Undirected Graphical Models II

Readings: K&F 4.4, 4.5, 4.6

Lecture 5 – Apr 11, 2011
CSE 515, Statistical Methods, Spring 2011

Instructor: Su-In Lee
University of Washington, Seattle

Last time

- Markov networks representation
  - Local factor models (potentials) \( \pi_1[D_1], \ldots, \pi_n[D_n] \)
  - Independence properties
    - Global, pairwise, local independencies
- I-Map \( \leftrightarrow \) Factorization
  \[ P(X_1, \ldots, X_n) = \frac{1}{Z} \prod \pi_i[D_i] \]

Today...

- Parameterization revisited
- Bayesian nets and Markov nets
- Partially directed graphs
- Inference 101
Factor Graphs

- From the Markov network structure, we do not know how it is parameterized.
  - Example: fully connected graph may have pairwise potentials or one large (exponential) potential over all nodes.

\[
P(A, B, C) = \frac{1}{Z} \prod \pi_i[D_i]
\]

\[
P_{\text{mc}}(A, B, C) = \frac{1}{Z_{\text{mc}}} \pi_{\text{abc}}[A, B, C]
\]

\[
P_{\text{pair}}(A, B, C) = \frac{1}{Z_{\text{pair}}} \pi_{\text{ab}}[A, B] \pi_{\text{bc}}[B, C] \pi_{\text{ac}}[A, C]
\]

Solution: Factor Graphs
- Undirected graph
- Two types of nodes: Variable nodes, Factor nodes
- Connectivity?

Factor Graphs: example

- Two types of nodes: Variable nodes, Factor nodes
- Connectivity
  - Each factor node is associated with exactly one factor \( \pi_i[D_i] \)
  - Scope of factor are all neighbor variables of the factor node

\[
P_{\text{mc}}(A, B, C) = \frac{1}{Z_{\text{mc}}} \pi_{\text{abc}}[A, B, C]
\]

\[
P_{\text{pair}}(A, B, C) = \frac{1}{Z_{\text{pair}}} \pi_{\text{ab}}[A, B] \pi_{\text{bc}}[B, C] \pi_{\text{ac}}[A, C]
\]

Markov network

Factor graph for joint parameterization

Factor graph for pairwise parameterization
Local Structure: Feature Representation

- Factor graphs still encode complete tables

- A feature $\phi[D]$ on variables $D$ is an indicator function that for some $d \in D$; for example,
  \[
  \phi[X,Y] = \begin{cases} 
  1 & \text{when } x = y \\
  0 & \text{otherwise}
  \end{cases}
  \]

- Several features can be defined on one clique
  \[
  \phi_1[D] = \begin{cases} 
  1 & \text{when } x = y \\
  0 & \text{otherwise}
  \end{cases} \quad \phi_2[D] = \begin{cases} 
  1 & \text{when } x > 50 \\
  0 & \text{otherwise}
  \end{cases}
  \]
  → Any factor can be represented by features, where in general case, we define a feature and weight for each entry in the factor

- Apply log-transformation: $\pi[D] = \exp(-w_1 \phi_1[D])$

Log-linear model

- A distribution $P$ is a log-linear model over $H$ if it has
  - Features $\phi_1[D_1],...,\phi_k[D_k]$ where each $D_i$ is a complete subgraph in $H$
  - A set of weights $w_1,...,w_k$ such that

\[
P(X_1,...,X_n) = \frac{1}{Z} \prod_{i=1}^k \pi_i[D_i] = \frac{1}{Z} \exp\left[\sum_{i=1}^k w_i \phi_i[D_i]\right]
\]

- Advantages
  - Log-linear model is more compact for many distributions especially with large domain variables
  - Representation is intuitive and modular – Features can be modularly added between any interacting sets of variables
Markov Network Parameterizations

- **Choice 1: Markov network**
  - Product over potentials
  - Right representation for discussing independence queries

- **Choice 2: Factor graph**
  - Product over potentials
  - Useful for inference (later)

- **Choice 3: Log-linear model**
  - Product over feature weights
  - Useful for discussing parameterizations
  - Useful for representing context specific structures

