CMPT 310
Assignment 4
Uncertainty

Jiadi Luo 301354107
Haoxuan Zhao 301385109

Instructions:

- You can do this assignment individually or in a team of two. If you are doing it in a group, only one submission per group is required.
- You may submit multiple times until the deadline. Grade penalties will be imposed for late submissions (see the course outline for the details).
- Always plan before coding.
- For coding questions - Use function-level and inline comments throughout your code. We will not be specifically grading documentation. However, remember that you will not be able to comment on your code unless sufficiently documented. Take the time to document your code as you develop it properly.
- We will carefully analyze the code submitted to look for plagiarism signs, so please do not do it! If you are unsure about what is allowed, please talk to an instructor or a TA.
- Make sure that in your answers, you clearly indicate the exact section you are answering.
- Check A3 rubric available on canvas.
- You will submit one .zip file.
    - o For theoretical questions - Your submission must be formatted as a single PDF file; other types of submissions (non-pdf, multiple files, etc.) will not be graded. Your name(s) and student number(s) must appear at the top of the first page of your submission.
    - o Make sure that in your answers, you clearly indicate the exact section you are answering.
- Please put all your files (pdf + code) in one folder. Compress these into a single archive named a3.zip and submit it on Canvas before the due date listed there.

Ques 1. [20 marks] - Probabilities

Tom has changed car's oil	Tom speeding	Tom crashes	p
F	F	F	0.18
F	F	T	0.12
F	T	F	0.06
F	T	T	0.24
T	F	F	0.12
T	F	T	0.08
T	T	F	0.04

T	T	T	0.16

Tom is sick today, but he has decided to drive his car. You can see the joint probability distribution for different events and situations in the above table.

a) [5 points] Compute the p(Tom crashes = T| Tom has changed car's oil = F).
   p(Tom crashes = T | Tom has changed car's oil = F)
   = (0.12 + 0.24)/(0. 18 + 0.12 + 0.06 + 0.24)
   = 0.6

b) [10 points] List all the independencies you can find. Prove it using formulas you have learned.
   1. Tom has changed car's oil is marginally independent of Tom speeding
      proof:
      p(Tom has changed car's oil = T | Tom speeding = T)
      = (0.04 + 0.16) / (0.06 + 0.24 + 0.04 + 0.16)
      = 0.4
      is equal to
      p(Tom has changed car's oil = T)
      = 0.12 + 0.08 + 0.04 + 0.16
      = 0.4

   2. Tom has changed car's oil is marginally independent of Tom crashes
      proof:
      p(Tom has changed car's oil = T | Tom crashes = T)
      = (0.08 + 0.16) / (0.12 + 0.24 + 0.08 + 0.16)
      = 0.4
      is equal to
      p(Tom has changed car's oil = T)
      = 0.12 + 0.08 + 0.04 + 0.16
      = 0.4

   3. Tom has changed car's oil is conditionally independent of Tom speeding given Tom crashes
      proof:
      p(Tom has changed car's oil = T | Tom speeding = T, Tom crashes = T)
      = 0.16 / (0.16 + 0.24)
      = 0.4
      is equal to
      p(Tom has changed car's oil = T | Tom crashes = T)
      = (0.08 + 0.16) / (0.12 + 0.24 + 0.08 + 0.16)
      = 0.4

   4. Tom has changed car's oil is conditionally independent of Tom crashes given Tom speeding

proof:
p(Tom has changed car's oil = T | Tom crashes = T, Tom speeding = T)
= 0.16 / (0.16 + 0.24)
= 0.4
is equal to
= **p(Tom has changed car's oil = T | Tom speeding = T)**
= (0.04 + 0.16) / (0.06 + 0.24 + 0.04 + 0.16)
= 0.4

c) [5 points] Examine the correctness of chain rule for calculating the $P(Tom\ has\ changed\ car's\ oil\ =\ T,\ Tom\ speeding\ =\ T,\ Tom\ crashes\ =\ T)$.

p(Tom has changed car's oil = T, Tom speeding = T, Tom crashes = T)
In chain rule:
p(Tom has changed car's oil = T) *
p(Tom speeding = T | Tom has changed car's oil = T) *
p(Tom crashes = T | Tom has changed car's oil = T, Tom speeding = T)
= (0.12 + 0.08 + 0.04 + 0.16) * [(0.04 + 0.16) / (0.12 + 0.08 + 0.04 + 0.16)] * [0.16 / (0.16 + 0.04)] = 0.4 * 0.5 * 0.8
= 0.16
is equal to
p(Tom has changed car's oil = T) * p(Tom speeding = T) *
p(Tom crashes = T | Tom speeding = T)
= 0.4 * 0.5 * (0.16 + 0.24) / (0.16 + 0.04 + 0.24 + 0.06) = 0.4 * 0.5 * (0.4 / 0.5)
= 0.16

