Machine learning: lecture 19

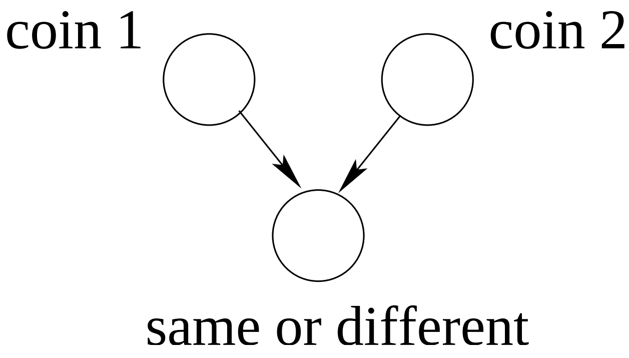
Tommi S. Jaakkola
MIT AI Lab
tommi@ai.mit.edu

Topics

- Gaphical models
  - Examples, specification
  - Bayesian networks
  - graph semantics
  - associated probability distribution
- Medical diagnosis example
  - three inference problems

Graphical models: two levels of description

1. Qualitative properties captured by a graph



x_1 = first coin toss

x_2 = second coin toss

x_3 = same?

2. Quantitative properties specified by the associated probability distribution

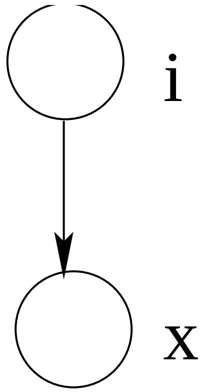
$$P(x_1, x_2, x_3) = P(x_1)\, P(x_2)\, P(x_3|x_1, x_2)$$

where, e.g.,

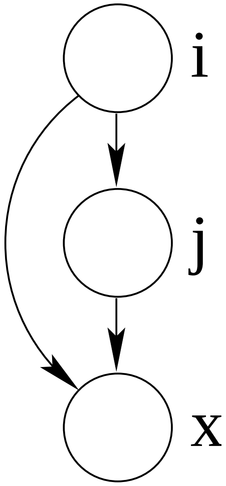
$$P(x_1 = heads) = 0.5$$

$$P(x_3 = same|x_1 = heads, x_2 = tails) = 0$$

Examples



Mixture model

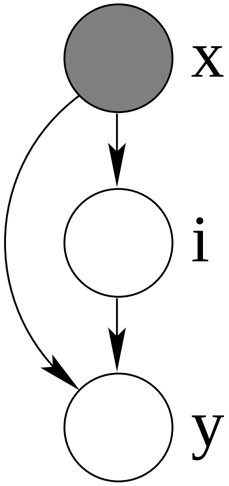


hierarchical mixture model

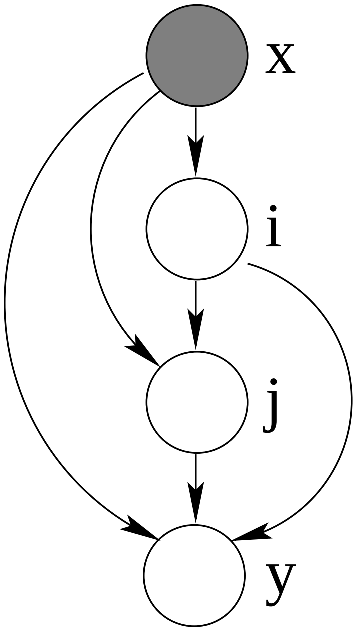
- i and j correspond to the discrete choices in the mixture model
- $\mathbf{x}$ is the (vector) variable whose density we wish to model

- We cannot tell what the component distributions $P(\mathbf{x}|i)$ are based on the graph alone

