Machine learning: lecture 20

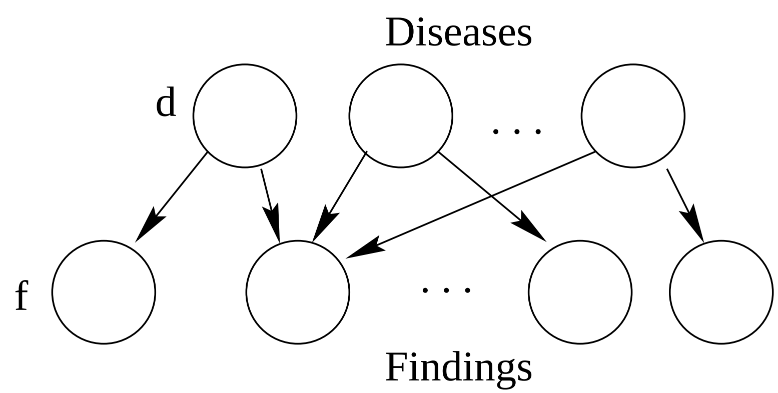
Tommi S. Jaakkola
MIT AI Lab
tommi@ai.mit.edu

Topics

- Medical diagnosis example cont'd
  - three inference problems
- Graphical models and decision theory
  - expected utility framework
  - influence diagrams
  - examples

Medical diagnosis example: review

- Model and assumptions:



1. Assumptions that are explicit in the graph:
   - marginal independence of diseases
   - conditional independence of findings
2. Assumptions about the underlying probability distribution:
   - causal independence assumptions (noisy-OR conditional probabilities)

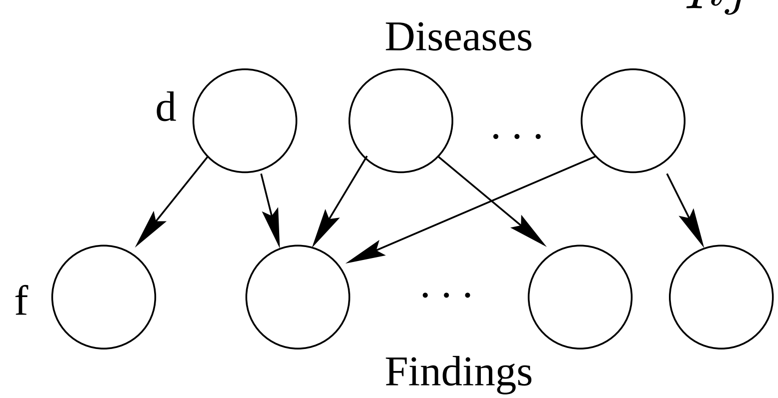
Joint distribution: review

- The assumptions imply the following joint distribution over n diseases and m findings

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

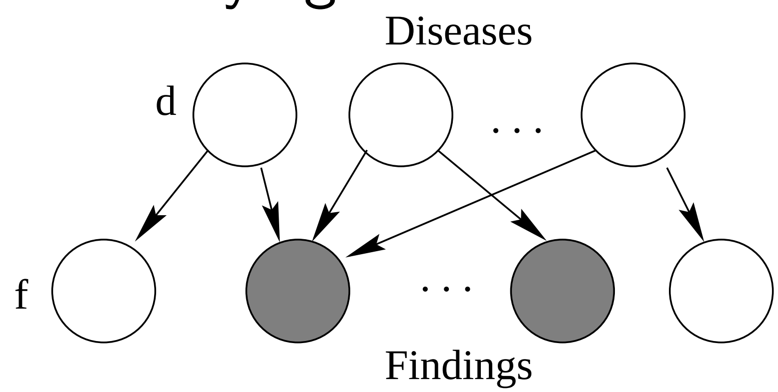
where $$P(f_i = 0|d_{pa_i}) = (1 - q_{i0}) \prod_{j \in pa_i} (1 - q_{ij})^{d_j}$$

