CS 6190: Probabilistic Modelling Spring 2019

Homework 3

Abhinav Kumar (u1209853)

Analytical problems [100 points + 40 bonus]

1. [13 points] The joint distribution over three binary variables are given in Table 1. Show by direct evaluation that this distribution has the property that a and b are marginally dependent, so that $p(a,b) \neq p(a)p(b)$, but that they become independent when c, so that $p(a,b|c) = p(a|c)p(b|c)$.

a	b	c	p(a,b,c)
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

Table 1: Joint distribution of a, b, c.

We make table 2 for calculation of joint distribution.

a	p(a)
0	0.6
1	0.4

b	p(b)
0	0.592
1	0.408

c	p(c)
0	0.48
1	0.52

a	b	p(a,b)
0	0	0.336
0	1	0.264
1	0	0.256
1	1	0.144

Table 2: Marginal and joint distribution of a, b.

Clearly, $p(a=1, b=1) = 0.144 \neq p(a=1)p(b=1)$ and hence they are not independent. Next we list out the marginals and joints when c is given.

a	p(a|c=0)	p(a|c=1)
0	0.5	0.69
1	0.5	0.31

b	p(b|c=0)	p(b|c=1)
0	0.8	0.4
1	0.2	0.6

a	b	p(a,b|c=0)	p(a,b|c=1)
0	0	0.4	0.277
0	1	0.1	0.415
1	0	0.4	0.123
1	1	0.1	0.185

Table 3: Marginal and joint distribution of a, b.

We can check from table 3 that $p(a,b|c) = p(a|c)p(b|c) \quad \forall a, b, c$

2. [12 points] Using the d-separation algorithm/criterion, show that the conditional distribution for a node x in a directed graph, conditioned on all of the nodes in the Markov blanket, is independent of the remaining variables in the graph.

Markov blanket for a node in a graphical model contains all the variables that shield the node from the rest of the network. This means that the Markov blanket of a node is the only knowledge needed to predict the behavior of that node and its children. It is the set of nodes composed of x's parents, x's children, and x's children's other parents.

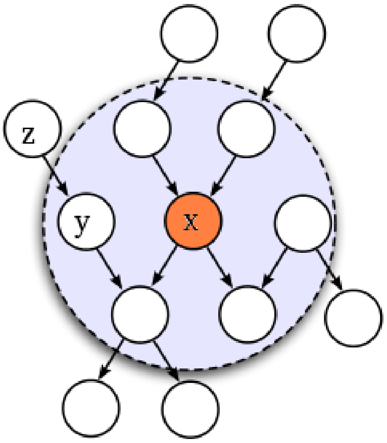


Figure 1: Markov Blanket. Courtesy- Wikipedia

We use the classic cases of d-separation. For grandparents of x, since the parents of x are in the Markov blanket, they will be independent. The grandchildren of x are independent with x since the children of x are in Markov Blanket. Now, the other grandparents of children of x such as z could make them dependent with x because of the case of the collider or head to head. But other parents of children of x such as y are also in Markov Blanket. So, the other grandparents of children of x such as z is independent with x. Hence, we have x is independent of all other nodes conditioned on nodes in the Markov Blanket.

3. [15 points] See the graphical model in Figure 2. Recall what we have discussed in the class. Show that $a \perp\!\!\!\perp b|\emptyset$. Suppose we have observed the variable d. Show that in general $a \not\perp\!\!\!\perp b|d$.

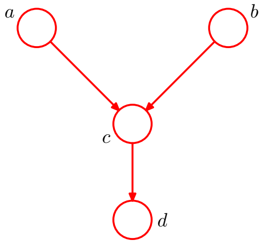


Figure 2: Graphical model.

We begin with the joint probability expression of the graphical model. The joint is given by

$$p(a,b,c,d) = p(d|c)p(c|a,b)p(a)p(b) \tag{1}$$

Now,

$$p(a,b) = \sum_c \sum_d p(a,b,c,d) \tag{2}$$

$$= \sum_c p(c|a,b)p(a)p(b) \sum_d p(d|c) \tag{3}$$

$$= \sum_c p(c|a,b)p(a)p(b) \tag{4}$$
$$= p(a)p(b) \tag{5}$$
$$\tag{6}$$

Hence, a and b are independent.