Examples cont'd



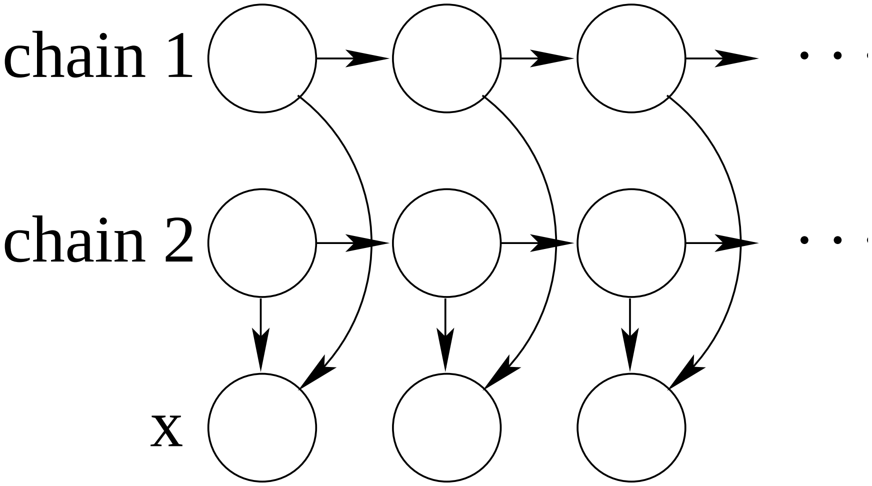
Mixture of experts



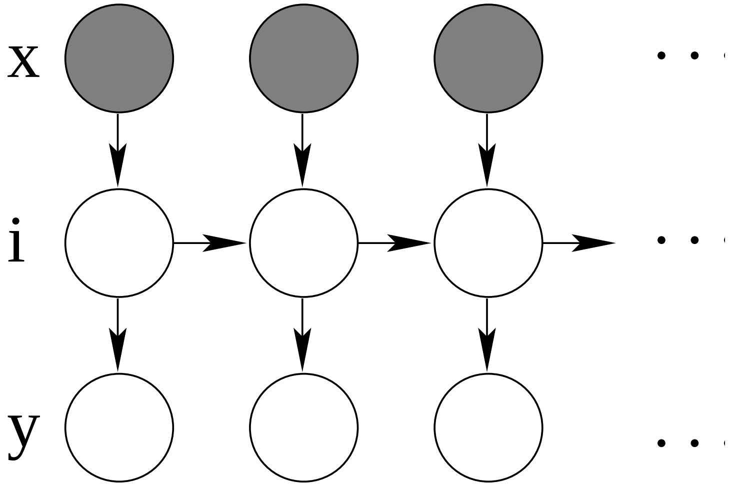
hierarchical mixture of experts

- In this case the choices of i and j and the output y depend on the input x
  (The shaded variables denote *observed* values; we do not need to model the density over $\mathbf{x}$)

Examples cont'd



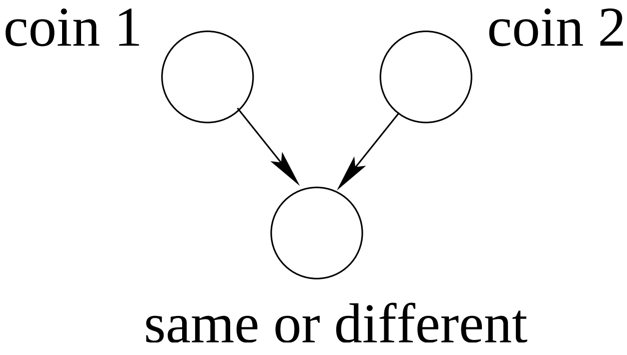
Factorial HMM



input-output HMM

- In factorial HMMs, independent processes conspire to generate the observed output sequence
- In input-output HMMs, any observed sequence of outputs y is accompanied by a corresponding sequence of *inputs* $\mathbf{x}$
  - the model tranforms any input sequence into an output sequence (markov?)

