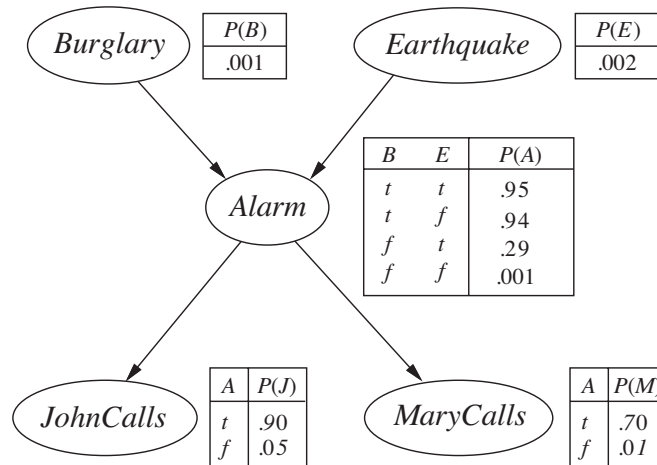


Worksheet 11: PGMs II

COMP90051 Statistical Machine Learning

Semester 1, 2021

Exercise 1 (Based on RN 14.15). Leo is a botanist who lives in the Bay Area.¹ His neighbourhood is a hotspot for burglars, so his house is fitted with an alarm system. Unfortunately, the alarm is not perfectly reliable: it doesn't always trigger during a home invasion, and it may be erroneously triggered during minor earthquakes, which occur occasionally. Leo has asked his neighbours John and Mary (who don't know each other) to call him if they hear the alarm.



- If no evidence is observed, are the *Burglary* and *Earthquake* nodes independent?
- Assume we observe $Alarm = t$, now are *Burglary* and *Earthquake* independent? Justify your answer by calculating whether the probabilities involved satisfy the definition of conditional independence.
- [Discussion]** Explain intuitively why knowledge of the *Alarm* random variable renders the random variables *Earthquake* and *Burglary* non-conditionally independent.

Solution.

- Marginalizing over unobserved variables in the joint distribution and simplifying the summation, we have

$$\begin{aligned}
 p(B, E) &= \sum_A \sum_J \sum_M p(A, B, E, J, M) \\
 &= \sum_A \sum_J \sum_M p(B)p(E)p(A|B, E)p(J|A)p(M|A) \\
 &= p(B)p(E) \sum_A p(A|B, E) \sum_J p(J|A) \sum_M p(M|A)
 \end{aligned}$$

¹Deja vu! Yes, this build on last week's exercise.

Note that every sum above is equal to 1 by the law of total probability—i.e. the variables A, J, M are irrelevant to the query. More generally, every variable that is not an ancestor of a query/evidence variable is irrelevant. So we have $p(B, E) = p(B)p(E)$ in this case.

2. By Bayes' Theorem

$$\begin{aligned} p(B|E, a) &= \frac{1}{Z} p(B, E, a) \\ &= \frac{1}{Z} p(a|B, E) p(B) p(E) \\ &= \frac{p(a|B, E) p(B)}{\sum_B p(a|B, E) p(B)} \end{aligned}$$

Where $Z = \sum_B p(a|B, E) p(B) p(E)$ is a normalizing constant (confusingly, also called the 'evidence'). For B and E to be conditionally independent given $A = a$, $p(B|E, a)$ should not depend on E . This is not possible as $p(a|B, E)$ depends on E (validate this numerically).

3. Under our assumptions, two possible explanations for the alarm going off ($Alarm = t$) are an earthquake occurring ($Earthquake = t$) and a burglary occurring ($Burglary = t$). If we know that $Alarm = t$ and $Earthquake = t$ then intuitively we think that the likelihood a burglary has occurred is low ($Burglary = f$) (Leo and the burglar would have to be very unlucky to have an earthquake and burglary occur at the same time):

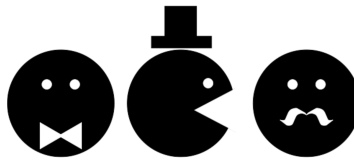
$$p(b|a, e) < p(\neg b|a, e) \quad (1)$$

This is reasonably intuitive: for observed data with two possible explanations, knowing one of the explanations has occurred decreases the probability that the other possible explanation has also occurred—there is now a conditional dependence between causes, given observation of the children.

