Exercise sheet for Session 4

Uncertain data management

Antoine Amarilli

December 12th, 2016

1 Exercise 1: From BID to pc-instances

The Bureau for Invasion and Domination in Westeros, using cutting-edge machine learning techniques, has determined an uncertain forecast of the future political state of key cities in the Seven Kingdoms, represented as the following BID instance. (The absence of a city in the table means that the city was sacked and burnt to the ground.)

Westeros

city	**house**	
Winterfell	Stark	0.3
Winterfell	Greyjoy	0.2
Winterfell	Bolton	0.4
King's Landing	Lannister	0.5
King's Landing	Baratheon	0.2
King's Landing	Targaryen	0.2
King's Landing	Stark	0.1

Question 1. Suppose that we know that King's Landing fell to house Stark. What is the probability, knowing this information, that Winterfell is also controlled by house Stark?

Answer. *The probability is unchanged, i.e.,* 0.3*, because probabilistic choices are done independently across blocks.*

Question 2. Using the technique shown in class, write a pc-table that represents the same probabilistic relation.

Answer. *The pc-table is:*

city	***house***	
Winterfell	*Stark*	$\neg x_1 \wedge \neg x_2$
Winterfell	*Greyjoy*	$\neg x_1 \wedge x_2$
Winterfell	*Bolton*	$x_1 \wedge \neg x'_2$
King's Landing	*Lannister*	$\neg y_1 \wedge \neg y_2$
King's Landing	*Baratheon*	$\neg y_1 \wedge y_2$
King's Landing	*Targaryen*	$y_1 \wedge \neg y'_2$
King's Landing	*Stark*	$y_1 \wedge y'_2$

The probability that each variable is true is:

- x_1*: 0.5*
- x_2*: 0.4*
- x_2'*: 0.2*
- y_1*: 0.3*
- y_2*: 2/7*
- y_2'*: 1/3*

Question 3. Describe in plain English the meaning of the variables occurring in the resulting pc-instance for the tuples containing "Winterfell".

Answer.

- x_1 *is true iff Winterfell is controlled by House Bolton or no longer stands*
- x_2 *is only relevant if* x_1 *is false, in which case it is true iff Winterfell is controlled by house Greyjoy*
- x_2' *is only relevant if* x_1 *is true, in which case it is true iff Winterfell no longer stands*

Question 4. Write in the relational algebra, then in the relational calculus, a query Q_1 that tests whether there is some house that controls two different cities.

Answer.

$$Q_1 : \Pi_{\emptyset} \left(\sigma_{\boldsymbol{house=house2, city \neq city2}}(\rho_{\boldsymbol{city \mapsto city2, house \mapsto house2}}(\mathit{Westeros}) \times \mathit{Westeros})\right)$$

$$Q_1 : \exists c\, c'\, h \;\; \mathit{Westeros}(c, h) \wedge \mathit{Westeros}(c', h) \wedge c \neq c'$$

Question 5. What is the probability that the pc-instance satisfies Q_1? How many possible worlds of the pc-instance instance satisfy Q_1? How many valuations of the pc-table satisfy Q_1?

Answer. *The probability of* Q_1 *is* $0.3 \times 0.1 = 0.03$*. The only world of the probabilistic instance that satisfies the query is:*

Westeros

city	***house***
Winterfell	*Stark*
King's Landing	*Stark*

To satisfy the query in the pc-instance, we must obtain the only possible world that satisfies the query, which means that y_1 *and* y_2' *must be true and* x_1 *and* x_2 *must be false; no constraints are imposed on* x_2' *and* y_2*. Conversely, all these valuations yield the required possible world. So there are four valuations of the pc-instance that make the query true.*

Question 6. Write in the relational algebra, then in the relational calculus, a query Q_2 that computes which houses control at least one city.

Answer.

$$Q_2 : \pi_{\boldsymbol{house}}(Westeros)$$

$$Q_2 : \exists c \ \ Westeros(c, h)$$

Question 7. Construct a pc-table to represent the output of the query Q_2(Westeros). How many possible worlds does it have? Compute the probability that house Stark holds some city.

Answer. *The pc-table is the following, with the same probabilities as before:*

house	
Stark	$\neg x_1 \wedge \neg x_2 \vee y_1 \wedge y_2'$
Greyjoy	$\neg x_1 \wedge x_2$
Bolton	$x_1 \wedge \neg x_2'$
Lannister	$\neg y_1 \wedge \neg y_2$
Baratheon	$\neg y_1 \wedge y_2$
Targaryen	$y_1 \wedge \neg y_2'$

There are fifteen different possible worlds. The probability that house Stark holds a city can be computed directly from the annotation, using the fact that the variables are independent*, as:*

$$1 - (1 - p(x_1) \times (1 - p(x_2))) \times (1 - p(y_1) \times p(y_2'))$$

This evaluates numerically to $\frac{37}{100}$.

