CS 476/676, Fall 2013
Problem Set #1b
(Due by 11:59pm on Wednesday, October 23rd)

---

1 Instructions

This assignment includes analytical questions. Please read the document carefully.

1.1 What to Hand In

All of your submission files should be handed in as a single archive named `hw1b-username.zip`, where `username` is replaced with your JHED ID. This zip archive will include:

- Written answers to questions should go in a single PDF document named `writeup.pdf`. Questions that require a written response are given under the section headings **Analytical Questions** or **Empirical Questions**. We recommend using LaTeX to typeset your writeup.

1.2 Submission Policies

Please note the following:

- **Collaboration:** You are allowed to work in groups of 1–3 people. If you work in a group, you can hand in a single submission that includes the names of every team member. If you do a joint submission, your file should be named `hw1-username1-username2-username3.zip`. Alternatively, each team member may turn in their own submission: however, if you work with others but hand in your own submission, your writeup **must** list the names of the other people (1–2 maximum) at the top of your `writeup.pdf`. **Uncredited collaboration is cheating.**
- **Late Submissions:** We allow each student to use up to 3 late days over the semester. You have late days, not late hours. This means that if your submission is late by any amount of time past the deadline, then this will use up a late day. If it is late by any amount beyond 24 hours past the deadline, then this will use a second late. **If you jointly submit an assignment as a team, then every team member will lose late days if the assignment is submitted late.** If you collaborate with team members but independently submit your own version, then late hours will only apply to you.

2 Problem Set [110 points]

2.1 Network Manipulation [20 points]

One operation on Bayesian networks that arises in many settings is the marginalization of some node in the network. Consider the network below, which models the flu prevalence in a population.

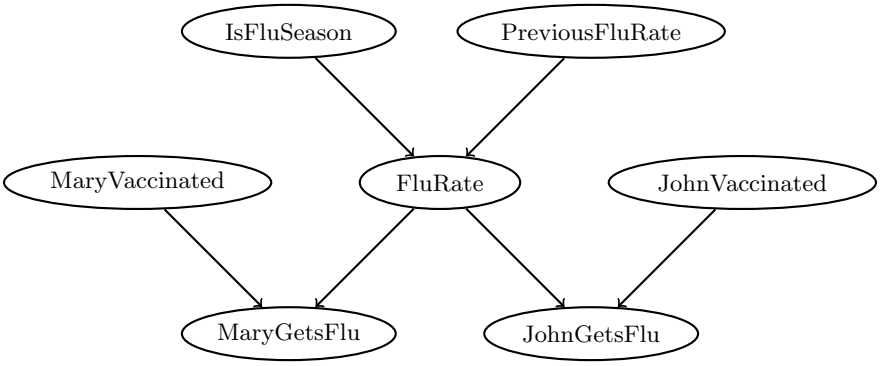


Figure 1: Network $\mathcal{F}$.

Let's call this network $\mathcal{F}$. Your job is to construct a Bayesian network $\mathcal{F}'$ over all of the nodes except for *FluRate* that is a minimal I-map for the marginal distribution $P_{\mathcal{F}}(IFS, PFR, MV, JV, MF, JF)$ defined by the above network. Be sure to get *all* dependencies that remain from the original network.

2.1.1 Drawing Questions [14 points]

Draw the network $\mathcal{F}'$ you constructed that satisfying the above requirements.

2.1.2 Analytical Questions [6 points]

1. **[6 points]** Generalize the procedure you used to solve the preceding problem into a node elimination algorithm. That is, define an algorithm that transforms the structure $\mathcal{G}$ into $\mathcal{G}'$ such that one of the nodes X_i of $\mathcal{G}$ is not in $\mathcal{G}'$ and $\mathcal{G}'$ is an I-map of the marginal distribution over the remaining variables as defined by $\mathcal{G}$.

