Machine Learning for Data Science (CS 4786)

Lecture 21: Belief Propagation

The text in black outlines main ideas to retain from the lecture. The text in blue give a deeper understanding of how we "derive" or get to the algorithm or method. The text in red are mathematical details for those who are interested. But is not crucial for understanding the basic workings of the method.

1 Belief Propagation

In class we looked at variable elimination algorithm which was a fancy way of saying to compute marginal of some bunch of variables, we sum over all remaining variables (other than ones we are computing marginal of). We saw that order of summation mattered to a great extent in terms of determining efficiency of the method. Often one might need to compute multiple marginals or conditional probabilities and using variable elimination to compute these one by one is very time consuming. As an example, consider the case of HMM. We could have used variable elimination there and done this for each variable multiple times. But this would have been exhausting! Instead what we did there was a single backward pass a single forward pass and with that we were ready to do the multiple inferences in a stroke. In fact the forward backward algorithm was your first example of message passing.

Belief propagation is a method for performing inference in graphical models where one can think of each vertex in the graph as a node in a network. Nodes send messages to neighboring nodes passing their beliefs along. The idea is that once the nodes have passed messages long enough, these messages hopefully converge and thus help us converge to the right answer that can be used for inference.

To this end, here are the rules of belief propagation game. Assume for simplicity that each nodes in our Bayesian network can take on K values. Now the messages each node passes to its neighbor will be a vector of length K.

1. Each node X_i has a so called evidence vector E_{X_i}. If a node X_i is unobserved, then $E_{X_i}(k) = 1$ for every $k \in \{1, \ldots, K\}$. On the other hand, if a node is observed, specifically if node X_i is observed to have value $X_i = x_i$, then, $E_{X_i}(x_i) = 1$ and $E_{X_i}(k) = 0$ for any $k \neq x_i$. Think of evidence as accounting for observations, if we don't observe anything, every value is equally likely, on the other hand, if a variable is observed, we set evidence to exactly indicate observation and set all other values to 0.

2. Messages a node sends to its children are belief about its own value. That is, the message that X_i sends to a child X_j is one that ranges over values of variable X_i.

3. Messages a node sends to its parent are belief about value of parent. That is, the message that X_j sends to a parent X_i is one that ranges over values of variable X_i.

Now the computations of messages on every round, is given as follows:

- Each message computed by each node are sum over product of three types of terms. First, the evidence of the node sending the message. Second, the conditional probability of that node given values of its parents. Finally, the third term is product of all messages from neighbors except the message received from the node we are computing the message for. Specifically we now provide the exact messages computed.
- If node X_i is a parent of X_j, then message at round t that X_i sends to X_j is computed as follows:

$$M^{(t)}_{X_i \mapsto X_j}(x_i) = \sum_{\{x_l : l \in \text{Parent}(X_i)\}} \left(E_{X_i}(x_i) \times P(X_i = x_i | \text{Parent}(X_i)) \times \prod_{k \in N(X_i) \setminus \{X_j\}} M^{(t-1)}_{X_k \mapsto X_i} \right)$$

- If node X_j is a child of X_i, then message at round t that X_j sends to X_i is computed as follows:

$$M^{(t)}_{X_i \mapsto X_j}(x_i) = \sum_{\{x_l : l \in \text{Parent}(X_i) \cup \{X_i\}\}} \left(E_{X_i}(x_i) \times P(X_i = x_i | \text{Parent}(X_i)) \times \prod_{k \in N(X_i) \setminus \{X_j\}} M^{(t-1)}_{X_k \mapsto X_i} \right)$$

The key thing to observe above is that the only portion of the message computation that changes on every round is the term involving product of messages received. Further, note that if all incoming messages to a node have converged on some round, then the message that this node sends out from that round onwards is the same (has converged).

