Causal Inference and Probabilistic Graphical Models

MoeKid101

0 Properties of Conditional Independence

- Symmetry: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z} \Leftrightarrow \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{Z}$.
- Decomposition: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}, \boldsymbol{W} | \boldsymbol{Z} \Rightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}$.
- Weak union: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}, \boldsymbol{W} | \boldsymbol{Z} \Rightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}, \boldsymbol{W}$.
- Contraction: $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} | \boldsymbol{Z}, \boldsymbol{Y}) \wedge (\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}) \Rightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}, \boldsymbol{W} | \boldsymbol{Z}$.
- Intersection: (for only positive distributions) $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}, \boldsymbol{W}) \wedge (\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} | \boldsymbol{Z}, \boldsymbol{Y}) \Rightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}, \boldsymbol{W} | \boldsymbol{Z}$.

Proof. From the two assumptions we have $\mathbf{Pr}\left[\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{Z}\right] = \mathbf{Pr}\left[\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}\right]$. So $\dfrac{\mathbf{Pr}\left[\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{Z}\right]}{\mathbf{Pr}\left[\boldsymbol{W}|\boldsymbol{Z}\right]} = \dfrac{\mathbf{Pr}\left[\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}\right]}{\mathbf{Pr}\left[\boldsymbol{Y}|\boldsymbol{Z}\right]}$.

$$\sum_{\boldsymbol{w} \in \mathrm{Val}(\boldsymbol{W})} \mathbf{Pr}\left[\boldsymbol{X}, \boldsymbol{W}|\boldsymbol{Z}\right] \mathbf{Pr}\left[\boldsymbol{Y}|\boldsymbol{Z}\right] = \sum_{\boldsymbol{w} \in \mathrm{Val}(\boldsymbol{W})} \mathbf{Pr}\left[\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}\right] \mathbf{Pr}\left[\boldsymbol{W}|\boldsymbol{Z}\right]$$

Then we have $\mathbf{Pr}\left[\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}\right] = \mathbf{Pr}\left[\boldsymbol{X}|\boldsymbol{Z}\right] \mathbf{Pr}\left[\boldsymbol{Y}|\boldsymbol{Z}\right]$, i.e. $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}$. Symmetrically, $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{W} | \boldsymbol{Z}$. □

1 Representing Directed PGMs

The goal is to find a compact representation of probabilities so that it would not cost 2^n storage for merely n binary random variables. The intuition is to utilizing conditional independence.

Definition 1. (**Naive Bayes Model**) *Assuming instances belong to a number of classes, i.e. $C \in \{c_1, ..., c_k\}$ for any sample and there are a number of features $X_1, ..., X_n$. The model is called a naive Bayes model iff. it satisfies the* ***naive Bayes assumption*** *that within each class, all features are conditionally independent ($X_i \perp\!\!\!\perp \boldsymbol{X} - \{X_i\} | C$).*

From the definition of Naive Bayes Model it is straightforward that this model is used for classification. $\mathbf{Pr}\left[C = c_i | x_1 \sim x_n\right] = \mathbf{Pr}\left[C = c_i\right] \prod_{j=1}^{n} \mathbf{Pr}\left[x_j | C = c_i\right]$ gives a very simple way to determine probability and confidence as well.

Definition 2. (**Bayesian Network Structures**) *The structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of a Bayesian network is a DAG with $\boldsymbol{pa}_{X_i}$ denoting the parents of X_i in $\mathcal{G}$ and $\boldsymbol{nd}_{X_i}$ denoting the vertices that are not descendants of X_i in G. These networks satisfy* ***local independency*** *assumptions that $\forall X_i \in \mathcal{V}$, $X_i \perp\!\!\!\perp \boldsymbol{nd}_{X_i} | \boldsymbol{pa}_{X_i}$. These independency assumptions are denoted by $\mathcal{I}(P)$ where P is the distribution of variables $X_1 \sim X_n$.*

Definition 3. (**I-Maps**) *$\mathcal{K}$ is any graph associated with independencies $\mathcal{I}(\mathcal{K})$. $\mathcal{K}$ is an I-map for a set of independencies $\mathcal{I}$ if $\mathcal{I}(\mathcal{K}) \subseteq \mathcal{I}$. (I-map ensures only correctness of $\mathcal{I}(\mathcal{K})$ but not sufficiency.)*

Definition 4. (**Factorization**) *If $\mathcal{G}$ is a BN graph with $\mathcal{V} = \{X_1 \sim X_n\}$ and P is the distribution of these variables, then P factorizes according to $\mathcal{G}$ when $P(X_1 \sim X_n) = \prod_{i=1}^{n} \boldsymbol{Pr}\left[X_i | \boldsymbol{pa}_{X_i}\right]$.*

Definition 5. (**Bayesian Network**) *A Bayesian network is $\mathcal{B} = (\mathcal{G}, P)$ where P factorizes over $\mathcal{G}$.*

Theorem 1. ***(Connecting I-map and factorization)*** *$\mathcal{G}$ is an I-map for P iff. P factorizes according to $\mathcal{G}$.*

