6.891 Machine learning and neural networks

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

LECTURE 18 (November 14, 2000)

Future lectures:

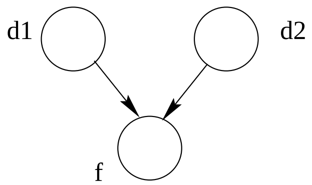
Lecture 19	–	Thu Nov 16
Lecture 20	–	Tue Nov 21
No lecture (Thanksgiving)	–	Thu Nov 23
No lecture	–	Tue Nov 28
Lecture 21 (Prof. Leslie Kaelbling)	–	Thu Nov 30

Topics

- Bayesian networks example (medical diagnosis)
  - model assumptions
  - the inference problem
  - sampling algorithm
- Markov random fields (undirected models)
  - graph semantics
  - associated probability distribution

Basics from last lecture

- A simple medical diagnosis example: two diseases marginally independent diseases d_1 and d_2 and one corresponding finding f

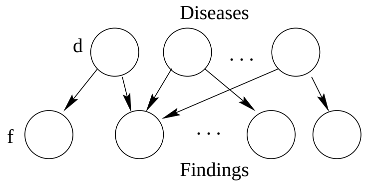


$P(d_1)$: $\boxed{\begin{array}{c} 0.9 \\ 0.1 \end{array}}$ $P(d_2)$: $\boxed{\begin{array}{c} 0.8 \\ 0.2 \end{array}}$ $P(f|d_1, d_2)$: $\boxed{\begin{array}{cccc} 0.5 & 0.6 & 0.2 & 0.9 \\ 0.5 & 0.4 & 0.8 & 0.1 \end{array}}$

(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)

A real medical diagnosis example

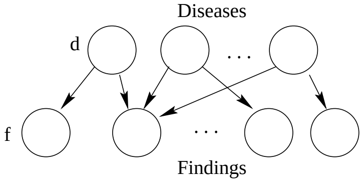
- The QMR-DT model (Shwe et al. 1991)



- 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



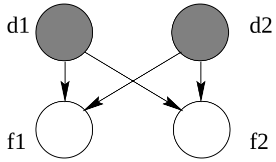
- Do we have all the relevant variables (e.g., significant diseases)?
- Assumptions that are explicit in the graph:
  - marginal independence of diseases
  - conditional independence of findings
- Assumptions about the underlying probability distribution:
  - causal independence assumptions

Assumptions in detail

- Diseases are marginally independent

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

- The findings are conditionally independent given the diseases



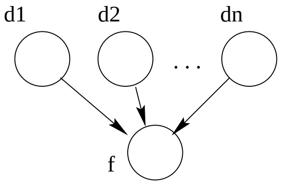
f_1 = Bone X-ray fracture
f_2 = ...

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

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



where 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 generated the finding

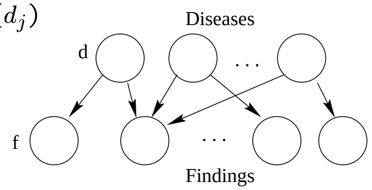
Joint distribution

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

$$\begin{aligned} P(f, d) &= \left[\prod_{i=1}^{m} P(f_i|d_{pa_i}) \right] \left[\prod_{j=1}^{n} P(d_j) \right] \\ P(f_i &= 0|d_{pa_i}) = (1 - q_{i0}) \prod_{j \in pa_i} (1 - q_{ij})^{d_j} \end{aligned}$$

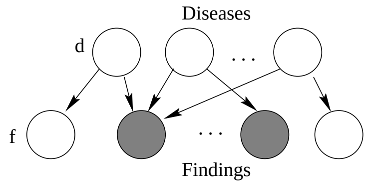
where d_{pa_i} is the set of diseases associated with finding f_i

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



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?