Graph model specification



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

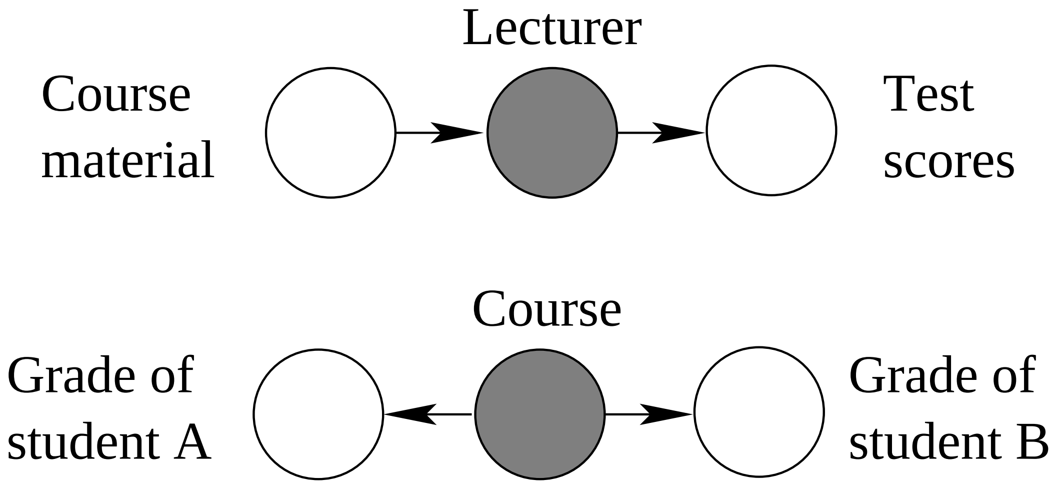
- We need to address the following questions

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?

- We will focus initially on *Bayesian networks* or directed acyclic graphs

Graph semantics: Bayesian networks

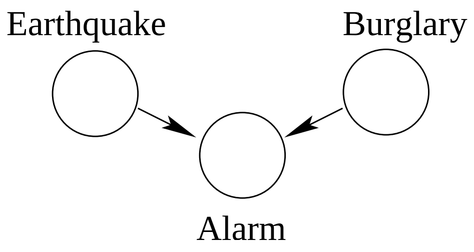
- The graph captures *independence properties* among the variables
- The independences can be read from the graph based on some notion of *graph separation*



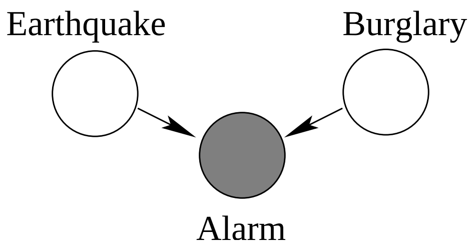
conditional independence

Graph semantics cont'd

- Here are the interesting cases...



x and y are marginally independent



x and y are conditionally dependent

- These capture the notion of *induced dependencies*. In other words, when you learn more you might make previously independent variables suddenly dependent

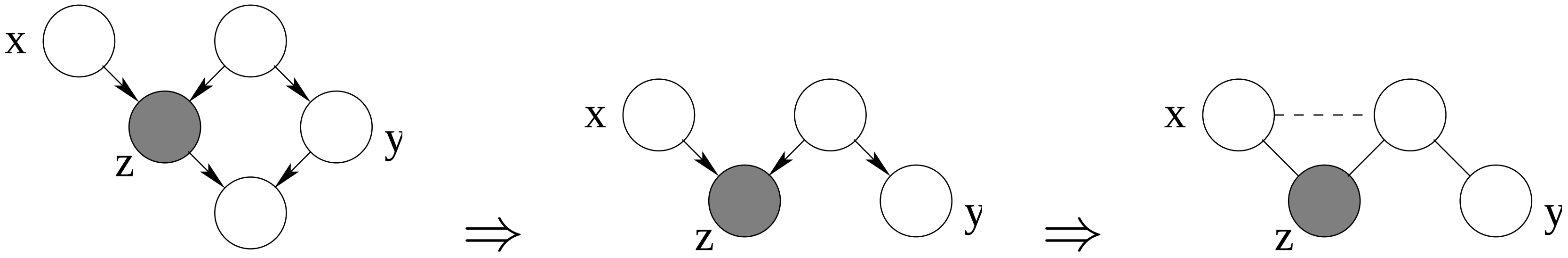
- Note that the "graph separation" measure must pay attention to the direction of the edges

