Exam

Uncertain Data Management
Université Paris-Saclay, M2 Data&Knowledge

February 1st, 2016

This is the final exam for the Uncertain Data Management class, which will determine your grade for this class. The duration of the exam is two hours. The exam consists of two independent exercises.

You are allowed up to two A4 sheets of personal notes (i.e., four page sides), printed or written by hand, with font size of 10 points at most. If you use such personal notes, you must hand them in along with your answers. You may not use any other written material.

Write your name clearly on the top right of every sheet used for your exam answers. Number every page. It is highly recommended to answer the exercises on separate sheets.

The exam is strictly personal: any communication or influence between students, or use of outside help, is prohibited. Any violation of the rules may result in a grade of 0 and/or disciplinary action.

Exercise 1: Choose Your Own Answers! (10 points)

This exercise is presented as a roleplaying game, where you will put yourself in the shoes of a hypothetical student, Alice, who is taking an exam. Unlike you, Alice did not pay much attention in class, so she is baffled by most of the questions she is facing[1]. Fortunately, just like you, Alice is an expert in uncertain data management, so she decides to use her skills to represent her uncertainty about the exam answers and optimize her chances.

The exam taken by Alice is a *multiple choice exam*: it consists of a number of questions, each question has a number of possible answers, and exactly one answer for each question is the correct one. Alice is unsure about what the correct answer to each question is, but she has her estimates about what the probabilities are.

Question 1 (0.5 points). Alice represents her probability distribution on the answers as the following BID instance:

[1] Alice is also confused by the problem statement: parts of it, especially some footnotes, seem entirely useless, unnecessarily and tediously long-winded, and oddly self-referential.

Answers

question	answer	
1	A	0.5
1	B	0.2
1	C	0.3
2	A	0.8
2	B	0.2
3	A	0.1
3	B	0.9

- What are the key attributes? (no justification expected)
- How many blocks are there, and what do they contain? (no justification expected)

Question 2 (1.5 points). In this question, we consider an arbitrary BID instance J where all probabilities are different and greater than 0, and the probabilities in each block sum up exactly to 1.

- What is the number of possible worlds of J, as a function of the number of tuples in each block of J? Explain your answer.
- Give a simple way to compute the most likely possible world of J and the probability of that world. Explain your answer.
- Apply this to compute the number of possible worlds of Answers, its most likely possible world, and the probability of that world.

Question 3 (1.5 points). To improve her chances, Alice wants to go a step further. She wishes to use the fact that exams are often badly designed and some answers give clues about answers to other questions. Alice thus determines that answering A to question 1 and answering B to question 2 are *mutually exclusive*: only one is possible at the same time. Likewise, answering A to 2 and A to 3 are mutually exclusive. She writes these question–answer pairs as the following deterministic relation Mutex:

Mutex

q1	a1	q2	a2
1	A	2	B
2	A	3	A

Alice now wishes to write a query that will compute *contradictions*, namely, compute what pairs of two questions with their answers occur both as a row of the Mutex table, and as two rows in the Answers table. Write this query Q:

- in the relational calculus (the query should have three atoms and four free variables);
- in the relational algebra;
- and in SQL.

Question 4 (2 points). Consider the probabilistic instance $\mathsf{R} := Q(\mathsf{Answers}, \mathsf{Mutex})$ defined by evaluating the query Q over the tables Answers and Mutex.

- Ignoring the probabilities, what are the tuples that may occur in possible worlds of the result? Explain why.
- Prove that the empty table is a possible world of R, by exhibiting one possible world of Answers that yields this result when evaluating Q.
- By reasoning on the tuples of Answers, prove that all possible worlds of R contain at most one tuple.

Question 5 (1 point).

- Prove that we cannot express R as a TID instance: there is no TID instance which defines the same probability distribution as R.

Question 6 (1.5 points).

- Compute the probability that R contains the tuple $(1, A, 2, B)$ (explain the steps).
- Compute the probability that R contains the tuple $(2, A, 3, A)$.
- Deduce a way to write R as a BID table. Explain your answer. What are the key attributes, and how many blocks are there?

Question 7 (2 points).

