IITM-CS6730 : Probabilistic Graphical Models | Release Date: Jan 25, 2020

Assignment 1 | Due Date : Feb 3, 23:59

Roll No: EE16B026 | Name: Muqeeth

Collaborators (if any):Gowri Shankar(EE16B021)

References (if any):

- Use LaTeX to write-up your solutions (in the solution blocks of the source LaTeX file of this assignment), and submit the resulting single pdf file at GradeScope by the due date. (Note: As always, **no late submissions** will be allowed, other than one-day late submission with 10% penalty! Within GradeScope, indicate the page number where your solution to each question starts, else we won't be able to grade it! You can join GradeScope using course entry code MG8P8R).
- Collaboration is encouraged, but all write-ups must be done individually and independently, and mention your collaborator(s) if any. Same rules apply for codes written for any programming assignments (i.e., write your own code; we will run plagiarism checks on codes).
- If you have referred a book or any other online material for obtaining a solution, please cite the source. Again don't copy the source *as is* - you may use the source to understand the solution, but write-up the solution in your own words.

1. (7 points) [PROBABILITY RAPID-FIRE (via David Blei's course)]

   (a) (1 point) Consider a probability density $p(x \mid \mu, \sigma^2) = \mathcal{N}(x \mid \mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$. (i) For some $x \in \mathbb{R}$, can $p(x) < 0$? (ii) For some $x \in \mathbb{R}$, can $p(x) > 1$?
   sol:
   (i) No.
   (ii) Yes

   (b) (1 point) In the exponential family of distributions, $p(x|\theta) = h(x)\exp\{\eta(\theta)^\top t(x) - a(\theta)\}$, for $x \in \mathbb{R}^n, \theta \in \mathbb{R}^d$, auxiliary measure $h(x) : \mathbb{R}^n \to \mathbb{R}$, natural parameter function $\eta(\theta) : \mathbb{R}^d \to \mathbb{R}^p$, sufficient statistics $t(x) : \mathbb{R}^n \to \mathbb{R}^p$, and $a(\theta) : \mathbb{R}^d \to \mathbb{R}$. What must $a(\theta)$ be for $p(x|\theta)$ to be a valid probability distribution? Why?
   sol:
   For $p(x|\theta)$ to be a valid probability distribution, we need $\int_x p(x|\theta)dx$ to be 1.

$$\int_x p(x|\theta)dx = \exp(-a(\theta)) \int_x h(x)\exp(\eta(\theta)^\top t(x)) = 1$$

$$\exp(a(\theta)) = \int_x h(x)\exp(\eta(\theta)^\top t(x))$$

$$a(\theta) = \ln(\int_x h(x)\exp(\eta(\theta)^\top t(x)))$$

(c) (3 points) The exponential family defined above can be used to represent many distributions. Show how to represent the following: Bernoulli, Poisson, and univariate Gaussian.
sol:
Bernoulli distribution

$$
\begin{aligned}
p(x|\theta) &= \theta^x(1-\theta)^{(1-x)} \\
&= \exp(x\ln(\theta) + (1-x)\ln(1-\theta)) \\
&= \exp(x\ln(\frac{\theta}{1-\theta}) + \ln(1-\theta))
\end{aligned}
$$

Bernoulli distribution is an exponential family distribution with:

$$h(x) = 1$$

$$t(x) = x$$

$$\eta(\theta) = \ln(\frac{\theta}{1-\theta})$$

$$a(\theta) = -\ln(1-\theta)$$

Poisson distribution

$$
\begin{aligned}
p(x|\lambda) &= \frac{\lambda^x \exp(-\lambda)}{x!} \\
&= \frac{1}{x!}\exp(x\ln(\lambda) - \lambda)
\end{aligned}
$$

poisson distribution is an exponential family distribution with:

$$h(x) = \frac{1}{x!}$$

$$t(x) = x$$

$$\theta = \lambda$$

$$\eta(\theta) = \ln(\theta)$$

$$a(\theta) = \theta$$

univariate gaussian distribution

$$
\begin{aligned}
p(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2) \\
&= \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2 - \ln(\mu))
\end{aligned}
$$

univariate gaussian distribution is an exponential family distribution with:

