3 hrs	10 / Nov / 2023	Venue	EXAMS OFFICE USE ONLY

University of the Witwatersrand, Johannesburg

Course or topic No(s)	COMS40XXA/70XXA
Course or topic name(s) Paper Number & title	Probabilistic Graphical Models
Examination to be held during the month(s) of	August 2023
Year of study	
Degrees/Diplomas for which this course is prescribed	BScHons (CS / BDA / CAM), MSc (AI / DS / CS / Robotics/ e-Science)
Faculties presenting candidates	Science
Internal examiner(s)	Prof. Ritesh Ajoodha x-76188
External examiner(s)	Prof. External Name (Ext Univ)
Special materials	Formula sheet and non-programmable calculator permitted
Time allowance	3 Hours
Instructions to candidates	Please answer all questions in this closed book test. A total of 100 marks are available, which corresponds to 100%. The test comprises of 20 pages.

Question 1 Multiple Choice Questions [10 Marks]

1. For each of the following MCQ questions, circle the correct answer label.

1.1 How does representing the joint distribution using the chain rule for probabilities make it intractable? [2]

(a) The distribution is computationally expensive to manipulate in memory.
(b) Probabilistic inference would take a long time.
(c) It is impossible to elicit priors for all the specified parameters from a human expert.
(d) A large amount of data is required because of fragmentation.
(e) All of the options above.

1.2. Suppose that you have a variable X with 2 dependencies (Y and Z). X, Y, and Z can all take one of 3 different values. Which of the following options specifies the length of the Tabular CPD for the variable X? [2]

(a) 8
(b) 9
(c) 27
(d) 30
(e) None of the above

1.3. Identify the assumption made in dynamic Bayesian networks that pertains to time from the options provided. [2]

(a) System is deterministic up to a point, then becomes completely random.
(b) Process being modeled remains statistically constant over time.
(c) Variable's rate of change determined by a time-based random number generator.
(d) Event probability at a time depends on bird-to-temperature ratio.

1.4. Which of the following statements is false when selecting a structure for a Bayesian network? [2]

(a) The structure should follow a causal ordering.
(b) The structure should be sparse
(c) The structure should contain as many edges as possible.
(d) The structure should be acyclic
(e) The structure should contain relevant variables

1.5. Which of the following statements is a limitation of variable elimination (VE)? [2]

(a) VE produces exact posterior distribution for any query.
(b) VE leads to a more compact factorization than other methods.
(c) VE is restricted to DAGs and not undirected graphs.
(d) VE can produce incorrect results if the network has continuous variables.
(e) VE is impractical for dense networks with many variables.

Question 2 Bayesian Networks [18 Marks]

2.1. Suppose you have a data set of 500 emails, where 100 are spam and 400 are not spam. You want to use the naïve Bayes model to classify a new email as either spam or not spam. If the word "free" appears in 80% of the spam emails and in 10% of the non-spam emails, and the word "money" appears in 50% of the spam emails and in 5% of the non-spam emails, what is the probability that an email containing both "free" and "money" is spam according to the naïve Bayes model? [5]

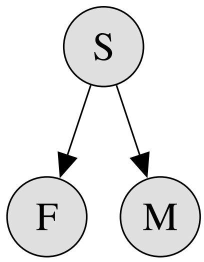
Answer:

- We have three variables:
  - E = whether the email is spam (s^0) or not spam (s^1).
  - F = whether the email contains the word "free" (f^1) or not (f^0)
  - M = whether the email contains the word "money" (m^1) or not (m^0)

  Using Bayes theorem:

$$P(S \mid F, M) = \frac{P(F, M \mid S) * P(S)}{P(F, M)}.$$

  where
  - $P(F, M \mid S)$ is the probability that an email is spam given that it contains both "free" and "money";
  - $P(S)$ is the prior probability that an email is spam (which is $\frac{100}{500} = 0.2$ in this case);
  - and $P(F, M)$ is the probability that an email contains both "free" and "money" regardless of whether it is spam or not.
- The naïve Bayes is provided by the following structural constraint:



- Therefore: $P(F, M \mid S) = P(F, M \mid S) = P(F \mid S) \times P(M \mid S) = 0.8 \times 0.5 = 0.4$
- Similarly: $P(F, M \mid \neg S) = P(F \mid \neg S) \times P(M \mid \neg S) = 0.1 \times 0.05 = 0.005$
- Then $P(F, M) = P(F, M \mid S) \times P(S) + P(F, M \mid \neg S) \times P(\neg S) = 0.4 \times 0.2 + 0.005 \times 0.8 = 0.081$
- Finally:

$$P(S \mid F, M) = \frac{0.4 \times 0.2}{0.081} = 0.9877.$$

- Therefore, the probability that an email containing both "free" and "money" is spam is approximately 0.9877.

2.2. Consider the Bayesian network $\mathcal{G}$ shown in Figure 1, and use it to answer the following questions.

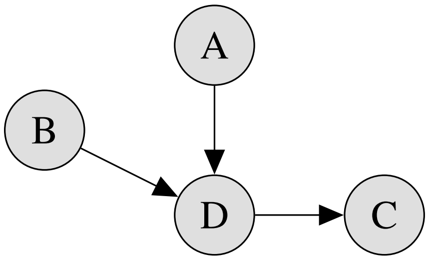


Figure 1: A simple Bayesian network with 4 variables

(a) Does the following set of independences that correspond to d-separation hold true in the context of the graph $\mathcal{G}$?

$$\mathcal{I}(\mathcal{G}) = \{(A, B \perp C \mid D) \ : \ \texttt{d-sep}_{\mathcal{G}}(A : C \mid D)\}$$

Explain your answer. [3]

Answer:

- Yes, $X = \{A, B\}$ and $Y = \{C\}$ are d-separated given D in $\mathcal{G}$. [1]
- There is a "Evidential trail" connecting C and B which is blocked when D is observed. There is also a "Causal trail" from B to C. [2]

(b) Does the following set of independences that correspond to d-separation hold true in the context of the graph $\mathcal{G}$?

$$\mathcal{I}(\mathcal{G}) = \{(A \perp B \mid C) \ : \ \texttt{d-sep}_{\mathcal{G}}(A : B \mid C)\}$$