Hence for any PGM with the structure $X \rightarrow Z \leftarrow Y$, knowledge of Z couples X and Y .

□

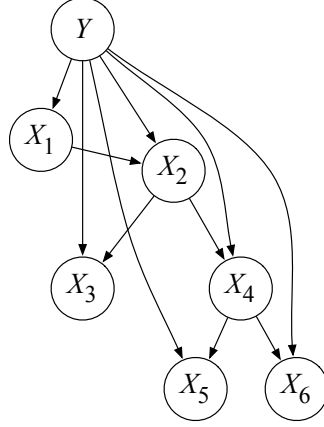
Exercise 2. Mr. and Ms. Pacman have been searching for each other in the Pacman world).² Ms. Pacman has been pregnant with a baby, and this morning she has given birth to Pacbaby (congratulations, Pacmans!). To train Pacbaby to avoid encountering ghosts in the maze,³ the Pacmans are trying to teach Pacbaby to distinguish Pacmen (pl.) from ghosts using discriminative visual features such as the presence of a bowtie, hat, mustache, etc.



Pacbaby has noticed that the features are not independent—nearly everyone who has a hat has a mustache, while those with bowties are always clean shaven. She decides to use a tree-augmented Naive Bayes model (TANB) to account for conditional dependencies. A TANB is an extension of a Naive Bayes model, where features are no longer assumed conditionally independent given the binary class $Y \in \{1, -1\}$ (Pacman or not-Pacman, respectively). Let X_1, X_2, \dots, X_6 be the random variables corresponding to the features that Pacbaby observes. The TANB model arranges vertices in a tree-structured Bayes net with Y at the root:

²Based on Berkeley CS188 section, http://ai.berkeley.edu/project_overview.html

³Ghosts are nice enough not to eat Pacbaby, but they will take all her money.



- (a) Assume all features X_1, \dots, X_6 are observed in the TANB model. What is the classification rule? Your answer should be in terms of the prior and conditional probabilities.

Solution. We need to find the value of the class $Y = y$ that maximises the posterior $p(y|X_1 = x_1, X_2 = x_2, \dots, X_6 = x_6)$. Using Bayes' Theorem and the Bayes net factorisation of the joint distribution:

$$\begin{aligned}
 y^* &= \arg \max_y p(y|x_1, \dots, x_6) \\
 &= \arg \max_y p(y, x_1, \dots, x_6) \\
 &= \arg \max_y p(y)p(x_1|y)p(x_2|x_1, y)p(x_3|x_2, y)p(x_4|x_2, y)p(x_5|x_4, y)p(x_6|x_4, y)
 \end{aligned}$$

□

When we perform a marginalisation operation—i.e. removing a variable from a joint distribution, we perform a sum over the product of all factors that include that random variable. For example, marginalising over X_4 in the joint distribution above involves a factor containing four random variables.

$$\sum_{X_4} \underbrace{p(X_4|X_2, Y)p(X_5|X_4, Y)p(X_6|X_4, Y)}_{\phi(X_2, X_4, X_5, X_6)}$$

This induces a dependency between all the random variables in the factor except the variable being marginalised—all subsequent operations will have to treat X_2, X_5, X_6 together (X_4 is summed out). Assuming there is no special algebraic structure in the summand that can be exploited, the complexity is exponential in the number of different random variables in the summand. Thus the overall complexity of the variable elimination algorithm is dominated by the number of variables in the largest elimination factor, $\phi(\dots)$. Determining the optimal (lowest-complexity) elimination ordering is intractable, but a useful heuristic is to find an ordering that minimises the size of the largest factor generated.

- (b) Specify an elimination order that is efficient for the query $p(Y|X_5 = x_5)$ in the TANB model above. How many variables are in the biggest factor induced by variable elimination with your ordering? Which variables are they?

