Probabilistic Graphical Models

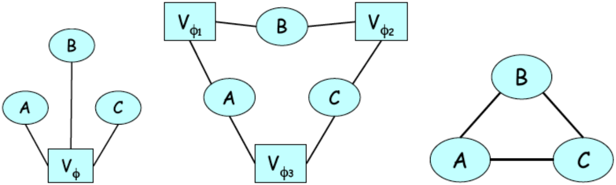
Raquel Urtasun and Tamir Hazan

TTI Chicago

April 8, 2011

Factor Graphs

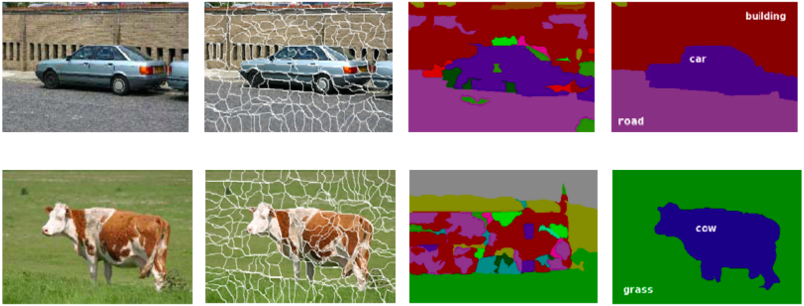
- $\mathcal{H}$ does not reveal the structure of the Gibbs parameterization: maximum cliques vs subsets of them.
- Example: For a complete graph, we could have one factor per edge or a single clique potential for the whole graph
- Factor graphs can distinguish these cases.
- A **factor graph** is an undirected graph containing variables nodes and factor nodes. There are only edges between the variable nodes and the factor nodes. Each factor node is associated with a single factor, which scope is the set of variables that are neighbors in the graph.



- What's the Gibbs distribution?

Example: segmentation

- The graph has only node potentials $\phi_i(X_i)$ and pairwise potentials $\phi_{i,j}(X_i, X_j)$
- Grids are particularly popular, e.g., pixels in an image with 4-connectivity



- What's the factor graph?

Energy-based models: log-linear models

- It is common to work in terms of energies: negative logs of the factors
- Where small energy means more probable

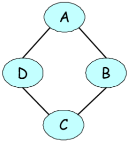
$$p(X_1, \cdots, X_n) = \frac{1}{Z} \exp\left[-\sum_{i=1}^{m} \epsilon_i(\mathbf{D}_i) \right]$$

  where $\epsilon(\mathbf{D}) = -\ln \phi(\mathbf{D})$ is called an **energy function**

- It is called *log-linear model* as the exponent is a linear function.
- Any Markov network parameterized using positive factors can be converted to this representation.

Misconception example

- Factor domain:



$\phi_1[A,B]$			$\phi_2[B,C]$			$\phi_3[C,D]$			$\phi_4[D,A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

- Log domain: $\epsilon(\mathbf{D}) = -\ln \phi(\mathbf{D})$. We see preference of D and A to have the same value.

$\epsilon_1[A,B]$			$\epsilon_2[B,C]$			$\epsilon_3[C,D]$			$\epsilon_4[D,A]$		
a^0	b^0	-3.4	b^0	c^0	-4.61	c^0	d^0	0	d^0	a^0	-4.61
a^0	b^1	-1.61	b^0	c^1	0	c^0	d^1	-4.61	d^0	a^1	0
a^1	b^0	0	b^1	c^0	0	c^1	d^0	-4.61	d^1	a^0	0
a^1	b^1	-2.3	b^1	c^1	-4.61	c^1	d^1	0	d^1	a^1	-4.61

Notion of feature

- Let $\mathbf{D}$ be a subset of variables. We define a **feature** $f(\mathbf{D})$ to be an indicator function for some event defined in $\mathbf{D}$, f takes value 1 for some values $y \in Val(\mathbf{D})$, and 0 otherwise.

$\epsilon_1[A,B]$			$\epsilon_2[B,C]$			$\epsilon_3[C,D]$			$\epsilon_4[D,A]$		
a^0	b^0	-3.4	b^0	c^0	-4.61	c^0	d^0	0	d^0	a^0	-4.61
a^0	b^1	-1.61	b^0	c^1	0	c^0	d^1	-4.61	d^0	a^1	0
a^1	b^0	0	b^1	c^0	0	c^1	d^0	-4.61	d^1	a^0	0
a^1	b^1	-2.3	b^1	c^1	-4.61	c^1	d^1	0	d^1	a^1	-4.61

$$\epsilon(C,D) = \begin{cases} -4.61 & \text{if } C \neq D \\ 0 & \text{otherwise} \end{cases}$$

- This can be represented with a feature $f(C,D)$ which takes value 1 when $C \neq D$.
- The energy is a constant multiply by $f(C,D)$