To understand the procedure better, lets look at the following example: To shorten notation,

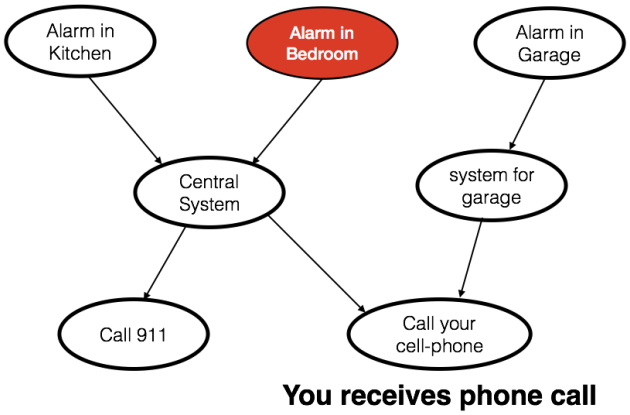

Figure 1: Mixture model

let us name variables as follows:

X_K: Alarm in Kitchen

X_B: Alarm in Bedroom

X_G: Alarm in Garage

X_{CS}: Central system alerted

X_{GS}: Garage system alerted

X_C: Cell phone call recieved

X_E: Emergency (911) call receieved

In the above example, all variables are binary having values 0 if switched off and 1 if switched on. The observation is that we have received a phone call on the cell phone. Hence we have the evidences as follows:
For node X_C, the evidence is given as $E_{X_C}(0) = 0$ and $E_{X_C}(1) = 1$. For all other nodes, the evidence is the vector $(1, 1)$.

Now we are ready to compute messages sent on various rounds. We will assume that all messages on round 0 are the all 1's vectors, that is $(1, 1)$. Now before we start working out the messages for each round let us simplify the computation a bit as follows: For any t,

$$\begin{aligned} M^{(t)}_{X_K \mapsto X_{CS}}(x_K) &= E_{X_K}(x_K) \times P(X_K = x_K) \\ &= P(X_K = x_k) \end{aligned}$$

$$\begin{aligned} M^{(t)}_{X_B \mapsto X_{CS}}(x_B) &= E_{X_B}(x_B) \times P(X_B = x_B) \\ &= P(X_B = x_B) \end{aligned}$$

$$\begin{aligned} M^{(t)}_{X_G \mapsto X_{GS}}(x_G) &= E_{X_G}(x_G) \times P(X_G = x_G) \\ &= P(X_G = x_G) \end{aligned}$$

$$\begin{aligned} M^{(t)}_{X_{GS} \mapsto X_C}(x_{GS}) &= \sum_{x_G} E_{X_{GS}}(x_{GS}) \times P(X_{GS} = x_{GS} | X_G = x_G) \times M^{(t-1)}_{X_G \mapsto X_{GS}}(x_G) \\ &= \sum_{x_G} P(X_{GS} = x_{GS} | X_G = x_G) \times M^{(t-1)}_{X_G \mapsto X_{GS}}(x_G) \end{aligned}$$

$$\begin{aligned} M^{(t)}_{X_{CS} \mapsto X_E}(x_{CS}) &= \sum_{x_B} \sum_{x_K} E_{X_{CS}}(x_{CS}) \times P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) \times M^{(t-1)}_{X_K \mapsto X_{CS}}(x_K) \times M^{(t-1)}_{X_B \mapsto X_{CS}}(x_B) \times M^{(t-1)}_{X_C \mapsto} \\ &= \sum_{x_B} \sum_{x_K} P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) \times M^{(t-1)}_{X_K \mapsto X_{CS}}(x_K) \times M^{(t-1)}_{X_B \mapsto X_{CS}}(x_B) \times M^{(t-1)}_{X_C \mapsto X_{CS}}(x_{CS}) \end{aligned}$$

$$\begin{aligned} M^{(t)}_{X_{CS} \mapsto X_C}(x_{CS}) &= \sum_{x_B} \sum_{x_K} E_{X_{CS}}(x_{CS}) \times P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) \times M^{(t-1)}_{X_K \mapsto X_{CS}}(x_K) \times M^{(t-1)}_{X_B \mapsto X_{CS}}(x_B) \times M^{(t-1)}_{X_E \mapsto} \\ &= \sum_{x_B} \sum_{x_K} P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) \times M^{(t-1)}_{X_K \mapsto X_{CS}}(x_K) \times M^{(t-1)}_{X_B \mapsto X_{CS}}(x_B) \times M^{(t-1)}_{X_E \mapsto X_{CS}}(x_{CS}) \end{aligned}$$

