CSE – 426 Homework 12

Griffin Kent

1) Prove that the right-hand side of the following equation sums to 1 over all X_i, given that it is a factorization of the joint probability distribution on the left-hand side:

$$\mathbb{P}(X_1, \dots, X_n) = \prod_{i=1}^{n} \mathbb{P}(X_i | Pa_i),$$

Where Pa_i is the set of parents of random variable X_i on the corresponding Bayesian network. Your proof should not just sum up both sides and use $1 = \sum_{x_1,\dots,x_n} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$ to make the claim. Rather, you're expected to manipulate the $\sum_{x_1,\dots,x_n} \prod_{i=1}^{n} \mathbb{P}(X_i | Pa_i)$ to obtain 1, relying on the fact that the factorization is represented by a directed acyclic graph (DAG).

(Proof): Summing over all X_i of the right-hand side, we have

$$\sum_{X_j} \prod_{i=1}^{n} \mathbb{P}(X_i | Pa_i) = \sum_{X_1} \sum_{X_2} \dots \sum_{X_n} \prod_{i=1}^{n} \mathbb{P}(X_i | Pa_i).$$

Since this is a Bayesian network, we know that there are no cycles and that at least the last node n must have no children. We also know that at least the first node X_1 will have no children. Using this property of Bayesian networks along with the property that $P(X|Y) = \frac{P(X,Y)}{P(Y)}$, we can rewrite the above expression as the following:

$$\sum_{X_1} \sum_{X_2} \dots \sum_{X_{n-1}} \prod_{i=1}^{n-1} \frac{\mathbb{P}(X_i, Pa_i)}{\mathbb{P}(Pa_i)} \sum_{X_n} \frac{\mathbb{P}(X_n, Pa_n)}{\mathbb{P}(Pa_n)}.$$

Since the parent nodes are independent of their children nodes, this becomes

$$\sum_{X_1} \sum_{X_2} \dots \sum_{X_{n-1}} \prod_{i=1}^{n-1} \frac{\mathbb{P}(X_i, Pa_i)}{\mathbb{P}(Pa_i)} \cdot \frac{1}{\mathbb{P}(Pa_n)} \sum_{X_n} \mathbb{P}(X_n, Pa_n)$$

$$= \sum_{X_1} \sum_{X_2} \dots \sum_{X_{n-1}} \prod_{i=1}^{n-1} \frac{\mathbb{P}(X_i, Pa_i)}{\mathbb{P}(Pa_i)} \cdot \frac{\mathbb{P}(Pa_n)}{\mathbb{P}(Pa_n)} = \sum_{X_1} \sum_{X_2} \dots \sum_{X_{n-1}} \prod_{i=1}^{n-1} \frac{\mathbb{P}(X_i, Pa_i)}{\mathbb{P}(Pa_i)}.$$

We can continue this process starting with node n and back to node 1. Thus, the initial summation can be rewritten as the following:

$$\sum_{X_1} \mathbb{P}(X_1) \frac{1}{\mathbb{P}(Pa_2)} \sum_{X_2} \mathbb{P}(X_2, Pa_2) \dots \frac{1}{\mathbb{P}(Pa_{n-1})} \sum_{X_{n-1}} \mathbb{P}(X_{n-1}, Pa_{n-1}) \frac{1}{\mathbb{P}(Pa_n)} \sum_{X_n} \mathbb{P}(X_n, Pa_n)$$

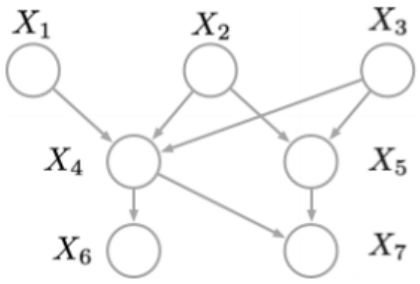
$$= \sum_{X_1} \mathbb{P}(X_1) \frac{\mathbb{P}(Pa_2)}{\mathbb{P}(Pa_2)} \cdot \dots \frac{\mathbb{P}(Pa_{n-1})}{\mathbb{P}(Pa_{n-1})} \cdot \frac{\mathbb{P}(Pa_n)}{\mathbb{P}(Pa_n)}$$

$$= \sum_{X_1} \mathbb{P}(X_1) = 1.$$

Therefore, completing the proof.

■

2) Let X_i be a discrete random variable taking values from $\{1, \dots, K\}$. Given the following Bayesian network, how many parameters are needed to fully specify the joint distribution $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$?