Proof. For a DAG $\mathcal{G}$, there is always a topological ordering. So wlog, assume $\mathcal{G}$ satisfies such ordering. Then we know that in $X_1, X_2, ..., X_n$, any edge $X_i \longrightarrow X_j$ satisfies $i < j$. Consequently, $\mathbf{pa}_{X_i} \subset \{X_1 \sim X_{i-1}\}$, and let $\mathbf{np}_{X_i} = \{X_1 \sim X_{i-1}\} - \mathbf{pa}_{X_i} \subset \mathbf{nd}_{X_i}$. We have

$$\begin{aligned} P(X_1 \sim X_n) &= \prod_{i=1}^{n} \mathbf{Pr}\left[X_i | X_1 \sim X_{i-1}\right] = \prod_{i=1}^{n} \mathbf{Pr}\left[X_i | \mathbf{pa}_{X_i}, \mathbf{np}_{X_i}\right] = \prod_{i=1}^{n} \frac{\mathbf{Pr}\left[X_i, \mathbf{np}_{X_i} | \mathbf{pa}_{X_i}\right] \mathbf{Pr}\left[\mathbf{pa}_{X_i}\right]}{\mathbf{Pr}\left[\mathbf{pa}_{X_i}\right] \mathbf{Pr}\left[\mathbf{np}_{X_i} | \mathbf{pa}_{X_i}\right]} \\ &= \prod_{i=1}^{n} \mathbf{Pr}\left[X_i | \mathbf{pa}_{X_i}\right] \end{aligned}$$

Conversely, we try proving $\mathbf{Pr}\left[X_i | \mathbf{pa}_{X_i}\right] = \mathbf{Pr}\left[X_i | \mathbf{pa}_{X_i}, \mathbf{nd}_{X_i}\right]$. Notice that we can always ignore the descendants of X_i, and in the induced subgraph $\mathcal{G}_i$ of $\mathcal{G}$ by $\{X_i\} \cup \mathbf{pa}_{X_i} \cup \mathbf{nd}_{X_i}$, X_i can always be the last term in topological ordering.

$$\mathbf{Pr}\left[X_i | \mathbf{pa}_{X_i}, \mathbf{nd}_{X_i}\right] = \frac{\mathbf{Pr}\left[X_i, \mathbf{pa}_{X_i}, \mathbf{nd}_{X_i}\right]}{\mathbf{Pr}\left[\mathbf{pa}_{X_i}, \mathbf{nd}_{X_i}\right]} = \frac{\prod_{j=1}^{k} \mathbf{Pr}\left[X'_j | \mathbf{pa}_{X'_j}\right]}{\prod_{j=1}^{k-1} \mathbf{Pr}\left[X'_j | \mathbf{pa}_{X'_j}\right]} = \mathbf{Pr}\left[X'_k | \mathbf{pa}_{X'_k}\right]$$

Here $X'_k = X_i$. The second equality holds because they are identical terms in different DAGs. □

For a Bayesian network, we can infer other dependencies and independencies according to the given structure $\mathcal{G}$. Considering basic dependency relationships, we have the following four empirical conclusion:

- X and Y are dependent if $(X, Y) \in \mathcal{E}$ or $(Y, X) \in \mathcal{E}$.
- X and Y are *likely* dependent if there exists a path between X and Y.
- X and Y are conditionally independent given Z if **(1)** there is only one path between X and Y; **(2)** Z intercepts the path.
- Y and Z are independent conditional on X if **(1)** X is a common cause of Y and Z; **(2)** there exists only one path between Y and Z in the undirected graph.
- X and Y are unconditionally independent but conditionally dependent given Z or any descendants of Z if **(1)** Z is the collision node between X and Y; **(2)** there exists only one path between X and Y in the undirected graph.

With these basic rules dealing with chains, forks and colliders, we can use **d-separation** as a formal expression of the above four empirical conclusions to measure in any graph whether variable nodes are connected (or dependent, termed as **d-connected** whose contrary is **d-separated**). Notice that being d-connected doesn't necessarily mean dependent, but being d-separated means independent.

Definition 6. (**Blocking**) *For DAG $\mathcal{G}$, a trail (not necessarily a path) p in $\mathcal{G}$ is blocked by a set of nodes Z iff.*
(1) p contains a chain $A \longrightarrow B \longrightarrow C$ or a fork $A \longleftarrow B \longrightarrow C$ and $B \in Z$;
(2) p contains a collider $A \longrightarrow B \longleftarrow C$ and none of B and its descendants are in Z.

Definition 7. (**d-Separated**) *$\boldsymbol{X}$ and $\boldsymbol{Y}$ as two sets of nodes are d-separated conditional on $\boldsymbol{Z}$ if $\boldsymbol{Z}$ blocks every trail between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in $\mathcal{G}$, annotated d-sep$_{\mathcal{G}}(\boldsymbol{X}; \boldsymbol{Y} | \boldsymbol{Z})$. The set of independence assertions corresponding to d-separation is* ***global Markov independencies*** *defined by $\mathcal{I}(\mathcal{G}) = \{(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}) | \text{d-sep}_{\mathcal{G}}(\boldsymbol{X}; \boldsymbol{Y} | \boldsymbol{Z})\}$.*

Theorem 2. ***(Soundness of d-separation)*** *If P factorizes over $\mathcal{G}$, then $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$.*

