6.891 Machine learning and neural networks

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

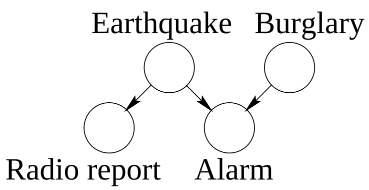
LECTURE 17 (November 9, 2000)

Topics

- Graph models
  - basics
  - examples
- Bayesian networks (directed models)
  - graph semantics
  - associated probability distribution
- Markov random fields (undirected models)
  - graph semantics
  - associated probability distribution

Graph models

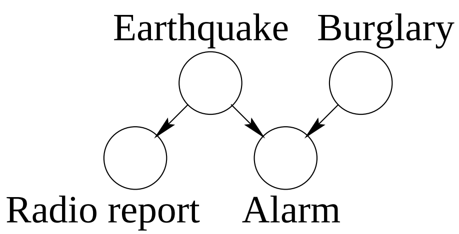
- In a graph model
  - the nodes in the graph correspond to variables
  - the edges signify dependencies among the variables
- A Bayesian network (acyclic directed graph model)



- We can think of the edges as relating "causes" and their "effects" (or "parents" to "children")
- The graph provides a qualitative description of the domain

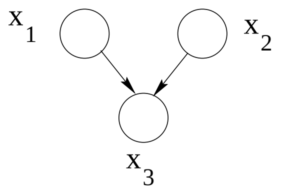
Qualitative inference

- Just by looking at the graph, we can determine what we can and cannot ignore
- Ability to ignore irrelevant information is at the heart of efficient reasoning
- Example:



Two levels of description

- Graphical models need two levels of specification
  1. Qualitative properties captured by a graph



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

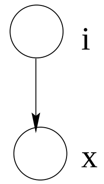
Here x_1 and x_2 are *marginally independent*

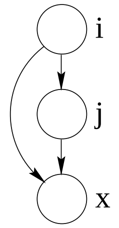
  2. Quantitative properties specified by the associated probability distribution

$$P(x_1, x_2, x_3) = P(x_1)\, P(x_2)\, P(x_3|x_1, x_2) \quad (1)$$

where, e.g., $P(x_1 = heads) = 0.5$,
$P(x_3 = same|x_1 = heads, x_2 = tails) = 0$

Examples

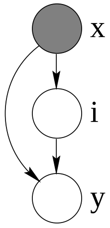


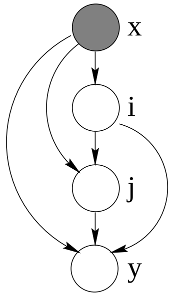


Mixture model hierarchical mixture model

- i and j correspond to the discrete choices in the mixture model
- $\mathbf{x}$ is the (vector) variable whose density we wish to model
- Note that we cannot tell what the component distributions $P(\mathbf{x}|i)$ are based on the graph

Examples cont'd

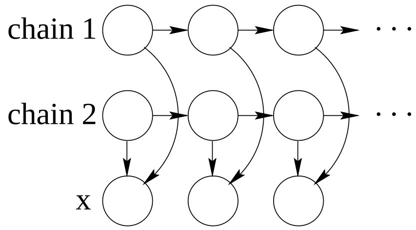


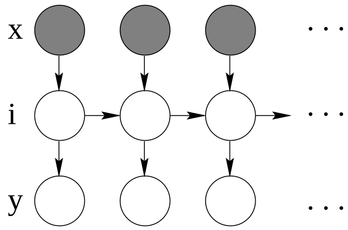


Mixture of experts hierarchical mixture of experts

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

Examples cont'd

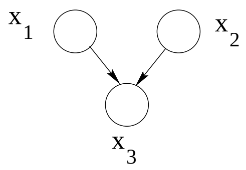




Factorial HMM input-output HMM

- In factorial HMMs, independent processes conspire to generate the observed output (here output is $\mathbf{x}$ rather than O as in earlier HMM examples)
- In input-output HMMs, any observed sequence of outputs y is accompanied by a corresponding sequence of *inputs* $\mathbf{x}$
  - the model tries to tranform the input sequence into the output sequence (markov?)

