6.867 Machine learning and neural networks

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

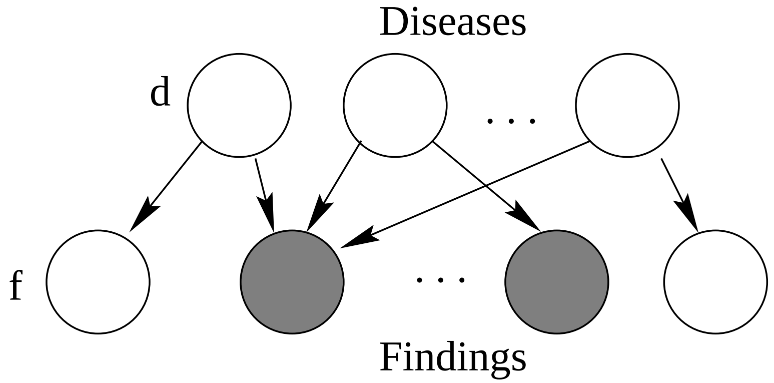
Lecture 21: graph models cont'd

Topics

- Medical diagnosis example cont'd
  - three inference problems
- Markov random fields
  - motivation, model semantics
  - associated distribution
  - pattern completion example

Review: 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?

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 configuration from the posterior $P(d|f^*)$

  But we cannot easily sample from $P(d|f^*)$...

Importance sampling

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

  Example:

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

  where the disease configurations d^t are drawn from the simple proposal distribution $Q(d)$ (which one?)

Second inference problem cont'd

- We can evaluate the relevant probabilities approximately by drawing samples from the simple proposal distribution $Q(d)$:

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

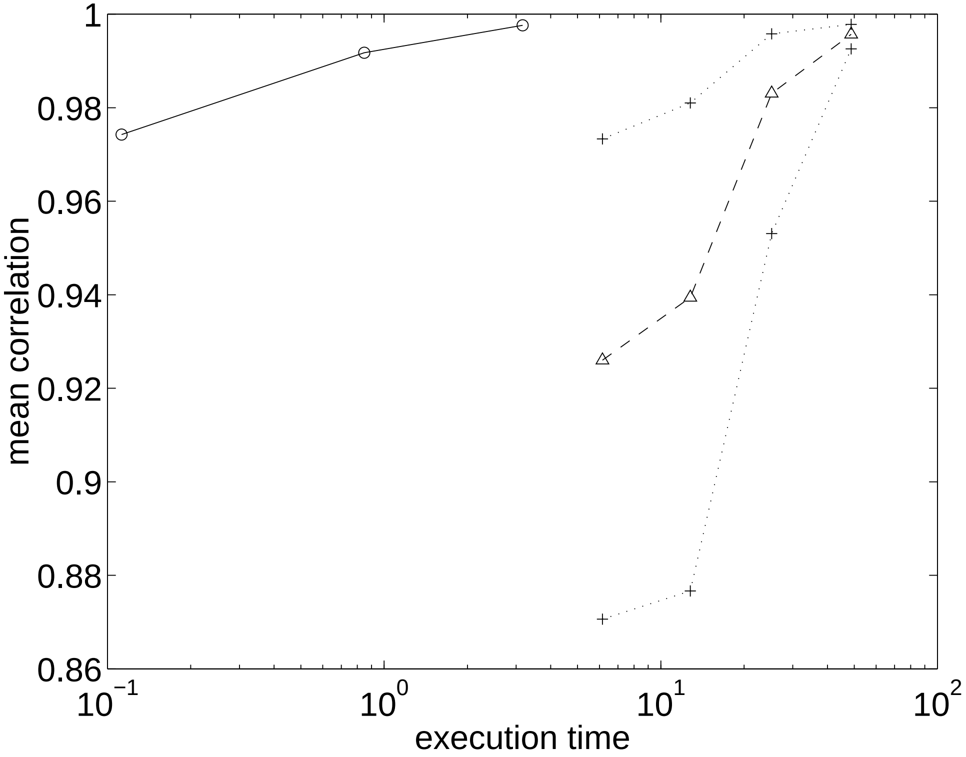
- The desired posterior marginals are obtained as ratios of these sampled estimates:

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

(likelihood weighted sampling)

Second inference problem cont'd

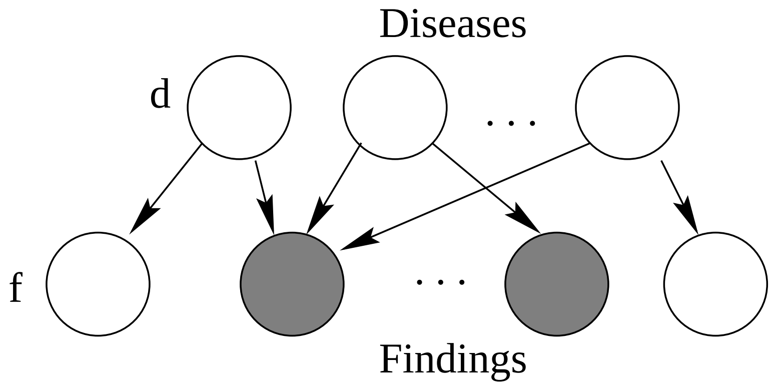
- This actually works...