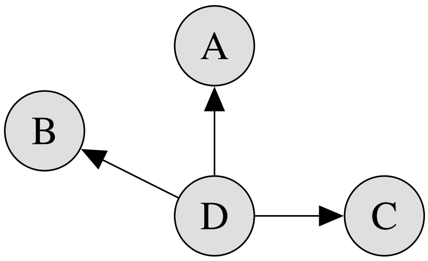
Explain your answer. [3]

Answer:

- No, A and B are not d-separated in $\mathcal{G}$ given D or C. [1]
- There is a "Common Effect Trail" (v-structure) connecting A to B which is activated when D (or any descendants of D is observed). Therefore, observing C renders A and B dependent. [2]

(c) Draw the Bayesian network resulting from the minimal I-map constructed using the independence properties observed in Figure 1, which has the variable ordering of D, B, A, and C. [4]

Answer:



(d) What does the term ‘I-equivalence’ mean in Bayesian network theory? Is the network in Figure 1 and the network in the answer to (c) above I-equivalent? [3]
Answer:

- The independencies that correspond to d-separation manifest in the structure of a Bayesian network, that is $\mathcal{I}(\mathcal{G})$. [1]
- When two Bayesian network structures, $\mathcal{G}_1$ and $\mathcal{G}_2$, have the same set of independencies that correspond to d-separation, $\mathcal{I}(\mathcal{G}_1) = \mathcal{I}(\mathcal{G}_2)$, then we say that $\mathcal{G}_1$ and $\mathcal{G}_2$ are I-Equivalent. [1]
- Yes, the network in Figure 1 and the answer to (c) are I-equivalent since they both encode the same independence assumptions. [1]

Question 3 Local Probability Models [16 Marks]

3.1. Consider the Tree-CPD, $\mathcal{T}$, shown in Figure 2, and use it to answer the following questions.

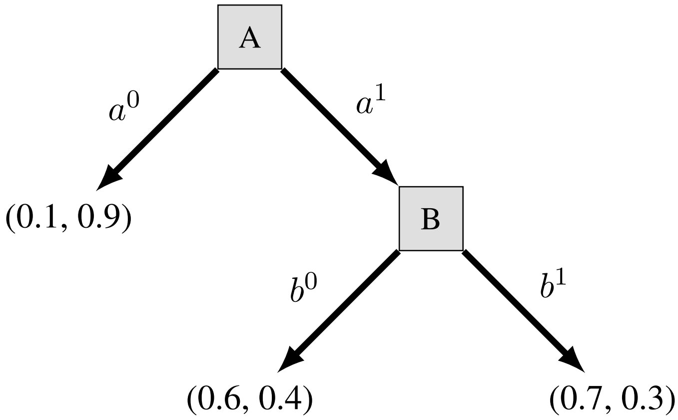


Figure 2: A Tree-CPD denoted $\mathcal{T}$

(a) Name one advantage and one disadvantage of using a Tabular-CPD over a Deterministic-CPD? [2]

Answer:

Advantages of Tabular-CPDs over Deterministic-CPDs:

- Capture more complex relationships between variables. [1]
- Easy to learn even when data is missing [1]
- Represent noisy data easier [1]

Disadvantages of Tabular CPDs over Deterministic CPDs:

- Require lots of memory [1]
- Probabilistic inference is harder [1]
- Difficult to scale [1]
- Sensitivity to sparsity [1]

(b) Draw $\mathcal{T}$ as a (normalised) Tabular-CPD which specifies the full joint distribution between the three variables A, B, and C. Round off two decimal places when specifying the joint distribution. [6]

Answer: $\frac{1}{2}$ marks for each correct probability. Two marks for listing all the correct assignments between 3 variables.

A	B	C	$P(A,B,C)$
a^0	b^0	c^0	0.03
a^0	b^0	c^1	0.23
a^0	b^1	c^0	0.03
a^0	b^1	c^1	0.23
a^1	b^0	c^0	0.15
a^1	b^0	c^1	0.10
a^1	b^1	c^0	0.18
a^1	b^1	c^1	0.08

(c) What are the context specific independence, $(\mathbf{X} \perp_c \mathbf{Y} \mid \mathbf{Z})$, that holds in $\mathcal{T}$? [2]
Answer:

- $(C \perp_c B \mid a^0)$ [2]

(d) List all of the rules, $\rho = \langle \mathbf{c}; p \rangle$, which together give you the Rule-CPD that hold in $\mathcal{T}$. [6]
Answer:

$\rho_1 = \langle a^0, c^0; 0.1 \rangle$ [1]
$\rho_2 = \langle a^0, c^1; 0.9 \rangle$ [1]
$\rho_3 = \langle a^1, b^0, c^0; 0.6 \rangle$ [1]
$\rho_4 = \langle a^1, b^0, c^1; 0.4 \rangle$ [1]
$\rho_5 = \langle a^1, b^1, c^0; 0.7 \rangle$ [1]
$\rho_6 = \langle a^1, b^1, c^1; 0.3 \rangle$ [1]

Question 4 Template-based Models [20 Marks]

4.1. Use the below parametrization of a Hidden Markov Model (HMM), λ, to answer the following questions.

i. Number of hidden states: 3.

ii. Number of observable symbols: 3.

iii. Initial state probabilities: $\pi_1 = 0.4, \pi_2 = 0.3, \pi_3 = 0.3$

iv. The transition probability matrix is: $\begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.4 & 0.5 & 0.1 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$

v. The observation model is provided by the following matrix: $\begin{pmatrix} 0.3 & 0.3 & 0.4 \\ 0.3 & 0.1 & 0.6 \\ 0.7 & 0.1 & 0.2 \end{pmatrix}$

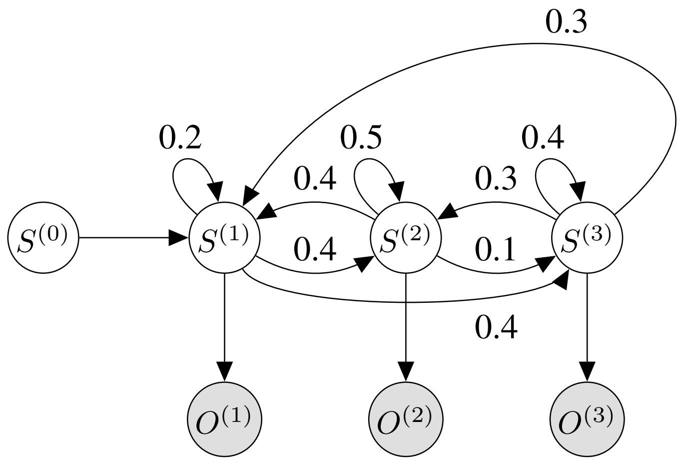
(a) Define the Markov assumption for temporal models. [2]