Graph model specification

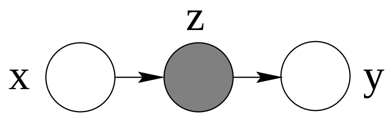


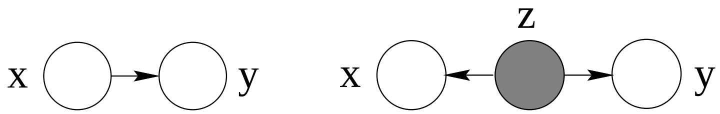
$x_1 =$ first coin toss
$x_2 =$ second coin toss
$x_3 =$ same?

- We need to address the following questions
  1. What is the graph semantics? In other words, how do we read the qualitative properties from the graph?
  2. What type of probability distribution can be associated with any specific graph? Surely the graph must be *consistent* with the associated probability distribution
  3. How can we exploit the graph in making quantitative inferences (computing posterior or other probabilities for various events)?

Graph semantics

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



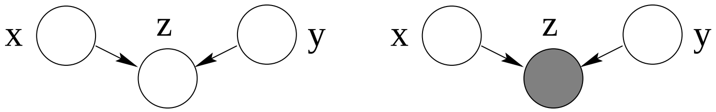


x and y are dependent

x and y are conditionally independent given z

Graph semantics cont'd

- Here are the interesting cases...



x and y are marginally independent

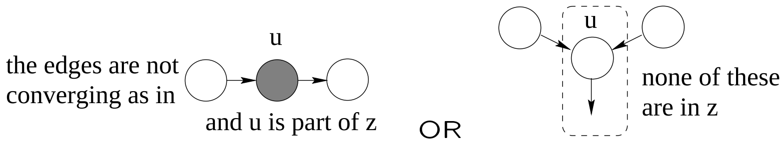
x and y are conditionally dependent

- 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

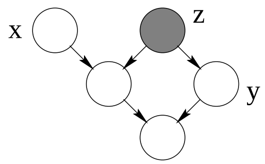
Graph separation criterion (briefly)

- D-separation criterion (D for Directed edges):

  Definition: sets of nodes (variables) x and y are D-separated (conditionally independent) given another set z if along any path between x and y we can find a node u such that either

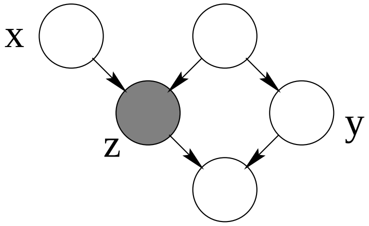


- "All paths between x and y must be blocked"

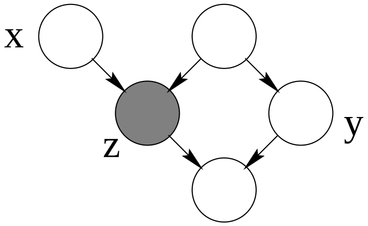


D-separation: example

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



Towards quantitative specification



- Separation properties in the graph imply independence properties between the associated variables
- Can we always find an underlying probability distribution that is *consistent* with all the independence properties we can derive from the graph?

