6.867 Machine learning and neural networks

Problem set 5

Solutions

Problem 1

In figure 1 we have two diseases d_1 and d_2 as well as two potential findings f_1 and f_2. The diseases and findings are assumed to be binary. We assume further that the conditional probabilities of $P(f_1|d_1, d_2)$ and $P(f_2|d_2)$ are noisy-OR models as described in the lectures.

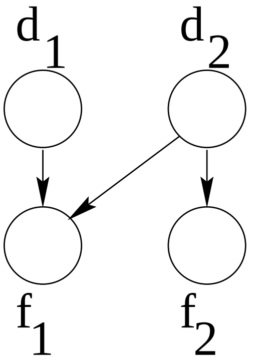


Figure 1: A Bayesian network for medical diagnosis. We assume that the conditional probabilities are noisy-OR models.

1. Now, suppose we know that the outcome of finding f_1 was positive ($f_1 = 1$). What happens to the disease probabilities? Justify your answer.

   Both diseases will have higher marginal posterior probabilities as a result: $P(d_1 = 1|f_1 = 1) > P(d_1 = 1)$ *and* $P(d_2 = 1|f_1 = 1) > P(d_2 = 1)$. *This is particular to the noisy-OR model.*

   Why would this happen? Well, we have to somehow explain the evidence that the finding was one. In a noisy-OR model, this can happen only if at least one of the diseases were present (or if the positive finding occured because of an unknown cause). The relative strength of which cause explained the finding depends on the actual probability values. Nevertheless, this doesn't change the fact that there's now more *evidence*

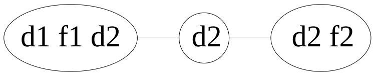


Figure 2: A junction tree representation of the disease/finding Bayesian network.

that each disease is present than before. This has to be reflected in the posterior probabilities.

You can verify that this is indeed what happens by plugging in any numerical values to the conditional probabilities

$$P(f_1 = 1|d_1, d_2) \quad = \quad 1 - (1 - q_0)\,(1 - q_1)^{d_1}\,(1 - q_2)^{d_2} \tag{1}$$

and the priors $P(d_1)$ and $P(d_2)$.

It is not hard though a bit lengthy to show this more mathematically.

2. Suppose later on we learn that finding f_2 was positive as well. How will the probability of disease d_1 being present change as a result? Justify your answer.

   Now something interesting happens. The posterior for $d_2 = 1$ further increases, $P(d_2 = 1|f_1 = 1, f_2 = 1) > P(d_2 = 1|f_1 = 1)$ but the posterior for $d_1 = 1$ actually decreases but still remains above the original prior probability:

   $$P(d_1 = 1) < P(d_1 = 1|f_1 = 1, f_2 = 1) < P(d_1 = 1|f_1 = 1) \tag{2}$$

   This is known as the explaining away *effect. We have acquired further evidence that d_2 is present (without saying anything about d_1) and this decreases our belief that d_1 is present. We would indeed now attribute $d_2 = 1$ as the main cause for observing $f_1 = 1$. This chips away the evidence from d_1 but not to zero. Thus there's still some evidence that d_1 might have been the cause of $f_1 = 1$ and thus its posterior remains higher than the prior.*

3. Construct a junction tree for the graph model in figure 1. Also initialiaze the potential functions for the junction tree. Are there many ways of initializing the potentials?

   In figure 2, we have a junction tree representation of the bayesian network. Note that we need one node for d_1, f_1 and d_2 because f_1 is conditionally dependent on d_1 and d_2. "Moralization" is the realization of this fact.

   There are many ways to initialize the potentials, for example

   $$\psi(d_1, f_1, d_2) = P(f_1|d_1, d_2)P(d_1)P(d_2) \qquad \psi(d_2) = 1 \qquad \psi(d_2, f_2) = P(f_2|d_2) \tag{3}$$

   $$\psi(d_1, f_1, d_2) = P(f_1|d_1, d_2)P(d_1)P(d_2) \qquad \psi(d_2) = P(d_2) \qquad \psi(d_2, f_2) = P(f_2, d_2) \tag{4}$$

4. Roughly speaking, how many operations do we need to perform to complete both "collect" and "distribute" steps? You can assume here that we will blindly apply both propagation steps whatever the evidence might be.