Proof. The proof utilizes the soundness of undirected separation and correspondence between Bayesian networks and Markov networks. For any d-sep$_{\mathcal{G}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$, we can construct $\mathcal{H} = \mathcal{M}[\mathcal{G}^+[\boldsymbol{U}]]$ and obtain sep$_{\mathcal{H}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$. By the soundness of undirected separation, $\boldsymbol{X}\perp\!\!\!\perp\boldsymbol{Y}|\boldsymbol{Z}$ in the Gibbs distribution $P_{\mathcal{G}^+[\boldsymbol{U}]}$ factorizing $\mathcal{H}$. Then we will only have to prove $P_{\mathcal{G}^+[\boldsymbol{U}]}(\boldsymbol{U}^*) = P_{\mathcal{B}}(\boldsymbol{U}^*)$. Notice the Gibbs distribution has exactly the same factors as $\mathbf{Pr}\left[X_i|\mathbf{pa}_{X_i}\right]$, and $U \in \boldsymbol{U}^* \Rightarrow \mathbf{pa}_U \subseteq \boldsymbol{U}^*$,

$$P_{\mathcal{B}}(\boldsymbol{U}^*) = \sum_{X_i\in\mathcal{V}-\boldsymbol{U}^*}\prod_{i=1}^{n}\mathbf{Pr}\left[X_i|\mathbf{pa}_{X_i}\right] = \left(\prod_{X_i\in\boldsymbol{U}^*}\mathbf{Pr}\left[X_i|\mathbf{pa}_{X_i}\right]\right)\left(\sum_{X_i\in\mathcal{V}-\boldsymbol{U}^*}\prod_{X_i\in\mathcal{V}-\boldsymbol{U}^*}\mathbf{Pr}\left[X_i|\mathbf{pa}_{X_i}\right]\right)$$
$$=P_{\mathcal{G}^+[\boldsymbol{U}]}(\boldsymbol{U}^*)\left(\sum_{X_i\in\mathcal{V}-\boldsymbol{U}^*}\mathbf{Pr}\left[\mathcal{V}-\boldsymbol{U}^*|\boldsymbol{U}^*\right]\right) = P_{\mathcal{G}^+[\boldsymbol{U}]}$$

Now we have proven that d-sep$_{\mathcal{G}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \Rightarrow \boldsymbol{X}\perp\!\!\!\perp\boldsymbol{Y}|\boldsymbol{Z}$. So $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$. □

Theorem 3. ***(Completeness of d-separation)*** *For any $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$, if not d-sep$_{\mathcal{G}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$, then $\neg(\boldsymbol{X}\perp\!\!\!\perp\boldsymbol{Y}|\boldsymbol{Z})$ in P factorizing over $\mathcal{G}$ (completeness).*

2 Representing Undirected PGMs

The Bayesian models aren't always useful. For example, when we want $A\perp\!\!\!\perp C|\{B,D\}$ and $B\perp\!\!\!\perp D|\{A,C\}$ without extraneous independencies, Bayesian model fails, due to its directions. Therefore, in undirectional PGMs, we use symmetric values to estimate the affinities between variables.

Definition 8. (**Factor**) *For the set $\boldsymbol{D} \subset \mathcal{X}$ of rvs, a factor is $\phi : Val(\boldsymbol{D}) \to \mathbb{R}$. $\boldsymbol{D}$ is the scope of ϕ, denoted $Scope[\phi]$.*

Definition 9. (**Factor Product**) *For disjoint sets $\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}$, suppose there are two factors $\phi_1(\boldsymbol{X},\boldsymbol{Y})$ and $\phi_2(\boldsymbol{Y},\boldsymbol{Z})$. The factor product $\psi = \phi_1 \times \phi_2 : Val(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) \to \mathbb{R}$ is defined by $\psi(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = \phi_1(\boldsymbol{X},\boldsymbol{Y}) \cdot \phi_2(\boldsymbol{Y},\boldsymbol{Z})$.*

Here factors describe local relationships between variables, which resembles the conditional probability used in Bayesian model. And similarly to the case of Bayesian model, we define $P(X_1 \sim X_n)$ as the product of multiple factors. Notably, joint distributions and conditional distributions in Bayesian models are both instances of factors.

Definition 10. (**Gibbs Distribution**) *A distribution P_Φ is a Gibbs distribution parameterized by factors $\Phi = \{\phi_1(\boldsymbol{D}_1) \sim \phi_K(\boldsymbol{D}_K)\}$ if $P_\Phi(X_1 \sim X_n) = \frac{1}{Z}\prod_{j=1}^m \phi_j(\boldsymbol{D}_j)$ with Z as a normalizing constant.*

Definition 11. (**Factorization**) *A Gibbs distribution P_Φ factorizes over a Markov network $\mathcal{H}$ if $\forall k \in [K]$, $\boldsymbol{D}_k$ is a complete subgraph of $\mathcal{H}$. Here the factors parameterizing the Markov network are* ***clique potentials****.*

Since every complete subgraph $\boldsymbol{D}_k$ is the subset of some maximal clique, we can let restrict for every factor that the scope can only be maximal subgraphs. However, in this case, the required number of parameters might exceed the original representation. For instance, a complete graph encoding pairwise interactions requires $O(n^2)$ entries, but maximal clique potentials require $O(2^n)$ entries.

