Markov networks, Factor graphs, and an unified view

Start approximate inference
If we are lucky...

Graphical Models – 10708
Carlos Guestrin
Carnegie Mellon University
October 27th, 2006

Factorization in Markov networks

- Given an undirected graph $H$ over variables $X = \{X_1, \ldots, X_n\}$
- A distribution $P$ factorizes over $H$ if $\exists$
  - subsets of variables $D_1 \subseteq X, \ldots, D_m \subseteq X$, such that the $D_i$ are fully connected in $H$
  - non-negative potentials (or factors) $\pi_1(D_1), \ldots, \pi_m(D_m)$
    - also known as clique potentials
  - such that
    $$P(X) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(D_i)$$
- Also called Markov random field $H$, or Gibbs distribution over $H$
Global Markov assumption in Markov networks

- A path $X_1 - \ldots - X_k$ is **active** when set of variables $\mathbf{Z}$ are observed if none of $X_i \in \{X_1, \ldots, X_k\}$ are observed (are part of $\mathbf{Z}$).
- Variables $\mathbf{X}$ are **separated** from $\mathbf{Y}$ given $\mathbf{Z}$ in graph $H$, $\text{sep}_H(\mathbf{X}, \mathbf{Y}|\mathbf{Z})$, if there is no active path between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ given $\mathbf{Z}$.
- The **global Markov assumption** for a Markov network $H$ is:

$$\mathcal{I}(H) = \left\{ \mathbf{Z} \mid \text{sep}_H(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) \right\}$$

Representation Theorem for Markov Networks

- **If joint probability distribution $\mathbf{P}$:**
  
  $$\mathbf{P}(X_1, \ldots, X_n) = \frac{1}{\mathbf{Z}} \prod_{i=1}^{m} \pi_i(\mathbf{D}_i)$$

  Then $\mathbf{H}$ is an I-map for $\mathbf{P}$.

- **OK** → if doing parameter learning and get a zero in a potential.

- **If $\mathbf{H}$ is an I-map for $\mathbf{P}$ and $\mathbf{P}$ is a positive distribution:**

  Then joint probability distribution $\mathbf{P}$:

  $$\mathbf{P}(X_1, \ldots, X_n) = \frac{1}{\mathbf{Z}} \prod_{i=1}^{m} \pi_i(\mathbf{D}_i)$$

  Can get in trouble → if structural learning & $\mathbf{P}$ not positive.
Local independence assumptions for a Markov network

- **Separation** defines global independencies $\mathcal{I}(H)$.

- **Pairwise Markov Independence**: $I_{pw}(H)$
  - Pairs of non-adjacent variables are independent given all others:
  $$A \perp B \mid X - \{A, B\}$$

- **Markov Blanket**: $I_{nb}(H)$
  - Variable independent of rest given its neighbors $N(A)$:
  $$A \perp X - \{N(A), A\} \mid N(A)$$

Equivalence of independencies in Markov networks

- **Soundness Theorem**: For all positive distributions $P$, the following three statements are equivalent:
  - $P$ entails the global Markov assumptions
    $$P \models \mathcal{I}(H)$$
  - $P$ entails the pairwise Markov assumptions
    $$P \models I_{pw}(H)$$
  - $P$ entails the local Markov assumptions (Markov blanket)
    $$P \models I_{nb}(H)$$
Minimal I-maps and Markov Networks

A fully connected graph is an I-map
Remember minimal I-maps?
- A "simplest" I-map → Deleting an edge makes it no longer an I-map

In a BN, there is no unique minimal I-map
Theorem: In a Markov network, minimal I-map is unique!!
Many ways to find minimal I-map, e.g.,
- Take pairwise Markov assumption:
- If \(P\) doesn’t entail it, add edge:

How about a perfect map?

Remember perfect maps?
- independencies in the graph are exactly the same as those in \(P\)
For BNs, doesn’t always exist
- counter example: Swinging Couples
How about for Markov networks?
Unifying properties of BNs and MNs

- **BNs:**
  - give you: V-structures, CPTs are conditional probabilities, can directly compute probability of full instantiation
  - but: require acyclicity, and thus no perfect map for swinging couples

- **MNs:**
  - give you: cycles, and perfect maps for swinging couples
  - but: don’t have V-structures, cannot interpret potentials as probabilities, requires partition function

- **Remember PDAGS???
  - skeleton + immoralities
  - provides a (somewhat) unified representation
  - see book for details

What you need to know so far about Markov networks

- **Markov network representation:**
  - undirected graph
  - potentials over cliques (or sub-cliques)
  - normalize to obtain probabilities
  - need partition function

- **Representation Theorem for Markov networks**
  - if P factorizes, then it’s an I-map
  - if P is an I-map, only factorizes for positive distributions

- **Independence in Markov nets:**
  - active paths and separation
  - pairwise Markov and Markov blanket assumptions
  - equivalence for positive distributions

- **Minimal I-maps in MNs are unique**
- **Perfect maps don’t always exist**
Some common Markov networks and generalizations

- Pairwise Markov networks
- A very simple application in computer vision
- Logarithmic representation
- Log-linear models
- Factor graphs

Pairwise Markov Networks

- All factors are over single variables or pairs of variables:
  - Node potentials $\prod_i (\tau_i)$
  - Edge potentials $\prod_{ij} (\tau_{ij})$

- Factorization:

$\mathcal{P}(x) = \frac{1}{Z} \prod_i \tau_i(x_i) \prod_{i,j \in E} \tau_{ij}(x_i, x_j)$

- Note that there may be bigger cliques in the graph, but only consider pairwise potentials
A very simple vision application

- Image segmentation: separate foreground from background
- Graph structure:
  - pairwise Markov net
  - grid with one node per pixel
- Node potential:
  - "background color" vs. "foreground color"
  \[ \Pi_i(x_i, y_i) = \begin{cases} \text{light} & \text{brown} \\ \text{dark} & \text{brown} \end{cases} \]
- Edge potential:
  - neighbors like to be of the same class
  \[ \lambda \text{ low } \Rightarrow \text{no smoothness (focus node pot.)} \]
  \[ \lambda \text{ high } \Rightarrow \text{smooth (ignore or less focus pot.)} \]

Logarithmic representation

- Standard model:
  \[ P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{i=1}^{m} \pi_i(D_i) \]
- Log representation of potential (assuming positive potential):
  - also called the energy function
  \[ \psi_i(D_i) = -\ln \Pi_i(D_i) \]
  \[ P(X) = \frac{1}{Z} \prod_{i=1}^{m} \Pi_i(D_i) = \frac{1}{Z} \exp \left\{ -\sum_{i=1}^{m} \psi_i(D_i) \right\} \]
- Log representation of Markov net:
  \[ \frac{1}{2} \exp \left\{ -\sum_{i=1}^{m} \psi_i(D_i) \right\} \]
Log-linear Markov network
(most common representation)

- **Feature** is some function $\phi[D]$ for some subset of variables $D$
  - e.g., indicator function $\phi[D] = \begin{cases} 1 & \text{if all of } D \text{ is true} \\ 0 & \text{otherwise} \end{cases}$

- **Log-linear model** over a Markov network $H$:
  - a set of features $\phi_1[D_1], \ldots, \phi_k[D_k]$
    - each $D_i$ is a subset of a clique in $H$
    - two $\phi$'s can be over the same variables
  - a set of weights $w_1, \ldots, w_k$
    - usually learned from data

- $P(x_1, \ldots, x_n) = \frac{1}{Z} \exp \left[ \sum_{i=1}^k w_i \phi_i(D_i) \right]$
  - sometimes defined as $\exp \left[ - \sum_{i=1}^k w_i \phi_i(D_i) \right]$

Structure in cliques

- Possible potentials for this graph:
  - $\text{A} \rightarrow \text{B} \rightarrow \text{C}$
  - How many branches
    - $K^{3-1} = K^2$
    - $3(K^2-1)$
Factor graphs

- Very useful for approximate inference
  - Make factor dependency explicit
- Bipartite graph:
  - variable nodes (ovals) for $X_1,\ldots,X_n$
  - factor nodes (squares) for $\phi_1,\ldots,\phi_m$
  - edge $X_i \rightarrow \phi_j$ if $X_i \in \text{Scope}[\phi_j]$
- More explicit representation, but exactly equivalent

Exact inference in MNs and Factor Graphs

- Variable elimination algorithm presented in terms of factors → exactly the same VE algorithm can be applied to MNs & Factor Graphs
- Junction tree algorithms also applied directly here:
  - triangulate MN graph as we did with moralized graph for BN
  - each factor belongs to a clique
  - same message passing algorithms
Summary of types of Markov nets

- Pairwise Markov networks
  - very common
  - potentials over nodes and edges

- Log-linear models
  - log representation of potentials
  - linear coefficients learned from data
  - most common for learning MNs

- Factor graphs
  - explicit representation of factors
    - you know exactly what factors you have
  - very useful for approximate inference

What you learned about so far

- Bayes nets
- Junction trees
- (General) Markov networks
- Pairwise Markov networks
- Factor graphs

How do we transform between them?

More formally:
- I give you an graph in one representation, find an l-map in the other
From Bayes nets to Markov nets

BNs $\rightarrow$ MNs: Moralization

- **Theorem**: Given a BN $G$ the Markov net $H$ formed by moralizing $G$ is the *minimal* $I$-map for $I(G)$

- **Intuition**:
  - in a Markov net, each factor must correspond to a subset of a clique
  - the factors in BNs are the CPTs $P(s|I,s)$
  - CPTs are factors over a node and its parents
  - thus node and its parents must form a clique

- **Effect**:
  - some independencies that could be read from the BN graph become hidden
From Markov nets to Bayes nets

MNs → BNs: Triangulation

**Theorem:** Given a MN $H$, let $G$ be the Bayes net that is a minimal $I$-map for $I(H)$ then $G$ must be **chordal**

**Intuition:**
- v-structures in BN introduce immoralities
- these immoralities were not present in a Markov net
- the triangulation eliminates immoralities

**Effect:**
- **many** independencies that could be read from the MN graph become hidden

Intuition:

- $v$-structures in BN introduce immoralities
- these immoralities were not present in a Markov net
- the triangulation eliminates immoralities

Effect:

- **many** independencies that could be read from the MN graph become hidden
Markov nets v. Pairwise MNs

- Every Markov network can be transformed into a Pairwise Markov net
  - introduce extra "variable" for each factor over three or more variables
  - domain size of extra variable is exponential in number of vars in factor

- Effect:
  - any local structure in factor is lost
  - a chordal MN doesn’t look chordal anymore

Overview of types of graphical models and transformations between them