Homework 6

CS 221 (Autumn 2012–2013)

Submission instructions: Write your answers in one PDF file named `hw6.pdf`. Remember to include your name and SUNet ID. Copy the PDF file onto `corn.stanford.edu`, ssh in to the machine, type `/usr/class/cs221/WWW/submit`, and follow the instructions.

1. Elimination (*5 points*)

Suppose we have a chain-structured Markov network with variables $X_1, \dots, X_n$ with domains $X_i \in \{1, \dots, r\}$, and factors $t_i(x_{i-1}, x_i)$ for each $i = 2, \dots, n$. If we wanted to compute $\mathbb{P}(X_i)$ (that is, $\mathbb{P}(X_i = v)$ for each $v \in \{1, \dots, r\}$) for a specific $i \in \{1, \dots, n\}$, we could eliminate all the variables except X_i from the two ends of the chain, and normalize the resulting weights (of possible values of X_i) to get a distribution over X_i.

a. (*1 point*) If we wanted to compute $\mathbb{P}(X_i)$ for every i, we could just repeat the above procedure for each X_i. What is the running time of this algorithm as a function of r and n?

b. (*1 point*) Let F_i be the set of factors produced by performing variable elimination (from the ends of the chain) on all variables except X_i for each i. If $n = 100$, which factors are in both F_3 and F_4? For example, the factor created by eliminating X_1 is in both. Hint: think about associating each new factor created with the set of variables whose elimination produced that factor.

c. (*3 points*) Describe an algorithm that computes $\mathbb{P}(X_i)$ for each $i = 1, \dots, n$ by re-using factors. Your algorithm should run in time $O(nr^2)$, and your description should be brief.

2. Markov networks to Bayesian networks (*8 points*)

We saw that Bayesian networks can be viewed as just Markov networks with a normalization constant of 1. Now we will show how an arbitrary Markov network can be converted into a Bayesian network.[1]

a. (*1 point*) Warm-up: consider a Markov network with two variables X_1 and X_2 with a single factor $f_{12}(x_1, x_2)$. Construct an equivalent Bayesian network (specify $p(x_1)$ and $p(x_2 \mid x_1)$ as a function of f_{12}). You must have $p(x_1)p(x_2 \mid x_1) \propto f_{12}(x_1, x_2)$.

b. (*1 point*) Now consider a Markov network with variables $X_1 \dots X_n$ with factors $f_i(x_i, x_{(i \bmod n)+1})$ for $i = 1, \dots, n$ (the factor graph looks like a ring). Recall that the weight of an assignment x is $\text{Weight}(x) = \prod_{i=1}^n f_i(x_i, x_{(i \bmod n)+1})$.

Let g_i be the new factor that is created when variables $X_{i+1}, \dots, X_n$ are eliminated. What variables (out of $X_1, \dots, X_i$) does g_i depend on? Write the expression for $\mathbb{P}(X_1 = x_1, \dots, X_i = x_i)$ as a function of $f_1, \dots, f_{i-1}, g_i$.

c. (*1 point*) Write an expression for the conditional distribution $\mathbb{P}(X_i = x_i \mid X_1 = x_1, \dots, X_{i-1} = x_{i-1})$ of the Markov network from part (b) as a function of some subset of the original factors $f_1, \dots, f_{i-1}$ and g_i.

[1] Note that this is not saying that each Markov network structure (which represents a set of possible Markov networks with that structure) can be represented by a Bayesian network structure (which represents a set of Bayesian networks with that structure). The sets are often overlapping but not exactly the same in general.

d. (*2 points*) Define a Bayesian network (i.e., what is the distribution $p_i(x_i \mid x_{\text{Parent}(i)})$ and specify the minimal set of parents Parent(i) for each node i) such that the Bayesian network defines the same joint distribution as the Markov network from parts (b) and (c), that is:

$$\text{(Bayesian network)} \quad \prod_{i=1}^{n} p(x_i \mid x_{\text{Parent}(i)}) = \frac{\text{Weight}(x)}{\sum_{x'} \text{Weight}(x)} \quad \text{(Markov network)},$$

where Weight(x) is defined above.

Hint: use induction. Assume that you've constructed a Bayesian network over $i-1$ variables. Use part (c) to construct a Bayesian network over i variables (you should be adding one local probability distribution $p(x_i \mid x_{\text{Parent}(i)})$ during each inductive step).

e. (*2 points*) Suppose you wanted to draw a set S of independent samples of assignments from the distribution $\mathbb{P}(X)$ defined by this Markov network. (Samples are useful for approximating queries; for example, the probability that $X_1 = X_5$ is estimated by $\frac{1}{|S|} \sum_{x \in S} [x_1 = x_5]$.)

Describe an algorithm that leverages Bayesian networks to draw independent samples. What is the running time of this algorithm as a function of n and $|S|$?

f. (*1 point*) Give a concrete example of a Markov network over $n = 3$ variables where Gibbs sampling fails to provide samples that yield correct estimates, but the above algorithm will work.

3. Chat room (*10 points*)

Suppose that there are K people (numbered 1 through K) who go in and out of a chat room. In the beginning, the room is empty. At each time step, the following occurs: (i) for each person in the chat room, he leaves with probability α and stays with probability $1-\alpha$; and (ii) for each person outside the chat room, he enters with probability α and stays out with probability $1-\alpha$.

If there are at least two people in the room, then one of them (uniformly at random), person j, will type in a utterance u with probability $p_j(u)$, where p_j is person j's distribution over utterances. If there are fewer than two people in the room, then no one types. Assume, for any person j, p_j is a distribution over a fixed set of utterances (including silence), $\mathcal{U}$, and is known to you.

You are not a member of this chat room, so you don't know exactly who is in the chat room at any time or who's talking, but do get to see the utterance u_i said by someone at each time step $i = 1, \dots, T$.

a. (*4 points*) Define a (dynamic) Bayesian network to model this scenario. What are the variables, domains of those variables, and local conditional probability distributions? All domain sizes should be linear in K, T, and $|\mathcal{U}|$.

b. (*2 points*) Suppose we're interested in a particular time step $i_0 \in \{1, \dots, T\}$. Given the observed utterances $u_1, \dots, u_T$, describe an algorithm to compute the probability that when person 1 and person K were both in the chat room at time i_0. Your algorithm can use variable elimination, but you must specify which variables you will eliminate, and write down explicitly how to combine the results of variable elimination (use equations).

c. (*2 points*) Describe an algorithm to compute the expected number of time steps that person 1 and person K were in the chat room *together* given the utterances $u_1, \dots, u_T$. Hint: recall that expectation is linear. The running time of your algorithm must be linear in T.

d. (*2 points*) Given the evidence $u_1, \dots, u_T$, you now want to compute the probability that there was *at least one* time step $i \in \{1, \dots, T\}$ that person 1 and person K were in the chat room together at time i. Change the variables and factors so that you can run variable elimination (plus a few simple operations) to compute the desired query.