(Solution): The joint distribution for this Bayesian network can be factored as follows:

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$$
$$= \mathbb{P}(X_1)\mathbb{P}(X_2)\mathbb{P}(X_3)\mathbb{P}(X_4|X_1, X_2, X_3)\mathbb{P}(X_5|X_2, X_3)\mathbb{P}(X_6|X_4)\mathbb{P}(X_7|X_4, X_5).$$

Each independent variable needs $(K-1)$ parameters to be specified. Each directed arrow to a node will multiply the base $(K-1)$ parameters of that node by K. Thus, we can see that each node will add the following number of parameters:

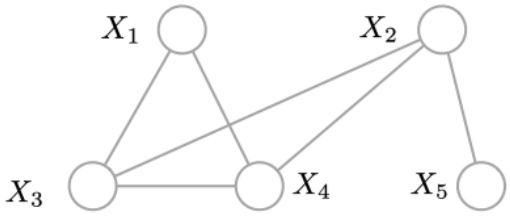
$$X_1 = (K-1)$$
$$X_2 = (K-1)$$
$$X_3 = (K-1)$$
$$X_4 = K^3(K-1)$$
$$X_5 = K^2(K-1)$$
$$X_6 = K(K-1)$$
$$X_7 = K^2(K-1).$$

Adding these together, we can see that the number of parameters that are needed to fully specify this joint distribution is

$$3(K-1) + K(K-1) + 2K^2(K-1) + K^3(K-1).$$

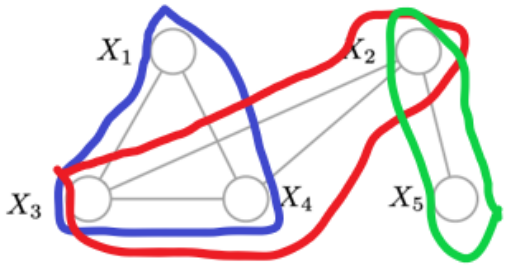
■

3) Identify the maximal cliques from the following Markov Random Field (MRF). Then factorize the joint distribution of the random variables on the MRF using the identified maximal cliques.



(Solution): A maximal clique is defined as a clique that, when adding any node outside of the clique, makes the augmented set of nodes no longer a clique. We can identify the

maximal cliques as the following sets: $\{X_1, X_3, X_4\}$, $\{X_2, X_3, X_4\}$, $\{X_2, X_5\}$. Drawing these on the graph, we have



The factorization equation for MRFs is defined as

$$\mathbb{P}(X_1, \dots, X_n) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(X_c)$$

Where $\psi_c(X_c)$ is a multivariate function on the variables in clique c and the normalization factor Z is defined as

$$Z = \sum_{x_1, .., x_n} \prod_{c \in \mathcal{C}} \psi_c(X_c).$$

Since every sub-clique can be absorbed into its maximal clique, we can use these equations for only the maximal cliques. Since we have three maximal cliques, we have the factorization form:

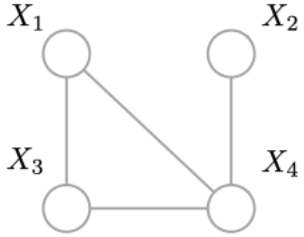
$$\mathbb{P}(X_1, X_2, X_3, X_4, X_5)$$

$$= \frac{\psi_1(X_1, X_3, X_4)\psi_2(X_2, X_3, X_4)\psi_3(X_2, X_5)}{\sum_{x_1, .., x_5} [\psi_1(X_1 = x_1, X_3 = x_3, X_4 = x_4)\psi_2(X_2 = x_2, X_3 = x_3, X_4 = x_4)\psi_3(X_2 = x_2, X_5 = x_5)]}$$

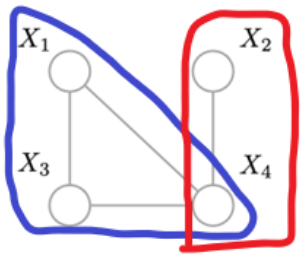
$$= \frac{\psi_1(X_1, X_3, X_4)\psi_2(X_2, X_3, X_4)\psi_3(X_2, X_5)}{\sum_{x_1, x_2, x_3, x_4, x_5} [\psi_1(X_1, X_3, X_4)\psi_2(X_2, X_3, X_4)\psi_3(X_2, X_5)]}.$$

■

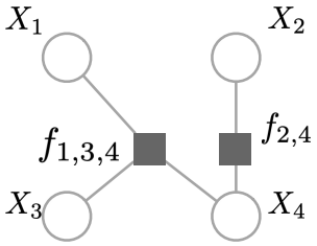
4) First, circle the maximal cliques in the MRF below. Then turn the MRF into a factor graph with the smallest number of factor nodes, using the identified maximal cliques. List the messages that are needed to compute the marginal $\mathbb{P}(X_2)$ when the node X_4 is used as the root. Explain when these messages are computed. (Refer to pages 409-410 of PRML)



(Solution): The maximal cliques are $\{X_1, X_3, X_4\}$ and $\{X_2, X_4\}$, and are shown here:



Using these maximal cliques, we can form the factor graph with the smallest number of factor nodes below:



The marginal distribution of a variable X_i is defined as the product of the incoming messages of the neighboring variables to X_i:

$$\mathbb{P}(X_i) = \prod_{s \in \mathcal{N}(i)} \mu_{s \to i}(X_i)$$

When X_4 is used as the root, using the sum-product algorithm will begin by computing all the messages from the leaf nodes to the root node; in this case, those messages will be:

$$\mu_{X_2 \to f_{(2,4)}}(X_2) = 1.$$

$$\mu_{f_{(2,4)} \to X_4}(X_4) = \sum_{X_2} f_{(2,4)}(X_2, X_4).$$

$$\mu_{X_1 \to f_{(1,3,4)}}(X_1) = 1.$$

$$\mu_{X_3 \to f_{(1,3,4)}}(X_3) = 1.$$

$$\mu_{f_{(1,3,4)} \to X_4}(X_4) = \sum_{X_1} \sum_{X_3} f_{(1,3,4)}(X_1, X_3, X_4).$$

Then, in order to calculate the marginals, we now propagate the messages out from the root node, which will be the following:

$$\mu_{X_4 \to f_{(2,4)}}(X_4) = 1.$$

$$\mu_{f_{(2,4)} \to X_2}(X_2) = \sum_{X_4} f_{(2,4)}(X_2, X_4).$$

$$\mu_{X_4 \to f_{(1,3,4)}}(X_4) = 1.$$

$$\mu_{f_{(1,3,4)} \to X_3}(X_3) = \sum_{X_4} f_{(1,3,4)}(X_3, X_4).$$

$$\mu_{f_{(1,3,4)} \to X_1}(X_1) = \sum_{X_4} f_{(1,3,4)}(X_1, X_4).$$

Now we can determine the messages that will be needed to calculate $\mathbb{P}(X_2)$ as the product of all incoming messages from the neighboring factor nodes; in this case it is only one:

$$\mathbb{P}(X_2) = \mu_{f_{(2,4)} \to X_2}(X_2).$$

This message will be calculated in the process as described above and will be determined after calculating the following messages in order:

$$\mu_{X_2 \to f_{(2,4)}}(X_2) = 1,$$

$$\mu_{f_{(2,4)} \to X_4}(X_4) = \sum_{X_2} f_{(2,4)}(X_2, X_4),$$

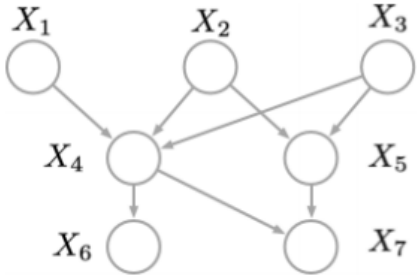
$$\mu_{X_4 \to f_{(2,4)}}(X_4) = 1,$$

$$\mu_{f_{(2,4)} \to X_2}(X_2) = \sum_{X_4} f_{(2,4)}(X_2, X_4).$$

■

5) (Graduate Only) D-Separation: In the Bayesian network in question 2, let $A = \{X_1\}$, $B = \{X_3\}$, and $C = \{X_2, X_5\}$. First, list all paths going from A to B. Then claim if each of the paths is blocked when C is observed nodes. Lastly, claim whether $A \perp\!\!\!\perp B|C$ or $A \ddagger B|C$ using the definition of D-Separation.
(*Note: In this problem the symbol "$\perp\!\!\!\perp$" is used to denote conditional independence and "$\ddagger$" is used to denote dependence*.)

(Solution): The Bayesian network we are working with is



From this graph we can see that all the paths that go from A to B are the following: $P_1 = \{X_1, X_4, X_3\}$, $P_2 = \{X_1, X_4, X_2, X_5, X_3\}$, $P_3 = \{X_1, X_4, X_7, X_5, X_3\}$.
According to the D-separation of a blocking, a path will be blocked if any of the following conditions hold:

- The path contains a node from C that meet wither head-to-tail or tail-to-tail at that node.
- A node is met head-to-head and neither that node nor any of its children are in the set C.

Path P_1: Path P_1 IS blocked by C because node X_4 is a head-to-head node which is not in C and none of its children are in C.

Path P_2: Path P_2 IS blocked by C because node X_4 is a head-to-head node which is not in C and none of its children are in C; it is also blocked because node X_2 is a tail-to-tail node that is observed in C.

Path P_3: Path P_3 IS blocked by C because node X_5 is a head-to-tail node which is in C.

Therefore, all three paths are blocked, and we can conclude by the definition of D-separation that $A \perp\!\!\!\perp B \mid C$ (A and B are conditionally independent given C).

■