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$$

$$t(x) = \begin{pmatrix} x^2 \\ x \end{pmatrix}$$

$$\eta(\theta) = \begin{pmatrix} -\frac{1}{2\theta_2} \\ \frac{\theta_1}{\theta_2} \end{pmatrix}$$

$$a(\theta) = \frac{\theta_1^2}{2\theta_2^2} + \ln(\theta_1)$$

(d) (2 points) You have a jar of $1,000$ coins. 999 are fair coins, and the remaining coin will always land heads. You take a single coin out of the jar and flip it 10 times in a row, all of which land heads. What is the probability your next toss with the same coin will land heads? Explain your answer. How would you call this probability in Bayesian jargon?
sol:
Let A_i be the event of tossing the coin i^{th} time.
Given that first 10 tosses resulted in head i.e $A_i = H$ for i = 1,2,..10, we need find the probability of 11^{th} i.e $A_{11} = H$ toss is head. This is called posterior probability in bayesian terms.
Let X be the event of picking fair coin and Y be the event of picking the coin which lands always heads.

$$\begin{aligned} P(A_{11} = H|A_i = H, i = 1,2..10) &= \frac{P(A_i = H, i = 1,2,..11)}{P(A_i = H, i = 1,2...10)} \\ &= \frac{P(A_i = H, i = 1,2,..11|X)P(X) + P(A_i = H, i = 1,2,...11|Y)P(Y)}{P(A_i = H, i = 1,2,..10|X)P(X) + P(A_i = H, i = 1,2,..10|Y)P(Y)} \\ &= \frac{(\frac{1}{2^{11}})(\frac{999}{1000}) + (1)(\frac{1}{1000})}{(\frac{1}{2^{10}})(\frac{999}{1000}) + (1)(\frac{1}{1000}}) \\ &= 0.753 \end{aligned}$$

2. (7 points) [CONDITIONAL FREEDOM] Let P be a joint distribution defined over n random variables $\mathscr{X} = X_1, X_2, \ldots, X_n$ Let X, Y, Z, W be sets of random variables, each being a subset of $\mathscr{X}$.

(a) (1 point) Prove from first principles (definition of conditional independence in KF book) that $(X \perp Y \mid Y)$?
sol:
if $P(Y) = 0$ then $X \perp Y$
else

To prove $P(X, Y \mid Y) = P(X \mid Y)P(Y \mid Y) = P(X \mid Y)$

$$\begin{aligned} P(X, Y \mid Y) &= \frac{P(X, Y, Y)}{P(Y)} \\ &= \frac{P(X, Y)}{P(Y)} \\ &= P(X \mid Y) \end{aligned}$$

Therefore $(X \perp Y \mid Y)$ is true.

(b) (1 point) Are the statements $(X \perp Y, Z \mid Z)$ and $(X \perp Y \mid Z)$ equivalent? Explain your answer.
if P(Z) = 0 then statements are equivalent.
else:

$$P(X, Y \mid Z) = P(X, Y, Z \mid Z)$$

Given $X \perp Y, Z \mid Z$, we have $P(X, Y, Z \mid Z) = P(X \mid Z)P(Y, Z \mid Z)$

$$\begin{aligned} P(X, Y \mid Z) &= P(X, Y, Z \mid Z) \\ &= P(X \mid Z)P(Y, Z \mid Z) \\ &= P(X \mid Z)P(Y \mid Z) \end{aligned}$$

Therefore, $X \perp Y, Z \mid Z$ implies $X \perp Y \mid Z$
Given $X \perp Y \mid Z$, we have $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$

$$\begin{aligned} P(X, Y, Z \mid Z) &= P(X, Y \mid Z) \\ &= P(X \mid Z)P(Y \mid Z) \\ &= P(X \mid Z)P(Y, Z \mid Z) \\ &= P(X \mid Z)P(Y, Z \mid Z) \end{aligned}$$

Therefore, $X \perp Y \mid Z$ implies $X \perp Y, Z \mid Z$
Hence Both statements are equivalent.

