CS 6190: Probabilistic Machine Learning Spring 2022

Homework 3

Handed out: 15 Mar, 2022
Due: 11:59pm, 29 Mar, 2022

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down your own solution. Please keep the class collaboration policy in mind.
- Feel free discuss the homework with the instructor or the TAs.
- Your written solutions should be brief and clear. You need to show your work, not just the final answer, but you do *not* need to write it in gory detail. Your assignment should be **no more than 10 pages**. Every extra page will cost a point.
- Handwritten solutions will not be accepted.
- The homework is due by **midnight of the due date**. Please submit the homework on Canvas.

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 conditioned on c, i.e., $p(a, b|c) = p(a|c)p(b|c)$.

   Answer

   Let's first write up a few probabilities that will be used to calculate joint distribution.

   From the table 1 we know that $p(a = 0) = 0.6$, $p(b = 0) = 0.592$ and $p(a = 0, b = 0) = 0.336$. If a and b were to be independent, $p(a, b) = p(a)p(b)$ but from these calculations we can see that this is not the case, hence they are not independent. Now to prove the marginal dependence, we have to list out of the marginal probabilities and joint probabilities, and then make the claim.

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.

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

(a) Marginal distribution of a.

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

(b) Marginal distribution of b.

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

(c) Joint distribution of a, b.

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

Table 2 shows the marginal and joint distribution of a, b when c is given. From this table we can see that $p(a, b|c) = p(a|c)p(b|c)$ for all a, b, c hence it's proved that they are marginally dependent.

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

Answer

Markov blanket for a node means that the blanket that covers the select node from rest of the network, i.e. it's parents, children and children's parents. This also means that rest of the nodes are independent of the select node and we don't need them to predict behaviour of the select node. Consider Figure 1 where we can see the blanket which shows which nodes are independent and which aren't with respect to node μ.

Using the cases of d-separation, the grandparents of μ are independent because the parents of μ are dependent. Similarly, the grandchildren will be independent because the children of μ are dependent. However if we have something like the case of head to head, we could get grandparents of the children of μ dependent on μ. This will not be the case because we have the parents of the children in blanket. Hence the grandparents of the children are independent of μ, and all other nodes conditioned independent as well 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 $a \not\perp\!\!\!\perp b|d$.

Answer

We know that $p(a, b, c, d) = p(d|c)p(c|a, b)p(a)p(b)$. The next thing to see is that if a and b are independent if nothing is observed. We do this by marginalising c and d.

$$p(a, b) = \sum_c \sum_d p(a, b, c, d)$$

$$p(a, b) = \sum_c p(c|a, b)p(a)p(b) \sum_d p(d|c)$$

$$p(a, b) = \sum_c p(c|a, b)p(a)p(b)$$

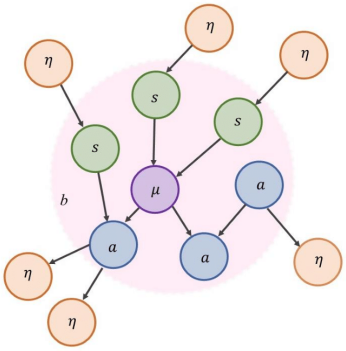


Figure 1: Markov Blanket, source- researchgate.net

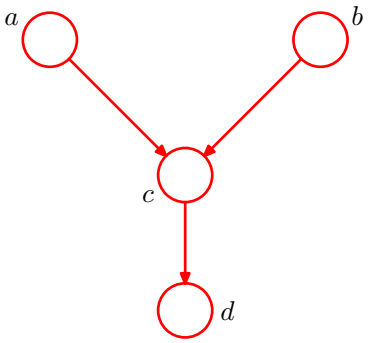


Figure 2: Graphical model.

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

This proves that a and b are independent conditioned on nothing.

Apart from this, we also have to show conditional independence, i.e. $p(a, b|d) = p(a|d)p(b|d)$.

Taking the left hand side,

$$p(a, b|d) = \frac{p(a, b, d)}{p(d)}$$

$$p(a, b|d) = \frac{\sum_c p(a, b, c, d)}{p(d)}$$

$$p(a, b|d) = \frac{\sum_c p(d|c)p(c|a, b)p(a)p(b)}{p(d)}$$

$$p(a, b|d) = \frac{p(d|a, b)p(a)p(b)}{p(d)} \neq p(a|d)p(b|d)$$

Hence we can say that they're 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.

