LEIDEN UNIVERSITY

Examination in Bayesian Networks **SESSION 2009–2010**

Wednesday, 12th May, 2010 **(11.15 – 13.15)**

*Answer the following **TWO** questions. Each question is worth 50 marks; the marks for each part are shown in brackets.*

Question 1

Consider the Bayesian network $\mathcal{B} = (G, P)$, with acyclic directed graph G as shown in Figure 1, and joint probability distribution P given below. All variables in the network are assumed

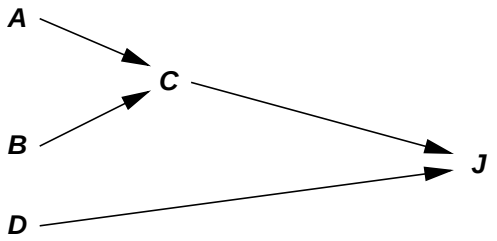

Figure 1: Bayesian network of Question 1.

to be binary. A variable V having the value $\top$ (true) is also indicated by v, whereas $\bot$ (false) is also denoted by $\neg v$. The following (local) probability distributions are defined for this network:

$$P(a) = 1,\ P(b) = 0.5,\ P(d) = 0.3$$

and

$$P(c \mid A, B) = \begin{cases} 0.8 & \text{if } A = B = \top \\ 0.3 & \text{if } A = \top \text{ and } B = \bot \\ 0 & \text{otherwise} \end{cases}$$

Finally, it is assumed that variable J models a noisy Boolean function f (e.g. a noisy OR) with the following parameters:

$$P(i_C \mid c) = 0.5 \qquad P(i_C \mid \neg c) = 0$$
$$P(i_D \mid d) = 0.3 \qquad P(i_D \mid \neg d) = 0$$

where I_C represents an intermediate variable between C and J and I_D an intermediate variable between D and J, respectively. These variables are not indicated in the network of Figure 1, but used to represent a *causal independence model* with respect to variable J. In other words, the family of probability distributions $P(J \mid C, D)$ is defined in terms of the Boolean function f, which corresponds to $P(J \mid I_C, I_D)$, and the parameters $P(I_C \mid C)$ and $P(I_D \mid D)$ respectively.

a. Compute the following probabilities (showing how you obtained your result): $P(c)$, $P(a \mid c)$, $P(a \mid d)$, and $P(a, \neg b \mid c)$. **[15]**

b. Briefly discuss why it is usually cumbersome to specify a family of (arbitrary) probability distributions $P(Y \mid X_1, \ldots, X_n)$ for a variable Y of a Bayesian network, and explain why causal independence models can help here. Also mention at least one other method that can help in determining $P(Y \mid X_1, \ldots, X_n)$. **[10]**

c. Suppose that the probability distribution $P(J \mid I_C, I_D)$, or equivalently Boolean function f, associated with the Bayesian network in Figure 1 represents a logical OR. Based on this assumption, compute the (marginal) probability distribution $P(J)$, i.e., the probabilities $P(j)$ and $P(\neg j)$. **[10]**

d. In principle any Boolean function could have been taken to model the interactions among the causes C and D giving rise to the effect J. Discuss the meaning of the causal independence model obtained by choosing a logical exclusive OR (XOR) as the Boolean function. Compute $P(J)$ (include the derivation of the result in your solution) for that case. **[15]**

[Total 50]

Question 2

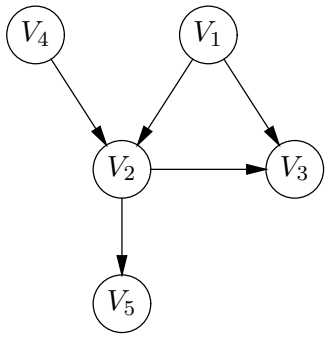

Figure 2: Bayesian network of Question 2.

Consider the Bayesian network $\mathcal{B} = (G, P)$ shown in Figure 2, where $G = (V(G), A(G))$ is the acyclic directed graph shown in the figure, and P is a probability distribution defined on the variables corresponding to the vertices $V(G) = \{V_1, V_2, V_3, V_4, V_5\}$. The following (local) probability distributions are defined for P:

$P(v_1) = 0.4$
$P(v_2 \mid v_1, v_4) = 0.5$ $\quad P(v_2 \mid \neg v_1, v_4) = 0.7$
$P(v_2 \mid v_1, \neg v_4) = 0.8$ $\quad P(v_2 \mid \neg v_1, \neg v_4) = 0.4$
$P(v_3 \mid v_1, v_2) = 0.6$ $\quad P(v_3 \mid \neg v_1, v_2) = 0.6$
$P(v_3 \mid v_1, \neg v_2) = 0.3$ $\quad P(v_3 \mid \neg v_1, \neg v_2) = 0.3$
$P(v_4) = 0.3$
$P(v_5 \mid v_2) = 0.2$ $\quad P(v_5 \mid \neg v_2) = 0.7$

a. Is the graph shown in Figure 2 a minimal or non-minimal directed I-map of P? Explain your answer. Construct the *undirected* graph that is an undirected I-map of P **[10]**

b. Give at least *5* conditional and unconditional independence relationships $\perp\!\!\!\perp$ which are represented in the graph G of the Bayesian network shown in Figure 2. **[15]**

c. The best structure of a Bayesian network $\mathcal{B} = (G, P)$ can be determined by computing $L_{\theta_G}(G) = \Pr(D \mid G, \theta_G)$, where Pr is the joint probability distribution on datasets D used for learning and Bayesian networks, where θ_G correspond to the probabilistic parameters P of $\mathcal{B}$. The measure $L_{\theta_G}(G)$ is called the *likelihood* of graph G. Assuming that the cases d in the database D are independent, finding the best structure amounts to optimising $L_{\theta_G}(G)$, or usually **[15]**

$$\log L_{\theta_G}(G) = l_{\theta_G}(G) = \sum_{d \in D} \log \Pr(d \mid G, \theta_G) = \sum_{d \in D} \log P(d),$$

i.e., we use the parameters P of $\mathcal{B}$ to compute the likelihood of each tuple d in database D.

Briefly discuss the validity of the assumptions behind this method. Would this measure give the same or different likelihoods for Markov equivalent Bayesian networks? Explain your answer.

d. Learning Bayesian network structure using a search-and-score method also requires the use of a search algorithm, typically greedy search. What are the limitations of greedy search in the structure learning context, and give at least one way by means of which this limitation can be circumvented. **[10]**

[Total 50]