Definition 12. (**Factor Reduction**) *To model conditional probabilities $\boldsymbol{Pr}[...|\boldsymbol{U} = \boldsymbol{u}]$ where $\boldsymbol{U} \subseteq \mathcal{V}$, the reduction $\phi[\boldsymbol{U} = \boldsymbol{u}]$ or $\phi[\boldsymbol{u}]$ with $\boldsymbol{U} \subseteq \boldsymbol{Y}$ is $\phi[\boldsymbol{u}](\boldsymbol{y}') = \phi(\boldsymbol{y}',\boldsymbol{u})$; while for $\boldsymbol{U} \nsubseteq \boldsymbol{Y}$, $\phi[u] = \phi[\boldsymbol{U} \cap \boldsymbol{Y} = \boldsymbol{u}']$.*

Definition 13. (**Reduced Gibbs Distribution**) *For P_Φ parameterized by $\Phi = \{\phi_1 \sim \phi_K\}$ and context $\boldsymbol{u}$, the reduced Gibbs distribution is $P_\Phi[u]$ defined by $\Phi[u] = \{\phi_1[\boldsymbol{u}] \sim \phi_K[\boldsymbol{u}]\}$.*

Definition 14. (**Reduced Markov Network**) *$\mathcal{H}$ is a Markov network over $\mathcal{V}$ and $\boldsymbol{U} = \boldsymbol{u}$ is a context. The reduced Markov network is the induced subgraph $\mathcal{H}[\boldsymbol{u}]$ from $\mathcal{H}$ by $\mathcal{V} - \boldsymbol{U}$. Here conditioning eliminates edges, while in Bayesian networks conditioning also creates edges.*

Definition 15. (**Blocking**) *For a Makov network structure $\mathcal{H}$, a path p is blocked by $\boldsymbol{Z} \subseteq \mathcal{V}$ if there exists $X_i \in p \cap \boldsymbol{Z}$.*

Definition 16. (**Separation**) *$\boldsymbol{Z}$ separates $\boldsymbol{X}$ and $\boldsymbol{Y}$ in $\mathcal{H}$ (denoted $sep_{\mathcal{H}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$) if all paths between $\boldsymbol{X}$ and $\boldsymbol{Y}$ are blocked by $\boldsymbol{Z}$. The corresponding global independency assertions are $\mathcal{I}(\mathcal{H}) = \{(\boldsymbol{X}\perp\!\!\!\perp\boldsymbol{Y}|\boldsymbol{Z})|sep_{\mathcal{H}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})\}$.*

Theorem 4. ***(Soundness of separation)*** *For distribution P and Markov network structure $\mathcal{H}$, if P is a Gibbs distribution factorizing over $\mathcal{H}$, then $\mathcal{I}(\mathcal{H}) \subseteq \mathcal{I}(P)$.*

Proof. Soundness is equivalent to saying $\text{sep}_{\mathcal{H}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \Rightarrow \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$ under the condition of factorization.

Lemma 5. *Suppose $\mathcal{X} = \boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$, then $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$ iff. $P(\mathcal{X}) = \phi_1(\boldsymbol{X},\boldsymbol{Z})\phi_2(\boldsymbol{Y},\boldsymbol{Z})$. (We should observe that when $\boldsymbol{Z}$ is not conditioned on, this decomposition won't lead to independence.)*

Using the lemma, we first prove soundness with assumption $\mathcal{X} = \boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$. Suppose $(X,Y) \in \boldsymbol{X} \times \boldsymbol{Y}$ satisfies $\{X,Y\} \in \boldsymbol{D}_j$, then X and Y are not separated because $\boldsymbol{D}_j$ is a complete subgraph. Consequently, any $\boldsymbol{D}_j$ contains either $\boldsymbol{X} \cup \boldsymbol{Z}$ or $\boldsymbol{Y} \cup \boldsymbol{Z}$. Then we can group $\boldsymbol{D}_j$ into those don't contain $\boldsymbol{Y}$ (and thus $\boldsymbol{D}_j \subseteq \boldsymbol{X} \cup \boldsymbol{Z}$) and those don't contain $\boldsymbol{X}$, which allows for a decomposition into $P(\mathcal{H}) = f(\boldsymbol{X},\boldsymbol{Z})g(\boldsymbol{Y},\boldsymbol{Z})$.

When $\mathcal{X} = \boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z} \cup \boldsymbol{U}$, then there doesn't exist a path passing through $\boldsymbol{U}$ while not passing through $\boldsymbol{Z}$. In the induced subgraph of $\mathcal{H}$ by $\boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{U}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not connected. Consequently, there exists partition $\boldsymbol{U} = \boldsymbol{U}_1 \cup \boldsymbol{U}_2$ such that in $\mathcal{H}$, there doesn't exist a path from $\boldsymbol{X} \cup \boldsymbol{U}_1$ to $\boldsymbol{Y} \cup \boldsymbol{U}_2$ not passing through $\boldsymbol{Z}$. Then the lemma is applied for $\boldsymbol{X} \cup \boldsymbol{U}_1$ and $\boldsymbol{Y} \cup \boldsymbol{U}_2$ given $\boldsymbol{Z}$. □

Theorem 6. ***(Completeness of separation)*** *If X and Y are not separated given $\boldsymbol{Z}$, then they are conditionally dependent in some distribution P factorizing over $\mathcal{H}$.*

Proof. Since X and Y are not separated, we can find a minimal path $p : X = U_0 \to U_1 \to \ldots \to U_k = Y$ where minimal means $\forall j \neq i \pm 1$, there doesn't exist an edge $U_i - U_j$, then any clique $\boldsymbol{C}_i$ containing $U_i - U_{i+1}$ must be different. Without loss of generality, we assume all variables are binary-valued, and $\phi_i(\boldsymbol{C}_i) = W$ (very large) when $u_i = u_{i+1}$, otherwise $\phi_i(\boldsymbol{C}_i) = 1$. Then the constructed probability yields dependent X and Y. □

Similar to the Bayesian models, we can also strengthen the conclusion to be: for almost all distributions except for some measure-zero distributions P that factorizes over $\mathcal{H}$, we have $\mathcal{I}(P) = \mathcal{I}(\mathcal{H})$.