- **All parameterizations are interchangeable**

Outline

- **Markov networks representation**
  - Local factor models: $\prod_{i=1}^n \pi_i(D_i)$
  - Independencies
    - global, pairwise, local independencies
  - I-Map $\leftrightarrow$ Factorization: $P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^n \pi_i(D_i)$

- **Today...**
  - Parameterization revisited
  - Bayesian nets and Markov nets
  - Partially directed graphs
  - Inference 101
From Bayesian nets to Markov nets

- **Goal:** build a Markov network $H$ capable of representing any distribution $P$ that factorizes over $G$
  - Equivalent to requiring $I(H) \subseteq I(G)$

- **Construction process**
  - Based on local Markov independencies
    - If $X$ is connected with $Y$ in $H$, $(X \perp U\setminus\{X\}-Y|Y)$.
  - Connect each $X$ to every node in the smallest set $Y$ s.t.:
    $$\{(X \perp U\setminus\{X\}-Y|Y) : X \in H\} \subseteq I(G)$$
  - How can we find $Y$ by querying $G$?
    - $Y = \text{Markov blanket of } X \text{ in } G$?

---

**Blocking Paths**

- **Active path:**
  - parents
  - descendants
  - v-structure

- **Block path:**
  - parents $\notin Y$
  - children $\in Y$
  - children’s parents
From Bayesian nets to Markov nets

- **Goal:** build a Markov network $H$ capable of representing any distribution $P$ that factorizes over $G$
  - Equivalent to requiring $I(H) \subseteq I(G)$

- **Construction process**
  - Based on local Markov independencies
    - If $X$ is connected with $Y$ in $H$, $(X \perp U\{X\}-Y|Y)$.
    - Connect each $X$ to every node in the smallest set $Y$ s.t.:
      $$\{ (X \perp U\{X\}-Y|Y) : X \in H \} \subseteq I(G)$$
  - How can we find $Y$ by querying $G$?
    - $Y = \text{Markov blanket of } X \text{ in } G \left( \text{parents, children, children's parents} \right)$

Moralized Graphs

- The **Moral graph** of a Bayesian network structure $G$ is the undirected graph that contains an undirected edge between $X$ and $Y$ if
  - $X$ and $Y$ are directly connected in $G$
  - $X$ and $Y$ have a common child in $G$
Parameterizing Moralized Graphs

- Moralized graph contains a full clique for every $X_i$ and its parents $\text{Pa}(X_i)$.
  - We can associate CPDs with a clique

- Do we lose independence assumptions implied by the graph structure?
  - Yes, immoral v-structures

From Markov nets to Bayesian nets

- Transformation is more difficult and the resulting network can be much larger than the Markov network

- Construction algorithm
  - Use Markov network as template for independencies $I(H)$
  - Fix ordering of nodes
  - Add each node along with its minimal parent set according to the independencies defined in the distribution
Chordal Graphs

- Let $X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_k \rightarrow X_1$ be a loop in the graph.
- A chord in the loop is an edge connecting $X_i$ and $X_j$ for two nonconsecutive nodes $X_i$ and $X_j$.

- An undirected graph is chordal if any loop $X_1 \rightarrow X_2 \rightarrow \ldots \rightarrow X_k \rightarrow X_1$ for $k \geq 4$ has a chord.
  - That is, longest minimal loop is a triangle.
  - Chordal graphs are often called triangulated.

- A directed graph is chordal if its underlying undirected graph is chordal.
From Markov Nets to Bayesian Nets

- Theorem: Let H be a Markov network structure and G be any minimal I-map for H. Then G is chordal.

- The process of turning a Markov network into a Bayesian network is called triangulation. The process loses independencies.

Last time

- Markov networks representation
  - Local factor models \( \pi_i[D_i], \ldots, \pi_i[D_n] \)
  - Independencies
    - global, pairwise, local independencies
  - I-Map \( \leftrightarrow \) Factorization
    \[ P(X_1, \ldots, X_n) = \frac{1}{Z} \prod \pi_i[D_i] \]

Today...

