COMP538: Introduction to Bayesian Networks

Lecture 3: Probabilistic Independence and Graph Separation

Nevin L. Zhang
lzhang@cse.ust.hk

Department of Computer Science and Engineering
Hong Kong University of Science and Technology

Fall 2008

Objective

- Objective:
  - Discusses the relationship between probabilistic independence and graph separation in Bayesian networks.
    - Given a BN structure, a DAG, what independence relationships are represented?
    - Given a joint distribution, under what conditions can the independence relationships it entails be represented using a DAG? How much?

- Reading: Zhang & Guo: Chapter 3
- Reference: Jensen (2001), Cowell *et al.* (1999), Chapter 5.

Outline

Intuitive Meaning of Independence

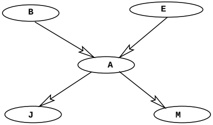
- Given: A Bayesian network and two variables X and Y.
- Question:
  - Are X and Y independent?
  - What are the (graph-theoretic) conditions under which X and Y are independent?
- We will try to answer this question based on intuition.
- This exercise will leads to the concept of d-separation.

- Intuitive meaning of independence:
  - X and Y are dependent under some condition C iff knowledge about one influences belief about the other under C.
  - X and Y are independent under some condition C iff knowledge about one does not influence belief about the other under C.

Case 1: Direction connection

- If X and Y are connected by an edge, then X and Y are dependent (under the empty condition).
- **Information can be transmitted over one edge.**

Example:



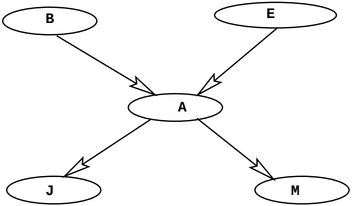
- Burglary and Alarm are dependent:
  - My knowing that a burglary has taken place increases my belief that the alarm went off.
  - My knowing that the alarm went off increases my belief that there has been a burglary.

Case 2: Serial connection

- If Z is not observed, X and Y are dependent.
- **Information can be transmitted between X and Y through Z if Z is not observed.**

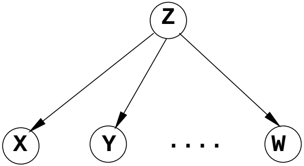
- If Z is observed, X and Y are independent.
- **Information cannot be transmitted between X and Y through Z if Z is observed. Observing Z blocks the information path.**

Case 2: Serial connection/Example



- If A not observed, B and M call are dependent:
  - My knowing that a burglary has taken place increases my belief on Marry call.
  - My knowing that Marry called increases my belief on burglary.
- If A is observed, B and M are conditionally independent :
  - If I already know that the alarm went off,
    - My further knowing that a burglary has taken place would not increases my belief on Marry call.
    - My further knowing that Marry called would not increases my belief on burglary.

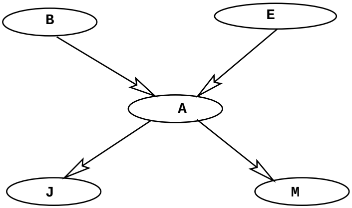
Case 3: Diverging connection (common cause)



- If Z is not observed, X and Y are dependent.
- **Information can be transmitted through Z among children of Z if Z is not observed.**

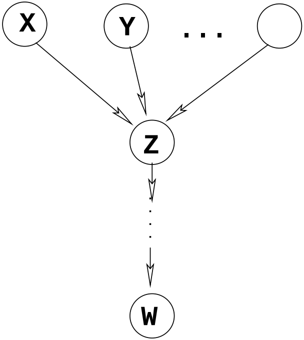
- If Z is observed, X and Y are independent.
- **Information cannot be transmitted through Z among children of Z if Z is observed. Observing Z blocks the information path.**

Case 3: Diverging connection/Example



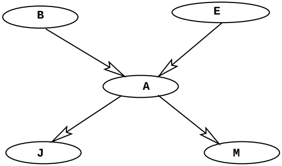
- If A is not observed, J and M are dependent:
  - My knowing that John called increases my belief on Marry call.
  - My knowing that Marry called increases my belief on John call.
- If A is observed, J and M are conditionally independent:
  - If I already know that the alarm went off,
    - My further knowing that John called would not increase my belief on Marry call.
    - My further knowing that Marry called would not increase my belief on John call.

