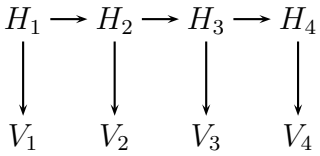
STAT 535 Homework 5
Out November 1, 2011
Due November 8 ,2011
©Marina Meilă
mmp@stat.washington.edu

Problem 1 - HMMs and variable elimination- Warmup, NOT GRADED



Consider the Hidden Markov Model (HMM) above. This problem studies some particular variable elimination orderings for the HMM. You are required to use the Bayes Net/conditional probabilities factorization of the HMM in this problem. Write the factorization of $P_{H_{1:4},V_{1:4}}$ as a product of conditional probability tables.

Variables $V_{1:4}$ are observed, i.e $V_1 = 0$, $V_2 = 1$, $V_3 = 1$, $V_4 = 0$. The elimination ordering is $\pi = V_1, H_1, V_2, H_2, V_3, H_3, V_4$.

a. In the table below (next page), fill in the potentials eliminated, the potentials created, and the sizes of the new potentials for each step of the elimination. Assume all variables are binary.

The table assumes that you use the algorithm in the notes, which reduces the potentials for an observed variable at the time of the elimination. But you can also use the algorithm given in class, where all potentials containing the observed variables are reduced before the elimination begins. If you do so, please write that clearly and explain what happens.

b. What (conditional) probability represents the potential created after the elimination of V_1?

c. Same question for the potential created after the elimination of V_2, V_3, V_4.

d. Denote by P' the product of the potentials at the end of the elimination, i.e $P' = \prod_{\phi \in \Phi} \phi$, where Φ is the set of potentials at the end of the elimination. What probability distribution represents P'? (no proof required)

e. Normalize P' to sum to 1 and denote by P'' the result. What probability distribution is P'' ? (no proof required)

f. What probabilities represent the potentials created when H_1, H_2, H_3 are eliminated? What algorithm does this elimination remind you of?

Variable eliminated	**Potentials eliminated**	**New potential**	**Size of new ϕ**	**[Optional New factorization/ graph]**
V_1:				
H_1:				
V_2:				
H_2:				
V_3:				
H_3:				
V_4:				

g. Now take a new elimination ordering $\pi = V_4, H_4, V_3, H_3, V_2, H_2, V_1$.

You will answer the same questions as for the previous ordering. Filling a table is *optional* though. Fill as much of the table as you need to answer the questions.

h. What (conditional) probability represents the potential created after the elimination of V_4?

i. Same question for the potential created after the elimination of V_3, V_2, V_1.

j. Denote by Q' the product of the potentials at the end of the elimination, i.e $Q' = \prod_{\phi \in \Phi} \phi$, where Φ is the set of potentials at the end of the elimination. What probability distribution represents Q'?

k. Normalize Q' to sum to 1 and denote by Q'' the result. What probability distribution is Q'' ?

More advanced topics
Many algorithms for manipulating graphical models with particular structure can be re-cast as variable elimination with a particular elimination order. Consider the **Forward-Backward algorithm** as applied to an HMM with known values for all observation variables, to calculate posterior marginal probabilities for each hidden state variable. The F-B algorithm is the name of the recursive computation of α_i^y (forward probabilities) and β_i^y (backward probabilities) described on page 9 in the notes. (You can also look it up in the Jordan chapter).

l. Explain how to use variable elimination and its intermediate results to arrive at the α_i^y, β_i^y and $\gamma_i^y = P(q_i|y)$ variables of the Forward-Backward algorithm.