$$d^* = \arg\max_d P(d|f^*) = \arg\max_d P(f^*, d) \quad (1)$$

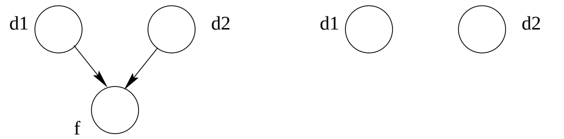
2. What are the marginal posterior probabilities

$$P(d_i = 1|f^*), \;\; i = 1, \ldots, n \quad (2)$$

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

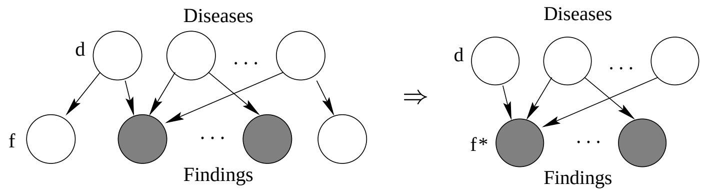
Inference problem cont'd

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



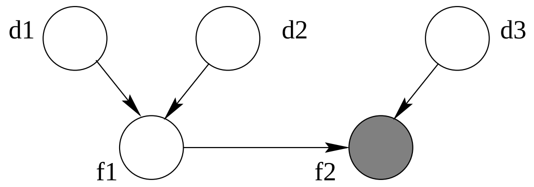
unobserved finding   no associated findings

- 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

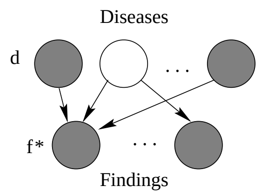
- We can try to find the most likely disease configuration given f^* via a search algorithm

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

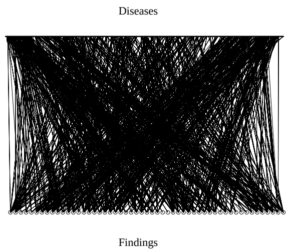
$$d_j^* \leftarrow \arg\max_{d_j} P(f^*, d_1^*, \ldots, d_{j-1}^*, d_j, d_{j+1}^*, \ldots, d_n^*) \quad (3)$$

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



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 they are present given the findings)

$$P(d_i = 1|f^*) = \frac{P(f^*, d_i = 1)}{P(f^*)} = \frac{\sum_d d_i P(f^*, d)}{\sum_d P(f^*, d)} \quad (4)$$

- 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 con'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?

$$P(d_i = 1|f^*) \approx \frac{1}{T}\sum_{t=1}^{T} d_i^t \quad (5)$$

  where each $d^t = \{d_1^t, \ldots, d_n^t\}$ is an independent sample from the posterior $P(d|f^*)$
- The problem is that we cannot easily sample from $P(d|f^*)$...

Second inference problem

- Importance sampling method:
  We draw samples from a much simpler proposal distribution $Q(d)$

$$P(f^*) = \sum_d P(f^*, d) \approx \frac{1}{T}\sum_{t=1}^{T} \frac{P(f^*, d^t)}{Q(d^t)} \quad (6)$$

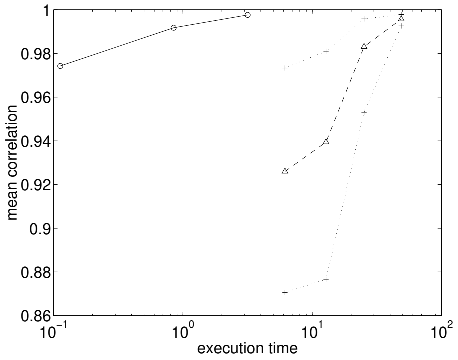
$$P(f^*, d_i = 1) = \sum_d d_i P(f^*, d) \approx \frac{1}{T}\sum_{t=1}^{T} d_i^t \frac{P(f^*, d^t)}{Q(d^t)} \quad (7)$$

  where $d^t = \{d_1^t, \ldots, d_n^t\}$ is now an independent sample from $Q(d)$
- We get the desired posterior marginals by ratios of these estimates (known as likelihood weighted sampling)

$$P(d_i = 1|f^*) = \frac{P(f^*, d_i = 1)}{P(f^*)} \approx \frac{\frac{1}{T}\sum_{t=1}^{T} d_i^t \frac{P(f^*, d^t)}{Q(d^t)}}{\frac{1}{T}\sum_{t=1}^{T} \frac{P(f^*, d^t)}{Q(d^t)}} \quad (8)$$

Second inference problem cont'd

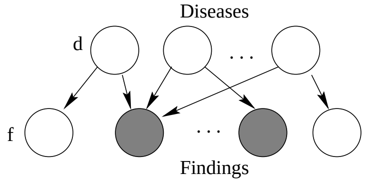
- This actually works...