Question 8. Prove that Q_2(Westeros) cannot be represented by a TID instance.

Answer. *Any TID instance has a* maximal possible world*, which contains all the tuples of the relation. However, to cover all required possible worlds, the TID would have to contain contain all the tuples of the pc-instance $Q_2(Westeros)$. However, the instance containing all of these tuples is not a possible world. Hence, we cannot represent $Q_2(Westeros)$ as a TID instance.*

Question 9. Prove that Q_2(Westeros) cannot be represented by a BID instance.

Answer. *Either the attribute* **house** *is a key, or it is not a key. If it is a key, then the BID instance amounts to a TID instance, and we conclude by the previous question. If it is not a key, then the BID instance only has possible worlds containing of a single tuple, so it would not represent the possible worlds of $Q_2(Westeros)$ that contain two tuples.*

2 Exercise 2: Adding probabilities to a c-table

Consider the (slightly modified) Boolean c-table obtained in the first question of Exercise 3 for Session 2:

Classes

session	**date**	**prof**	**room**	
2	Nov 30	Antoine	C017	
3	Dec 7	Antoine	C47	$\neg x_1$
4	Dec 14	Silviu	C47	$\neg x_1 \wedge \neg x'_3$
5	Jan 4	Silviu	C47	$\neg x_1 \wedge x_2$
6	Jan 11	Silviu	C47	$\neg x_1 \wedge x_2$

Recall also the semantics of the events:

- x_1: Room C47 collapses. All UDM classes in room C47 must be canceled.
- x_2: D&K students accept to return from vacation. If this does *not* happen, all UDM classes in January are cancelled.
- x'_3: Silviu is sick on December 14, we must cancel this class.

Question 1. Assign the following probabilities that the variables are true:

- x_1: 0.01
- x_2: 0.2
- x'_3: 0.1

Consider the resulting pc-instance. For each tuple, compute the probability that this tuple is present, i.e., the total mass of the possible worlds where it occurs. Construct the TID instance Classes2 on the same tuples, where each tuple carries this probability.

Answer. *The TID instance is:*

Classes2

session	***date***	***prof***	***room***	
2	*Nov 30*	*Antoine*	*C017*	1
3	*Dec 7*	*Antoine*	*C47*	0.99
4	*Dec 14*	*Silviu*	*C47*	0.99×0.9
5	*Jan 4*	*Silviu*	*C47*	0.99×0.2
6	*Jan 11*	*Silviu*	*C47*	0.99×0.2

Question 2. Consider the Boolean query Q that asks whether Silviu teaches a class. Write it in the relational algebra and in the relational calculus.

Answer.

$$Q : \pi_{\emptyset}(\sigma_{\textbf{\textit{prof}}= \text{“}Silviu\text{”}}(Classes))$$

$$Q : \exists s\, d\, r \;\; Classes(s, d, \text{“}Silviu\text{”}, r)$$

Question 3. Consider the result of evaluating Q on the pc-table. What would be the Boolean formula that would annotate the one empty tuple of this result?

Answer.

$$\neg x_1 \wedge \neg x_3' \vee \neg x_1 \wedge x_2$$

Question 4. What is the probability of this Boolean formula? What to conclude about Q on the pc-instance?

Answer. *The probability of this annotation is:*

$$0.99 \times (1 - 0.1 \times (1 - 0.2))$$

This evaluates to $\frac{9108}{10000}$. *We deduce that Q is true on the pc-instance with this probability.*

Question 5. How does this compare to the probability that the TID instance satisfies query Q? What to conclude about the pc-instance and the TID instance?

Answer. *The probability that the TID instance satisfies Q is:*

$$1 - (1 - .99 \times .9) \times (1 - .99 \times .2)^2$$

This is clearly much smaller. This implies that the TID instance and pc-instance capture different possible worlds, even though the probability of each individual tuple is the same.

Question 6. We observe that the December 7th class has taken place. We formally define the distribution on possible worlds *conditioned* by this observation, where the universe of possible worlds consists of those consistent with the observation, and the probability of each world is its probability in the initial instance, divided by the total probability of the possible world of the initial instance that satisfy the observation.

Write a pc-instance Classes3 describing the conditioned distribution.

Answer. *The pc-instance, with the same probabilities as before, is:*

Classes3

session	***date***	***prof***	***room***	
2	*Nov 30*	*Antoine*	*C017*	
3	*Dec 7*	*Antoine*	*C47*	
4	*Dec 14*	*Silviu*	*C47*	$\neg x_3'$
5	*Jan 4*	*Silviu*	*C47*	x_2
6	*Jan 11*	*Silviu*	*C47*	x_2