For the conditional independence, we need to show that

$$p(a,b|d) = p(a|d)p(b|d)$$

Now, we have

$$\begin{aligned} LHS = p(a,b|d) &= \frac{p(a,b,d)}{p(d)} \\ &= \frac{\sum_c p(a,b,c,d)}{p(d)} \\ &= \frac{\sum_c p(d|c)p(c|a,b)p(a)p(b)}{p(d)} \\ &= \frac{p(d|a,b)p(a)p(b)}{p(d)} \\ &\neq p(a|d)p(b|d) = RHS \end{aligned} \tag{7}$$

Hence, they are not conditionally independent.

4. [10 points] Convert the directed graphical model in Figure 2 into an undirected graphical model. Draw the structure and write down the definition of the potential functions.

As told in the class, we use moralization to convert to undirected graphical model Figure 3.

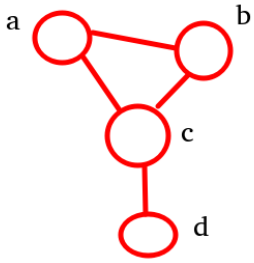


Figure 3: Undirected Graphical model

The potential function is then given by

$$p(a,b,c,d) = \frac{\psi(a,b,c)\psi(c,d)}{Z} \tag{8}$$

where $\psi(a,b,c) = p(c|a,b)p(a)p(b)$, $\psi(c,d) = p(d|c)$ and Z is the normalization constant.

5. [15 points] Write down every step of the sum-product algorithm for the graphical model shown in Figure 4. Note that you need to first choose a root node, and write down how to compute each message. Once all your messages are ready, please explain how to compute the marginal distribution $p(x_4, x_5)$.

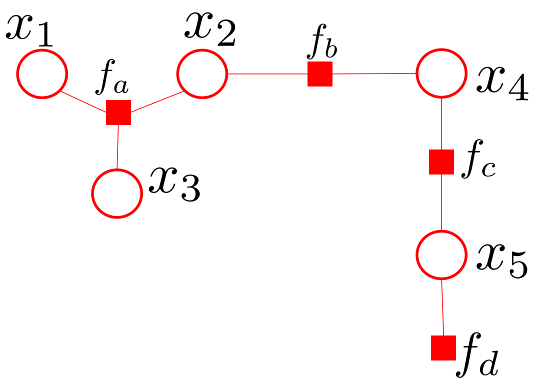


Figure 4: Factor graph.

We follow the lecture 10 for the notations. Let us assume that x_1 is the root node and f_d is the leaf node in Figure 4. The factor nodes are used to carry the optimization of the messages while the variable nodes simply pass the message obtained from other nodes.

We first compute the message from leaf to root and then we compute from the root to leaf. We initialize the leaf as 1. The former messages are

$$\begin{aligned}
\mu_{f_d \to x_5} &= f_d(x_5) \\
\mu_{x_5 \to f_c} &= f_d(x_5) \\
\mu_{f_c \to x_4} &= \sum_{x_5} f_c(x_4, x_5)\mu_{f_d \to x_5} \\
\mu_{x_4 \to f_b} &= \mu_{f_c \to x_4} \\
\mu_{f_b \to x_2} &= \sum_{x_4} f_b(x_4, x_2)\mu_{x_4 \to f_b} \\
\mu_{x_3 \to f_a} &= 1 \\
\mu_{x_2 \to f_a} &= \mu_{f_b \to x_2} \\
\mu_{f_a \to x_1} &= \sum_{x_2, x_3} f_a(x_1, x_2, x_3)\mu_{x_2 \to f_a}\mu_{x_3 \to f_a}
\end{aligned}$$