(c) (3 points) Prove the intersection property holds for any positive distribution P, assuming X, Y, Z, W are disjoint. Show where you used positivity of the distribution and set disjoint requirements in the proof.
sol: given

$$X \perp Y \mid Z, W \Rightarrow P(X \mid Y, W, Z) = P(X \mid W, Z)$$

$$X \perp W \mid Z, Y \Rightarrow P(X \mid Y, W, Z) = P(X \mid Y, Z)$$

$$P(X \mid W, Z) = P(X \mid Y, Z)$$

From baye's rule and since distributions are positive

$$P(X, W, Z)P(Y, Z) = P(X, Z, Y)P(W, Z)$$

since X, Y, Z, W are disjoint

$$\sum_Y P(X, W, Z)P(Y, Z) = \sum_Y P(X, Z, Y)P(W, Z)$$

$$P(X, W, Z)P(Z) = P(X, Z)P(W, Z)$$

$$P(X \mid W, Z) = P(X \mid Z)$$

Now consider

$$\begin{aligned} P(X, Y, W \mid Z) &= P(X, Y \mid Z, W)P(W \mid Z) \\ &= P(X \mid Z, W)P(Y \mid Z, W)P(W \mid Z) \\ &= P(X \mid Z)P(Y, W \mid Z) \end{aligned}$$

Thefore $X \perp Y, W \mid Z$ is true for the given condition.

(d) (2 points) Provide an example non-positive distribution P where the intersection property doesn't hold.
Note that $P(X|Z = z)$ is undefined if $P(Z = z) = 0$, so be sure to consider the definition of conditional independence when proving these properties.

3. (10 points) [TAKING A CHANCE WITH GRAPHS]

(a) (3 points) Prove that every Directed Acyclic Graph (DAG) G admits at least one topological ordering.
sol:
First we will show that there exists atleast one node in DAG with in degree 0 by contradiction.

Assume that every node in DAG has indegree >=1. Since indegree is greater than 1, every node will have a parent. We start with a node, visit its parent and repeat the process of visiting parent from given node. After n iterations we would definitely visit a node which was already visited, thus creating a cycle and therefore the graph is not DAG.

Therefore, our assumption is false and there always exist a node with indegree 0 in DAG.

We pick a node with indegree 0 and remove the node and its edges from graph thus resulting in another DAG. This process is repeated until there are no nodes in graph. The order of picking nodes gives topological ordering.

(b) (3 points) Prove that any P that factorizes according to a DAG G is in fact a valid probability distribution (i.e., is non-negative and adds up to one over all values that its variables can take, given that the local conditional probability distributions are all valid).

sol:
Let $X_1, X_2, ..X_n$ be the topological ordering of n nodes in a graph.

$$P(X_1, X_2, X_3, ..X_n) = \prod_{i=1}^{n} P(X_i \mid Pa_i)$$

Each term in the product is non-negative.

$$\begin{aligned} \sum_{x_n} P(X_1, X_2, .., X_n) &= \prod_{i=1}^{n-1} P(X_i \mid Pa_i) \sum_{x_n} P(X_n \mid Pa_n) \\ &= \sum_{i=1}^{n-1} \prod P(X_i \mid Pa_i) \end{aligned}$$

Repeating above with X_{n-1}, X_{n-2},...,X_1 we get

$$\begin{aligned} \sum_{X_1} .... \sum_{x_n} P(X_1, X_2, .., X_n) &= \sum_{X_1} P(X_1) \\ &= 1 \end{aligned}$$

Therefore

$$P(X_1, X_2, X_3, ..X_n) = \prod_{i=1}^{n} P(X_i \mid Pa_i)$$

is a valid probability distribution.