Answer: The Markov assumption states that the future is independent of the past given the present. The associated joint probability using the Markov assumption for a period 0:T is:

$$P(\mathcal{X}^{(0:T)}) = P(\mathcal{X}^{(0)}) \prod_{t=0}^{T-1} P(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(t)})$$

(b) Draw the structure of $\mathcal{H}$ with the associated probabilities labelled at the correct edges. Remember to label each node. [4]

Answer:



(c) Calculate the initial state probabilities for the forward variables $\alpha_1(1)$, $\alpha_1(2)$, and $\alpha_1(3)$ using λ for the sequence $O = \{0, 2\}$. [3]

Answer:

- $\alpha_1(1) = \pi_i \times b_{1,0} = 0.4 \times 0.3 = 0.12$
- $\alpha_1(2) = \pi_i \times b_{2,0} = 0.3 \times 0.3 = 0.09$
- $\alpha_1(3) = \pi_i \times b_{3,0} = 0.3 \times 0.7 = 0.21$

(d) Compute the forward variables $\alpha_2(1)$, $\alpha_2(2)$, and $\alpha_2(3)$ using λ for the first induction step using the initial state probabilities from the previous question. [6]

Answer:

- From the previous question: $\alpha_1(1) = 0.12$, $\alpha_1(2) = 0.09$, and $\alpha_1(3) = 0.21$.
- Two marks for each forward variable.
- Induction ($t = 2$)

$$\begin{aligned}
\alpha_2(1) &= \sum_{i=1}^{3} \alpha_1(i) a_{i,1} b_{1,2} \\
&= \alpha_1(1) a_{1,1} b_{1,1} + \alpha_1(2) a_{2,1} b_{1,1} + \alpha_1(3) a_{3,1} b_{1,1} \\
&= \Big(0.12 \times 0.2 \times 0.4\Big) + \Big(0.09 \times 0.4 \times 0.4\Big) + \Big(0.21 \times 0.3 \times 0.4\Big) \\
&= 0.0492
\end{aligned}$$

$$\begin{aligned}
\alpha_2(2) &= \sum_{i=1}^{3} \alpha_1(i) a_{i,2} b_{2,2} \\
&= \alpha_1(1) a_{1,2} b_{2,2} + \alpha_1(2) a_{2,2} b_{2,2} + \alpha_1(3) a_{3,2} b_{2,2} \\
&= \Big(0.12 \times 0.4 \times 0.6\Big) + \Big(0.09 \times 0.5 \times 0.6\Big) + \Big(0.21 \times 0.3 \times 0.6\Big) \\
&= 0.0936
\end{aligned}$$

$$\begin{aligned}
\alpha_2(3) &= \sum_{i=1}^{3} \alpha_1(i) a_{i,3} b_{3,2} \\
&= \alpha_1(1) a_{1,3} b_{3,2} + \alpha_1(2) a_{2,3} b_{3,2} + \alpha_1(3) a_{3,3} b_{3,2} \\
&= \Big(0.12 \times 0.4 \times 0.2\Big) + \Big(0.09 \times 0.1 \times 0.2\Big) + \Big(0.21 \times 0.4 \times 0.2\Big) \\
&= 0.0282
\end{aligned}$$

(e) Using the previous question, calulate the $P(O \mid \lambda)$ for the sequence $O = \{0, 2\}$. [2]

Answer:

$$\begin{aligned}
P(O \mid \lambda) &= P(0, 2 \mid \lambda) \\
&= \alpha_2(1) + \alpha_2(2) + \alpha_2(3) \\
&= 0.0492 + 0.0936 + 0.0282 \\
&= 0.171
\end{aligned}$$

(f) Unroll the plate model illustrated by Figure 3 and specify the variable name and scope as a function in each node. On the unrolled model indicate the shared parameters for all nodes. [3]

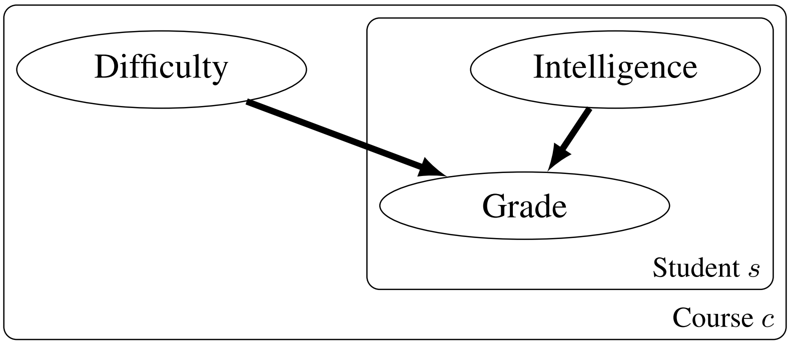
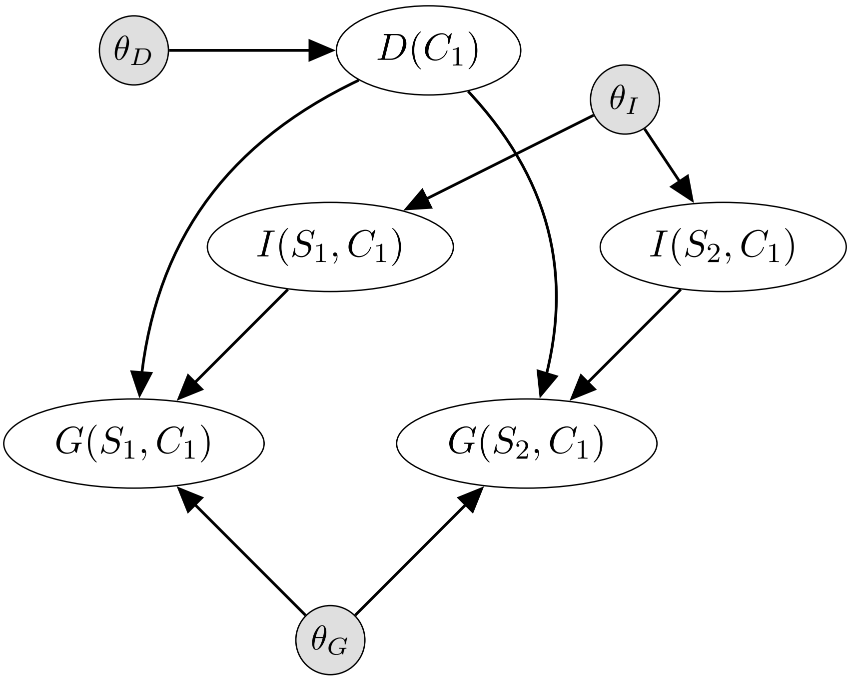


