67800: Probabilistic Methods in AI **Spring 2017/18**

Recitation 3: More on Bayesian and Markov Networks

Teaching Assistant: Eitan Richardson

Recommended reading:

- PGM Book, Chapters 4, (5) [1]

3.1 Reminder – independencies and I-maps

Definition 3.1 *I-map (a general definition)*
A graph $\mathcal{K}$ associated with a set of independencies $\mathcal{I}(\mathcal{K})$ is an I-map for a set of independencies $\mathcal{I}$ if $\mathcal{I}(\mathcal{K}) \subseteq \mathcal{I}$.

Definition 3.2 $\mathcal{I}(P)$
We define $\mathcal{I}(P)$ to be the set of all independencies of the form $(\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$ that hold in P.

3.1.1 In directed models – BNs

We defined the local Markov independencies:

$$\mathcal{I}_{LM}(\mathcal{G}) = \forall i : (X_i \perp \text{NonDescendants}_{X_i} \mid Pa^{\mathcal{G}}_{X_i}))$$

And the global independencies:

$$\mathcal{I}(\mathcal{G}) = \{(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) : \text{d-sep}_{\mathcal{G}}(\mathbf{X}\,; \mathbf{Y} \mid \mathbf{Z})\}$$

We proved the following theorems:

1. $\mathcal{I}_{LM}(\mathcal{G}) \subseteq \mathcal{I}(\mathcal{G})$ – every conditional independence in $\mathcal{I}_{LM}(\mathcal{G})$ appears in $\mathcal{I}(\mathcal{G})$
2. $P \models \mathcal{I}_{LM}(\mathcal{G}) \iff P \models \mathcal{I}(\mathcal{G})$ – they are equivalent (imply each other)
3. $\mathcal{G}$ is an I-map for $P \iff P$ factorizes according to $\mathcal{G}$

☞ *Claim 3 is true with respect to both $\mathcal{I}_{LM}(\mathcal{G})$ and $\mathcal{I}(\mathcal{G})$, because of claim 2.*

☞ *Property 1 is less important since the two sets are equivalent.*

3.1.2 In undirected models – MNs

We defined three sets of independencies associated with the MN graph:

$$\mathcal{I}(\mathcal{H}) = \{(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}) \ : \ \texttt{Sep}_{\mathcal{H}}(\mathbf{X}\,; \mathbf{Y} \mid \mathbf{Z})\}$$

[1] *Original LaTeX template courtesy of UC Berkeley.*

$$\mathcal{I}_P(\mathcal{H}) = \{(X \perp Y \mid \mathcal{X} - \{X, Y\}) \ : \ X - Y \notin \mathcal{H})\}$$

$$\mathcal{I}_{LM}(\mathcal{H}) = \{(X \perp \mathcal{X} - \{X\} - MB_{\mathcal{H}}(X) \mid MB_{\mathcal{H}}(X)) \ : \ X \in \mathcal{X})\}$$

We proved (partially) the following theorems:

1. $P \models \mathcal{I}(\mathcal{H}) \implies P \models \mathcal{I}_{LM}(\mathcal{H}) \implies P \models \mathcal{I}_P(\mathcal{H})$
2. P factorizes according to $\mathcal{H} \implies \mathcal{H}$ is an I-map for P

☞ *We proved claim 2 w.r.t. the global independence $\mathcal{I}(\mathcal{H})$ and therefor it implies the other independencies.*

For **positive** distributions, the following is true as well:

1. $P \models \mathcal{I}(\mathcal{H}) \iff P \models \mathcal{I}_{LM}(\mathcal{H}) \iff P \models \mathcal{I}_P(\mathcal{H})$
2. $\mathcal{H}$ is an I-Map for $P \implies P$ factorizes according to $\mathcal{H}$

3.2 I-maps

A is a graph (MN or BN) with an associated set of independencies $\mathcal{I}(A)$, B is a graph or a distribution with an associated set of independencies $\mathcal{I}(B)$:

Definition	Meaning	Always exists
A is an I-map for B	$\mathcal{I}(A) \subseteq \mathcal{I}(B)$	✓
A is a minimal I-map for B	$\mathcal{I}(A) \subseteq \mathcal{I}(B), \forall \mathcal{E} : \mathcal{I}(A \setminus \mathcal{E}) \nsubseteq \mathcal{I}(B)$	✓
A is a P-map (perfect map) for B	$\mathcal{I}(A) = \mathcal{I}(B)$	No

We discussed an algorithm for constructing a directed (BN) minimal I-map given the set of independencies $\mathcal{I}(B)$ and a **predefined topological order** – when adding X_i, pick the minimal required subset of nodes as parents s.t. X_i is independent of already-added nodes given the set of parents (see Algorithm 3.2 in the book).