(c) (4 points) Prove for a Bayesian network in general that the factorization view implies the (local conditional) independences view.
sol:
We need to prove that $X_i \perp Non - Des_{X_i} \mid X_{Pa_i} \forall i$.
Let us consider node X_i we marginalize out the descendants of X_i to make a new graph G'
New set of nodes $X' = X_i \cup Non - Des_{X_i} \cup X_{Pa_i}$. We know that $P_G(X') = P'_G(X')$.

Now let X_1,X_2,..X_i is topological ordering of graph G' with X_i being a leaf node.

Factorization tells us that

$$P(X_1, X_2, ..X_i) = \prod_{k=1}^{i} P(X_k \mid X_{Pa_k})$$

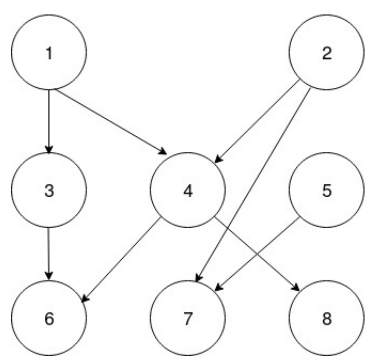
We will prove for a given X_i node and proof can be extended to for all nodes.

To prove $X_i \perp Non - Des_{X_i} \mid X_{Pa_i}$

$$
\begin{aligned}
P(X_i \mid Non - Des_{X_i}, X_{Pa_i}) &= \frac{P(X_i, Non - Des_{X_i}, X_{Pa_i})}{P(Non - Des_{X_i}, X_{Pa_i})} \\
&= \frac{P(X_1, X_2, ...X_i)}{P(X_1, X_2, ..X_{i-1})} \\
&= \frac{P(X_1, X_2, ...X_i)}{\sum_{x_i} P(X_1, X_2, ..X_{i-1}, X_i)} \\
&= \frac{\prod_{k=1}^{i} P(X_k \mid X_{Pa_k})}{\sum_{x_i} \prod_{k=1}^{i-1} P(X_k \mid X_{Pa_k}) P(X_i \mid X_{Pa_i})} \\
&= \frac{\prod_{k=1}^{i} P(X_k \mid X_{Pa_k})}{\prod_{k=1}^{i-1} P(X_k \mid X_{Pa_k}) \sum_{x_i} P(X_i \mid X_{Pa_i})} \\
&= \frac{\prod_{k=1}^{i} P(X_k \mid X_{Pa_k})}{\prod_{k=1}^{i-1} P(X_k \mid X_{Pa_k})} \\
&= P(X_i \mid X_{Pa_i})
\end{aligned}
$$

Therefore $X_i \perp Non - Des_{X_i} \mid X_{Pa_i}$. Similarly we can prove all other nodes.

4. (3 points) [TRYING OUT D-SEPARATION] Consider a distribution over 8 random variables $X_1, ..., X_8$ for the Bayesian network given below:



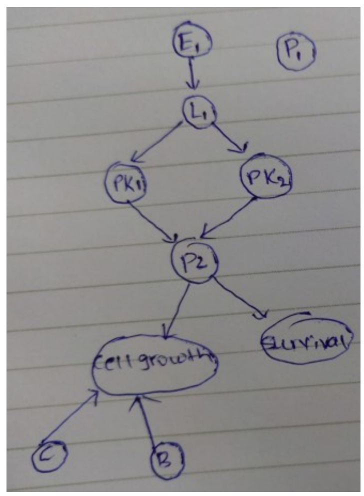
Give the largest set of random variables that is independent of the random variable:

(a) (1 point) X_3.
sol:
X_2, X_5, X_7

(b) (1 point) X_3, conditioned on X_1.
sol:
$X_1, X_2, X_4, X_5, X_7, X_8$

(c) (1 point) X_3, conditioned on X_1 and X_4.
sol:
$X_1, X_2, X_4, X_5, X_7, X_8$

5. (3 points) [FROM DOMAIN TO STRUCTURE] Construct a Bayesian network for the following signaling pathway. "The enzyme E1 is capable of phosphorylation and activates the lipid L1 without any signalling from protein P1. P1 is encoded by the PTEN gene. Mutations of this gene are a major step in the development of many cancers. L1 recruits oncoprotein P2, which is activated by the kinases PK1 and PK2. PK1 and PK2 are generated by L1. P2 activates downstream anabolic signaling pathways required for cell growth and survival. Once activated, P2 inactivates the pro-apoptotic proteins C and B, whose correct functioning is required for the normal growth of cells in the central nervous system." Provide the BN network DAG, and write down the expression for the joint probability density of random variables {E1, L1, P1, P2, PK1, PK2, C, B, CellGrowth}.



