Bayesian networks

Andrea Passerini
passerini@disi.unitn.it

Machine Learning

Why Bayesian Networks?

Many applications involve several uncertain quantities that interact.

Examples

- diseases, symptoms, and tests;
- failures and sensor alarms;
- weather, irrigation, and wet grass;
- genes, proteins, and phenotypes.

The modeling problem

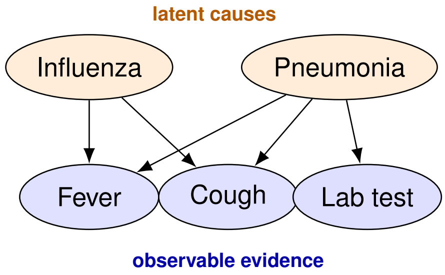
We want one structured model that can

- encode dependencies;
- answer many probability queries;
- avoid an exponentially large table.

Central challenge

Represent a complex joint distribution through local relationships.

A Running Example: Medical Diagnosis



Latent causes: influenza, pneumonia.

Observable evidence: fever, cough, test.

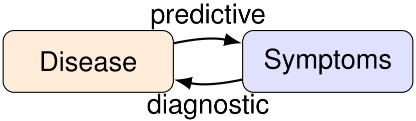
Questions we may ask

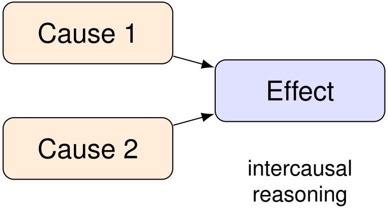
- How likely is pneumonia after observing fever and cough?
- How does a negative test change the diagnosis?
- Which disease best explains the evidence?

One model, many queries

The same joint distribution supports prediction, diagnosis, and explanation.

Different Directions of Reasoning





Predictive
Cause $\rightarrow$ effect

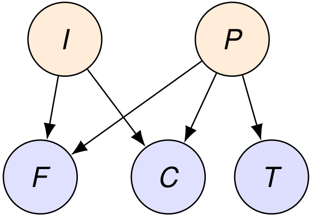
$$p(F \mid I)$$

Diagnostic
Effect $\rightarrow$ cause

$$p(I \mid F)$$

Intercausal
Observing a common effect couples beliefs about its possible causes.

Bayesian Networks: Graphical Semantics



I: influenza, P: pneumonia
F: fever, C: cough, T: test

A Bayesian network consists of

- a *directed acyclic graph* $\mathcal{G}$;
- one random variable X_i for each node;
- one local model

$$p(X_i \mid \mathrm{Pa}(X_i))$$

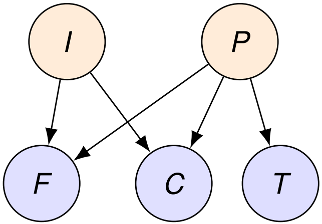
  for each variable.

Meaning of an arrow

$X_j \to X_i$ identifies X_j as a direct probabilistic parent of X_i.

Caution. A causal reading requires additional assumptions and domain knowledge.

Bayesian Networks: Probabilistic Semantics



A Bayesian network represents a joint probability distribution that factorizes according to the graph:

$$\boxed{p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i \mid \mathrm{Pa}(X_i))}$$

Each factor describes the distribution of one variable conditioned only on its parents.

Medical network

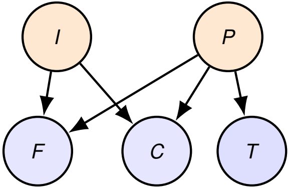
For the graph on the left,

$$\begin{aligned} p(I, P, F, C, T) = p(I)\, p(P) p(F \mid I, P) \\ p(C \mid I, P) p(T \mid P). \end{aligned}$$

The graph specifies the structure of the factorization; the local probability models specify its numerical values.

Bayesian Networks: A Complete Probabilistic Model

Structure



Local probability models

$p(I)$

I	$p(I)$
1	.08
0	.92

$p(P)$

P	$p(P)$
1	.06
0	.94

$p(T = + \mid P)$

P	$p(T = +)$
1	.92
0	.08

$p(F = 1 \mid I, P)$

I	P	$p(F = 1)$
0	0	.05
1	0	.82
0	1	.73
1	1	.97

$p(C = 1 \mid I, P)$

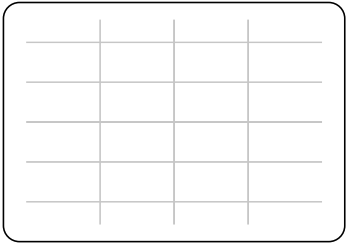
I	P	$p(C = 1)$
0	0	.04
1	0	.58
0	1	.76
1	1	.88

Key idea

The graph specifies which local models are needed; the conditional probability tables specify their numerical values.

