INFEREN CE IN BAYES I AN NETWORKS
- BELIEF PROPAGATION

Instructor: Sael Lee

Combination of slides:
“Pearl’s algorithm” by Tomas Singliar & Daniel Lowd’s slide for UW CSE 573 & “Belief Propagation” by Jakob Metzler & “Generalized BP” by Jonathan Yedidia
OUTLINE

- Motivation
- Pearl’s BP Algorithm
- Generalized Belief Propagation
Computing the a posteriori belief of a variable in a general Bayesian Network is NP-hard

- Approximate inference
  - MCMC sampling
  - Belief Propagation
BELIEF PROPAGATION

BP is a message passing algorithm that solves approximate inference problems in graphical model, including Bayesian networks and Markov random fields.

Calculates marginal distribution for each of the unobserved variable, conditional on any observed variables.

It was first proposed by Judea Pearl in 1982 for trees (exact) and later extended to polytrees and general graphs (approximate).
(G, P) directed acyclic graph with the joint p.d. P
each node is a variable of a multivariate distribution
links represent causal dependencies
  + CPT in each node
Polytree
  + What is a polytree?
    × A Bayesian network graph is a polytree if (an only if) there is at most one path between any two nodes, \( V_i \) and \( V_k \)
    × implies each node separates the graph into two disjoint components
  + Why do we care about polytrees?
    × Exact BN inference is NP-hard...
    × ...but on polytrees, takes linear time.
EXAMPLES: POLYTREE OR NOT?

\[ \begin{align*}
V_1 & \rightarrow V_3 & V_2 & \rightarrow V_4 \\
V_3 & \rightarrow V_5 & V_4 & \rightarrow V_6 \\
V_5 & \rightarrow V_6 & V_2 & \rightarrow V_4 \\
V_3 & \rightarrow V_5 & V_6 & \rightarrow V_6 \\
V_4 & \rightarrow V_3 & V_4 & \rightarrow V_2 \\
V_3 & \rightarrow V_2 & V_4 & \rightarrow V_4
\end{align*} \]
We know the values of some evidence variables $E$: $V_{e_1}, \ldots, V_{e_{|E|}}$

We wish to compute the posterior probability $P(X_i \mid E)$ for all non-evidence variables $X_i$. 
PEARL’S BELIEF PROPAGATION

- We have the evidence $E$
- Local computation for one node $V$ desired
- Information flows through the paths of $G$
  - flows as messages of two types – $\lambda$ and $\pi$
- $V$ splits network into two disjoint parts
  - Strong independence assumptions induced – crucial!
- Denote $E_V^+$ the part of evidence accessible through the parents of $V$ (causal)
  - passed downward in $\pi$ messages
- Analogously, let $E_V^-$ be the diagnostic evidence
  - passed upwards in $\lambda$ messages
THE Π MESSAGES

- What are the messages?
- For simplicity, let the nodes be binary

| V₁=T | 0.8 |
| V₁=F | 0.2 |

The message passes on information.

What information? Observe:

\[
P(V₂) = P(V₂| V₁=T)P(V₁=T) + P(V₂| V₁=F)P(V₁=F)
\]

The information needed is the CPT of \( V₁ = \pi_V(V₁) \)

\( \pi \) Messages capture information passed from parent to child
Evidence – values of observed nodes
+ $V_3 = T, V_6 = 3$

Our belief in what the value of $V_i$ ‘should’ be changes.

This belief is propagated

As if the CPTs became

<table>
<thead>
<tr>
<th>$V_3$</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_3$</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P$</th>
<th>$V_2 = T$</th>
<th>$V_2 = F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_6 = 1$</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$V_6 = 2$</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>$V_6 = 3$</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
THE λ MESSAGES

- We know what the π messages are
- What about λ?

Assume $E = \{ V_2 \}$ and compute by Bayes rule:

$$P(V_1 | V_2) = \frac{P(V_1)P(V_2 | V_1)}{P(V_2)} = \alpha P(V_1)P(V_2 | V_1)$$

The information not available at $V_1$ is the $P(V_2 | V_1)$. To be passed upwards by a λ-message. Again, this is not in general exactly the CPT, but the belief based on evidence down the tree.

