CMPUT 366, Winter 2021
Assignment #2

Due: Friday, Mar. 5, 2021, 11:59pm
Total points: 98

For this assignment use the following consultation model:

1. you can discuss assignment questions and exchange ideas with other *current* CMPUT 366 students;
2. you must list all members of the discussion in your solution;
3. you may **not** share/exchange/discuss written material and/or code;
4. you must write up your solutions individually;
5. you must fully understand and be able to explain your solution in any amount of detail as requested by the instructor and/or the TAs.

Anything that you use in your work and that is not your own creation must be properly cited by listing the original source. Failing to cite others' work is plagiarism and will be dealt with as an academic offense.

First name: Vicky

Last name: Zhao

CCID: ziwei11@ualberta.ca

Collaborators: ______________

These are example solutions for assignment 2.

1. **(Probability theory)**

   (a) **[20 points]**

   Consider the following scenario. 2% of the people who walk through a specific metal detector at YEG are carrying a gun. 30% of the people who walk through the same metal detector are carrying coins. The remaining 68% are carrying nothing made of metal. Everyone carries either nothing, coins, or a gun through the detector; never both coins and a gun.

   If someone carries a gun through this metal detector, it will beep with probability 95%. If someone carries coins through this same metal detector, it will beep with probability 80%. If someone carries nothing made of metal through the detector, it will still beep about 25% of the time.

   Suppose that the metal detector beeps when someone walks through it. With what probability is that person carrying a gun? Show how you calculated your answer.

$$P(gun \mid beeps) = \frac{P(gun \cap beeps)}{P(beeps)}$$

$$= \frac{P(gun \cap beeps)}{P(gun \cap beeps) + P(coins \cap beeps) + P(nothing \cap beeps)}$$

$$= \frac{P(beeps \mid gun) \times P(gun)}{P(beeps \mid gun) \times P(gun) + P(beeps \mid coins) \times P(coins) + P(beeps \mid nothing) \times P(nothing)}$$

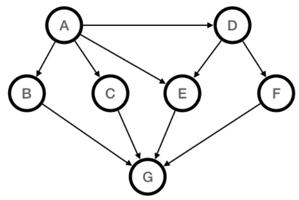
$$= \frac{95\% \times 2\%}{95\% \times 2\% + 80\% \times 30\% + 25\% \times 68\%}$$

$$= \frac{0.019}{0.019 + 0.24 + 0.17}$$

$$= \frac{0.019}{0.429} \approx 0.0442890 = 4.4289\%$$

2. (**Belief networks**)

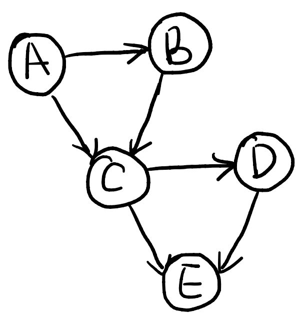
   (a) [**5 points**] What factorization of the joint distribution $P(A, B, C, D, E, F, G)$ does the network below represent?

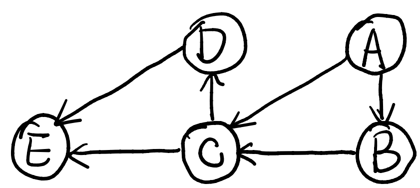


$$P(A,B,C,D,E,F,G)$$
$$= P(A)P(D|A)P(B|A)P(C|A)P(E|A,D)P(F|D)P(G|B,C,E,F)$$

   (b) [**5 points**] Draw a belief network that is consistent with a joint distribution that factors as $P(A, B, C, D, E) = P(E|C, D)P(D|C)P(C|A, B)P(B|A)P(A)$.

   For **5 bonus marks**, draw another, different belief network that is *also* consistent with this factoring.





   (c) [**3 points**] Suppose that every random variable in the joint distribution of question (2b) has a domain containing 10 elements. How many rows are needed to list the full joint distribution in an explicit table?

   There are 5 variables, so 10^5 rows are needed to list the full joint distribution.

   (d) [**7 points**] Suppose that every random variable in the joint distribution of question (2b) has a domain containing 10 elements. How many rows in total are needed to list the conditional probability tables for your belief network representation?

A: $P(A) \Longrightarrow 10$ rows

B: $P(B|A) \Longrightarrow 10 \times 10$ rows $= 100$ rows.

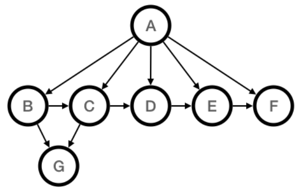
C: $P(C|A,B) \Longrightarrow 10^3$ rows $= 1000$ rows.

