MAC6916 PROBABILISTIC GRAPHICAL MODELS
LECTURE 3: REPRESENTING INDEPENDENCIES

DENIS D. MAUÁ

1. Motivation

Bayesian networks can represent a probability function compactly by exploiting a set of Markov conditions in a digraph. This set of Markov conditions imply (and are implied) by other independences. In this lecture we will see how graphs can be used to derive a sound and nearly complete representation system for independencies. The key to achieving this feature is the concept of d-separation. In fact, a major contribution of the theory of Bayesian networks was to bring forth a representational system that associates (in)dependencies with graph (dis)connectedness.

2. Semi-Graphoids

Independency satisfies a number of useful properties:

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z} \Rightarrow \mathcal{Y} \perp\!\!\!\perp \mathcal{X} \mid \mathcal{Z}\,, \qquad \text{[symmetry]}$$
$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \cup \mathcal{W} \mid \mathcal{Z} \Rightarrow \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z}\,, \qquad \text{[decomposition]}$$
$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \cup \mathcal{W} \mid \mathcal{Z} \Rightarrow \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z} \cup \mathcal{W}\,, \qquad \text{[weak union]}$$
$$\mathcal{X} \perp\!\!\!\perp \mathcal{W} \mid \mathcal{Y} \cup \mathcal{Z} \text{ and } \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z} \Rightarrow \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \cup \mathcal{W} \mid \mathcal{Z}\,. \qquad \text{[contraction]}$$

The above conditions hold for any probability function and are called *semi-graphoid* axioms (for technical reasons we assume that any variable is conditionally independent of the empty set). When the probability function is positive, a third also holds:

semi-graphoid

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{Z} \cup \mathcal{W} \text{ and } \mathcal{X} \perp\!\!\!\perp \mathcal{W} \mid \mathcal{Z} \cup \mathcal{Y} \Rightarrow \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \cup \mathcal{W} \mid \mathcal{Z}\,. \qquad \text{[intersection]}$$

A semi-graphoid satisfying intersection is a *graphoid*.

graphoid

Decomposition is what allowed us to derive the factorization property of Bayesian networks from the (local) Markov property by using the *ordered Markov property*

ordered Markov property

$$X \perp\!\!\!\perp \{Y : Y < X\} | \text{pa}(X) \quad \text{for any topological ordering } < .$$

Decomposition also allows us to derive the *pairwise Markov property*:

pairwise Markov property

$$X \perp\!\!\!\perp Y | \text{pa}(X) \quad \text{for all } Y \in \text{nd}(X) - \text{pa}(X)\,.$$

The opposite direction (i.e., composition) however does not hold in general. Thus, the the pairwise Markov property does not suffice to establish the local Markov

Date: March 8, 2016.

property (unless the distribution is positive, then intersection allows us to prove equivalence). The weak union property allows us to derive the following fact:

$$X \perp\!\!\!\perp \mathrm{nd}(X) - \mathcal{S} \mid \mathrm{pa}(X) \cup \mathcal{S}$$

for any $\mathcal{S} \subseteq \mathrm{nd}(X)$. Using this fact we can show that the ordered Markov property implies the local Markov property (for each variable select a topological ordering where X is greater than all its nondescendants).

Pearl wrongly conjectured that the semi-graphoid formed a complete system for the independency relation in the sense that any other property can be derived from them. However, Studený later proved that there is no finite characterization of probabilistic independency, which implies that semi-graphoids cannot be used as a axiomatization of the independency relation. Yet, semi-graphoid properties remained as set of useful properties for understanding the concept of d-separation.

