Inference and Representation

David Sontag

New York University

Lecture 3, Sept. 22, 2015

Bayesian networks

Reminder of first lecture

- A **Bayesian network** is specified by a directed *acyclic* graph $G = (V, E)$ with:
  1. One node $i \in V$ for each random variable X_i
  2. One conditional probability distribution (CPD) per node, $p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$, specifying the variable's probability conditioned on its parents' values
- Corresponds 1-1 with a particular factorization of the joint distribution:

$$p(x_1, \ldots x_n) = \prod_{i \in V} p(x_i \mid \mathbf{x}_{\mathrm{Pa}(i)})$$

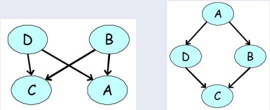
- Powerful framework for designing *algorithms* to perform probability computations

Bayesian networks have limitations

- Recall that G is a **perfect map** for distribution p if $I(G) = I(p)$
- **Theorem:** Not every distribution has a perfect map as a DAG

Proof.

(By counterexample.) There is a distribution on 4 variables where the only independencies are $A \perp C \mid \{B, D\}$ and $B \perp D \mid \{A, C\}$. This cannot be represented by any Bayesian network.



(a) (b)

Both (a) and (b) encode $(A \perp C|B, D)$, but in both cases $(B \not\perp D|A, C)$. $\square$

Example

- Let's come up with an example of a distribution p satisfying $A \perp C \mid \{B, D\}$ and $B \perp D \mid \{A, C\}$
- A=Alex's hair color (red, green, blue)
  B=Bob's hair color
  C=Catherine's hair color
  D=David's hair color
- Alex and Bob are friends, Bob and Catherine are friends, Catherine and David are friends, David and Alex are friends
- Friends *never* have the same hair color!

Undirected graphical models

- An alternative representation for joint distributions is as an **undirected graphical model**
- As in BNs, we have one node for each random variable
- Rather than CPDs, we specify (non-negative) **potential functions** over sets of variables associated with cliques C of the graph,

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \phi_c(\mathbf{x}_c)$$

  Z is the **partition function** and normalizes the distribution:

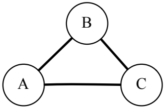
$$Z = \sum_{\hat{x}_1, \ldots, \hat{x}_n} \prod_{c \in C} \phi_c(\hat{\mathbf{x}}_c)$$

- Like CPD's, $\phi_c(\mathbf{x}_c)$ can be represented as a table, but it is *not normalized*
- Also known as **Markov random fields** (MRFs) or Markov networks

Undirected graphical models

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \phi_c(\mathbf{x}_c), \qquad Z = \sum_{\hat{x}_1, \ldots, \hat{x}_n} \prod_{c \in C} \phi_c(\hat{\mathbf{x}}_c)$$

Simple example (potential function on each edge encourages the variables to take the same value):



$$\phi_{A,B}(a,b) = \begin{array}{cc|c|c|} & & \multicolumn{2}{c}{B} \\ & & 0 & 1 \\ \cline{3-4} \multirow{2}{*}{A} & 0 & 10 & 1 \\ \cline{3-4} & 1 & 1 & 10 \\ \cline{3-4} \end{array} \qquad \phi_{B,C}(b,c) = \begin{array}{cc|c|c|} & & \multicolumn{2}{c}{C} \\ & & 0 & 1 \\ \cline{3-4} \multirow{2}{*}{B} & 0 & 10 & 1 \\ \cline{3-4} & 1 & 1 & 10 \\ \cline{3-4} \end{array} \qquad \phi_{A,C}(a,c) = \begin{array}{cc|c|c|} & & \multicolumn{2}{c}{C} \\ & & 0 & 1 \\ \cline{3-4} \multirow{2}{*}{A} & 0 & 10 & 1 \\ \cline{3-4} & 1 & 1 & 10 \\ \cline{3-4} \end{array}$$