Case 4: Converging connection (common effect)



- If neither Z nor any of its descendant are observed, X and Y are independent.
- **Information cannot be transmitted through Z among parents of Z.**
  It leaks down Z and its descendants.
- If Z or any of its descendant is observed, X and Y are dependent.
- **Information can be transmitted through Z among parents of Z if Z or any of its descendants are observed. Observing Z or its descendants opens the information path.**

Case 4: Converging connection/Example



- E and B are conditionally dependent if A is observd:
  - If I already know that the alarm went off,
    - My further knowing that there has been a earthquake decreases my belief on Burglary.
    - My further knowing that there has been a burglary decreases my belief on earthquake.

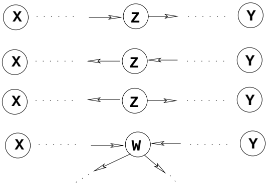
  Explaining away.
- E and B are conditionally dependent if M is observed:
  - Observing Marry call gives us some information about Alarm. So we are back to the previous case.
- E and B are marginally independent (if A, M and J not observed).

Hard Evidence and Soft Evidence

- **Hard evidence** on a variable: The value of the variable is directly observed.
- **Soft evidence** on a variable: The value of the variable is NOT directly observed. However the value of a descendant is observed.

- The rules restated:
  - Hard evidence blocks information path in the case of serial and diverging connection
  - Both hard and soft evidence are enough for opening of information path in the case of converging connection.

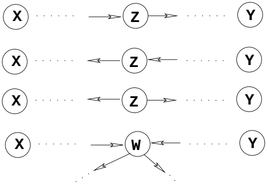
Blocked Paths



A path between X and Y is **blocked** by a set $\mathbf{Z}$ of nodes if

1. Either that path contains a node Z that is in $\mathbf{Z}$ and the connection at Z is either serial or diverging.
2. Or that the path contains a node W such that W and its descendants are not in $\mathbf{Z}$ and the connection at W is a converging connection.

Blocked Paths



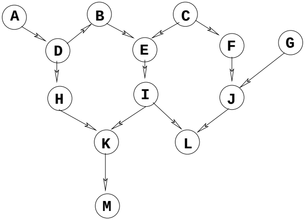
- Suppose all variables in **Z** are the observed variables.
- Then a path between X and Y being blocked by **Z** implies:
  1. Either information cannot be transmitted through Z because observing Z blocks that path.
  2. Or information cannot be transmitted through W, it leaks through W.

  In both cases, information cannot be transmitted between X and Y along the path.
- If path is not blocked, on the other hand, information CAN flow between X and Y.

D-separation

- Two nodes X and Y are **d-separated** by a set $\mathbf{Z}$ if
  - All paths between X and Y are blocked by $\mathbf{Z}$.
- Theorem 3.1:
  - If X and Y are d-separated by $\mathbf{Z}$, then $X \perp Y | \mathbf{Z}$.
- It should be pointed out that this conclusion is derived from intuition.
- One of the main tasks in this lecture is to rigorously show that the conclusion is indeed true.

Examples

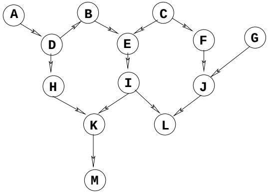


- A d-separated (by empty set) from C, F, G, J
- A d-separated by $\{M, B\}$ from G
- A d-separated by $\{E, K, L\}$ from M
- I d-separated by $\{E, K, L\}$ from M

Exercise: Try more examples on your own.

Outline

Ancestral Sets



- Let **X** be a set of nodes in a Bayesian network.
- The **ancestral set** $an(\mathbf{X})$ of **X** consists of
  - All nodes in **X** and all the ancestors of nodes in **X**.
- Example: The ancestral set of $\{I, G\}$ consists of

$$\{I, G, A, B, C, D, E\}$$

- We say that **X** is **ancestral** if

$$\mathbf{X} = an(\mathbf{X})$$

- A **leaf node** is one without children. Examples: M, L

A Lemma

Lemma (3.1)

Suppose $\mathcal{N}$ is a Bayesian network, and Y is a leaf node. Let $\mathcal{N}'$ be the Bayesian network obtained from $\mathcal{N}$ by removing Y. Let $\mathbf{X}$ be the set of all nodes in $\mathcal{N}'$. Then

$$P_{\mathcal{N}}(\mathbf{X}) = P_{\mathcal{N}'}(\mathbf{X}).$$

Proof

$$\begin{aligned}
P_{\mathcal{N}}(\mathbf{X}) &= \sum_{Y} P_{\mathcal{N}}(\mathbf{X}, Y) \\
&= \sum_{Y} [\prod_{W \in \mathbf{X}} P(W|pa(W))] P(Y|pa(Y)) \\
&= \prod_{W \in \mathbf{X}} P(W|pa(W)) \sum_{Y} P(Y|pa(Y)) \\
&= \prod_{W \in \mathbf{X}} P(W|pa(W)) \\
&= P_{\mathcal{N}'}(\mathbf{X})
\end{aligned}$$

