6.891 Machine learning and neural networks

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

LECTURE 19 (November 16, 2000)

Problem set #4	–	available Tue Nov 21 due Tue Dec 5

Future lectures:

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

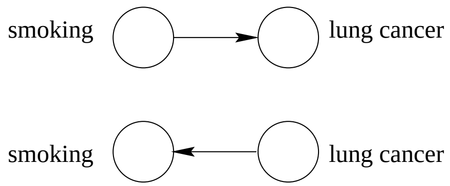
Topics

- Why do we need anything other than Bayesian networks?
- Markov random fields (undirected models)
  - graph semantics
  - associated probability distribution
- Image reconstruction example
  - Boltzmann machine
  - inference (Gibbs' sampling)
  - learning

Bayesian networks: interpretation

- "Probabilistic" semantics of Bayesian networks does not equal "causal interpretation"

  For example: the following two Bayesian networks are *identical* in the sense that they make the exact same independence assumptions about the variables (i.e., none)

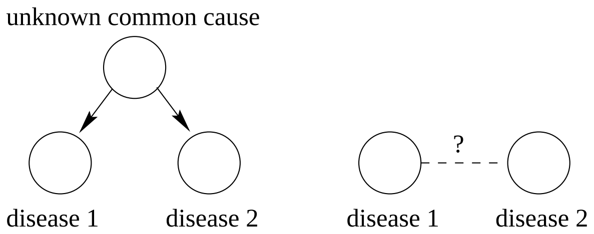


By Bayes rule

$$P(\text{smoking})P(\text{lung cancer}|\text{smoking}) = P(\text{smoking}|\text{lung cancer})P(\text{lung cancer}) \quad (1)$$

Unknown causes

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

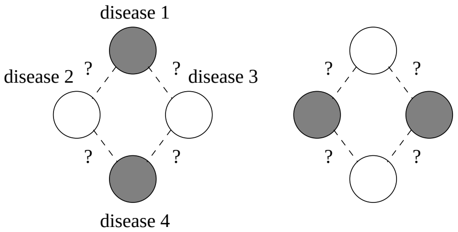


- Such symmetric interactions are better modeled with undirected graph models (Markov random fields)

Representational limitations of Bayesian networks

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

  For example: how should I draw the arrows in a Bayesian network

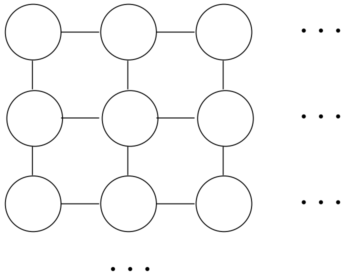


such that

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

Markov random fields

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

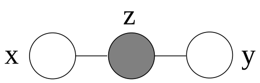


- As before, we have to
  - define graph semantics
  - associated probability distribution

Graph semantics

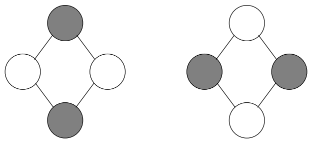
- The (conditional) independence properties can read from the graph via simple graph separation:

  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

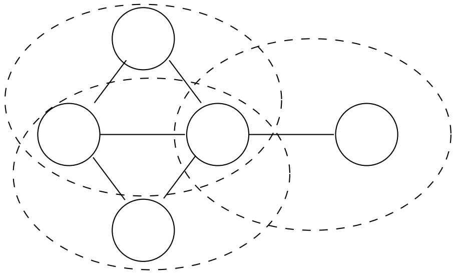
- This graph semantics captures our previous example



- 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

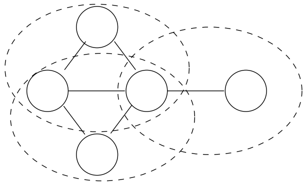
- Hammersley-Clifford factorization theorem:

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

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{c \in cliques} \psi_c(\mathbf{x}_c) \tag{2}$$

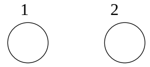
  where Z is a *global normalization* constant and $\mathbf{x}_c$ is the set of variables (nodes) associated with clique c.

