UNIVERSITY OF SASKATCHEWAN
Department of Computer Science
176 Thorvaldson Building
110 Science Place, Saskatoon, SK, S7N 5C9, Canada
Telephine: (306) 966-4886, Facimile: (306) 966-4884

Assignment 3

Bayesian Networks

Date Due: March 16, 2020, 5pm **Total Marks: 57**

Notation

The language of probability is mathematical, and notational conventions vary. The following conventions are useful and practical for some concepts, but they may be unfamiliar.

- Capital letters will represent random variables, e.g., "Let X denote the result of a coin toss, $X{=}0$ for tails, $X{=}1$ for heads." One might call X a variable, but that's imprecise, because it is not varying.

  Specific outcomes will be represented by assertions of the form $X{=}0$. Typically, discrete outcomes are represented by integers, but sometimes short strings are used, e.g., $X{=}'\mathrm{tails}'$.

- We will use the probability notation $P()$ in two ways. The most familiar is $P(X{=}1)$. This is a probability, a non-negative value between 0 and 1.

  We will also write $P(X)$, by which we mean a probability distribution over the outcome space of X. It is short hand notation for $\forall x\, P(X{=}x)$. Similarly, $P(XY)$ is a joint probability distribution over the joint outcome space of X and Y, and $P(X|Y)$ is a conditional probability distribution over the joint outcome space of X and Y.

  We prefer to write $P(X|Y)$ because $\forall x \forall y\, P(X{=}x|Y{=}y)$ is cluttered and contributes nothing to any derivations.

- One of the important operations is summation or integration over a probability distribution. To cut down on clutter, we'll write formulae like this:

$$P(X) = \sum_{Y} P(X|Y)P(Y)$$

  instead of the more precise but cluttered version:

$$\begin{aligned} P(X) &= \sum_{y \in Y} P(X|Y{=}y)P(Y{=}y) \\ &= P(X|Y{=}0)P(Y{=}0) + P(X|Y{=}1)P(Y{=}1) \end{aligned}$$

  where $y \in Y$ means "All the outcomes for Y" (here, just 0 and 1). The same informal notation works for integration of continuous probability mass functions.

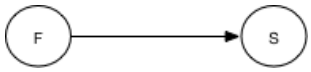
- When lots of outcome labels are involved, we will use adjacency to imply a joint outcome. Sometimes a comma will be used to help clarify. For example, $P(XY) = P(X, Y)$.

Question 1 (6 points):

Purpose: Simple calculations as a warm-up.

Competency Level: Basic

The following diagram represents a simple, 2 node Bayesian Network.



The nodes represent outcomes F and S. The Conditional Probability Distributions (CPDs) for these nodes are as follows:

			$P(S{=}0\|F{=}0)$	=	0.82
$P(F{=}0)$	=	0.13	$P(S{=}1\|F{=}0)$	=	0.18
$P(F{=}1)$	=	0.87	$P(S{=}0\|F{=}1)$	=	0.07
			$P(S{=}1\|F{=}1)$	=	0.93

Using this information, answer the following questions.

(a) Calculate $P(S{=}0|F{=}0)$. Hint: Look up the value in the CPD.

Solution: We must look this up directly from the tables provided. $P(S{=}0|F{=}0) = 0.82$
Grading: has to be the right number to get the mark.

(b) Calculate $\sum_S P(S|F{=}0)$.

Solution:

$$\begin{aligned}\sum_S P(S|F{=}0) &= P(S{=}0|F{=}0) + P(S{=}1|F{=}0)\\ &= 0.82 + 0.18\\ &= 1\end{aligned}$$

This formula works for any single variable in front of the conditioning bar, assuming the evidence behind the bar is constant.

(c) Calculate $P(F{=}0)$. Hint: Look up the value.

Solution: $P(F{=}0) = 0.13$
Grading: has to be the right number to get the mark.

(d) Calculate $\sum_S P(S|F{=}0)P(F{=}0)$. Use your calculations from previous steps.

Note: Take care in the algebra here. There is a correct way and an incorrect way to get the same answer. Remember that multiplication has higher precedence than addition!

Solution: There are a couple of ways to go about it. Because $P(F{=}0)$ does not involve S, we can factor immediately:

$$\sum_S P(S|F{=}0)P(F{=}0) = P(F{=}0)\sum_S P(S|F{=}0)$$

then substitute, or we could expand first and then substitute:

$$\sum_S P(S|F{=}0)P(F{=}0) \quad = \quad P(S{=}0|F{=}0)P(F{=}0) + P(S{=}1|F{=}0)P(F{=}0)$$

Either way we get to the same result:

$$\begin{aligned}\sum_S P(S|F{=}0)P(F{=}0) &= (0.82)(0.13) + (0.18)(0.13) \\ &= (0.13)(0.82 + 0.18) \\ &= 0.13\end{aligned}$$

Grading: has to be the right number to get the mark.

