Graphical Models

Lecture 4:
Undirected Graphical Models

Andrew McCallum
mccallum@cs.umass.edu

Thanks to Noah Smith and Carlos Guestrin for some slide materials.
Administrivia

• HW#1:
  – Source code was due Tuesday by 5pm.
  – Reports due today (Thursday) 5pm.
  – Heard of some successes.
  – Comments?
Motivating Example:  
No Bayesian Network is a P-Map

- *Misunderstanding students*

$I(P)$:
- $A \perp C \mid B, D$
- $B \perp D \mid A, C$
- $\neg B \perp D$
- $\neg A \perp C$

Want to represent preference for pairwise affinities
Undirected Graphical Models

• Also known as **Markov networks** and **Markov random fields**.

• Alternative representation of graphs with probability distributions.
  – Motivation
  – Definition
  – Independence
  – Representation theorems
  – I-maps and P-maps
Motivating Example:
This Markov Network is a P-Map!

- *Misunderstanding students*

$I(P)$:
- $A \perp C \mid B, D$
- $B \perp D \mid A, C$
- $\neg B \perp D$
- $\neg A \perp C$

(Will explain soon why it is a P-Map.)

“affinity functions” or “compatibility functions” on edges.
Call these “factors”
Markov Networks

• Each random variable is a vertex.
• Undirected edges.
• **Factors** are associated with edges (or more generally subsets of nodes that form cliques).
  – A factor maps assignments of its nodes to nonnegative “compatibility” values.
Markov Networks

- In this example, associate a factor with each edge.
  - Could also have factors for single nodes!
Markov Networks

- Probability distribution:

\[ P(a, b, c, d) \propto \phi_1(a, b) \phi_2(b, c) \phi_3(c, d) \phi_4(a, d) \]

\[ P(a, b, c, d) = \frac{\phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(a, d)}{\sum_{a',b',c',d'} \phi_1(a', b')\phi_2(b', c')\phi_3(c', d')\phi_4(a', d')} \]

\[ Z = \sum_{a',b',c',d'} \phi_1(a', b')\phi_2(b', c')\phi_3(c', d')\phi_4(a', d') \]

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>( \phi_1(A, B) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B</th>
<th>C</th>
<th>( \phi_2(B, C) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C</th>
<th>D</th>
<th>( \phi_3(C, D) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>D</th>
<th>( \phi_4(A, D) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>
Markov Networks

- Probability distribution:

\[
P(a, b, c, d) \propto \phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(a, d)
\]

\[
P(a, b, c, d) = \frac{\phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(a, d)}{\sum_{a', b', c', d'} \phi_1(a', b')\phi_2(b', c')\phi_3(c', d')\phi_4(a', d')}
\]

\[
Z = \sum_{a', b', c', d'} \phi_1(a', b')\phi_2(b', c')\phi_3(c', d')\phi_4(a', d')
\]

\[
= 7,201,840
\]
Markov Networks

• Probability distribution:

\[ P(a, b, c, d) \propto \phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(a, d) \]

\[ P(a, b, c, d) = \frac{\phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(a, d)}{\sum_{a', b', c', d'} \phi_1(a', b')\phi_2(b', c')\phi_3(c', d')\phi_4(a', d')} \]

\[ Z = \sum_{a', b', c', d'} \phi_1(a', b')\phi_2(b', c')\phi_3(c', d')\phi_4(a', d') \]

\[ = 7,201,840 \]

\[ P(0, 1, 1, 0) = \frac{5,000,000}{Z} = 0.69 \]
Markov Networks

• Probability distribution:

\[ P(a, b, c, d) \propto \phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(a, d) \]

\[ P(a, b, c, d) = \frac{\phi_1(a, b)\phi_2(b, c)\phi_3(c, d)\phi_4(a, d)}{\sum_{a', b', c', d'} \phi_1(a', b')\phi_2(b', c')\phi_3(c', d')\phi_4(a', d')} \]

\[ Z = \sum_{a', b', c', d'} \phi_1(a', b')\phi_2(b', c')\phi_3(c', d')\phi_4(a', d') \]

= 7,201,840

Are factors on edges enough?
No!
Think about P with no ind.
Fully connected.
Think about # parameters.
Up to us to define alternative.

\[ P(1, 1, 0, 0) = \frac{10}{Z} = 0.0000014 \]
Administrivia

• CSCF mailing lists for HW submission were a disaster.