D: $P(D|C) \Longrightarrow 10^2$ rows $= 100$ rows

E: $P(E|C,D) \Longrightarrow 10^3$ rows $= 1000$ rows

In total: $10 + 100 + 1000 + 100 + 1000 = 2210$ rows

3. **(Variable Elimination)** Consider the belief network below.



(a) [**15 points**] List the factors that would be created, and the operations used to create them, by running the variable elimination algorithm on this belief network to answer the query $P(B|G, E)$. Use the variable ordering G, E, A, B, C, D, F.

(b) [**15 points**] List the factors that would be created, and the operations used to create them, by running the variable elimination algorithm on this belief network to answer the query $P(B|G, E)$. Use the variable ordering G, E, F, D, C, B, A

(c) [**5 points**] Which of the two given variable orderings is more efficient for this query? Justify your answer. You may assume that the domain of each variable is the same size.

(a) $P(A,B,C,D,E,F,G)$

$$= P(A)\,P(B|A)\,P(C|A,B)\,P(D|A,C)\,P(E|A,D)\,P(F|A,E)\,P(G|B,C)$$

Construct factors for each table:

$\{f_0(A), f_1(B,A), f_2(C,A,B), f_3(D,A,C), f_4(E,A,D), f_5(F,A,E), f_6(G,B,C)\}$

Condition on G : $f_7 = (f_6)_{G=true}$

$\{f_0(A), f_1(B,A), f_2(C,A,B), f_3(D,A,C), f_4(E,A,D), f_5(F,A,E), f_7(B,C)\}$

Condition on E: $f_8 = (f_4)_{E=true}$, $f_9 = (f_5)_{E=true}$

$\{f_0(A), f_1(B,A), f_2(C,A,B), f_3(D,A,C), f_8(A,D), f_9(F,A), f_7(B,C)\}$

Variable ordering: ~~G, E~~, ~~A, B, C, D, F~~

$\{f_0(A), f_1(B,A), f_2(C,A,B), f_3(D,A,C), f_8(A,D), f_9(F,A), f_7(B,C)\}$

Sum out A from product of $f_0, f_1, f_2, f_3, f_8, f_9$:

$$f_{10} = \Sigma_A (f_0 \times f_1 \times f_2 \times f_3 \times f_8 \times f_9)$$

$\{f_{10}(B,C,D,F), f_7(B,C)\}$

Sum out C from product of f_{10}, f_7 : $f_{11} = \Sigma_C (f_{10} \times f_7)$

$\{f_{11}(B,D,F)\}$

Sum out D from product of f_{11} : $f_{12} = \Sigma_D (f_{11})$

$\{f_{12}(B,F)\}$

Sum out F from product of f_{12} : $f_{13} = \Sigma_F (f_{12})$

$\{f_{13}(B)\}$

Normalize by division:

$query(B) = f_{13}(B) / (\Sigma_B f_{13}(B))$

(b) $P(A, B, C, D, E, F, G)$

$= P(A) P(B|A) P(C|A,B) P(D|A,C) P(E|A,D) P(F|A,E) P(G|B,C)$

Construct factors for each table:

$\{f_0(A), f_1(B,A), f_2(C,A,B), f_3(D,A,C), f_4(E,A,D), f_5(F,A,E), f_6(G,B,C)\}$

Condition on G : $f_7 = (f_6)_{G=true}$

$\{f_0(A), f_1(B,A), f_2(C,A,B), f_3(D,A,C), f_4(E,A,D), f_5(F,A,E), f_7(B,C)\}$

Condition on E: $f_8 = (f_4)_{E=true}$, $f_9 = (f_5)_{E=true}$

$\{f_0(A), f_1(B,A), f_2(C,A,B), f_3(D,A,C), f_8(A,D), f_9(F,A), f_7(B,C)\}$

Variable ordering: ~~G, E~~, ~~F, D, C, B, A~~

$\{f_0(A), f_1(B,A), f_2(C,A,B), f_3(D,A,C), f_8(A,D), f_9(F,A), f_7(B,C)\}$

Sum out F from product of f_9 : $f_{10} = \sum_F (f_9)$

$\{f_0(A), f_1(B,A), f_2(C,A,B), f_3(D,A,C), f_8(A,D), f_{10}(A), f_7(B,C)\}$

Sum out D from product of f_3, f_8 : $f_{11} = \sum_D (f_3 \times f_8)$

$\{f_0(A), f_1(B,A), f_2(C,A,B), f_{11}(A,C), f_{10}(A), f_7(B,C)\}$

Sum out C from product of f_2, f_{11}, f_7 : $f_{12} = \sum_C (f_2 \times f_{11} \times f_7)$

