Machine learning: lecture 21

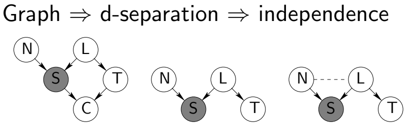
Tommi S. Jaakkola
MIT CSAIL
tommi@csail.mit.edu

Outline

- Bayesian networks cont'd
  - graphs and consistency
- Undirected graphical models (Markov random fields)
  - graphs, independence, consistency, associated distribution
  - Bayesian networks as undirected models
- Quantitative probabilistic inference
  - medical diagnosis example
  - basic algorithms and problems

Bayesian networks: review

- Graph ⇒ d-separation ⇒ independence



  - conditional independence properties provide the basis for qualitative inferences
- Graph ⇒ associated probability distribution

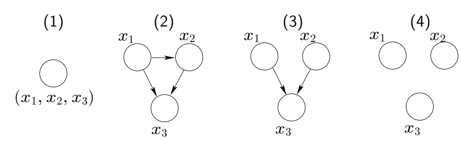
$$P(N)\,P(L)\,P(S|N,L)\,P(T|L)\,P(C|S,T)$$

(any distribution that factors in this manner is consistent with all the independence properties implied by the graph)

Graphs, probabilities, and consistency

- Suppose x_1, x_2, and x_3 represent three independent coin tosses so that the probability distribution can be written as a product $P(x_1)P(x_2)P(x_3)$

  This distribution is consistent with *all* the following graphs in the sense that all the independence properties we can infer from the graphs also hold for this distribution:



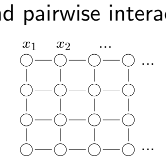
  Moreover, (1) and (2) are consistent with *any* distribution over x_1, x_2, and x_3

Outline

- Bayesian networks cont'd
  - graphs and consistency
- Undirected graphical models (Markov random fields)
  - graphs, independence, consistency, associated distribution
  - Bayesian networks as undirected models
- Quantitative probabilistic inference
  - medical diagnosis example
  - basic algorithms and problems

Undirected graphical models

- For example: a simple lattice model with binary variables $x_i \in \{1, 1\}$ (spins) and pairwise interactions (edges E)

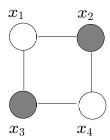


$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{(i,j)\in E} \exp(J_{ij} x_i x_j)$$

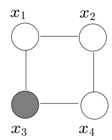
where J_{ij} specifies the "interaction strength" between nearby variables x_i and x_j.

Undirected graphical models: graph semantics

- Graph semantics of undirected graphical models comes from simple graph separation



x_1 and x_4 are independent given x_2 and x_3

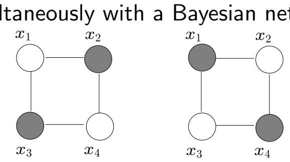


x_1 and x_4 are not independent given x_3

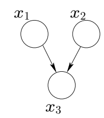
Graph semantics: comparison

- Directed and undirected graphs are complementary

The following two independence properties cannot be captured simultaneously with a Bayesian network:

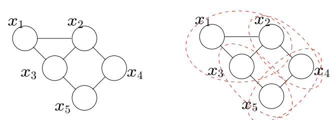


Marginal but not conditional independence cannot be captured with an undirected graph:



Undirected graphs: associated distribution

- The simple graph separation properties again impose independence (or *Markov*) properties on the associated distribution



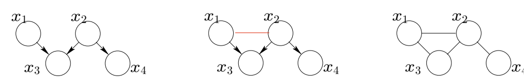
Theorem: (Hammersley-Clifford) Any distribution consistent with an undirected graph has to factor according to the (maximal) cliques in the graph

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c)$$

where $\mathbf{x}_c$ denotes the variables in clique c.

Graph transformations

- We can transform directed graphical models (Bayesian networks) into undirected graphical models simply via moralization



$P(x_1)P(x_2)P(x_3|x_1,x_2)P(x_4|x_2)$ $\quad$ $[P(x_1)P(x_2)P(x_3|x_1,x_2)] \cdot [P(x_4|x_2)]$

(only the graph representation changes, not the distribution)

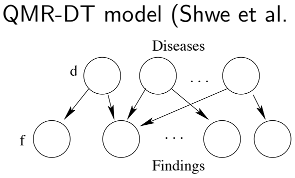
- The resulting undirected graph will be consistent with the distribution associated with the original directed graph

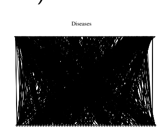
Outline

- Bayesian networks cont'd
  - graphs and consistency
- Undirected graphical models (Markov random fields)
  - graphs, independence, consistency, associated distribution
  - Bayesian networks as undirected models
- Quantitative probabilistic inference
  - medical diagnosis example
  - basic algorithms and problems

Example setting: medical diagnosis

- The QMR-DT model (Shwe et al. 1991)

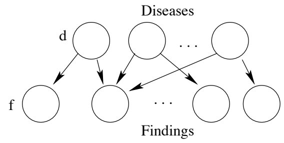




  - about 600 binary (0/1) disease variables representing diseases that are "present" or "absent"
  - about 4000 associated binary (0/1) findings; findings may be either "positive" or "negative"

Example cont'd

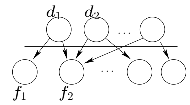
- The model is based on a number of simplifying assumptions