Ques 2. [20 marks] You are catching Pokemons in the wild, there are only 4 species of Pokemons available to catch. Pokemons can flee during a catch, so you might need to try multiple times before successfully catching a Pokemon. Some species of Pokemons flee faster, thus are harder to be caught, than other species.

Species	Proportion of all Pokemons	Probability of a successful catch at first try
Pikachu	0.46	0.2
Charmander	0.16	0.3
Eevee	0.18	0.4
Snorlax	0.20	0.5

a). [5 marks] What is the probability that you successfully caught a Pokemon at your first try?

p(successfully caught a Pokemon at your first try)

= p(successful catch at first try| Pikachu) * p(Pikachu) +

p(successful catch at first try| Charmander) * p(Charmander) +

p(successful catch at first try| Eevee) * p(Eevee) +

p(successful catch at first try| Snorlax) * p(Snorlax)

= 0.2 * 0.46 + 0.3 * 0.16 + 0.4 * 0.18 + 0.5 * 0.2

= 0.312

b). [15 marks] Each species of Pokemon has a special magic power depending on its type. A charmander is a fire-type Pokemon, a Pikachu is an electric-type Pokemon, Eevee and Snorlax are both normal-type Pokemons. If a Pokemon was successfully caught at first try, what is the probability that it was a normal-type Pokemon?

p(normal-type | successful catch at first try)

= p(normal-type, successful catch at first try) / p(successful catch at first try)

= p((Eevee or Snorlax), successful catch at first try) / p(successful catch at first try)

= [p(Eevee, successful catch at first try) + p(Snorlax, successful catch at first try)] /

p(successful catch at first try)

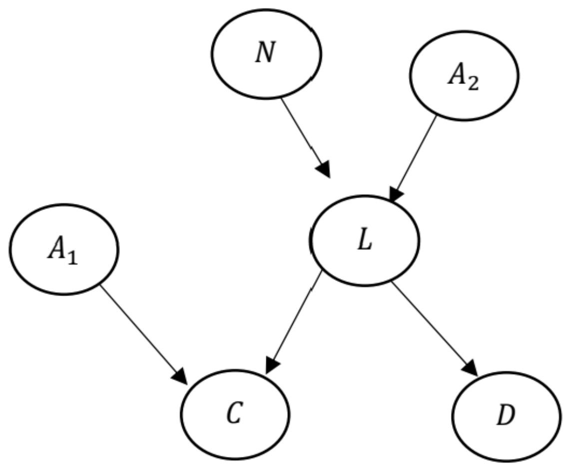
= [p(successful catch at first try | Eevee) * p(Eevee) +

p(successful catch at first try | Snorlax) * p(Snorlax)] / p(successful catch at first try)

= [(0.4 * 0.18) + (0.5 * 0.2)] / 0.312

= 0.55128205

Ques 3. [10 marks] Consider the Bayesian Network below representing a problem with random variables. Let A_1, A_2, and D be Boolean variables and C, N, and L have 31 possible values each. Between the Bayesian network and the joint distribution, compute the **representational savings** (Number of fewer values that need to be stored). Show your work.



JPD needs to store at least:

2^3 * 31^3 - 1= 238327 entries.

Bayesian Network needs to store at least:

