CSE 515: Statistical Methods in Computer Science

Homework #2

Due at noon on February 11th

Guidelines: You can brainstorm with others, but please solve the problems and write up the answers by yourself. You may use textbooks (Koller & Friedman, Russel & Norvig, etc.) and lecture notes from class. Please do NOT use any other resources or references (e.g., example code, online problem solutions, etc.) without asking.

Submission instructions: Submit this assignment either by email to Chloé Kiddon (chloe@cs) or in person at the start of class on February 11th. If submitting by email, the attachment should be a PDF. Typed answers are highly preferred, but if this is a hardship, then handwritten answers are fine as long as they are completely legible.

1. Let's suppose we want to model the weather with an HMM. We can look outside our window in the morning each day and easily determine whether it is Sunny (S), Rainy (R) or Cloudy (C). We know that the weather is dependent on the current pressure system (for simplicity let's assume the weather is only dependent on that); however, we don't have access to any fancy meteorology equipment. Let's assume there is either a High Pressure (H) system or a Low Pressure (L) system on any given day. For our HMM, we will assume uniform start probabilities. The day-to-day transition probabilities and the emission probabilities are as follows:

π_i \ π_{i+1}	H	L
H	0.6	0.4
L	0.4	0.6

π_i \ x_i	S	R	C
H	0.7	0.1	0.2
L	0.2	0.7	0.1

   (a) Compute the most likely sequence of the five hidden states for the observed sequence R, C, C, S, S by stepping through the Viterbi algorithm by hand. (You don't need to show every mathematical operation; however, you may want to show at least one or two sample calculations to allow for partial credit if the final answer is incorrect.)

   (b) Use the forward-backwards algorithm to compute the probability distribution over states at the time $i = 3$. (Again, you don't need to show every mathematical operation; however, you may want to show at least one or two sample calculations to allow for partial credit if the final answer is incorrect.)

   (c) Is the most likely state the same as the state in the most likely sequence? Will this always be the case? Why?

(d) Now let's assume we wanted to create a different model of the weather. This time, instead of using the pressure systems as the hidden states, we wanted to use the four seasons (Spring, Summer, Fall, and Winter) as our hidden states. Assume the probability that a season transitions to its following season (e.g., Spring $\rightarrow$ Summer, Summer $\rightarrow$ Fall, etc.) is p and the probability that it stays in the same season is $(1-p)$.

i. If at time t it is Summer, what is the distribution over the number of days d until it first becomes Fall (that is, the smallest number d such that the season at time $t+d$ is not Summer)?

ii. Based on your previous answer, why is an HMM not an ideal model for modeling weather with the seasons as hidden states?

2. Assume a 1-D linear Gaussian dynamical model defined over a set of state variables $\mathbf{X}$ and a set of observation variables $\mathbf{Z}$ as follows:
$x_i \sim \mathcal{N}(x_{i-1}, \sigma_x)$
$z_i \sim \mathcal{N}(x_i, \sigma_z)$
Assume the initial mean and standard deviation of the belief state are μ_0 and σ_0, respectively.

(a) Assume $\sigma_x = 0$, $\sigma_z = 1$, $\mu_0 = z_0$, and $\sigma_0 = 1$. If the observations at the first three time steps are z_1, z_2, z_3, what are the mean and standard deviation at time $t = 3$ (that is, what are μ_3 and σ_3) in terms of z_0, z_1, z_2, and z_3? Briefly explain why the parameters and/or initial conditions specified cause the optimal estimate for x (i.e., the mean) to have this form.

(b) What if we use all the same parameters, initial conditions, and observations as the previous question except that now $\sigma_x = 1$. What is the mean and standard deviation at time $t = 3$? How does the mean compare to the answer from the previous question and why did the change of σ_d cause this behavior?

3. A group of thieves are planning a bank heist. Here are the details:

- The heist's success depends on how much cash they steal and whether or not they get away.
- It is easier for the thieves to get away if they remembered to gas up the car.
- The more the group practices and the more coffee they drink, the sneakier they are.
- The sneakier the group, the more easily it can get past the alarm.
- If the alarm sounds, the vault bolts closed, which makes it harder to steal cash.
- The group is better at remembering things if they've had some coffee.

(a) Draw the graph of a Bayesian network consistent with the statements above using variables Success, GetAway, GetCash, Coffee, Gas, Practice, Sneakiness, and Alarm.

(b) In this network, what is the Markov blanket of Coffee?

(c) According to this network, if the team is very sneaky is the car more likely to be gassed?

(d) According to this network, are the amount of practice and the group getting away independent?

(e) According to this network, are the amount of practice and the group getting away independent given the team doesn't succeed?

(f) According to this network, are the amount of practice and the group getting away independent given the team had coffee?

(g) Convert the Bayesian network to a Markov network using moralization.