- Assumptions explicit in the graph:
  - relevant variables
  - marginal independence of diseases
  - conditional independence of findings
- Further assumptions about the probability distribution:
  - causal independence

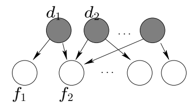
Assumptions in detail

- Diseases are marginally independent



$$\begin{aligned} d_1 &= \text{Hodgkins disease} \\ d_2 &= \text{Plasma cell myeloma} \\ d_3 &= \ldots \end{aligned}$$

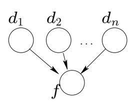
- Findings are conditionally independent given the diseases



$f_1 =$ Bone X-ray fracture
$f_2 = \ldots$

Assumptions in detail

- We have to specify how n (potentially 100 or more) underlying diseases conspire to influence any finding

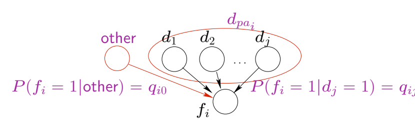


The size of the conditional probability table for $P(f|d_1, d_2, d_3, \ldots)$ would increase exponentially with the number of associated diseases

$\Rightarrow$ e.g, causal independence assumption

Causal independence: noisy-or

- We assume that each finding is negative if all the associated diseases (if present) *independently* fail to produce a positive outcome



$$\begin{aligned} P(f_i = 0|d_{pa_i}) &= P(f_i = 0|\text{other}) \prod_{j \in pa_i} P(f_i = 0|d_j) \\ &= (1 - q_{i0}) \prod_{j \in pa_i} (1 - q_{ij})^{d_j} \end{aligned}$$

and $P(f_i = 1|d_{pa_i}) = 1 - P(f_i = 0|d_{pa_i})$.

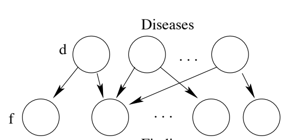
Joint distribution

- After all these assumptions, we can write down the following joint distribution over n diseases and m findings

$$P(f, d) = \left[\prod_{i=1}^{m} P(f_i|d_{pa_i}) \right] \left[\prod_{j=1}^{n} P(d_j) \right]$$

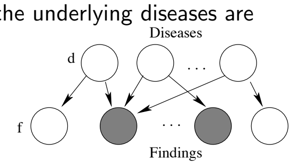
where $\quad P(f_i = 0|d_{pa_i}) = (1 - q_{i0}) \prod_{j \in pa_i} (1 - q_{ij})^{d_j}$

The only adjustable parameters in this model are q_{ij} and $P(d_j)$



Three inference problems

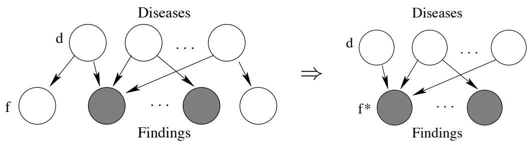
- Given a set of observed findings $f^* = \{f_1^*, \ldots, f_k^*\}$, we wish to infer what the underlying diseases are



1. What are the marginal posterior probabilities over the diseases?
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?

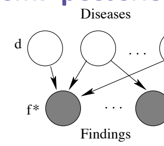
Inference problem cont'd

- For the purposes of inferring the presence or absence of the underlying diseases, we can ignore any findings that remain unobserved (as if they were not in the model to begin with)

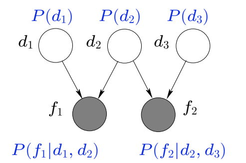


First inference problem: posterior marginals

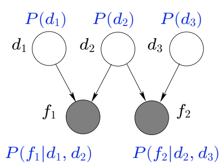
- Given the observations we already have all the information, only implicitly



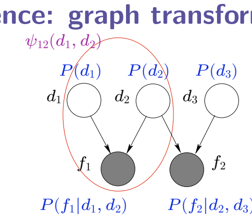
- What messages (if any) do the disease variables have to share for them to be able to compute the posterior marginals locally?



Inference: graph transformation

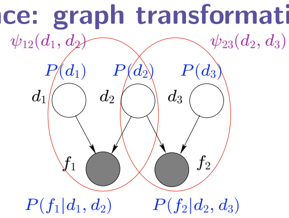


Inference: graph transformation



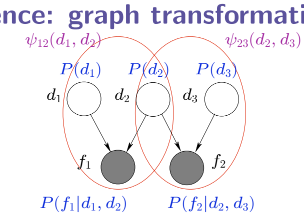
$$\psi_{12}(d_1, d_2) = P(d_1)P(d_2)P(f_1^*|d_1, d_2)$$

Inference: graph transformation



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

Inference: graph transformation



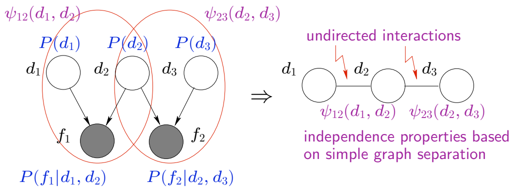
$$\psi_{12}(d_1, d_2) = P(d_1)P(d_2)P(f_1^*|d_1, d_2)$$
$$\psi_{23}(d_2, d_3) = P(d_3)P(f_2^*|d_2, d_3)$$

- Joint distribution as a product of "interaction potentials"

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

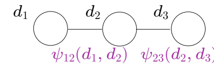
Inference: graph transformation

- 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

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

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

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

These will readily yield the posterior probabilities of interest:

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