- Parameterization revisited
- Bayesian nets and Markov nets
- Partially directed graphs
- Inference 101
Conditional Random Fields (CRFs)

- Special case of partially directed models
- A conditional random field is an undirected graph $H$ whose nodes correspond to $X \cup Y$; the network is annotated with a set of factors $\phi_1(D_1), \ldots, \phi_m(D_m)$ such that each $D_i \in X$. The network encodes a conditional distribution as follows:

$$
\tilde{P}(Y, X) = \prod_{i=1}^{m} \phi_i(D_i)
$$

$$
P(Y | X) = \frac{1}{Z(X)} \tilde{P}(Y, X)
$$

$$
Z(X) = \sum_Y \tilde{P}(Y, X)
$$

- Two variables in $H$ are connected by an undirected edge whenever they appear together in the scope of some factor $\phi$.

Why Conditional?

- Why $P(Y | X)$, not $P(Y, X)$?
  - The network explicitly does not encode any distribution over the variables in $X$.
  - One of the main strengths of the CRF representation.

- This flexibility allows us to do many things:
  - Incorporating into the model a rich set of observed variables $X$ whose dependencies may be quite complex or even poorly understood.
  - Including continuous variables $X$ whose distribution may not have a simple parametric form.
  - Using domain knowledge in order to define a rich set of features characterizing our domain, without worrying about modeling their joint distribution.

- Many applications: Computer vision (detail later), text analysis, part-of-speech labeling, many more.
Conditional Random Fields
- Directed and undirected dependencies.
- A CRF defines conditional distribution of \( Y \) on \( X \), \( P(Y|X) \).
  - It can be viewed as a partially directed graph, where we have an undirected component over \( Y \), which has the variables in \( X \) as parents.
- Any difference with Bayesian networks?

Chain Networks
- Combines Markov networks and Bayesian networks
- Partially directed graph (PDAG)
- As for undirected graphs, we have three distinct interpretations for the independence assumptions implied by a P-DAG

Example:
Pairwise Independencies

- Every node $X$ is independent from any node which is not its descendant given all non-descendants of $X$

- Formally:
  \[ I_P(K) = \{(X \perp Y | ND(X)-\{X,Y\}) : X \not\rightarrow Y \notin K, Y \in ND(X)\} \]

Example:
- \((D \perp A | B,C,E)\)
- \((C \perp E | A)\)

Local Markov Independencies

- Let $\text{Boundary}(X)$ be the union of the parents of $X$ and the neighbors of $X$

- Local Markov independencies state that a node $X$ is independent of its non-descendants given its boundary

- Formally:
  \[ I_L(K) = \{(X \perp ND(X) \setminus \text{Boundary}(X) | \text{Boundary}(X)) : X \in \mathcal{U}\} \]

Example:
- \((D \perp A,E | B,C)\)
Global Independencies

- \( I(K) = \{(X, Y | Z) : X, Y, Z, X \text{ is } c\text{-separated from } Y \text{ given } Z\} \)

- \( X \) is \( c\)-separated from \( Y \) given \( Z \) if \( X \) is separated from \( Y \) given \( Z \) in the undirected \( \text{moralized graph } M[K] \)

- The \text{moralized graph} of a P-DAG \( K \) is an undirected \( M[K] \) by
  - Connecting any pair of parents of a given node
  - Converting all directed edges to undirected edges

  For positive distributions: \( I(K) \leftrightarrow I_L(K) \leftrightarrow I_P(K) \)

Domain Application: Vision

- The \text{image segmentation} problem
  - Task: Partition an image into distinct parts of the scene
  - Example: separate water, sky, background
Markov Network for Segmentation

- Grid structured Markov network (CRF)
- Random variables \((X_i, Y_i)\) correspond to pixel \(i\)
  - \(X_i\): Input image for pixel \(i\) (always given)
  - Color, texture, location ...
  - \(Y_i\): Domain is \(\{1, \ldots, K\}\) e.g. \(1: \) road, \(2: \) car, \(3: \) bldg) (generally not given)
  - Value represents region assignment to pixel \(i\)
- Neighboring pixels are connected in the network

Node potentials (appearance distribution)
- Introduce node potential \(\exp(-w_k Y_i = \{1\{Y_i = k\})\)
- \(w_k\) – extent to which pixel \(i\) “fits” region \(k\) (e.g., based on \(X_i\) containing various info such as color, location, texture on pixel \(i\))