- Construct a table Mutex2 such that the result of evaluating $Q(\mathsf{Answers}, \mathsf{Mutex2})$ cannot be represented as a BID instance. Prove that your choice of Mutex2 satisfies this property.
- Is there a Mutex3 table such that $Q(\mathsf{Answers}, \mathsf{Mutex3})$ can be represented as a TID instance? (no need to justify)

Exercise 2: Commuting to Work (10 points)

Two drivers, Pauline and Jean, are commuting to work by car. During rush hours, they run the risk of being stuck in traffic and being late for work. They have a choice of taking a few possible paths to work, via three roads: the N118, the A10, or the A6. Sometimes, they have the possibility of working from home, hence not commuting to work. They would like to estimate their chance of getting to the office on time; when they need to commute to work, we consider that they are on time if there is *at least* one road that is not congested.

They represent their probability of having to commute to work (instead of staying home) by a Commuters relation that indicates who has to commute to work, and they use a Roads relation to represent the probability, for each worker, of being able to use a specific path to be on time. The probabilities $p_1, p_2, q_1, \ldots, q_4$ are the probabilities of *independent* events, and the tables are as follows:

Commuters	
name	
Pauline	p_1
Jean	p_2

Roads		
name	**road**	
Pauline	N118	q_1
Jean	A6	q_2
Pauline	A10	q_3
Jean	N118	q_4

Question 1 (1 point). Consider the query Q_1: "Is there someone who commutes to work and arrives on time?", where someone who has to commute arrives on time if there is *some* path that they can use to arrive on time. Write the corresponding query in the relational calculus and in the relational algebra.

Question 2 (1.5 points).

- Let Q'_1 be the query "Does Pauline decide to go to work and arrives in time?" Write Q'_1 in the relational algebra.
- From the relational algebra expression, write an *extensional query plan* for Q'_1
- Evaluate the plan on the instance above, detailing the steps.
- Is your plan safe, and why?

Question 3 (2 points). We add variables X_i and Y_i to the instance above to refer to its individual tuples as follows:

Commuters	
name	
Pauline	X_1
Jean	X_2

Roads		
name	**road**	
Pauline	N118	Y_1
Jean	A6	Y_2
Pauline	A10	Y_3
Jean	N118	Y_4

- Write the formula for the *lineage* of Q_1 as a function of the X_i and Y_i.
- Can the resulting formula be rewritten as a *read-once* formula? Explain how it can or why it cannot.
- Draw a FBDD (Free Binary Decision Diagram) of *minimal size*[2] for this formula.
- Explain what is the relation between the size of the diagram and the size of the lineage.

Question 4 (1 point). Compute the probability that Q_1 holds, as a function of p_1, p_2, q_1,... q_4, and using *intensional rules*.

[2]The size of a FBDD/OBDD is the number of conditional gates it has.

Question 5 (2 points). Consider now that Jean and Pauline use the roads at the same time. Each road (instead of each path) has a probability of being congested and thus unusable by the commuters. The updated relations for this scenario are detailed below.

Commuters'

name	
Pauline	p_1
Jean	p_2

Roads'

name	**road**
Pauline	N118
Jean	A6
Pauline	A10
Jean	N118

Free'

road	
N118	q_1
A6	q_2
A10	q_3

We wish to answer the query[3] Q_2: "Does anyone use the car and arrives at work in time?", on the relations Commuters', Roads', and Free'.

- Write the lineage of Q_2 as a function of the variables X_i and Y_i, where we assign these variables to the tuples of the relations above in the following way:

Commuters'

name	
Pauline	X_1
Jean	X_2

Roads'

name	**road**
Pauline	N118
Jean	A6
Pauline	A10
Jean	N118

Free'

road	
N118	Y_1
A6	Y_2
A10	Y_3

- Can the resulting formula be rewritten as a *read-once* formula? If it can, explain how, or state whether it cannot.

Question 6 (1.5 points). Derive the probability formula for the lineage of Q_2. What is the difference between computing the probability of the lineage of Q_1 and of the lineage of Q_2, in terms of the set of intensional rules used for each query?

Question 7 (1 point). Is the query Q_1 safe? What about Q_2? Justify your answers.

[3]Note that this is the same query as Q_1, but it is asked on a different database.