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



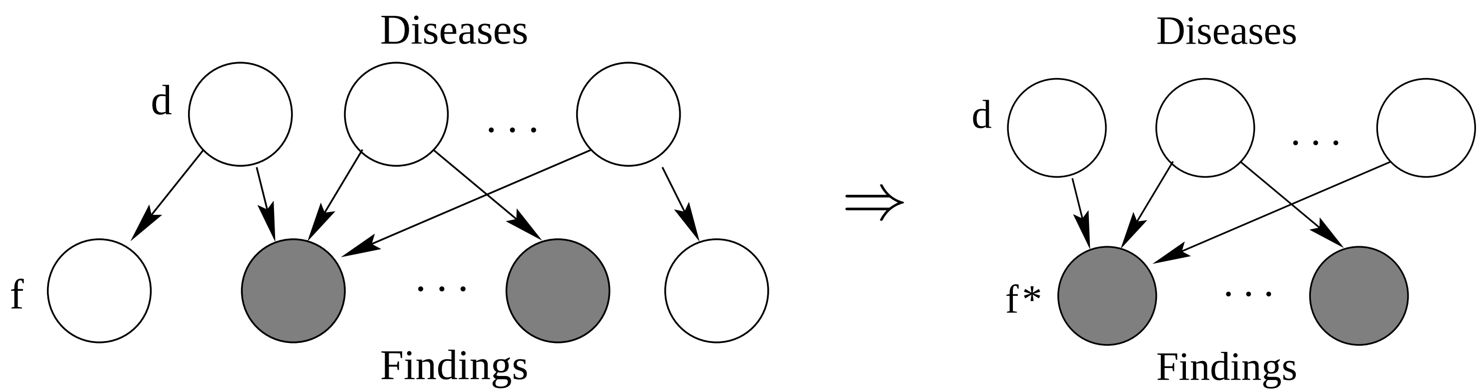
1. What is the most likely setting of all the underlying disease variables?

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?

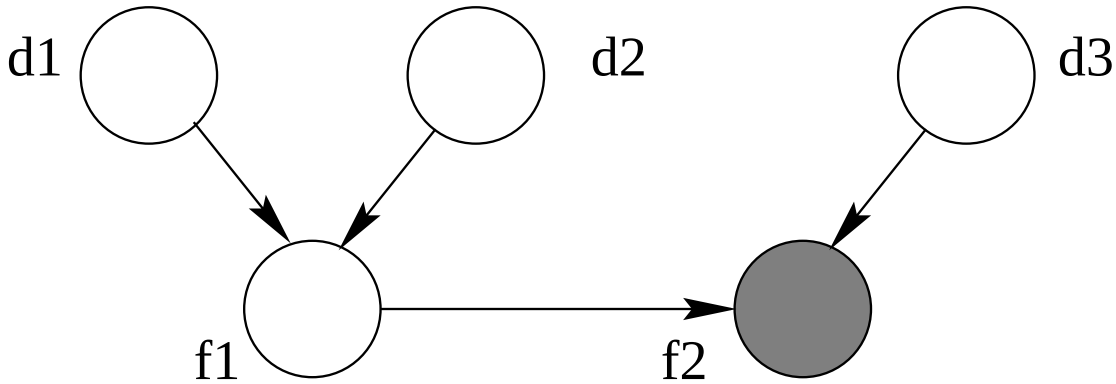
Inference problem cont'd

- 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)



Inference problem cont'd

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



First inference problem

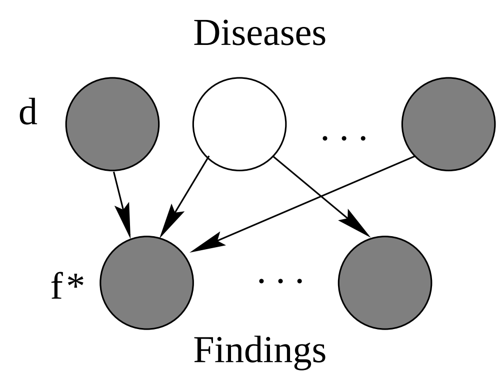
- We can try to find the most likely disease configuration given f^* via a search algorithm

  A naive search algorithm:

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

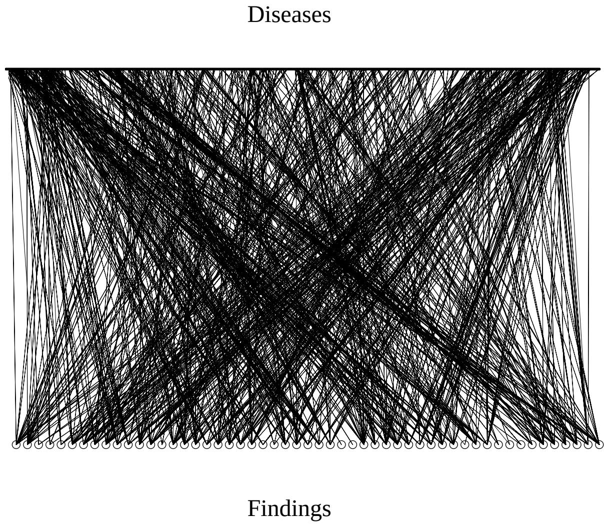
$$d_j^* \leftarrow \arg\max_{d_j} P(f^*, d_1^*, \ldots, d_{j-1}^*, d_j, d_{j+1}^*, \ldots, d_n^*)$$

Why is this naive approach likely to fail?