Edge potentials (contiguity preference)
- Encodes contiguity preference by edge potential \(\exp(\lambda Y_i = Y_j)\) for \(\lambda > 0\)
Markov Network for Segmentation

Solution: inference on the pairwise Markov network

- Find most likely assignment $k$ (=sky, building, etc) to $Y_i$ variables

$$\pi_i [Y_i, X_i] = \exp(-w_i k \mathbf{1} \{Y_i = k\})$$

$$\pi_{i,j} [Y_i, Y_j] = \exp(\lambda_{i,j} \mathbf{1} \{Y_i = Y_j\})$$

$Y=1$: sky, $Y=2$: car, $Y=3$: building

Example Results

Baseline (a simple classifier): Result of segmentation using node potentials alone, so that each pixel is classified independently

Result of segmentation using a pairwise Markov network encoding interactions between adjacent pixels
Last time

- Markov networks representation
  - Local factor models \( \pi_i(D_i), \ldots, \pi_n(D_n) \)
  - Independencies
    - global, pairwise, local independencies
  - I-Map \( \leftrightarrow \) Factorization
    \[ P(X_1, \ldots, X_n) = \frac{1}{Z} \prod \pi_i(D_i) \]

Today...

- Parameterization revisited
- Bayesian nets and Markov nets
- Partially directed graphs
- Inference 101

Inference

- Markov networks and Bayesian networks represent a joint probability distribution

↓

- Networks contain information needed to answer any query about the distribution
- **Inference** is the process of answering such queries

- Direction between variables does not restrict queries
- Inference combines evidence from all network parts
Likelihood Queries

- Compute probability (=likelihood) of the evidence
  - Evidence: subset of variables $E$ and an assignment $e$
  - Task: compute $P(E=e)$

- Computation

$$P(E = e) = \sum_{Z \in U - E} P(Z = z, E = e)$$

Conditional Probability Queries

- Conditional probability queries
  - Evidence: subset of variables $E$ and an assignment $e$
  - Query: a subset of variables $Y$
  - Task: compute $P(Y \mid E=e)$

- Applications
  - Medical and fault diagnosis
  - Genetic inheritance

- Computation

$$P(Y = y \mid E = e) = \frac{P(Y = y, E = e)}{P(E = e)} = \frac{\sum_{W, Z} P(W = w, Y = y, E = e)}{\sum_{Z} P(Z = z, E = e)}$$
Maximum A Posteriori Assignment

- **Maximum A Posteriori Assignment (MAP)**
  - **Evidence:** subset of variables \( E \) and an assignment \( e \)
  - **Query:** a subset of variables \( Y \)
  - **Task:** compute \( \text{MAP}(Y \mid E = e) = \arg\max_y P(Y = y \mid E = e) \)
  - **Note 1:** there may be more than one possible solution
  - **Note 2:** equivalent to computing \( \arg\max_y P(Y = y, E = e) / P(E = e) \)

- **Computation**
  \[
  \text{MAP}(Y = y \mid e) = \arg\max_y \sum_{w \in U - Y - E} P(W = w, Y = y' \mid E = e)
  \]

Most Probable Assignment: MPE

- **Most Probable Explanation (MPE)**
  - **Evidence:** subset of variables \( E \) and an assignment \( e \)
  - **Query:** all other variables \( Y \ (Y = U - E) \)
  - **Task:** compute \( \text{MPE}(Y \mid E = e) = \arg\max_y P(Y = y \mid E = e) \)
  - **Note:** there may be more than one possible solution

- **Applications**
  - Decoding messages: find the most likely transmitted bits
  - Diagnosis: find a single most likely consistent hypothesis
Most Probable Assignment: MPE

- Note: We are searching for the most likely joint assignment to all variables
  - May be different than most likely assignment (MAP) of each variable.
  - Any example?
  - Given $E = \phi$
    - $P(a^1) > P(a^0) \rightarrow \text{MAP}(A) = a^1$
    - $\text{MPE}(A, B) = \{a^0, b^1\}$
      - $P(a^0, b^0) = 0.04$
      - $P(a^0, b^1) = 0.36$
      - $P(a^1, b^0) = 0.3$
      - $P(a^1, b^1) = 0.3$