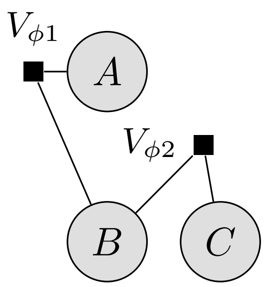
Figure 3: A plate model.

Answer:



Question 5 Undirected Graphical Models [16 Marks]

5.1. Use the below factor graph, $\mathcal{H}$, to answer the following questions.



$\phi_1(A, B)$

a^0	b^0	4
a^0	b^1	2
a^1	b^0	10
a^1	b^1	15

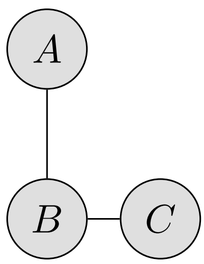
$\phi_1(B, C)$

b^0	c^0	8
b^0	c^1	2
b^1	c^0	7
b^1	c^1	10

Figure 4: A Markov network with associated factors

(a) Draw the Markov network structure associated with $\mathcal{H}$. [2]

Answer:



(b) Does the following set of independences that correspond to variable separation hold true in the context of the graph $\mathcal{H}$?

$$\mathcal{I}(\mathcal{H}) = \{(A \perp C \mid MB_{\mathcal{H}}(A))\},$$

where $MB_{\mathcal{H}}(A)$ is the Markov blanket of A. Explain your answer. [2]
Answer:

- Yes, $(A \perp C \mid MB_{\mathcal{H}}(A))$ is true. [1]
- There is a active trail between A and C in $\mathcal{H}$. Observing the Markov blanket of A, that is $MB_{\mathcal{H}}(A) = \{B\}$, blocks the influence between A and C which makes them independent of each other. [1]

(c) Write the joint distribution of $\mathcal{H}$ using corresponding maximal clique potentials, Φ. [2]

Answer: The joint distribution for $\mathcal{H}$ is

$$P(A, B, C) = \frac{1}{Z}\Big(\phi_1(A, B) \times \phi_2(B, C)\Big),$$

where

$$Z = \sum_{A,B,C}\Big(\phi_1(A, B) \times \phi_2(B, C)\Big)$$

(d) Compute the factor product $\psi_3(A, B, C) = \phi_1(A, B) \times \phi_2(B, C)$. [4]

Answer: Half a mark for each correct probability and assignment in $\psi_3(A, B, C)$.

a^0	b^0	4
a^0	b^1	2
a^1	b^0	10
a^1	b^1	15

$\phi(A, B)$

$\times$

b^0	c^0	8
b^0	c^1	2
b^1	c^0	7
b^1	c^1	10

$\phi(B, C)$

$=$

a^0	b^0	c^0	32
a^0	b^0	c^1	8
a^0	b^1	c^0	14
a^0	b^1	c^1	20
a^1	b^0	c^0	80
a^1	b^0	c^1	20
a^1	b^1	c^0	105
a^1	b^1	c^1	150

$\psi(A, B, C)$

(e) Calculate the value of the partition function. [2]
Answer:

$$\begin{aligned} Z &= \sum A, B, C\Big(\phi_1(A, B), \phi_2(B, C)\Big) \\ &= \sum A, B, C\Big(\psi_3(A, B, C)\Big) \\ &= 429 \end{aligned}$$

(f) Calculate $\psi[B = b^1](A, C)$. [2]

Answer: Half a mark for each correct probability and assignment in $\psi[B = b^1](A, C)$.

a^0	b^0	c^0	32
a^0	b^0	c^1	8
a^0	b^1	c^0	14
a^0	b^1	c^1	20
a^1	b^0	c^0	80
a^1	b^0	c^1	20
a^1	b^1	c^0	105
a^1	b^1	c^1	150

$\psi(A, B, C)$

$\rightarrow$

a^0	b^1	c^0	14
a^0	b^1	c^1	20
a^1	b^1	c^0	105
a^1	b^1	c^1	150

$\psi[B = b^1](A, C)$

(g) Calculate $P(b^1)$. Round off two decimal places. [2]
Answer:

$$\begin{aligned} P(b^1) &= \frac{1}{Z}\left(\sum_{A,C} \psi_3[B = b^1](A, C)\right) \\ &= \frac{1}{429}\left(289\right) \\ &= 0.67 \end{aligned}$$

Question 6 Exact and Approximate Inference [20 Marks]

6.1. Use the Bayesian network, $\mathcal{B}$, illustrated in Figure 5 to answer the following questions.

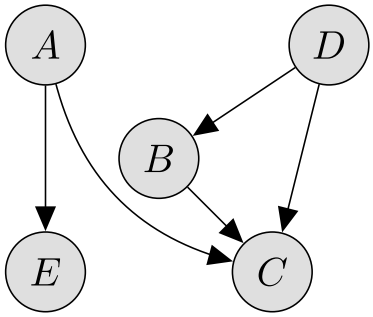


Figure 5: A Bayesian network

(a) Factorise the joint distribution using the chain rule for Bayesian networks in a way that corresponds to $\mathcal{B}$. [2]
Answer:

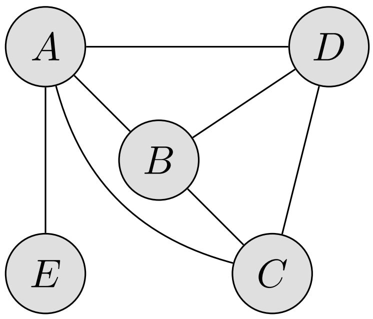
$$P(A, B, C, D, E) = P(A)P(E \mid A)P(D)P(B \mid D)P(C \mid A, B, D)$$

(b) Using variable elimination algorithm compute $P(A)$. Show all the required steps of the algorithm using the elimination ordering: $\prec = \{C, B, D, E\}$. [10]
Answer: One mark for each step.

$$\begin{aligned}
P(A) = &\sum_{D,E,B,C} P(A) \cdot P(E \mid A) \cdot P(D) \cdot P(B \mid D) \cdot P(C \mid A, B, D) \\
&= \phi_A(A) \sum_E \phi_E(A, E) \sum_D \phi_D(D) \sum_B \phi_B(B, D) \sum_C \phi_C(A, B, C, D) \\
&= \phi_A(A) \sum_E \phi_E(A, E) \sum_D \phi_D(D) \sum_B \phi_B(B, D)\tau_1(A, B, D) \\
&= \phi_A(A) \sum_E \phi_E(A, E) \sum_D \phi_D(D) \sum_B \psi_1(A, B, D) \\
&= \phi_A(A) \sum_E \phi_E(A, E) \sum_D \phi_D(D)\tau_2(A, D) \\
&= \phi_A(A) \sum_E \phi_E(A, E) \sum_D \psi_2(A, D) \\
&= \phi_A(A) \sum_E \phi_E(A, E)\tau_3(A) \\
&= \phi_A(A) \sum_E \psi_3(A, E) \\
&= \phi_A(A)\tau_4(A) \\
&= \psi_4(A)
\end{aligned}$$

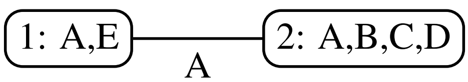
(c) Draw the induced graph $\mathcal{I}_{\mathcal{B},\prec}$ which results in this ordering of variable elimination. [2]

Answer:



(d) Draw the clique tree for $\mathcal{I}_{\mathcal{B},\prec}$ [2]

Answer:



(e) What does it mean for a cluster graph to have the running intersection property? Does the clique tree form the previous question satisfy the running intersection property? [2]

Answer:

- For a cluster graph to contain the running intersection property then for any any factor X, the set of clusters and sepsets containing X should form a tree. [1]

- Yes, the clique tree from the previous question does satisfy the running intersection property. [1]
- Variables B,C,D,E are all contained within a node and is therefore a tree. Variable A is contained in both nodes and is a sepset connecting the two nodes which is also a tree.

(f) Suppose that you are in the middle of a message-passing procedure using the clutser graph $\mathcal{C}$ as shown in Figure 6. The first message is $\delta_{1\to 2}(B)$, what would be the message $\delta_{2\to 3}(C)$? [2]

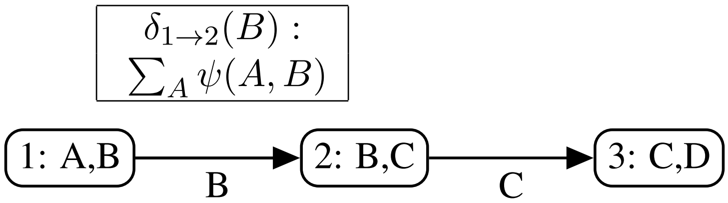


Figure 6: A cluster graph $\mathcal{C}$

Answer:

$$\delta_{2\to 3}(C) = \sum_{B} \psi(B, C) \times \delta_{1\to 2}$$

END OF TEST

Working out

Working out

Probabilistic Graphical Models
Formula Sheet

Probability Theory

Chain Rule for Probabilities:
$P(X_1, \ldots, X_n) = P(X_1) \cdots P(X_n | X_1, X_{n-1})$

Bayes Rule:
$P(\alpha \mid \beta) = \frac{P(\beta \mid \alpha)}{P(\alpha)P(\beta)}$

Probability Density Function: $p : \mathbb{R} \to \mathbb{R}$ is a probability density function (PDF) for $\mathcal{X}$ if it is a non-negative integrable function such that:
$\int_{Val(X)} p(x)\, dx = 1.$

Uniform Distribution: $X \sim \text{Unif}[a, b]$ if it has the PDF:

$$p(x) = \begin{cases} \frac{1}{b-a} & b \geq x \geq a \\ 0 & \text{otherwise.} \end{cases}$$

Gaussian Distribution: X has a Gaussian distribution: $X \sim \mathcal{N}(\mu; \sigma^2)$ if it has the PDF:
$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$

Joint Density Function: Let P be a joint distribution over $X_1, \ldots, X_n$. A function $p(x_1, \ldots, x_n)$ is a joint density function of $X_1, \ldots, X_n$ if:

1. $p(x_1, \ldots, x_n) \geq 0 \; \forall x_1, \ldots, x_n \in X_1, \ldots, X_n$.
2. p is integratable.
3. For any choice of $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$:

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} p(x_1, \ldots, x_n)\, dx_1 \ldots dx_n$$

Conditional Density Function: Suppose you would like to condition over the event:
$x - \epsilon \leq X \leq x + \epsilon$. Then
$P(Y \mid x) = \lim_{\epsilon \to 0} P(Y \mid x - \epsilon \leq X \leq x + \epsilon)$. If there is a continuous joint density function p(x,y) then
$= P(a \leq Y \leq b \mid x - \epsilon \leq X \leq x + \epsilon)$
$= \frac{P(a \leq Y \leq b, x-\epsilon \leq X \leq x+\epsilon)}{P(x-\epsilon \leq X \leq x+\epsilon)} = \frac{\int_a^b \int_{x-\epsilon}^{x+\epsilon} p(x', y)\, dy dx'}{\int_{x-\epsilon}^{x+\epsilon} p(x')\, dx'}$

Expectation of X under P is:
$\mathbb{E}_P[X] = \sum_x x.P(x).$

Expectation if X is Continuous:
$\mathbb{E}_P[X] = \int x.p(x)\, dx.$

Linearity of Expectation:
$\mathbb{E}_P[X + Y] = \mathbb{E}_P[X] + \mathbb{E}_P[Y].$

Conditional Expectation:
$\mathbb{E}_P[X \mid \mathbf{y}] = \sum_x x.P(x \mid \mathbf{y}).$

Variance of X:
$\mathbb{V}ar_P[X] = \mathbb{E}_P[(X - \mathbb{E}_P[X])^2].$

Standard Deviation:
$\sigma_X = \sqrt{\mathbb{V}ar_P[X]}.$

Expectation and Variance of Gaussian distribution $X \sim \mathcal{N}(\mu; \sigma^2)$, then $\mathbb{E}[X] = \mu$ and $\mathbb{V}ar[X] = \sigma^2$.