3. A Few Concepts from Graph Theory

Fix a directed graph $G = (\mathcal{V}, \mathcal{E})$. We will represent an arc $(X, Y) \in A$ as $X \to Y$. A *trail* is a sequence of nodes $X_1, \ldots, X_k$ such that for $i = 1, \ldots, k-1$ either $X_i \to X_{i+1}$ or $X_{i+1} \to X_i$, and each arc appears at most once. That is, a trail is a way of going from a node to another by following arcs in any direction (without passing twice in the same arc). The *length of a trail* is the number of nodes it contains. A *directed path* is a trail which follows the direction of the arcs (i.e., X_i, X_{i+1} is in the path only if $X_i \to X_{i+1}$). Two nodes are *connected* if there is a trail starting at one of the nodes and ending at the other one, otherwise they are *separated*. A (directed) cycle is a path starting and ending at the same node, e.g. $A \to B \to C \to A$.

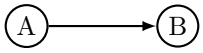
trail

directed path

4. D-Separation

Before formally defining d-separation, let us motivate its definition by discussing some simple scenarios that a (directed) graphical representation of (in)dependence should encode. Recall that our goal is to represent (in)dependences using an acyclic directed graph. We say that X depends on Y if X and Y are probabilistic dependent. We associate an arc with a *dependence relation*, so that

dependence relation

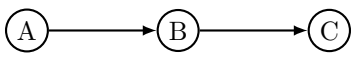


represents the relation $A \not\perp\!\!\!\perp B$. A graph interpreted under this semantics is called a *dependency graph*. Since (in)dependency is symmetric, a (direct) dependence can be locally represented in any direction (however other dependencies may constrain the direction).

dependency graph

The first scenario is *chain reasoning*, when C depends on A through B, meaning that C depends on B ($C \not\perp\!\!\!\perp B$), B depends on A ($B \not\perp\!\!\!\perp A$), C is independent of A ($C \perp\!\!\!\perp A \mid B$) given B, and C (unconditionally) depends on A ($C \not\perp\!\!\!\perp A$). B is called an intermediary or mediating variable of the phenomenon. This can be graphically captured by a *serial connection*:

chain reasoning



We say that the trail from A to C in the serial connection is *unblocked* when B is not given and *blocked* otherwise, and that B blocks the trail from A to C. All the (in)dependencies states can be readily read off the graph by equation unblocked

unblocked trail

blocked trail

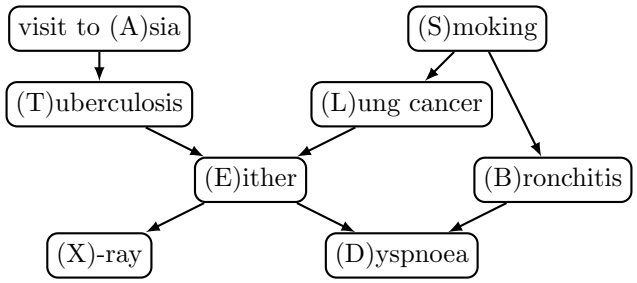
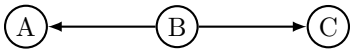


FIGURE 1. The structure (graph) of the Asia Bayesian network.

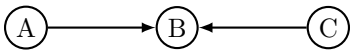
connection (resp., blocked connection) and dependence (resp., independence): A and C are connected (dependent), and A and C are separated (independent) given B. Inducing on this connection (i.e., serial connections with more than 3 variables) shows that any variable is unconditionally dependent on its ancestors (and by symmetry on its descendants). So for example in the Asia network (Fig. 1) E depends on A, S, X and D (and obviously on T and L).

Another interesting scenario is when A is a common cause of unrelated events B and C (so that B and C are conditionally independent given A). Clearly, B depends on A ($B \not\perp\!\!\!\perp A$), and C depends on A ($C \not\perp\!\!\!\perp A$). Since A is a common cause, B and C must be unconditionally dependent on each other ($B \not\perp\!\!\!\perp C$): observing B increases the chances that A occurred, which in turn increases the chances that C occurred. Graphically, this can be captured by a *divergent connection*: divergent connection



The trail between A and C is blocked given B and unblocked otherwise. A consequence of this connection is that any variable depends on its non-descendants with a common ancestor. For example, E depends on B in the Asia network.