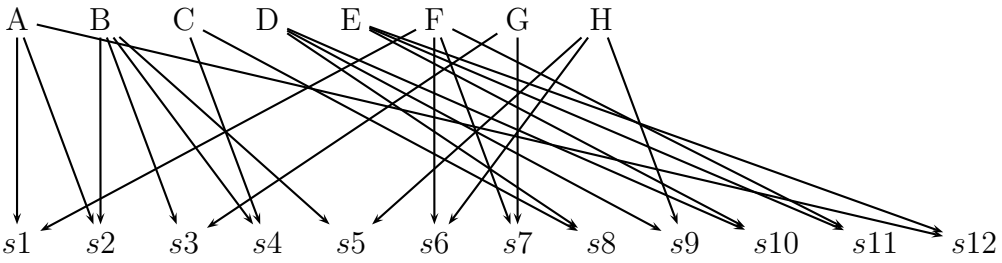
m. In a sentence or two, what computational deficiency of VE does this highlight? (Why is Forward-Backward used for computing the posterior marginal probabilities $P(q_i|y)$, $i = 1 : n$ in HMMs and not the VE algorithm? You need only consider the standard VE algorithm as presented in the notes or in class; it is possible that "clever" adaptations of VE can become as efficient as Forward-Backward)

Problem 2 - Looking inside the box

The (famous) QMR-DT graphical model is a probability distribution over diseases and symptoms created using expert knowledge, used for medical diagnosis; the basic operation of interest is to calculate the posterior probability of particular diseases given measurements of a subset of symptoms. Diseases and symptoms are both binary variables. The model assumes:

- diseases are independent ($D_i \perp D_j$)
- symptoms are conditionally independent given diseases ($s_i \perp s_j|D$)

Graphically, therefore, QMR-DT is bipartite and looks something like this (where $A, B, C, \ldots$ are the diseases and s_j are the symptoms):



Each disease D_i has a known marginal distribution P_{D_i}, and each symptom s_j a conditional distribution given a subset of diseases $P_{s_j|\text{pa}(s_j)}$. The conditional distribution of a symptom given its parent diseases takes a particularly simple form known as a "noisy-OR",which has one parameter q_{ij} per symptom-disease pair:

$$\begin{aligned} P(s_i = 0|D_1 \ldots D_n) &= \prod_{j=1\ldots n} q_{ij}^{D_j} \\ P(s_i = 1|D_1 \ldots D_n) &= 1 - P(s_i = 0|D_1 \ldots D_n) \end{aligned}$$

Notice that $P(s_i = 0|D_1 = 0 \ldots D_n = 0) = 1$: symptoms are never present in absence of disease.[1] If exactly one disease D_j is present ($D_j = 1$,$D_{k\neq j} = 0$), then $P(s_i = 0) = q_{ij}$. Thus q_{ij} is, roughly, the probability that disease j fails to provoke symptom i and $(1 - q_{ij})$ is the probability of disease j provoking symptom i independently of other diseases.

If symptom s_1 has two parents A and F, and $q_{A,1} = .1, q_{F,1} = .2$ then

$$\begin{aligned} P(s_1 = 0|A = 0, F = 0) &= 1 \\ P(s_1 = 0|A = 1, F = 0) &= .1 = q_{A,1} \\ P(s_1 = 0|A = 0, F = 1) &= .2 = q_{F,1} \\ P(s_1 = 0|A = 1, F = 1) &= .1 \times .2 = .02 \end{aligned}$$

The noisy-OR parameterization is very convenient for this medical domain: it has few parameters and if diseases are rare and independent (by assumption), so that in most cases at most one parent disease is active for a given symptom, it is easy to estimate q_{ij}.

This problem investigates how the variable elimination (VE) procedure for undirected graphs can be dramatically improved by using knowledge of the form of potentials. It's a lesson that algorithms that rely only on conditional independencies are often much weaker than those tailored to numeric properties of distributions.

For questions **a, b, c** you are required to use the Bayes Net/conditional probabilities factorization of the joint distribution (and we will show that the potentials created will be smaller than for the moralized and triangulated MRF factorization). For questions **d, e** you will replace the conditional probability tables $P_{X|Y}$ with the Noisy-OR parametrizations described here. You will find that additional savings in computation are possible for this particular case.

a. For the disease-symptom graph given above the values of s_1, s_2, s_5, s_6 have been observed. Graphically execute the VE procedure to calculate $P(A|s_1, s_2, s_5, s_6)$ using elimination order $s_{12}, s_{11}, s_{10}, s_9, s_8, s_7, s_6, s_5, \ldots, s_1, C, D, E, G, H, F, B$. For this question, ignore the specific noisy-OR form of the probability tables, i.e work with literals of the form $P_{s_i|XY}$ (conditional probability tables).