- The messages are $\pi(V) = P(V | E^+)$ and $\lambda(V) = P(E^- | V)$
**COMBINATION OF EVIDENCE**

- Let $E_V = E_V^+ \cup E_V^-$ and let us compute

  
  $$P(V | E) = P(V | E_V^+, E_V^-) = \alpha' P(E_V^+, E_V^- | V)P(V) = \alpha' P(E^-_V | V)P(E^+_V | V)P(V) = \alpha P(E^-_V | V)P(V | E^+_V) = \alpha \lambda(V) \pi(V) = \text{BEL}(V)$$

- $\alpha$ is the normalization constant
- normalization is not necessary (can do it at the end)
- but may prevent numerical underflow problems
MESSAGES

- Assume $X$ received $\lambda$-messages from neighbors
- How to compute $\lambda(X) = p(E^- | X)$?
- Let $Y_1, \ldots, Y_c$ be the children of $X$
- $\lambda_{XY}(x)$ denotes the $\lambda$-message sent between $X$ and $Y$

$$\lambda(X) = \prod_{j=1}^{c} \lambda_{Y_jX}(X)$$
• Assume X received $\pi$-messages from neighbors
• How to compute $\pi(X) = p(X | E^+)$?
• Let $U_1, ..., U_p$ be the parents of $X$
• $\pi_{XY}(x)$ denotes the $\pi$-message sent between $X$ and $Y$
• summation over the CPT

$$\pi(X) = \sum_{u_1, ..., u_p} P(X | U_1, ..., U_p) \prod_{j=1}^p \pi_{U_jX}(U_j)$$
MESSAGES TO PASS

- We need to compute $\pi_{XY}(x)$

$$
\pi_{XY}(x) = \alpha \pi_X(x) \prod_{k \neq j} \lambda_{Y_kX}(x)
$$

- Similarly, $\lambda_{XY}(x)$, $X$ is parent, $Y$ child

- Symbolically, group other parents of $Y$ into $V = V_1, \ldots, V_q$

$$
\lambda_{Y_jX}(x) = \sum_{y_j} \lambda_{Y_j}(y_j) \sum_{v_1, \ldots, v_q} p(y \mid v_1, \ldots, v_q) \prod_{k=1}^{q} \pi_{V_kY_j}(v_k)
$$
PEARL’S BP ALGORITHM

- Initialization
  - For nodes with evidence $e$
    - $\lambda(x_i) = 1$ wherever $x_i = e_i$; 0 otherwise
    - $\pi(x_i) = 1$ wherever $x_i = e_i$; 0 otherwise
  - For nodes without parents
    - $\pi(x_i) = p(x_i)$ - prior probabilities
  - For nodes without children
    - $\lambda(x_i) = 1$ uniformly (normalize at end)
THE PEARL BELIEF PROPAGATION ALGORITHM

- Iterate until no change occurs
  - (For each node $X$) if $X$ has received all the $\pi$ messages from its parents, calculate $\pi(x)$
  - (For each node $X$) if $X$ has received all the $\lambda$ messages from its children, calculate $\lambda(x)$
  - (For each node $X$) if $\pi(x)$ has been calculated and $X$ received all the $\lambda$-messages from all its children (except $Y$), calculate $\pi_{XY}(x)$ and send it to $Y$.
  - (For each node $X$) if $\lambda(x)$ has been calculated and $X$ received all the $\pi$-messages from all parents (except $U$), calculate $\lambda_{XU}(x)$ and send it to $U$.

- Compute Belief $\text{BEL}(X) = \lambda(x)\pi(x)$
- and normalize
PROPERTIES OF BP

- Exact for polytrees
  - Each node separates Graph into 2 disjoint components
- On a polytree, the BP algorithm converges in time proportional to diameter of network – at most linear
- Work done in a node is proportional to the size of CPT
  - Hence BP is linear in number of network parameters
- For general BBNs
  - Exact inference is NP-hard
  - Approximate inference is NP-hard
Most graphs are not polytrees
  + Cutset conditioning
  + Clustering
    - Join Tree Method
  + Approximate Inference
    - Loopy BP
LOOPY BELIEF PROPAGATION

- If BP is used on graphs with loops, messages may circulate indefinitely
- Empirically, a good approximation is still achievable
  - Stop after fixed # of iterations
  - Stop when no significant change in beliefs
  - If solution is not oscillatory but converges, it usually is a good approximation
LOOPY BELIEF PROPAGATION

- Just apply BP rules in spite of loops
- In each iteration, each node sends all messages in parallel
- Seems to work for some applications
TROUBLE WITH LBP

- May not converge
  + A variety of tricks can help
- Cycling Error – old information is mistaken as new
- Convergence Error – unlike in a tree, neighbors need not be independent. However, LBP treats them as if they were.