A Lemma

- The third equality is true because, being a leaf node, Y is not in $\mathbf{X}$ and cannot be in any $pa(W)$ for any $W \in \mathbf{X}$.
- The fourth equality is true because probability sum to one. Q.E.D

First Proposition

Proposition (3.1)

Let $\mathbf{X}$ *be a set of nodes in a Bayesian network* $\mathcal{N}$*. Suppose* $\mathbf{X}$ *is ancestral. Let* $\mathcal{N}'$ *be the Bayesian network obtained from* $\mathcal{N}'$ *by removing all nodes outside* $\mathbf{X}$*.Then,*

$$P_{\mathcal{N}}(\mathbf{X}) = P_{\mathcal{N}'}(\mathbf{X}).$$

Proof:

- Consider the following procedure
  - While there are nodes outside $\mathbf{X}$,
    - Find a leaf node. (There must be one. Exercise.)
    - Remove it.
- Afterwards, we get $\mathcal{N}'$.
- And according to Lemma 3.1,the probability distribution of $\mathbf{X}$ remains unchanged throughout the procedure.
- The proposition is hence proved. Q.E.D.

Second Proposition

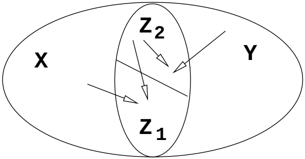
Proposition (3.2)

Let $\mathbf{X}$, $\mathbf{Y}$, *and* $\mathbf{Z}$ *be three disjoint sets of nodes in a Bayesian network such that their union is the set of all nodes.*

- *If* $\mathbf{Z}$ *d-separates* $\mathbf{X}$ *and* $\mathbf{Y}$, *then*

$$\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$$

Proof:



- Let $\mathbf{Z}_1$ be the set of nodes in $\mathbf{Z}$ that have parents in $\mathbf{X}$. And let $\mathbf{Z}_2 = \mathbf{Z} \setminus \mathbf{Z}_1$.
- Because $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$,
  - For any $W \in \mathbf{X} \cup \mathbf{Z}_1$, $pa(W) \subseteq \mathbf{X} \cup \mathbf{Z}$.
  - For any $W \in \mathbf{Y} \cup \mathbf{Z}_2$, $pa(W) \subseteq \mathbf{Y} \cup \mathbf{Z}$.

Proof Second Proposition (cont'd)

- Consider

$$
\begin{aligned}
P(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) &= \prod_{W \in \mathbf{X} \cup \mathbf{Z} \cup \mathbf{Y}} P(W|pa(W)) \\
&= [\prod_{W \in \mathbf{X} \cup \mathbf{Z}_1} P(W|pa(W))][\prod_{W \in \mathbf{Z}_2 \cup \mathbf{Y}} P(W|pa(W))]
\end{aligned}
$$

- Note that
  - $\prod_{W \in \mathbf{X} \cup \mathbf{Z}_1} P(W|pa(W))$ is a function of **X** and **Z**
  - $\prod_{W \in \mathbf{Z}_2 \cup \mathbf{Y}} P(W|pa(W))$ is a function of **Z** and **Y**.
- It follows from Proposition 1.1 (of Lecture 1) that

$$\mathbf{X} \perp \mathbf{Y} | \mathbf{Z}$$

Q.E.D

Global Markov property

Theorem (3.1)

Given a Bayesian network, let X and Y be two variables and $\mathbf{Z}$ be a set of variables that does not contain X or Y. If $\mathbf{Z}$ d-separates X and Y, then

$$X \perp Y | \mathbf{Z}$$

Proof:

- Because of Proposition 3.1, we can assume that $an(\{X, Y\} \cup \mathbf{Z})$ equals the set of all nodes.
    - $X \perp Y | \mathbf{Z}$ in original network iff it is true in the restriction onto the ancestral set.
    - $\mathbf{Z}$ d-separates X and Y in original network iff it is true in the restriction onto the ancestral set. (Exercise)

Proof of Global Markov property (cont'd)