The latter messages are now

$$\begin{aligned}
\mu_{x_1 \to f_a} &= 1 \\
\mu_{f_a \to x_2} &= \sum_{x_1, x_3} f_a(x_1, x_2, x_3)\mu_{x_1 \to f_a}\mu_{x_3 \to f_a} \\
\mu_{f_a \to x_3} &= \sum_{x_1, x_3} f_a(x_1, x_2, x_3)\mu_{x_1 \to f_a}\mu_{x_2 \to f_a} \\
\mu_{x_2 \to f_b} &= \mu_{f_a \to x_2} \\
\mu_{f_b \to x_4} &= \sum_{x_2} f_b(x_4, x_2)\mu_{x_2 \to f_b} \\
\mu_{x_4 \to f_c} &= \mu_{f_b \to x_4} \\
\mu_{f_c \to x_5} &= \sum_{x_4} f_c(x_4, x_5)\mu_{x_4 \to f_c} \\
\mu_{x_5 \to f_d} &= \mu_{f_c \to x_5}
\end{aligned}$$

Now, the marginal is computed as

$$p(x_4, x_5) = \frac{\mu_{f_b \to x_4} \mu_{f_d \to x_5}}{Z} \tag{9}$$

6. [10 points] Now if x_2 in Figure 4 is observed, explain how to conduct the sum-product algorithm, and compute the posterior distribution $p(x_4, x_5|x_2)$.

Let us assume that $x_2 = a$. In that case, we replace the distribution of x_2 with the observed values.

We first compute the message from leaf to root and then we compute from the root to leaf. We initialize the leaf as 1. The former messages are

$$\begin{aligned}
\mu_{f_d \to x_5} &= f_d(x_5) \\
\mu_{x_5 \to f_c} &= f_d(x_5) \\
\mu_{f_c \to x_4} &= \sum_{x_5} f_c(x_4, x_5) \mu_{f_d \to x_5} \\
\mu_{x_4 \to f_b} &= \mu_{f_c \to x_4} \\
\mu_{f_b \to x_2} &= \sum_{x_4} f_b(x_4, x_2) \mu_{x_4 \to f_b} \\
\mu_{x_3 \to f_a} &= 1 \\
\mu_{x_2 \to f_a} &= \mu_{f_b \to x_2} \\
\mu_{f_a \to x_1} &= \sum_{x_3} f_a(x_1, x_2 = a, x_3) \mu_{x_3 \to f_a}
\end{aligned}$$

The latter messages are now

$$\begin{aligned}
\mu_{x_1 \to f_a} &= 1 \\
\mu_{f_a \to x_2} &= \sum_{x_1, x_3} f_a(x_1, x_2, x_3) \mu_{x_1 \to f_a} \mu_{x_3 \to f_a} \\
\mu_{f_a \to x_3} &= \sum_{x_1, x_3} f_a(x_1, x_2, x_3) \mu_{x_1 \to f_a} \mu_{x_2 \to f_a} \\
\mu_{x_2 \to f_b} &= \mu_{f_a \to x_2} \\
\mu_{f_b \to x_4} &= f_b(x_4, x_2 = a) \mu_{x_2 \to f_b} \\
\mu_{x_4 \to f_c} &= \mu_{f_b \to x_4} \\
\mu_{f_c \to x_5} &= \sum_{x_4} f_c(x_4, x_5) \mu_{x_4 \to f_c} \\
\mu_{x_5 \to f_d} &= \mu_{f_c \to x_5}
\end{aligned}$$

Now,

$$\begin{aligned}
p(x_4, x_5 | x_2 = a) &= \frac{p(x_4, x_5, x_2 = a)}{p(x_2)} \qquad (10) \\
&= \frac{\mu_{f_a \to x_2} \mu_{f_d \to x_5}}{\mu_{f_a \to x_2} \mu_{f_b \to x_2}} \qquad (11) \\
&= \frac{\mu_{f_d \to x_5}}{\mu_{f_b \to x_2}} \qquad (12)
\end{aligned}$$

7. [10 points] Suppose all the random variables in Figure 4 are discrete, and no one has been observed. Now we want to find the configuration of the $x_1, \ldots, x_5$ to maximize the joint probability. Write done every step of the max-sum algorithm to calculate the maximum joint probability and to find the corresponding configurations of each random variable.

We follow the lecture 10 for the notations. The factor nodes are used to carry the optimization of the messages while the variable nodes simply pass the message obtained from other nodes. The figure is shown in Figure 5.