The cases discussed so far suggest that dependence is better represented without directions (i.e., by an undirected graph). However, directing the arcs plays a fundamental role in representing a common scenario known as *explaining away*, which occurs when A and C are unrelated causes of a common effect B. Then A and C are unconditionally independent ($A \perp\!\!\!\perp C$) but conditionally dependent given B ($A \not\perp\!\!\!\perp C \mid B$). To convince yourself why the causes must be conditionally dependent, consider the case when B is the exclusive-or of A and C and we observe $B = 1$. Since either $A = 0$ or $C = 0$ any increase in the probability of $A = 0$ must be followed by a decrease in the probability of $B = 0$. We can represent this type of reasoning as a *convergent connection*: convergent connection



The trail from A to C (and viceversa) is unblocked when B is given and blocked otherwise. Note that this is the opposite of the previous cases (e.g., divergent connection). This case can only be distinguished by considering the direction of arcs. In the Asia network, we have that A and S are unconditionally independent and conditionally dependent given E.

We now formalize the definition of (un)blocked trails.

Definition 1. We say that a *trail* from X to Y is *blocked* by a set of variables $\mathcal{B}$ if blocked trail

(1) X or Y are in $\mathcal{B}$, or
(2) there is a serial or divergent connection A, B, C blocked by some $B \in \mathcal{B}$, or
(3) there is a convergent connection $A \to B \leftarrow C$ and $(\{B\} \cup \text{nd}(B)) \not\subseteq \mathcal{B}$ (i.e., neither B nor any of its descendants are in $\mathcal{B}$).

unblocked trail

Otherwise the trail is *unblocked*.

Note that a convergent connection is unblocked even by nodes outside the trail. In the Asia network, the trail A, T, E, L, S by $\mathcal{B} = \{L\}$ and unblocked by $\mathcal{B} = \{D\}$.

d-separation

Definition 2. Two sets of variable $\mathcal{X}$ and $\mathcal{Y}$ are *d-separated* by a set of variables $\mathcal{Z}$ if every trail from a variable $X \in \mathcal{X}$ to a variable $Y \in \mathcal{Y}$ is blocked by $\mathcal{Z}$. Otherwise, $\mathcal{X}$ and $\mathcal{Y}$ are *d-connected*.

d-connection

We denote that $\mathcal{X}$ and $\mathcal{Y}$ are d-separated by $\mathcal{Z}$ as

$$\mathcal{X} \perp_d \mathcal{Y} \mid \mathcal{Z}\,.$$

Proposition 1. *Any node X is d-separated from its nondescedants by its parents.*

Proof. Consider a trail from a from a nondescendant nonparent node Y to X. If the trail contains a parent $Z \in \text{pa}(X)$ then it either contains a serial connection $W \to Z \to X$ (when Y is an ancestor of X), which is blocked by $\text{pa}(X)$ or a divergent connection $W \leftarrow Z \to X$ (when Y is not an ancestor of X), which is also blocked by $\text{pa}(X)$. So consider there is no parent of X in the trail. In this case the trail must contain a descendant $Z \in \text{de}(X)$ which is also an descendant of Y. This shows that the trail contains a convergent connection $W_1 \to Z \leftarrow W_2$, where $W_1 \text{de}(X)$ and $W_2 \in \text{de}(Y)$. Since this connection is unblocked by $\text{pa}(X)$, this completes the proof. □

Thus, d-separation is consistent with the local Markov properties. Ideally, we would like to equate d-separation (resp., d-connection) with independency (resp., dependency). That is, we would like to have

$$\mathcal{X} \perp_d \mathcal{Y} \mid \mathcal{X} \Rightarrow \mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{X}\,, \tag{1}$$ [soundness]

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} \mid \mathcal{X} \Rightarrow \mathcal{X} \perp_d \mathcal{Y} \mid \mathcal{X}\,. \tag{2}$$ [completeness]

The first desiderata indeed holds:

Theorem 1. *D-separation is sound.*

global Markov property

The soundness condition of d-separation is known as the *global Markov property*. By Proposition 1, the global Markov property implies the local Markov property. And the theorem above shows that the converse also holds (so local and global Markov properties coincide).