  Consistency here means

  Separated in the graph $\Rightarrow$ independent in $P(x, y, z, \ldots)$

Bayesian networks

- Factorization theorem:

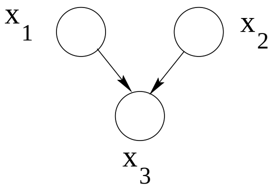
  Theorem: The most general form of the probability distribution consistent with the graph factors according to "node given its parents":

$$P(\mathbf{x}) = \prod_{i=1}^{d} P(x_i|\mathbf{x}_{pa_i}) \qquad (2)$$

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

Bayesian networks: example

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



  is given by

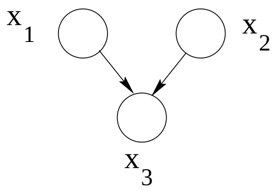
$$P(x_1, x_2, x_3) = P(x_1)\, P(x_2)\, P(x_3|x_1, x_2) \qquad (3)$$

- Note that this includes, e.g.,

$$P(x_1, x_2, x_3) = P(x_1)\, P(x_2)\, P(x_3), \text{ or} \qquad (4)$$
$$P(x_1, x_2, x_3) = P(x_1)\, P(x_2)\, P(x_3|x_1) \qquad (5)$$

Bayesian networks

Some additional properties:



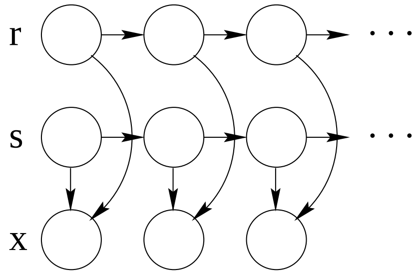
$$P(x_1, x_2, x_3) = P(x_1)\, P(x_2)\, P(x_3|x_1, x_2) \tag{6}$$

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

  $P(x_3|x_1, x_2)$ can be a full probability table or a logistic regression model

Bayesian network: example

- A factorial HMM



- The associated probability model has the following form

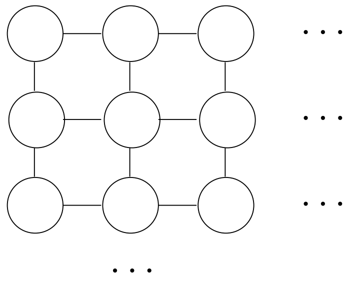
$$\begin{aligned} P(r_0, s_0, \mathbf{x}_0, r_1, s_1, \mathbf{x}_1, \ldots) &= P_0(r_0)\, P_1(r_1|r_0) \cdots \\ &\quad \times P_0(s_0)\, P_1(s_1|s_0) \cdots \\ &\quad \times P_o(\mathbf{x}_0|r_0, s_0)\, P_o(\mathbf{x}_1|r_1, s_1) \cdots \end{aligned}$$

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

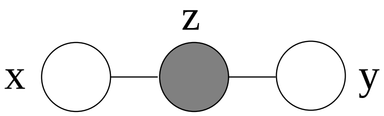


- Often used as models of physical processes, in computer vision, etc.

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

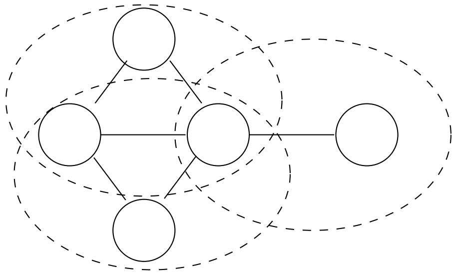


x and y are conditionally independent given z

- 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

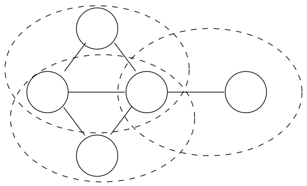
- Hammersley-Clifford factorization theorem:

  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

$$P(\mathbf{x}) = \frac{1}{Z} \prod_{c \in cliques} \psi_c(\mathbf{x}_c) \tag{7}$$

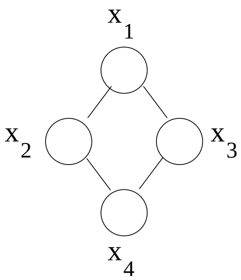
  where Z is a *global normalization* constant and $\mathbf{x}_c$ is the set of variables (nodes) associated with clique c.

- 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

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