Overall correlation between the estimated and exact posterior marginals (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 how *uncertain* the outcomes of other findings are given the observed ones f^*

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

as well as the (hypothetical) effect of observing $f_i = 0, 1$ on the diseases

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

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

$$I(f_i; d) = \sum_{f_i = 0,1} P(f_i|f^*) \underbrace{\left[\sum_d P(d|f_i, f^*) \log \frac{P(d|f_i, f^*)}{P(d|f^*)} \right]}_{\substack{\text{comparison of disease uncer-} \\ \text{tainties before and after ob-} \\ \text{serving } f_i = 0,1}}$$

(individual terms here could be approximated as before)

- Other criteria?

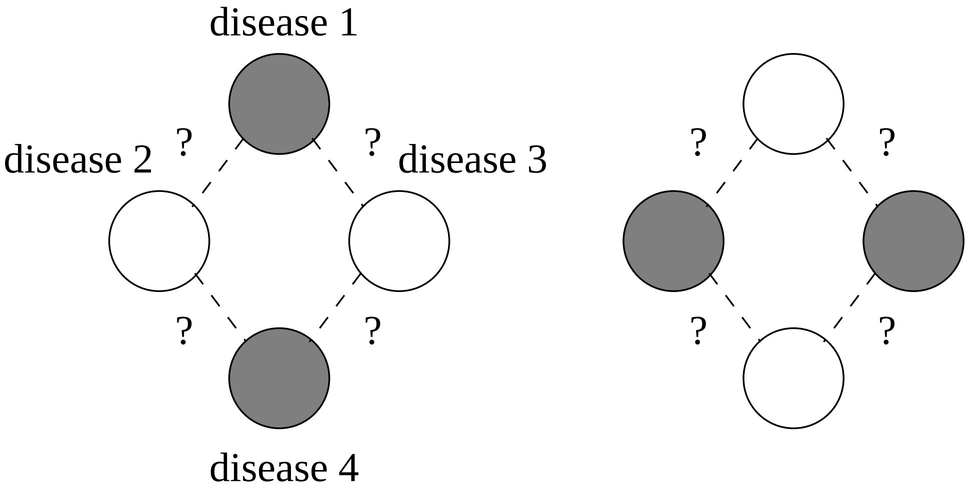
Topics

- Markov random fields
  - motivation, model semantics
  - associated distribution
  - pattern completion example

Limitations of Bayesian networks

- The graph should *explicitly* capture the independence properties among the variables

