Probabilistic Graphical Models

Raquel Urtasun and Tamir Hazan

TTI Chicago

April 4, 2011

Bayesian Networks and independences

Not every distribution independencies can be captured by a directed graph

- Regularity in the parameterization of the distribution that cannot be captured in the graph structure, e.g., XOR example

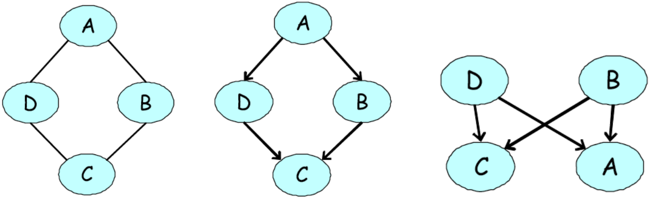
$$P(x, y, z) = \begin{cases} 1/12 & \text{if } x \oplus y \oplus z = \text{false} \\ 1/6 & \text{if } x \oplus y \oplus z = \text{true} \end{cases}$$

    - $(X \perp Y) \in \mathcal{I}(P)$
    - Z is not independent of X given Y or Y given X.
    - An I-map is the network $X \rightarrow Z \leftarrow Y$.
    - This is not a perfect map as $(X \perp Z) \in \mathcal{I}(P)$

- Symmetric variable-level independencies that are not naturally expressed with a Bayesian network.

- Independence assumptions imposed by the structure of the DBN are not appropriate, e.g., misconception example

Misconception example

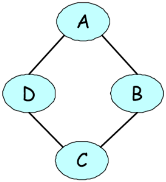


(a) (b) (c)

- (a) Two independencies: $(A \perp C | D, B)$ and $(B \perp D | A, C)$
- Can we encode this with a BN?
- (b) First attempt: encodes $(A \perp C | D, B)$ but it also implies that $(B \perp D | A)$ but dependent given both A, C
- (c) Second attempt: encodes $(A \perp C | D, B)$, but also implies that B and D are marginally independent.

Undirected graphical models I

- So far we have seen directed graphical models or Bayesian networks
- BN do not captured all the independencies, e.g., misconception example,

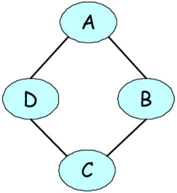


- We want a representation that does not require directionality of the influences. We do this via an undirected graph.
- *Undirected graphical models*, which are useful in modeling phenomena where the interaction between variables does not have a clear directionality.
- Often simpler perspective on directed models, in terms of the independence structure and of inference.

Undirected graphical models II

- As in BN, the **nodes** in the graph represent the variables
- The **edges** represent direct probabilistic interaction between the neighboring variables
- How to parametrize the graph?
    - In BN we used CPD (conditional probabilities) to represent distribution of a node given others
    - For undirected graphs, we use a more symmetric parameterization that captures the affinities between related variables.
- Given a set of random variables $\mathbf{X}$ we define a **factor** as a function from $Val(\mathbf{X})$ to $\Re$.
- The set of variables $\mathbf{X}$ is called the **scope** of the factor.
- Factors can be negative. In general, we restrict the discussion to positive factors

Misconception example once more...



$\phi_1[A,B]$			$\phi_2[B,C]$			$\phi_3[C,D]$			$\phi_4[D,A]$		
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

- We can write the joint probability as

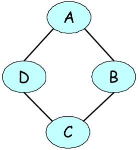
$$p(A,B,C,D) = \frac{1}{Z}\phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(A,D)$$

- Z is the **partition function** and is used to normalized the probabilities

$$Z = \sum_{A,B,C,D} \phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(A,D)$$

- It is called function as it depends on the parameters: important for learning.
- For positive factors, the higher the value of ϕ, the higher the compatibility.
- This representation is very flexible.

Query about probabilities

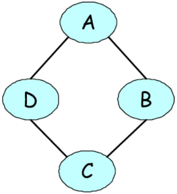


$\phi_1[A,B]$			$\phi_2[B,C]$			$\phi_3[C,D]$			$\phi_4[D,A]$		
a^0	b^0	30	b^0	c^0	100	c^0	d^0	1	d^0	a^0	100
a^0	b^1	5	b^0	c^1	1	c^0	d^1	100	d^0	a^1	1
a^1	b^0	1	b^1	c^0	1	c^1	d^0	100	d^1	a^0	1
a^1	b^1	10	b^1	c^1	100	c^1	d^1	1	d^1	a^1	100

Assignment				*Unnormalized*	*Normalized*
a^0	b^0	c^0	d^0	300000	0.04
a^0	b^0	c^0	d^1	300000	0.04
a^0	b^0	c^1	d^0	300000	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	5000000	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	1000000	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	100000	0.014
a^1	b^1	c^1	d^0	100000	0.014
a^1	b^1	c^1	d^1	100000	0.014

- What's the $p(b^0)$? Marginalize the other variables!