Definition log-linear model

A distribution P is a log linear model over a Markov network $\mathcal{H}$ if it is associated with:

- a set of features $\Phi = \{f_1(\mathbf{D}_1), \cdots, f_m(\mathbf{D}_m)\}$, where each $\mathbf{D}_i$ is a complete subgraph in $\mathcal{H}$.
- A set of weights $w_1, \cdots, w_m$ such that

$$p(X_1, \cdots, X_n) = \frac{1}{Z} \exp\left[-\sum_{i=1}^{m} w_i f_i(\mathbf{D}_i)\right]$$

- Importantly, we can have several features over the same scope.
- This representation is more compact for many distributions, especially with variables with large domains.

Example: Ising model

- Captures the energy of a set of interacting atoms.
- Each atom $X_i \in \{-1, +1\}$, whose value is the direction of the atom spin.
- The energy of the edges is symmetric and makes a contribution when $X_i = X_j$ (both atoms with the same spin).
- Also individual node potentials that encode the bias of the individual atoms
- The energy associated is

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

- The energy can be written as

$$\epsilon(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{W} (\mathbf{x} - \boldsymbol{\mu}) + c$$

  with $\boldsymbol{\mu} = -\mathbf{W}^{-1}\mathbf{u}, \quad c = \frac{1}{2}\boldsymbol{\mu}^T \mathbf{W} \boldsymbol{\mu}$
- Often modulated by a temperature $p(\mathbf{x}) = \frac{1}{Z} \exp(-\epsilon(\mathbf{x})/T)$
- T small makes distribution picky

What is the factor graph of an Ising model?

- The energy associated is

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

- What's the factor graph?
- What are the features?

Example: Bolzmann machine I

- Is a type of Ising model, i.e., same energy function
- The nodes are taken to have values $\{0, 1\}$.
- The energy then reduces to

$$\epsilon(\mathbf{x}) = \sum_i \epsilon_i(x_i) + \sum_{(i,j)\in\varepsilon} \epsilon_{i,j}(x_i, x_j)$$

  with ε the set of edges

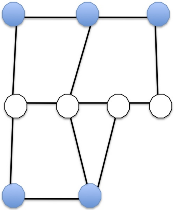
- The probability of each variable given its neighbors is sigmoid(z), with

$$z = -\left(\sum_j w_{i,j} x_j\right) - w_i$$

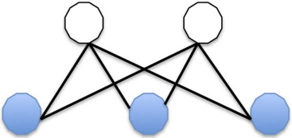
- Which is the simplest model of activation of a neuron

Example: Bolzmann machine II

- Bolzmann machines are usually defined in terms of visible units and hidden units.



(BM)



(RBM)

- A restricted Bolzmann machine does not have connections

Representation of Markov Networks

We have seen 3 representations:

- Markov networks $\mathcal{H}$: involves product over potentials or cliques.
- Factor graphs: product of factors.
- Set of features: product over weighted features.

Usefulness:

- Markov networks are useful for defining independencies
- Factor graphs are useful for inference
- Set of features are useful for learning

Over-parameterization

- Markov network parameterizations are over-parameterized.
- There are multiple choices of parameters that describe the same distribution

$\epsilon_1[A,B]$

a^0	b^0	-3.4
a^0	b^1	-1.61
a^1	b^0	0
a^1	b^1	-2.3

$\epsilon_2[B,C]$

b^0	c^0	-4.61
b^0	c^1	0
b^1	c^0	0
b^1	c^1	-4.61

$\epsilon_3[C,D]$

c^0	d^0	0
c^0	d^1	-4.61
c^1	d^0	-4.61
c^1	d^1	0

$\epsilon_4[D,A]$

d^0	a^0	-4.61
d^0	a^1	0
d^1	a^0	0
d^1	a^1	-4.61

(Original Parameterization)

$\epsilon'_1[A,B]$

a^0	b^0	-4.4
a^0	b^1	1.61
a^1	b^0	-1
a^1	b^1	2.3

$\epsilon'_2[B,C]$

b^0	c^0	-3.61
b^0	c^1	$+1$
b^1	c^0	0
b^1	c^1	4.61

(New Parameterization)

- What's the energy of a particular configuration $\epsilon(a^0, b^0, c^0)$ in both cases?