Graph Theory

A **Graph** is a data structure $\mathcal{K} = (\mathcal{X}, \mathcal{E})$ consisting of a set of nodes, denoted $\mathcal{X} = X_1, \ldots, X_n$, and edges, denoted $\mathcal{E}$.

Induced Subgraph: Let $\mathcal{K} = (\mathcal{X}, \mathcal{E})$, and $\mathbf{X} \in \mathcal{X}$, then an induced subgraph, denoted $\mathcal{K}[\mathbf{X}]$ is a graph $(\mathbf{X}, \mathcal{E}')$ where $\mathcal{E}'$ are all the edges $X \rightleftharpoons Y \in \mathcal{E}'$ such that $X, Y \in \mathbf{X}$.

Complete Graph (Clique): A subgraph over $\mathbf{X}$ is complete if every two nodes in $\mathbf{X}$ are connected by some edge. The set $\mathbf{X}$ is called a clique. A clique $\mathbf{X}$ is maximal if for any superset of nodes $\mathbf{Y} \supset \mathbf{X}$, $\mathbf{Y}$ is not a clique.

Upward Closure: A subset of nodes $\mathbf{X} \in \mathcal{X}$ is upwardly closed in $\mathcal{K}$ if, for any $\mathbf{X} \in \mathcal{X}$, we have that the $\text{Boundary}_X \subset \mathbf{X}$. We define upward closure of $\mathbf{X}$ to be the minimally upward closed subset $\mathbf{Y}$ that contains $\mathbf{X}$.

Topological ordering: An ordering of the nodes $X_1, \ldots, X_n$ is a topological ordering if when we have $(X_i \to X_j) \in \mathcal{E}$, then $i < j$.

Chordal Graph: Let $X_1 - X_2 - \ldots - X_k - X_1$ be a loop in a graph. A chord in a loop is an edge connecting X_i and X_j for two nonconsecutive nodes X_i, X_j. An undirected graph $\mathcal{H}$ is said to be chordal if and loop $X_1 - X_2 - \ldots - X_k - X_1$ for $k > 4$ has a chord. A directed graph $\mathcal{K}$ is said to be chordal if its underlying undirected graph is chordal.

Bayesian Networks

Naïve Bayes:
$P(C, X_1, \ldots, X_n) = P(C) \prod_{i=0}^n P(X_i \mid C)$

Bayesian Network:
$P(X_1, \ldots, X_n) = \prod_{i=1}^n P(X_i \mid \mathtt{Pa}_{X_i}^{\mathcal{G}})$

Deterministic CPD: $f : Val(Pa_X) \mapsto Val(X)$ s.t.:

$$P(x \mid pa_x) = \begin{cases} 1 & \text{if } x = f(pa_X) \\ 0 & \text{if } x \texttt{ otherwise} \end{cases}$$

Time Granularity Assumption:
$P(\mathcal{X}^{(0:T)}) = P(\mathcal{X}^{(0)}) \prod_{t=0}^{T-1} P(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(0:t)})$

Markov Assumption:
$P(\mathcal{X}^{(0:T)}) = P(\mathcal{X}^{(0)}) \prod_{t=0}^{T-1} P(\mathcal{X}^{(t+1)} \mid \mathcal{X}^{(t)})$

Time Invariance Assumption:
$P(\mathcal{X}^{(t+1)} = \xi' \mid \mathcal{X}^{(t)} = \xi) = P(\mathcal{X}' = \xi' \mid \mathcal{X} = \xi)$

Two-TBN:
$P(\mathcal{X}' \mid \mathcal{X}) = P(\mathcal{X}' \mid \mathcal{X}_I) = \prod_{i=1}^n P(X_i' \mid Pa_{X_i'})$

Linear Dynamical Systems:
$P(\mathbf{X}^{(t)} \mid \mathbf{X}^{(t-1)}) = \mathcal{N}(A\mathbf{X}^{(t-1)}; Q)$
$P(O^{(t)} \mid \mathbf{X}^{(t)}) = \mathcal{N}(H\mathbf{X}^{(t)}; R)$

Gibbs Distribution: A distribution P_Φ is a Gibbs distribution parameterised by a set of factors $\Phi = \{\phi_1(\mathbf{D_1}), \ldots, \phi_K(\mathbf{D_k})\}$ if it is defined as:
$P_\Phi(X_1, \ldots, X_n) = \frac{1}{Z} \tilde{P}_\Phi(X_1, \ldots, X_n)$

Inference

Inference:
$P(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \frac{P(\mathbf{Y}, \mathbf{e})}{P(\mathbf{e})} = \frac{\sum_w P(\mathbf{y}, \mathbf{e}, \mathbf{w})}{\sum_{y,w} P(\mathbf{e})}$

Sum-Product Message Passing:
$\delta_{i \to j} = \sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} (\psi_i \times \prod_{k \in (Nb_i - \{j\})} \delta_{k \to i})$

Tree Calibration:
$\sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i(\mathbf{C}_i) = \sum_{\mathbf{C}_j - \mathbf{S}_{i,j}} \beta_j(\mathbf{C}_j)$

Graph Calibration:
$\sum_{\mathbf{C}_i - \mathbf{S}_{i,j}} \beta_i = \sum_{\mathbf{C}_j - \mathbf{S}_{i,j}} \beta_j$

MAP:
$\text{MAP}(\mathbf{Y} = \mathbf{y} \mid \mathbf{E} = \mathbf{e})$
$= \text{argmax}_y P(\mathbf{Y} = \mathbf{y} \mid \mathbf{E} = \mathbf{e})$