$$M^{(t)}_{X_E \mapsto X_{CS}}(x_{CS}) = \sum_{x_E} E_{X_E}(x_E) \times P(X_E = x_E | X_{CS} = x_{CS}) = \sum_{x_E} P(X_E = x_E | X_{CS} = x_{CS})$$
$$= 1$$

$$M^{(t)}_{X_C \mapsto X_{GS}}(x_{GS}) = \sum_{x_C} \sum_{x_{CS}} E_{X_C}(x_C) \times P(X_C = x_C | X_{CS} = x_{CS}, X_{GS} = x_{GS}) \times M^{(t-1)}_{X_{CS} \mapsto X_C}(x_{CS})$$
$$= \sum_{x_{CS}} P(X_C = 1 | X_{CS} = x_{CS}, X_{GS} = x_{GS}) \times M^{(t-1)}_{X_{CS} \mapsto X_C}(x_{CS})$$

$$M^{(t)}_{X_C \mapsto X_{CS}}(x_{CS}) = \sum_{x_C} \sum_{x_{GS}} E_{X_C}(x_C) \times P(X_C = x_C | X_{CS} = x_{CS}, X_{GS} = x_{GS}) \times M^{(t-1)}_{X_{GS} \mapsto X_C}(x_{GS})$$
$$= \sum_{x_{GS}} P(X_C = 1 | X_{CS} = x_{CS}, X_{GS} = x_{GS}) \times M^{(t-1)}_{X_{GS} \mapsto X_C}(x_{GS})$$

$$M^{(t)}_{X_{CS} \mapsto X_K}(x_K) = \sum_{x_B} \sum_{x_{CS}} E_{X_{CS}}(x_{CS}) P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) M^{(t-1)}_{X_E \mapsto X_{CS}}(x_{CS}) M^{(t-1)}_{X_C \mapsto X_{CS}}(x_{CS}) M^{(t-1)}_{X_B \mapsto X_{CS}}(x_B)$$
$$= \sum_{x_B} \sum_{x_{CS}} P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) M^{(t-1)}_{X_E \mapsto X_{CS}}(x_{CS}) M^{(t-1)}_{X_C \mapsto X_{CS}}(x_{CS}) M^{(t-1)}_{X_B \mapsto X_{CS}}(x_B)$$

$$M^{(t)}_{X_{CS} \mapsto X_B}(x_B) = \sum_{x_K} \sum_{x_{CS}} E_{X_{CS}}(x_{CS}) P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) M^{(t-1)}_{X_E \mapsto X_{CS}}(x_{CS}) M^{(t-1)}_{X_C \mapsto X_{CS}}(x_{CS}) M^{(t-1)}_{X_K \mapsto X_{CS}}(x_K)$$
$$= \sum_{x_K} \sum_{x_{CS}} P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) M^{(t-1)}_{X_E \mapsto X_{CS}}(x_{CS}) M^{(t-1)}_{X_C \mapsto X_{CS}}(x_{CS}) M^{(t-1)}_{X_K \mapsto X_{CS}}(x_K)$$

$$M^{(t)}_{X_{GS} \mapsto X_G}(x_G) = \sum_{x_{GS}} E_{X_{GS}}(x_{GS}) \times P(X_{GS} = x_{GS} | X_G = x_G) \times M^{(t-1)}_{X_C \mapsto X_{GS}}(x_{GS})$$
$$= \sum_{x_{GS}} P(X_{GS} = x_{GS} | X_G = x_G) \times M^{(t-1)}_{X_C \mapsto X_{GS}}(x_{GS})$$

Now all we need to do is, in each of the equations above substitute the messages from previous rounds to compute the new messages. To this end, note that,

Round $t = 1$: the following messages already converge.