That is,

- Write the expression of P_V in the Bayes Net representation.
- Then, for each variable eliminated, list: the variable, the edges appearing, the new potential formed, the eliminated potentials. Find which of the potential formed are equal to 1. These potentials do not need to be carried over to the next step.

[1] The possibility of a "false positive" can be accounted for by introducing a fictitious disease, always present.

- Draw the graph over the diseases only (no symptoms) again, with all the new edges from moralization and elimination (let this graph be called $\mathcal{G}^{VE}$). Draw also a simplified graph (over diseases, without the symptoms), where no edges are added if the resulting potential equals 1 (let this graph be called $\mathcal{G}^{VEs}$.
- Triangulate the original graph by the Tarjan Elimination algorithm with the elimination order above (eliminate A last). Draw the resulting triangulated graph $\mathcal{G}^T$ (again, ignoring the symptom nodes and their edges).

  Compare $\mathcal{G}^T$, with $\mathcal{G}^{VE}$ and $\mathcal{G}^{VEs}$ the full and the simplified graphs obtained above. (No more than 1-2 sentences expected).

What is the largest potential during the elimination?

b. Consider the elimination of the unobserved symptom variables. In each case, what special numeric form does the new potential over neighbors have and what optimization to VE does it allow? Give a general rule for optimizing VE of unobserved variables in directed graphs.

c. Consider the elimination of the disease variables with unobserved symptoms. What special form does this potential have and what optimization of VE deoes it allow. Give a general rule.

d. Assume that the observations are $s_1 = 0, s_2 = 1, s_5 = 0, s_6 = 0$. Also denote $P_X(1) = p_X, P_X(0) = q_X$ for $X = A, B, C \ldots$. Calculate the expression of the intermediate potentials obtained in question **a** in terms of the q's and p's, and obtain the expression of $P(A, \text{observations})$ in terms of these parameters. [Hint: look at the next question]

e. Consider the elimination (in the noisy-OR model) of observed symptoms with value 0. What special numeric form does the new potential over neighbors have and what optimization to VE does it allow? Can the same simplification be performed if the symptom has value 1?

Problem 3 – Trees as Junction Trees

Consider a connected junction tree with maximal clique size 2. We called this a *tree graphical model* (or a **tree MRF**). Denote the set of edges of the tree by $\mathcal{E}$.

a. Prove that the P_V in the junction tree factorization is expressed as

$$P_V = \frac{\prod_{XY \in \mathcal{E}} P_{XY}}{\prod_{X \in V} P_X^{\deg X - 1}}$$

where $\deg X$ is the *degree* of node X, i.e. the number of edges incident to X.

b. We observe a set of nodes E taking value e. Assume that all nodes in E are *leaves* of the tree (i.e. they have degree 1). Consider now the joint distribution of the remaining variables given $E = e$, i.e $P_{V \setminus E|E=e}$ (which is proportional to $P_{V \setminus E, E=e}$ as you recall). Prove that $P_{V \setminus E|E=e}$ admits a factorization as a spanning tree over $V \setminus E$.

c. Now assume that E is a single node which is not a leaf. Prove that $P_{V \setminus E|E=e}$ admits a factorization as a *forest* tree over $V \setminus E$ (a forest is a graph with no cycles that is not connected). How many connected components will the forest have?

[Problem 4 - Exploiting conditional independencies in VE– OPTIONAL, for extra credit]

We are interested in $P_{X|E=e}$ where $X \in V$ is a single variable. Assume now that there is another variable $Y \in V \setminus E$ that is conditionally independent of X given E. Intuitively, we should not need consider Y in any way when computing $P_{X|E=e}$ since Y is not relevant; however, the VE algorithm given in class will eliminate (sum over) the values of Y.

Write a modified VE algorithm that does not perform "useless" calculations involving variables Y for which $X \perp Y \,|\, E$. Give a proof that your algorithm is correct.

Notes: 1. There is a short proof, and the modification to the VE is a small one. 2. You can assume that you know when you start the elimination which are the variables Y, as Y can always be found from the graph.