From a Large Joint Table to Local Models

Full joint table



$2^m - 1$ parameters
for m binary variables

graphical assumptions →

Local probability tables

$p(I)$ $p(P)$

$p(F \mid I, P)$ $p(C \mid I, P)$

$p(T \mid P)$

smaller, interpretable
local mechanisms

What the graph buys us

Conditional-independence assumptions replace one enormous table with several smaller conditional probability tables.

Why Factorization Saves Parameters

Assume all five medical variables are binary.

Parameters in the local tables

$p(I),\ p(P)$ $1+1$

$p(F \mid I, P)$ $2^2 = 4$

$p(C \mid I, P)$ $2^2 = 4$

$p(T \mid P)$ $2^1 = 2$

Total: 12 **parameters**

Unrestricted joint model

$$2^5 - 1 = 31$$

independent parameters.

Bayesian-network model

$$1 + 1 + 4 + 4 + 2 = 12$$

parameters.

How the Graph Simplifies the Chain Rule

Ordering node variable names in topological order, the chain rules says that:

$$p(x_1, \ldots, x_m) = \prod_{i=1}^{m} p(x_i \mid x_1, \ldots, x_{i-1}).$$

$$\boxed{p(T \mid I, P, F, C)} \xrightarrow{\text{graph structure}} \boxed{p(T \mid P)} \longrightarrow \boxed{T \perp\!\!\!\perp \{I, F, C\} \mid P}$$

chain-rule conditional

local conditional

What the graph asserts

Each variable depends on its predecessors only through its parents:

$$p(x_i \mid x_1, \ldots, x_{i-1}) = p(x_i \mid \mathrm{Pa}(x_i)).$$

The omitted predecessors correspond to conditional-independence assumptions.

A Generative Interpretation

The factorization also defines a procedure for generating complete observations.

1. Sample roots $I \sim p(I), \ P \sim p(P)$ → 2. Sample symptoms $F, C \sim p(\cdot \mid I, P)$ → 3. Sample test $T \sim p(T \mid P)$

parents are sampled before their children

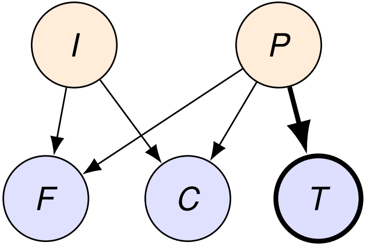
1. Sample root variables from their marginal distributions.
2. Visit the remaining nodes in a topological order.
3. Sample each node conditional on the already sampled values of its parents.

Ancestral sampling

Multiplying the probabilities used along this procedure gives exactly $\prod_i p(x_i \mid \mathrm{Pa}(x_i))$.

What Can We Read Directly from the Graph?

The graph is useful even before the numerical probabilities are known.



$$T \perp\!\!\!\perp \{I, F, C\} \mid P$$

The medical graph suggests that

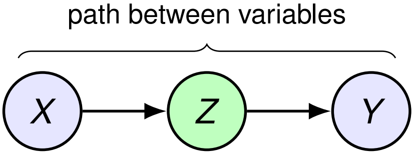
- the test depends directly on pneumonia;
- once pneumonia is known, influenza adds no direct information about the test;
- symptoms can transmit information between possible causes.

Next question

How do paths and observed variables determine whether information can flow between two nodes?

Conditional Independence in Bayesian Networks

Bayesian networks are useful because the graph represents assumptions about which variables become irrelevant once other variables are known.



Once Z is known, information from Y may no longer change beliefs about X.

Conditional independence

Random variables X and Y are conditionally independent given Z if

$$p(x \mid y, z) = p(x \mid z).$$

Equivalently,

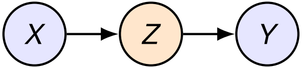
$$p(x, y \mid z) = p(x \mid z)p(y \mid z).$$

- Knowing Y gives no additional information about X once Z is known.
- The graph helps us read such statements systematically.

Three Basic Patterns

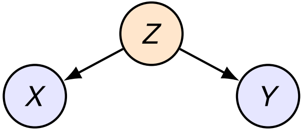
Every path in a Bayesian network is built from three local patterns.