2.1 Converting Bayesian Networks into Markov Networks

There are two approaches, one given probability $P_{\mathcal{B}}$, we represent $P_{\mathcal{B}}$ as a Markov network (find minimal I-map for distribution); the other given graph $\mathcal{G}$, we represent independencies in $\mathcal{G}$ using undirected graph $\mathcal{H}$ (find minimal I-map for independencies).

When given $P_{\mathcal{B}}$ parameterized over $\mathcal{G}$, this $P_{\mathcal{B}}$ is exactly a Gibbs distribution. Moreover, conditioning on evidence $\boldsymbol{E} = \boldsymbol{e}$, the distribution is exactly the reduced Gibbs distribution with context $\boldsymbol{E} = \boldsymbol{e}$.

When building Markov network from independencies $\mathcal{G}$, we want for each X_i to be connected to $\mathbf{pa}_{X_i}$ and $\mathbf{pa}_{X_i}$ forms a complete subgraph. The resulting graph is called moralized graph.

Definition 17. (**Moralized Graph**) *The moral graph $\mathcal{M}[\mathcal{G}]$ is the undirected graph containing an edge $(X,Y) \in \mathcal{E}_{\mathcal{M}[\mathcal{G}]}$ if $(X,Y) \in \mathcal{E}_{\mathcal{G}}$ or $(Y,X) \in \mathcal{E}_{\mathcal{G}}$ or X,Y are the parents of the same node.*

Theorem 7. *The moralized graph $\mathcal{M}[\mathcal{G}]$ is a minimal I-map (thus expressing maximal independencies) for $\mathcal{G}$.*

Proof. Consider the **Markov blanket** (the parents, children and other parents of children) of X, denoted $\text{MB}_{\mathcal{G}}(X)$. Then $\text{MB}_{\mathcal{G}}(X)$ is exactly the set d-separating X from other variables in $\mathcal{G}$ (no subset has this property, requiring parents of children because conditioning on common children opens the path). □

Definition 18. (**Induced Subgraph of Upward Closure**) *For $\boldsymbol{U} \subseteq \boldsymbol{V}$ in $\mathcal{G}$, $\boldsymbol{U}^* = \boldsymbol{U} \cup Ancestors_{\boldsymbol{U}}$ is the upward closure. $\mathcal{G}^+[\boldsymbol{U}]$ is the induced subgraph from $\mathcal{G}$ by $\boldsymbol{U}^*$.*

Theorem 8. *For disjoint nodes $\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}$ in a Bayesian network $\mathcal{G}$, if $\boldsymbol{U} = \boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{Z}$, let $\mathcal{H} = \mathcal{M}[\mathcal{G}^+[\boldsymbol{U}]]$, then d-sep$_{\mathcal{G}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z}) \Leftrightarrow$ sep$_{\mathcal{H}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$.*

Proof. First assume $\text{d-sep}_{\mathcal{G}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$ and $\neg\text{sep}_{\mathcal{H}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$, then all trails connecting $\boldsymbol{X}$ and $\boldsymbol{Y}$ in $\mathcal{G}$ are blocked by $\boldsymbol{Z}$ but there exists a path $p_0 \in \mathcal{H}$ satisfying $\boldsymbol{X}$ is connected to $\boldsymbol{Y}$ without passing $\boldsymbol{Z}$. There are two possibilities, either being a new edge that doesn't exist in $\mathcal{G}$ or $X \leftrightsquigarrow A \to B \leftarrow C \leftrightsquigarrow Y$ with B's descendants not in $\boldsymbol{Z}$. The latter case implies that B is an ancestor of $\boldsymbol{X}$ (otherwise $B \notin \mathcal{H}$), which means the trail $X \leftarrow B \leftarrow C \leftrightsquigarrow Y$ is blocked by $\boldsymbol{Z}$. This procedure is done infinitely, which is clearly impossible. The former case is $X \leftrightsquigarrow A - B \leftrightsquigarrow Y$ with A and B being parents of C, which is reduced to the latter case. Contradiction!

Then assuming $\text{sep}_{\mathcal{H}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$, every path in $\mathcal{H}$ passes through Z. Moreover, all $A \rightarrow Z \leftarrow B$ are illegal without other presence of $\boldsymbol{Z}$ in any path because it creates shortcut $A - B$ in $\mathcal{H}$. Consequently, $\text{d-sep}_{\mathcal{G}^+[\boldsymbol{U}]}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$ holds. However, if there exists a trail passing through $\boldsymbol{G} - \boldsymbol{G}^+[\boldsymbol{U}]$, then it is $X \leftrightsquigarrow A \rightarrow p' \leftarrow B \leftrightsquigarrow Y$ (otherwise vertices in p' is in $\mathcal{H}$). Therefore there always exist $A' \rightarrow B' \leftarrow C'$ in p', which is blocked. Therefore, $\text{d-sep}_{\mathcal{G}}(\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{Z})$ holds. □

2.2 Converting Markov Networks into Bayesian Networks

Definition 19. (**Chordal Graph**) *For any loop $X_1 - X_2 - ... - X_k - X_1$, a chord is an edge connecting nonconsecutive X_i and X_j. An undirected graph $\mathcal{H}$ is chordal if any loop with $k \geq 4$ has a chord.*

Theorem 9. *Suppose $\mathcal{G}$ is minimal I-map for $\mathcal{H}$, then $\mathcal{G}$ can have no immoralities, and consequently $\mathcal{G}$ is chordal. (In inversed conversion, we have also eliminated immoralities)*

The conclusion is that only chordal graphs can represent independencies without loss.