Graph separation criterion (briefly)

- D-separation criterion for Bayesian networks (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:



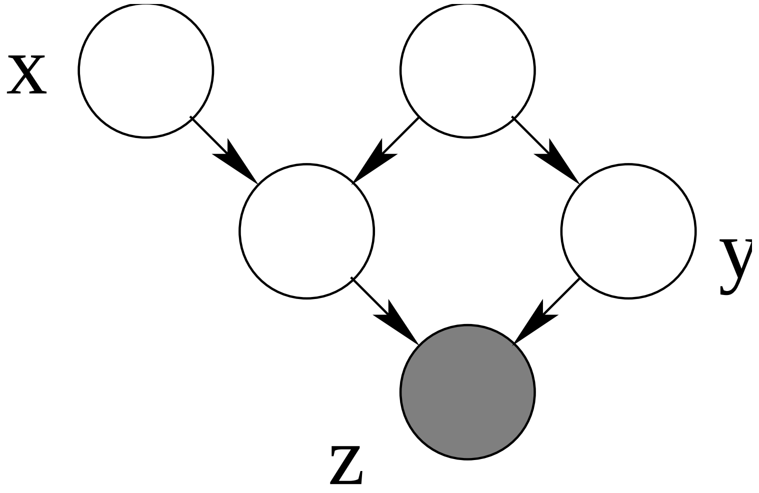
original graph

ancestral

moral ancestral

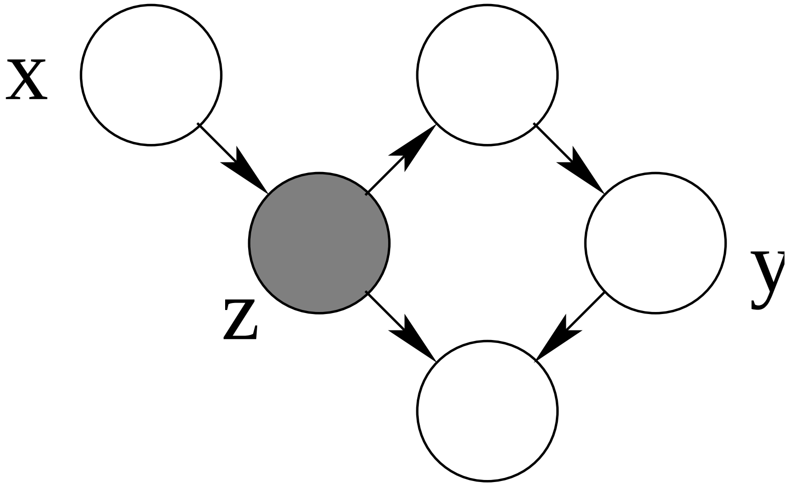
D-separation: example

- Example: are x and y D-separated given z?



Towards quantitative specification

- Separation properties in the graph imply independence properties about the associated variables
- For the graph to be useful any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents



For example, if x and y are D-separated given z then the underlying distribution should satisfy

$$P(x, y|z) = P(x|z)P(y|z)$$

Bayesian networks

- Factorization theorem:

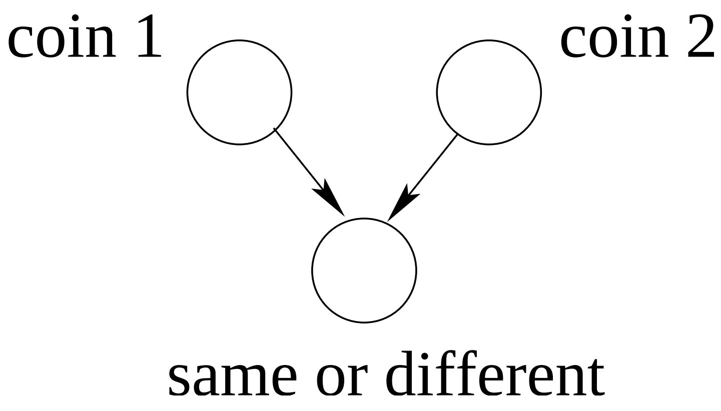
  Theorem: The most general form of the probability distribution that is consistent with the graph factors according to "node given its parents":