Misconception example once more...



$\phi_1[A,B]$			$\phi_2[B,C]$			$\phi_3[C,D]$			$\phi_4[D,A]$		
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

- We can write the joint probability as

$$p(A,B,C,D) = \frac{1}{Z}\phi_1(A,B)\phi_2(B,C)\phi_3(C,D)\phi_4(A,D)$$

- Use the joint distribution to query about conditional probabilities by summing out the other variables.
- Tight connexion between the factorization and the independence properties

$$\mathbf{X} \perp \mathbf{Y} | \mathbf{Z} \qquad \text{iff} \qquad p(\mathbf{X},\mathbf{Y},\mathbf{Z}) = \phi_1(\mathbf{X},\mathbf{Z})\phi_2(\mathbf{Y},\mathbf{Z})$$

- We see that in the example, $(A \perp C | D, B)$ and $(B \perp D | A, C)$

Factors

- A factor can represent a joint distribution over D by defining $\phi(\mathbf{D})$.
- A factor can represent a CPD $p(X|\mathbf{D})$ by defining $\phi(\mathbf{D} \cup X)$
- But joint and CPD are more restricted, i.e., normalization constraints.
- Associating parameters over edges is not enough
- We need to associate factors over sets of nodes, i.e., higher order terms

Factor product

- Given 3 disjoint set of variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, and factors $\phi_1(\mathbf{X}, \mathbf{Y})$, $\phi_2(\mathbf{Y}, \mathbf{Z})$, the factor product is defined as

$$\psi(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \phi_1(\mathbf{X}, \mathbf{Y})\phi_2(\mathbf{Y}, \mathbf{Z})$$

a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

a^1	b^1	c^1	0.5·0.5
a^1	b^1	c^2	0.5·0.7
a^1	b^2	c^1	0.8·0.1
a^1	b^2	c^2	0.8·0.2
a^2	b^1	c^1	0.1·0.5
a^2	b^1	c^2	0.1·0.7
a^2	b^2	c^1	0·0.1
a^2	b^2	c^2	0·0.2
a^3	b^1	c^1	0.3·0.5
a^3	b^1	c^2	0.3·0.7
a^3	b^2	c^1	0.9·0.1
a^3	b^2	c^2	0.9·0.2

Gibbs distributions and Markov networks

- A distribution P_ϕ is a **Gibbs distribution** parameterized with a set of factors $\phi_1(\mathbf{D}_1), \cdots, \phi_m(\mathbf{D}_m)$ if it is defined as

$$P_\phi(X_1, \cdots, X_n) = \frac{1}{Z}\phi_1(\mathbf{D}_1) \times \cdots \times \phi_m(\mathbf{D}_m)$$

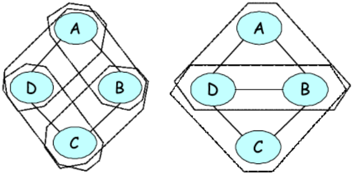
  and the partition function is defined as

$$Z = \sum_{X_1, \cdots, , X_n} \phi_1(\mathbf{D}_1) \times \cdots \times \phi_m(\mathbf{D}_m)$$

- The factors do NOT represent marginal probabilities of the variables of their scope. A factor is only one contribution to the joint.
- A distribution P_ϕ with $\phi_1(\mathbf{D}_1), \cdots, \phi_m(\mathbf{D}_m)$ factorizes over a Markov network $\mathcal{H}$ if each $\mathbf{D}_i$ is a complete subgraph of $\mathcal{H}$
- The factors that parameterize a Markov network are called **clique potentials**

Maximal cliques

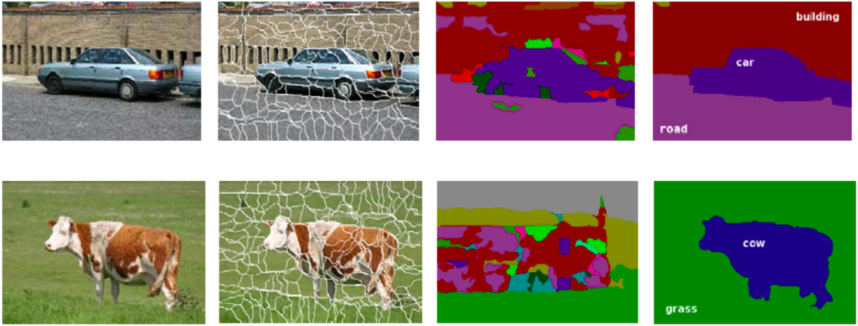
- One can reduce the number of factors by using factors of the maximal cliques



- This obscures the structure
- What's the P_ϕ on the left?
- And on the right?
- What's the relationship between the factors?