Conversion between representations

- From BN to Markov networks via **moralization**
- From Markov networks to BN via **triangulation**

From Bayesian Networks to Markov Networks I

- We are interested in finding a minimal I-map from a distribution $P_{\mathcal{B}}$.
- The parameterization of $\mathcal{B}$ can also be viewed as a Gibbs distribution: each CPD $P(X_i|Pa_{X_i})$ is a factor.
- The factor satisfies additional normalization properties, and $(Z = 1)$. Why?
- A BN conditioned on evidence $\mathbf{E}$ also induces a Gibbs distribution: defined by the original factors reduced to the context $\mathbf{E} = \mathbf{e}$.
- Let $\mathcal{B}$ be a BN over $\mathcal{X}$, with $\mathbf{E}$ an observation and $\mathbf{W} = \mathcal{X} - \mathbf{E}$. Then $P_{\mathcal{B}}(\mathbf{W}|\mathbf{e})$ is a Gibbs distribution defined by the factors $\Phi = \{\phi_{X_i}\}_{X_i \in \mathcal{X}}$ with

$$\phi_{X_i} = P_{\mathcal{B}}(X_i|Pa_{X_i})[\mathbf{E} = \mathbf{e}]$$

  and the partition function for this distribution is $P(\mathbf{e})$.
- To create a Markov network we need to create an edge between X_i and each of its parents, as well as between the parents of X_i.

From Bayesian Networks to Markov Networks II

- The **moral graph** $\mathcal{M}[\mathcal{G}]$ of a BN $\mathcal{G}$ over $\mathcal{X}$ is an undirected graph over $\mathcal{X}$ that contains an undirected edge between X and Y if
  1. there is a directed edge between them (in either direction)
  2. X and Y are both parents of the same node.

Let's show some examples on the board !!

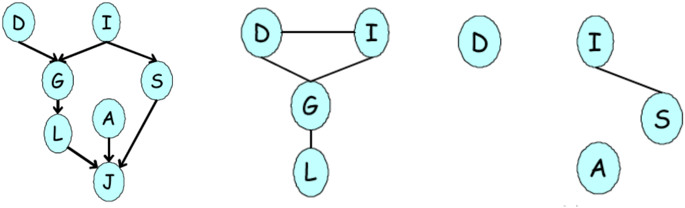
From Bayesian Networks to Markov Networks III

- The **moral graph** $\mathcal{M}[\mathcal{G}]$ of a BN $\mathcal{G}$ over $\mathcal{X}$ is an undirected graph over $\mathcal{X}$ that contains an undirected edge between X and Y if
  1. there is a directed edge between them (in either direction)
  2. X and Y are both parents of the same node.
- For any distribution $P_{\mathcal{B}}$ such that $\mathcal{B}$ is a parameterization of $\mathcal{G}$, then $\mathcal{M}[\mathcal{G}]$ is an I-map for $P_{\mathcal{B}}$.
- The moralized graph $\mathcal{M}[\mathcal{G}]$ is a minimal I-map for $\mathcal{G}$.
- The addition of the moralizing edges leads to the loss of some independence information, e.g., $X \rightarrow Z \leftarrow Y$, where $X \perp Y$ is lost.
- Moralization causes lost of independence if it introduces new edges.
- If $\mathcal{G}$ is moral, then $\mathcal{M}[\mathcal{G}]$ is a perfect map of $\mathcal{G}$.
- If the v-structure can be short cut then it preserves the independencies

D-separation

- Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ be three disjoint sets of nodes in a Bayesian network $\mathcal{G}$. Let $\mathbf{U} = \mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$, and let $\mathcal{G}'$ be the induced Bayesian network over $\mathbf{U} \cup Ancestor_{\mathbf{U}}$. Let $\mathcal{H}$ the moralized graph $\mathcal{M}[\mathcal{G}']$. Then

$$d - sep_{\mathcal{G}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z}) \quad \text{iff} \quad sep_{\mathcal{H}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$$

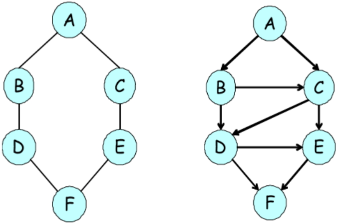


(BN) $\quad d - sep_{\mathcal{G}}(D; I | L) \quad d - sep_{\mathcal{G}}(D; I | L)$

- (center) Moralized graph for $d - sep_{\mathcal{G}}(D; I | L)$, $\mathbf{U} = \{D, I, L\}$.
- (right) Moralized graph for $d - sep_{\mathcal{G}}(D; I | L)$, $\mathbf{U} = \{D, I, L, A\}$.
- If a distribution $P_{\mathcal{B}}$ factorizes according to $\mathcal{G}$, then $\mathcal{G}$ is an I-map fo P.

From Markov Networks to Bayesian Networks

- More difficult transformation, and the BN can be considerably larger.



- Order was $\{A, B, C, D, E, F\}$, but different ordering has the same problems
- It must add edges so that the resulting graph is **chordal**, i.e., all loops have been partitioned into triangles.
- This process is called **triangulation**.
- The addition of edges leads to the loss of independence information, i.e., in the example $(C \perp D | A, F)$.