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



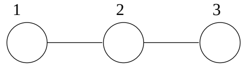
Markov random fields

- Simple independent example:



$$P(x_1, x_2) = \underbrace{\frac{1}{Z}}_{1} \; \underbrace{\psi_1(x_1)}_{P(x_1)} \; \underbrace{\psi_2(x_2)}_{P(x_2)} \tag{3}$$

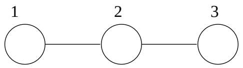
- A Markov chain



$$P(x_1, x_2, x_3) = \underbrace{\frac{1}{Z}}_{\substack{1 \\ 1}} \; \underbrace{\psi_{12}(x_1, x_2)}_{\substack{P(x_1, x_2) \\ P(x_1|x_2)}} \; \underbrace{\psi_{23}(x_2, x_3)}_{\substack{P(x_3|x_2) \\ P(x_2, x_3)}} \tag{4}$$

  The potential functions are not unique.

Potential functions and energy



- We can think of the potential functions associated with the cliques in terms of *energies*

$$\psi_{12}(x_1, x_2) = \exp\left(-E_{12}(x_1, x_2) \right) \tag{5}$$

  where $E_{12}(x_1, x_2)$ is the *energy* associated with a particular configuration of variables x_1 and x_2.

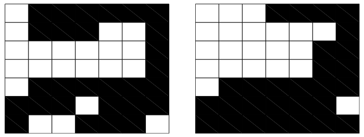
- Physical systems prefer low energy configurations (low energy $\Rightarrow$ high probability)
- For example, for binary variables x_1 and x_2 we might have

$$E_{12}(x_1, x_2) = -J_{12}\, x_1\, x_2 \tag{6}$$

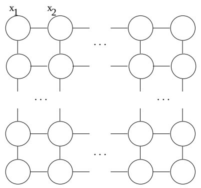
  where J_{12} is the *connection strength*. For large values of the connection strength, x_1 and x_2 become highly dependent.

Image reconstruction example

- Modeling images with *Boltzmann machines*
- nearby pixels in images should be correlated

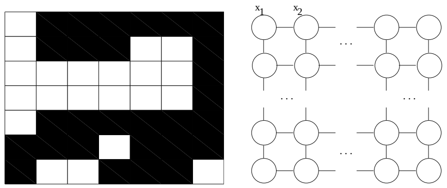


- we can capture such *nearest neighbor* dependences with the following lattice model



Example cont'd

- Each variable x_i in the model indicates whether a pixel is on ($x_i = 1$) or off ($x_i = 0$)



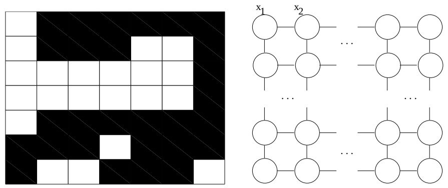
- For any pair of pixels linked with an edge in the graph we have the following potential function

$$\psi_{ij}(x_i, x_j) = \exp(\ J_{ij}\, x_i\, x_j + h_i\, x_i + h_j\, x_j\ ) \qquad (7)$$

where J_{ij} defines the *connection strength* between pixels i and j (large values of J_{ij} imply highly correlated pixels)

h_i permit us to bias the pixel values towards on or off

Example cont'd

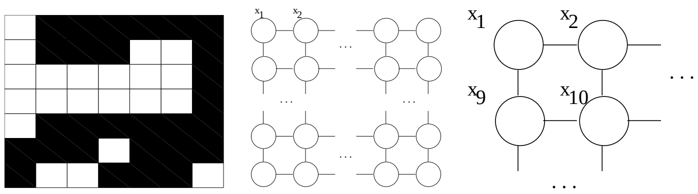


- The full joint distribution over the pixels in the image is given by

$$\begin{aligned} P(x_1, \ldots, x_{64} | J, h) &= \frac{1}{Z(J,h)} \prod_{\text{edges}\, ij} \exp(\ J_{ij}\, x_i\, x_j + h_i\, x_i + h_j\, x_j\ ) \\ &= \frac{1}{Z(J,h)} \exp\left(\sum_{\text{edges}\, ij} J_{ij}\, x_i\, x_j + \sum_i \tilde{h}_i\, x_i \right) \qquad (8) \end{aligned}$$

(here, e.g., $\tilde{h}_i = 4h_i$ for pixels away from the borders)

Example cont'd



- The graph structure implies a number *Markov properties* such as

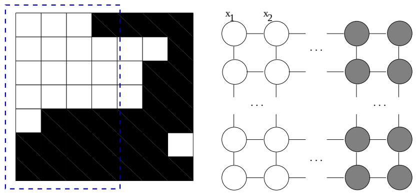
$$P(x_1 | x_2, \ldots, x_{64}, J, h) = P(x_1 | x_2, x_9, J, h) \qquad (9)$$

- The specific potential functions we have used also turn these conditional probabilities into logistic regression models (verify)

$$P(x_1 = 1 | x_2, x_9, J, h) = g(J_{12} x_2 + J_{19} x_9 + \tilde{h}_1) \qquad (10)$$

where $g(z) = (1 + e^{-z})^{-1}$ is the logistic function.