Example: Pairwise MRF

- Undirected graphical model very popular in applications such as computer vision: segmentation, stereo, de-noising
- 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



Reduced Markov Networks I

- Conditioning on an assignment **u** to a subset of variables **U** can be done by
  - Eliminating all entries that are inconsistent with the assignment
  - Re-normalizing the remaining entries so that they sum to 1

a^1	b^1	0.5
a^1	b^2	0.8
a^2	b^1	0.1
a^2	b^2	0
a^3	b^1	0.3
a^3	b^2	0.9

b^1	c^1	0.5
b^1	c^2	0.7
b^2	c^1	0.1
b^2	c^2	0.2

a^1	b^1	c^1	0.5·0.5
a^1	b^1	c^2	0.5·0.7
a^1	b^2	c^1	0.8·0.1
a^1	b^2	c^2	0.8·0.2
a^2	b^1	c^1	0.1·0.5
a^2	b^1	c^2	0.1·0.7
a^2	b^2	c^1	0·0.1
a^2	b^2	c^2	0·0.2
a^3	b^1	c^1	0.3·0.5
a^3	b^1	c^2	0.3·0.7
a^3	b^2	c^1	0.9·0.1
a^3	b^2	c^2	0.9·0.2

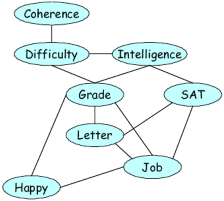
(Original)

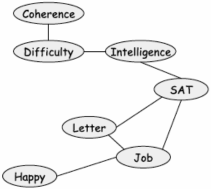
a^1	b^1	c^1	0.25
a^1	b^2	c^1	0.08
a^2	b^1	c^1	0.05
a^2	b^2	c^1	0
a^3	b^1	c^1	0.15
a^3	b^2	c^1	0.09

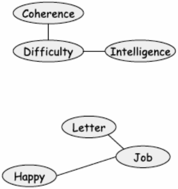
(Cond. on c^1)

Reduced Markov Networks

- Let $\mathcal{H}$ be a Markov network over $\mathbf{X}$ and let $\mathbf{U} = u$ be the context. The reduced network $\mathcal{H}[u]$ is a Markov network over the nodes $\mathbf{W} = \mathbf{X} - \mathbf{U}$ where we have an edge between X and Y if there is an edge between then in $\mathcal{H}$



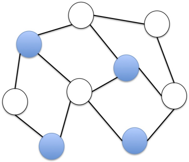




- If $\mathbf{U} = Grade$?
- If $\mathbf{U} = \{Grade, SAT\}$?

Markov Network Independencies I

- As in BN, the graph encodes a set of independencies.
- Probabilistic influence flows along the undirected paths in the graph.
- It is blocked if we condition on the intervening nodes
- A path $X_1 - \cdots - X_k$ is "active" given the observed variables $\mathbf{E} \subseteq \mathcal{X}$ if none of the X_i is in $\mathbf{E}$.



- A set of nodes $\mathbf{Z}$ separates $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{H}$, i.e., $sep_{\mathcal{H}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$, if there exists no active path between any node $X \in \mathbf{X}$ and $Y \in Y$ given $\mathbf{Z}$.
- The definition of separation is monotonic

$$\text{if} \quad sep_{\mathcal{H}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z}) \quad \text{then} \quad sep_{\mathcal{H}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z}') \quad \text{for any} \quad Z' \supseteq Z$$

Markov Network Independencies II

- If P is a Gibbs distribution that factorizes over $\mathcal{H}$, then $\mathcal{H}$ is an I-map for P, i.e., $I(H) \subseteq I(P)$ (soundness of separation)
- Proof: Suppose $\mathbf{Z}$ separates $\mathbf{X}$ from $\mathbf{Y}$. Then we can write

$$p(X_1, \cdots, X_n) = \frac{1}{Z} f(\mathbf{X}, \mathbf{Z}) g(\mathbf{Y}, \mathbf{Z})$$

- A distribution is **positive** if $P(x) > 0$ for all x.
- **Hammersley-Clifford** theorem: If P is a positive distribution over $\mathcal{X}$ and $\mathcal{H}$ is an I-map for P, then P is a Gibbs distribution that factorizes over $\mathcal{H}$

$$p(\mathbf{x}) = \frac{1}{Z} \prod_c \phi_c(\mathbf{x}_c)$$

- It is not the case that every pair of nodes that are not separated in $\mathcal{H}$ are dependent in every distribution which factorizes over $\mathcal{H}$
- If X and Y are not separated given $\mathbf{Z}$ in $\mathcal{H}$, then X and Y are dependent given Z in some distribution that factorizes over $\mathcal{H}$.