- Let $\mathbf{X}$ be the set of all nodes that are NOT d-separated from X by $\mathbf{Z}$.
- Let $\mathbf{Y}$ be the set of all nodes that are neither in $\mathbf{X}$ or $\mathbf{Z}$.
- Because of Proposition 3.2, $\mathbf{X} \perp \mathbf{Y}|\mathbf{Z}$.
- Because of Proposition 1.1, there must exist functions $f(\mathbf{X}, \mathbf{Z})$ and $g(\mathbf{Z}, \mathbf{Y})$ such that

$$P(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) = f(\mathbf{X}, \mathbf{Z})g(\mathbf{Z}, \mathbf{Y})$$

- Note that $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$.
- Let $\mathbf{X}' = \mathbf{X} \setminus \{X\}$ and $\mathbf{Y}' = \mathbf{Y} \setminus \{Y\}$.
- We have

$$P(X, \mathbf{X}', \mathbf{Z}, Y, \mathbf{Y}') = f(X, \mathbf{X}', \mathbf{Z})g(\mathbf{Z}, Y, \mathbf{Y}')$$

Proof of Global Markov property (cont'd)

- Consequently

$$
\begin{aligned}
P(X, Y, \mathbf{Z}) &= \sum_{\mathbf{X}', \mathbf{Y}'} P(X, \mathbf{X}', \mathbf{Z}, Y, \mathbf{Y}') \\
&= \sum_{\mathbf{X}', \mathbf{Y}'} f(X, \mathbf{X}', \mathbf{Z}) g(\mathbf{Z}, Y, \mathbf{Y}') \\
&= [\sum_{\mathbf{X}'} f(X, \mathbf{X}', \mathbf{Z})][\sum_{\mathbf{Y}'} g(\mathbf{Z}, Y, \mathbf{Y}')] \\
&= f'(X, \mathbf{Z}) g'(\mathbf{Z}, Y)
\end{aligned}
$$

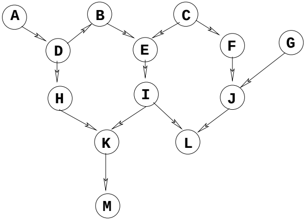
That is

$$X \perp Y | \mathbf{Z}$$

Q.E.D

Markov blanket

- In a Bayesian network, the **Markov blanket** of a node X is the set consisting of
  - Parents of X
  - Children of X
  - Parents of children of X
- Example:



The Markov blanket of I is $\{E, H, J, K, L\}$

Markov blanket

- Corollary (3.1)

In a Bayesian network, a variable X is conditionally independent of all other variables given its Markov blanket.(This is why it is so called.)

  - **Proof**:
    - Because of Theorem 3.1, it suffices to show that
      - The Markov blanket of X d-separates X from all other nodes.
    - This is true because, in any path from X to outside its Markov blanket, the connection at that last node before leaving the blanket is either serial or diverging. Q.E.D

Local Markov property

Corollary (3.2)

(Local Markov property) *In a Bayesian network, a variable X is independent of all its non-descendants given its parents.*

Proof:

- Because of Theorem 3.1, it suffices to show that
  - $pa(X)$ d-separates X from the non-descendants of X.
- Consider a path between X and a non-descendant Y. Let Z be the neighbor of X on the path.
  - Case 1: $Z \in pa(X)$,
    - The connection at Z is not converging because we have $Z \rightarrow X$.
    - Hence, path is blocked by $pa(X)$.
  - Case 2: $Z \notin pa(X)$:
    - Moving downward from Z, we can reach a converging node on the path.
    - The converging node and its descendants are not in $pa(X)$.
    - The path is blocked by $pa(X)$.

Some Notes

- The local Markov property was first mentioned in Lecture 2, when introducing the concept of BN. It is now proved.

- This also explains why we need to make the causal Markov assumption when we causality to build BN structure (slide 36 of Lecture 2):
  - If you use a causal network as a Bayesian network, then we are assuming that causality implies the local Markov property.

Outline

Representing independence using DAG

- A joint distribution $P(\mathbf{V})$ entails conditional independence relationships among variables:
  - Use $\mathbf{X} \perp_P \mathbf{Y}|\mathbf{Z}$ denotes the fact that, under P, $\mathbf{X}$ and $\mathbf{Y}$ are conditional independent given $\mathbf{Z}$, i.e.,

$$P(\mathbf{X}, \mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})P(\mathbf{Y}|\mathbf{Z}) \text{ whenever } P(\mathbf{Z}) > 0$$

- In a DAG $\mathcal{G}$, there D-separation relationships:
  - Use $S_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ denotes that the fact that $\mathbf{Z}$ d-separates $\mathbf{X}$ and $\mathbf{Y}$ in $\mathcal{G}$.