We saw in class a method for constructing an undirected minimal I-map – add an edge between every pair of variables that are not independent in P given all other variables. We saw that the resulting minimal I-map is unique.

☞ *We didn't learn how to construct a perfect map. See book section 3.4.3*

3.3 From Markov to Bayesian Networks

Reminder: Last time we saw how to construct a MN that will be a minimal I-map for a given BN – the *Moral Graph* (undirected skeleton of the BN plus edges between co-parents).

Converting a MN to a BN is more challenging – we will see that converting an undirected MN $\mathcal{H}$ to a directed BN $\mathcal{G}$ might add many edges and dependencies.

3.3.1 Constructing a minimal I-map

Definition 3.3 *Chordal Graph*
In a chordal (or triangulated) graph, there are no minimal loops (i.e. undirected cyclic trails without short-cuts) longer than three edges.

Theorem 3.4 *$\mathcal{G}$ is a minimal I-map for $\mathcal{H} \implies \mathcal{G}$ has no immoralities.*
This is true for every topological order.

Proof: Lets assume, by contradiction, that the following immorality exists in $\mathcal{G}$:
$X_i \rightarrow X_j \leftarrow X_k$ with no edge between X_i and X_k, and assuming $i < k < j$.

The minimal I-map construction chooses as Pa_{X_j} the minimal set s.t. $X_j \perp X_{<j} \mid Pa_{X_j}$,
therefore $(X_j \perp X_i \mid Pa_{X_j} - X_i) \notin \mathcal{I}(\mathcal{H})$,
and therefore a path $X_j \ldots X_i$ exists in $\mathcal{H}$, which is not cut by other parents of X_j.

Similarly, a path $X_j \ldots X_k$ exists in $\mathcal{H}$, which is not cut by other parents of X_j,
and so, a path $X_i \ldots X_j \ldots X_k$ exists in $\mathcal{H}$.

Lets consider X_k's parents Pa_{X_k}. Because of the assumed immorality, $X_i \notin Pa_{X_k}$, therefore:
Pa_{X_k} cuts the path $X_i \ldots X_j \ldots X_k$.

Lets assume WLOG that Pa_{X_k} separates (in $\mathcal{H}$) X_j from X_k. This would cause (some variable in) Pa_{X_k} to replace X_k as a parent of X_j – a contradiction.

■

☞ *We use the regular minimal I-map construction algorithm given some topological order. The result will always be a chordal BN.*

Claim 3.5 *$\mathcal{G}$ is a minimal I-map for $\mathcal{H} \implies \mathcal{G}$ is* chordal.

Proof: Any minimal loop larger than three edges will cause an immorality. ■

3.3.2 Constructing a P-map*

Theorem 3.6 *For $\mathcal{H}$ to have a directed P-map $\mathcal{G}$, $\mathcal{H}$ must be chordal.*

Proof: Any minimal I-map for $\mathcal{H}$ must be chordal. If $\mathcal{H}$ is not chordal, the skeleton of any I-map $\mathcal{G}$ will contain edges that are not in $\mathcal{H}$, thus eliminating (pairwise) independencies that $\mathcal{H}$ encodes. ■

☞ ** The proof for the opposite direction requires the definition of a* Clique Tree *– optional material.*

Definition 3.7 *Clique Tree*
A tree $\mathcal{T}$ is a clique tree *for a MN $\mathcal{H}$ if:*

- *Each node in $\mathcal{T}$ corresponds to a clique in $\mathcal{H}$ (and each maximal clique has a node)*
- *For each sepset, we have:* $\texttt{Sep}_{\mathcal{H}}(W_{<(i,j)}\,; W_{<(j,i)} \mid S_{ij})$

Where the sepset is defined as $S_{ij} = C_i \cap C_j$, C_i *and* C_j *are connected by an edge* $i - j$ *in* $\mathcal{T}$.

$W_{<(i,j)}$ *is the set of all variables to the* C_i *side of the edge* $i - j$ *in* $\mathcal{T}$.

$\mathcal{X} = (W_{<(i,j)} - S_{ij}) \cup (W_{<(j,i)} - S_{ij}) \cup S_{ij}$

Theorem 3.8 $\mathcal{H}$ *is chordal* $\implies$ $\mathcal{H}$ *has a clique tree.*

Proof: By induction... ■

Theorem 3.9 *If* $\mathcal{H}$ *is chordal, a BN which is a P-map for* $\mathcal{H}$ *exists.*

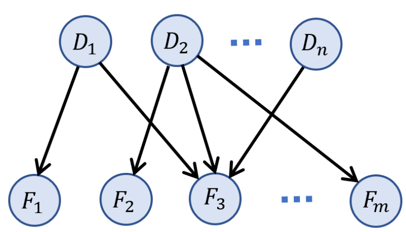
Proof: (Proof sketch)
We first induce a topological order based on the clique tree:
We select some clique C_1 in the clique tree $\mathcal{T}$ to be the root and then define an order for the other cliques so that $i < j \implies C_i$ is closer to the root than C_j.
We then define a topological order for the variables $\mathcal{X}$ that is consistent with the cliques order and construct the BN $\mathcal{G}$ using the minimal I-map algorithm.

