Probabilistic Graphical Models

Raquel Urtasun and Tamir Hazan

TTI Chicago

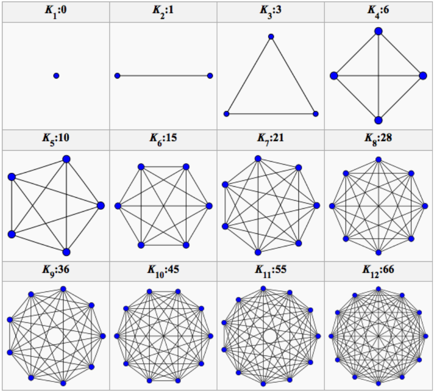
April 11, 2011

Undirected Chordal Graphs

- When can I represent perfectly a set of independencies by both a BN and a Markov network?
- This class is the class of **undirected chordal graphs**
- Let $\mathcal{H}$ be a non-chordal Markov network. Then there is no BN $\mathcal{G}$ which is a perfect map for $\mathcal{H}$, i.e., $\mathcal{I}(\mathcal{H}) = \mathcal{I}(\mathcal{G})$.
- Proof: in the book, chapter 4.5.3.
- Chordal graphs and their properties play a central roll in the derivation of exact inference algorithms.
- We restrict ourselves to connected graphs; trivial extension to disconnected.

Reminder on complete graphs

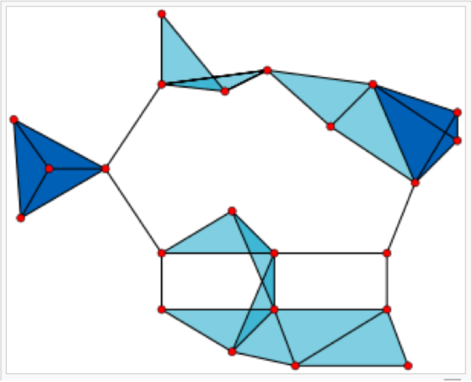
- The complete graph K_n of order n is a simple graph with n vertices in which every vertex is adjacent to every other vertex.
- The complete graph on n has $n(n-1)/2$ edges



Reminder on cliques

- A **clique** in an undirected graph $G = (V, E)$ is a subset of the vertex set $C \subseteq V$, such that for every two vertices in C, there exists an edge connecting the two.
- This is equivalent to saying that the subgraph induced by C is complete.
- A **maximal clique** is a clique which does not exist exclusively within the vertex set of a larger clique.
- A **maximum clique** is a clique of the largest possible size in a given graph.
- The **clique number** of a graph G is the number of vertices in a maximum clique in G.

Example of cliques



- What's the clique number?
- What's the maximum clique? and the maximal cliques?

Tree of cliques I

- We can decompose any connected chordal graph $\mathcal{H}$ into a tree of cliques.
- A **tree of cliques** $\mathcal{T}$ is one whose nodes are the maximal cliques in $\mathcal{H}$, $\mathbf{C}_1, \cdots, \mathbf{C}_k$.
- The structure of the clique precisely encodes the independencies in $\mathcal{H}$.
- For disconnected graph, we have a set of forests , one per component.
- Let $\mathbf{C}_i, \mathbf{C}_j$ be two cliques in the tree that are directly connected by an edge.
- We define $\mathbf{S}_{i,j} = \mathbf{C}_i \cap \mathbf{C}_j$ a **sepset** between $\mathbf{C}_i$ and $\mathbf{C}_j$.
- Let $\mathbf{W}_{<(i,j)}$ ($\mathbf{W}_{<(j,i)}$) be all of the variables that appear on the $\mathbf{C}_i$ ($\mathbf{C}_j$) side of the edge.
- We can decompose $\mathcal{X}$ into disjoint sets $\mathbf{W}_{<(i,j)}, \mathbf{W}_{<(j,i)}, \mathbf{S}_{i,j}$

Tree of cliques II

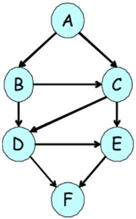
We can now say that a tree $\mathcal{T}$ is a clique tree for $\mathcal{H}$ if

- Each node corresponds to a clique in $\mathcal{H}$, and each maximal clique $\mathbf{C}_1, \cdots, \mathbf{C}_k$ in $\mathcal{H}$ is a node in $\mathcal{T}$
- Each sepset $\mathbf{S}_{i,j}$ separates $\mathbf{W}_{<(i,j)}$ and $\mathbf{W}_{<(j,i)}$

Note that this implies that each separator $\mathbf{S}_{i,j}$ renders its two sides conditionally independent in $\mathcal{H}$.