This calculation illustrates the meaning of irrelevance. The rule for irrelevance in a Bayesian network allows us to leave out variables that just will not affect the answer. If you leave irrelevant variables in your calculations, you get the same answer, but you do more work.

(e) Calculate $P(F{=}0|S{=}0)$. Use Bayes Rule. Pay attention specifically to the denominator.

Solution: Applying Bayes rule:

$$P(F{=}0|S{=}0) \quad = \quad \frac{P(S{=}0|F{=}0)P(F{=}0)}{P(S{=}0)}$$

We'll look at the denominator first:

$$\begin{aligned}P(S{=}0) &= \sum_F P(S{=}0|F)P(F) \\ &= P(S{=}0|F{=}0)P(F{=}0) + P(S{=}0|F{=}1)P(F{=}1) \\ &= (0.82)(0.13) + (0.07)(0.87) \\ &= 0.1675\end{aligned}$$

The numerator is one of the terms used in the calculation of the denominator:

$$\begin{aligned}P(F{=}0|S{=}0) &= \frac{P(S{=}0|F{=}0)P(F{=}0)}{P(S{=}0)} \\ &= \frac{(0.82)(0.13)}{(0.82)(0.13) + (0.07)(0.87)} \\ &= \frac{0.1066}{0.1675} \\ &\approx 0.6364\end{aligned}$$

Basically, before any observation of S, the probability of $F = 0$ was pretty low. The observation of $S{=}0$ made $F{=}0$ much more likely.

While no interpretation of the variables were given, $F{=}0$ means "fire", and $S{=}0$ means "smoke". Seeing smoke makes fire much more likely than before.

Grading: has to be the right number to get the mark.

(f) Calculate $P(F{=}1|S{=}0)$. Use Bayes Rule. Pay attention specifically to the denominator.

Solution: The same calculations, with one change in substitutions. Most importantly, the denominator is the same as the previous question.

$$P(S{=}0) \quad = \quad 0.1675$$

This numerator is the other term used in the calculation of the denominator:

$$\begin{aligned} P(F{=}1|S{=}0) &= \frac{P(S{=}0|F{=}1)P(F{=}1)}{P(S{=}0)} \\ &= \frac{(0.07)(0.87)}{(0.82)(0.13)+(0.07)(0.87)} \\ &= \frac{0.0609}{0.1675} \\ &\approx 0.3636 \end{aligned}$$

By no coincidence, $P(F{=}0|S{=}0) + P(F{=}1|S{=}0) = 1$. I tend to round off probabilities at 3 or 4 significant digits, because there's almost no point in using higher precision than that.

Grading: has to be the right number to get the mark.

Notice the similarity in the formulae of this answer and the previous answer. This is why we prefer to derive a symbolic formula, e.g. $P(F|S)$: the same formula can be applied to any outcome for F or S.

What to Hand In

- Six small calculations.

Add these answers to a document for questions 1-5, named `A3.PDF`. You can use Jupyter notebooks, or if you prefer, a LaTeX document.

Evaluation

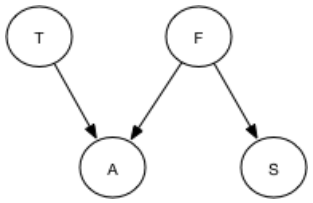
- 1 mark each. The numeric answer has to be correct to get the mark.

Question 2 (6 points):

Purpose: A few more complicated calculations, to make the formulae concrete.

Competency Level: Basic

The following diagram represents a simple, 4 node Bayesian Network.



The nodes represent outcomes F, S, T, A. The Conditional Probability Distributions (CPDs) for these nodes are as follows:

$P(T{=}0)$	=	0.99		$P(F{=}0)$	= 0.13
$P(T{=}1)$	=	0.01		$P(F{=}1)$	= 0.87
$P(A{=}0\|F{=}0,T{=}0)$	=	0.93			
$P(A{=}1\|F{=}0,T{=}0)$	=	0.07			
$P(A{=}0\|F{=}0,T{=}1)$	=	0.02	$P(S{=}0\|F{=}0)$	=	0.82
$P(A{=}1\|F{=}0,T{=}1)$	=	0.98	$P(S{=}1\|F{=}0)$	=	0.18
$P(A{=}0\|F{=}1,T{=}0)$	=	0.006	$P(S{=}0\|F{=}1)$	=	0.07
$P(A{=}1\|F{=}1,T{=}0)$	=	0.994	$P(S{=}1\|F{=}1)$	=	0.93
$P(A{=}0\|F{=}1,T{=}1)$	=	0.25			
$P(A{=}1\|F{=}1,T{=}1)$	=	0.75			

Using this information, answer the following questions.