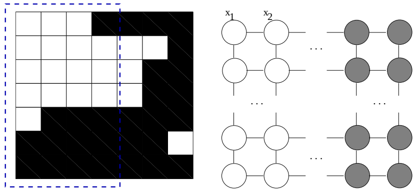
The reconstruction problem



- Part of the image may be corrupted or missing (pixel values are unknown)
- We can use the probability model to infer what the missing pixels are (complete the image)
- Two types of inference problems (as in medical diagnosis):
  1. Find the most likely configuration of the missing pixels given the observed ones
  2. Evaluate the marginal posterior probabilities for the missing pixels

  (we'll only look at the second one here)

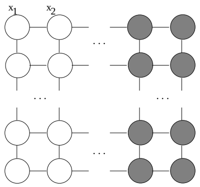
Second inference problem



- Assuming half the image is missing, we have 2^{32} possible configurations of the missing pixels
- We could apply the sampling method we already know about (importance sampling + likelihood weighting) to evaluate the posterior probabilities as in the context of medical diagnosis
- For Boltzmann machines there's a more natural sampling approach: Gibbs' sampling

Gibbs' sampling

- Given a set of known pixel values $x^* = \{x_i^*\}$ we wish to evaluate the posterior probabilities $P(x_i|\mathbf{x}^*, J, h)$ for all the missing pixels



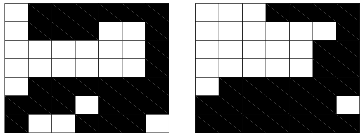
- We can do this via Gibbs' sampling:
  1. Initialize all the missing pixels (e.g., set $x_i = 0$)
  2. Sequentially sample a new value for each missing pixel x_i based on the current setting of its *neighbors* (those connected to it)
  $$x_i \sim P(x_i|x_{nb_i}, J, h)$$
  In our case, these conditional probabilities are logistic regression models
  $$P(x_i = 1|x_{nb_i}, J, h) = g\left(\sum_{j \in nb_i} J_{ij}x_j + \tilde{h}_i\right) \tag{11}$$

Gibbs' sampling cont'd

- Gibbs' sampling method generates a sequence of new pixel values
$$x_i^1, x_i^2, \ldots, x_i^t, x_i^{t+1}, \ldots \tag{12}$$
where x_i^t is the sampled value for pixel i at the t^{th} iteration of the sampler (we assume that one iteration means generating a new value for each missing pixel)
- The posterior mean (expected value) of pixel i can be evaluated approximately from these samples
$$E\{\, x_i \,|\, x^*, J, h\} \approx \frac{1}{T}\sum_{t=1}^{T} x_i^t \tag{13}$$

Learning problem

- How do we set the connection strengths J_{ij} in the Boltzmann machine to get a good reconstruction?
- Given a training set of images $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ we can set J_{ij} via maximum likelihood estimation



The connection strengths can be updated via gradient ascent
$$J_{ij} \leftarrow J_{ij} + \epsilon \frac{\partial}{\partial J_{ij}} \log P(\mathbf{x}_t|J, h) \tag{14}$$
so long as we can evaluate the gradients for any particular training image $\mathbf{x}_t$ (vector of known pixel values)

Learning problem cont'd

$$\begin{aligned}
\frac{\partial}{\partial J_{ij}} \log P(\mathbf{x}_t|J,h) &= \frac{\partial}{\partial J_{ij}} \left[\sum_{\text{edges}\, ij} J_{ij} x_i^t x_j^t + \sum_i \tilde{h}_i\, x_i^t - \log Z(J,h) \right] \\
&= x_i^t x_j^t - \frac{\partial}{\partial J_{ij}} \log Z(J,h) \\
&\dots \\
&= x_i^t x_j^t - E\{\, x_i x_j \mid J,h \,\} \qquad (15)
\end{aligned}$$

where $E\{\, x_i x_j \mid J,h \,\}$ is the expected co-occurence of the pixel values based on our current model (setting of J and h).

- We can evaluate $E\{\, x_i x_j \mid J,h \,\}$ via Gibbs' sampling analogously to the posterior expectations $E\{x_i|x^*, J, h\}$