CMSI 485 – Classwork 2

SOLUTION

Instructions:
This worksheet will not only provide you with practice problems for your upcoming exam, but will add to your toolset as initiate data scientists taking various probabilistic inference problems from start to finish. Specific notes:

- Provide answers to each of the following questions and write your responses in the blanks. If you are expected to show your work in arriving at a particular solution, space will be provided for you.
- Place the names of your group members below:

Group Members:

1. _____
2. _____
3. _____
4. _____

Problem 1 – Probabilistic Reasoning

Forney Industries is branching further into the medical domain and has made the discovery of a new, rare disease, Schistoforneymiosis. It is believed that roughly 1/1000 of the population suffers from the disease [D], experiencing a consistently wet left foot and sentient freckles.

To test for the condition, the test [T] is an elaborate procedure involving multiple probes, and returns an end result that is either positive (T = True) or negative (T = False). The problem is that the tests are not perfect, but 95% of people who have the disease will test positive, and 2% of people who do *not* have the disease will test positive.

Answer the following questions to answer an important query: the probability that someone has the disease If they test positive. (NB: 80% of MDs fail to answer this question correctly)

Part 1.1*:* Sometimes, the query that we are interested in answering is of the format P(A|B), but the information we have is in the format P(B|A). Towards this end, we can use *Bayes' Theorem,* (the namesake of Bayesian Networks) which states:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Below, prove Bayes' Theorem using the conditioning rule from probabilistic logic and the fact that $P(A,B) = P(B,A)$.

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

Part 1.2: Now that we have Bayes' Theorem under our belt, we'll encode the claims of probabilities from the problem statement into the syntax of probabilistic reasoning. Use binary variables $D = \{0, 1\}$ for whether or not someone has the disease, and $T = \{0,1\}$ for whether or not someone tests positive.

1. "It is believed that roughly 1/1000 of the population suffers from the disease"

   P(D = 1) = 0.001

2. “95% of people who have the disease will test positive”

   $P(T = 1 \mid D = 1) = 0.95$

3. “2% of people who do *not* have the disease will test positive”

   $P(T = 1 \mid D = 0) = 0.02$

4. Compute, using the law of total probability, the chance that an arbitrary individual will have tested positive for Schistoforneymiosis.

$$\begin{aligned} P(T=1) &= \sum_d P(T=1, D=d) \\ &= \sum_d P(T=1|D=d)P(D=d) \\ &= P(T=1|D=0)P(D=0) + P(T=1|D=1)P(D=1) \\ &= 0.02 * 0.999 + 0.95 * 0.001 \end{aligned}$$

   $P(T = 1) \approx 0.021$

We now have all of the parts needed to solve our problem.

Part 1.3: Compute, using your work from parts 1.1 and 1.2, the probability that a patient has the disease given that they tested positive for it.

$$\begin{aligned} P(D=1|T=1) &= \frac{P(T=1|D=1)P(D=1)}{P(T=1)} \\ &= \frac{0.95 * 0.001}{0.021} \end{aligned}$$

$P(D = 1 \mid T = 1) \approx 0.045$

Note: that’s pretty rare given what is quite the accurate test! The importance of knowing probability calculus!

Problem 2 – Simple Bayesian Network Inference

Forney Industries are committed to the treatment of Schistoforneymiosis. Towards this end, their other genetic mapping division, FornGene, has attempted to isolate certain genetic expressions that contribute to whether or not an individual has a hereditary disposition to the disease. They conduct a study measuring whether or not an individual's female parent [F] and male parent [M] have the genetic marker, and subsequently, if their child [C] has the disease.

In this section, we will attempt to craft a Bayesian Network that explains the data collected by the FornGene group, and use that to answer inference queries related to the disease.

Part 2.1: In the following table, we track each combination of M, F, and C witnessed from the group's study. From this data, construct the joint distribution over $P(M, F, C)$ in the table below. [This gives you some fundamental experience with data science!]

M	F	C	# of data points	$\boldsymbol{P(M, F, C)}$
0	0	0	864	0.432
0	0	1	96	0.048
0	1	0	72	0.036
0	1	1	168	0.084
1	0	0	448	0.224
1	0	1	192	0.096
1	1	0	32	0.016
1	1	1	128	0.064

Part 2.2: We now want a model to interpret this data. In particular, we have a pretty good idea of the nature of cause and effect implicit in this environment (from what we know of the science behind hereditary genetics). Encode these assumptions as independence statements:

1. "The genetic expression of one parent is unaffected by the other"

   $M \perp\!\!\!\perp F$

2. "Whether or not a child has the disease provides information between the parents"

   $M \not\perp\!\!\!\perp F \mid C$

Part 2.3: With the independence claims you're making above, draw the *structure* of a Bayesian Network that encodes these claims.