(a) Calculate $P(A{=}1|F{=}1)$. Hint: Which nodes are not relevant to the query?

Solution: The relevant nodes are the ancestors of A and F, which includes T but not S. If we follow the Variable Elimination procedure, we work as follows:

$$\begin{aligned}
P(A{=}0|F{=}0) &= \frac{P(A{=}0, F{=}0)}{P(F{=}0)} \\
&= \frac{\sum_T P(A{=}0, F{=}0, T)}{P(F{=}0)} \\
&= \frac{\sum_T P(A{=}0|F{=}0, T)P(F{=}0)P(T)}{P(F{=}0)} \\
&= \frac{P(F{=}0)\sum_T P(A{=}0|F{=}0, T)P(T)}{P(F{=}0)} \\
&= \sum_T P(A{=}0|F{=}0, T)P(T) \\
&= P(A{=}0|F{=}0, T{=}0)P(T{=}0) + P(A{=}0|F{=}0, T{=}1)P(T{=}1) \\
&= (0.93)(0.99) + (0.02)(0.01) \\
&= 0.9209
\end{aligned}$$

Notice that the denominator cancelled one of the factors in the numerator.

It turns out that Variable Elimination takes some steps that are not strictly necessary. We could have reasoned as follows, just by using the sum rule for probability:

$$\begin{aligned} P(A{=}0|F{=}0) &= \sum_T P(A{=}0, T|F{=}0) \\ &= \sum_T P(A{=}0|T, F{=}0)P(T|F{=}0) \\ &= \sum_T P(A{=}0|T, F{=}0)P(T) \\ &= P(A{=}0|F{=}0, T{=}0)P(T{=}0) + P(A{=}0|F{=}0, T{=}1)P(T{=}1) \\ &= (0.93)(0.99) + (0.02)(0.01) \\ &= 0.9209 \end{aligned}$$

The key step is noticing that T and F are independent, so $P(T|F{=}0) = P(T)$ for all values of T. We get the same answer. Someone who is reasonably experienced might have done the second derivation; but the first is general enough that we can encode it into a computer program.

Grading: The arithmetic is not important, as long as the values from the CPDs are substituted into a formula derived by using the Variable Elimination procedure. Give two marks if the method is right, even if the calculations or substitutions had errors. No part marks.

(b) Calculate $P(F{=}0|S{=}0)$. Hint: Which nodes are not relevant to the query?

Solution: The relevant nodes are the ancestors of S and F, which leaves out A and T. In other words, the model is more complex, but this particular query is exactly the same as before (Q1.e — even the probability values are the same). You could, if you were bored, include A and T in the variable elimination procedure, and you'd get the exact same result, but you'd have done a lot more work.

$$\begin{aligned} P(F{=}0|S{=}0) &= \frac{P(S{=}0|F{=}0)P(F{=}0)}{P(S{=}0)} \\ &= \frac{(0.82)(0.13)}{(0.82)(0.13) + (0.07)(0.87)} \\ &= \frac{0.1066}{0.1675} \\ &\approx 0.6364 \end{aligned}$$

Grading: The arithmetic is not important, as long as the values from the CPDs are substituted into a formula derived by using the Variable Elimination procedure. Give two marks if the method is right, even if the calculations or substitutions had errors. No part marks.

(c) Calculate $P(A{=}1|S{=}1)$. Hint: Which nodes are not relevant to the query?