Chain



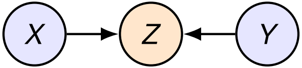
$X \rightarrow Z \rightarrow Y$

Fork



$X \leftarrow Z \rightarrow Y$

Collider



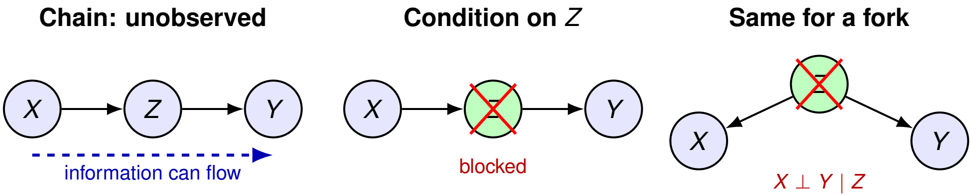
$X \rightarrow Z \leftarrow Y$

Reading paths

A path can transmit probabilistic information unless it is blocked by the evidence. Chains, forks, and colliders behave differently under conditioning.

Chains and Forks: Observing the Middle Blocks the Path

For chains and forks, information can flow through the middle node. Conditioning on the middle node blocks the path.



Chain

$$X \rightarrow Z \rightarrow Y \quad \text{implies} \quad X \perp Y \mid Z.$$

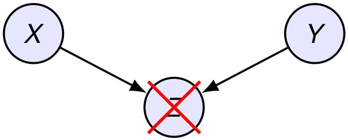
Fork

$$X \leftarrow Z \rightarrow Y \quad \text{implies} \quad X \perp Y \mid Z.$$

Colliders: Observing the Middle Opens the Path

Colliders behave in the opposite way. Without observing the effect, the causes are independent along the path. Observing the effect makes the causes dependent.

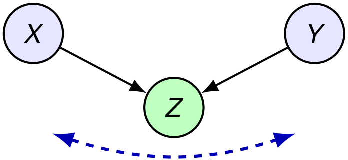
Unobserved collider



path blocked

$X \perp Y$ along this path

Observed collider



path opens

X and Y can become dependent

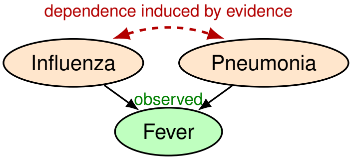
Collider pattern

$$X \to Z \leftarrow Y.$$

Normally the path is blocked, but observing Z *or one of its descendants* opens the path.

Explaining Away

A collider explains why evidence for one cause can reduce belief in another cause.



Two possible causes compete to explain the same evidence.

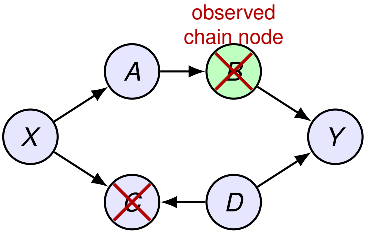
1. Initially, influenza and pneumonia may be approximately independent.
2. Observing fever makes both causes more plausible.
3. If pneumonia is then confirmed, fever is partly explained.
4. The posterior probability of influenza may decrease.

Intuition

Competing explanations for the same observation become dependent after that observation is known.

D-Separation

D-separation generalizes the three local patterns to arbitrary paths.



collider at C is unobserved: lower path blocked

All paths blocked $\Rightarrow X \perp Y \mid B$

A path is *blocked* by observed variables E if it contains

- a chain or fork whose middle node is observed;
- a collider whose middle node is not observed and has no observed descendant.

If all paths from X to Y are blocked by E, then

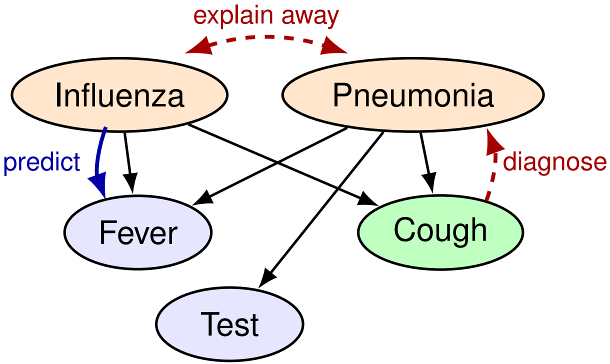
$$X \perp Y \mid E.$$

Graphical criterion

D-separation lets us read conditional-independence statements directly from the graph.

Evidence and Explanation in the Medical Network

The same network supports different kinds of reasoning depending on which variables are observed.



Prediction: disease $\Rightarrow$ symptoms **Diagnosis:** symptoms $\Rightarrow$ diseases
Explanation: one cause can explain away another.

