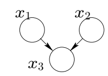
Machine learning: lecture 22

Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Review: graphs and probabilities

- Directed graphical models (Bayesian networks)



$\Rightarrow$ d-sep. $\Rightarrow$ indep. $\Rightarrow$ assoc. distribution

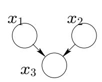
x_1 d-sep. from x_2 given nothing

x_1 indep. from x_2 given nothing

$P(x_1) \cdot P(x_2) \cdot P(x_3|x_1, x_2)$

Review: graphs and probabilities

- Directed graphical models (Bayesian networks)



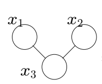
$\Rightarrow$ d-sep. $\Rightarrow$ indep. $\Rightarrow$ assoc. distribution

x_1 d-sep. from x_2 given nothing

x_1 indep. from x_2 given nothing

$P(x_1) \cdot P(x_2) \cdot P(x_3|x_1, x_2)$

- Undirected graphical models (Markov Random Fields)



$\Rightarrow$ separation $\Rightarrow$ indep. $\Rightarrow$ assoc. distribution

x_1 separated from x_2 given x_3

x_1 indep. from x_2 given x_3

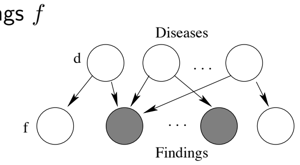
$\frac{1}{Z} \cdot \psi_{13}(x_1, x_3) \cdot \psi_{23}(x_2, x_3)$

Outline

- Quantitative probabilistic inference
  - diagnosis example
  - marginalization and message passing
  - cliques, clique trees, and the junction tree algorithm

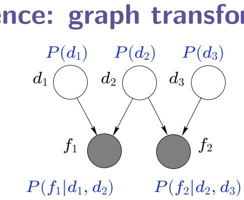
Three inference problems

- Medical diagnosis example: binary disease variables d and possible findings f

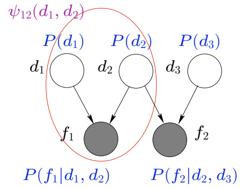


1. What are the marginal posterior probabilities over the diseases given observed findings $f^* = \{f_1^*, \ldots, f_k^*\}$?
2. What is the most likely setting of all the underlying disease variables?
3. Which test should we carry out next in order to get the most information about the diseases?

Inference: graph transformation

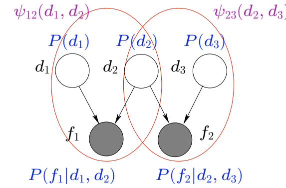


Inference: graph transformation



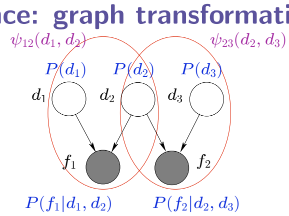
$$\psi_{12}(d_1, d_2) = P(d_1)P(d_2)P(f_1^*|d_1, d_2)$$

Inference: graph transformation



$$\begin{aligned}\psi_{12}(d_1, d_2) &= P(d_1)P(d_2)P(f_1^*|d_1, d_2)\\ \psi_{23}(d_2, d_3) &= P(d_3)P(f_2^*|d_2, d_3)\end{aligned}$$

Inference: graph transformation



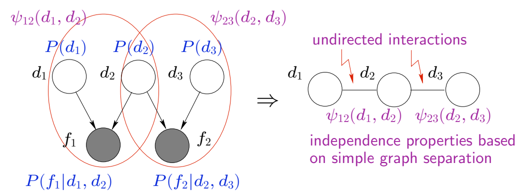
$$\begin{aligned}\psi_{12}(d_1, d_2) &= P(d_1)P(d_2)P(f_1^*|d_1, d_2)\\ \psi_{23}(d_2, d_3) &= P(d_3)P(f_2^*|d_2, d_3)\end{aligned}$$