Solution: The ancestors of A and F include T and F, which means no variables are irrelevant. Applying the Variable Elimination procedure, we calculate as follows:

$$P(A{=}0|S{=}0) = \frac{P(A{=}0, S{=}0)}{P(S{=}0)}$$

The denominator was calculated earlier (Q1.e). We'll just work with the numerator for a bit:

$$P(A{=}0, S{=}0) = \sum_{F,T} P(A{=}0, S{=}0, F, T)$$

$$
\begin{aligned}
&= \sum_{F,T} P(A{=}0|F,T)P(S{=}0|F)P(T)P(F) \\
&= \sum_{F} P(S{=}0|F)P(F) \sum_{T} P(A{=}0|F,T)P(T)
\end{aligned}
$$

Now we have to expand.

$$
\begin{aligned}
&\sum_{F} P(S{=}0|F)P(F) \sum_{T} P(A{=}0|F,T)P(T) \\
&= P(S{=}0|F{=}0)P(F{=}0)\left(P(A{=}0|F{=}0,T{=}0)P(T{=}0) + P(A{=}0|F{=}0,T{=}1)P(T{=}1)\right) \\
&\quad + P(S{=}0|F{=}1)P(F{=}1)\left(P(A{=}0|F{=}1,T{=}0)P(T{=}0) + P(A{=}0|F{=}1,T{=}1)P(T{=}1)\right) \\
&= (0.82)(0.13)\left((0.93)(0.99) + (0.02)(0.01)\right) \\
&\quad + (0.07)(0.87)\left((0.006)(0.99) + (0.25)(0.01)\right) \\
&\approx 0.09817 + 0.000514 \\
&\approx 0.0987
\end{aligned}
$$

That was the numerator. Bringing in the denominator:

$$
\begin{aligned}
P(A{=}0|S{=}0) &= \frac{P(A{=}0, S{=}0)}{P(S{=}0)} \\
&\approx \frac{0.0987}{0.1675} \\
&\approx 0.5892
\end{aligned}
$$

It's very important to understand why this question was assigned as homework. You need to see how messy and ultimately pointless it is to work with numbers. While it may seem comforting to work with concrete data, it's appallingly tedious. For this kind of work, we want computer programs to do the calculations. But we do need to understand the formulae that we derive!

Grading: The arithmetic is not important, as long as the values from the CPDs are substituted into a formula derived by using the Variable Elimination procedure. Give two marks if the method is right, even if the calculations or substitutions had errors. No part marks.

Hopefully you have the idea that numbers can make calculations seem less abstract, but they don't actually buy us anything, and they're a pain in the neck. From now on, we'll just do the algebra.

What to Hand In

- Three calculations, made odious by keeping track of arithmetic.

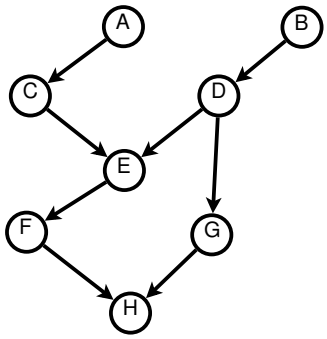
Add these answers to a document for questions 1-5, named `A3.PDF`. You can use Jupyter notebooks, or if you prefer, a LaTeX document.

Evaluation

- 2 marks each. The arithmetic is not important, as long as the values from the CPDs are substituted into a formula derived by using the Variable Elimination procedure. Give two marks if the method is right, even if the calculations or substitutions had errors. No part marks.

Bayesian network structure for Q3-6

Consider the Bayesian network given in the following diagram. Notice that there are families defined by the graph, but I have not given you the numbers for the tables. That's because doing mindless arithmetic is for computers, not people.



This network will form the basis for questions 3-6 in this assignment.

Question 3 (10 points):

Purpose: To practice reading conditional independence from a Bayesian Network.

Competency Level: Basic

Consider the Bayesian network given above. Answer the following statements about conditional independence. For each one, show the paths, and say whether the path is active or blocked.

1. (2 marks) Given $\{\}$ (i.e., the empty set), is A conditionally independent of D? (Because the set is empty, we say "unconditionally independent," or just "independent.")

   Solution: There are two paths between A and D.

   (a) $A \rightarrow C \rightarrow E \leftarrow D$. Since there is no evidence, the path is blocked at E, a head-to-head node on this path.

   (b) $A \rightarrow C \rightarrow E \rightarrow F \rightarrow H \leftarrow G \leftarrow D$. Sine there is no evidence at H, a head-to-head node on this path, the path is blocked.

   Since both paths have at one node blocking them, all paths are blocked, and we can say A is independent of D.
   Notice that E is head-to-head on one path, but not on the other path!

2. (2 marks) Given H, is C conditionally independent of B?

   Solution: There are two paths between C and B.

   (a) $C \rightarrow E \rightarrow F \rightarrow H \leftarrow G \leftarrow D \leftarrow B$. Normally, head-to-head nodes block paths, unless they are part of the evidence. The evidence at H makes the path through H unblocked (or active). Since we found one path is not blocked, we can answer saying "No."