We can try to improve inference by taking into account higher-order interactions among the variables.

An intuitive way to do this is to define messages that propagate between groups of nodes rather than just single nodes.

This is the intuition in Generalized Belief Propagation (GPB).
GBP ALGORITHM

1) Split the graph into basic clusters

[1245], [2356],
[4578], [5689]
GBP ALGORITHM

2) Find all intersection regions of the basic clusters, and all their intersections
[25], [45], [56], [58], [5]
3) Create a hierarchy of regions and their direct sub-regions
GBP ALGORITHM

4) Associate a message with each line in the graph
e.g. message from
[1245]→[25]:
\( m_{14→25}(x_2, x_5) \)
5) Setup equations for beliefs of regions
   - remember from earlier:
     \[ b_i(x_i) = k \phi_i(x_i) \prod_{j \in N(i)} m_{ji}(x_i) \]
   - So the belief for the region containing [5] is:
     - for the \( b_5 = k \left[ \phi_5 \right] \left[ m_{2\rightarrow 5} m_{4\rightarrow 5} m_{6\rightarrow 5} m_{8\rightarrow 5} \right] \)
   - etc.
     \[ b_{45} = k \left[ \phi_4 \phi_5 \phi_{45} \right] \left[ m_{12\rightarrow 45} m_{78\rightarrow 45} m_{2\rightarrow 5} m_{6\rightarrow 5} m_{8\rightarrow 5} \right] \]
Generalized Belief Propagation
Generalized Belief Propagation

\[ b_5 \propto m_{2 \rightarrow 5} m_{4 \rightarrow 5} m_{6 \rightarrow 5} m_{8 \rightarrow 5} \]
Generalized Belief Propagation

\[ b_{45} \propto \left[ f_{45} \right] \left[ m_{12 \rightarrow 45} m_{78 \rightarrow 45} m_{2 \rightarrow 5} m_{6 \rightarrow 5} m_{8 \rightarrow 5} \right] \]
Generalized Belief Propagation

\[ b_{1245} \propto \prod_{f_{12}, f_{14}, f_{25}, f_{45}} \prod_{m_{36 \to 25}, m_{78 \to 45}, m_{6 \to 5}, m_{8 \to 5}} \]
Generalized Belief Propagation

Use Marginalization Constraints to Derive Message-Update Rules

\[ b_5(x_5) = \sum_{x_4} b_{45}(x_4, x_5) \]
Generalized Belief Propagation

Use Marginalization Constraints to Derive Message-Update Rules

\[ b_5(x_5) = \sum_{x_4} b_{45}(x_4, x_5) \]
Generalized Belief Propagation

Use Marginalization Constraints to Derive Message-Update Rules

\[ b_5(x_5) = \sum_{x_4} b_{45}(x_4, x_5) \]
Generalized Belief Propagation

Use Marginalization Constraints to Derive Message-Update Rules

\[ m_{4 \to 5}(x_5) \propto \sum_{x_4} f_{45}(x_4, x_5) m_{12 \to 45}(x_4, x_5) m_{78 \to 45}(x_4, x_5) \]
6) Setup equations for updating messages by enforcing marginalization conditions and combining them with the belief equations:

\[ b_5(x_5) = \sum_{x_4} b_{45}(x_4, x_5) \]

yields, with the previous two belief formulas, the message update rule

\[ m_{4\rightarrow 5}(x_5) \leftarrow k \sum_{42} \phi_4(x_4) \psi_{45}(x_4, x_5) m_{12\rightarrow 45}(x_4, x_5) m_{78\rightarrow 25}(x_2, x_5) \]
REFERENCES

  × Derivations shown in class are from this book, except that we worked with $\pi$ instead of $\rho$ messages. They are related by factor of $p(e^+)$. 
+ www.cs.kun.nl/~peterl/teaching/CS45CI/bbn3-4.ps.gz
+ reason.cs.uiuc.edu/eyal/classes/.../lec18-BeliefPropagation.ppt
+ www.cs.pitt.edu/~tomas/cs3750/pearl.ppt