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)

Answer.

- The key attribute is the **question** attribute.
- There are three blocks: the first three tuples, the next two tuples, and the last two tuples. *(0.5 points)*

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.

Answer.

- The number of possible worlds of J is $\prod_i n_i$, where n_i is the number of tuples in the i-th block. This is because a possible world is defined by choosing, from each block, one of its tuples. Indeed, any way to choose tuples in this way yields a possible world, as tuples have non-zero probability by our assumption on J. Conversely, any two different such ways of choosing tuples cannot lead to the same possible world, because it is impossible for the same tuple to occur multiple times in J. *(0.5 points)*
- First, notice that the probability of a possible world of J is obtained as the product of the probabilities of its tuples in J, because the only way to obtain that world is to pick exactly its tuples, and the probabilities of doing so, by independence across blocks, is the product of the tuple probabilities.

  Now, by this characterization, it is clear that the most likely possible world of J is obtained by picking, in each block, the tuple with the highest probability (which is unique by our assumption on J). *(0.5 points)*

- We conclude that Answers has $3 \times 2 \times 2 = 12$ possible worlds, and that the possible world with the highest probability is:

question	answer
1	A
2	A
3	B

The probability of this world is $0.5 \times 0.8 \times 0.9 = 0.36$. *(0.5 points)*

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.

Answer.

Relational calculus. *(0.5 point)*

$$\mathsf{Mutex}(q, a, q', a') \wedge \mathsf{Answers}(q, a) \wedge \mathsf{Answers}(q', a')$$

Relational algebra. *(0.5 point)*

$$(\rho_{\mathbf{question} \mapsto \mathbf{q1}, \mathbf{answer} \mapsto \mathbf{a1}}(\mathsf{Answers}) \times \rho_{\mathbf{question} \mapsto \mathbf{q2}, \mathbf{answer} \mapsto \mathbf{a2}}(\mathsf{Answers})) \bowtie \mathsf{Mutex}$$

SQL. *(0.5 point)*

```
SELECT M.q1, M.a1, M.q2, M.a2 FROM Answers A1, Answers A2, Mutex M
    WHERE M.q1 = A1.question AND M.a1 = A1.answer
    AND M.q2 = A2.question AND M.a2 = A2.answer;
```

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.

Answer.

- The tuples that may occur in possible worlds of R are those that occur in Mutex. Indeed, all tuples on R must by definition occur in Mutex. Conversely, it is easy to see that each individual tuple of Mutex can be obtained in a possible world of R obtained by choosing tuples in Answer that match this tuple in Mutex. *(0.5 point)*
- The most probable possible world of Answers computed in Question 2 yields the empty table when evaluating Q. *(0.5 point)*
- From the answer to the first part of this question, the only way for a possible world of R to contain two tuples is that it contains the two tuples of Mutex. Let A be a possible world of Answers that yields this result when evaluating Q. As the result contains the first tuple of Mutex, we know that A contains the tuple $(2, B)$. But as the result contains the second tuple of Mutex, we know that A contains the tuple $(2, A)$. However, by definition of the BID instance Answers, these two tuples are mutually exclusive, so we have reached a contradiction. *(1 point)*

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.

Answer. Assume by contradiction that we can represent R as a TID instance. We know from the previous question that R has two possible tuples, so the TID representation of R must include these two tuples with non-zero probability. Now, by the independence of tuples in a TID instance, the TID representation has a possible world where both tuples occur. However, we also know from the previous question that R has no possible world with two tuples. Hence, we have reached a contradiction.

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?

Answer.

- R contains the tuple $(1, A, 2, B)$ whenever Answers contains the tuples $(1, A)$ and $(2, B)$. As these tuples occur in different blocks, they are independent, so the probability that they occur is $0.5 \times 0.2 = 0.1$. *(0.5 points)*
- Likewise, the probability that R contains $(2, A, 3, A)$ is $0.8 \times 0.1 = 0.08$. *(0.5 points)*
- From Question 4 we know that R has three possible worlds: the empty table, the table containing only $(1, A, 2, B)$, and the table containing $(2, A, 3, A)$. We have computed the probability of the last two worlds, and we know that the probabilities must sum up to 1. Hence, a BID representation of R is the following:

R

q1	**a1**	**q2**	**a2**	
1	A	2	B	0.1
2	A	3	A	0.08

In this BID, the set of key attributes is the empty set, there is a single block containing both tuples. *(0.5 points)*

Question 7 (2 points).

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

Answer.

- Take the following table:

Mutex2

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

Assume by contradiction that the result R' of evaluating Q can be represented as a BID table. Clearly R' should contain the two tuples of Mutex2. There are two possibilities:

– The two tuples are in the same block in R', so they are mutually exclusive. This is not possible, as there is a possible world containing both tuples, as witnessed by the possible world of Answers where we keep the first tuple of each block.

– The two tuples are in different blocks in R', so R' is equivalent to a TID table. Now, the first tuple should have probability $0.5 \times 0.8 = 0.4$ to reflect the probability of the possible worlds that make it appear, and the second tuple should have probability $0.8 \times 0.1 = 0.08$ for the same reason. This entirely defines the BID representation. However, the possible world where both tuples appear should have probability $0.5 \times 0.8 \times 0.1 = 0.04$, which is different from 0.4×0.08. Hence, the TID representation does not give the correct probability to that world, so we have a contradiction.