   (b) $C \rightarrow E \leftarrow D \leftarrow B$. Normally, head-to-head nodes block paths, unless they are part of the evidence, or unless they have a descendant in the evidence. Since H is a descendant of E, the path through E is not blocked, and since there is no other evidence, the path is not blocked.

Since both paths are not blocked, we can say that A is **not** conditionally independent of D given E. In general, we don't need to find more than one unblocked (active) path, but here we're being complete, looking at both.

3. (2 marks) Given F, is H conditionally independent of A?

Solution: There are two paths between A and H.

(a) $A \rightarrow C \rightarrow E \rightarrow F \rightarrow H$. The evidence at F blocks this path.

(b) $A \rightarrow C \rightarrow E \leftarrow D \rightarrow G \rightarrow H$. The node E is head-to-head on this path. The evidence F is a descendant of E, which makes the path through E active (unblocked). Therefore this path is not blockerd.

Since one path is not blocked, we can say that A is **not** conditionally independent of H given F.

4. (2 marks) Given C, is F conditionally independent of B?

Solution: There are two paths between B and F.

- $F \leftarrow E \leftarrow D \leftarrow B$. This path is sequential, and the evidence is not on this path, so this path is not blocked.
- $F \rightarrow H \leftarrow G \leftarrow D \leftarrow B$. The node H is head-to-head on this path, and H is not evidence, so the path is blocked.

Since one path is not blocked, F is **not** conditionally independent of B, given C. We only need to remark on the unblocked path; for completeness I looked at both.

5. (2 marks) Given F and D, is C conditionally independent of G?

Solution: There are two paths between C and G.

(a) $C \rightarrow E \rightarrow F \rightarrow H \leftarrow G$. The node F blocks this path. The node H also blocks the path. All we need is one node to block a path, but either of these can be used as the reason.

(b) $C \rightarrow E \leftarrow D \rightarrow G$. The evidence node F is a descendant of E, which is a head-to-head node. Therefore the evidence F causes the path through E to be unblocked. But the evidence at D blocks the path.

Since both paths are blocked, we can say that C is conditionally independent of G given F and D.

What to Hand In

- Five statements about conditional independence, accompanied by path analysis for justification.

Add these answers to a document for questions 1-5, named `A3.PDF`. You can use Jupyter notebooks, or if you prefer, a LaTeX document.

Evaluation

- 2 marks each: one for the right answer, and one for a correct justification.

Question 4 (5 points):

Purpose: To practice reading ancestors in a Bayesian network.

Competency Level: Basic

Consider the Bayesian network given above. For each of the following queries, which nodes are relevant, according to the rule for relevance given in class?

1. $P(C|G)$

   Solution: The query and evidence nodes are C, G. The relevant nodes are C, A, G, D, B. It's okay if the query and evidence nodes are omitted. The question asked for them, but it's not worth deducting marks if they are the only ones missing.
   However, it is correct to observe (from the previous question), that C and G are independent, so $P(C|G) = P(C)$, so only C, A are relevant. Both answers are acceptable.

2. $P(D|E, A)$

   Solution: The query and evidence nodes are D, E, A. The relevant nodes are D, B, E, C, A. It's okay if the query and evidence nodes are omitted.

3. $P(F)$ (i.e., no evidence)

   Solution: The query and evidence nodes are just F. The relevant nodes are F, E, D, C, A, B. It's okay if the query and evidence nodes are omitted.

4. $P(B|D, F)$

   Solution: The query and evidence nodes are B, D, F. The relevant nodes are F, E, C, A, D, B. It's okay if the query and evidence nodes are omitted.

5. $P(C|F, G)$

   Solution: The query and evidence nodes are C, F, G. The relevant nodes are F, E, C, A, D, B, G. It's okay if the query and evidence nodes are omitted.

Include in your answer the query and evidence variables (just to avoid ambiguity).

What to Hand In

- Five lists of variables (or nodes).

Add these answers to a document for questions 1-5, named `A3.PDF`. You can use Jupyter notebooks, or if you prefer, a LaTeX document.

Evaluation

- 1 mark each.

Question 5 (16 points):

Purpose: To practice deriving formulae for any query on a Bayesian Network.

Consider the Bayesian network given above. Write out a formula for each of the following queries, using technique called variable elimination shown in class. You cannot do any arithmetic, so leave your solution in terms of the CPDs implied in the Bayesian network. Choose any order for the marginalization; there is no need to determine the "optimal" order. Be sure to consider your answers from the previous questions, as some of them are either relevant directly, or indirectly.