Next, we show that $\mathcal{G}$ is a P-map for $\mathcal{H}$ (see proof for theorem 4.13 in the book).

■

3.4 The BN2O model

A *BN2O* network is a two-layer BN, where the top layer corresponds to *causes* (e.g. diseases) and the bottom layer to symptoms (e.g. medical findings):



We assume all variables are binary and each bottom-layer CPD is defined by a *noisy-or* model:

$$P(f_i^0 \mid \mathbf{Pa}_{F_i}) = (1 - \lambda_{i,0}) \prod_{D_j \in \mathbf{Pa}_{F_i}} (1 - \lambda_{i,j})^{d_j}$$

where $\lambda_{i,j}$ is the noise parameter associated with parent D_j of variable F_i.

The model is simple and intuitive:

- An edge indicate that a disease D_j can cause a symptom F_i
- The parameter $\lambda_{i,j}$ defined the probability that $d_j = 1$ causes $f_i = 1$ (in isolation)

We will prove two more very useful properties of BN2O:

1. The parents of a symptom F_i are independent given f_i^0.

2. The posterior distribution with a negative observation $P_B(\cdot \mid f_i^0)$ can be encoded by a BN B' which has an identical structure to B, except that F_i is omitted.

Regarding the first property:

Generally in BNs, an observation of a child makes the parents dependent (v-structure), like in the *explaining away* example. In general CPD tables, parents become dependent on either positive or negative child observation.

The noisy-or CPD causes *context-specific independence*. Intuitively, if someone doesn't have a fever, the fact that he has a flu does not change the probability that he has strep.

Proof: The parents in a *noisy-or* BN are independent given a negative child observation.
Notation: D – the set of k parents, F – the child.

$$
\begin{aligned}
P(D \mid f^0) &= \frac{P(D, f^0)}{P(f^0)} = \frac{\left[\prod_{i=1}^{k} P(D_i)\right](1-\lambda_0)\prod_{i=1}^{k}(1-\lambda_i)^{D_i}}{P(f^0)} \\
&= \frac{(1-\lambda_0)\prod_{i=1}^{k}\left[P(D_i)(1-\lambda_i)^{D_i}\right]}{\sum_D P(D, f^0)} \\
&\propto \prod_{i=1}^{k}\left[P(D_i)(1-\lambda_i)^{D_i}\right] = \prod_{i=1}^{k}\psi(D_i)
\end{aligned}
$$

If a joint distribution can be written as a product of functions of single variables, then the variables are independent. ∎

☞ *We can also prove explicitly that* $P(D \mid f^0) = \prod_i (D_i \mid f^0)$ *... a few more lines.*

The second property is useful, since most symptoms are typically negative.

Proof:

$$
\begin{aligned}
P(D, F - F_i \mid f_i^0) &= \frac{P(D, F - F_i, f_i^0)}{P(f_i^0)} \\
&= \frac{\left[\prod_{j=1}^{n} P(D_j)\right]\left[\prod_{k\neq i} P(F_k \mid Pa_{F_k})\right] P(f_i^0 \mid Pa_{F_i})}{P(f_i^0)} \\
&= \left[\prod_{j=1}^{n} P(D_j)\right]\left[\prod_{k\neq i} P(F_k \mid Pa_{F_k})\right]\frac{P(Pa_{F_i} \mid f_i^0)\cancel{P(f_i^0)}}{P(Pa_{F_i})\cancel{P(f_i^0)}} \quad \text{(Bayes rule)} \\
&= \left[\prod_{j\notin Pa_{F_i}} P(D_j)\right]\left[\prod_{k\neq i} P(F_k \mid Pa_{F_k})\right]\left[\cancel{\prod_{j\in Pa_{F_i}} P(D_j)}\right]\frac{P(Pa_{F_i} \mid f_i^0)}{\cancel{P(Pa_{F_i})}} \\
&= \left[\prod_{j\notin Pa_{F_i}} P(D_j)\right]\left[\prod_{k\neq i} P(F_k \mid Pa_{F_k})\right]\left[\prod_{j\in Pa_{F_i}} P(D_j \mid f_i^0)\right]
\end{aligned}
$$

The parameters $P(D_j)$ for the parents of F_i are changed to the posterior parameters $P(D_j \mid f_i^0)$. Other parameters are unchanged. ∎

References

[1] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.