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

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

Examples

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



is given by

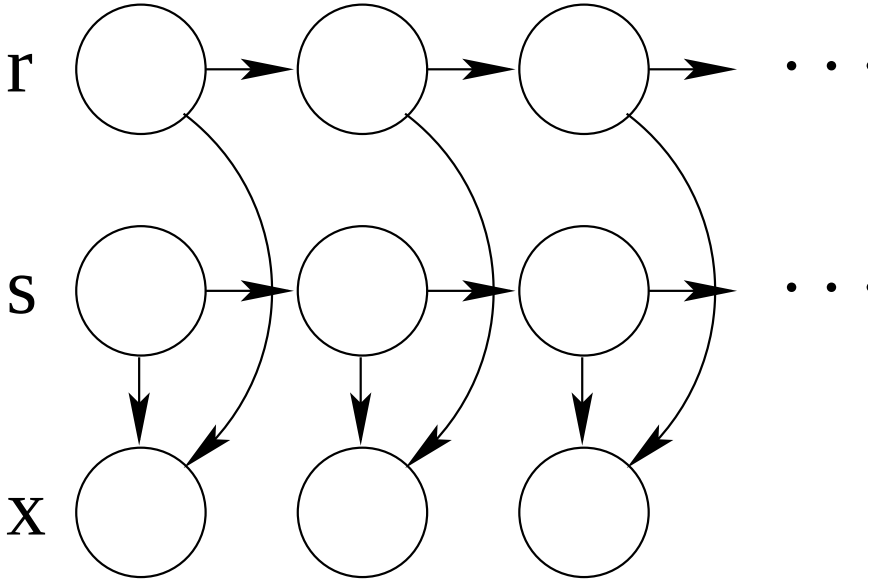
$$P(x_1, x_2, x_3) = P(x_1)\, P(x_2)\, P(x_3|x_1, x_2)$$

- Note that this still includes, e.g.,

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

Examples cont'd

- A factorial HMM

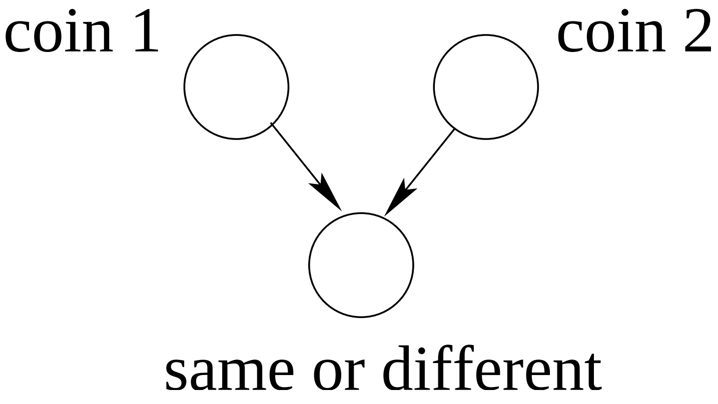


- The associated probability model has the following form

$$
\begin{aligned}
P(r_0, s_0, \mathbf{x}_0, r_1, s_1, \mathbf{x}_1, \ldots) &= P_0(r_0)\, P_1(r_1|r_0)\, \cdots \\
&\quad \times P_0(s_0)\, P_1(s_1|s_0)\, \cdots \\
&\quad \times P_o(\mathbf{x}_0|r_0, s_0)\, P_o(\mathbf{x}_1|r_1, s_1)\, \cdots
\end{aligned}
$$

Bayesian networks

Some additional properties:

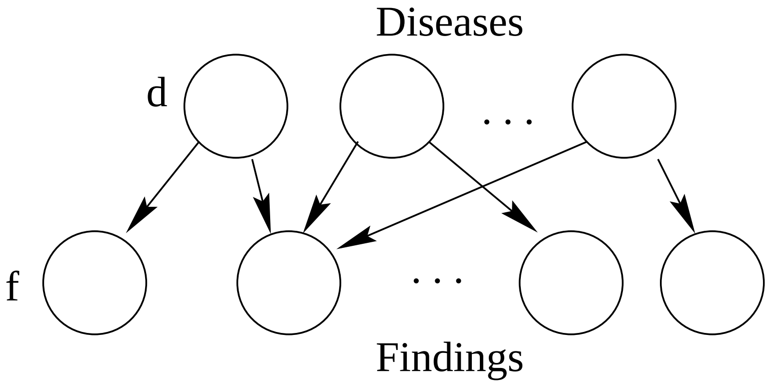


$$P(x_1, x_2, x_3) = P(x_1)\, P(x_2)\, P(x_3|x_1, x_2)$$

- The normalization is *local* in the sense that each of the components in the factorization is normalized to one
- We still have a lot of freedom to choose, e.g., $P(x_3|x_1, x_2)$ and be consistent with the graph; $P(x_3|x_1, x_2)$ can be a full probability table, logistic regression model, etc

Medical diagnosis example

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

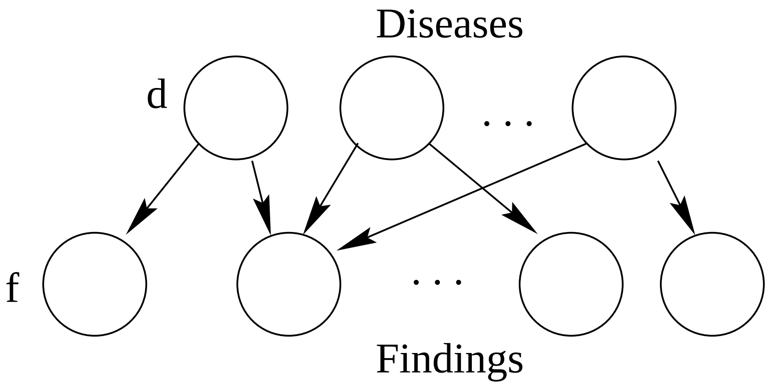


- The model contains about 600 significant diseases
  - the diseases can be either "present" or "absent" ($d = 1$ or $d = 0$)

- There are about 4000 associated findings
  - the outcome of the findings are either "positive" or "negative" ($f = 1$ or $f = 0$)

Medical diagnosis example cont'd

- There are a number of simplifying assumptions in the model



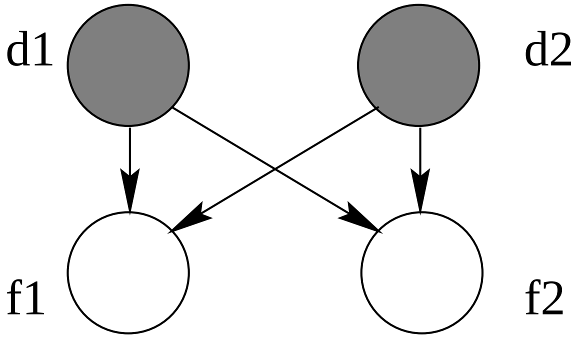
- Do we have all the relevant variables (e.g., significant diseases)?
- Assumptions that are explicit in the graph:
  - marginal independence of diseases
  - conditional independence of findings
- Assumptions about the underlying probability distribution:
  - causal independence assumptions