1. $P(C)$ (i.e., no evidence)

Solution: When there's no evidence, we just use the sum rule for probabilities, considering only the relevant variables. One solution to the problem uses the relevant variables A, G, D, B.

$$\begin{aligned} P(C) &= \sum_{ABDG} P(ABCDG) \\ &= \sum_{ABDG} P(A)P(B)P(C|A)P(D|B)P(G|D) \\ &= \sum_{A}\sum_{B}\sum_{D}\sum_{G} P(A)P(B)P(C|A)P(D|B)P(G|D) \\ &= \sum_{A} P(A)P(C|A) \sum_{B} P(B) \sum_{D} P(D|B) \sum_{G} P(G|D) \end{aligned}$$

Any order of the relevant variables gives a correct formula, provided that the factoring in the last step is correct.
That's as far as I expect students to go with this. But there is a second solution, which uses the insight that C and G are independent. Because of that, G, D, B are irrelevant, and we onlly need to consider A as the relevant variable. WE proceed as follows:

$$\begin{aligned} P(C) &= \sum_{A} P(AC) \\ &= \sum_{A} P(A)P(C|A) \end{aligned}$$

Much shorter.
It must be emphasized that these two formulae would result in the same answer, if worked through with numbers as in Q1 and Q2. One is just more work. But how can we see the equality?
Let's start with the last line of the longer solution:

$$P(C) = \sum_{A} P(A)P(C|A) \sum_{B} P(B) \sum_{D} P(D|B) \sum_{G} P(G|D)$$

We will note that $\sum_G P(G|D)$ is summing a single CPD over the child variable. The result of the summation is a vector (or 1D table) consisting of 1s only. This table cannot change future result, so we can continue:

$$P(C) = \sum_{A} P(A)P(C|A) \sum_{B} P(B) \sum_{D} P(D|B)$$

Again, the last sum $\sum_D P(D|B)$ is a sum over the CPD for the child variable D, and again the resulting sum is a 1D table containing 1s only. So:

$$P(C) = \sum_{A} P(A)P(C|A) \sum_{B} P(B)$$

Again the last sum $\sum_B P(B)$ is a trivial factor, but it's just the scalar value 1.
Finally, we are left with:

$$P(C) = \sum_A P(A)P(C|A)$$

which is the result from the solution that considered independence. One might wonder why we can't just continue to cancel every factor out like this. The answer is that the the last sum is the sum of a product, not the sum of a single CPD. We can only cancel a sum if there is a single CPD being summed over its child variable.
This sequence of cancellations depends on the order of the variables. A different order for the nuisance variables might result in similar cancellations, but not all orderings will. Finally, the fact that C and G are independent implies (mathematically) that there is an ordering that will result in such cancellations, but it does not tell us which ordering that might be. On the other hand, if there is an ordering that will cancel the sums, then we needn't look for the ordering; we just omit the variable G from the query, and calculate relevant nodes without it.
Grading: Give 4 marks if all of the following were done correctly:

- Summing the relevant variables. Either set of relevant variables is acceptable. Deduct one mark if for some reason a conditional probability (involving a fraction) is being calculated. There is no conditional here, so no fraction.
- Factorizing the JPD using the CPDs.
- Distributing the summation without error.

Give 3 marks if any one of the steps above is incorrect. Give only 1 mark if more than one of the steps above is incorrect.

2. $P(B|C)$

Solution: There are two ways to approach this question. First, the short way. We can notice that the path between C and G is blocked. That means that C is independent of G, in which case:

$$P(B|C) = P(B)$$

This is a much simplified query, with no additional relevant variables. In other words, we're done.
The longer answer does not make use of the independence, but simply works with the variables relevant to C and C both, namely A, B, C.
Using the Variable Elimination technique, we work as follows, starting with the definition of conditional probability:

$$P(B|C) = \frac{P(BC)}{P(C)}$$

Working with the numerator:

$$\begin{aligned} P(BC) &= \sum_A P(ABC) \\ &= \sum_A P(C|A)P(A)P(B) \\ &= P(B)\sum_A P(C|A)P(A) \end{aligned}$$

Now, let's turn to the denominator:

$$P(C) = \sum_B P(BC)$$

To wrap the solution up we could write:

$$\begin{aligned} P(B|C) &= \frac{P(BC)}{P(C)} \\ &= \frac{P(B)\sum_A P(C|A)P(A)}{\sum_B P(BC)} \end{aligned}$$

We could leave it there, but again, we have two different formulae that have to be equal. It's not magic. We could reason as follows. Let's look at $P(C)$ just in the network, and not as a part of the query $P(B|C)$. We would notice that A is the relevant nuisance variable, and write:

$$\begin{aligned} P(C) &= \sum_A P(AC) \\ &= \sum_A P(A)P(C|A) \end{aligned}$$

And this sum appears in the numerator, and so we can cancel, leaving the simpler formula above. Again, noticing conditional independence early can simplify calculations a great deal.