Convergence Bound:
$\hat{\mathbb{E}}_{\mathcal{D}}(f) = \frac{1}{M} \sum_{m=1}^M f(\xi[m]).$

Hoeffding Bound:
$P_{\mathcal{D}}(\hat{P}(\mathbf{y}) \notin [P(\mathbf{y}) - \epsilon, P(\mathbf{y}) + \epsilon]) \leq 2e^{-2M\epsilon^2}$

Chernoff Bound:
$P_{\mathcal{D}}(\hat{P}(\mathbf{y}) \notin [P(\mathbf{y})(\pm\epsilon)]) \leq 2e^{-MP(\mathbf{y})\epsilon^2/3}$
$M \geq 3 \frac{\ln(2/\delta)}{P(\mathbf{y})\epsilon^2}.$

Likelihood Weighting:

$$\hat{P}_{\mathcal{D}}(\mathbf{y} \mid \mathbf{e}) = \frac{\sum_{m=1}^{M} w[m]\mathbb{1}\{\mathbf{y}[m]=\mathbf{y}\}}{\sum_{m=1}^{M} w[m]}.$$

MCMC Sampling:

$$P^{(t+1)}(\mathbf{X}^{(t+1)} = \mathbf{x'}) = \sum_{\mathbf{x}\in Val(\mathbf{X})} P^{(t)}(\mathbf{X}^{(t)} = \mathbf{x})\mathcal{T}(\mathbf{x} \to \mathbf{x'})$$

Stationary Distribution:

$$\pi(\mathbf{X} = \mathbf{x'}) = \sum_{\mathbf{x}\in Val(\mathbf{X})} \pi(\mathbf{X} = \mathbf{x})\mathcal{T}(\mathbf{x} \to \mathbf{x'})$$

Detailed Balance Equation:

$$\pi(x)\mathcal{T}(x \to x') = \pi(x')\mathcal{T}(x' \to x)$$

Acceptance Probability:

$$\mathcal{A}(x \to x') = \min[1, \frac{\pi(x')\mathcal{T}^Q(x'\to x)}{\pi(x)\mathcal{T}^Q(x\to x')}].$$

Metropolis-Hastings Acceptance Probability:

$$\mathcal{A}(x_{-i}, x_i \to x_{-i}, x'_i) = \min[1, \frac{P_\Phi(x'_i,x_{-i})\mathcal{T}_i^Q(x_{-i},x'_i\to x_{-i},x_i)}{P_\Phi(x_i,x_{-i})\mathcal{T}_i^Q(x_{-i},x_i\to x_{-i},x'_i)}].$$

Learning

Relative Entropy:

$$\mathbb{D}(P^* \,||\, \tilde{P}) = \mathbb{E}_{\xi\sim P^*}[\log(\frac{P^*(\xi)}{\tilde{P}(\xi)})],$$

Negative Empirical Log-loss:

$$\log P(\mathcal{D} : \mathcal{M}) = \sum_{m=1}^{M} \log P(\xi[m] : \mathcal{M}).$$

Bayesian Parameter Estimation:

$$P(\theta \,| x[1], \ldots, x[M]) = \frac{P(x[1],\ldots,x[M] \;|\theta)P(\theta)}{P(x[1],\ldots,x[M])}$$

Expected Sufficient Statistics:

$$\bar{M}_\theta[\mathbf{y}] = \sum_{m=1}^{M} \sum_{\mathbf{h}[m]\in Val(\mathbf{H}[m])} Q(\mathbf{h}[m])\mathbb{1}\{\xi[m]\langle\mathbf{Y}\rangle = \mathbf{y}\}$$

Maximisation of Expected Parameter:

$$\tilde{\theta}_{d^1|c^0} = \frac{\bar{M}_\theta[d^1,c^0]}{\bar{M}_\theta[c^0]}$$

Bayesian Clustering:

$$\bar{M}_\theta[c] = \frac{\bar{M}_\theta[c]}{M}$$

$$\bar{M}_\theta[x_i \mid c] = \frac{\bar{M}_\theta[x_i,c]}{\bar{M}_\theta[c]}$$

K-means Clustering:

$$c[m] = \mathrm{argmax}_c\, P(c \mid x[m], \theta^t)$$

Hypothesis Testing:

$$d_{\mathbb{I}}(\mathcal{D}) = \sum_{x,y} \frac{M[x,y]}{M} \log \frac{M[x,y]/M}{M[x]/M\cdot M[y]/M}$$

$$\mathrm{R}_{d,t}(\mathcal{D}) \begin{cases}\text{Accept} & \text{if } d(\mathcal{D}) \le t\\ \text{Reject} & \text{if } d(\mathcal{D}) > t\end{cases}$$

$$\text{p-value}(t) = P(\{\mathcal{D} : d(\mathcal{D}) > t\} \mid H_0, M)$$

Likelihood:

$$\mathbb{I}_{\hat{P}_{\mathcal{D}}}(X_i; Pa^{\mathcal{G}}_{X_i}) = \sum_{\mathbf{u}_i} \sum_{\mathbf{x}_i} \hat{P}(x_i, \mathbf{u}_i) \log \frac{\hat{P}(x_i,\mathbf{u}_i)}{\hat{P}(x_i)\hat{P}(\mathbf{u}_i)}$$

Entropy:

$$\mathbb{H}_{\hat{P}_{\mathcal{D}}}(X_i) = \sum_{x_i} \hat{P}(x_i) \log \frac{1}{\hat{P}(x_i)}$$

Bayesian Structure Learning:

$$P(\mathcal{G} \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \mathcal{G})P(\mathcal{G})}{P(\mathcal{D})}$$

$$\mathrm{score}_B(\mathcal{G} : \mathcal{D}) = \log P(\mathcal{D} \mid \mathcal{G}) + \log P(\mathcal{G})$$

$$P(\mathcal{D} \mid \mathcal{G}) = \int_{\Theta_{\mathcal{G}}} P(\mathcal{D} \mid \boldsymbol{\theta}_{\mathcal{G}}, \mathcal{G})P(\boldsymbol{\theta}_{\mathcal{G}} \mid \mathcal{G})d\boldsymbol{\theta}_{\mathcal{G}}$$