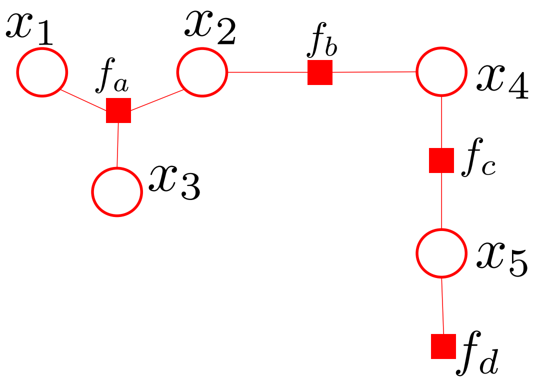


Figure 5: Factor graph.

$$p(x_1, x_2, .., x_5) = f_a(x_1, x_2, x_3) f_b(x_2, x_3) f_c(x_4, x_5) f_d(x_5) \tag{13}$$

The joint probability is maximised even by taking the ln since log is monotonically increasing function.

$$\max \ln p(x_1, x_2, .., x_5) = \max_{x_1,x_2,x_3} \ln f_a(x_1, x_2, x_3) + \max_{x_2,x_3} \ln f_b(x_2, x_3) + \max_{x_4,x_5} \ln f_c(x_4, x_5) + \max_{x_5} \ln f_d(x_5) \tag{14}$$

We initialize the leaf as 0. The former messages are

$$\begin{aligned}
\mu_{f_d \to x_5} &= \max \ln f_d \\
\mu_{x_5 \to f_c} &= \max \ln f_d \\
\mu_{f_c \to x_4} &= \max_{x_5} [\ln f_c(x_4, x_5) + \mu_{f_d \to x_5}] \\
\mu_{x_4 \to f_b} &= \mu_{f_c \to x_4} \\
\mu_{f_b \to x_2} &= \max_{x_4} [\ln f_b(x_4, x_2) + \mu_{x_4 \to f_b}] \\
\mu_{x_3 \to f_a} &= 0 \\
\mu_{x_2 \to f_a} &= \mu_{f_b \to x_2} \\
\mu_{f_a \to x_1} &= \max_{x_2,x_3} [\ln f_a(x_1, x_2, x_3) + \mu_{x_2 \to f_a} + \mu_{x_3 \to f_a}]
\end{aligned}$$

The latter messages are now

$$\begin{aligned}
\mu_{x_1 \to f_a} &= 0 \\
\mu_{f_a \to x_2} &= \max_{x_1,x_3} [\ln f_a(x_1, x_2, x_3) + \mu_{x_1 \to f_a} + \mu_{x_3 \to f_a}] \\
\mu_{f_a \to x_3} &= \max_{x_1,x_3} [\ln f_a(x_1, x_2, x_3) + \mu_{x_1 \to f_a} + \mu_{x_2 \to f_a}]
\end{aligned}$$

$$\begin{aligned}
\mu_{x_2 \to f_b} &= \mu_{f_a \to x_2} \\
\mu_{f_b \to x_4} &= \max_{x_2} [\ln f_b(x_4, x_2) + \mu_{x_2 \to f_b}] \\
\mu_{x_4 \to f_c} &= \mu_{f_b \to x_4} \\
\mu_{f_c \to x_5} &= \max_{x_4} [\ln f_c(x_4, x_5) + \mu_{x_4 \to f_c}] \\
\mu_{x_5 \to f_d} &= \mu_{f_c \to x_5}
\end{aligned}$$

Once we run belief propagation through multiple iterations, we get the stable values of all the nodes. Once we have obtained stability, we infer the optimal states of the nodes using the final beliefs at each of the nodes. We then substitute these values in equation 14 to get the maximum joint probability.

8. [**Bonus**][20 points] Show the message passing protocol we discussed in the class is always valid on the tree-structured graphical models— whenever we compute a message (from a factor to a variable or a variable to a factor), the dependent messages are always available.

From graph theory, a tree always has a root node and is always connected. Other than the root, every children has a unique parent node and therefore the tree has no cycles.