Grading: Give 4 marks for:

- Noticing the independence and jumping right to the answer as in the short solution above
- Or all of the following:
  - Expressing the query in terms of a fraction (definition of conditional probability).
  - Summing the relevant variables.
  - Factorizing the JPD using the CPDs.
  - Distributing the summation without error.

  Give 3 marks if any one step is done incorrectly, and 1 mark only if more than one step done incorrectly.

3. $P(A|H,C)$

Solution: Again, there is a short solution, and a long solution. The short solution notices that A is independent of H given C, because C blocks both paths between A and H. So:

$$\begin{aligned} P(A|H,C) &= P(A|C) \\ &= \frac{P(AC)}{P(C)} \end{aligned}$$

Working with the numerator:

$$P(AC) = P(C|A)P(A)$$

Working with the denominator, in the standard way, by leaning on the numerator:

$$P(C) = \sum_A P(AC)$$

This is the way I typically leave things. But to summarize:

$$\begin{aligned} P(A|H,C) &= P(A|C) \\ &= \frac{P(AC)}{P(C)} \\ P(AC) &= P(C|A)P(A) \\ P(C) &= \sum_A P(AC) \end{aligned}$$

I don't prefer to do the substitution back into the fraction, because it doesn't reflect the computational aspect. The denominator is (almost) always calculated from the numerator.
The longer derivation is quite long, because of the number of nuisance variables.

$$P(A|H,C) = \frac{P(ACH)}{P(CH)}$$

Working with the numerator, and realizing that every variable is a relevant variable A, B, C, D, E, F, G, H:

$$\begin{aligned} P(ACH) &= \sum_{BDEFG} P(ABCDEFGH) \\ &= \sum_{BDEFG} P(A)P(C|A)P(B)P(D|B)P(G|D)P(E|DC)P(F|E)P(H|F,G) \\ &= \sum_{BDEFG} P(A)P(C|A)\sum_B P(B)\sum_D P(D|B)\sum_E P(E|DC)\sum_F P(F|E)\sum_G P(G|D)P(H|F,G) \end{aligned}$$

I just used the alphabetical order, which is not particularly good. But this demonstretes the method, and choosing a good order for the variables is not part of the exercise.
Now the denominator (which is easy to forget, because it is so forgettable):

$$P(CH) = \sum_A P(ACH)$$

and that's where to leave that. The denominator could be expanded mathematically, similar to the numerator, but if we treat the formulae not as mathematics, but as a script in some programming language, then we can consider the value for $P(CFG)$ as a table (3 dimensions), that we've gone through the trouble to work out. Let's store is somewhere for future use, and then compute the denominator from this table, because we've stored it. We can calculate the denominator from the numerate in every case; so storing the numerator as an intermediate result is simply the way to go by default. It seems unusual only because we're used to solving everything and writing expressions as far as they can go in math. We're taught to avoid duplication of work in a program though!
Again, we have two answers, and they have to be equivalent. The way to show equivalence is to use a different order for the nuisance variables in the numerator. We can try this:

$$P(ACH) = \sum_{EFGDB} P(ABCDEFGH)$$

And from there, there are cancellations that get rid of almost all of the CPDs in the numerator. Almost. To get the last step, we have to work through the denominator using the same order, and then do some cancellation within the fraction. Again, we could try to give the computer smarts enough to solve algebra, but Bayes Ball is simpler, and working only with relevant variables is the way to go here.
Grading: Give 4 marks if all of the following are done correctly.

- Expressing the query in terms of a fraction (definition of conditional probability).
- Summing the relevant variables (using the short or long form).
- Factorizing the JPD using the CPDs.
- Distributing the summation without error.

Give 3 marks if any one step is done incorrectly, and 1 mark only if more than one step was done incorrectly.