The second desiderata however is not true for all distributions. In fact, the incompleteness of d-separation is a necessary requirement. To prove that d-separation is incomplete for Bayesian networks consider the following simple model with graph $A \to B$, $p(A) = 1/2$ and $p(B|A)$ given below.

	B=0	B=1
A=0	0.4	0.6
A=1	0.4	0.6

Clearly, A and B are independent w.r.t. $p(A, B) = p(A)p(B|A)$. However, the graph $A \to B$ implies that A and B are d-connected. This is generally true:

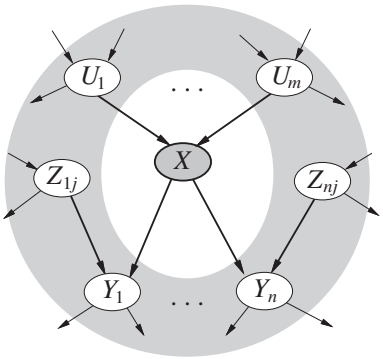


FIGURE 2. The Markov blanket of a variable.

Theorem 2. *For any graph G there is a Bayesian network with graph G and variables X and Y such that $X \perp\!\!\!\perp Y$ and $X \not\perp_d Y$.*

The proof is based on the example given. Let $Y \in \mathrm{pa}(X)$, and specify p such that $p(X|\mathrm{pa}(X)) = p(X|\mathrm{pa}(X) - Y)$. This distribution satisfies all the Markov properties in G, and has $X \perp\!\!\!\perp Y$ (but $X \not\perp_d Y$). This shows that d-separation cannot be complete in general. The following result shows that d-separation is "as complete as possible".

Theorem 3. *If G is a DAG and X and Y are d-connected by $\mathcal{Z}$ in G, then there is a Bayesian network $(G, \mathbb{P})$ such that X and Y are* dependent *given $\mathcal{Z}$.*

Proof sketch. Since X and Y are d-connected by $\mathcal{Z}$, there is an unblocked trail $X, X_1, \ldots, X_m, Y$ given $\mathcal{Z}$ in G. We construct a joint distribution by specifying the conditional distributions (local models) of every variable along this trail so as to ensure that consecutive variables are dependent. For the convergent connections $A \to B \leftarrow C$ in the trail, we also set the distributions of the descendants of B so as to make each dependent on its parents that are descendant of a variable in the trail. We finally set the distributions of the remaining variable as uniforms (this prevents canceling of the dependencies specified thus far). □

The above result shows that for every graph there is a distribution that makes the d-separation complete for the corresponding Bayesian network. Hence, any other sound and complete system would have to coincide with d-separation in these networks; in particular, any system that is based exclusively on the graph structure would coincide with d-separation for all distributions lest it be incomplete.

Definition 3. The *Markov blanket* of a variable X is a set $\mathcal{B}$ such that Markov blanket

$$X \perp\!\!\!\perp \mathcal{V} - \mathcal{B} - \{X\} | \mathcal{B}\,.$$

Proposition 2. *The set formed by the parents, the children and the parents of the children of a variable X is a markov blanket of X. If the distribution is positive it is* minimal. minimal Markov blanket

4.1. **D-separation and semi-graphoids.** We have seen that independency satisfies a number of basic properties known as semi-graphoid axioms. Moreover, independency does not generally satisfy composition and intersection. It can be shown

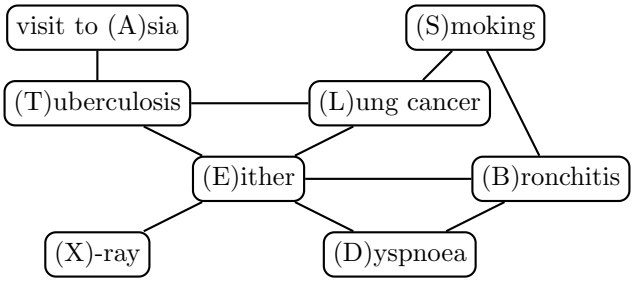


FIGURE 3. Moral graph of the Asia Bayesian network.

