Assignment 3 - Part 1

Probabilistic Reasoning

Deadline: April 14th, 11:55pm.
Perfect score: 50.

Assignment Instructions:

Teams: Assignments should be completed by teams of up to three students. You can work on this assignment individually if you prefer. No additional credit will be given for students that complete an assignment individually. Make sure to write the name and RUID of every member of your group on your submitted report.

Submission Rules: Submit your reports electronically as a PDF document through Canvas (`canvas.rutgers.edu`). For programming questions, you need to also submit a compressed file via Canvas, which contains your code. Do not submit Word documents, raw text, or hardcopies etc. Make sure to generate and submit a PDF instead. Each team of students should submit only a single copy of their solutions and indicate all team members on their submission. Failure to follow these rules will result in lower grade for this assignment.

Program Demonstrations: You will need to demonstrate your program to the TAs on a date after the deadline. The schedule will be coordinated by the TAs. During the demonstration you have to use the file submitted on Canvas and execute it on your personal computer. You will also be asked to describe the architecture of your implementation and key algorithmic aspects of the project. You need to make sure that you are able to complete the demonstration and answer the TAs' questions within the allotted 10 minutes of time for each team. If your program is not directly running on the computer you are using and you have to spend time to configure your computer, this counts against your allotted time.

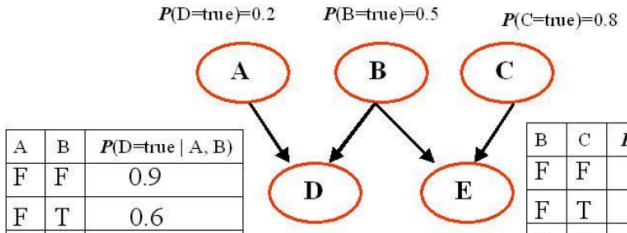
Late Submissions: No late submission is allowed. 0 points for late assignments.

Extra Credit for LaTeX: You will receive 10% extra credit points if you submit your answers as a typeset PDF (using LaTeX, in which case you should also submit electronically your source code). If you want to submit a handwritten report, scan it and submit a PDF via Canvas. We will not accept hardcopies. If you choose to submit handwritten answers and we are not able to read them, you will not be awarded any points for the part of the solution that is unreadable.

Precision: Try to be precise. Have in mind that you are trying to convince a very skeptical reader (and computer scientists are the worst kind...) that your answers are correct.

Collusion, Plagiarism, etc.: Each team must prepare its solutions independently from other teams, i.e., without using common notes, code or worksheets with other students or trying to solve problems in collaboration with other teams. You must indicate any external sources you have used in the preparation of your solution. **Do not plagiarize online sources** and in general make sure you do not violate any of the academic standards of the department or the university. Failure to follow these rules may result in failure in the course.

Problem 1 (*15 points):* Consider the following Bayesian network, where variables A through E are all Boolean valued. Note: there is a typo in the image, it should be $P(A = true) = 0.2$ instead of $P(D = true) = 0.2$.



A	B	P(D=true | A, B)
F	F	0.9
F	T	0.6
T	F	0.5
T	T	0.1

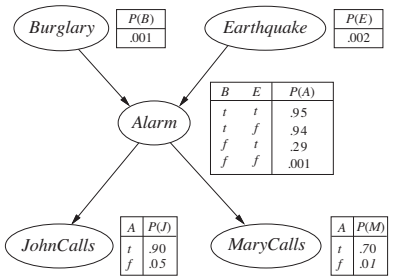
B	C	P(E=true | B, C)
F	F	0.2
F	T	0.4
T	F	0.8
T	T	0.3

a) What is the probability that all five of these Boolean variables are simultaneously true?
[Hint: You have to compute the joint probability distribution. The structure of the Bayesian network suggests how the joint probability distribution is decomposed to the conditional probabilities available.]

b) What is the probability that all five of these Boolean variables are simultaneously false?
[Hint: Answer similarly to above.]

c) What is the probability that A is false given that the four other variables are all known to be true?

Problem 2 (*15 points):*



a) Calculate $P(Burglary|JohnsCalls = true, MaryCalls = true)$ and show in detail the calculations that take place. Use your book to confirm that your answer is correct.

b) Suppose a Bayesian network has the from of a *chain*: a sequence of Boolean variables $X_1, \ldots X_n$ where $Parents(X_i) = \{X_{i-1}\}$ for $i = 2, \ldots, n$. What is the complexity of computing $P(X_1|X_n = true)$ using enumeration? What is the complexity with variable elimination?

Problem 3 (*20 points):* Suppose you are working for a financial institution and you are asked to implement a fraud detection system. You plan to use the following information:

- When the card holder is travelling abroad, fraudulent transactions are more likely since tourists are prime targets for thieves. More precisely, 1% of transactions are fraudulent when the card holder is travelling, where as only 0.4% of the transactions are fraudulent when she is not travelling. On average, 5% of all transactions happen while the card holder is travelling. If a transaction is fraudulent, then the likelihood of a foreign purchase increases, unless the card holder happens to be travelling. More precisely, when the card holder is not travelling, 10% of the fraudulent transactions are foreign purchases where as only 1% of the legitimate transactions are foreign purchases. On the other hand, when the card holder is travelling, then 90% of the transactions are foreign purchases regardless of the legitimacy of the transactions.
- Purchases made over the internet are more likely to be fraudulent. This is especially true for card holders who don't own any computer. Currently, 75% of the population owns a computer or smart phone and for those card holders, 1% of their legitimate transactions are done over the internet, however this percentage increases to 2% for fraudulent

transactions. For those who don't own any computer or smart phone, a mere 0.1% of their legitimate transactions is done over the internet, but that number increases to 1.1% for fraudulent transactions. Unfortunately, the credit card company doesn't know whether a card holder owns a computer or smart phone, however it can usually guess by verifying whether any of the recent transactions involve the purchase of computer related accessories. In any given week, 10% of those who own a computer or smart phone purchase (with their credit card) at least one computer related item as opposed to just 0.1% of those who don't own any computer or smart phone.

a) Construct a Bayes Network to identify fraudulent transactions.
What to hand in: Show the graph defining the network and the Conditional Probability Tables associated with each node in the graph. This network should encode the information stated above. Your network should contain exactly six nodes, corresponding to the following binary random variables:
OC : card holder owns a computer or smart phone.
Fraud : current transaction is fraudulent.
Trav : card holder is currently travelling.
FP : current transaction is a foreign purchase.
IP : current purchase is an internet purchase.
CRP : a computer related purchase was made in the past week.
The arcs defining your Network should accurately capture the probabilistic dependencies between these variables.

b) What is the prior probability (i.e., before we search for previous computer related purchases and before we verify whether it is a foreign and/or an internet purchase) that the current transaction is a fraud? What is the probability that the current transaction is a fraud once we have verified that it is a foreign transaction, but not an internet purchase and that the card holder purchased computer related accessories in the past week?
What to hand in: Indicate the two queries (i.e., $Pr(variables|evidence)$) you used to compute those two probabilities. Show each step of the calculation