$$M^{(1)}_{X_K \mapsto X_{CS}}(x_K) = P(X_K = x_K), \qquad M^{(1)}_{X_B \mapsto X_{CS}}(x_B) = P(X_B = x_B), \qquad M^{(1)}_{X_G \mapsto X_{GS}}(x_G) = P(X_G = x_G)$$
$$M^{(1)}_{X_E \mapsto X_{CS}}(x_{CS}) = 1$$

Round $t = 2$: In second round, given the messages converged in round 1, note that all but one incoming messages to nodes X_{CS} and X_{GS} have converged and so in this round the following

messages converge:

$$\begin{aligned}
M^{(2)}_{X_{CS}\mapsto X_C}(x_{CS}) &= \sum_{x_B}\sum_{x_K} P(X_{CS}=x_{CS}|X_K=x_K,X_B=x_B)\times M^{(1)}_{X_K\mapsto X_{CS}}(x_K)\times M^{(1)}_{X_B\mapsto X_{CS}}(x_B)\times M^{(1)}_{X_E\mapsto X_{CS}}(x_{CS}) \\
&= \sum_{x_B}\sum_{x_K} P(X_{CS}=x_{CS}|X_K=x_K,X_B=x_B)\times P(X_K=x_K)\times P(X_B=x_B)\times 1 \\
&= \sum_{x_B}\sum_{x_K} P(X_{CS}=x_{CS},X_K=x_K,X_B=x_B) \\
&= P(X_{CS}=x_{CS})
\end{aligned}$$

$$\begin{aligned}
M^{(2)}_{X_{GS}\mapsto X_C}(x_{GS}) &= \sum_{x_G} P(X_{GS}=x_{GS}|X_G=x_G)\times M^{(1)}_{X_G\mapsto X_{GS}}(x_G) = \sum_{x_G} P(X_{GS}=x_{GS}|X_G=x_G)\times P(X_G=x_G) \\
&= \sum_{x_G} P(X_{GS}=x_{GS},X_G=x_G) = P(X_{GS}=x_{GS})
\end{aligned}$$

Round $t=3$: Note that give the messages converged up to round 2, on round 3 we can show that messages sent from node X_C to X_{CS} and to X_{GS} will converge. Let us calculate these messages:

$$\begin{aligned}
M^{(3)}_{X_C\mapsto X_{CS}}(x_{CS}) &= \sum_{x_{GS}} P(X_C=1|X_{CS}=x_{CS},X_{GS}=x_{GS})\times M^{(2)}_{X_{GS}\mapsto X_C}(x_{GS}) \\
&= \sum_{x_{GS}} P(X_C=1|X_{CS}=x_{CS},X_{GS}=x_{GS})P(X_{GS}=x_{GS}) \\
&= \sum_{x_{GS}} P(X_C=1|X_{GS}=x_{GS},X_{CS}=x_{CS})\frac{1}{P(X_{CS}=x_{CS})}P(X_{GS}=x_{GS})P(X_{CS}=x_{CS})
\end{aligned}$$

where in the above we simply multiplied and divided by $P(X_{CS}=x_{CS})$. Next note that nodes X_{CS} and X_{GS} are marginally independent and so $P(X_{GS}=x_{GS})P(X_{CS}=x_{CS}) = P(X_{GS}=x_{GS},X_{CS}=x_{CS})$ and so,

$$\begin{aligned}
M^{(3)}_{X_C\mapsto X_{CS}}(x_{CS}) &= \sum_{x_{GS}} P(X_C=1|X_{GS}=x_{GS},X_{CS}=x_{CS})\frac{1}{P(X_{CS}=x_{CS})}P(X_{GS}=x_{GS})P(X_{CS}=x_{CS}) \\
&= \sum_{x_{GS}} P(X_C=1|X_{GS}=x_{GS},X_{CS}=x_{CS})\frac{1}{P(X_{CS}=x_{CS})}P(X_{GS}=x_{GS},X_{CS}=x_{CS}) \\
&= \sum_{x_{GS}} P(X_{GS}=x_{GS},X_{CS}=x_{CS},X_C=1)\frac{1}{P(X_{CS}=x_{CS})} \\
&= P(X_{CS}=x_{CS},X_C=1)\frac{1}{P(X_{CS}=x_{CS})} \\
&= P(X_C=1|X_{CS}=x_{CS})
\end{aligned}$$