that d-separation satisfies all the semi-graphoid axioms. However, d-separation satisfies composition and intersection:

$$\mathcal{X} \perp_d \mathcal{Y} \mid \mathcal{X} \text{ and } \mathcal{X} \perp_d \mathcal{W} \mid \mathcal{Z} \Rightarrow \mathcal{X} \perp_d \mathcal{Y} \cup \mathcal{W} \mid \mathcal{Z}\,, \qquad \text{[composition]}$$

$$X \perp_d Y \mid Z \cup W \text{ and } X \perp_d W \mid Z \cup Y \Rightarrow X \perp_d Y \cup W \mid Z \cup W\,. \qquad \text{[intersection]}$$

Since not all distributions satisfy composition and intersection, there are distributions whose set of independencies cannot be fully characterized by a Bayesian network.

5. M-separation

Another graphical criteria for verifying independence is based on undirected graphs.

moral graph

Definition 4. The *moral graph* of an acyclic directed graph G is the undirected graph obtained by connecting nodes with a common child and then discarding arc directions.

Let $M[G]$ denote the moral graph of a fixed a acyclic directed graph G. Denote by $G[\mathcal{X}] = G - (\mathcal{V} - \mathcal{X})$ the subgraph of M obtained by deleting all nodes outside $\mathcal{X}$ and their corresponding edges. Also, denote by $\overline{\text{an}(\mathcal{X})}$ the set containing $\mathcal{X}$ and each ancestor of a node in $\mathcal{X}$.

m-separation

m-connectedness

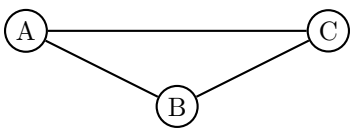
Definition 5. We say that the sets of variables $\mathcal{X}$ and $\mathcal{Y}$ are *m-separated* by the set of variables $\mathcal{Z}$, and write $\mathcal{X} \perp_m \mathcal{Y} \mid \mathcal{Z}$, if they are disconnected in $M[G[\overline{\text{an}(\mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z})}]] - \mathcal{Z}$. Otherwise they are *m-connected*, denoted by $\mathcal{X} \not\perp_m \mathcal{Y} \mid \mathcal{Z}$.

M-separation and d-separation are equivalent.

Theorem 4.

$$\mathcal{X} \perp_d \mathcal{Y} \mid \mathcal{Z} \textit{ if and only if } \mathcal{X} \perp_m \mathcal{Y} \mid \mathcal{Z}\,.$$

Proof sketch. Consider the transformation of a directed graph into $M' = M[G[\overline{\text{an}(\mathcal{X} \cup \mathcal{Y}\mathcal{Z})}]] - setZ$. The serial and divergent connections are transformed into regular 3-long trails, and are blocked if and only if the intermediary node is in $\mathcal{Z}$. Thus, d-separation and m-separation coincide. Consider a convergent connection $A \to B \leftarrow C$. This is transformed into a clique



Thus, the variables A and C are connected in the moral graph. In principle, this could introduce a dependence if neither B nor any of its descendants are in $\mathcal{Z}$. However, this connection is only present if there B is in $\mathcal{Z}$. □

6. Exercises

(4.1) Prove the first part of Proposition 2 (that the set of parents, children and spouses form a Markov Blanket of a node).
(4.2) Solve exercises 4.1, 4.2 and 4.3 in [Darwiche 2009].

7. Reading

- Bayes Ball paper.

8. Assignment

Write code to decide d-separation of sets of nodes. Use it to speed up your current (enumerative) inference algorithm. Hint: implement the Bayes Ball algorithm. An alternative is the algorithm described in section 3.3.3 of [Koller & Friedman 2009].