2.2 Network Queries [20 points]

Let's consider the sensitivity of a particular query $P(X|\boldsymbol{Y})$ to the CPD of a particular node Z. Let X and Z be nodes (which are not directly connected) and $\boldsymbol{Y}$ be a set of nodes. We say that Z has a *requisite CPD* for answering the query $P(X|\boldsymbol{Y})$ if there are two networks $\mathcal{B}_1$ and $\mathcal{B}_2$ that have identical graph structure $\mathcal{G}$ and identical CPDs everywhere except at the node Z, and where $P_{\mathcal{B}_1}(X|\boldsymbol{Y}) \neq P_{\mathcal{B}_2}(X|\boldsymbol{Y})$; in other words, the CPD of Z affects the answer to this query.

This type of analysis is useful in various settings, including determining which CPDs we need to acquire for a certain query.

Suppose we modify $\mathcal{G}$ into a graph $\mathcal{G}'$ which is identical to $\mathcal{G}$ except contains a new "dummy" node $\hat{Z}$ which is a parent of Z (thereby altering the CPD of Z). One way to test whether Z is a requisite probability node for $P(X|\boldsymbol{Y})$ is to test whether $\hat{Z}$ has an active trail to X given $\boldsymbol{Y}$ in $\mathcal{G}'$—if so, you can conclude that altering the CPD of Z would affect the result of $P(X|\boldsymbol{Y})$.

2.2.1 Analytical Questions [20 points]

1. **[10 points]** Prove that the above strategy is a sound criterion for determining whether Z is a requisite probability node for $P(X|\boldsymbol{Y})$ in $\mathcal{G}$. Specifically, show that if this criterion fails to

identify Z as a requisite node, then for all pairs of networks $\mathcal{B}_1$, $\mathcal{B}_2$, $P_{\mathcal{B}_1}(X|\boldsymbol{Y}) = P_{\mathcal{B}_2}(X|\boldsymbol{Y})$.

2. **[10 points]** Show that this criterion is weakly complete (like d-separation) in the sense that, if it identifies Z as a requisite in $\mathcal{G}$, there exists some pair of networks $\mathcal{B}_1$, $\mathcal{B}_2$ such that $P_{\mathcal{B}_1}(X|\boldsymbol{Y}) \neq P_{\mathcal{B}_2}(X|\boldsymbol{Y})$.

2.3 Markov Blankets and Separation [20 points]

Let $\text{MB}_{\mathcal{G}}(X)$ denote the Markov blanket of node X in an **undirected** graph $\mathcal{G}$, whose set of nodes is denoted $\mathcal{X}$. The Markov blanket of X is the set of X's neighbors.

1. **[16 points]** Prove the following:

   (a) **[8 points]** For any variable X, let $\boldsymbol{W} = \mathcal{X} - \{X\} - \text{MB}_{\mathcal{G}}(X)$. Then X and $\boldsymbol{W}$ are separated given $\text{MB}_{\mathcal{G}}(X)$.

   (b) **[8 points]** The set $\text{MB}_{\mathcal{G}}(X)$ is the minimal set for which this property holds.

2. **[4 points]** Now suppose that $\mathcal{G}$ is **directed**. In a directed graph, the Markov blanket of X is the following set of nodes: X's parents, X's children, and X's children's parents. We would like to prove the two statements in question 1 above in the directed case, i.e. d-separation. It turns out that you can do this straightforwardly by utilizing the proofs you have already constructed for the undirected case. Please sketch a proof of these two statements for a directed graph $\mathcal{G}$ which relies on the proof for the undirected case. Your answer should be brief—if your solution is complicated, you are probably on the wrong track.

2.4 Comparing Network Types: Disease severity over time [22 points]

In this section, we will consider modeling the severity of a disease over time using *linear chain* graphical models, which are commonly used to model discrete time series data.