Answer

Figure 3 shows the undirected version of graph in Figure 2, and the functions are as follows:

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

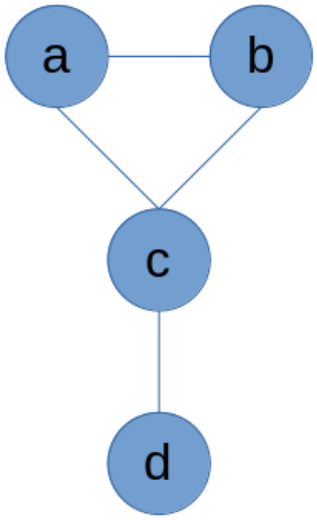


Figure 3: Graphical model.

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 $p(x_4, x_5)$.

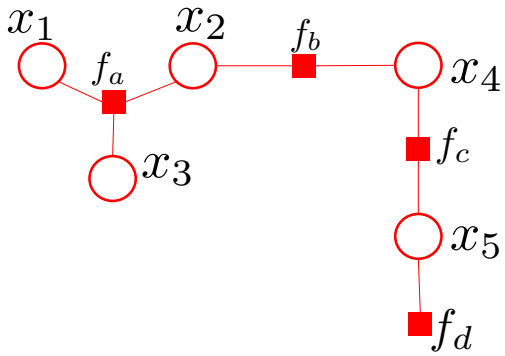


Figure 4: Factor graph.

Answer

From Figure 4, let's say that x_1 is the root note and f_d is the leaf node. Now, we have to compute the message from leaf to root and vice-versa.

Note that leaf is initialised as 1.

$$\mu_{f_d \to x_5} = f_d(x_5)$$

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

Now the messages in reverse direction

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

The marginal is computed by taking the desired terms from the above.

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

Z in above is the normalising constant.

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 $p(x_4, x_5|x_2)$.

Answer

The messages in this case is same as above with changes to x_2 only as it is a constant now. Let's say this constant as k.

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

$$\mu_{f_a \to x_1} = \sum_{x_3} f_a(x_1, x_2 = k, x_3) \mu_{x_3 \to f_a}$$

Now the messages in reverse direction

$$\mu_{x_1 \to f_a} = 1$$

$$\mu_{f_a \to x_2} = \sum_{x_1} \sum_{x_3} f_a(x_1, x_2, x_3) \mu_{x_1 \to f_a} \mu_{x_3 \to f_a}$$

$$\mu_{f_a \to x_3} = \sum_{x_1} \sum_{x_3} f_a(x_1, x_2, x_3) \mu_{x_1 \to f_a} \mu_{x_2 \to f_a}$$

$$\mu_{x_2 \to f_b} = \mu_{f_a \to x_2}$$

$$\mu_{f_b \to x_4} = f_b(x_4, x_2 = k) \mu_{x_2 \to f_b}$$

$$\mu_{x_4 \to f_c} = \mu_{f_b \to x_4}$$

$$\mu_{f_c \to x_5} = \sum_{x_4} f_c(x_4, x_5) \mu_{x_4 \to f_c}$$

$$\mu_{x_5 \to f_d} = \mu_{f_c \to x_5}$$

After this we compute the conditional as follows

$$p(x_4, x_5 | x_2 = k) = \frac{p(x_4, x_5, x_2 = k)}{p(x_2)}$$

$$p(x_4, x_5 | x_2 = k) = \frac{\mu_{f_a \to x_2} \mu_{f_d \to x_5}}{\mu_{f_a \to x_2} \mu_{f_b \to x_2}}$$

$$p(x_4, x_5 | x_2 = k) = \frac{\mu_{f_d \to x_5}}{\mu_{f_b \to x_2}}$$

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.

Answer

Looking at Figure 4, we can write joint probability as $p(x_1, x_2, \cdot, x_5) = f_a(x_1, x_2, x_3) f_b(x_2, x_3) f_c(x_4, x_5) f_d(x_5)$. This can be maximised by taking log on both sides.

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

Here, initialising the leaf as 1 as well.