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 how *uncertain* the outcomes of the findings are given the observed ones f^*

$$P(f_i|f^*) = \sum_d P(f_i|d_{pa_i})\, P(d|f^*) \tag{9}$$

as well as what the effect of observing f_i would be on the diseases

$$P(d|f_i, f^*) = \frac{P(d, f_i, f^*)}{P(f_i, f^*)} \tag{10}$$

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 results in the lowest entropy of the diseases (on average) given a hypothetical finding

$$\sum_{f_i=0,1} P(f_i|f^*)\, H(d|f_i, f^*) \tag{11}$$

where $H(d|f_i, f^*)$ is the entropy of the hypothetical posterior $P(d|f_i, f^*)$ (assuming we would observe a particular value for f_i)

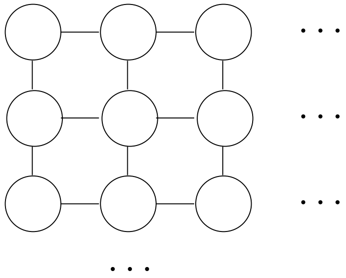
- The individual terms could be approximated similarly to the previous sampling method

Topics

- Markov random fields (undirected models)
  - graph semantics
  - associated probability distribution

Markov random fields

- Markov random fields are *undirected* graph models where the edges signify symmetric dependencies
- Example: a spin lattice with nearest neighbor dependencies

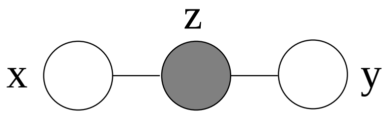


- Often used as models of physical processes, in computer vision, etc.

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

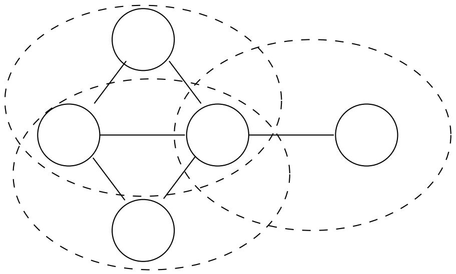


x and y are conditionally independent given z

- Again we'd like to find the most general form of the joint distribution that is consistent with the graph

Preliminaries: cliques

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



(cliques are circled in the figure)

Markov random fields

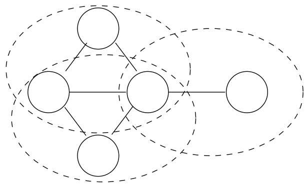
- Hammersley-Clifford factorization theorem:

  Theorem: The most general from of the probability distribution consistent with all the independence properties in the undirected graph must factor according to the cliques in the graph

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{c \in cliques} \psi_c(\mathbf{x}_c) \tag{12}$$

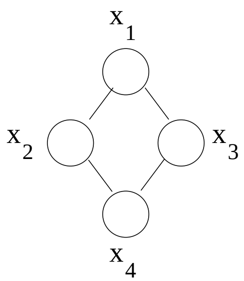
  where Z is a *global normalization* constant and $\mathbf{x}_c$ is the set of variables (nodes) associated with clique c.

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



Markov random fields

- Simple example:



  In this simple graph we have only pairwise cliques (cliques containing only two variables)

- The probability distribution consistent with the graph factors as

  $$P(x_1, x_2, x_3, x_4) = \psi_{12}(x_1, x_2)\, \psi_{13}(x_1, x_3)\, \psi_{24}(x_2, x_4)\, \psi_{34}(x_3, x_4)$$