Example

- Let's revisit the previous example.
- The undirected graph $\mathcal{H}$ is chordal and has the same number of edges as the BN which is also chordal

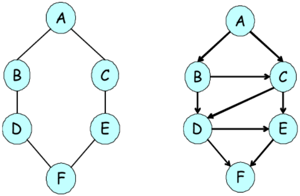


- The clique tree is simply a chain $\{A, B, C\} \rightarrow \{B, C, D\} \rightarrow \{C, D, E\} \rightarrow \{D, E, F\}$.
- What's $\mathbf{S}_{1,2}$? and $\mathbf{W}_{<(1,2)}$? and $\mathbf{W}_{<(2,3)}$?

A bit more on chordal graphs

- Every undirected chordal graph $\mathcal{H}$ has a clique tree $\mathcal{T}$.
- Proof: Book, page 132
- This means that the independencies in an undirected graph $\mathcal{H}$ can be captured perfectly in a BN if and only if $\mathcal{H}$ is chordal.
- Let $\mathcal{H}$ be a chordal Markov network. Then there is a BN $\mathcal{G}$ such that $\mathcal{I}(\mathcal{H}) = \mathcal{I}(\mathcal{G})$.
- Proof: Book, page 133.

An example



(a) (b)

- The graph in (b) and it's moralized network encode precisely the same independencies
- There is no BN that encodes precisely the independencies in the non-chordal graph (a)

Conclusion on chordal graphs

- Thus chordal graphs are the intersection between Markov networks and BN.
- The independencies in the graph can be represented exactly in both types of models if and only if the graph is chordal.

Partially Directed Models

So far we have talked about 2 distinct types of graphical models

- Bayesian networks
- Undirected graphical models or Markov networks

Both representations allow us to incorporate directed and undirected dependencies.

We can unify both representations by allowing models that represent both types of dependencies, e.g., Conditional Random Fields.

Conditional Random Fields

- Markov networks encode a joint distribution over $\mathcal{X}$.
- The same undirected graph can be used to describe conditional distributions $P(\mathbf{Y}|\mathbf{X})$.
- $\mathbf{Y}$ is a set of **target variables**.
- $\mathbf{X}$ is a set of **observable variables**.
- Both $\mathbf{X}$ and $\mathbf{Y}$ are disjoint.
- This representation is usually called a **Conditional Random Field (CRF)**

More on CRFs

- CRF is an undirected graph whose nodes correspond to $\mathbf{Y} \cup \mathbf{X}$.
- The graph is parameterized with a set of factors $\phi_1(\mathbf{D}_1), \cdots, \phi_m(\mathbf{D}_m)$.
- The graph can also be encoded with features $f(\mathbf{D}_i)$.
- The difference: instead of encoding $P(\mathbf{Y}, \mathbf{X})$ it encodes $P(\mathbf{Y}|\mathbf{X})$.
- This is important, as it avoid representing $P(\mathbf{X})$.
- This is done by disallowing potentials that encode only variables on $\mathbf{X}$.

Formal definition

- A CRF is an undirected graph $\mathcal{H}$ whose nodes correspond to $\mathbf{X} \cap \mathbf{Y}$; and the network is annotated with factors $\phi(\mathbf{D}_1), \cdots, \phi_m(\mathbf{D}_m)$ such that $\mathbf{D}_i \not\subseteq \mathbf{X}$. The network encodes the conditional distribution

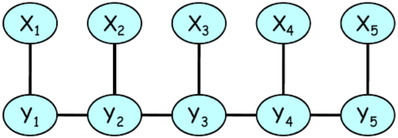
$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z(\mathbf{X})}\hat{P}(\mathbf{X},\mathbf{Y}) = \frac{1}{Z(\mathbf{X})}\prod_{i=1}^{m}\phi_i(\mathbf{D}_i)$$

  with partition function

$$Z(\mathbf{X}) = \sum_{Y}\hat{P}(\mathbf{X},\mathbf{Y}) = \sum_{Y}\prod_{i=1}^{m}\phi_i(\mathbf{D}_i)$$

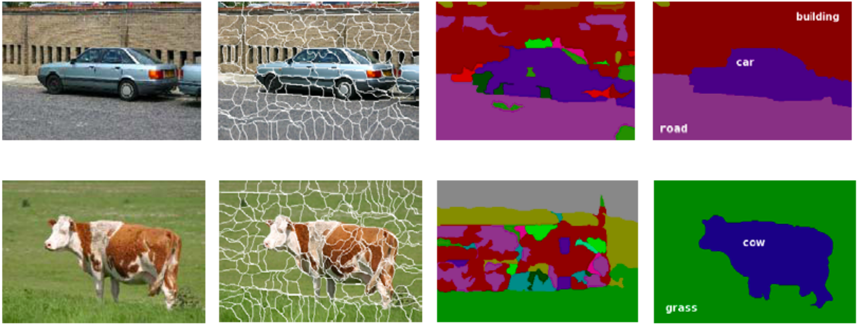
- Two variables in $\mathcal{H}$ are connected with an undirected edge if they appear together in the scope of some factor.
- The only difference with a standard Markov network is the partition function.
- Before marginalized over $\mathbf{X}$ and $\mathbf{Y}$, now not.