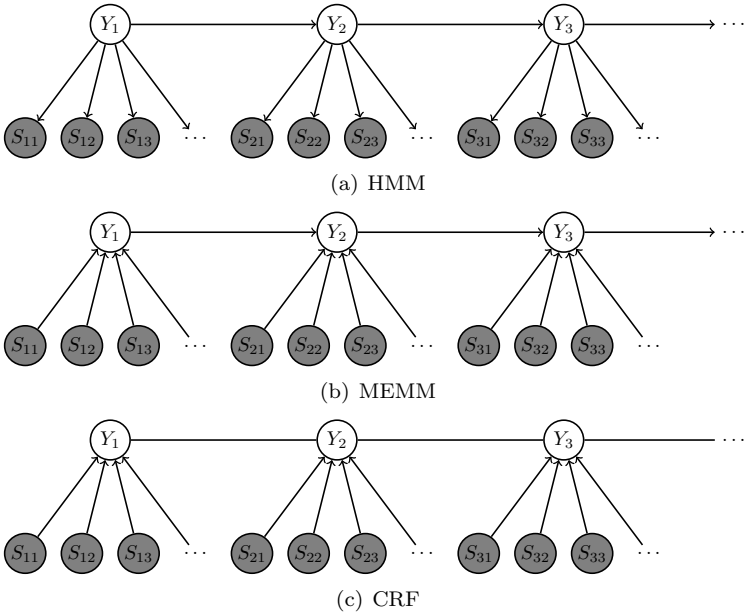
The random variable Y_i denotes the disease severity (None, Mild, Severe) on day i. The random variable $S_{i,j}$ indicates a symptom j on day i, such as the patient's temperature or whether or not she has a cough (for example, if Mary had a case of the flu, we could model it day-by-day using such a model). The symptoms $\boldsymbol{S}$ are observed. The severity $\boldsymbol{Y}$ is not observed, but it can be inferred from the observed symptoms.

Figure 2 shows three different types of networks to model this. The first is directed graph which encodes $P(\boldsymbol{Y}, \boldsymbol{S})$, called a hidden Markov model (HMM). The second is a directed graph which encodes $P(\boldsymbol{Y}|\boldsymbol{S})$ called a maximum-entropy Markov model (MEMM). The third is a type of conditional random field (CRF), which also models $P(\boldsymbol{Y}|\boldsymbol{S})$. CRFs can be partially directed or undirected; a partially directed version is shown here.

The differences between these models are subtle, and even in practice it is not always clear which model to use. As you'll learn later in the semester, the choice of model has implications and tradeoffs regarding learning and inference. In this assignment, we want you to think about the subtle differences in assumptions made by these models.

2.4.1 Analytical Questions [22 points]

1. **[12 points]** This question will investigate differences in conditional independence and the "explaining away" property, which will be discussed again in 2.5. Suppose we are modeling Mary's flu over a series of days and suppose that on day 1 and 2, Mary's symptoms are



(a) HMM

(b) MEMM

(c) CRF

Figure 2: Three types of linear chain models, described in Section 2.4. Gray nodes are observed variables; white nodes are unobserved.

consistent with the flu. However, on day 3, Mary begins vomiting. Vomiting is not typically associated with the flu, so we suspect that she actually has some other illness, and therefore it is unlikely that she has had the flu on any of the days. More precisely, we want to say that if we observe that the vomiting symptom on day 3 is active (denoted $S_{3,1}$) then the probability of a mild or severe flu case on day 1 (Y_1) should decrease.

(a) **[4 points]** Describe whether each of the three networks can or cannot model this influence, and why. Specifically, is it possible to define the CPDs such that $P(Y_1 = \texttt{Severe}|S_{3,1} = \texttt{Yes}) < P(Y_1 = \texttt{Severe}|S_{3,1} = \texttt{No})$? To use the terminology of 2.2, this asks whether $S_{3,1}$ is a requisite probability node for $P(Y_1)$.

(b) **[4 points]** Repeat the previous question assuming that Y_3 is observed. That is, for each of the three models, determine if $S_{3,1}$ is a requisite node for $P(Y_1|Y_3)$.

(c) **[4 points]** Repeat the previous question assuming both Y_3 and Y_2 are observed. That is, for each of the three models, determine if $S_{3,1}$ is a requisite node for $P(Y_1|Y_3, Y_2)$.

2. **[10 points]** This question will investigate how your set of features (the $\boldsymbol{S}$ variables) might influence your choice of model. Your answers to the following questions should be brief.