31 - 1 = 30 entries for N

2 - 1 = 1 entry for A2

2 - 1 = 1 entry for A1

2 * 31 * (31 - 1) entries for L | N, A2

2 * 31 * (31 - 1) entries for C | A1, L

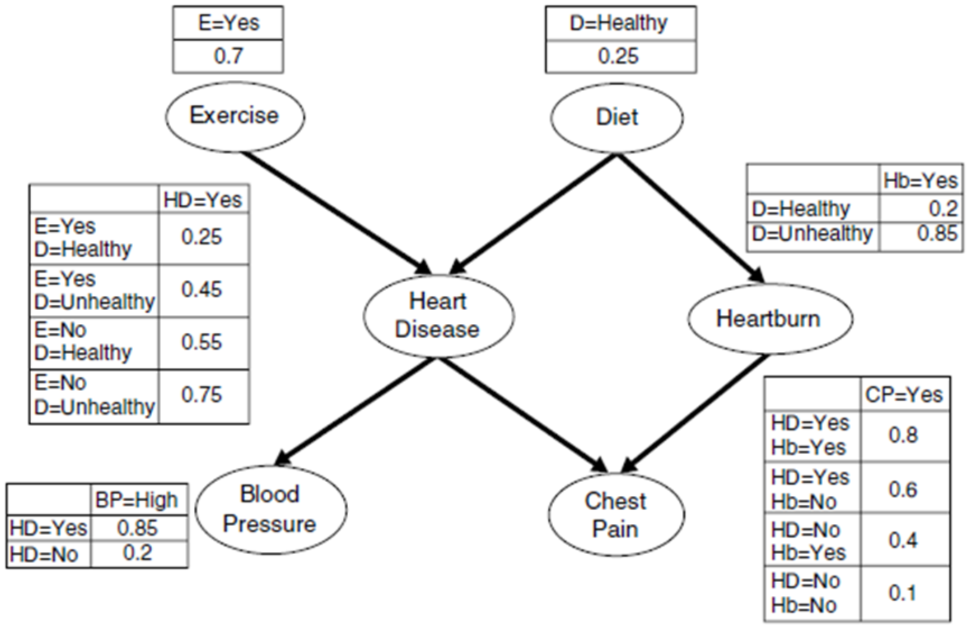
31 * (2 - 1) entries for D | L

In total: 30 + 2 * 1 + 2 * (2 * 31 * 30) + 31 = 3783 entries

Savings = 238327 - 3783 = 234544 entries

Bonus question

(Coding) Ques 4 [50 marks]. In this question, you are going to write code which can perform inference by calculating any conditional probability given the observation of some variables. Consider the following Bayesian Network which shows the connection between different lifestyles and heart-related diseases:



Variable *E* determines if the patient does exercises regularly.
Variable *D* shows that the patient eats healthy meals.
Variable *HD* means that the patient has heart disease.
Variable *Hb* shows that the patient is experiencing heartburn.
Variable *BP* determines if the patent is experiencing high blood pressure.
Variable *CP* means that the patent feels pain in their chest.

a). [10 marks] As you can see the tables in the above Bayesian Network are in compact forms. Rewrite the tables in such a way that each variable has its own separate column (both parent and child variables) beside the probability column, which comes last.

Exercise (E):

E	P(E)
Yes	0.7
No	0.3

Diet (D):

D	P(D)
Healthy	0.25
Unhealthy	0.75

Heart Disease (HD):

E	D	HD	P(HD | E, D)
Yes	Healthy	Yes	0.25

Yes	Healthy	No	0.75
Yes	Unhealthy	Yes	0.45
Yes	Unhealthy	No	0.55
No	Healthy	Yes	0.55
No	Healthy	No	0.45
No	Unhealthy	Yes	0.75
No	Unhealthy	No	0.25

Heartburn (Hb):

D	Hb	P(Hb | D)
Healthy	Yes	0.2
Healthy	No	0.8
Unhealthy	Yes	0.85
Unhealthy	No	0.15

Blood Pressure (BP):

HD	BP	P(BP | HD)
Yes	High	0.85
Yes	Low	0.15
No	High	0.2
No	Low	0.8

Chest Pain (CP):

HD	Hb	CP	P(CP | HD, Hb)
Yes	Yes	Yes	0.8
Yes	Yes	No	0.2
Yes	No	Yes	0.6
Yes	No	No	0.4
No	Yes	Yes	0.4
No	Yes	No	0.6

No	No	Yes	0.1
No	No	No	0.9

b). [10 marks] Manually perform inference by variable elimination to calculate the query *P(BP=High)*.

$$P(BP) = \sum_{E,D,HD,Hb,CP} P(E, D, HD, Hb, CP, BP)$$

$$= \sum_{E,D,HD,Hb,CP} P(E)\,P(D)\,P(HD|E,D)\,P(Hb|D)\,P(CP|HD,Hb)\,P(BP|HD)$$

$$= \sum_{E,D,HD,Hb,CP} f_0(E)\,f_1(D)\,f_2(HD,E,D)\,f_3(Hb,D)\,f_4(CP,HD,Hb)\,f_5(BP,HD)$$

f_i, $0 \leq i \leq 5$ are tables in Part a.