$$M \rightarrow C \leftarrow F$$

Part 2.4: With the independence claims and structure from Parts 2.2 & 2.3, encode the Bayesian Network's *semantics,* i.e., the Conditional Probability Tables (CPTs), for each variable in the network. [Hint: you will only need the joint distribution, conditioning, and law of total probability to complete each of the following]

1. CPT for *M*

Calculations

$$P(M) = \sum_{f} \sum_{c} P(M, F, C)$$

CPT

M	P(M)
0	0.6
1	0.4

2. CPT for *F*

Calculations

$$P(F) = \sum_{m} \sum_{c} P(M, F, C)$$

CPT

F	P(F)
0	0.8
1	0.2

3. CPT for *C*

Calculations

$$P(F,M) = \sum_c P(M,F,C)$$

$$P(C|F,M) = \frac{P(M,F,C)}{P(M,F)} = \frac{P(M,F,C)}{\sum_c P(M,F,C=c)} \quad \text{OR} \quad = \frac{P(M,F,C)}{P(M)P(F)}$$

CPT

M	F	C	P(C|M, F)
0	0	0	0.9
0	0	1	0.1
0	1	0	0.3
0	1	1	0.7
1	0	0	0.7
1	0	1	0.3
1	1	0	0.2
1	1	1	0.8

Part 2.5: For each of the following inference queries, determine which are *immediately answerable* from the previous two sections, and for those that are, provide their value.

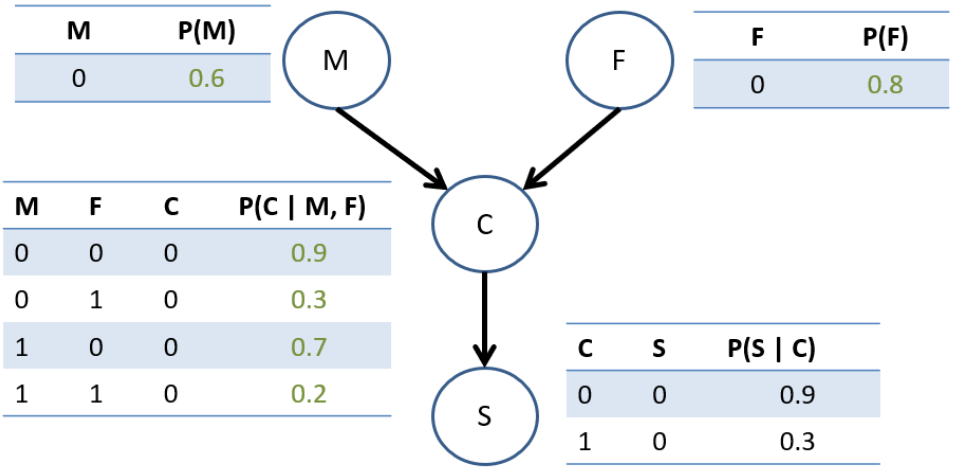
Query	Immediately Answerable?	If so, value:
$P(M = 0)$	Yes	0.6
$P(F = 1 \mid M = 1)$	Yes	0.2
$P(C = 1)$	No	-
$P(C = 0 \mid M = 1)$	No	-
$P(C = 0 \mid M = 1, F = 1)$	Yes	0.2
$P(C = 1 \mid M = 1, F = 1)$	Yes	0.8

Part 2.6: Given the above, suppose a patient comes in who we do *not* know whether or not they have the disease, but we know that both parents possess the predictive genetic trait (i.e., $M = 1, F = 1$). Should we give this patient the (expensive) vaccine for Schistoforneymiosis?

Yes, since $P(C = 0 \mid M = 1, F = 1) < P(C = 1 \mid M = 1, F = 1)$ by quite a bit.

Problem 3 – Exact Bayesian Network Inference

Suppose we now add the presence of a child's symptomology to the network established in the previous section (you didn't read ahead did you?). Note we can model this as an effect of the disease's presence in the child, as follows.



M	P(M)
0	0.6

F	P(F)
0	0.8

M	F	C	P(C | M, F)
0	0	0	0.9
0	1	0	0.3
1	0	0	0.7
1	1	0	0.2

C	S	P(S | C)
0	0	0.9
1	0	0.3

1. Fill in the missing CPT likelihoods from your work in Problem 2.
2. Using these, compute the likelihood that a patient's father had the disease, given that the mother did and the child exhibits symptoms, i.e., find $P(F = 1|M = 1, S = 1)$

Step 1: Label Vars:

$$Q = \{F\}, e = \{M = 1, S = 1\}, Y = \{C\}$$

Step 2: Compute $\boldsymbol{P(Q, e) = P(F, M = 1, S = 1)}$

$$P(F, M = 1, S = 1) = \sum_c P(F, M = 1, S = 1, C = c)$$

$$= P(M = 1)P(F) \sum_c P(C = c|M = 1, F)\, P(S = 1|C = c)$$

- **For F = 0:** $P(F = 0, M = 1, S = 1) = 0.4 * 0.8 * [0.7 * 0.1 + 0.3 * 0.7] = 0.0896$
- **For F = 1:** $P(F = 1, M = 1, S = 1) = 0.4 * 0.2 * [0.2 * 0.1 + 0.8 * 0.7] = 0.0464$