   The collection and distribution steps for this junction tree are relatively simple because we only have two nodes (the d_2 is not counted here). Let's choose the $d_2 f_2$ node as the root node. One can consider the number of values that need to be summed over as the "operations". Here, the collection operation requires summing over the values of d_1 and f_1. However, to get $\psi'(d_2)$ for all values of d_2 we have to perform 8 operations (calculations involve three binary valued variables). Updating $\psi(f_2, d_2)$ as a result would take another 4 operations for the same reason.

   Next, the distribution operation consists of propagating changes made to $\psi(f_2, d_2)$ to the $d_2 f_1 d_1$ node. Similarly, we have to obtain $\psi'(d_2)$ again and this takes four operations (involves one less variable than before). Updating $\psi(d_1, f_1, d_2)$ for all configurations would require another 8 operations, however.

   This rough calculation that ignores additional multiplications or divisions that we need merely highlights the fact that the number of operations needed increases exponentially with the clique size (the number of configurations of the associated variables increases exponentially in the clique size).

5. Suppose we would like to determine which finding to query (which of the corresponding tests to carry out). For this we need to evaluate the mutual information between the disease configurations and the possible values of each of the findings. In other words, we have to compute $I(f_1; d_1, d_2)$ and $I(f_2; d_1, d_2)$, where, for example,

$$I(f_1; d_1, d_2) = \sum_{d_1, d_2, f_1 = 0,1} P(d_1, d_2, f_1) \log \frac{P(d_1, d_2, f_1)}{P(d_1, d_2) P(f_1)} \tag{5}$$

   Now, show that the graph structure (original or the junction tree) implies that $I(f_2; d_1, d_2) = I(f_2; d_2)$.

$$\begin{aligned} I(f_2; d_1, d_2) &= \sum P(d_1, d_2, f_2) \log \frac{P(d_1, d_2, f_2)}{P(d_1, d_2) P(f_2)} && (6) \\ &= \sum P(d_1, d_2, f_2) \log \frac{P(f_2|d_2) P(d_1) P(d_2)}{P(d_1) P(d_2) P(f_2)} && (7) \\ &= \sum P(d_2, f_2) \log \frac{P(f_2, d_2)}{P(d_2) P(f_2)} = I(f_2; d_2) && (8) \end{aligned}$$

   In equation 7, we used the fact that f_2 is conditionally independent of d_1 and that the diseases are marginally independent. In equation 8, we marginalize d_1 out since the log *term is independent of d_1.*

6. Based on your results for 3) and 5) show that the marginal probabilities that we compute in the junction tree suffice for evaluating which test we should query next ("active learning").

   The question asks whether the junction tree contains all the probabilities that we will need access to. In other words, does it explicitly *contain the probabilities we need. We need $P(d_1, d_2, f_1)$, $P(d_1, d_2)$ and $P(f_1)$ in order to calculate $I(f_1; d_1, d_2)$. As we showed in part e), we need $P(d_2, f_2)$, $P(d_2)$ and $P(f_2)$ in order to calculate $I(f_2; d_1, d_2)$. The junction tree explicitly holds the probabilities $P(d_1, f_1, d_2)$ and $P(d_2, f_2)$. The only other probabilities that we need are contained within these two. We can achieve $P(d_2)$, $P(f_2)$, $P(d_1, d_2)$ and $P(f_1)$ by simply summing (marginalizing) over probabilities in the junction tree. Hence, all of the probabilites are available in the sense that there is no need to invoke, e.g., Bayes Law to calculate any of the necessary probabilities.*

   To get a sense of why it is important to only use probabilities that are explicit in the junction tree, imagine trying to calculate $P(x_1, x_{100})$ in a 100 node markov chain where the probabilities $P(x_i, x_{i+1})$ and $P(x_i)$ are the only distributions explicit in the junction tree. This does not mean that we couldn't obtain $P(x_1, x_{100})$ efficiently but just that this requires additional work.

Problem 2

Your task here is to identify the relevant variables and the graph structure that captures the following (imaginary) setting. There may be multiple "correct" answers.

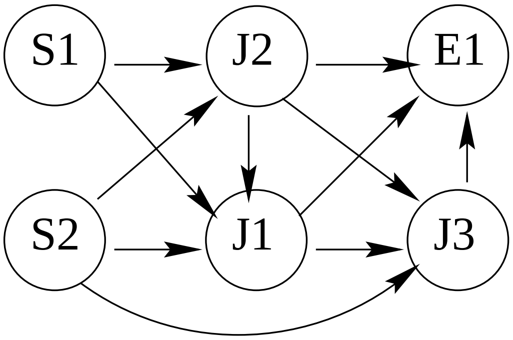
"A panel of three judges determines the outcome of presidential elections. Each judge can vote for one of the two possible candidates and the outcome is obtained by a majority rule. Two of the judges are impartial in the sense that they will listen to arguments from two spokespersons each working for one of the candidates while the remaining judge consistently pays attention to only one of the spokespersons. Each spokesperson will ask a judge to vote for a specific candidate. The spokespersons never talk nor listen to each other directly."

