Bayesian networks

Overview

1. Directed Acyclic Graphs
2. Directed Graphical Models
3. Independence and Markov Properties

Graph Terminology

$\mathcal{G} = (V, E)$ **directed graph.** V **vertices.** E **edges** (ordered pairs of vertices)

X, Y **adjacent** if $X \to Y$ edge. Y **child** of X. X **parent** of Y.

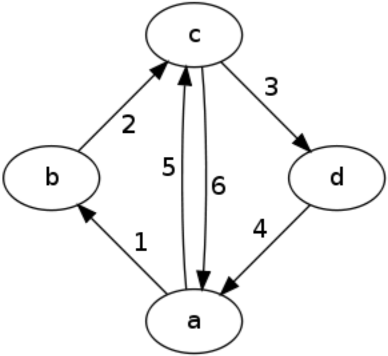
$X \to \cdots \to Y$ **directed path.** Y **descendant** of X. X **ancestor** of Y.

$X \leftarrow \cdots \to Y$ **undirected path** (ignore direction of arrows).

$X \to Y \leftarrow Z$ **collider.** $X \to Y \to Z$, $X \leftarrow Y \leftarrow Z$, $X \leftarrow Y \to Z$ **non-colliders.**

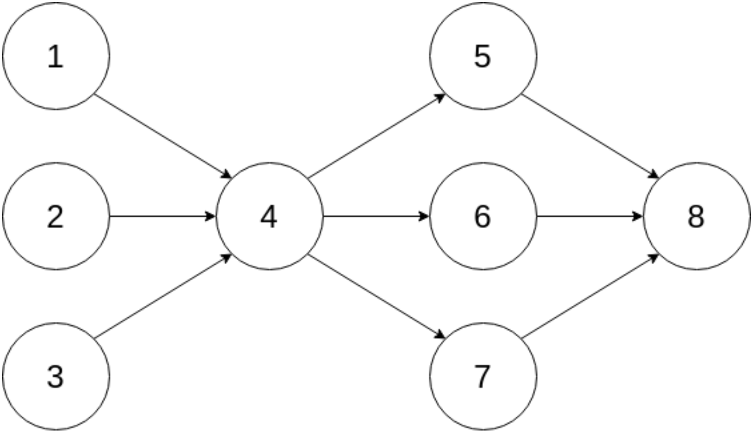
$X \to \cdots \to X$ **cycle.** $\mathcal{G}$ **directed acyclic graph** if no cycles.

Directed Graphs with Cycles



$\{1, 2, 3, 4\}$ and $\{3, 4, 5\}$ are cycles.
$\{3, 4, 6\}$ is not a cycle.

Directed Acyclic Graphs (DAG)



Directed Graphical Models

A directed graphical model or Bayesian network consists of a multivariate random variable $\boldsymbol{X} = (X_1, \ldots, X_n)$ and a corresponding graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where

- $\mathcal{V} = \{1, \ldots, n\}$, where the variable X_i is represented by node i,
- $(i, j) \in \mathcal{E}$ is denoted by an arrow connecting i to j,
- the probability mass (density) function of $\boldsymbol{X}$ satisfies the factorization property.

Factorization

We denote the parents of node i by $\mathrm{Pa}(i)$.

A probability mass function $P(\boldsymbol{X} = \boldsymbol{x})$ satisfies the factorization property with respect to a DAG if

$$P(\boldsymbol{X} = \boldsymbol{x}) = \prod_{i=1}^{n} P(X_i = x_i \mid \mathrm{Pa}(i)). \tag{1}$$

Example

A Set of Tables for Each Node

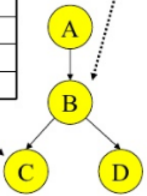
A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

Each node X_i has a conditional probability distribution $P(X_i \mid \text{Parents}(X_i))$ that quantifies the effect of the parents on the node

The parameters are the probabilities in these conditional probability tables (CPTs)

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1



B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

Using a Bayesian Network Example

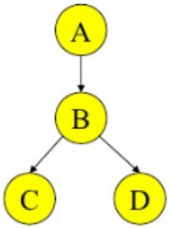
Using the network in the example, suppose you want to calculate:

