6.867 Machine learning and neural networks

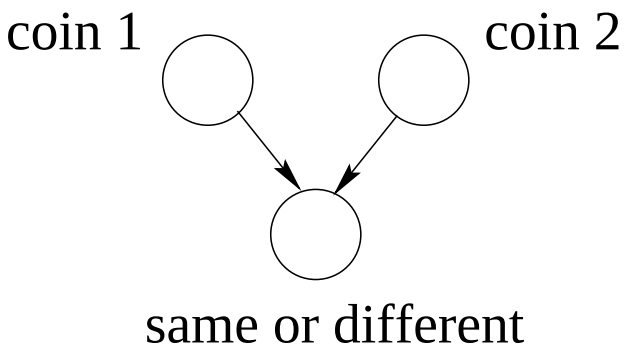
Tommi Jaakkola
MIT AI Lab
tommi@ai.mit.edu

Lecture 20: graph models

Topics

- Bayesian networks
  - graph semantics
  - associated probability distribution
- Medical diagnosis example
  - model specification
  - three inference problems

Graph model specification



$x_1 =$ first coin toss
$x_2 =$ second coin toss
$x_3 =$ same?

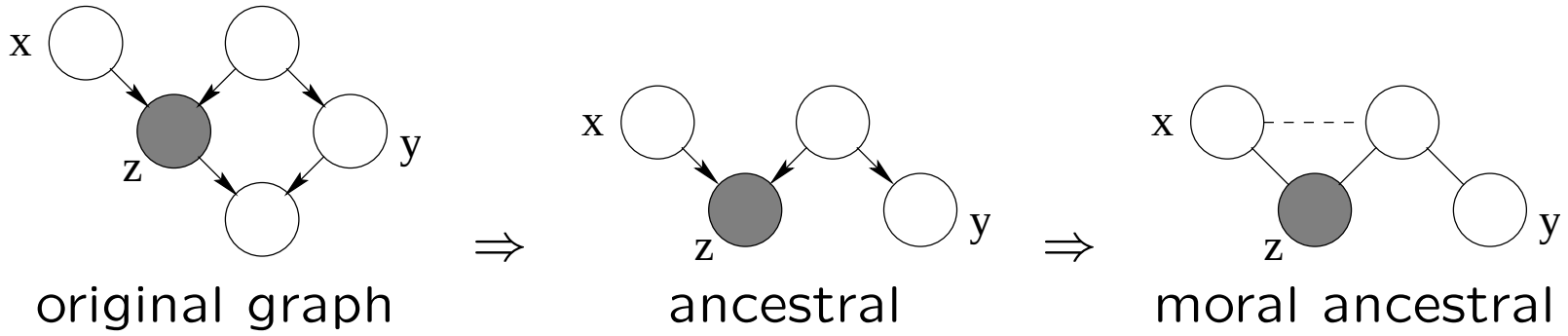
- We need to address the following questions
  1. What is the graph semantics?
  2. What type of probability distribution can be associated with any specific graph?
  3. How can we exploit the graph in making quantitative inferences?

Review: D-separation

- D-separation criterion (D for Directed edges):

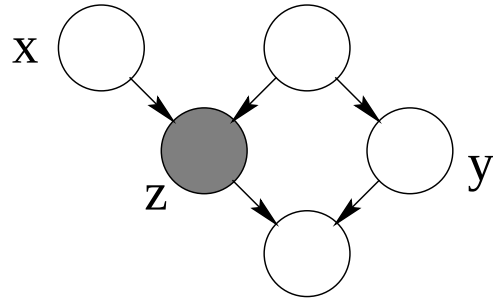
  Definition: variables x and y are D-separated (conditionally independent) given z if they are separated in the *moralized ancestral graph*

- Example:



original graph ⇒ ancestral ⇒ moral ancestral

Towards quantitative specification



- We can derive a number of conditional independence properties among the variables
- Can we always find a probability distribution that is *consistent* with all such independence properties?