3 Causal Model

Causal chains typically have multiple latent variables that aren't observed in reality. In purely probabilistic graphical models, the correlation contributed by latent variables can always be represented by simply calculating the joint distribution or conditional distribution, despite leading to more complicated models. However, in causal models, causal relationships and correlations due to confounding factors are significantly different.

An important type of latent variables is **selection bias**. For example, when trying to model the causality between athletic activity participation and high GPA, one might easily find a negative correlation. However, the reason might be the student did well in neither sports nor grades tend not to respond the survey.

Definition 20. (**Exogenous Variables**) *The set of variables that are determined outside the model and act as the driving force of the model.*

Definition 21. (**Endogenous Variables**) *The set of variables that are influenced by other variables inside the model.*

Definition 22. (**Cause / Direct Cause**) *X is a cause of Y if X is a direct cause of Y or the cause of any cause of Y.*

A **structural causal model** (SCM) has two sets of variables, exogeous variables U, endogenous variables V and a set of stochastic functions f used to determine the values of endogenous variables based on other variables (when variables in U are determined, f gives all distributions in V). Each SCM is associated with a **graphical causal model** represented by a Bayesian network structure $G = (\mathcal{V}, \mathcal{E})$ where $(X, Y) \in \mathcal{E}$ represents direct causality. Consequently, in a graphical causal model G, $\mathbf{Pr}\left[x_1, ..., x_n\right] = \prod_i \mathbf{Pr}\left[x_i|pa_i\right]$ where $X_1, ..., X_n$ are variable nodes.

Usually the goal of statistical studies are to study the effect of **interventions** (through random controlled experiments), but such experiments aren't always practical. Intervening on a variable $\mathbf{Pr}\left[Y = y|do(X = x)\right]$ and conditioning on a variable $\mathbf{Pr}\left[Y = y|X = x\right]$ are different in that intervening changes the system, through which other variables are changed, however in conditioning, we are simply narrowing the focus to a subset of the system. Specifically, intervention changes the graphical causal model while conditioning doesn't. Mathematically, intervening a set of variables $\boldsymbol{S} = \boldsymbol{s}$ are expressed as

$$\mathbf{Pr}_m\left[X_i|\boldsymbol{pa}_i\right] = \begin{cases} \mathbf{Pr}\left[X_i|\boldsymbol{pa}_i\right], & X_i \notin \boldsymbol{S} \\ 1, & X_i \in \boldsymbol{S} \wedge x_i = s_i \\ 0, & X_i \in \boldsymbol{S} \wedge x_i \neq s_i \end{cases}$$

Considering the existence of latent variables, in most cases we can obtain only marginal distribution over observable variables. Consequently, the problem lies in finding out what intervention queries are **identifiable** using only conditional probabilities from observed variables.

3.1 Identifying Causality

Definition 23. (**Augmented Causal Model**) *In $\mathcal{G}$ we define for every variable Z the* ***decision variable*** $\hat{Z} \in \{\epsilon\} \cup Val(Z)$ *determining whether we intervene Z and the intervention value. Then resulting augmented graph is $\mathcal{G}^\dagger$. $\mathcal{G}^\dagger_{\hat{\boldsymbol{Z}}}$ is the graph obtained from $\mathcal{G}^\dagger$ with every $Z \in \boldsymbol{Z}$ having only parent $\hat{Z}$.*

Theorem 10. ***(Intervention Query Simplification Rule 1)*** *For a causal model over $\mathcal{G}$, $\boldsymbol{Pr}\left[\boldsymbol{Y}|do(\boldsymbol{Z}),\boldsymbol{X},\boldsymbol{W}\right] = \boldsymbol{Pr}\left[\boldsymbol{Y}|do(\boldsymbol{Z}),\boldsymbol{X}\right]$ if d-sep$_{\mathcal{G}^\dagger_{\hat{\boldsymbol{Z}}}}(\boldsymbol{W};\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{X})$.*

Proof. Trivial. $\mathbf{Pr}_{\mathcal{G}}\left[\boldsymbol{Y}|do(\boldsymbol{Z}),\boldsymbol{X}\right] = \mathbf{Pr}_{\mathcal{G}^\dagger_{\hat{\boldsymbol{Z}}}}\left[\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{X}\right]$. □

Theorem 11. ***(Intervention Query Simplification Rule 2)*** *For a causal model over $\mathcal{G}$, $\boldsymbol{Pr}\left[\boldsymbol{Y}|do(\boldsymbol{Z}),do(\boldsymbol{X}),\boldsymbol{W}\right] = \boldsymbol{Pr}\left[\boldsymbol{Y}|do(\boldsymbol{Z}),\boldsymbol{X},\boldsymbol{W}\right]$ if d-sep$_{\mathcal{G}^\dagger_{\hat{\boldsymbol{Z}}}}(\hat{\boldsymbol{X}};\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{X},\boldsymbol{W})$.*