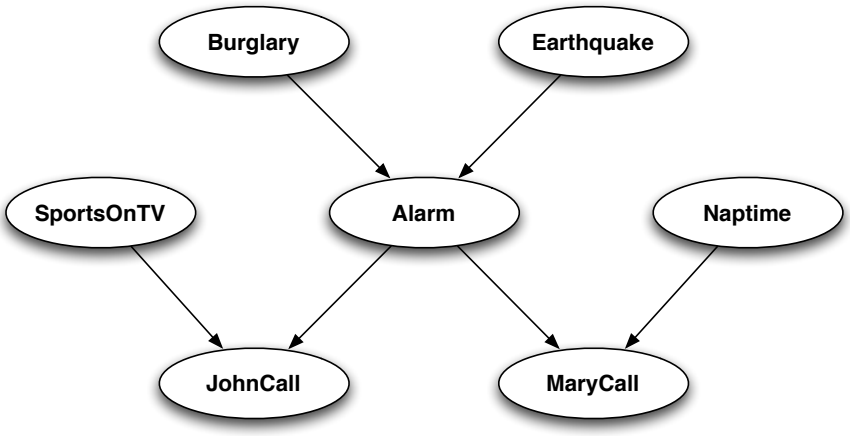
4. A group of thieves has been arrested and is sitting in holding: Al, Bugsy, Carmine, and Danny. When interrogated, each thief tells his story independently of the rest, unless he conspired with someone else while in the holding pen to tell the same story. Carmine refuses to talk Bugsy ever since Bugsy forgot to gas the car, and Al won't speak to Danny since Danny tripped the alarm. However, every other pairing of two thieves spends some time talking while in holding. (They were in there a long time.) Assume we have four variables representing the story each of the thieves tell the police.

   (a) List the set of conditional independencies that the distribution of these four variables should satisfy.

   (b) Can you create a Bayesian network that captures all the independence statements? If so, draw the network. If not, explain why.

   (c) Can you create a Markov network that captures all the independence statements? If so, draw the network. If not, explain why.

5. (a) Consider the alarm network show below:

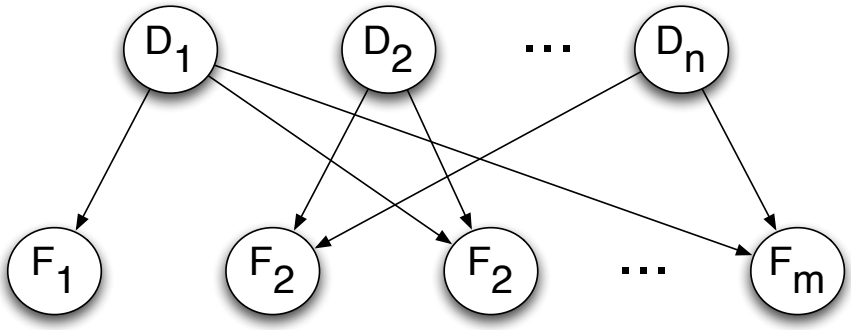


Construct a Bayesian network structure with nodes Burglary, Earthquake, JohnCall, MaryCall, SportsOnTV, and Naptime that is consistent with any probability distribution over the network after marginalizing out Alarm. Your network should

be the minimal consistent network, i.e., preserve as many independencies as possible while consistently representing all necessary dependencies. (Formally, you are constructing a minimal I-map for the marginal distribution over those variables defined by the above network. For more information about I-maps, refer to Section 3.2.3 of Koller & Friedman.)

(b) Generalize the procedure you used above to an arbitrary network. More precisely, assume we are given a network BN, an ordering $X_1, \ldots, X_n$ that is consistent with the ordering of the variables in BN, and a node X_i to be removed. Specify a network BN' such that BN' is consistent with this ordering, and such that BN' is the minimal consistent network of $P_{BN}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots X_n)$. Your answer must be an explicit specification of the set of parents for each variable in BN'.

6. Consider the network shown below:



We assume that all variables are binary, and that the F_i variables in the second layer all have *noisy or* CPDs. Specifically, the CPD of F_i is given by:

$$P(f_i^0 \mid \mathbf{Pa}_{F_i}) = (1 - \lambda_{i,0}) \prod_{D_j \in \mathbf{Pa}_{F_i}} (1 - \lambda_{i,j})^{d_j}$$

where $\lambda_{i,j}$ is the noise parameter associated with parent D_j of variable F_i. (This network architecture, called a *BN2O network*, is characteristic of several medical diagnosis applications, where the D_i variables represent diseases (e.g., flu, pneumonia), and the F_i variables represent medical findings (e.g., coughing, sneezing). For more information about BN2O networks see box 5.C of Koller & Friedman.)

Our general task is medical diagnosis: We obtain evidence concerning some of the findings, and we are interested in the resulting posterior probability over some subset of diseases. Since we are only interested in computing the probability of a particular subset of the diseases, we wish (for reasons of computational efficiency) to remove from the network those disease variables that are not of interest at the moment.

(a) Begin by considering a particular variable F_i, and assume (without loss of generality) that the parents of F_i are $D_1, \ldots, D_k$, and that we wish to maintain only the parents $D_i, \ldots, D_\ell$ for $\ell < k$. Show how we can construct a new *noisy or* CPD for F_i that preserves the correct joint distribution of $D_1, \ldots, D_\ell, F_i$.

(b) We now remove some fixed set of disease variables D from the network, executing this pruning procedure for all the finding variable F_i, removing all parents $D_j \in D$. Is this transformation exact? In other words, if we compute the posterior probability over some variable $D_i \notin D$, will we get the correct posterior probability (relative to our original model)? Justify your answer.