  Consistency here means

  Separated in the graph $\Rightarrow$ independent in $P(x, y, z, \ldots)$

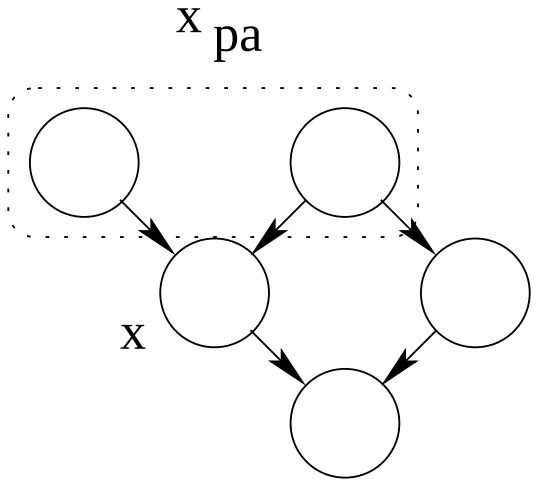
Bayesian networks

- Factorization theorem:

  Theorem: The most general form of the probability distribution consistent with the graph factors according to "node given its parents" :

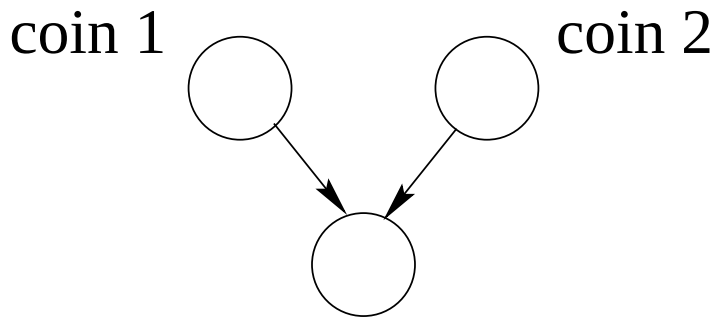
$$P(\mathbf{x}) = \prod_{i=1}^{d} P(x_i|\mathbf{x}_{pa_i})$$

  where $\mathbf{x}_{pa_i}$ is the set of *parents* of x_i. d is the number of nodes (variables) in the graph.



Bayesian networks: example

- The most general form of the probability distribution consistent with the graph



is given by

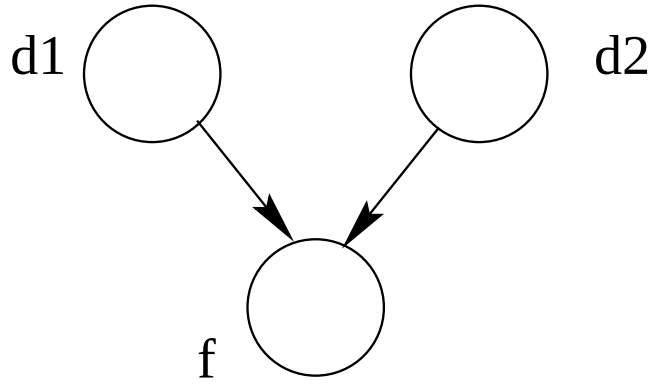
$$P(x_1, x_2, x_3) = P(x_1)\, P(x_2)\, P(x_3|x_1, x_2)$$

- Note that this includes, e.g.,

$$\begin{aligned} P(x_1, x_2, x_3) &= P(x_1)\, P(x_2)\, P(x_3), \text{ or} \\ P(x_1, x_2, x_3) &= P(x_1)\, P(x_2)\, P(x_3|x_1) \end{aligned}$$

Medical diagnosis example

- Two marginally independent diseases d_1 and d_2 and one corresponding finding f



$P(d_1)$:

0.9
0.1

$P(d_2)$:

0.8
0.2

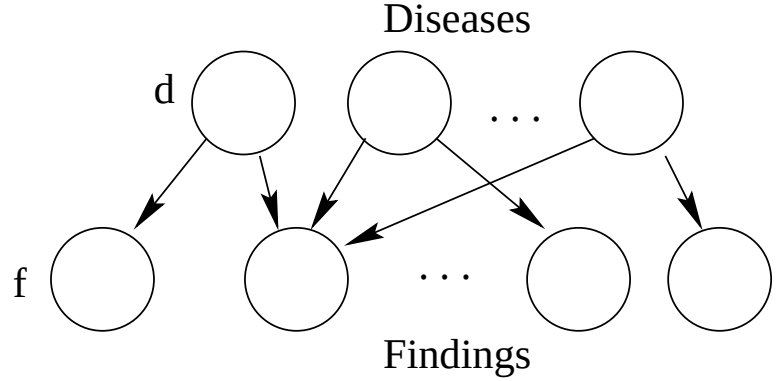
$P(f|d_1, d_2)$:

0.5	0.6	0.2	0.9
0.5	0.4	0.8	0.1

(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)

Medical diagnosis example cont'd

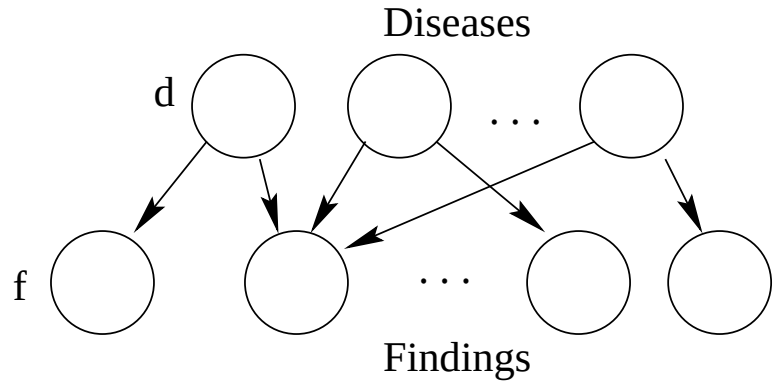
- The QMR-DT model (Shwe et al. 1991)



- The model contains about 600 significant diseases
  - the diseases can be either "present" $d = 1$ or "absent" $d = 0$
- There are about 4000 associated findings
  - the outcome of the findings are either "positive" $f = 1$ or "negative" $f = 0$ (e.g., results of laboratory tests or physician observations)

Medical diagnosis example cont'd

- There are a number of simplifying assumptions in the model



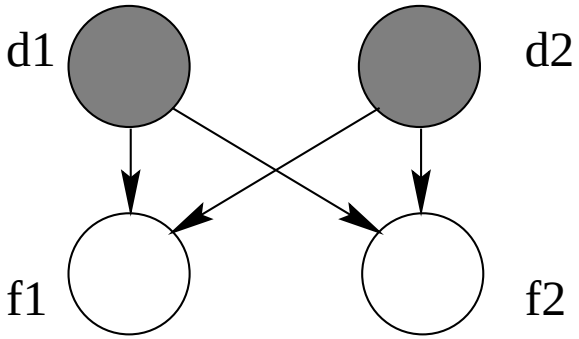
- Do we have all the relevant variables (e.g., significant diseases)?
- Assumptions that are explicit in the graph:
  - marginal independence of diseases
  - conditional independence of findings
- Assumptions about the underlying probability distribution:
  - causal independence assumptions

Assumptions in detail

- Diseases are marginally independent

d_1 = Hodgkins disease
d_2 = Plasma cell myeloma

- The findings are conditionally independent given the diseases



f_1 = Bone X-ray fracture
f_2 = ...