The messages can now be written as

$$\mu_{f_d \to x_5} = \max \ln f_d(x_5)$$

$$\mu_{x_5 \to f_c} = \max \ln f_d(x_5)$$

$$\mu_{f_c \to x_4} = \max_{x_5} (\ln f_c(x_4, x_5) \mu_{f_d \to x_5})$$

$$\mu_{x_4 \to f_b} = \mu_{f_c \to x_4}$$

$$\mu_{f_b \to x_2} = \max_{x_4} (\ln f_b(x_4, x_2) \mu_{x_4 \to f_b})$$

$$\mu_{x_3 \to f_a} = 1$$

$$\mu_{x_2 \to f_a} = \mu_{f_b \to x_2}$$

$$\mu_{f_a \to x_1} = \max_{x_2, x_3} (\ln f_a(x_1, x_2, x_3) \mu_{x_2 \to f_a} \mu_{x_3 \to f_a})$$

Now the messages in reverse direction

$$\mu_{x_1 \to f_a} = 1$$

$$\mu_{f_a \to x_2} = \max_{x_1, x_3} (_a(x_1, x_2, x_3) \mu_{x_1 \to f_a} \mu_{x_3 \to f_a})$$

$$\mu_{f_a \to x_3} = \max_{x_1, x_3} (\ln f_a(x_1, x_2, x_3) \mu_{x_1 \to f_a} \mu_{x_2 \to f_a})$$

$$\mu_{x_2 \to f_b} = \mu_{f_a \to x_2}$$

$$\mu_{f_b \to x_4} = \max_{x_2} (\ln f_b(x_4, x_2) \mu_{x_2 \to f_b})$$

$$\mu_{x_4 \to f_c} = \mu_{f_b \to x_4}$$

$$\mu_{f_c \to x_5} = \max_{x_4} (\ln f_c(x_4, x_5) \mu_{x_4 \to f_c})$$

$$\mu_{x_5 \to f_d} = \mu_{f_c \to x_5}$$

We can get stable values of all nodes after running propagation through several iterations and then infer the optimal states of these nodes. These optimal values are then substituted in joint probability distribution equation to get the maximum of that equation.

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. Hint: use induction.

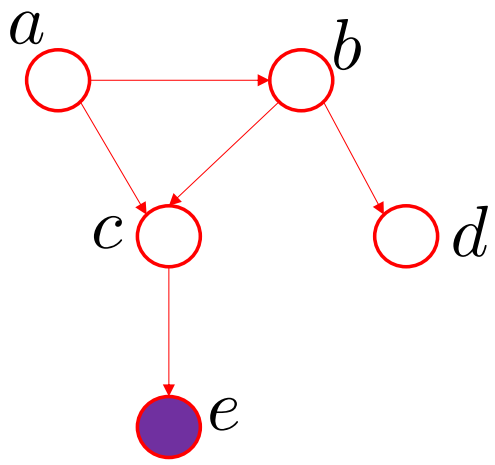


Figure 5: Model 1.

Answer

From Graph Theory, we know that the tree is acyclic and always have a root node. Let say that all leaf nodes are denoted by x_i, i.e. these are at the bottom of the tree and are independent. Consider the message from these variables to be a constant 1. We also know that in tree structure, each factor node have a unique parent, hence this parent node will receive a message $\sum_x f_{ai}(x_j) \mu_{x_j \to f_c}$. The components making this message are always available since they're coming from down the tree (coming from children), and this will get passed up the tree to the root node. At the root node, we can start passing the message back and reuse the messages from the previous pass to send the messages to other root node.

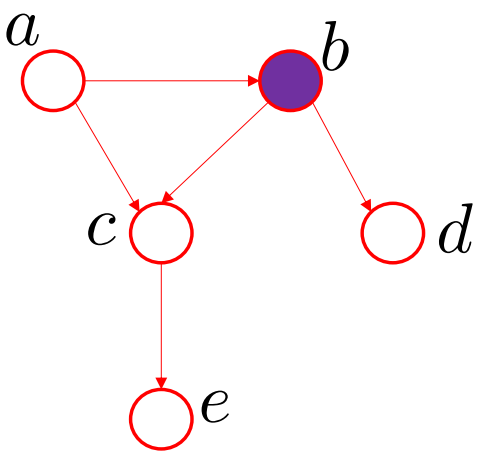


Figure 6: Model 2.