  For example: how can we draw the arrows in a Bayesian network



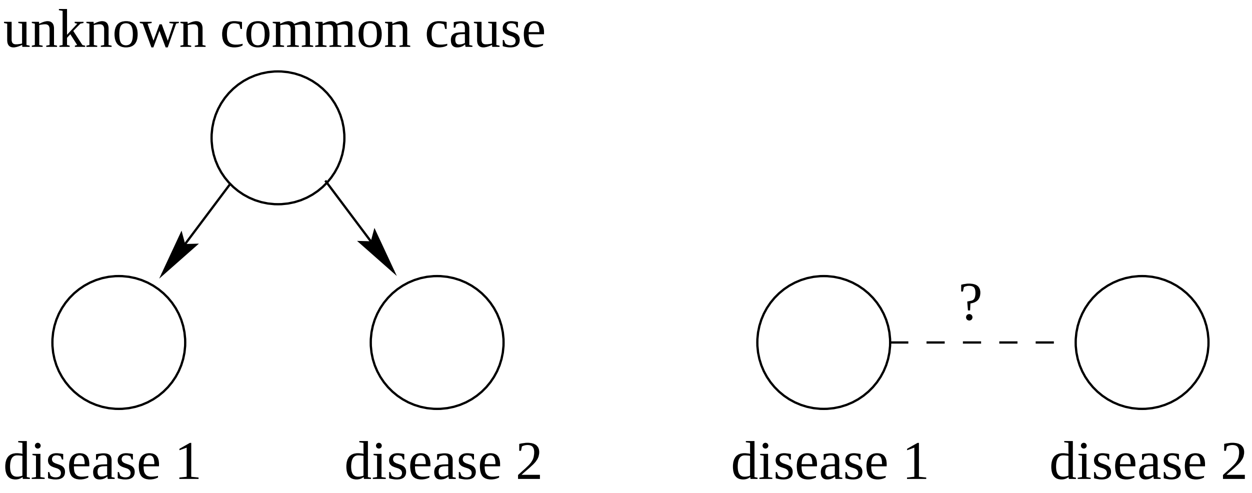
  such that

  diseases 2 and 3 are cond. indep. given 1 and 4

  diseases 1 and 4 are cond. indep. given 2 and 3

Limitations of Bayesian networks cont'd

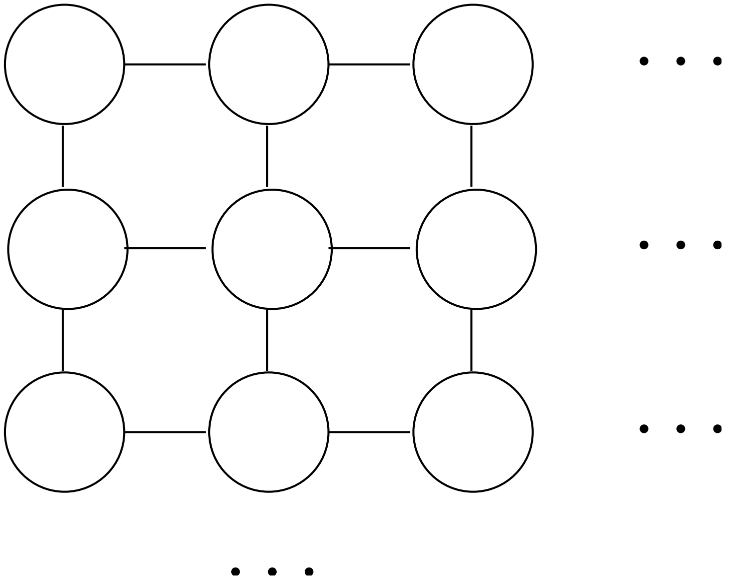
- How can we model *symmetric* interactions between two variables (e.g., diseases) with a Bayesian network?



- Such symmetric interactions are better modeled with undirected graph models (Markov random fields)

Markov random fields

- Markov random fields are complementary graph models that try to capture symmetric dependencies
- Example: a spin lattice with nearest neighbor dependencies

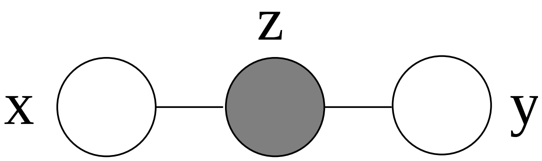


- As before, we have to
  - define graph semantics
  - associated probability distribution

Graph semantics

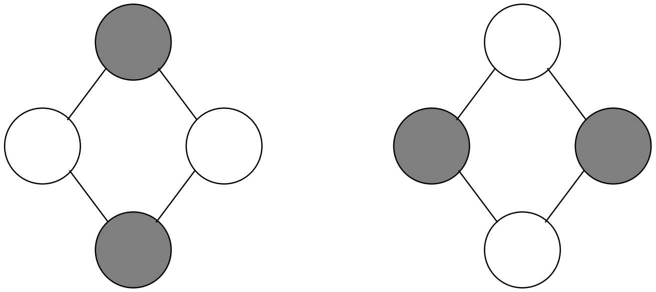
- The (conditional) independence properties can read from the graph via simple graph separation:

  x and y are conditionally independent given z if all paths between x and y go through z



x and y are conditionally independent given z

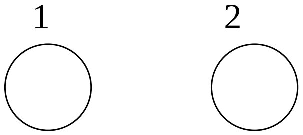
- This graph semantics captures our previous example



- We still need to determine what type of distributions are consistent with the graph...