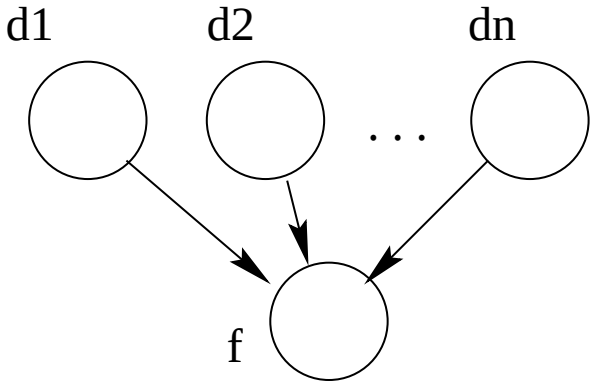
Assumptions cont'd

- We have to specify how n underlying diseases associated with a particular finding conspire to generate the outcome
- Causal independence assumption (Noisy-OR):

  The outcome is negative ($f = 0$) if all the diseases that are present ($d = 1$) independently fail to induce a positive outcome

$$P(f = 0|d_{pa}) = (1 - q_0) \prod_{j \in pa} (1 - q_j)^{d_j}$$

$$P(f = 1|d_{pa}) = 1 - P(f = 0|d_{pa})$$



  - d_{pa} is the set of diseases associated with finding f and q_j is the probability that disease j alone, if present, can generate a positive outcome
  - q_0 is the probability that an unknown disease would cause a positive finding

Joint distribution

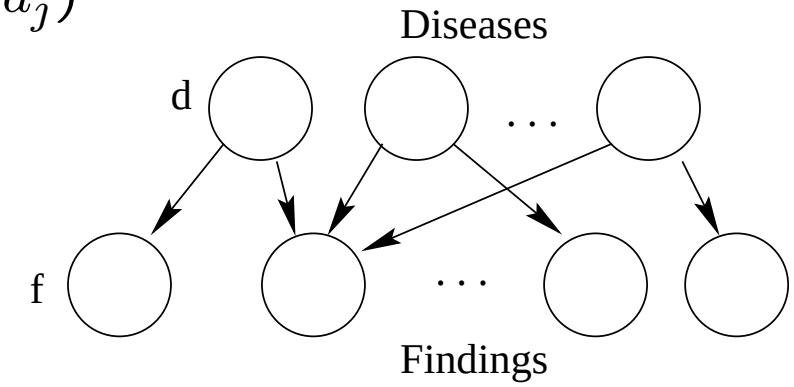
- After all these assumptions, we can write down the following joint distribution over n diseases and m findings

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

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

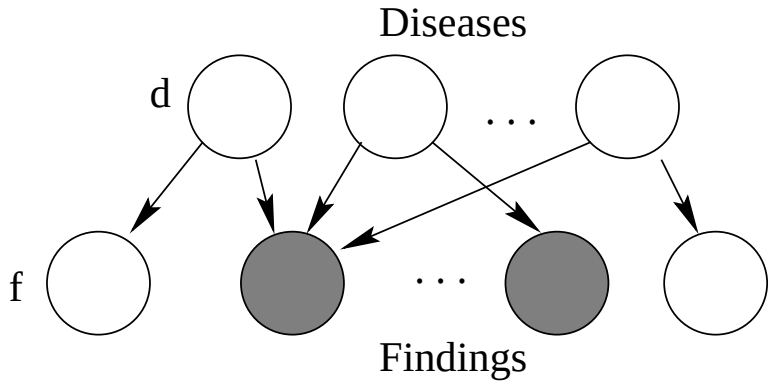
  where d_{pa_i} is the set of diseases associated with finding f_i

- To fully specify the underlying probability distribution (given the graph structure), we have to choose q_{ij} and the priori disease probabilities $P(d_j)$



Three inference problems

- Given a set of observed findings $f^* = \{f_2^*, \ldots, f_k^*\}$, we wish to infer what the underlying diseases are