Proof. Notice that the d-separation condition doesn't hold when there exists causal chain $\boldsymbol{Y} \rightsquigarrow \boldsymbol{X}$. D-separation ⇒ all trails $p : \hat{X} \to X \leftrightsquigarrow Y$ are blocked by X. Therefore, $p : \hat{X} \to X \to \ldots \to U \leftarrow V \leftrightsquigarrow Y$. Since $\mathbf{Pr}\left[V|X\right] = \frac{1}{\mathbf{Pr}[X]}\sum_u \mathbf{Pr}\left[V\right]\mathbf{Pr}\left[X\right]\mathbf{Pr}\left[U|V,X\right] = \mathbf{Pr}\left[V\right]$, all marginal distributions of $\boldsymbol{Y}$'s ancestors (when $\boldsymbol{X}$ are observed) are unchanged compared to intervening $\boldsymbol{X}$ (where those marginal distributions are naturally unchanged because none of those nodes are descendants of $\boldsymbol{X}$). □

Theorem 12. ***(Intervention Query Simplification Rule 3)*** *For a causal model over $\mathcal{G}$, $\boldsymbol{Pr}\left[\boldsymbol{Y}|do(\boldsymbol{Z}),do(\boldsymbol{X}),\boldsymbol{W}\right] = \boldsymbol{Pr}\left[\boldsymbol{Y}|do(\boldsymbol{Z}),\boldsymbol{W}\right]$ if d-sep$_{\mathcal{G}^\dagger_{\hat{\boldsymbol{Z}}}}(\hat{\boldsymbol{X}};\boldsymbol{Y}|\boldsymbol{Z},\boldsymbol{W})$.*

Proof. Similarly, all trails connecting $\hat{X}$ and $\boldsymbol{Y}$ are blocked by $\varnothing$. That is, $p : \hat{X}\ldots \to U \leftarrow \ldots Y$ (here U can be X). The same reasoning shows that changing causal mechanism of $\boldsymbol{X}$ changes nothing about $\boldsymbol{Y}$. □

Notably, despite being complete (and thus stated as three axioms in **do-calculus**), these three rules have limited applicability, while the more general rules apply under **back-door criterion** or **front-door criterion**.

Definition 24. (**The Back-Door Criterion**) *Given $\boldsymbol{X},\boldsymbol{Y}$ in $\mathcal{G}$, $\boldsymbol{Z}$ satisfies back-door criterion relative to $\boldsymbol{X},\boldsymbol{Y}$ if no node in $\boldsymbol{Z}$ is a descendant of $\boldsymbol{X}$ in $\mathcal{G}$, and $\boldsymbol{Z}$ blocks every trail between X and $\boldsymbol{Y}$ that points to $\boldsymbol{X}$ (these trails are called* ***back-door paths****). (Intuitively, $\boldsymbol{Z}$ blocks all spurious paths (those with arrows into $\boldsymbol{X}$) and creates no new spurious paths (by not including descendants of $\boldsymbol{X}$).)*

Theorem 13. *If $\boldsymbol{Z}$ satisfies back-door criterion relative to $\boldsymbol{X},\boldsymbol{Y}$, then*

$$\boldsymbol{Pr}\left[\boldsymbol{Y}=\boldsymbol{y}|do(\boldsymbol{X}=\boldsymbol{x})\right] = \sum_{z}\boldsymbol{Pr}\left[\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}=z\right]\boldsymbol{Pr}\left[\boldsymbol{Z}=z\right]$$

Proof.

$$\mathbf{Pr}_{\mathcal{G}}\left[\boldsymbol{Y}|do(\boldsymbol{X})\right] = \mathbf{Pr}_{\mathcal{G}^\dagger_{\hat{\boldsymbol{X}}}}\left[\boldsymbol{Y}|\hat{\boldsymbol{X}}=do(\boldsymbol{x})\right] = \sum_{z}\mathbf{Pr}_{\mathcal{G}^\dagger_{\hat{\boldsymbol{X}}}}\left[\boldsymbol{Y}|\hat{\boldsymbol{X}}=do(\boldsymbol{x}),\boldsymbol{Z}=z\right]\mathbf{Pr}_{\mathcal{G}^\dagger_{\hat{\boldsymbol{X}}}}\left[\boldsymbol{Z}=z|\hat{\boldsymbol{X}}=do(\boldsymbol{x})\right]$$

Since $\boldsymbol{Z}$ are not descendants of $\boldsymbol{X}$, there must exist a collider in the trail $p : \hat{X} \to X \leftrightsquigarrow Z$, or equivalently, d-sep$_{\mathcal{G}^\dagger_{\hat{\boldsymbol{X}}}}(\hat{\boldsymbol{X}};\boldsymbol{Z}|\varnothing)$. Therefore, applying the 3rd rule, $\mathbf{Pr}_{\mathcal{G}^\dagger_{\hat{\boldsymbol{X}}}}\left[\boldsymbol{Z}=z|\hat{\boldsymbol{X}}=do(\boldsymbol{x})\right] = \mathbf{Pr}_{\mathcal{G}^\dagger_{\hat{\boldsymbol{X}}}}\left[\boldsymbol{Z}=z\right] = \mathbf{Pr}_{\mathcal{G}}\left[\boldsymbol{Z}=z\right]$.