Step 3: Compute $\boldsymbol{P(e) = P(S = 1, M = 1)}$

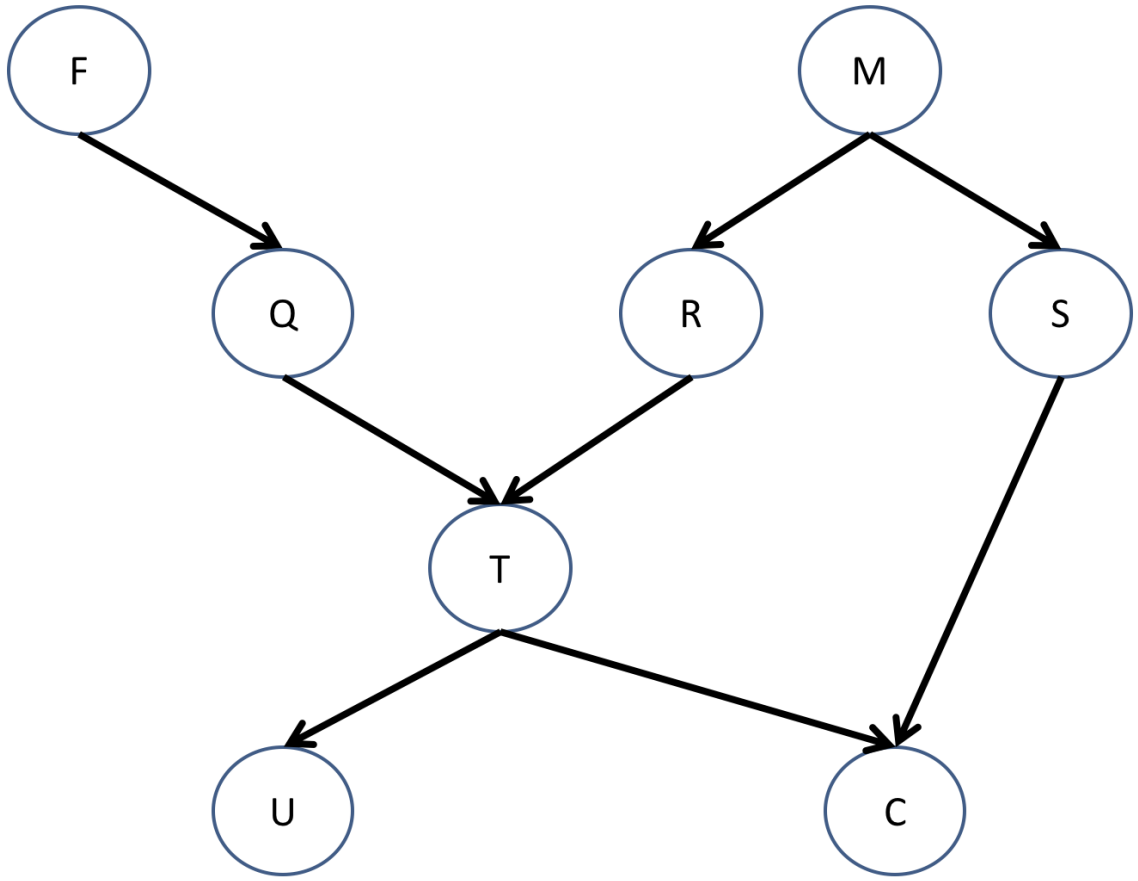
$$P(S = 1, M = 1) = \sum_f P(F = f, S = 1, M = 1) \; = 0.0896 + 0.0464 = 0.136$$

Step 4: Normalize to solve: $\boldsymbol{P(Q|e) = P(Q, e)/P(e)}$

$$P(F = 1|M = 1, S = 1) = \frac{P(F = 1, M = 1, S = 1)}{P(M = 1, S = 1)} = \frac{0.0464}{0.136} \approx 0.341$$

Problem 4 – d-separation

The FornGene group further investigated the genetic relationship between parent and child with regards to the disease's expression and revealed a more complex network than originally thought. In particular, they discovered that a variety of other environmental factors would contribute to an arbitrary patient's risk of attaining the disease, as given by the structure below:



Determine if the following independence relationships hold in the network above:

X	**Y**	**{Z}**	**X ⫫ Y | {Z}?**
F	T	{Q}	True / Yes
Q	R	{T}	False / No
F	M	{C}	False / No
U	S	{C, M}	False / No
Q	C	{T}	False / No
Q	C	{T, S}	True / Yes