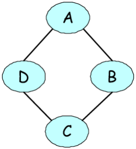
$$p(a, b, c) = \frac{1}{Z} \phi_{A,B}(a, b) \cdot \phi_{B,C}(b, c) \cdot \phi_{A,C}(a, c),$$

where

$$Z = \sum_{\hat{a}, \hat{b}, \hat{c} \in \{0,1\}^3} \phi_{A,B}(\hat{a}, \hat{b}) \cdot \phi_{B,C}(\hat{b}, \hat{c}) \cdot \phi_{A,C}(\hat{a}, \hat{c}) = 2 \cdot 1000 + 6 \cdot 10 = 2060.$$

Hair color example as a MRF

- We now have an **undirected** graph:



- The joint probability distribution is parameterized as

$$p(a,b,c,d) = \frac{1}{Z}\phi_{AB}(a,b)\phi_{BC}(b,c)\phi_{CD}(c,d)\phi_{AD}(a,d)\ \phi_A(a)\phi_B(b)\phi_C(c)\phi_D(d)$$

- **Pairwise potentials** enforce that no friend has the same hair color:

$$\phi_{AB}(a,b) = 0 \text{ if } a = b, \quad \text{and 1 otherwise}$$

- **Single-node potentials** specify an affinity for a particular hair color, e.g.

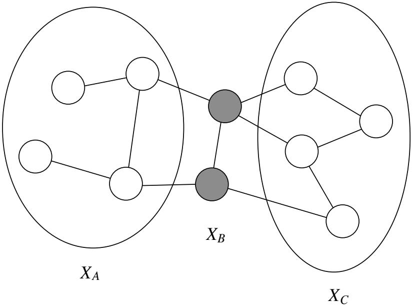
$$\phi_D(\text{"red"}) = 0.6, \quad \phi_D(\text{"blue"}) = 0.3, \quad \phi_D(\text{"green"}) = 0.1$$

The normalization Z makes the potentials **scale invariant**! Equivalent to

$$\phi_D(\text{"red"}) = 6, \quad \phi_D(\text{"blue"}) = 3, \quad \phi_D(\text{"green"}) = 1$$

Markov network structure implies conditional independencies

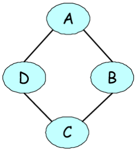
- Let G be the undirected graph where we have one edge for every pair of variables that appear together in a potential
- Conditional independence is given by **graph separation**!



- $X_{\mathbf{A}} \perp X_{\mathbf{C}} \mid X_{\mathbf{B}}$ if there is no path from $a \in \mathbf{A}$ to $c \in \mathbf{C}$ after removing all variables in $\mathbf{B}$

Example

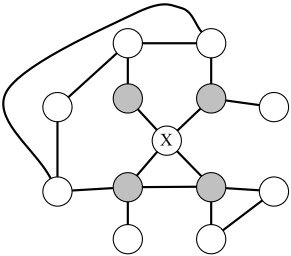
- Returning to hair color example, its undirected graphical model is:



- Since removing A and C leaves no path from D to B, we have $D \perp B \mid \{A, C\}$
- Similarly, since removing D and B leaves no path from A to C, we have $A \perp C \mid \{D, B\}$
- No other independencies implied by the graph

Markov blanket

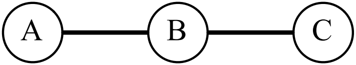
- A set **U** is a **Markov blanket** of X if $X \notin \mathbf{U}$ and if **U** is a minimal set of nodes such that $X \perp (\mathcal{X} - \{X\} - \mathbf{U}) \mid \mathbf{U}$
- In undirected graphical models, the Markov blanket of a variable is precisely its **neighbors** in the graph:



- In other words, X is independent of the rest of the nodes in the graph given its immediate neighbors