Let us conveniently assume that the leaf node are all variables x_i. (if the leaf node are all factors, there will be a single message exchange between the variables connected to leaf factors and therefore the situation is equivalent). We initialize the variables to all 1. The messages from $\mu_{x_i \to f_{a_i}} = 1$. Now since the factor node has a unique parent (property of tree), the parent variable node will receive a message $\mu_{f_{a_i} \to x_i} = \sum_{x_j = x \; x_i} f_{a_i}(x_j) \mu_{x_j \to f_c}$. These messages are all available since they are received from the children nodes. The variable node will then again pass the messages to their parent factor node and this continues till the root node. This is shown in figure 6.

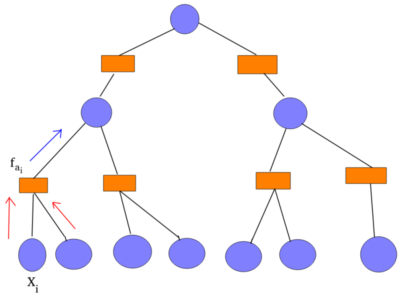


Figure 6: Tree graphical model with factor nodes (orange) and variable nodes (violet). When blue messages are sent in direction towards the root.

Once we are at the root node, we start passing the messages from the root node and we can use the messages from the previous pass to send messages to the other root node.

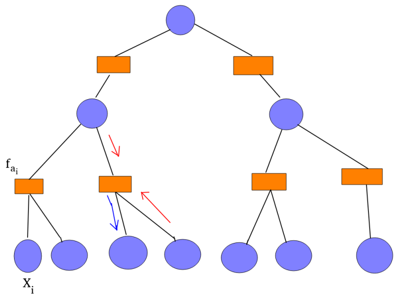


Figure 7: Tree graphical model with factor nodes (orange) and variable nodes (violet). When blue messages are sent in direction of root towards the leaves

9. [15 points] Use d-separation algorithm to determine if $a \perp\!\!\!\perp d|e$ in the graphical model shown in Figure 8, and if $a \perp\!\!\!\perp d|b$ in the graphical model shown in Figure 9.

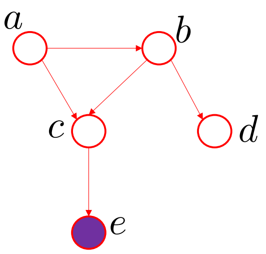


Figure 8: Model 1.

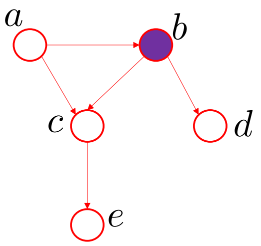


Figure 9: Model 2.

To check d-separation, we marry the parents of the variables concerned and then convert it to undirected graph. We then delete the nodes and the paths associated with the nodes that are observed and see if there is a path to reach from one node to another. If a path exists, then the nodes are not independent. We use the handout at http://web.mit.edu/jmn/www/6.034/d-separation.pdf

(a) For Figure 8, after marrying, disorientation and removing the given node e and connections from the graph, we see that there is a path from a to d via node b and hence they are not conditionally independent. That is $a \not\perp\!\!\!\perp d|e$

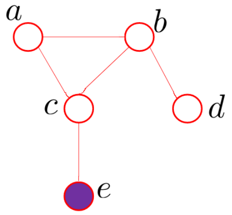


Figure 10: Marrying ancestors and disorientation

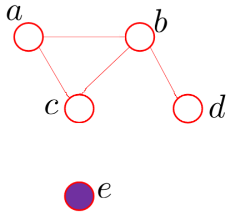


Figure 11: Pruning

(b) For Figure 9, after marrying, disorientation and removing the node b from the graph, we see that there is no path from a to d and hence they are now conditionally independent. That is $a \perp\!\!\!\perp d|b$

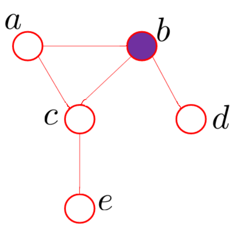


Figure 12: Marrying ancestors and disorientation

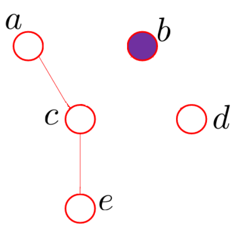
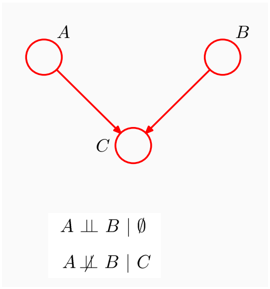


Figure 13: Pruning

10. [**Bonus**][20 points] We have listed two examples in the class to show that in terms of the expressiveness (i.e., conditional independence) of the directed and undirected graphical models , there is not a guarantee that who is better than who.