Solution. Here the query variable (the variable of interest) is Y while the evidence (observed variables) is $X_5 = x_5$. We saw last week that to find the posterior:

- We use Bayes' Theorem to express the posterior as proportional to the joint over the query and evidence.

- We express the joint over the query and evidence as the full joint, marginalised over the remaining variables.
- We use the Bayes net structure to factorise the full joint as the product of simpler conditional probability tables.

$$\begin{aligned}
p(Y|x_5) &\propto p(Y, x_5) = \sum_{X_1, X_2, X_3, X_4, X_6} p(Y, X_1, X_2, X_3, X_4, x_5, X_6) \\
&= \sum_{X_1, X_2, X_3, X_4, X_6} p(Y)p(X_1|Y)p(X_2|X_1, Y)p(X_3|X_2, Y)p(X_4|X_2, Y)p(x_5|X_4, Y)p(X_6|X_4, Y)
\end{aligned}$$

Note that marginalising X_3 and X_6 in the factored distribution above sums to one. Hence we only have to consider the sums over X_1, X_2, X_4 . Also note that X_1 and X_4 are connected to one undetermined variable (X_2), but X_2 is connected to two undetermined variables (X_1, X_4). This suggests we should use the variable elimination algorithm to eliminate either of X_1, X_4 first before eliminating X_2 , to minimise the number of variables in the largest factor, for example, walking up the tree:

$$\begin{aligned}
p(Y|x_5) &= \sum_{X_1, X_2, X_4} p(Y)p(X_1|Y)p(X_2|X_1, Y)p(X_4|X_2, Y)p(x_5|X_4, Y) \\
&= p(Y) \sum_{X_1} p(X_1|Y) \sum_{X_2} p(X_2|X_1, Y) \underbrace{\sum_{X_4} p(X_4|X_2, Y)p(x_5|X_4, Y)}_{\phi(X_2, X_4, Y)} \\
&= p(Y) \sum_{X_1} p(X_1|Y) \underbrace{\sum_{X_2} p(X_2|X_1, Y)m(X_2)}_{\phi(X_1, X_2, Y)} \\
&= p(Y) \sum_{X_1} p(X_1|Y)m(X_1) \\
&= p(Y)m_1
\end{aligned}$$

The largest factor size is 3. Convince yourself the reverse order (X_1, X_2, X_4) is similar. What happens if we attempt to eliminate X_2 before X_1 or X_4 ?

$$p(Y|x_5) = \sum_{X_1, X_4} p(Y)p(X_1|Y)p(x_5|X_4, Y) \underbrace{\sum_{X_2} p(X_2|X_1, Y)p(X_4|X_2, Y)}_{\phi(X_1, X_2, X_4, Y)}$$

Notice we get a factor of size 4. Remember that the time complexity of inference is exponential in the size of the largest factor, so we shouldn't eliminate X_2 first. Any ordering that eliminates one of X_1 or X_4 before X_2 is acceptable, and the size of the largest factor should be 3. By the variable elimination algorithm, the last variable to be eliminated should be Y , which is needed to normalise the final result. \square

- (c) Specify an elimination order that is efficient for the query $p(X_3|X_5 = x_5)$ in the TANB model above. How many variables are in the biggest factor induced by variable elimination with your ordering? Which variables are they?

Solution. Repeating the arguments above with query X_3 and evidence $X_5 = x_5$:

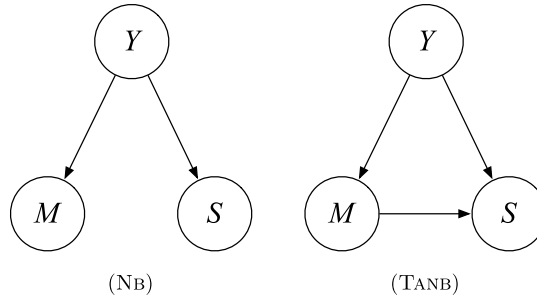
$$\begin{aligned}
p(X_3|x_5) &\propto p(X_3, x_5) = \sum_{Y, X_1, X_2, X_4, X_6} p(Y, X_1, X_2, X_3, X_4, x_5, X_6) \\
&= \sum_{Y, X_1, X_2, X_4} p(Y)p(X_1|Y)p(X_2|X_1, Y)p(X_3|X_2, Y)p(X_4|X_2, Y)p(x_5|X_4, Y)
\end{aligned}$$