Similarly note that:

$$\begin{aligned}
M^{(3)}_{X_C\mapsto X_{GS}}(x_{GS}) &= \sum_{x_{CS}} P(X_C=1|X_{CS}=x_{CS},X_{GS}=x_{GS})\times M^{(2)}_{X_{CS}\mapsto X_C}(x_{CS}) \\
&= \sum_{x_{CS}} P(X_C=1|X_{CS}=x_{CS},X_{GS}=x_{GS})P(X_{CS}=x_{CS}) \\
&= \sum_{x_{CS}} P(X_C=1|X_{GS}=x_{GS},X_{CS}=x_{CS})\frac{1}{P(X_{GS}=x_{GS})}P(X_{GS}=x_{GS})P(X_{CS}=x_{CS})
\end{aligned}$$

Again $P(X_{GS} = x_{GS})P(X_{CS} = x_{CS}) = P(X_{GS} = x_{GS}, X_{CS} = x_{CS})$ and so,

$$\begin{aligned}
M^{(3)}_{X_C \mapsto X_{GS}}(x_{GS}) &= \sum_{x_{CS}} P(X_C = 1 | X_{GS} = x_{GS}, X_{CS} = x_{CS}) \frac{1}{P(X_{GS} = x_{GS})} P(X_{GS} = x_{GS}, X_{CS} = x_{CS}) \\
&= \sum_{x_{CS}} P(X_{GS} = x_{GS}, X_{CS} = x_{CS}, X_C = 1) \frac{1}{P(X_{GS} = x_{GS})} \\
&= P(X_{GS} = x_{GS}, X_C = 1) \frac{1}{P(X_{GS} = x_{GS})} \\
&= P(X_C = 1 | X_{GS} = x_{GS})
\end{aligned}$$

Round $t = 4$: Finally on round 4 all the messages converge. Specifically, we have,

$$\begin{aligned}
M^{(4)}_{X_{CS} \mapsto X_K}(x_K) &= \sum_{x_B} \sum_{x_{CS}} P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) M^{(3)}_{X_E \mapsto X_{CS}}(x_{CS}) M^{(3)}_{X_C \mapsto X_{CS}}(x_{CS}) M^{(3)}_{X_B \mapsto X_{CS}}(x_B) \\
&= \sum_{x_B} \sum_{x_{CS}} P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) \times 1 \times P(X_C = 1 | X_{CS} = x_{CS}) P(X_B = x_B) \\
&= \sum_{x_B} \sum_{x_{CS}} P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) P(X_C = 1 | X_{CS} = x_{CS}) P(X_B = x_B) P(X_K = x_K) \frac{1}{P(X_K = x_K)}
\end{aligned}$$

Now since X_K and X_B are marginally independent, we have that $P(X_B = x_B)P(X_K = x_K) = P(X_B = x_B, X_K = x_K)$ and so,

$$\begin{aligned}
M^{(4)}_{X_{CS} \mapsto X_K}(x_K) &= \sum_{x_B} \sum_{x_{CS}} P(X_{CS} = x_{CS} | X_K = x_K, X_B = x_B) P(X_C = 1 | X_{CS} = x_{CS}) P(X_B = x_B, X_K = x_K) \frac{1}{P(X_K = x_K)} \\
&= \sum_{x_B} \sum_{x_{CS}} P(X_{CS} = x_{CS}, X_K = x_K, X_B = x_B) P(X_C = 1 | X_{CS} = x_{CS}) \frac{1}{P(X_K = x_K)} \\
&= \sum_{x_{CS}} P(X_{CS} = x_{CS}, X_K = x_K) P(X_C = 1 | X_{CS} = x_{CS}) \frac{1}{P(X_K = x_K)} \\
&= \sum_{x_{CS}} P(X_{CS} = x_{CS}, X_K = x_K) P(X_C = 1 | X_{CS} = x_{CS}, X_K = x_K) \frac{1}{P(X_K = x_K)} \qquad \text{(local Markov)} \\
&= \sum_{x_{CS}} P(X_C = 1, X_{CS} = x_{CS}, X_K = x_K) \frac{1}{P(X_K = x_K)} \\
&= P(X_C = 1, X_K = x_K) \frac{1}{P(X_K = x_K)} \\
&= P(X_C = 1 | X_K = x_K)
\end{aligned}$$