• http://nescai.cs.umass.edu/cs691/
  – username: cs691
  – password: *******
Markov Networks
(General Form)

- Let $D_i$ denote the set of variables (subset of $X$) in the $i$th clique.
- Probability distribution is a **Gibbs** distribution:

$$P(X) = \frac{U(X)}{Z}$$

$$U(X) = \prod_{i=1}^{m} \phi_i(D_i)$$

$$Z = \sum_{x \in \text{Val}(X)} U(x)$$

Compare to directed! $Z = 1$
Notes

• Z might be hard to calculate.
  – “Normalization constant”
  – “Partition function”

• Can get efficient calculation in some cases.
  – This is an inference problem; it’s equivalent to marginalizing over everything.

• Ratios of probabilities are easy.

\[
\frac{P(x)}{P(x')} = \frac{U(x)/Z}{U(x')/Z} = \frac{U(x)}{U(x')}
\]
Pairwise Markov Networks

• All factors associated with one node or one pair (connected by an edge).

\[
P(X) = \prod_{i} \phi_i(X_i) \prod_{(i,j) \in H} \phi_{i,j}(X_i, X_j)
\]

• The graph may have cliques with more than two nodes, but they do not have factors.
Markov Networks Can Always Be Made Pairwise

- For any factor over three or more variables, introduce a new variable.
- \( \text{Val}(X) \) has a size that is the number of values the factor can take (exponential in values of neighbors).
- Local factor structure is lost.

Show example on board
Pairwise Markov Network Example

- Classify each pixel as foreground or background.

\[ \phi_i(X_i = \text{fg}, C_i) = \exp \left( -\frac{||C_i - \mu_{\text{fg}}||^2}{\sigma^2} \right) \]

\[ \phi_i(X_i = \text{bg}, C_i) = \exp \left( -\frac{||C_i - \mu_{\text{bg}}||^2}{\sigma^2} \right) \]

\[ \phi_{i,j}(X_i, X_j) = \begin{cases} 10 & \text{if } X_i = X_j \\ 1 & \text{otherwise} \end{cases} \]
Application: Image Segmentation

Figure 4.B.1 — Two examples of image segmentation results (a) The original image. (b) An oversegmentation known as superpixels; each superpixel is associated with a random variable that designates its segment assignment. The use of superpixels reduces the size of the problems. (c) Result of segmentation using node potentials alone, so that each superpixel is classified independently. (d) Result of segmentation using a pairwise Markov network encoding interactions between adjacent superpixels.
Reducing Factors

• Given some variables’ values, we can reduce the factors to that context.

• Resulting conditional distribution is still Gibbs. (New Z.)
Reducing Factors

• Given some variables’ values, we can **reduce** the factors to that context.

• Resulting conditional distribution is still Gibbs. (New Z.)
Reduced Markov Networks

coherence

difficulty

intelligence

grade

SAT

letter

job

happy
Reduced Markov Networks

coherence

difficulty

intelligence

SAT

letter

job

happy

Condition on “grade”
Reduced Markov Networks

coherence

difficulty

intelligence

Condition on “grade” and “SAT”

letter

job

happy
Independence in Markov Networks

• Given a set of observed nodes $Z$, a path $X_1$-$X_2$-$X_3$-$\ldots$-$X_k$ is active if no nodes on the path are observed.

• Two sets of nodes $X$ and $Y$ in $\mathcal{H}$ are separated given $Z$ if there is no active path between any $X_i \in X$ and any $Y_i \in Y$.
  
  – Denoted: $\text{sep}_{\mathcal{H}}(X, Y \mid Z)$

• Global Markov assumption:
  
  $\text{sep}_{\mathcal{H}}(X, Y \mid Z) \Rightarrow X \perp Y \mid Z$
Representation Theorems

• Bayesian networks ...

\[
P(X) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i))
\]

- Independencies give you the Bayesian network.
- Bayesian network reveals independencies.

The Bayesian network graph is an I-map for \(P\).
Representation Theorems

- Bayesian networks ...

\[ P(X) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i)) \]

The Bayesian network graph is an I-map for \( P \).

- Markov networks ...
Representation Theorems

• Bayesian networks ...

The Bayesian network graph is an I-map for $P$.

\[
P(X) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i))
\]

• Markov networks ...