Proof of independence through separation

- We will show that $A \perp C \mid B$ for the following distribution:



$$p(a,b,c) = \frac{1}{Z}\phi_{AB}(a,b)\phi_{BC}(b,c)$$

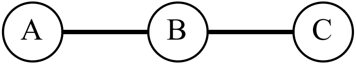
- First, we show that $p(a \mid b)$ can be computed using only $\phi_{AB}(a,b)$:

$$\begin{aligned} p(a \mid b) &= \frac{p(a,b)}{p(b)} \\ &= \frac{\frac{1}{Z}\sum_{\hat{c}}\phi_{AB}(a,b)\phi_{BC}(b,\hat{c})}{\frac{1}{Z}\sum_{\hat{a},\hat{c}}\phi_{AB}(\hat{a},b)\phi_{BC}(b,\hat{c})} \\ &= \frac{\phi_{AB}(a,b)\sum_{\hat{c}}\phi_{BC}(b,\hat{c})}{\sum_{\hat{a}}\phi_{AB}(\hat{a},b)\sum_{\hat{c}}\phi_{BC}(b,\hat{c})} = \frac{\phi_{AB}(a,b)}{\sum_{\hat{a}}\phi_{AB}(\hat{a},b)}. \end{aligned}$$

- More generally, the probability of a variable conditioned on its Markov blanket depends *only* on potentials involving that node

Proof of independence through separation

- We will show that $A \perp C \mid B$ for the following distribution:



$$p(a, b, c) = \frac{1}{Z}\phi_{AB}(a, b)\phi_{BC}(b, c)$$

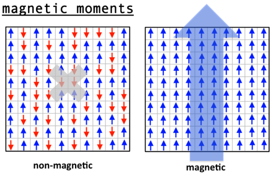
Proof.

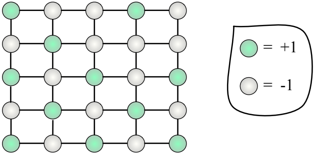
$$\begin{aligned}
p(a, c \mid b) = \frac{p(a, c, b)}{\sum_{\hat{a},\hat{c}} p(\hat{a}, b, \hat{c})} &= \frac{\phi_{AB}(a, b)\phi_{BC}(b, c)}{\sum_{\hat{a},\hat{c}} \phi_{AB}(\hat{a}, b)\phi_{BC}(b, \hat{c})} \\
&= \frac{\phi_{AB}(a, b)\phi_{BC}(b, c)}{\sum_{\hat{a}} \phi_{AB}(\hat{a}, b) \sum_{\hat{c}} \phi_{BC}(b, \hat{c})} \\
&= p(a \mid b)p(c \mid b)
\end{aligned}$$

$\square$

Example: Ising model

- Invented by the physicist Wilhelm Lenz (1920), who gave it as a problem to his student Ernst Ising
- Mathematical model of ferromagnetism in statistical mechanics
- The spin of an atom is biased by the spins of atoms nearby on the material:

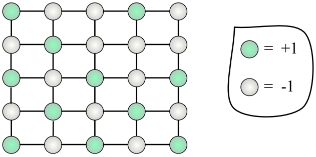




- Each atom $X_i \in \{-1, +1\}$, whose value is the direction of the atom spin
- If a spin at position i is $+1$, what is the probability that the spin at position j is also $+1$?
- Are there phase transitions where spins go from "disorder" to "order"?

Example: Ising model

- Each atom $X_i \in \{-1, +1\}$, whose value is the direction of the atom spin
- The spin of an atom is biased by the spins of atoms nearby on the material:



$$p(x_1, \cdots, x_n) = \frac{1}{Z} \exp\left(\sum_{i<j} w_{i,j} x_i x_j - \sum_i u_i x_i \right)$$

- When $w_{i,j} > 0$, nearby atoms encouraged to have the same spin (called **ferromagnetic**), whereas $w_{i,j} < 0$ encourages $X_i \neq X_j$
- Node potentials $\exp(-u_i x_i)$ encode the bias of the individual atoms
- Scaling the parameters makes the distribution more or less spiky