Similarly we have that

$$\begin{aligned}
M^{(4)}_{X_{CS}\mapsto X_B}(x_B) &= \sum_{x_K}\sum_{x_{CS}} P(X_{CS}=x_{CS}|X_K=x_K, X_B=x_B) M^{(3)}_{X_E\mapsto X_{CS}}(x_{CS}) M^{(3)}_{X_C\mapsto X_{CS}}(x_{CS}) M^{(3)}_{X_K\mapsto X_{CS}}(x_K) \\
&= \sum_{x_K}\sum_{x_{CS}} P(X_{CS}=x_{CS}|X_K=x_K, X_B=x_B) P(X_C=1|X_{CS}=x_{CS}) P(X_K=x_K) \\
&= \sum_{x_K}\sum_{x_{CS}} P(X_{CS}=x_{CS}|X_K=x_K, X_B=x_B) P(X_C=1|X_{CS}=x_{CS}) P(X_K=x_K) P(X_B=x_B) \frac{1}{P(X_B=x_B}} \\
&= \sum_{x_K}\sum_{x_{CS}} P(X_{CS}=x_{CS}|X_K=x_K, X_B=x_B) P(X_C=1|X_{CS}=x_{CS}) P(X_K=x_K, X_B=x_B) \frac{1}{P(X_B=x_B)} \\
&= \sum_{x_{CS}} P(X_{CS}=x_{CS}, X_B=x_B) P(X_C=1|X_{CS}=x_{CS}) \frac{1}{P(X_B=x_B)} \\
&= \sum_{x_{CS}} P(X_C=1, X_{CS}=x_{CS}, X_B=x_B) \frac{1}{P(X_B=x_B)} \\
&= P(X_C=1, X_B=x_B) \frac{1}{P(X_B=x_B)} \\
&= P(X_C=1|X_B=x_B)
\end{aligned}$$

Finally note that,

$$\begin{aligned}
M^{(4)}_{X_{GS}\mapsto X_G}(x_G) &= \sum_{x_{GS}} P(X_{GS}=x_{GS}|X_G=x_G) M^{(3)}_{X_C\mapsto X_{GS}}(x_{GS}) \\
&= \sum_{x_{GS}} P(X_{GS}=x_{GS}|X_G=x_G) P(X_C=1|X_{GS}=x_{GS}) \\
&= \sum_{x_{GS}} P(X_{GS}=x_{GS}|X_G=x_G) P(X_C=1|X_{GS}=x_{GS}, X_G=x_G) \qquad \text{(local markov)} \\
&= \sum_{x_{GS}} P(X_C=1, X_{GS}=x_{GS}|X_G=x_G) \\
&= P(X_C=1|X_G=x_G)
\end{aligned}$$

Thus on the 4th round all messages converge. More generally this algorithm is guaranteed to converge on any tree and takes twice the diameter number of rounds to converge.

Now for instance to finally compute

$$\begin{aligned}
P(X_K=x_K|X_C=1) &\propto E_{X_K}(x_k) \times P(X_K=x_k) \times \text{product} - \text{of} - \text{messages} \\
&= E_{X_K}(x_K) \times P(X_K=x_k) \times M^{(4)}_{X_{CS}\mapsto X_K}(x_K) \\
&= P(X_K=x_k) P(X_C=1|X_K=x_K) \\
&= P(X_C=1, X_K=x_K)
\end{aligned}$$

Clearly it is true that $P(X_K=x_K|X_C=1) \propto P(X_C=1, X_K=x_K)$ and hence normalizing we can indeed complete inference of $P(X_K=x_K|X_C=1)$.