Marginal Likelihood for Binomials:

$$P(x[1], \ldots, x[M]) = P(x[1]) \cdot \ldots \cdot P(x[m] \mid x[1], \ldots, x[M-1])$$

Marginal Likelihood for Multinomials:

$$P(x[1], \ldots, x[M]) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+M)} \cdot \prod_{i=1}^{k} \frac{\Gamma(\alpha_i+M[x^i])}{\Gamma(\alpha_i)}$$

Bayesian Score:

$$P(\mathcal{D} \mid \mathcal{G}) = \prod_i \prod_{\mathbf{u}_i\in Val(Pa^{\mathcal{G}}_{X_i})} \frac{\Gamma(\alpha^{\mathcal{G}}_{X_i|\mathbf{u}_i})}{\Gamma(\alpha^{\mathcal{G}}_{X_i|\mathbf{u}_i}+M[\mathbf{u}_i])} \cdot \prod_{\mathbf{x}_i^j\in Val(X_i)} \left[\frac{\Gamma(\alpha^{\mathcal{G}}_{x_i^j|\mathbf{u}_i}+M[x_i^j,\mathbf{u}_i])}{\Gamma(\alpha^{\mathcal{G}}_{x_i^j|\mathbf{u}_i})}\right]$$

BIC Score:

$$\mathrm{score}_{BIC}(\mathcal{G} : \mathcal{D}) = M\sum_{i=1}^{n} \mathbb{I}_{\hat{P}_{\mathcal{D}}}(X_i; Pa^{\mathcal{G}}_{X_i}) - \frac{\log M}{2}\dim[\mathcal{G}]$$

Decomposibility:

$$\mathrm{score}(\mathcal{G} : \mathcal{D}) = \sum_i \mathrm{FamScore}(X_i \mid Pa^{\mathcal{G}}_{X_i} : \mathcal{D})$$

Tree weight:

$$w_{i\to j} = \mathrm{FamScore}(X_i \mid X_j : \mathcal{D}) - \mathrm{FamScore}(X_i : \mathcal{D})$$

Learning Graphs:

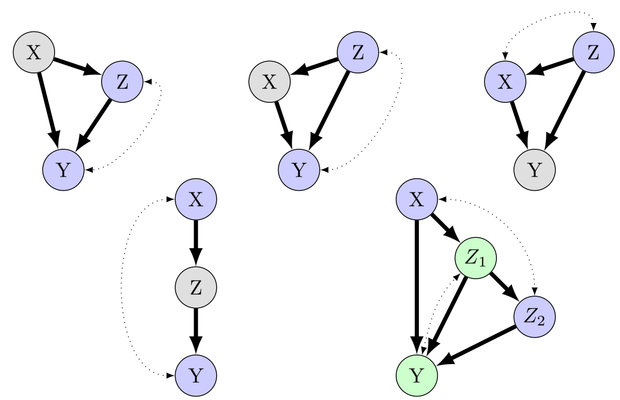
$$\mathcal{G}^* = \mathrm{argmax}_{\mathcal{G}\in\boldsymbol{\mathcal{G}}}\, \mathrm{score}(\mathcal{G} : \mathcal{D})$$

Causality

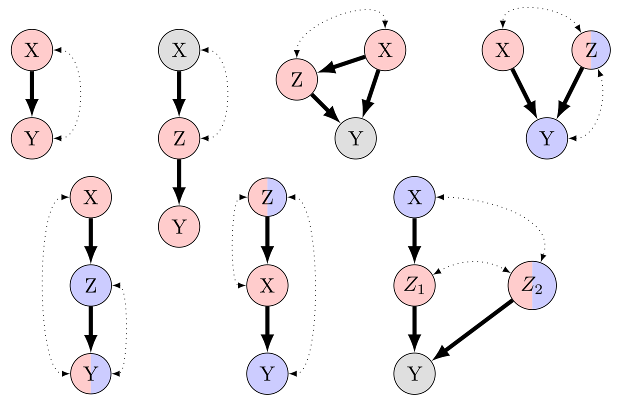
Intervention Query:

$$P_{\mathcal{C}}(\mathbf{Y} \mid do(z), \mathbf{x}) = P_{\mathcal{C}_{\mathbf{Z=z}}}(\mathbf{Y} \mid \mathbf{x})$$

Identifiable when $P(Y \mid do(X))$:



Not Identifiable when $P(Y \mid do(X))$:



Learning with Intervention Data:

$$P(\xi \mid do(\mathbf{Z}{:=}\,\mathbf{z}), \mathcal{C}) = \prod_{X_i\notin\mathbf{Z}} P(x_i \mid \mathbf{u}_i)$$

Sufficient Statistics (Intervention Data):

$$M[x_i; \mathbf{u}_i] = \sum_{m:X_i\notin\mathbf{Z}[m]} \mathbb{1}\{X_i[m] = x_i, Pa_{X_i}[m] = \mathbf{u}_i\}$$

Likelihood of Data (Intervention):

$$L(\mathcal{C} : \mathcal{D}) = \prod_{i=1}^{n} \prod_{x_i\in Val(X_i),\mathbf{u}_i\in Val(Pa_{X_i})} \theta_{x_i|\mathbf{u}_i}^{M[x_i;\mathbf{u}_i]}$$

Decision Theory

Expected Utility:

$$\mathrm{EU}[D[a]] = \sum_{\mathbf{x}} P(\mathbf{x} \mid a)U(\mathbf{x}, a)$$

Maximum Expected Utility:

$$a^* = \mathrm{argmax}_a\, \mathrm{EU}[D[a]] = \mathrm{argmax}_a \sum_{\mathbf{x}} P(\mathbf{x} \mid a)U(\mathbf{x}, a)$$

Expected Utility with Information:

$$\mathrm{EU}[D[\delta_A]] = \sum_{\mathbf{x},a} P_{\delta_A}(\mathbf{x}, a)U(\mathbf{x}, a)$$

Maximal Expected Utility (MEU) Strategy:

$$\mathrm{argmax}_{\delta_{D_1},\ldots,\delta_{D_k}}\, \mathrm{EU}[\mathcal{I}[\delta_{D_1}, \ldots, \delta_{D_k}]]$$

Value of Information:

$$\mathrm{VPI}(A \mid X) := \mathrm{MEU}(D_{X\to A}) - \mathrm{MEU}(D)$$