P(A = true, B = true, C = true, D = true)

= P(A = true) * P(B = true | A = true) *

P(C = true | B = true) P(D = true | B = true)

= (0.4)*(0.3)*(0.1)*(0.95)



Examples

EXAMPLES

Smoking

$$
\begin{aligned}
& f(\text{overweight}, \text{smoking}, \text{heart disease}, \text{cough}) \\
&= f(\text{overweight}) \times f(\text{smoking}) \\
&\times f(\text{heart disease} \mid \text{overweight}, \text{smoking}) \\
&\times f(\text{cough} \mid \text{smoking}).
\end{aligned}
$$

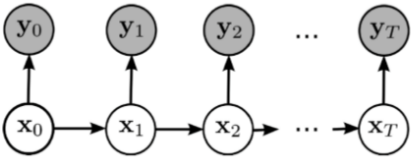
joint probability

root probabilities

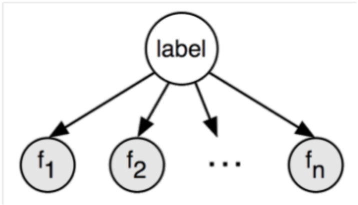
transition probabilities

Example

Hidden Markov Model

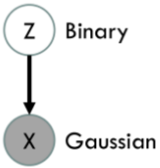


Naïve Bayes

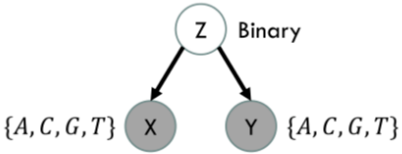


Example

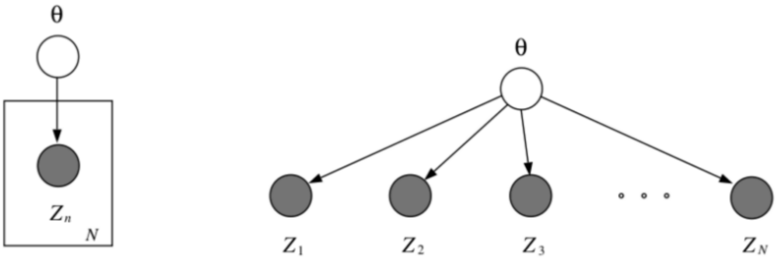
Gaussian Mixtures



Phylogenetic Models



Examples

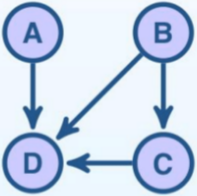


When N is large, the joint distribution is difficult to analyze. A directed graphical model helps us to factorize the model so that analysis can be done more efficiently.

$$P(\boldsymbol{Z} = \boldsymbol{z}, \theta = x) = P(\theta = x) \prod_{i=1}^{N} P(Z_i = z_i \mid \theta = x). \tag{2}$$

Directed Gaussian Graphical Model

Parametrize the model using *structural equation modelling* (SEM).



Gaussian

$$
\begin{aligned}
A &= \varepsilon_A, \qquad \varepsilon_A, \varepsilon_B, \varepsilon_C, \varepsilon_D \sim \mathcal{N}(0,1)\\
B &= \varepsilon_B\\
C &= \lambda_{BC} B + \varepsilon_C\\
D &= \lambda_{AD} A + \lambda_{BD} B + \lambda_{CD} C + \varepsilon_D
\end{aligned}
$$

$$
\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -\lambda_{BC} & 1 & 0 \\ -\lambda_{AD} & -\lambda_{BD} & -\lambda_{CD} & 1 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \\ D \end{pmatrix} \sim \mathcal{N}(0, \mathrm{Id})
$$

Directed Gaussian Graphical Model

Let $\boldsymbol{K}$ be the matrix of the left hand side of the previous expression, and let $\boldsymbol{A} = (A, B, C, D)$ be the multivariate random variable. Then the graphical model in the previous slide implies that $\boldsymbol{KA}$ is the standard normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, which further implies that

$$\boldsymbol{A} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{K}^{-1}\left(\boldsymbol{K}^{-1}\right)^{T}\right). \tag{3}$$