First inference problem cont'd

- Exact search may not be very 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)

Second inference problem cont'd

- What if we just sampled disease configurations from the posterior distribution $P(d|f^*)$ and computed the fraction of times disease d_i were present?

$$P(d_i = 1|f^*) \approx \frac{1}{T}\sum_{t=1}^{T} d_i^t$$

where each $d^t = \{d_1^t, \ldots, d_n^t\}$ is an independent sample drawn from the posterior $P(d|f^*)$

- The problem is that we cannot easily sample from $P(d|f^*)$...

Importance sampling

- We can approximate the infeasible summations over exponentially many disease configurations via *importance sampling*

$$
\begin{aligned}
P(f^*) = \sum_d P(f^*, d) &= \sum_d Q(d)\,\frac{P(f^*, d)}{Q(d)} \\
&= E_{d\sim Q}\left\{ \frac{P(f^*, d)}{Q(d)} \right\} \\
&\approx \frac{1}{T}\sum_{t=1}^{T} \frac{P(f^*, d^t)}{Q(d^t)}
\end{aligned}
$$

where the disease configurations d^t are drawn from the simple distribution $Q(d)$ (known as a *proposal distribution*).

Second inference problem cont'd

- We draw samples from a much simpler proposal distribution $Q(d)$ and approximate

$$
\begin{aligned}
P(f^*) &= \sum_d P(f^*, d) \approx \frac{1}{T} \sum_{t=1}^{T} \frac{P(f^*, d^t)}{Q(d^t)} \\
P(f^*, d_i = 1) &= \sum_d d_i \, P(f^*, d) \approx \frac{1}{T} \sum_{t=1}^{T} d_i^t \, \frac{P(f^*, d^t)}{Q(d^t)}
\end{aligned}
$$

where each configuration $d^t = \{d_1^t, \ldots, d_n^t\}$ is an independent sample from $Q(d)$

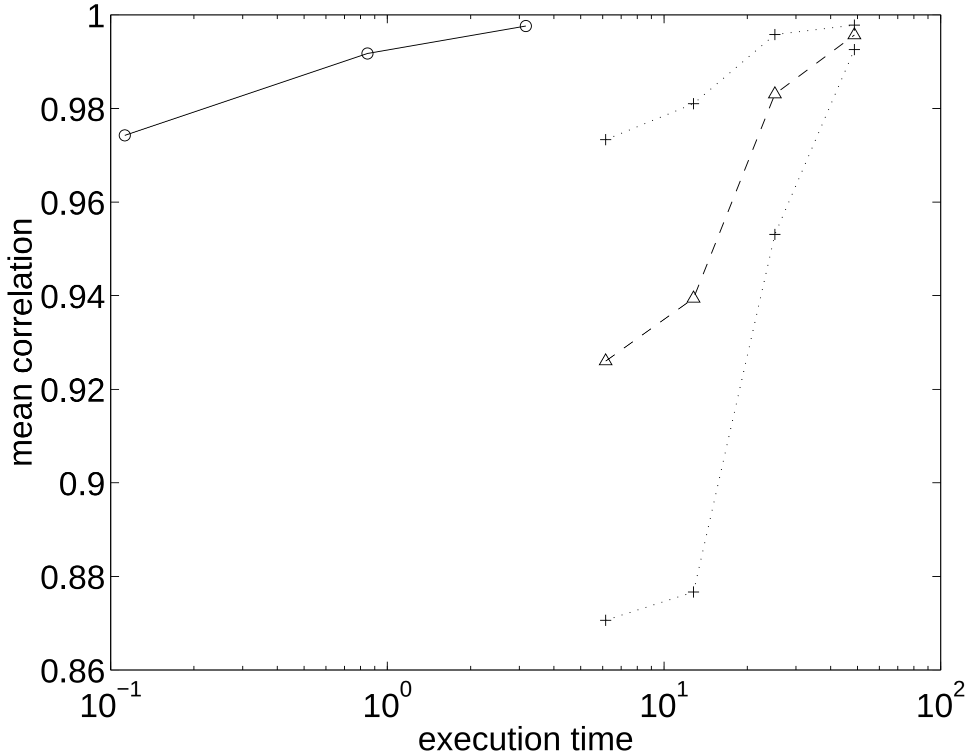
Second inference problem cont'd

- We get the desired posterior marginals by taking ratios of these estimates (known as likelihood weighted sampling)

$$P(d_i = 1 | f^*) = \frac{P(f^*, d_i = 1)}{P(f^*)} \approx \frac{\frac{1}{T}\sum_{t=1}^{T} d_i^t \frac{P(f^*, d^t)}{Q(d^t)}}{\frac{1}{T}\sum_{t=1}^{T} \frac{P(f^*, d^t)}{Q(d^t)}}$$