Independence assumptions I

In a BN we specify local Markov assumptions and d-separation. In Markov networks we have

1. **Global assumption**: A set of nodes $\mathbf{Z}$ separates $\mathbf{X}$ and $\mathbf{Y}$ if there is no active path between any node $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$ given $\mathbf{Z}$.

$$\mathcal{I}(\mathcal{H}) = \{(\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}) \quad : \quad sep_{\mathcal{H}}(\mathbf{X};\mathbf{Y}|\mathbf{Z})\}$$

2. **Pairwise Markov assumption**: X and Y are independent give all the other nodes in the graph if no direct connection exists between them

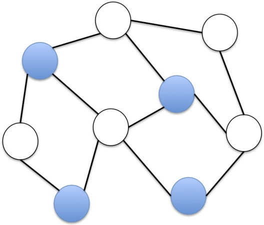
$$\mathcal{I}_p(\mathcal{H}) = \{(X \perp Y|\mathcal{X} - \{X, Y\}) \quad : \quad X - Y \notin \mathcal{X}\}$$

3. **Markov blanket assumption**: X is independent of the rest of the nodes given its neighbors

$$\mathcal{I}_l(\mathcal{H}) = \{(X \perp \mathcal{X} - \{X\} - MB_{\mathcal{H}}(X)|MB_{\mathcal{H}}(X)) \quad : \quad X \in \mathcal{X}\}$$

A set $\mathbf{U}$ is a Markov blanket of X if $X \notin \mathbf{U}$ and if $\mathbf{U}$ is a minimal set of nodes such that $(X \perp \mathcal{X} - \{X\} - \mathbf{U}|\mathbf{U}) \in \mathcal{I}$

Independence assumptions II



(Markov blanket)

- In general $\mathcal{I}(\mathcal{H}) \subseteq \mathcal{I}_l(\mathcal{H}) \subseteq \mathcal{I}_p(\mathcal{H})$
- If P satisfies $\mathcal{I}(\mathcal{H})$, then it satisfies $\mathcal{I}_l(\mathcal{H})$
- If P satisfies $\mathcal{I}_l(\mathcal{H})$, then it satisfies $\mathcal{I}_p(\mathcal{H})$
- If P is a positive distribution and satisfies $\mathcal{I}_p(\mathcal{H})$ then it satisfies $\mathcal{I}_l(\mathcal{H})$

From distributions to graphs I

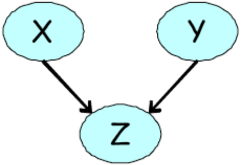
- The notion of I-map is not enough: as in BN the complete graph is an I-map for any distribution, but does not imply any independencies
- For a given distribution, we want to construct a minimal I-map based on the local indep. assumptions
  1. Pairwise: Add an edge between all pairs of nodes that do NOT satisfy $(X \perp Y | \mathcal{X} - \{X, Y\})$
  2. Markov blanket: For each variable X we define the neighbors of X all the nodes that render X independent of the rest of nodes. Define a graph by introducing and edge for all X and all $Y \in MB_P(X)$.
- If P is positive distribution, there is a unique Markov blanket of X in $\mathcal{I}(P)$, denoted $MB_P(X)$

From distributions to graphs II

- If P is a positive distribution, and let $\mathcal{H}$ be the graph defined by introducing an edge $\{X, Y\}$ for which P does NOT satisfied $(X \perp Y | \mathcal{X} - \{X, Y\})$, then $\mathcal{H}$ is the unique minimal I-map for P.
- Minimal I-map is the one that if we remove one edge is not an I-map.
- Proof:
  - $\mathcal{H}$ is an I-map for P since P by construction satisfies $\mathcal{I}_p(P)$ which for positive distributions equals $\mathcal{I}(P)$.
  - To prove that it's minimal, if we eliminate an edge $\{X, Y\}$ the graph would imply $(X \perp Y | \mathcal{X} - \{X, Y\})$, which is false for P, otherwise edge omitted when constructing $\mathcal{H}$.
  - To prove that it's unique: any other I-map must contain the same edges, and it's either equal or contain additional edges, and thus it is not minimal
- If P is a positive distribution, and for each node let $MB_P(X)$ be a minimal set of nodes $\mathbf{U}$ satisfying $(X \perp \mathcal{X} - \{X\} - \mathbf{U} | \mathbf{U}) \in \mathcal{I}$. Define $\mathcal{H}$ by introducing an edge $\{X, Y\}$ for all X and all $Y \in MB_P(X)$. Then $\mathcal{H}$ is the unique minimal I-map of P.

Examples of deterministic relations

- Not every distribution has a perfect map
- Even for positive distributions
- Example is the V-structure, where minimal I-map is the fully connected graph



- It fails to capture the marginal independence $(X \perp Y)$ that holds in P