The Half Way Point

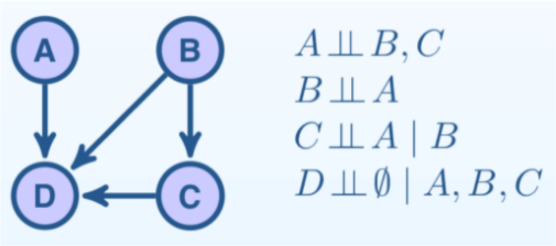
5 Minutes Break

Local Markov Property

We say that a distribution $\mathbb{P}$ satisfies the **local Markov property** with respect to $\mathcal{G}$ if for all variables W,

$$W \perp \widetilde{W} \mid \pi_W$$

where π_W are the parents of W, and $\widetilde{W}$ are the variables which are neither parents nor descendants of W.



D-Separation

Consider the following undirected path from X to Z:

$$X - Y_1 - Y_2 - \cdots - Y_n - Z.$$

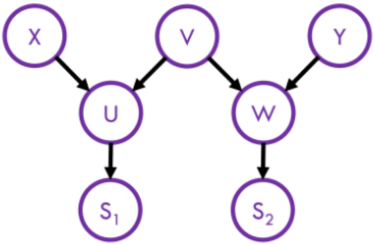
Let W be some subset of vertices that do not contain X or Z.

Think of each intermediate vertex as a gate, and W a set of keys.

1. Collider gates are usually closed; all other gates are usually open.
2. If a collider or one of its descendants is in W, then that gate is opened.
3. If a non-collider is in W, then that gate is closed.

If all the gates are open, we say that X and Z are **d-connected** given W.
If we cannot find any such path, then X and Z are **d-separated** given W.
Sets S and T are **d-separated** given W if it is true for all $X \in S, Z \in T$.

EXAMPLES



Three layer DAG

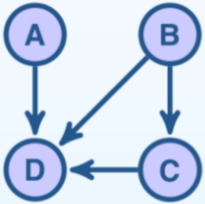
X and Y are d-separated (given the empty set);
X and Y are d-connected given $\{S_1, S_2\}$;
X and Y are d-separated given $\{S_1, S_2, V\}$.

Global Markov Property

We say that a distribution $\mathbb{P}$ satisfies the **global Markov property** with respect to $\mathcal{G}$ if

$$S \perp T \mid W$$

for all disjoint subsets S, T, W such that S and T are d-separated given W.



$$A \perp\!\!\!\perp B, C$$
$$B \perp\!\!\!\perp A$$
$$C \perp\!\!\!\perp A \mid B$$
$$D \perp\!\!\!\perp \emptyset \mid A, B, C$$

$$A \perp B \mid C$$
$$A \perp C$$

Hammersley-Clifford Theorem

The following are equivalent:

1. $\mathbb{P}$ satisfies the **factorization property** with respect to $\mathcal{G}$.
2. $\mathbb{P}$ satisfies the **local Markov property** with respect to $\mathcal{G}$.
3. $\mathbb{P}$ satisfies the **global Markov property** with respect to $\mathcal{G}$.

Explaining Away

EXPLAINING AWAY

Why does conditioning on a collider lead to dependence?

If you don't know your friend is late:

$$\mathbb{P}(\text{Aliens}|\text{Watch}) = \mathbb{P}(\text{Aliens}) \qquad \text{Aliens} \perp \text{Watch}$$

If you now know your friend is late:

$$\mathbb{P}(\text{Aliens}|\text{Watch}, \text{Late}) < \mathbb{P}(\text{Aliens}|\text{Late}) \qquad \text{Aliens} \not\perp \text{Watch} \mid \text{Late}$$

Knowing his broken watch made him late **explains away** the possibility that he is late because he was abducted by aliens.

Advantages of Bayesian Network Representations

Conditional Independence relations can be read of the underlying DAG.

Can describe causal relationships between variables.

Can handle incomplete data or latent variables.