Second inference problem cont'd

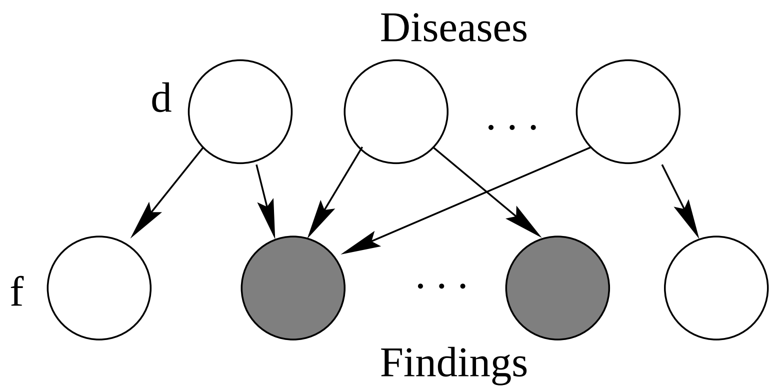
- This actually works... sort of



The figure shows the overall correlation between the estimated and exact posterior marginals (in simple cases)

Third inference problem

- We would like to find out which tests to carry out next in order to get the most information about the underlying diseases



- For this we need to know
  1. how uncertain the outcomes of the other findings are given the information we have so far
  2. the hypothetical effect of actually observing either outcome of the new findings

Third inference problem cont'd

- how uncertain the outcomes of the other findings are given the information we have so far (f^*)

$$P(f_i|f^*) = \sum_d P(f_i|d_{pa_i})\, P(d|f^*)$$

- the hypothetical effect of actually observing either outcome of the new findings

$$P(d|f_i, f^*) = \frac{P(d, f_i, f^*)}{P(f_i, f^*)}$$

for $f_i = 0, 1$.

Third inference problem cont'd

- We select the test that has the best chance of reducing the uncertainty about the underlying diseases
- This is the test that has the highest mutual information with the diseases

$$I(f_i; d|f^*) = \sum_{f_i=0,1} P(f_i|f^*) \underbrace{\left[\sum_d P(d|f_i, f^*) \log \frac{P(d|f_i, f^*)}{P(d|f^*)} \right]}_{\text{the difference between our uncertainty about the diseases before and after observing } f_i}$$

- The individual terms could be approximated similarly to the previous sampling method

Topics

- Graphical models and decision theory
  - expected utility framework
  - influence diagrams
  - examples

Decision theory: expected utility framework

- To select among alternative courses of action we need to balance the potential costs/benefits of actions with the uncertainty that we have about the outcomes of actions, current situation, etc.

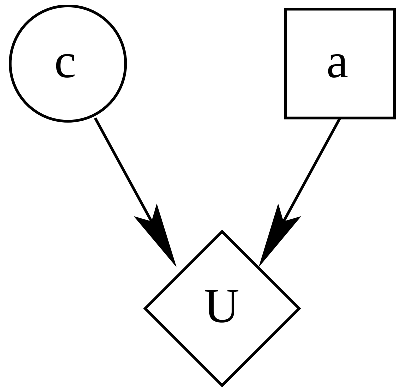
- Expected utility framework: given a probability distribution $P(c)$ over the uncertain quantities c, we select the action that maximizes the *expected utility*:

$$a^* = \arg\max_{a \in A} E_{c \sim P}\left\{U(c,a)\right\} = \arg\max_{a \in A} \sum_{c} P(c)\, U(c,a)$$

  where A is the set of available actions and $U(c,a)$ is the (subjective) utility of choosing action a in context c.

- Influence diagrams are graphical models that incorporate the notion of actions/decisions and their utility.

Influence diagrams: basics



$$a^* = \arg\max_{a \in A} E_{c \sim P} \left\{ U(c, a) \right\}$$

- Basic components:
  - decision node(s) (squares) specify the actions we can take
  - chance nodes (circles) specify our belief about the values of variables relevant for decisions
  - utility node (diamond) specifies the utility of any decision in a contex c