Local and Global Independences

A BN structure G encodes local assumptions

$$\mathcal{I}_\ell(G) = \{X_i \perp\!\!\!\perp \mathrm{NonDesc}(X_i) \mid \mathrm{Pa}(X_i)\},$$

and global assumptions obtained through d-separation

$$\mathcal{I}(G) = \{A \perp\!\!\!\perp B \mid C : \mathrm{dsep}(A; B \mid C)\}.$$

Connection

For distributions that factorize over G, the local Markov property and d-separation are sound descriptions of conditional independence.

Inference in Bayesian Networks

Once a Bayesian network has been constructed, it can answer many probabilistic queries.

- What is the probability that a patient has pneumonia?
- How does observing fever change this probability?
- Which disease best explains the observed symptoms?
- Which additional test would be most informative?

Common inference tasks

$P(X)$	marginal probability
$P(X \mid E = e)$	posterior probability after observing evidence
$\arg\max_x P(x \mid E = e)$	most probable explanation

Inference Example (head-to-head connection)

Setting

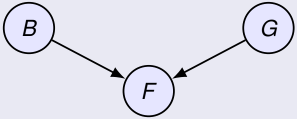
- A fuel system in a car:

  battery B, either charged ($B = 1$) or flat ($B = 0$)
  fuel tank F, either full ($F = 1$) or empty ($F = 0$)
  electric fuel gauge G, either full ($G = 1$) or empty ($G = 0$)

Conditional probability tables (CPT)

- Battery and tank have independent prior probabilities:

$$P(B = 1) = 0.9 \qquad P(F = 1) = 0.9$$



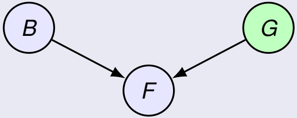
- The fuel gauge is conditioned on both (unreliable!):

$$P(G = 1|B = 1, F = 1) = 0.8 \quad P(G = 1|B = 1, F = 0) = 0.2$$
$$P(G = 1|B = 0, F = 1) = 0.2 \quad P(G = 1|B = 0, F = 0) = 0.1$$

Inference Example (head-to-head connection)

Probability of empty tank

- Prior:



$$P(F=0) = 1 - P(F=1) = 0.1$$

- Posterior after observing empty fuel gauge:

$$P(F=0|G=0) = \frac{P(G=0|F=0)P(F=0)}{P(G=0)} \simeq 0.257$$

Note

The probability that the tank is empty *increases* from observing that the fuel gauge reads empty (not as much as expected because of strong prior and unreliable gauge)

Inference Example (head-to-head connection)

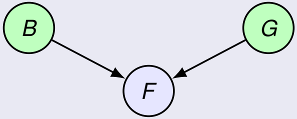
Derivation

$$\begin{aligned} P(G=0|F=0) &= \sum_{B\in\{0,1\}} P(G=0,B|F=0) \\ &= \sum_{B\in\{0,1\}} P(G=0|B,F=0)P(B|F=0) \\ &= \sum_{B\in\{0,1\}} P(G=0|B,F=0)P(B) = 0.81 \end{aligned}$$

$$\begin{aligned} P(G=0) &= \sum_{B\in\{0,1\}} \sum_{F\in\{0,1\}} P(G=0,B,F) \\ &= \sum_{B\in\{0,1\}} \sum_{F\in\{0,1\}} P(G=0|B,F)P(B)P(F) \end{aligned}$$

Inference Example (head-to-head connection)

Probability of empty tank



- Posterior after observing that the battery is also flat:

$$P(F=0|G=0,B=0) = \frac{P(G=0|F=0,B=0)P(F=0|B=0)}{P(G=0|B=0)} \simeq 0.111$$

Note

- The probability that the tank is empty *decreases* after observing that the battery is also flat
- The battery condition *explains away* the observation that the fuel gauge reads empty
- The probability is still greater than the prior one, because the fuel gauge observation still gives some evidence in favour of an empty tank

Building Bayesian Networks

1. Define variables for entities that can be *observed* or that you can be interested in *predicting* (latent variables can also be sometimes useful)
2. Work with domain experts to identify plausible direct dependencies.
3. Prefer sparse, interpretable structures; causal knowledge can guide directions when justified.
4. Avoid assigning exact zero probabilities unless they are logically certain.
5. Use data to estimate parameters and, when appropriate, help learn structure.

Model checking

A useful graph should support the intended queries and make defensible conditional-independence assumptions.

Take-Home Messages

1. Bayesian networks combine a DAG with local conditional probability models.
2. Factorization replaces a large joint table with smaller local factors.
3. D-separation connects graph structure to conditional independence.
4. Evidence supports predictive, diagnostic, and intercausal reasoning.