- Joint distribution as a product of "interaction potentials"

$$P(d_1, d_2, d_3, \text{data}) = \psi_{12}(d_1, d_2) \cdot \psi_{23}(d_2, d_3)$$

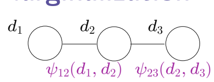
Inference: graph transformation

- We have transformed the Bayesian network into an undirected graph model (Markov random field):



$$P(d_1, d_2, d_3, \text{data}) = \psi_{12}(d_1, d_2) \cdot \psi_{23}(d_2, d_3)$$

Marginalization



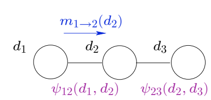
- It suffices to evaluate the following probabilities

$$\begin{aligned}P(d_1, \text{data}) &= \sum_{d_2, d_3} P(d_1, d_2, d_3, \text{data})\\ P(d_2, \text{data}) &= \sum_{d_1, d_3} P(d_1, d_2, d_3, \text{data})\\ P(d_3, \text{data}) &= \sum_{d_1, d_2} P(d_1, d_2, d_3, \text{data})\end{aligned}$$

These will readily yield the posterior probabilities of interest:

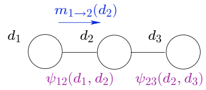
$$P(d_1|\text{data}) = P(d_1, \text{data}) / \sum_{d_1'} P(d_1', \text{data})$$

Marginalization and messages



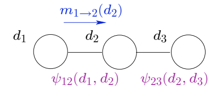
$$P(d_2, d_3, \text{data}) = \sum_{d_1} P(d_1, d_2, d_3, \text{data})$$

Marginalization and messages



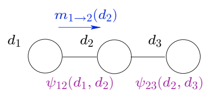
$$\begin{aligned}P(d_2,d_3,\mathsf{data}) &= \sum_{d_1} P(d_1,d_2,d_3,\mathsf{data})\\ &= \sum_{d_1} \psi_{12}(d_1,d_2)\cdot\psi_{23}(d_2,d_3)\end{aligned}$$

Marginalization and messages



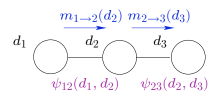
$$\begin{aligned}P(d_2,d_3,\mathsf{data}) &= \sum_{d_1} P(d_1,d_2,d_3,\mathsf{data})\\ &= \sum_{d_1} \psi_{12}(d_1,d_2)\cdot\psi_{23}(d_2,d_3)\\ &= \left[\sum_{d_1} \psi_{12}(d_1,d_2)\right]\cdot\psi_{23}(d_2,d_3)\end{aligned}$$

Marginalization and messages



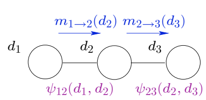
$$\begin{aligned}P(d_2,d_3,\mathsf{data}) &= \sum_{d_1} P(d_1,d_2,d_3,\mathsf{data})\\ &= \sum_{d_1} \psi_{12}(d_1,d_2)\cdot\psi_{23}(d_2,d_3)\\ &= \left[\sum_{d_1} \psi_{12}(d_1,d_2)\right]\cdot\psi_{23}(d_2,d_3)\\ &= m_{1\to2}(d_2)\cdot\psi_{23}(d_2,d_3)\end{aligned}$$

Marginalization and messages



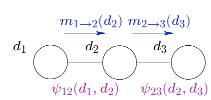
$$P(d_3,\mathsf{data}) = \sum_{d_2} P(d_2,d_3,\mathsf{data})$$

Marginalization and messages



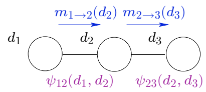
$$\begin{aligned}P(d_3,\mathsf{data}) &= \sum_{d_2} P(d_2,d_3,\mathsf{data})\\ &= \sum_{d_2} m_{1\to2}(d_2)\cdot\psi_{23}(d_2,d_3)\cdot 1\end{aligned}$$