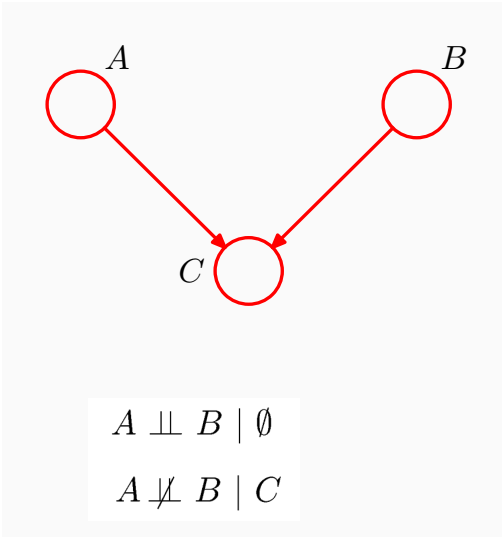


Figure 7: Directed.

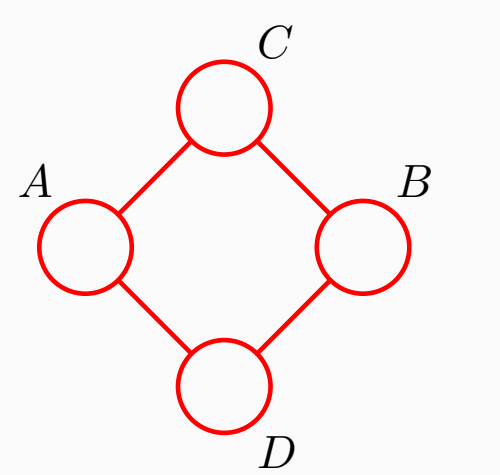
9. [15 points] Use d-separation algorithm (Bayes ball) to determine if $a \perp\!\!\!\perp d|e$ in the graphical model shown in Figure 5, and if $a \perp\!\!\!\perp d|b$ in the graphical model shown in Figure 6.

   Answer

   We know that if there exists a path from a to d after deleting the givens nodes, a and d are not required to be conditionally independent given the node. For example we can see this in Figure 7. The answer is shown in 9 where we can see that after moralising, disorientating and removing the given nodes from the graph, there's still a path.

   We also know that if there exists no path from a to d after deleting the givens nodes, a and d are conditionally independent given the node. For example we can see this in Figure 8. The answer is shown in 10 where we can see that after moralising, disorientating and removing the given nodes from the graph, there's no path from a to d.

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.



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

$$A \perp\!\!\!\perp B \mid C \cup D$$

$$C \perp\!\!\!\perp D \mid A \cup B$$

Figure 8: Undirected.

(a) [10 points] Now show that for the directed graphical model in Figure 7, 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 8, we cannot find an equivalent directed graphical model to express the same set of conditional independence.

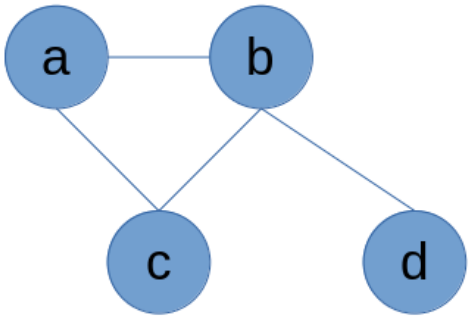


Figure 9: Graph after moralising, disorientating and removing the given nodes

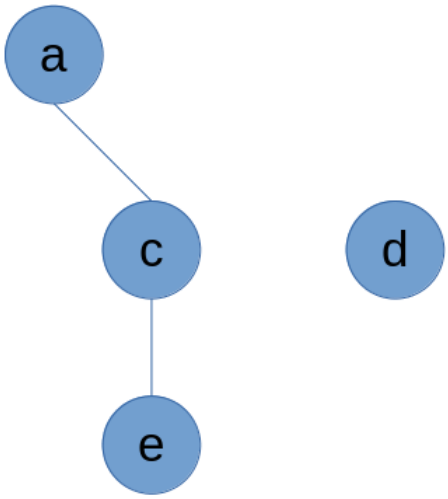


Figure 10: Graph after moralising, disorientating and removing the given nodes