Representing independence using DAG

- $P(\mathbf{V})$ obeys the **global Markov property** according to $\mathcal{G}$ if for any three disjoint subsets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$.

$$S_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \text{ implies } \mathbf{X} \perp_P \mathbf{Y} | \mathbf{Z}$$

- When it is the case, we say that $\mathcal{G}$ **represents** some of the independence relationships entailed by P:
  - We can identify independence under P by examining $\mathcal{G}$.

- When can we use a DAG $\mathcal{G}$ to represent independence relationships entailed by a joint distribution P?

Factorization

- $P(\mathbf{V})$ **factorizes** according to $\mathcal{G}$ if there exists a Bayesian network such that
  - Its network structure us $\mathcal{G}$
  - The joint probability it represents is $P(\mathbf{V})$.

Local Makov properties

- $P(\mathbf{V})$ obeys the **local Markov property** according to $\mathcal{G}$ if for any variable X

$$X \perp_P nd_{\mathcal{G}}(X) | pa_{\mathcal{G}}(X)$$

where $nd(X)$ stands for the set of non-descendants of X.

Factorization and independence

Theorem (3.2)

Let $P(\mathbf{V})$ be a joint probability and $\mathcal{G}$ be a DAG over a set of variables $\mathbf{V}$. The following statements are equivalent:

1. *$P(\mathbf{V})$ factorizes according to $\mathcal{G}$.*
2. *$P(\mathbf{V})$ obeys the global Markov property according to $\mathcal{G}$.*
3. *$P(\mathbf{V})$ obeys the local Markov property according to $\mathcal{G}$*

Proof:

- $1 \Rightarrow 2$: Theorem 3.1.
- $2 \Rightarrow 3$: Corollary 3.2.

Proof of Theorem 3.2 (cont'd)

- $3 \Rightarrow 1$:
  - Induction on the number of nodes.
  - Trivially true where there is only one node.
  - Suppose true in the case of $n-1$ nodes.
  - Consider the case of n nodes.
    - Let X be a leaf node in $\mathcal{G}$, $\mathbf{V}' = \mathbf{V} \setminus \{X\}$.
    - By (3), X is independent of all other nodes given $pa(X)$.
    - Hence
$$P(\mathbf{V}) = P(\mathbf{V}')P(X|\mathbf{V}') = P(\mathbf{V}')P(X|pa(X)$$
    - Let $\mathcal{G}'$ be obtained from $\mathcal{G}$ by removing X.
    - Then $P(\mathbf{V}')$ obeys the local Markov property according to $\mathcal{G}'$.
    - Since there are only $n-1$ nodes in $\mathbf{V}'$, $P(\mathbf{V}')$ factorizes according to $\mathcal{G}'$.
    - Hence $P(\mathbf{V})$ factorizes according to $\mathcal{G}$.
  - The theorem is proved. Q.E.D

I-Map and D-Map

- $\mathcal{G}$ is an I-map of $P(\mathbf{V})$ if for any three disjoint subsets of variables $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$:

  $$S_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \text{ implies } \mathbf{X} \perp_P \mathbf{Y} | \mathbf{Z}$$

  i.e. d-Separation in DAG implies independence.

- $\mathcal{G}$ is an **D-map** of $P(\mathbf{V})$ if

  $$\mathbf{X} \perp_P \mathbf{Y} | \mathbf{Z} \text{ implies } S_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$$

  i.e. Independence implies separation in DAG. Non-separation implies dependence.

- $\mathcal{G}$ is an **perfect map** of $P(\mathbf{V})$ if
  - it is both an I-map and a D-map.

  This is ideal case. But there are joint distributions that do not have perfect maps. (Can you think of one?)

I-Map and D-Map

- Adding an edge in an I-map results in another I-map. (Exercise)
- Deleting an edge in a D-map results in another D-Map. (Exercise)
- A **minimal I-map** of $P(\mathbf{V})$ is an I-map such that deletion of one edge will render the graph a non-I-map.
- When constructing BN structure following the procedure given on Slide 24 of Lecture 2,
  - If $pa(X_i)$ is selected to be minimal, then resulting network is an I-map of P.