Intelligent Systems: Reasoning and Recognition

James L. Crowley

MOSIG M1 Second Semester 2020/2021
Lesson 20 15 April 2021

Reasoning with Bayesian Networks

Sources:

1. Koller, D., and Friedman, N., Probabilistic graphical models: principles and techniques. MIT press, 2009.
2. NEIL, Martin, FENTON, Norman, and NIELSON, Lars. Building large-scale Bayesian networks. *The Knowledge Engineering Review*, 2000, vol. 15, no 3, p. 257-284.

Evidential Reasoning

In lesson 18 we saw that a situation model could enable reasoning with evidence to confirm partially observable narratives. For example, let S_1, S_2 and S_3 be a narrative composed of 3 situations in which S_2 is not observable. Recall that the situation S_2 is a conjunction of N predicates $\{R\}=(r_1()\wedge...\wedge r_N())$ over the K entities in working memory. The predicates $r_n(\text{-})$ represent relations between entities (observable phenomena).

To perform Bayesian reasoning, we must replace the Boolean predicates $r_n()$ with probabilistic predicates. Probabilistic predicates are predicate functions that return a probability as a truth-value instead of a Boolean {T, F},

We can then use the probabilities of the predicates, $r_n()$, to determine the probability of a situation. To simplify, we will note $r_n()$ as simply r_n. $\{R\}$ is the conjunction of predicates that defines the situation. The probability of a situation given the probability of relations.

$$P(S\,|\,\{R\}) = \prod_{r_n\in\{R\}} P(S\,|\,r_n)$$

This is made easier if we reason with odds. Consider the probability of a Situation, S, given the probability for a relation, r.

From Baye's Rule: $P(r)\cdot P(S\,|\,r) = P(S)\cdot P(r\,|\,S)$

and $P(r)\cdot P(\neg S\,|\,r) = P(\neg S)\cdot P(r\,|\,\neg S)$

The ratio is $\frac{P(r)}{P(r)}\cdot\frac{P(S\,|\,r)}{P(\neg S\,|\,r)} = \frac{P(S\,|\,r)}{P(\neg S\,|\,r)} = \frac{P(S)\cdot P(r\,|\,S)}{P(\neg S)\cdot P(r\,|\,\neg S)}$

This ratio is referred to as "odds" and used in betting:

The a-priori odds of a situation S are defined as $Odds(S:\neg S) = \frac{P(S)}{P(\neg S)}$

The conditional odds for the situation are $Odds((S:\neg S)\,|\,r) = \frac{P(S\,|\,r)}{P(\neg S\,|\,r)}$

Thus: $Odds((S:\neg S)\mid r) = Odds((S:\neg S)\cdot \frac{P(r\mid S)}{P(r\mid \neg S)}$

The ratio $\frac{P(r\mid S)}{P(r\mid \neg S)}$ is called the conditional likelihood of r from S.

$$L_r = \frac{P(r\mid S)}{P(r\mid \neg S)}$$

Written this way, the conditional odds for a situation are

$$Odds((S:\neg S)\mid r) = Odds((S:\neg S)\cdot L_r$$

Products of probabilities are inconvenient as they tend toward very small numbers. We can reformulate evidence accumulation using addition in place of multiplication using logarithms. This is commonly done using Log-odds. Thus allows us to convert the problem to an additive process for accumulating evidence using logarithms.

$$Log(Odds((S:\neg S)\mid r)) = Log(Odds((S:\neg S)) + Log(L_r)$$

Evidence for S by r is then defined as the Log of the conditional likelihood.

$$E_r = Log(L_r) = Log\left(\frac{P(r\mid S)}{P(r\mid \neg S)}\right) = Log(P(r\mid S)) - Log(P(r\mid \neg S))$$

This favors detection of relations that are unique to certain situations. Discovery of such relations provides evidence that the situation occurred. Relations that occur in all situations are dicarded.

The problem is that relations may not be directly observable. In many cases, the non-observable relations can be inferred using causal reasoning with Bayesian Networks. The key is to model each relation as a Random Variable and to determine the probability of the variable using Bayesian Networks.

Bayesian Networks

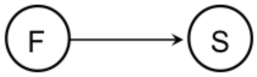
Bayesian Networks are graphical models for reasoning about random variables.

In a Bayesian Network, the nodes represent random variables (discrete or continuous) and the arcs represent relations between variable. Arcs are often causal connections but can be other forms of association. Bayesian networks allow probabilistic beliefs about random variables to be updated automatically as new information becomes available.

The nodes in a Bayesian network represent the probability of random variables, X from the domain.