Assumptions in detail

- Diseases are marginally independent

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

- The findings are conditionally independent given the diseases

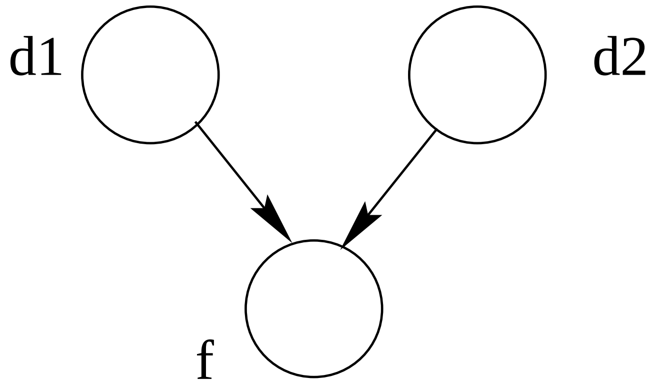


$f_1 =$ Bone X-ray fracture

$f_2 = \ldots$

Assumptions cont'd

- We have to specify how n underlying diseases associated with a particular finding conspire to generate the outcome



$P(d_1)$:

0.9
0.1

$P(d_2)$:

0.8
0.2

$P(f|d_1,d_2)$:

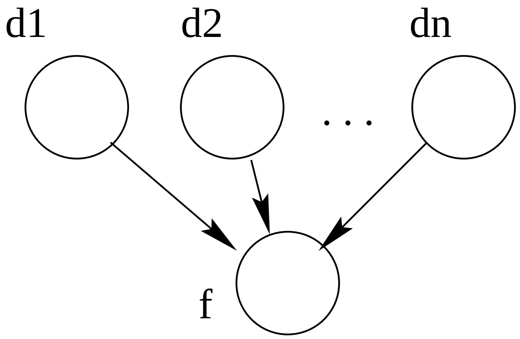
0.5	0.6	0.2	0.9
0.5	0.4	0.8	0.1

(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)

Assumptions cont'd

- 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

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



  - 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

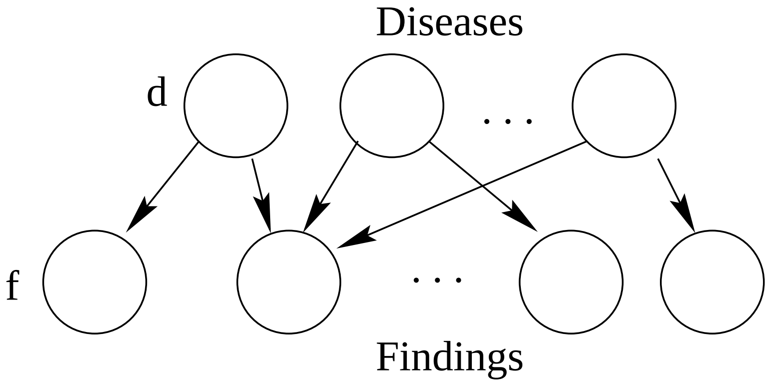
Joint distribution

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

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

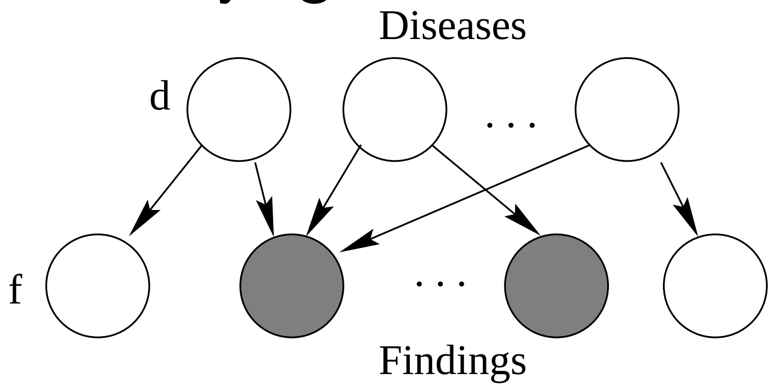
where $$P(f_i = 0|d_{pa_i}) = (1 - q_{i0}) \prod_{j \in pa_i} (1 - q_{ij})^{d_j}$$

and d_{pa_i} is the set of diseases associated with finding f_i. The adjustable parameters of this model are q_{ij} and $P(d_j)$



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?

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

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