This concludes the proof, so that the result R' of evaluating Q with Mutex2 cannot be represented as a BID table. *(1.5 points)*

- Taking the empty table for Mutex3 ensures that evaluating Q yields an empty table, which can be represented as a TID instance. *(0.5 points)*.

  Other choices for Mutex3 are also possible, for instance making Mutex3 contain a single tuple.

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, ..., 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.

Answer. Relational calculus: $Q_1 : \exists x, y\ \mathsf{Commuters}(x) \wedge \mathsf{Roads}(x, y)$

Relational algebra: $\pi_\emptyset(\mathsf{Commuters} \bowtie \mathsf{Roads})$

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?

Answer. The query in relational algebra is:

$$\sigma_{\text{name=Pauline}}(\pi_{\text{name}}(\mathsf{Commuters} \bowtie \mathsf{Roads})).$$

There are several possible plans, but the difference between a safe and an unsafe plan is made by the order of the projection.

If the projection is done first on Roads, i.e.,

$$\sigma_{\text{name=Pauline}}(\mathsf{Commuters} \bowtie \pi_{\text{name}}\mathsf{Roads}),$$

then the final probability of Q_1' is equal to

$$p_1(1 - (1 - q_1)(1 - q_3)).$$

This is the correct probability, and hence the plan is safe.

However, if the projection is done after the join, i.e,,

$$\sigma_{\text{name=Pauline}}(\pi_{\text{name}}(\mathsf{Commuters} \bowtie \mathsf{Roads})),$$

then the final probability is

$$1 - (1 - p_1q_1)(1 - p_1q_3)$$

This is an incorrect probability, and hence the plan is unsafe.

Placing the select operator anywhere in the plan has no effect on whether the plan is safe or not. This is because the select only eliminates tuples from a relation, and does not recompute probabilities.

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.

Answer. The lineage is

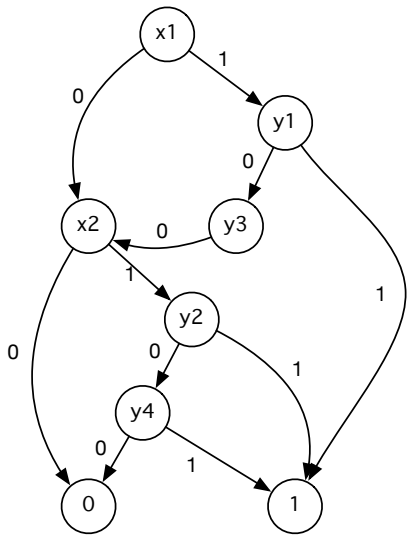
$$X_1Y_1 \vee X_2Y_2 \vee X_1Y_3 \vee X_2Y_4.$$

We can simplify it further to:

$$X_1(Y_1 \vee Y_3) \vee X_2(Y_2 \vee Y_4).$$

To evaluate this formula, we only need to read each variable once. Hence, it is a read-once formula.

A possible FBDD is



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

The diagram has the same size as the number of variables in the lineage, because it is compiling a read-once formula. That means it has 6 intermediate gates for the variables (the same as the total number of variables) and 2 for the values 1 and 0.

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*.

Answer. Starting for the lineage in Question 3, and by combining the formulas for independent OR and independent AND we obtain the probability:

$$1 - (1 - p_1(1 - (1 - q_1)(1 - q_3)))\,(1 - p_2(1 - (1 - q_2)(1 - q_4)))\,.$$

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.

Answer. The lineage formula is:

$$\Phi = X_1Y_1 \vee X_2Y_2 \vee X_1Y_3 \vee X_2Y_1.$$

The formula is not read-once, because there does not exist any rewriting that ensures that a variable is only read once in the evaluation.

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

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?

Answer. The lineage Φ in the answer to Question 5 can be rewritten as:

$$\Phi = X_1(Y_1 \vee Y_3) \vee X_2(Y_1 \vee Y_2).$$

As we can see, there are no independent parts of the lineage, due to Y_1. To solve this, we can do a Shannon expansion on the variable Y_1:

$$\Pr(\Phi) = \Pr(Y_1 = 1)\Pr(X_1 \vee X_2) + \Pr(Y_1 = 0)\Pr(X_1Y_3 \vee X_2Y_2).$$

We can use independent OR and independent AND on the resulting two formulas for the two "sides" of the expansion. This leads to the following probability formula:

$$\begin{aligned}\Pr(\Phi) = {} & q_1(1 - (1 - p_1)(1 - p_2))+ \\ & + (1 - q_1)(1 - (1 - p_1q_3)(1 - p_2y_2))\end{aligned}$$

The evaluation of Q_1 can be done linearly in the size of its lineage and only using independent intensional formulas, since it has a read-once formula and all variables are independent. On the other hand, the lineage of Q_2 needs a Shannon expansion step. After the Shannon expansion, we need to evaluate two read-once formulas.

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

Answer. The query for Q_1 is safe, since there is a safe plan for it: we can use the same safe plan as in the answer to Question 2 for each tuple in Commuters. We thus get the final probability for each tuple in Commuters. These probabilities are *independent*, so we can use independent OR to compute the final probability of Q_1. The probability is the same as the one shown in the answer to Question 4.

The query for Q_2 in relational calculus is:

$$\mathsf{Commuters'}(X), \mathsf{Roads'}(X, Y), \mathsf{Free'}(Y).$$

This is the same as $H_0 = R(X), S(X, Y), T(Y)$ and hence the query for Q_2 is unsafe.