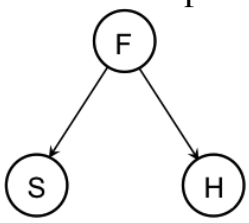
Directed arcs (or links) connect pairs of nodes, $X_1 \rightarrow X_2$, representing the direct dependencies between random variables.

For example: Fire causes Smoke. Let F=Fire, S=Smoke

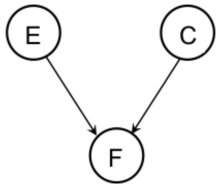


We can use graphical models to represent causal relations.

For example add a third random variable, H=Heat.
Then Fire causes Smoke and Heat would be expressed as:



Graphical models can also express multiple possible causes for diagnostic reasoning. For example, Fire can be caused by an Electrical problem (E) or by a Cigarette (C)



The strength of the relationship between variables is quantified by conditional probability distributions associated with each node. These are represented by Conditional Probability Tables.

Probability Distribution Tables

Recall that a **Probability Distribution Table** gives the relative frequency of occurrence for all possible values for a random variable (or property or feature or attribute or probabilistic predicate) from a set of observations (training set). Random variables can be Boolean, symbolic or numeric (natural, integer or real). The set of possible values for a variable must be (1) Mutually Exclusive and (2) Complete.

The Probability Distribution Table gives the relative frequency of occurrence for each value of the variable, and can be computed by simply counting the number of occurrences in the training set. To be a valid probability, the values must be normalized to sum to 1.

Joint Probability Distributions Tables

Distribution tables can be easily generalized to multiple random variables. For example consider a training set of observations, $\{\vec{X}_m\}$ of M observations of 2 random variables, A_m, and B_m.

$$\forall m = 1, M: \; h(A_m, B_m) \leftarrow h(A_m, B_m) + 1;$$

we commonly write this as

$$\forall m = 1, M: \; h(\vec{X}_m) \leftarrow h(\vec{X}_m) + 1;$$

Then for any pair of values of A=*a*, *B=b* $P(a,b) = \frac{1}{M} h(a,b)$

The complete table must sum to 1. $\sum_{a,b} P(a,b) = 1$

We can eliminate a class from the table by summing a column:

$$P(A) = \sum_{x \in B} P(A,x)$$

All this can be generalized to multiple features. For three features A, B, C

$P(A,B,C) = \frac{1}{M} h(A,B,C)$ and $P(A,B) = \sum_{x \in C} P(A,B,x)$

Graphically, probability distribution tables are displayed as a table:

P(G,C)	Brown	Blue	Green
Male	0.4	0.1	0.0
Female	0.3	0.1	0.1

Conditional Probability Tables (CPT)

Bayes Rule provides a definition of conditional probability tables.

For a probability distribution *P(A,B)* the Conditional probability can be defined as

$$P(A \mid B) = \frac{P(A,B)}{\sum_x P(x,B)} = \frac{P(A,B)}{P(B)}$$

With multiple features; $$P(A,B \mid C) = \frac{P(A,B,C)}{\sum_{x \in C} P(A,B,x)} = \frac{P(A,B,c)}{P(A,B)}$$

For example, consider the Boolean values F=Fire and S=Smoke

P(Fire, Smoke) = P(Smoke|Fire) P(Fire)

P(Smoke|Fire)	Smoke	¬Smoke
Fire	0.9	0.1
¬Fire	0.001	0.999

Each row sums to one. Columns are independent.

Note that with Boolean features, some authors omit the columns for False.

Suppose we know a joint table P(F, S, H) and we wish to compute P(F | S).

$$P(F \mid S) = \frac{\sum_{x \in H} P(F,S,x)}{\sum_{x \in S} \sum_{y \in H} P(F,x,y)}$$ This is clumsy and expensive.

The calculation is even worse if our table includes possible causes such as an electrical Fire (E) or a Cigarette fire (C): P(F, S, H, E, C). To compute P(F|S) we first have to sum out all the other terms.

Bayesian networks gives a way to simplify the calculation by factoring the distribution table P(F,S,H) into components.

Conditional Independence

Conditional independence allows us to factor a Probability Distribution Table into a product of much smaller Conditional Probability Tables.

Independent Random Variables

Two random variables are Independent if $P(A,B) = P(A) \cdot P(B)$
This is written: $A \perp B$. $A \perp B$ implies that $P(A \mid B) = P(A)$

Demonstration: $P(A \mid B) = \frac{P(A,B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$

Conditional Independence

Conditional independence occurs when observations A and B are independent given a third observations C. Conditional independence tells us that when we know C, evidence of B does not change the likelihood of A.