Exact Inference in Graphical Models

- Graphical models can be used to answer
  - Conditional probability queries
  - MAP queries
  - MPE queries

- Naïve approach
  - Generate joint distribution
  - Depending on query, compute sum/max
    - Exponential blowup

- Exploit independencies for efficient inference
Summary: Markov network representation

- Markov Networks – undirected graphical models
  - Like Bayesian networks, define independence assumptions
  - Three definitions exist, all equivalent in positive distributions
  - Factorization is defined as product of factors over complete sub-graph

- Alternative parameterizations
  - Factor graphs
  - Log-linear models

- Relationship to Bayesian networks
  - Represent different families of independencies
    - Triangulation – transforming Markov networks to Bayesian networks.

- Partially directed graphs
  - Conditional random fields (CRFs)
  - Application to image segmentation

Announcements

- Feedback on the course
  - Email your comments anonymously.
  - See the course website.

- Additional OH
  - Tuesday in the morning 9-10am

- Slightly modified course outline
Where are we? What next?

<table>
<thead>
<tr>
<th>Week</th>
<th>Dates</th>
<th>Topics and Lecture Notes</th>
<th>Readings</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>2/18</td>
<td>Probabilistic Graphical Models Representation</td>
<td>2.1, 2.2, 3.3</td>
</tr>
<tr>
<td>I</td>
<td>2/19</td>
<td>Bayesian model representation</td>
<td>3.1, 3.2, 3.3</td>
</tr>
<tr>
<td>I</td>
<td>2/24</td>
<td>Local probability models</td>
<td>3.4, 3.5</td>
</tr>
<tr>
<td>I</td>
<td>3/1</td>
<td>Undirected graphical model 1</td>
<td>4.1, 4.2, 4.3</td>
</tr>
<tr>
<td>I</td>
<td>3/8</td>
<td>Undirected graphical model 2 + PGDAG</td>
<td>4.4, 4.5, 4.6</td>
</tr>
<tr>
<td>II</td>
<td>3/15</td>
<td>Inference: exact inference</td>
<td>5.1, 6.2, 5.3</td>
</tr>
<tr>
<td>II</td>
<td>3/16</td>
<td>Exact inference in BNs</td>
<td>9.4, 9.5, 9.6</td>
</tr>
<tr>
<td>II</td>
<td>3/20</td>
<td>Exact inference: Clique Trees</td>
<td>10.1, 10.2, 10.3, 10.4</td>
</tr>
<tr>
<td>III</td>
<td>4/25</td>
<td>Learning: parameter estimation</td>
<td>17</td>
</tr>
<tr>
<td>III</td>
<td>4/27</td>
<td>Parameter learning in BNs</td>
<td>17</td>
</tr>
<tr>
<td>IV</td>
<td>5/2</td>
<td>Structure learning in BNs</td>
<td>19</td>
</tr>
<tr>
<td>IV</td>
<td>5/4</td>
<td>Familiy observed data learning with missing data</td>
<td>19</td>
</tr>
<tr>
<td>IV</td>
<td>5/9</td>
<td>More on learning (TED)</td>
<td>19</td>
</tr>
<tr>
<td>V</td>
<td>5/11</td>
<td>Approximate Inference: particle-based I</td>
<td>12</td>
</tr>
<tr>
<td>V</td>
<td>5/16</td>
<td>Approximate Inference: particle-based II</td>
<td>12</td>
</tr>
<tr>
<td>V</td>
<td>5/18</td>
<td>Global approximate inference</td>
<td>11</td>
</tr>
<tr>
<td>V</td>
<td>5/25</td>
<td>Global approximate inference II</td>
<td>11</td>
</tr>
<tr>
<td>V</td>
<td>5/30</td>
<td>Special Topics &amp; Applications</td>
<td>20</td>
</tr>
<tr>
<td>V</td>
<td>5/30</td>
<td>Decision Process (Instructor Message)</td>
<td>20</td>
</tr>
<tr>
<td>V</td>
<td>6/1</td>
<td>Temporal models (BNs, HMMs)</td>
<td>20</td>
</tr>
</tbody>
</table>

Acknowledgement

- These lecture notes were generated based on the slides from Prof Eran Segal.