The Markov network’s graph is an I-map for $P$.

\[
P(X) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(D_i)
\]
Representation Theorem (I)

\[ P(X) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(D_i) \]

- Gibbs distribution satisfies the independencies associated with the graph

- Factorization into \( D_i \) gives a simple way to build the graph:
  - put an edge between \( X \) and \( Y \) iff
    \[ \exists D_i \text{ such that } X, Y \in D_i. \]

The Markov network’s graph is an I-map for \( P \).
Representation Theorem (I)

• Assume Gibbs.
• Consider three disjoint sets of variables, $W$, $Y$, and $Z$, such that $\text{dsep}_{\mathcal{H}}(W, Y \mid Z)$.
  – For now assume these comprise all of $X$; general case is not hard.
• No edges between $W$ and $Y$, so any clique is either in $W \cup Z$ or $Y \cup Z$.

$$
P(X) = \frac{1}{Z} \left( \prod_{i:D_i \subseteq W \cup Z} \phi_i(D_i) \right) \left( \prod_{i:D_i \subseteq Y \cup Z} \phi_i(D_i) \right)
= \frac{1}{Z} \Phi_1(W, Z) \Phi_2(Y, Z)
$$

• It follows that $W \perp Y \mid Z$. 
Representation Theorem (II)

- Other direction?

The Markov network’s graph is an I-map for $P$.

$$P(X) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(D_i)$$
Representation Theorem (II)

- Fails!

The Markov network’s graph is an I-map for P.

\[ P(X) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(D_i) \]
**Example**

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.125</td>
</tr>
</tbody>
</table>

I-map for $P$
Example

I-map for $P$

$X_1 \perp X_3 \mid X_2, X_4$ and others hold in $P$. 

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.125</td>
</tr>
</tbody>
</table>

$P$
### Example

The distribution does not factorize into the graph’s cliques!

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.125</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.125</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.125</td>
</tr>
</tbody>
</table>
Representation Theorem (II)

- Succeeds if $P(x) > 0$ for all $x$.
- Hammersley-Clifford Theorem

The Markov network’s graph is an I-map for $P$ and $P$ is nonnegative.

$$P(X) = \frac{1}{Z} \prod_{i=1}^{m} \phi_i(D_i)$$
## Graphs and Independencies

<table>
<thead>
<tr>
<th></th>
<th>Bayesian Networks</th>
<th>Markov Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>local independencies</strong></td>
<td>local Markov assumption</td>
<td>?</td>
</tr>
<tr>
<td><strong>global independencies</strong></td>
<td>d-separation</td>
<td>separation</td>
</tr>
</tbody>
</table>
Local Independence Assumptions in Markov Networks

- **Separation** defines global independencies.
Local Independence Assumptions in Markov Networks

- **Pairwise** Markov independence: pairs of non-adjacent variables are independent given everything else.
Local Independence Assumptions in Markov Networks

- **Markov blanket**: each variable is independent of the rest given its *neighbors*.

Define *neighbors* on the board
Local Independence Assumptions in Markov Networks

- **Separation:**
  \[ \text{sep}_\mathcal{H}(W, Y \mid Z) \Rightarrow W \perp Y \mid Z \]

- **Pairwise Markov:**
  \[ A \perp B \mid X \setminus \{A, B\} \]

- **Markov blanket:**
  \[ A \perp X \setminus \text{Neighbors}(A) \mid \text{Neighbors}(A) \]
I-Maps

• Fully connected graph is an I-map (like BNs)
• *Minimal* I-maps (delete edge → not an I-map)
  – Not unique for Bayesian networks.
  – *What about Markov Networks?*...
  – Unique for Markov networks (if positive distribution)!
• Simple way to construct a minimal I-map:
  – Check each pairwise Markov assumption.
  – If it’s not entailed by P, add edge.
P-Maps

• Want: independencies from the graphical model are exactly the same as in P.
• Doesn’t always exist for Bayesian networks (misunderstanding students).
• What about Markov Networks?...
• Doesn’t always exist for Markov networks.

A ⊥ B
¬ A ⊥ B | C

Minimal I-map is not a P-map.
## Bayesian Networks and Markov Networks

<table>
<thead>
<tr>
<th></th>
<th>Bayesian Networks</th>
<th>Markov Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>local independencies</strong></td>
<td>local Markov assumption</td>
<td>pairwise; Markov blanket</td>
</tr>
<tr>
<td><strong>global independencies</strong></td>
<td>d-separation</td>
<td>separation</td>
</tr>
<tr>
<td><strong>relative advantages</strong></td>
<td>• v-structures handled elegantly</td>
<td>• cycles allowed</td>
</tr>
<tr>
<td></td>
<td>• CPDs are conditional probabilities</td>
<td>• perfect maps for misunderstanding students</td>
</tr>
<tr>
<td></td>
<td>• probability of full instantiation is easy (no partition function)</td>
<td></td>
</tr>
</tbody>
</table>
Markov Networks So Far

• Markov network: undirected graph, potentials over cliques (or sub-cliques), normalization via partition function

• Representation theorems

• Independence: active paths/separation; pairwise; Markov blanket

• Minimal I-maps are unique

• P-maps don’t always exist
HW#2

• Due Feb 17 and Feb 22