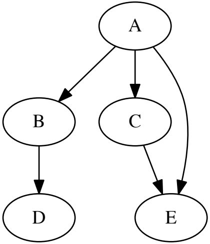
Bayesian Networks
CSC 591 Week 14

November 19, 2015

Background – Graph Theory

- Graph: consists of a set of *vertices* V and *edges* E

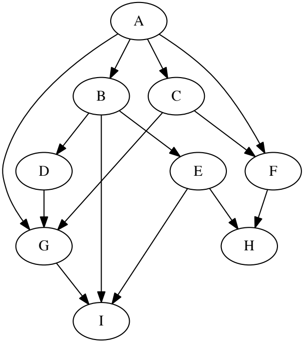


$$V = \{A, B, C, D, E\}$$

$$E = \{(A, B), (A, C), (A, E), (B, D), (C, E)\}$$

- For directed graphs, edge $(A, B) \neq (B, A)$

Background – Graph Theory



- Child: If the graph contains the edge $A \rightarrow B$ then B is a *child* of A
- Parent: ... and A is a *parent* of B
- Path: There is a *path* from A to Z if there exists a sequence of edges $A \rightarrow B \rightarrow \ldots \rightarrow Y \rightarrow Z$
- Descendant: If there is a path from A to Z then Z is a *descendant* of A
- Ancestor: ... and A is an *ancestor* of Z
- Trail: There is a *trail* from A to Z if there exists a sequence of edges $A \leftrightarrow B \leftrightarrow \ldots \leftrightarrow Y \leftrightarrow Z$
  - $A \leftrightarrow B$ means $A \rightarrow B$ *or* $B \rightarrow A$, not necessarily both

Background – Graph Theory

Topological ordering: an ordering of the vertices in a graph such that whenever the graph contains an edge $A \rightarrow B$, A appears before B in the ordering

- When iterating through a graph in a topological order, each time you reach a vertex, you have already seen all of its parents

Background – Independence

- (Marginal) Independence: $X \perp Y$
  - Learning the value of Y doesn't tell us anything about X
  - $P(X, Y) = P(X)P(Y)$
  - $P(X \mid Y) = P(X)$
- Conditional Independence: $X \perp Y \mid Z$
  - If we already know Z, learning the value of Y doesn't tell us anything about X
  - $P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$
  - $P(X \mid Y, Z) = P(X \mid Z)$
- Conditional independence *does not* imply marginal independence
- Marginal independence *does not* imply conditional independence

Motivation

- List all of the relevant variables in your problem (observed or unobserved)
- Student example from *Probabilistic Graphical Models* by Koller and Friedman:
  - Difficulty
  - Intelligence
  - SAT
  - Grade
  - Letter
- Suppose we knew the entire joint distribution $P(D, I, S, G, L)$
- Questions about problem can be formulated as probability queries
  - $P(G)$
  - $P(G \mid S)$
  - $P(L \mid D, S)$
- In principle, we can compute these easily from the joint distribution:

$$P(L \mid D, S) = \frac{P(L,D,S)}{P(D,S)} = \frac{\sum_{i,g} P(D,i,S,g,L)}{\sum_{i,g,l} P(D,i,S,g,l)}$$

- What about in practice?
  - Computational difficulty – can't store entire joint, and sums have an exponential number of terms
  - Need an exponential amount of data to learn $P(D, I, S, G, L)$
- Can we somehow reduce the size?

Independence Assumptions

- For discrete random variables X and Y, each with four possible values, the fully specified joint distribution $P(X, Y)$ has 16 parameters:

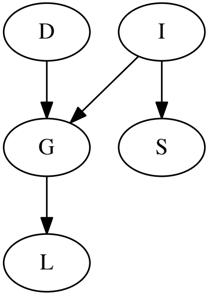
	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$
$\mathbf{y_1}$	...	...	...	...
$\mathbf{y_2}$	...	...	...	...
$\mathbf{y_3}$	...	...	...	...
$\mathbf{y_4}$	...	...	...	...

  (only 15 *independent* parameters, because the entries must sum to one)
- If we assume that X and Y are independent, we can factor the joint as $P(X, Y) = P(X)P(Y)$
- We now have only 6 independent parameters (3 for each variable)
- But marginal independence is a strong assumption, which usually does not hold
- Most variables interact in most problems, but many interactions between variables are indirect
- From student example: Difficulty affects Grade, Grade affects Letter
  - $D \not\perp L$
  - $D \perp L \mid G$
- Interactions between many interrelated variables can be difficult to reason about

Bayesian Networks

A *Bayesian network*, or Bayes net, consists of:

- A directed acyclic graph G where vertices correspond to variables and edges represent dependence between them
  - Direction of edges can informally be viewed as indicating the direction of causation
- Conditional distributions $P(X_i \mid \mathrm{Pa}_G(X_i))$ for each variable X_i, where $\mathrm{Pa}_G(X_i)$ is the parents of X_i in G



