines the same relation as the head of the rule:

   a. P(x, y) ← Q(x, z) AND (R(z, y)
   b. P(x, y) ← Q(x, z) AND Q(z, y)
   c. P(x, y) ← Q(x, z) AND R(z, y) AND x < y

6.

a. In data mining, is a typical fact table in 4NF? Explain?

b. Consider the table below for a binary classification problem. This is the same example we saw in class.
   i. Compute Gini index for the overall collection of training examples.
   ii. Compute Gini index for the **Married** attribute.
   iii. Computer Gini index for the **PreviousDefaul**t attribute.
   iv. Compute Gini index for the **Income** attribute setting the threshold for default at less than 50.
   v. Which attribute (**Married** or **PreviousDefault** or **Income**) is better? Is there any difference between these results and the Entropy results?

Id	***Married***	***PreviousDefault***	***Income***	***Default*** (outcome)
C1	Yes	No	50	No
C2	Yes	No	100	No
C3	No	Yes	135	Yes
C4	Yes	No	125	No
C5	Yes	No	50	No
C6	No	No	30	No
C7	Yes	Yes	10	No
C8	Yes	No	10	Yes
C9	Yes	No	75	No
C10	Yes	Yes	45	No
C11	Yes	No	60	Yes
C12	No	Yes	125	Yes
C13	Yes	Yes	20	No
C14	No	No	15	No
C15	No	No	60	No
C16	Yes	No	15	Yes
C17	Yes	No	35	No
C18	No	Yes	160	Yes
C19	Yes	No	40	No
C20	Yes	No	30	No