Elimination order: E, D, Hb, CP, HD

$$= \sum_{HD} f_5(BP,HD) \sum_{CP} \cancel{f_4(CP,HD,Hb)} \sum_{Hb} f_4(CP,HD,Hb), \sum_{D} f_1(D)\,\cancel{f_3}\, f_3(Hb,D) \cdot \sum_{E} f_0(E)\,f_2(HD,E,D)$$

$$= \sum_{HD} f_5(BP,HD) \sum_{CP} \sum_{Hb} f_4(CP,HD,Hb) \sum_{D} f_1(D)\,f_3(Hb,D)\,f_6(HD,D)$$

f_6:

D	HD	Val
Healthy	Yes	0.34
Healthy	No	0.66
Unhealthy	Yes	0.54
Unhealthy	No	0.46

$$= \sum_{HD} f_5(BP,HD) \sum_{CP} \sum_{Hb} f_4(CP,HD,Hb) \sum_{D} f_7(Hb,D)\,f_6(HD,D)$$

f_7:

D	Hb	Val
Healthy	Yes	0.05
Healthy	No	0.2
Unhealthy	Yes	0.6375
Unhealthy	No	0.1125

$$= \sum_{HD} f_5(BP,HD) \sum_{CP} \sum_{Hb} f_4(CP,HD,Hb) \sum_{D} f_8(Hb,D,HD)$$

f_8:

D	Hb	HD	Val
Healthy	Yes	Yes	0.017
Healthy	Yes	No	0.033
Healthy	No	Yes	0.068
Healthy	No	No	0.132
Unhealthy	Yes	Yes	0.34425
Unhealthy	Yes	No	0.29325
Unhealthy	No	Yes	0.06075
Unhealthy	No	No	0.05175

$$= \sum_{HD} t_5(BP,HD) \sum_{CP} \sum_{Hb} t_4(CP,HD,Hb)\, t_9(HD,Hb)$$

t_9:

Hb	HD	Val
Yes	Yes	0.36125
Yes	No	0.32625
No	Yes	0.12875
No	No	0.18375

$$= \sum_{HD} t_5(BP,HD) \sum_{CP} \sum_{Hb} t_{10}(CP,HD,Hb)$$

t_{10}:

Hb	HD	CP	Val
Yes	Yes	Yes	0.289
Yes	Yes	No	0.07225
Yes	No	Yes	0.1305
Yes	No	No	0.19575
No	Yes	Yes	0.07725
No	Yes	No	0.0515
No	~~Yes~~ No	Yes	0.018375
No	No	No	0.165375

$$= \sum_{HD} t_5(BP,HD)\, t_{11}(HD)$$

t_{11}:

HD	Val
Yes	0.49
No	0.51

$$= t_{12}(BP)$$

t_{12}:

BP	Val
High	0.5185
Low	0.4815

Since the sum of rows of $t_{12}(BP) = 1$, we don't need to normalize it.
Thus, $P(BP=High) = 0.5185$.

c).[30 marks] Now write a program that inputs a query and does inference by variable elimination. The first part of the input consists of the query and evidence variables in the first and second lines respectively. The second part of the input is the value of the variables in the same order. For example, if the desired query is *P(HD=Yes | BP=High, Hb=No),* then the input should be in the following format:

HD
BP, Hb
Yes
High, No

The output is a number between 0 and 1, which represents the probability of the input query. Your program should be able to process any query with arbitrary variables of query and evidence. It is recommended to write the probabilities of the bayesian network in a separate file so that your program can read from it.

Your report should include the screenshots of the inputs and outputs of your program for the following queries:

P(HD=Yes | BP=High, Hb=No)

```
Query Variable: HD
Evidence Variable: BP, Hb
Query Variable Value: Yes
Evidence Variable Value: High, No
Result:
0.412
```

P(Hb=Yes | CP=No)

```
Query Variable: Hb
Evidence Variable: CP
Query Variable Value: Yes
Evidence Variable Value: No
Result:
0.5527197731
```

P(BP=High)

```
Query Variable: BP
Evidence Variable:
Query Variable Value: High
Evidence Variable Value:
Result:
0.5185
```

P(CP=Yes)

```
Query Variable: CP
Evidence Variable:
Query Variable Value: Yes
Evidence Variable Value:
Result:
0.515125
```

P(BP=Low | D=Healthy)

```
Query Variable: BP
Evidence Variable: D
Query Variable Value: Low
Evidence Variable Value: Healthy
Result:
0.579
```