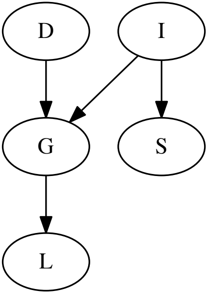
- $P(D)$
- $P(I)$
- $P(G \mid D, I)$
- $P(S \mid I)$
- $P(L \mid G)$

Local Independencies

- The structure of a Bayes net defines a set of independencies
- The *local independencies* have the form

$$X_i \perp \text{NonDescendants}_G(X_i) \mid \text{Pa}_G(X_i)$$

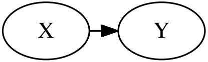
for each X_i



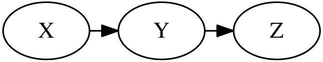
- $D \perp I, S$
- $I \perp D$
- $G \perp S \mid D, I$
- $S \perp D, G, L \mid I$
- $L \perp D, I, S \mid G$

D-Separation

- Directed separation (separation in a directed graph)
- Can think of probabilistic influence as something that can flow through a graph
- If influence can flow from one variable to another, they are dependent
- If variables are d-separated, influence can not flow and they are independent
  - Depends on which variables are observed (being conditioned on)
- Simplest case:

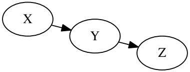
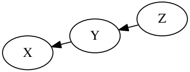
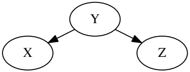
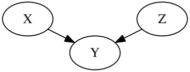


  - X and Y are *never* d-separated
- Indirect influence?

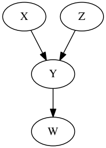


D-Separation

Are X and Z independent?

	Y not observed	Y observed
	No	Yes
	No	Yes
	No	Yes
	Yes*	No*

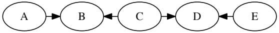
* Almost – if Y has descendants, conditioning on them has the same effect as conditioning on Y



$$X \not\perp Z \mid W$$

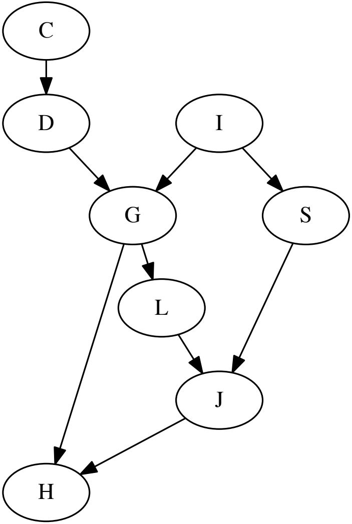
D-Separation

- If X and Y are not independent (flow of influence is not blocked) we say the trail between them is *active*
- A longer trail is active if all of its subtrails are active



  - Examine trails $A \rightarrow B \leftarrow C$, $B \leftarrow C \rightarrow D$ and $C \rightarrow D \leftarrow E$
  - If all are active, trail from A to E is active
- If there are multiple trails between two variables, influence can flow if at least one is active
- For independence statements involving more than two variables, must have that each variable on one side is d-separated from each on the other – for example, $W, X \perp Y, Z \mid V$:
  - $\text{d-sep}_G(W; Y \mid V)$
  - $\text{d-sep}_G(W; Z \mid V)$
  - $\text{d-sep}_G(X; Y \mid V)$
  - $\text{d-sep}_G(X; Z \mid V)$

D-Separation



$D \perp I$?	Yes
$C \perp L$?	No
$C \perp S$?	Yes
$C \perp S \mid H$?	No
$C \perp S \mid L, H$?	No
$D \perp S \mid J$?	No
$D \perp S \mid L, J$?	No
$D \perp S \mid I, L, J$?	Yes

Global Independencies

For all sets $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, if $\mathcal{X}$ and $\mathcal{Y}$ are d-separated given $\mathcal{Z}$ then $\mathcal{X} \perp \mathcal{Y} \mid \mathcal{Z}$

- More general than local independencies – local independencies are also implied by d-separation, but d-separation implies more than just local independencies
- The set of all independencies implied by d-separation is the *global independencies*

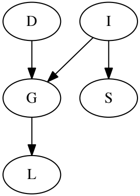
Calculating Probabilities

To calculate entries of the joint from the conditional parameterization of a Bayes net:

- Expand the joint using the chain rule in a topological ordering:

  $$P(D, I, G, S, L) = P(D)P(I \mid D)P(G \mid D, I)P(S \mid D, I, G)P(L \mid D, I, G, S)$$