(a) **[5 points]** Suppose you have a large number of features (observed symptoms or test results) with which to make inferences about disease status. There are many different situations that can lead you to prefer certain model types over others. You might have sparse feature vectors (only a few `Yes` symptoms or tests for a given patient on a given day) or dense feature vectors. You might expect that only a few features will matter in predicting the disease status, or you might believe that many features be important. You might have a lot of prior knowledge about what symptoms you expect

to observe given a particular disease state (because you are a biology expert) or you might have little knowledge and will need to learn from historical data. Keeping in mind these possibilities, for each of the three models, describe circumstances under which you might favor this model over the others. (There is not necessarily one "correct" answer; we will grade your response based on your justification.)

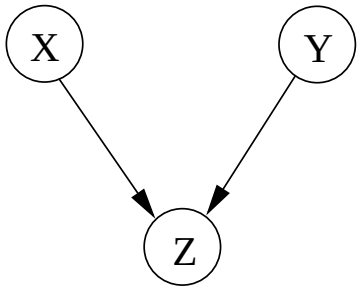
(b) **[5 points]** Sometimes we might want to use features that are strongly correlated. For example, suppose one symptom variable represents a "dry cough" and another variable represents any type of cough: if a person exhibits a dry cough, then by definition they are also exhibiting a cough, so if the value of the former is `Yes`, the value of the latter will also be `Yes`. For each of the three models, describe whether having correlated symptom variables is a violation of the model's assumptions of independence or conditional independence.

2.5 Noisy-OR and Explaining Away [28 points]

One property of Bayesian networks is something called *explaining away* which occurs when evidence that establishes a cause for an event reduces the likelihood of other possible causes.

2.5.1 Analytical Questions [28 points]

1. **[14 points]** Consider the following three-node Bayesian network:



Assume that X, Y, and Z are binary random variables, and that there is a Noisy-Or interaction between Z and X, Y; i.e., the CPD for Z is:

Z	x^0, y^0	x^0, y^1	x^1, y^0	x^1, y^1
z^0	1	λ_Y	λ_X	$\lambda_X \lambda_Y$
z^1	0	$1 - \lambda_Y$	$1 - \lambda_X$	$1 - \lambda_X \lambda_Y$

(a) **[6 points]** Show that this network must satisfy the explaining-away property: $P(x^1 \mid z^1) \geq P(x^1 \mid y^1, z^1)$. (Hint: the solution to this problem can fit in half a page.)

(b) **[4 points]** Show different CPDs for the same 3-node network structure where the reverse type of intercausal reasoning occurs, i.e., where: both causes increase the probability of the effect, i.e., $P(z^1|x^1) > P(z^1)$ and $P(z^1|y^1) > P(z^1)$, and also each cause increases the probability of the other, i.e., $P(x^1|z^1) < P(x^1|y^1, z^1)$, and similarly $P(y^1|z^1) < P(y^1|x^1, z^1)$.

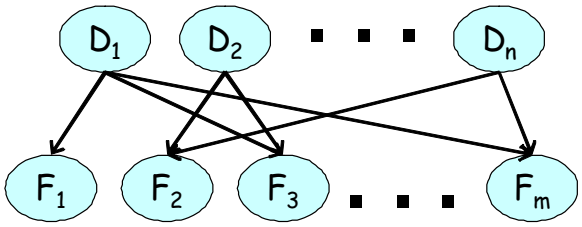
(c) **[4 points]** One method we would like to use in probabilistically modeling our world is context-specific independence. In class, we showed how such independencies arise in

the context of tree-structured CPTs. In fact, they also arise in the Noisy-Or model. Consider the Noisy-Or network in (a). Find a form of context-specific independence that necessarily arises in this network, regardless of the values of the Noisy-Or parameters.

2. **[14 points]** Consider the network shown below, where 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:

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

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).



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. However, we are only interested in computing the probability of a particular subset of the diseases, so that we wish (for reasons of computational efficiency) to remove from the network those disease variables that are not of interest at the moment.

(a) **[8 points]** 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 $D_1, \ldots, D_\ell$ for $\ell < k$. Show how we can construct a new *noisy or* CPD for F_i that preserves the correct joint distribution over $D_1, \ldots, D_\ell, F_i$.

(b) **[6 points]** We now remove some fixed set of disease variables D from the network, executing this pruning procedure for all the finding variables 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.