Where marginalising over X_6 sums to 1 and does not effect inference. Notice that Y and X_2 are connected to 3 vertices in the TANB, which means that eliminating either early on will result in a factor of size 4 (try it for yourself). Hence we should eliminate one of X_1 or X_4 before eliminating X_2 or Y —any ordering of the remaining three random variables will be fine. For example:

$$\begin{aligned}
p(X_3|x_5) &\propto \sum_Y p(Y) \sum_{X_1} p(X_1|Y) \sum_{X_2} p(X_2|X_1, Y) p(X_3|X_2, Y) \sum_{X_4} \underbrace{p(X_4|X_2, Y) p(x_5|X_4, Y)}_{\phi(X_2, X_4, Y)} \\
&= \sum_Y p(Y) \sum_{X_1} p(X_1|Y) \sum_{X_2} \underbrace{p(X_2|X_1, Y) p(X_3|X_2, Y) m(X_2, Y)}_{\phi(X_1, X_2, Y)} \\
&= \sum_Y p(Y) \sum_{X_1} \underbrace{p(X_1|Y) m(X_1, Y)}_{\phi(X_1, Y)} \\
&= \sum_Y p(Y) m(Y)
\end{aligned}$$

The largest factor ϕ generated is of size 3 for $(X_2, X_4, Y), (X_1, X_2, Y)$. □

Exercise 3. Consider the Bayes nets below over the nodes Y (Pacbaby sees Pacman or not), M (Pacbaby sees a moustache), and S (Pacbaby sees sunglasses).



Empirically:

- Pacbaby observes $Y = 1$ or $Y = -1$ (Pacman or not) 50% of the time.
- Given $Y = 1$, Pacbaby observes $M = 1$ (moustache) 50% of the time and $S = 1$ (sunglasses) 50% of the time.
- When Pacbaby observes $Y = -1$, the frequency of observations are identical (equal probabilities of $M = 1, -1, S = 1, -1$).
- When Pacbaby observes $Y = 1$, anyone with a moustache wears sunglasses and anyone without a moustache does not wear sunglasses.
- If $Y = -1$ the presence/absence of a moustache has no influence on sunglasses.

(a) Based on the above information, fill in Pacbaby's conditional probability tables.

Solution.