  This does *not* make any independence assumptions.
- Apply the local independencies represented by the network:



$$D \perp I, S \qquad S \perp D, G, L \mid I$$

$$G \perp S \mid D, I \qquad L \perp D, I, S \mid G$$

$$\begin{aligned} P(D, I, G, S, L) &= P(D)P(I \mid \cancel{D})P(G \mid D, I)P(S \mid \cancel{D}, I, \cancel{G})P(L \mid \cancel{D}, \cancel{I}, G, \cancel{S}) \\ &= P(D)P(I)P(G \mid D, I)P(S \mid I)P(L \mid G) \end{aligned}$$

  - These factors are exactly the conditional distributions that define the Bayes net

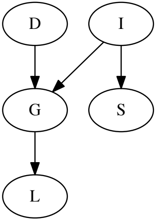
Calculating Probabilities

What about marginals?

$$\begin{aligned}
P(L) &= \sum_d \sum_i \sum_g \sum_s P(d, i, g, s, L) \\
&= \sum_d \sum_i \sum_g \sum_s P(d)P(i)P(g \mid d, i)P(L \mid g)P(s \mid i) \\
&= \sum_d \sum_i \sum_g P(d)P(i)P(g \mid d, i)P(L \mid g) \sum_s P(s \mid i) \\
&= \sum_d \sum_i \sum_g P(d)P(i)P(g \mid d, i)P(L \mid g) \\
&= \sum_d \sum_i P(d)P(i) \sum_g P(g \mid d, i)P(L \mid g) \\
&= \sum_d P(d) \sum_i P(i) \sum_g P(g \mid d, i)P(L \mid g)
\end{aligned}$$

Calculating Probabilities

Given the Bayes net:



$D \in \{0, 1\}$
$I \in \{0, 1\}$
$G \in \{A, B, C\}$
$S \in \{0, 1\}$
$L \in \{0, 1\}$

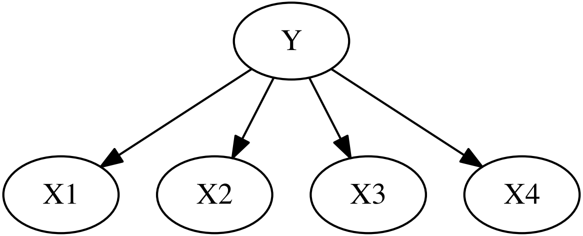
$P(D = 1) = 0.4$
$P(S = 1 \mid I = 0) = 0.05$
$P(S = 1 \mid I = 1) = 0.8$
$P(L = 1 \mid G = A) = 0.9$
$P(L = 1 \mid G = B) = 0.6$
$P(L = 1 \mid G = C) = 0.01$
$P(I = 1) = 0.3$

$P(G = A \mid D = 0, I = 0) = 0.3$
$P(G = A \mid D = 0, I = 1) = 0.9$
$P(G = A \mid D = 1, I = 0) = 0.05$
$P(G = A \mid D = 1, I = 1) = 0.5$
$P(G = B \mid D = 0, I = 0) = 0.4$
$P(G = B \mid D = 0, I = 1) = 0.08$
$P(G = B \mid D = 1, I = 0) = 0.25$
$P(G = B \mid D = 1, I = 1) = 0.3$

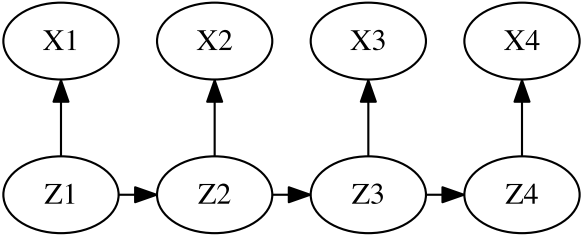
Compute $P(G = A)$:

$$
\begin{aligned}
P(G = A) &= \sum_d \sum_i \sum_l \sum_s P(d)P(i)P(G = A \mid d, i)P(l \mid G = A)P(s \mid i) \\
&= \sum_d \sum_i \sum_l P(d)P(i)P(G = A \mid d, i)P(l \mid G = A) \\
&= \sum_d \sum_i P(d)P(i)P(G = A \mid d, i) \\
&= \sum_d [P(d)P(I = 0)P(G = A \mid d, I = 0) + P(d)P(I = 1)P(G = A \mid d, I = 1)] \\
&= P(D = 0)P(I = 0)P(G = A \mid D = 0, I = 0) + P(D = 0)P(I = 1)P(G = A \mid D = 0, I = 1) + \\
&\quad P(D = 1)P(I = 0)P(G = A \mid D = 1, I = 0) + P(D = 1)P(I = 1)P(G = A \mid D = 1, I = 1) \\
&= 0.6 * 0.7 * 0.3 + 0.6 * 0.3 * 0.9 + 0.4 * 0.05 * 0.3 + 0.4 * 0.3 * 0.5 = 0.354
\end{aligned}
$$

Naive Bayes



Hidden Markov Model



Why Bayes Nets?

- Tractable
- Interpretable
- Declarative representation – separation of knowledge from reasoning
- Generalization of many other models