$\{f_0(A), f_1(B,A), f_{12}(A,B), f_{10}(A)\}$

Sum out A from product of f_0, f_1, f_{12}, f_{10} : $f_{13} = \sum_A (f_0 \times f_1 \times f_{12} \times f_{10})$

$\{f_{13}(B)\}$

Normalize by division:

$query(B) = f_{13}(B) / (\sum_B f_{13}(B))$

(C) For order 1:

Step Condition on G : 3 variables

Step Condition on E : 4 variables

Step Sum of A : 5 variables

Step sum of C : 4 variables

Step sum of D : 3 variables

Step sum of F : 2 variables

Total : $3+4+5+4+3+2 = 21$

For order 2:

Step Condition on G : 3 variables

Step Condition on E : 4 variables

Step Sum of F : 2 variables

Step sum of D : 3 variables

step sum of C : 3 variables

Step sum of A : 2 variables

Total : $3+4+2+3+3+2=17$

$17 < 21$

Thus, order 2 : G, E, F, D, C, B, A is more efficient for this query.

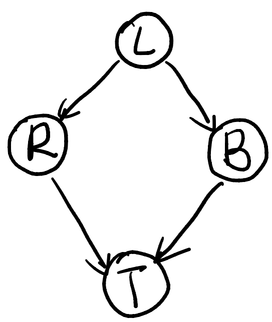
4. **(Causal inference)**

   Consider the following causal model containing random variables {L, R, B, T}, with dom(T) = $\{high, low\}$, dom(B) = $\{many, few\}$, and all other variables having domain $\{true, false\}$.

   The variable L indicates that the parents in a house like to read. The variable R indicates that the parents in a house read to the children in the house. The variable B indicates whether there are few or many books in the house. The variable T indicates whether the children in the house get high or low scores on reading tests.

   Parents who like to read (L) are more likely to read to (R). Parents who like to read (L) are also more likely to have lots of books (B) in their house. Both of being read to (R) and having lots of books (B) in the house have a causal influence on a child's performance on reading tests (T).

   (a) [**10 points**] Draw a directed graph representing the causal model.



   (b) [**3 points**] What factorization is represented by the causal model of question (4a)?

$$P(L,R,B,T) = P(T \mid R,B)\,P(B \mid L)\,P(R \mid L)\,P(L)$$

   (c) [**2 points**] Give an expression for the observational query

$$P(T = high \mid B = many)$$

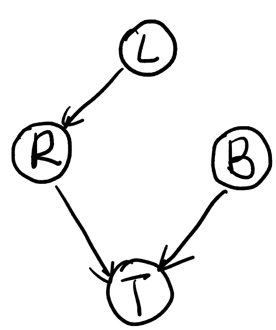
   using the factors listed in question (4b).

$$P(T=high \mid B=many) = \frac{P(T=high, B=many)}{P(B=many)} = \frac{\sum_{L,R} P(L,R,B=many,T=high)}{\sum_{L,R,T} P(L,R,B=many,T=high)}$$

$$= \frac{\sum_{L,R} P(T=high \mid R, B=many)\,P(B=many \mid L)\,P(R \mid L)\,P(L)}{\sum_{L,R,T} P(T=high \mid R, B=many)\,P(B=many \mid L)\,P(R \mid L)\,P(L)}$$

   (d) [**5 points**] Draw a directed graph representing the post-intervention distribution for the causal query

$$P(T = high \mid do(B = many)).$$



(e) **[3 points]** Give an expression for the causal query

$$P(T = high \mid do(B = many))$$

using the factors listed in question (4b).

$P(T=high \mid do(B=many))$

$= \hat{P}(T=high \mid B=many)$

$$= \frac{\sum_{L,R} P(T=high \mid R, B=many)\, P(R \mid L)\, P(L)}{\sum_{L,R,T} P(T=high \mid R, B=many)\, P(R \mid L)\, P(L)}$$

Submission

The assignment you downloaded from eClass is a single ZIP archive which includes this document as a PDF *and* its LaTeX source.

Each assignment is to be submitted electronically via eClass by the due date. **Your submission must be a a single PDF file containing your answers.**

To generate the PDF file with your answers you can do any of the following:

- insert your answers into the provided LaTeX source file between `\begin{answer}` and `\end{answer}`. Then run the source through LaTeX to produce a PDF file;
- print out the provided PDF file and legibly write your answers in the blank spaces under each question. Make sure you write as legibly as possible for we cannot give you any points if we cannot read your hand-writing. Then scan the pages and include the scan in your ZIP submission to be uploaded on eClass;
- use your favourite text processor and type up your answers there. Make sure you number your answers in the same way as the questions are numbered in this assignment.