Simple Example of CRF



- What's the probability distribution in this case?
- And the partition function?

Segmentation example



- If we represent segmentation with a CRF, what's the prob. distribution?
- And the partition function?

More on CRFs

- Unlike the definition of a conditional Bayesian network, the structure of the CRF might contain edges between the $\mathbf{X}$ variables.
- This happens when two such variables appear together in a factor that also contains a target variable.
- However, the edges between $\mathbf{X}_i$ do not encode anything about the distribution $p(\mathbf{X})$.
- Not encoding $p(\mathbf{X})$ is the main strength of this technique, e.g., if $\mathbf{X}$ is the image, then we will need to encode the distribution of natural images!
- We can now encode a rich set of features, without worrying of their distribution.

CRF as a Partially Directed Model

- We can view it as a Markov network where some of the edges are directed, i.e., those from **X** to **Y**.
- Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ and $\mathbf{Y} = \{Y\}$ be binary variables with potentials defined via a log-linear model

$$\phi_i(X_i, Y) = \exp\{w_i \mathbb{1}\{X_i = 1, Y = 1\}\} \qquad \phi_0(Y) = \exp\{w_0 \mathbb{1}\{Y = 1\}\}$$

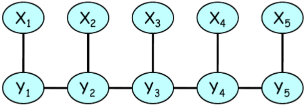
- We can compute the unnormalized conditional probability $\hat{P}(Y = 1|x_1, \cdots, x_n)$ and $\hat{P}(Y = 0|x_1, \cdots, x_n)$.
- Computing the partition function we can show that

$$P(Y = 1|x_1, \cdots, x_n) = sigmoid\left(w_0 + \sum_{i=1}^{n} w_i x_i\right)$$

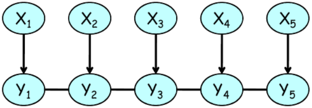
- The CPD is not a table, but it is induced by a set of parameters w_i whose number is linear, instead of exponential with the number of parents.
- This type of CPD is called **logistic CPD**.

CRF as a Partially Directed Model II

- Any CRF can be viewed as a partially directed graph, where the edges from **X** to **Y** are directed and parameterized using logistic CPDs.

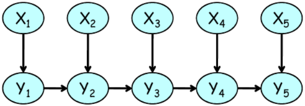


(CRF)



(Partially Directed CRF)

- This is not equivalent to a fully directed model, as the unnormalized marginal measure of **Y** depends on the entire parameterization of the chain.
- A conditional BN for this distribution would require edges from all of the variables in **X** to each of the variables $\mathbf{Y}_i$, loosing much of the structure.



NLP example I

- Name entity recognition task: one target variable for each word, which encodes the possible labels of that word.
- Each word represent with X_i.
- Each target variable is connected to a set of features variables that capture properties relevant to the task.
- Entities sometimes span multiple words.
- Type not always obvious from e.g., New York vs New York Times.
- The targets are for example "B-person, I-person, B-location, I-location, B-organization, I-organization, others".
- Having beginning (B) and outcome (I) allows the model to segment adjacent entities.
- How can we represent this?

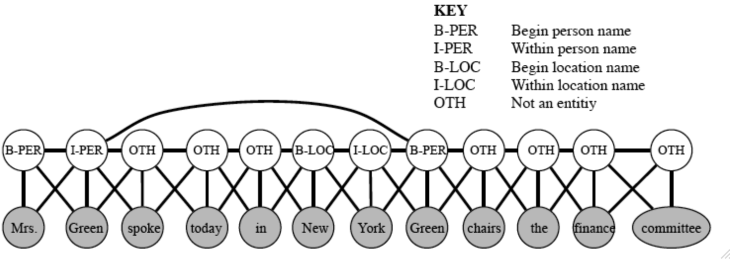
NLP example II

This is typically represented having two factors for each word

- $\phi_t^1(Y_t, Y_{t+1})$ represents dependencies between neighboring target variables.
- $\phi_t^2(Y_t, X_1, \cdots, X_T)$ represents dependencies between a target and its context in the word sequence.

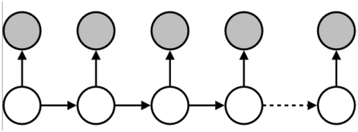
Usually encoded using a log-linear model and features for example

$$f_t(Y_t, X_t) = \mathbb{1}\{Y_t = B - organization, X_t = times\}$$

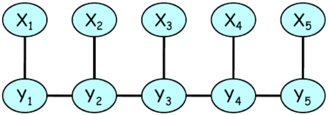


Equivalences

- Logistic regression is the conditional analog of naive Bayes classifier
- The linear-chain CRF is the conditional analog of a hidden Markov model (HMM).



(HMM)



(Linear-Chain CRF)

Next Class

- Tamir will talk about the exponential family
- And the concepts of entropy and relative entropy
- As well as distances between distribution