1. Identify the relevant variables based on the above description. For each variable state the possible values that it can take. If you use abbreviations to identify the variables make sure they are not ambiguous.

   The variables are:

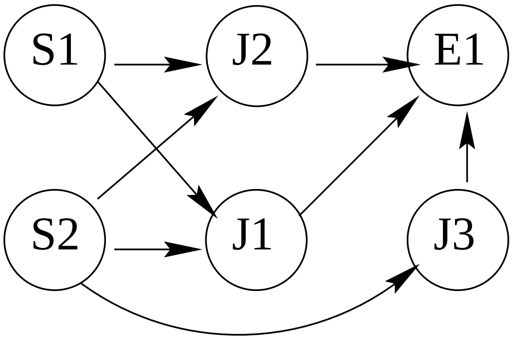
Election outcome	*E1*	$\{1, 2\}$	*(candidate 1 or 2)*
Spokesperson 1	*S1*	$\{1, 2\}$	*(argument has the effect of supporting 1 or 2)*
Spokesperson 2	*S2*	$\{1, 2\}$	*(argument has the effect of supporting 1 or 2)*
Judge 1 (impartial)	*J1*	$\{1, 2\}$	*(votes for 1 or 2)*
Judge 2 (impartial)	*J2*	$\{1, 2\}$	*(votes for 1 or 2)*
Judge 3 (partial)	*J3*	$\{1, 2\}$	*(votes for 1 or 2; listens to S2 only, say)*

2. Draw a Bayesian network that captures the interactions between the variables. *Avoid any assumptions that you cannot make on the basis of the above description.* Please indicate which variables correspond to which nodes.



The description does not tell us whether the judges discuss the case amongst themselves. We therefore cannot make any independence assumptions between the judges conditionally on the spokespersons. We therefore draw all possible arrows between the judges so long as the resulting graph is acyclic. *The directions of the arrows that connect the judges are irrelevant; they are all equivalent. You could also draw an undirected edge between each pair of judges but not bi-directional edges.*

3. The graph might change if the above description had started with "A panel of three *independent* judges...". If the graph would change, please draw the new graph. Otherwise state that there are no changes.



If the judges make their decisions independently of each other (but still contingent on the spokespersons), we simply remove all the arrows between the judges.

4. Explain under what circumstances (setting of some of the variables etc.) we might observe "explaining away" in the graph you just drew. If none exists, briefly explain why not.

   (For your convience, here's a brief description of "explaining away": *When we have multiple possible causes for a single known effect, explaining away refers to the phenomenon where acquiring further evidence about the presence of one of the causes* *makes the other ones less likely.*)

There is a natural explaining away effect here. Suppose we know the outcome of the election (E1 = 1). This increases the probability that each judge individually voted for candidate 1. If we now learn that the first two judges voted for candidate 1 (i.e., J1 = 1 and J2 = 1), then there's no remaining evidence supporting that J3 voted for candidate 1 (since the election outcome is a majority vote). The additional evidence (J1 = 1 and J2 = 1) now fully explains the initial observation (E1 = 1).

5. **(T/F – 2 points)** The graph structure is useful *only if* it captures all the independence properties present in the underlying probability distribution

   F

   Graph structure is useful if the properties that we can derive from the graph are true for the underlying probability distribution. It is often the case that we cannot capture all the independence properties with a graph.

6. **(T/F – 2 points)** Given any probability distribution, we can find a Bayesian network as well as a Markov random field that is consistent with the distribution

   T

   A fully connected undirected graph (or its directed quivalent) is consistent with any distribution as it makes no independence assumptions whatsoever.

7. **(T/F – 2 points)** A Boltzmann machine where *all* the variables are observable can *only* capture second order statistics (means and covariances) between the variables

   T

   When all the variables are observed, Boltzmann machines care only about the second order statistics (recall the estimation equations in the lecture notes). This is no longer true if there are unobserved variables as such variables can correlate more than two observed variables (this is analogous to the case of one underlying but unknown cause and multiple effects).