Marginalization and messages



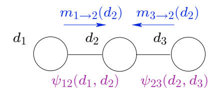
$$\begin{aligned}P(d_3,\mathsf{data}) &= \sum_{d_2} P(d_2,d_3,\mathsf{data})\\ &= \sum_{d_2} m_{1\to2}(d_2)\cdot\psi_{23}(d_2,d_3)\cdot 1\\ &= \left[\sum_{d_2} m_{1\to2}(d_2)\cdot\psi_{23}(d_2,d_3)\right]\cdot 1\end{aligned}$$

Marginalization and messages



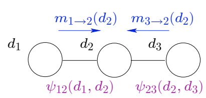
$$
\begin{aligned}
P(d_3, \text{data}) &= \sum_{d_2} P(d_2, d_3, \text{data}) \\
&= \sum_{d_2} m_{1\to2}(d_2) \cdot \psi_{23}(d_2, d_3) \cdot 1 \\
&= \left[\sum_{d_2} m_{1\to2}(d_2) \cdot \psi_{23}(d_2, d_3) \right] \cdot 1 \\
&= m_{2\to3}(d_3) \cdot 1
\end{aligned}
$$

Marginalization and messages



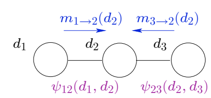
$$
P(d_2, \text{data}) = \sum_{d_3} P(d_2, d_3, \text{data})
$$

Marginalization and messages



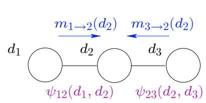
$$
\begin{aligned}
P(d_2, \text{data}) &= \sum_{d_3} P(d_2, d_3, \text{data}) \\
&= \sum_{d_3} m_{1\to2}(d_2) \cdot \psi_{23}(d_2, d_3)
\end{aligned}
$$

Marginalization and messages



$$
\begin{aligned}
P(d_2, \text{data}) &= \sum_{d_3} P(d_2, d_3, \text{data}) \\
&= \sum_{d_3} m_{1\to2}(d_2) \cdot \psi_{23}(d_2, d_3) \\
&= m_{1\to2}(d_2) \cdot \left[\sum_{d_3} \psi_{23}(d_2, d_3) \right]
\end{aligned}
$$

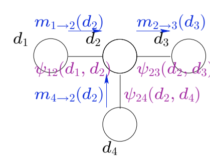
Marginalization and messages



$$
\begin{aligned}
P(d_2, \text{data}) &= \sum_{d_3} P(d_2, d_3, \text{data}) \\
&= \sum_{d_3} m_{1\to2}(d_2) \cdot \psi_{23}(d_2, d_3) \\
&= m_{1\to2}(d_2) \cdot \left[\sum_{d_3} \psi_{23}(d_2, d_3) \right] \\
&= m_{1\to2}(d_2) \cdot m_{3\to2}(d_2)
\end{aligned}
$$

Message passing and trees

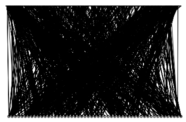
- The same message passing approach (belief propagation) works for any tree structured model



$$
\begin{aligned}
m_{2\to3}(d_3) &= \sum_{d_2} m_{1\to2}(d_2) m_{4\to2}(d_2) \psi_{23}(d_2, d_3) \\
P(d_2, \text{data}) &= \;?
\end{aligned}
$$

Back to the diagnosis problem

- This does not look like a tree...



Outline

- Quantitative probabilistic inference
  - diagnosis example
  - marginalization and message passing
  - cliques, clique trees, and the junction tree algorithm

Exact inference

- All exact inference algorithms for Bayesian networks perform essentially the same calculations but operate on different representations
- The junction tree algorithm is a simple message passing algorithm over *clusters of variables*

  Preliminary steps:

1. transform the Bayesian network into an undirected model via moralization ("marry parents")
2. triangulate the resulting undirected graph (add edges)
3. identify the cliques (clusters) of the resulting triangulated graph
4. construct the junction tree from the cliques