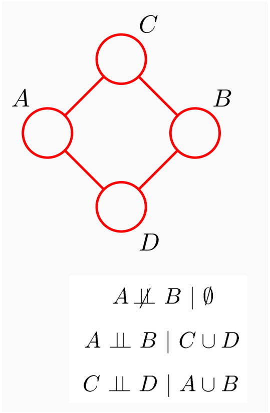
    (a) [10 points] Now show that for the directed graphical model in Figure 14, we cannot find an equivalent undirected graphical model to express the same set of conditional independence.

    (b) [10 points] Show that for the undirected graphical model in Figure 15, we cannot find an equivalent directed graphical model to express the same set of conditional independence.



A
B
C
$A \perp\!\!\!\perp B \mid \emptyset$
$A \not\perp\!\!\!\perp B \mid C$

Figure 14: Directed.



C
A
B
D
$A \not\perp\!\!\!\perp B \mid \emptyset$
$A \perp\!\!\!\perp B \mid C \cup D$
$C \perp\!\!\!\perp D \mid A \cup B$

Figure 15: Undirected.

(a) For Figure 14, when we convert from directed to undirected graphical model, we also have an edge between A and B Hence, $A \not\perp B$ irrespective of the fact that C was observed or not. And so there is no other graph possible as shown in Figure 16.

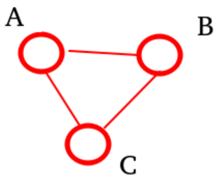


Figure 16: Correponding undirected graphical model for Figure 14

(b) We will show this by contradiction. Let us assume that there exists a directed graphical model. The first condition $A \not\perp B | \phi$ means that A and B are not parent nodes and there exists atleast one directed path from $A \rightarrow B$ or $B \rightarrow A$. Since, the conditional independence equations listed in Figure 15 is symmetric about A and B, with no loss of generality, we assume that the path is $A \rightarrow B$ which we show in Figure 17.

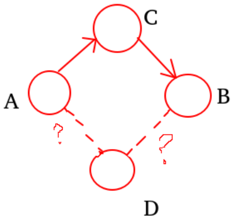


Figure 17: Making a directed acyclic graphical model after deciding one of the paths being $A \rightarrow B$.

Next, using Figure 17, there can be only four possibilities of directed graphs.

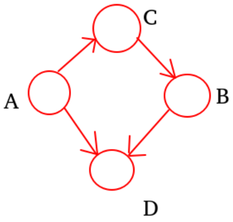


Figure 18

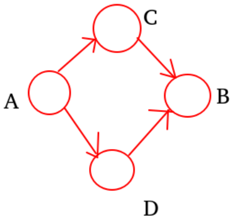


Figure 19

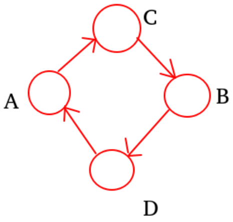


Figure 20

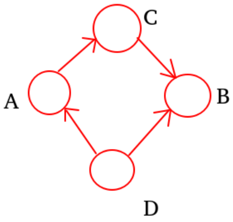


Figure 21

For Figure 18, we do not satisfy $A \not\perp B | C \cup D$ since D becomes a head to head connection or collider.

For Figure 19, we do not satisfy $C \not\perp D | A \cup B$ since B becomes a head to head connection or collider.

For Figure 20, we have a cyclic graph which is not allowed for a Bayesian network. The problem wants us to say on the directed cyclic graph and therefore this graphical model is not correct.

For Figure 21, we do not satisfy $C \not\perp D | A \cup B$ since B becomes a head to head connection or collider.

Hence, all the four cases are either invalid or do not satisfy conditional independence assumption. Hence, our assumption that there is an directed acyclic graph satisfying all conditional independence assumption is wrong. Thus, we do not have any directed acyclic graph for the same.