Meanwhile, $\mathbf{Pr}_{\mathcal{G}^\dagger_{\hat{\boldsymbol{X}}}}\left[\boldsymbol{Y}|\hat{\boldsymbol{X}}=do(\boldsymbol{x}),\boldsymbol{Z}\right] = \sum_{\boldsymbol{x}'}\mathbf{Pr}_{\mathcal{G}^\dagger_{\hat{\boldsymbol{X}}}}\left[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}',\hat{\boldsymbol{X}}=do(\boldsymbol{x}),\boldsymbol{Z}\right] = \mathbf{Pr}_{\mathcal{G}^\dagger_{\hat{\boldsymbol{X}}}}\left[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\hat{\boldsymbol{X}}=do(\boldsymbol{x}),\boldsymbol{Z}\right]$. Note that $\hat{\boldsymbol{X}} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}$ because all trails connecting $\hat{\boldsymbol{X}}$ and $\boldsymbol{Y}$ have either X as a collider (blocked by $\boldsymbol{Z}$) or blocked by $\boldsymbol{X}$ itself. Consequently,

$$\mathbf{Pr}_{\mathcal{G}}\left[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})\right] = \sum_{z}\mathbf{Pr}_{\mathcal{G}^\dagger_{\hat{\boldsymbol{X}}}}\left[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\hat{\boldsymbol{X}}=\text{idle},\boldsymbol{Z}\right]\mathbf{Pr}_{\mathcal{G}}\left[\boldsymbol{Z}\right] = \sum_{z}\mathbf{Pr}_{\mathcal{G}}\left[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{Z}\right]\mathbf{Pr}_{\mathcal{G}}\left[\boldsymbol{Z}\right]$$

□

Back-door adjustment are not applicable for $X \leftarrow U \rightarrow Y$ and $X \rightarrow Y$ because the backdoor path $X \rightarrow U \rightarrow Y$ can't be blocked by anything. However, if we introduce another observed variable Z intercepting the edge $X \rightarrow Y$, we can measure the causality $\mathbf{Pr}\left[Y|do(X=x)\right]$.

Definition 25. (**The Front-Door Criterion**) *$\boldsymbol{Z}$ satisfies the front-door criterion relative to $\boldsymbol{X}, \boldsymbol{Y}$ if **(1)** $\boldsymbol{Z}$ intercepts all paths $p : \boldsymbol{X} \rightsquigarrow \boldsymbol{Y}$; **(2)** there is no unblocked back-door path from $\boldsymbol{X}$ to $\mathbf{Z}$; **(3)** all back-door paths from $\mathbf{Z}$ to $\boldsymbol{Y}$ are blocked by $\boldsymbol{X}$.*

Theorem 14. *If $\mathbf{Z}$ satisfies the front-door criterion relative to $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Pr}\left[\boldsymbol{x}, z\right] > 0$, then*

$$\boldsymbol{Pr}\left[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})\right] = \sum_{z} \boldsymbol{Pr}\left[\mathbf{Z} = z|\boldsymbol{X} = \boldsymbol{x}\right] \sum_{\boldsymbol{x}'} \boldsymbol{Pr}\left[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}', \mathbf{Z} = z\right] \boldsymbol{Pr}\left[\boldsymbol{X} = \boldsymbol{x}'\right]$$

Proof.

$$\begin{aligned}\mathbf{Pr}_{\mathcal{G}}\left[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})\right] &= \sum_{z} \mathbf{Pr}_{\mathcal{G}^{\dagger}_{\bar{\boldsymbol{X}}}}\left[\boldsymbol{Y}|\mathbf{Z} = z\right] \mathbf{Pr}_{\mathcal{G}^{\dagger}_{\bar{\boldsymbol{X}}}}\left[\mathbf{Z} = z|\boldsymbol{X} = \boldsymbol{x}\right] = \sum_{z} \mathbf{Pr}_{\mathcal{G}^{\dagger}_{\bar{\boldsymbol{Z}}}}\left[\boldsymbol{Y}|\mathbf{Z} = z\right] \mathbf{Pr}_{\mathcal{G}}\left[\mathbf{Z} = z|\boldsymbol{X} = \boldsymbol{x}\right] \\ &= \sum_{z} \mathbf{Pr}\left[\boldsymbol{Y}|do(\mathbf{Z} = z)\right] \mathbf{Pr}\left[\mathbf{Z} = z|\boldsymbol{X} = \boldsymbol{x}\right] \\ &= \sum_{z} \mathbf{Pr}\left[\mathbf{Z} = z|\boldsymbol{X} = \boldsymbol{x}\right] \left(\sum_{\boldsymbol{x}'} \mathbf{Pr}\left[\boldsymbol{Y}|\mathbf{Z} = z, \boldsymbol{X} = \boldsymbol{x}'\right] \mathbf{Pr}\left[\boldsymbol{X} = \boldsymbol{x}'\right]\right)\end{aligned}$$

Here, the first equality requires **(1)** to omit $\boldsymbol{X} = \boldsymbol{x}$ as a condition in the first term. The second equality requires $\mathbf{Pr}_{\mathcal{G}^{\dagger}_{\bar{\boldsymbol{X}}}}\left[\boldsymbol{Y}|\mathbf{Z}\right] = \mathbf{Pr}_{\mathcal{G}^{\dagger}_{\bar{\boldsymbol{Z}}}}\left[\boldsymbol{Y}|\mathbf{Z}\right]$ and $\hat{\boldsymbol{X}} \perp\!\!\!\perp \mathbf{Z}|\boldsymbol{X}$, which are satisfied by **(2)** and **(3)** respectively. □