$P(E1, L1, P1, P2, PK1, PK2, C, B, \mathit{CellGrowth}) =$
$P(P1)P(E1)P(L1 \mid E1)P(PK1 \mid L1)P(PK2 \mid L1)P(P2 \mid PK1, PK2)P(\mathit{CellGrowth} \mid P2, C, B)P(C)P(B)$

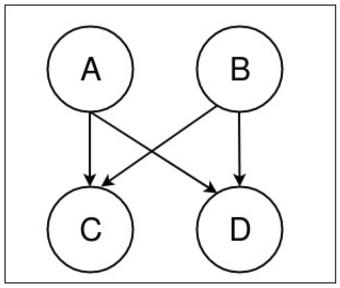
6. (6 points) [GRAPH ↔ CIs ↔ DISTRIBUTION] Consider the directed graphs and the probability distribution shown below. A, B, C, D are all binary random variables.

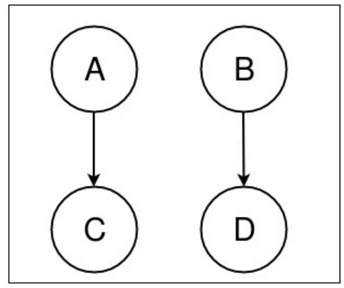
(a) (3 points) For each graph shown below, list all possible conditional independence statements implied by it (Note: no need to write CIs trivially implied by the Decomposition property i.e., given two CIs that hold $(X \perp Y, W \mid Z)$ and $(X \perp Y \mid Z)$, simply write only the first one).
sol:
first graph

$$C \perp D \mid A, B$$

$$A \perp B$$

second graph $(A \perp B, D)$, $(B \perp A, C)$, $(C \perp D, B)$, $(D \perp C, A)$
$(A \perp B, D|C)$, $(B \perp A, C|D)$, $(C \perp D, B|A)$, $(D \perp C, A|B)$





(b) (3 points) For the distribution given below (as a table of the joint probabilities of all the 16 configurations) list all conditional independences satisfied by it. Which of the graphs G above is an I-map for this distribution P(i.e., $I(G) \subseteq I(P)$)?

	$C, D = 0, 0$	$C, D = 0, 1$	$C, D = 1, 0$	$C, D = 1, 1$
$A, B = 0, 0$	0.5	0	0	0
$A, B = 0, 1$	0	0	0	0
$A, B = 1, 0$	0	0	0	0
$A, B = 1, 1$	0	0	0	0.5

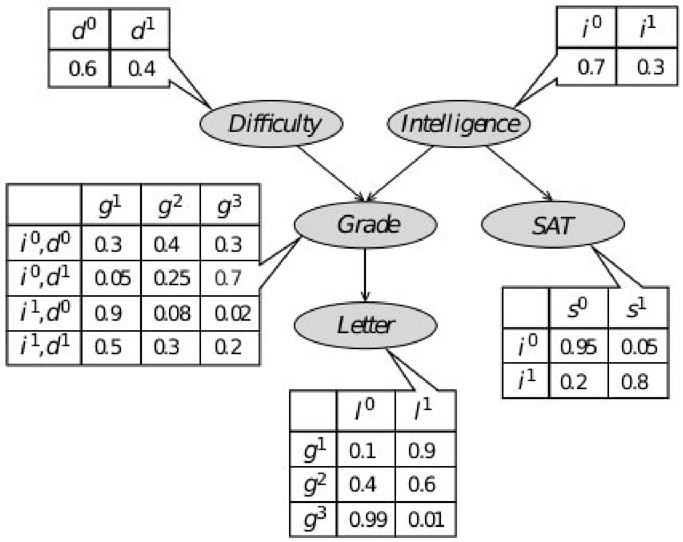
sol:
Here random variables A,B,C,D are symmetric w.r.t to each other.
$A \perp B \mid C$ and possible combinations from set A,B,C,D
$A \perp B \mid C, D$ and possible combinations from set A,B,C,D
$A \perp B$ is not true in this distribution. So the above given graphs which have $A \perp B$ as true are not I-map for this distribution.