1. What is the most likely setting of all the underlying disease variables?

$$d^* = \arg\max_d P(d|f^*) = \arg\max_d P(f^*, d)$$

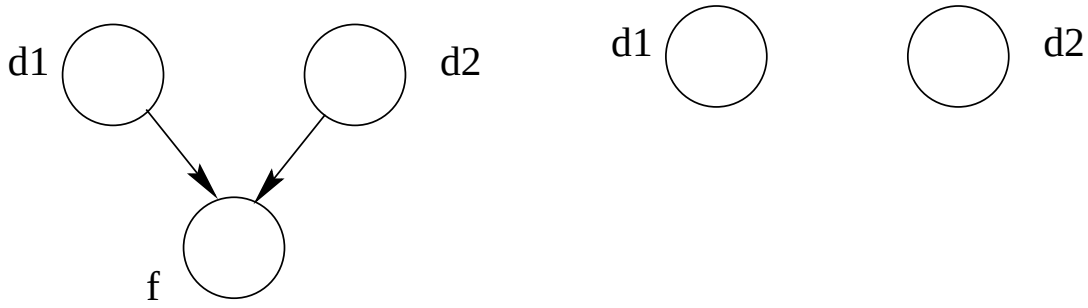
2. What are the marginal posterior probabilities

$$P(d_i = 1|f^*), \;\; i = 1, \ldots, n$$

3. Which test should we carry out next in order to get the most information about the diseases?

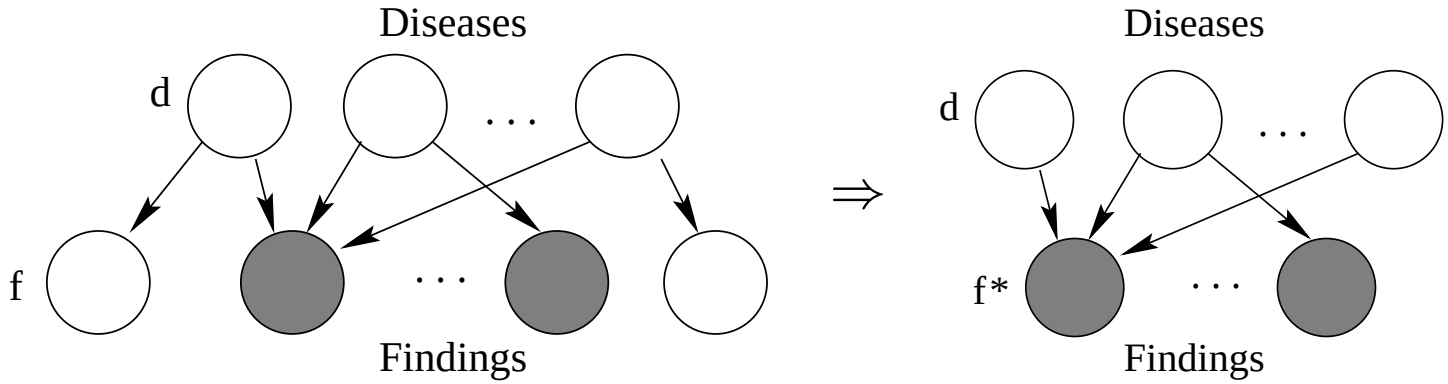
Inference problem cont'd

- For the purposes of inferring the presence or absence of the underlying diseases, the following two cases are equivalent



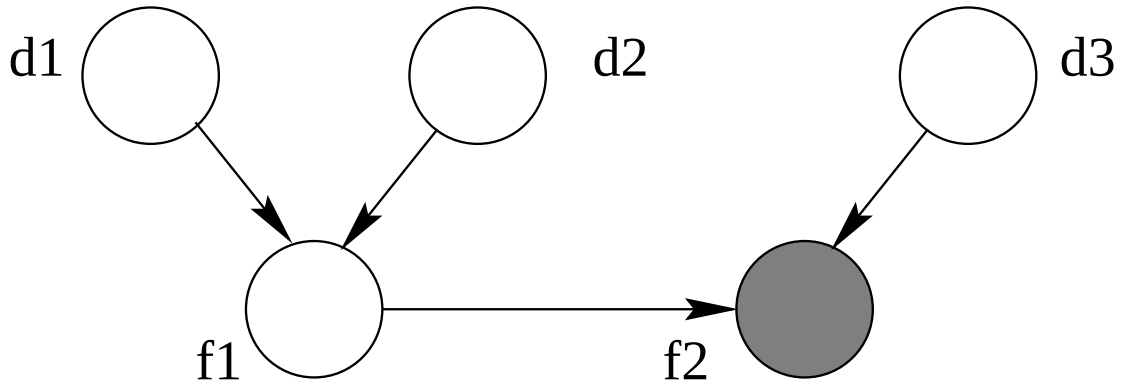
unobserved finding no associated findings