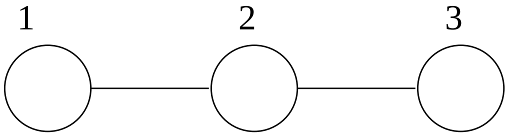
Markov random fields

- Simple independent example:



$$P(x_1, x_2) = \underbrace{\frac{1}{Z}}_{1} \; \underbrace{\psi_1(x_1)}_{P(x_1)} \; \underbrace{\psi_2(x_2)}_{P(x_2)}$$

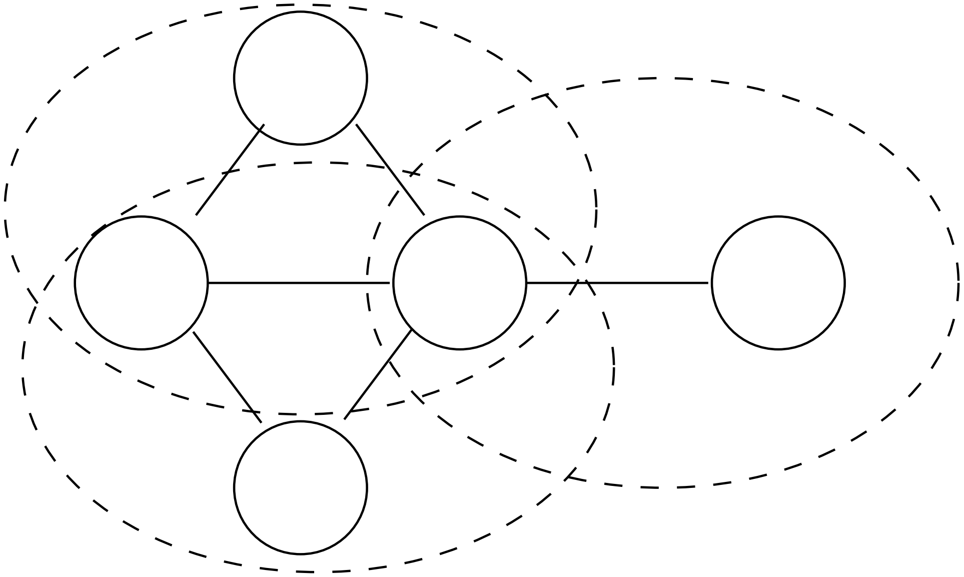
- A Markov chain



$$P(x_1, x_2, x_3) = \underbrace{\frac{1}{Z}}_{\substack{1 \\ 1}} \; \underbrace{\psi_{12}(x_1, x_2)}_{\substack{P(x_1, x_2) \\ P(x_1|x_2)}} \; \underbrace{\psi_{23}(x_2, x_3)}_{\substack{P(x_3|x_2) \\ P(x_2, x_3)}}$$

Preliminaries: cliques

- A *clique* is any maximal fully connected subset of nodes in the graph



(cliques are circled in the figure)

Markov random fields

- Hammersley-Clifford factorization theorem:

  Theorem: Any distribution consistent with the undirected graph must factor according to the cliques in the graph

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

  where Z is a *global normalization* constant and $\mathbf{x}_c$ is the set of variables (nodes) associated with clique c.

- The non-negative factors $\psi_c(\mathbf{x}_c)$ that depend only on variables within each clique are known as *potential functions*

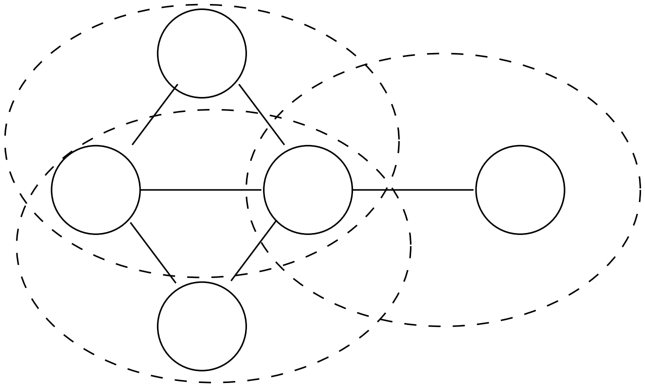
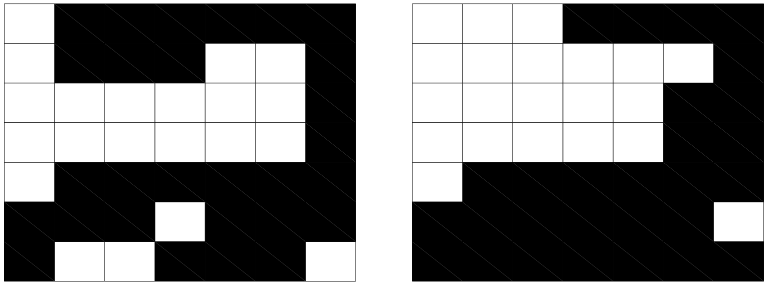


Image reconstruction example

- Modeling images with *Boltzmann machines*

- nearby pixels in images should be correlated



- we can capture such *nearest neighbor* dependences with the following lattice model