For NB (left model)			For TANB (right model)																																									
<table><tr><th>y</th><th colspan="2">$\mathbb{P}(Y = y)$</th></tr><tr><td>1</td><td colspan="2">.5</td></tr><tr><td>-1</td><td colspan="2">.5</td></tr></table>			y	$\mathbb{P}(Y = y)$		1	.5		-1	.5		<table><tr><th>y</th><th colspan="4">$\mathbb{P}(Y = y)$</th></tr><tr><td>1</td><td colspan="4">.5</td></tr><tr><td>-1</td><td colspan="4">.5</td></tr></table>					y	$\mathbb{P}(Y = y)$				1	.5				-1	.5																
y	$\mathbb{P}(Y = y)$																																											
1	.5																																											
-1	.5																																											
y	$\mathbb{P}(Y = y)$																																											
1	.5																																											
-1	.5																																											
<table><tr><th></th><th colspan="2">$\mathbb{P}(M = m \mid Y = y)$</th></tr><tr><th></th><th>$y = 1$</th><th>$y = -1$</th></tr><tr><td>$m = 1$</td><td>.5</td><td>.5</td></tr><tr><td>$m = -1$</td><td>.5</td><td>.5</td></tr></table>				$\mathbb{P}(M = m \mid Y = y)$			$y = 1$	$y = -1$	$m = 1$.5	.5	$m = -1$.5	.5	<table><tr><th></th><th colspan="4">$\mathbb{P}(M = m \mid Y = y)$</th></tr><tr><th></th><th colspan="2">$y = 1$</th><th colspan="2">$y = -1$</th></tr><tr><td>$m = 1$</td><td>.5</td><td>.5</td><td>.5</td><td>.5</td></tr><tr><td>$m = -1$</td><td>.5</td><td>.5</td><td>.5</td><td>.5</td></tr></table>						$\mathbb{P}(M = m \mid Y = y)$					$y = 1$		$y = -1$		$m = 1$.5	.5	.5	.5	$m = -1$.5	.5	.5	.5					
	$\mathbb{P}(M = m \mid Y = y)$																																											
	$y = 1$	$y = -1$																																										
$m = 1$.5	.5																																										
$m = -1$.5	.5																																										
	$\mathbb{P}(M = m \mid Y = y)$																																											
	$y = 1$		$y = -1$																																									
$m = 1$.5	.5	.5	.5																																								
$m = -1$.5	.5	.5	.5																																								
<table><tr><th></th><th colspan="2">$\mathbb{P}(S = s \mid Y = y)$</th></tr><tr><th></th><th>$y = 1$</th><th>$y = -1$</th></tr><tr><td>$s = 1$</td><td>.5</td><td>.5</td></tr><tr><td>$s = -1$</td><td>.5</td><td>.5</td></tr></table>				$\mathbb{P}(S = s \mid Y = y)$			$y = 1$	$y = -1$	$s = 1$.5	.5	$s = -1$.5	.5	<table><tr><th></th><th colspan="4">$\mathbb{P}(S = s \mid Y = y, M = m)$</th></tr><tr><th></th><th colspan="2">$y = 1$</th><th colspan="2">$y = -1$</th></tr><tr><th></th><th>$m = 1$</th><th>$m = -1$</th><th>$m = 1$</th><th>$m = -1$</th></tr><tr><td>$s = 1$</td><td>1</td><td>0</td><td>.5</td><td>.5</td></tr><tr><td>$s = -1$</td><td>0</td><td>1</td><td>.5</td><td>.5</td></tr></table>						$\mathbb{P}(S = s \mid Y = y, M = m)$					$y = 1$		$y = -1$			$m = 1$	$m = -1$	$m = 1$	$m = -1$	$s = 1$	1	0	.5	.5	$s = -1$	0	1	.5	.5
	$\mathbb{P}(S = s \mid Y = y)$																																											
	$y = 1$	$y = -1$																																										
$s = 1$.5	.5																																										
$s = -1$.5	.5																																										
	$\mathbb{P}(S = s \mid Y = y, M = m)$																																											
	$y = 1$		$y = -1$																																									
	$m = 1$	$m = -1$	$m = 1$	$m = -1$																																								
$s = 1$	1	0	.5	.5																																								
$s = -1$	0	1	.5	.5																																								

□

- (b) Pacbaby sees someone with a moustache and wearing a pair of sunglasses. What prediction does the NB model make? What probability does it assign to its prediction? What prediction does Pacbaby's TANB model make? What probability does it assign to its prediction?

Solution. For the NB model, Pacbaby assigns probability 1/2 to each label and is undecided.

$$\begin{aligned}
y_{NB}^* &= \arg \max_y p(Y = y | M = 1, S = 1) \\
&= \arg \max_y p(Y = y) p(M = 1 | Y = y) p(S = 1 | Y = y) \\
&= \arg \max_y \begin{cases} (1/2)^3, & \text{for } Y = 1 \\ (1/2)^3, & \text{for } Y = -1 \end{cases}
\end{aligned}$$

For the TANB model, features are no longer assumed conditionally independent given observation of y :

$$\begin{aligned}
y_{TANB}^* &= \arg \max_y p(Y = y | M = 1, S = 1) \\
&= \arg \max_y p(Y = y) p(M = 1 | Y = y) p(S = 1 | M = 1, Y = y) \\
&= \arg \max_y \begin{cases} (1/2)^2, & \text{for } Y = 1 \\ (1/2)^3, & \text{for } Y = -1 \end{cases}
\end{aligned}$$

Normalising, Pacbaby assigns a probability of 2/3 that she saw a Pacman.

□