- In other words, we can ignore any findings that remain unobserved (as if they were not in the model to begin with)



Inference problem cont'd

- What if the findings were not conditionally independent given the diseases?



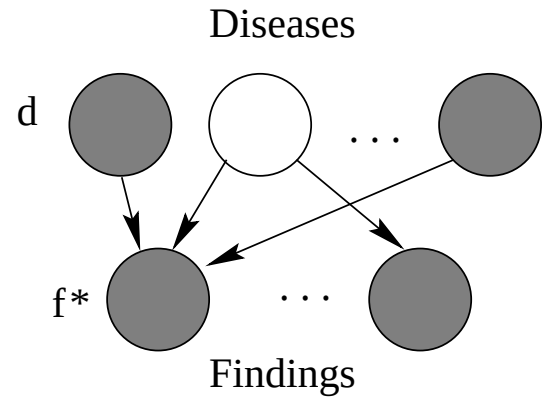
First inference problem

- We can try to find the most likely disease configuration given f^* via a search algorithm

  A very simple (naive) algorithm:
  1. Start with all diseases absent $d_1^* = 0, \ldots, d_n^* = 0$
  2. Successively update each disease variable to increase the probability $P(f^*, d^*)$ of diseases and the observed findings

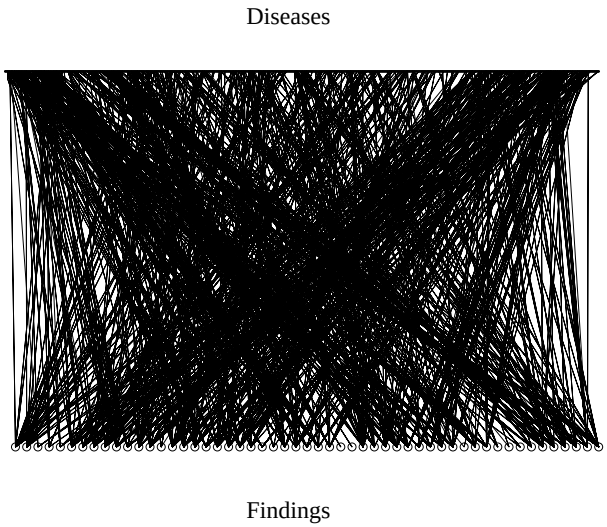
$$d_j^* \leftarrow \arg\max_{d_j} P(f^*, d_1^*, \ldots, d_{j-1}^*, d_j, d_{j+1}^*, \ldots, d_n^*)$$

Are there any better algorithms that we could use? Dynamic programming?



First inference problem cont'd

- The search may not be that easy...



(this is a small portion of the real QMR-DT)

Second inference problem

- We wish to find the marginal posterior probabilities of the diseases given the findings (i.e., the overall probability that individual diseases are present given the findings)

$$P(d_i = 1|f^*) = \frac{P(f^*, d_i = 1)}{P(f^*)} = \frac{\sum_d d_i \, P(f^*, d)}{\sum_d P(f^*, d)}$$

- This involves summing over all configurations of diseases...

  ... there are 2^{600} such disease configurations

- Two possible ways around this:

  1. Exploit the model structure (later)
  2. Approximate inference (sampling)