7. (4 points) [REASONING/INFERENCE BY PENCIL] Write the expression for the joint probability density of $\{I, D, G, S, L\}$ for the BN below and use it to find:

(a) (1 point) What is the probability of a low IQ student to get a strong recommendation by taking an easy course?

$$
\begin{aligned}
P(L = l^1 \mid D = d^0, I = i^0) &= \frac{P(I = i^0, L = I^1, D = d^0)}{P(I = i^0, D = d^0)} \\
&= \frac{0.7 * 0.6 * \sum_G P(G \mid D = d^0, I = i^0) P(L = l^1 \mid G)}{0.7 * 0.6} \\
&= 0.3 * 0.9 + 0.4 * 0.6 + 0.3 * 0.01 \\
&= 0.513
\end{aligned}
$$

(b) (1 point) Given that a student has a weak recommendation letter and g^3 grade, what is the



probability for him/her to have high IQ?

$$\begin{aligned}
P(I = i^1 \mid L = l^0, G = g^3) &= \frac{P(I = i^1, L = l^0, G = g^3)}{P(L = l^0, G = g^3)} \\
&= \frac{0.3 * 0.99 * \sum_D P(D)P(G = g^3 \mid D, I = i^1)}{P(G = g^3)P(L = l^0 \mid G = g^3)} \\
&= \frac{0.3 * 0.99 * (0.6 * 0.02 + 0.4 * 0.2)}{0.99 * (0.3 * 0.6 * 0.7 + 0.7 * 0.7 * 0.4 + 0.02 * 0.3 * 0.6 + 0.2 * 0.3 * 0.4)} \\
&= 0.0789
\end{aligned}$$

(c) (2 points) Find probability of IQ being i^1 given that grade is g^2. What will be this probability if you also know that the course difficulty is d^1.
Note that these queries are examples of causal, evidential and intercausal reasoning respectively.

$$\begin{aligned}
P(I = i^1 \mid G = g^2) &= \frac{P(I = i^1, G = g^2)}{P(G = g^2)} \\
&= \frac{0.3 * \sum_D P(D)P(G = g^2 \mid D, I = i^1)}{\sum_D \sum_I P(G = g^2, D, I)} \\
&= \frac{0.3 * (0.6 * 0.08 + 0.4 * 0.3)}{0.4 * 0.7 * 0.6 + 0.25 * 0.7 * 0.4 + 0.08 * 0.3 * 0.6 + 0.3 * 0.3 * 0.4} \\
&= 0.17475
\end{aligned}$$

$$
\begin{aligned}
P(I = i^1 \mid G = g^2, D = d^1) &= \frac{P(I = i^1, G = g^2, D = d^1)}{P(G = g^2, D = d^1)} \\
&= \frac{P(G = g^2 \mid I = i^1, D = d^1)P(I = i^1)P(D = d^1)}{0.4 * \sum_I P(I)P(G = g^2 \mid D = d^1, I)} \\
&= \frac{0.3 * 0.4 * 0.3}{0.4 * (0.7 * 0.25 + 0.3 * 0.3)} \\
&= 0.33962
\end{aligned}
$$

(d) (3 points) (Bonus) As seen in class, "explaining away" (in c above) is a special case of the more general intercausal reasoning where different causes of the same effect can interact in any general way. Can you come up with examples, either in this BN or any other BN that is toy yet realistic, where the reverse of "explaining away" is true? To be precise, solve Exercise 3.4 in KF book to claim the bonus.