4. $P(C|D,F)$

Solution: This one is pretty standard. There is no conditional independence to exploit in the initial query. We start with the definition of conditional probability:

$$P(C|D,F) = \frac{P(CDF)}{P(DF)}$$

Working with the numerator, and the relevant variables $A, B, (C), (D), E, (F)$:

$$\begin{aligned} P(CDF) &= \sum_{ABE} P(ABCDEF) \\ &= \sum_{ABE} P(A)P(B)P(C|A)P(D|B)P(E|CD)P(F|E) \\ &= \sum_{A} P(A)P(C|A) \sum_{B} P(B)P(D|B) \sum_{E} P(E|CD)P(F|E) \end{aligned}$$

Now for the denominator:

$$P(DF) = \sum_{C} P(CDF)$$

And that's calculating the denominator from the pre-computed numerator.
Grading: Give 4 marks if all of the following are dine correctly:

- Expressing the query in terms of a fraction (definition of conditional probability).
- Summing the relevant variables.
- Factorizing the JPD using the CPDs.
- Distributing the summation without error.

Give 3 marks if any one step is done incorrectly, and 1 mark only if more than one step was done incorrectly.

What to Hand In

- Four derivations, showing the application of Variable Elimination, making use of JPD factorization, conditional independence and marginalization of the relevant variables.

Add these answers to a document for questions 1-5, named `A3.PDF`. You can use Jupyter notebooks, or if you prefer, a LaTeX document.

Evaluation

- 4 marks each. Full marks if the answers are correct and use the graph structure to inform the algebra as in questions 3 and 4.

Question 6 (14 points):

Purpose: To work by hand with a simple example of inferring Bayesian Network structure, and model selection.

Competency Level: Intermediate.

On the Assignment 2 page, you'll find a Jupyter Notebook named `A2Q6.ipynb`, which walks through the process of inferring Bayesian Network structure from data. The document has the following steps:

1. Load a categorical dataset (`diamond.csv`)
2. Create 2 (related) Bayesian network models, fitting their parameters to the data.
3. Compare the 2 structures, using log-likelihood.

What to Hand In

- Your Jupyter notebook, exported and named `A3Q6.pdf`

Evaluation

- 5 marks: Step 1: Your notebook calculates the 5 tables correctly, and presents them neatly.
- 4 marks: Step 2: Your calculations are correct.
- 1 mark: Conclusions: Your conclusions are correct.

Solution: A solution for this problem can be found in the file `A3Q6_Solution.ipynb` (exported as `A3Q6_Solution.pdf`).

The main issue here is convincing Python and Pandas to do the needed work. My solution is very short in terms of code, but required some time (browsing through the API) to devise.

Having exactly the numbers in the solution is not really the point of the exercise. I used "add-one-smoothing" (Also known as "Beta(1,1) priors"). This is ideal, but not necessary. The given data set had 1000 samples of a total of 16 possible combinations of values for the features. Every combination was observed at least a few times. If the data set had more features per sample (say 10 variables instead of 4), then there might easily have been combinations that are not present in a data set with only 1000 samples.

For step 2, I used a Python loop to work through the dataset. There might be a way to use Pandas so that the loop does not have to be in Python. Maybe a function to calculate the probability of each row, passed to a Pandas method. The main issue here is using the tables from Step 1 to calculate the probability that the data comes from each of the models M_1 and M_2, i.e., the likelihood of the data. The model solution shows 2 ways to do it. The first uses the entire model, and all of the data. The second way is to use only the parts of the model that differ. In this case, only the CPT for variable D is different, so that's the only part that can make an effect on the required likelihood. Either method is fine.

For step three, using comparison of log-likelihood, we can conclude that Model M_1 has the greater value, and is therefore the better model.

The main problem with this whole exercise is checking correctness. If the tools are complicated, it might be hard to understand exactly what they are doing.

Grading:

Step 1. 5 tables.

- Any technique to do the counting is acceptable.
- 3 marks: Correctness. Give full marks for correctness if all of the following are true:
  - They have the right number of entries.
  - They are normalized in the right way.
  - There are indeed 5 tables.

  If any of the above are not true, make a deduction.
- 2 marks. Neatness. The tables have to have minimal formatting at least.

Step 2. Calculating Likelihood of the data using the two models.

- 4 marks. Correctness. To be correct the script has to
  - Loop through each row in the data set.
  - Calculate the probability (likelihood) of the combination of values on the row, using both models (a partial calculation here is fine)
  - Calculate the likelihood of the entire dataset by accumulating the likelihood from each row. This has to be done using log likelihoods, because otherwise the calculation will result in underflow.

  Deduct a mark for every step not done. The numeric results don't have to be exactly the same as mine, but they have to be plausible given Step 1.

Step 3. Conclusions.

- By comparison of log-likelihood, Model M_1 has the greater value, and is therefore the better model.
- 1 mark. Give the mark if the correct model was named as the better model.