If A and B are independent given C then $P(A \mid B, C) = P(A \mid C)$.
Formally: $A \perp B \mid C \Leftrightarrow P(A \mid B, C) = P(A \mid C)$
Note that $A \perp B \mid C = B \perp A \mid C \Leftrightarrow P(B \mid A, C) = P(B \mid C)$

A typical situation is that both A and B result from the same cause, C.
For example, Fire causes Smoke and Heat.

When A is conditionally independent from B given C, we can also write:

$$P(A, B \mid C) = P(A \mid B, C) \cdot P(B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

Chain Rule

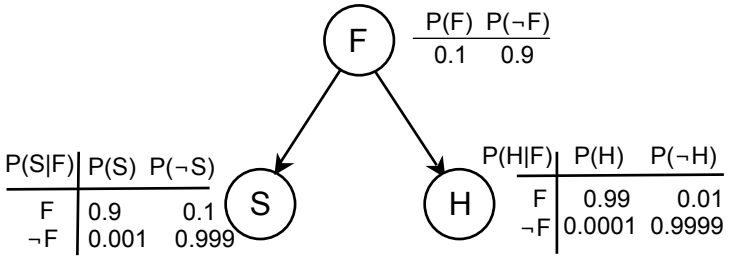
Bayesian networks explicitly express conditional independencies in probability distributions and allows computation of probabilities distributions using the chain rule. When A and B are conditionally independent given C,

$$P(A \mid B, C) = P(A \mid C) \quad \text{and} \quad P(A, B \mid C) = P(A \mid C) \cdot P(B \mid C)$$

When conditioned on C, the probability distribution table $P(A, B)$ factors into a product of marginal distributions, $P(A|C)$ and $P(B|C)$.

Factoring Distribution Tables with Bayesian Networks

Bayesian Networks factor a large Probability Distribution Table (PDT) into a set of much smaller Conditional Probability Tables (CPTs).



Factoring a PDT requires that the variables be conditionally independent.

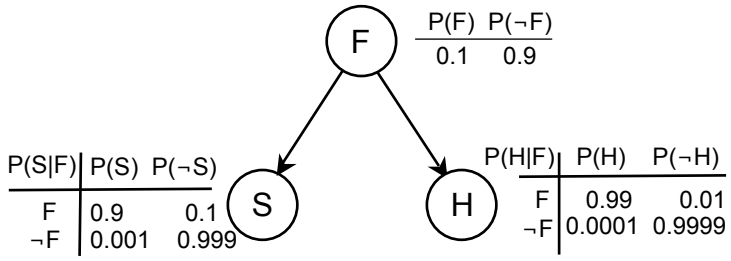
Computing with Conditional Probability Tables

Conditional independence allows us to factor a Probability Distribution into a product of much smaller Conditional Probability Tables.

For example, let F=Fire, S=Smoke and H=Heat.

$P(S, H, F) = P(S \mid F)\, P(H \mid F)\, P(F)$ Factors into
$P(S, F) = P(S \mid F)\, P(F)$ and $P(H, F) = P(H \mid F)\, P(F)$

Each factor is described by a Conditional Probability Table.



Each row of the table must sum to 1. To simplify the table, most authors do not include the last column. The values for last column are determined by subtracting the sum of the other columns from 1.

Arcs link a "Parent node" to a "Child Node). F → S Fire is Parent to Smoke

This is written Parent(S) = F
The set of all parents of a node x is the function Parents(x).

In General $P(X_1, X_2, \cdots, X_D) = \prod_n P(X_n \mid \text{parents}(X_n))$

We can use the network to answer questions. For example:

What is the probability of fire if we see smoke?

$$P(F \mid S) = \frac{P(F,S)}{P(S)}$$

For this we need the joint probability of fire and smoke, *P(F,S)* as well as *P(S)*

If we use the full PDT, we would be required to compute the joint probability by summing out terms H and F in the table *P(F,S,H)*.

$$P(F,S) = \sum_H P(F,S,H) \quad \text{and} \quad P(S) = \sum_F \sum_H P(F,S,H)$$

The graph provides a direct solution using only *P(F,S)*

$$P(F,S) = P(S \mid F)P(F) = 0.9 \cdot 0.1 = 0.09$$

and

$$P(S) = P(F,S) + P(\neg F,S) = 0.9 \cdot 0.1 + 0.001 \cdot 0.9 = 0.0909$$

Thus

$$P(F \mid S) = \frac{P(F,S)}{P(S)} = \frac{0.09}{0.0909} = 0.99$$

In a larger problem the full PDT would have been MUCH larger.

A Joint Distribution in Structured Form

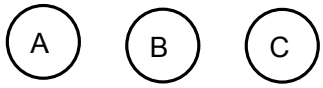
A Bayesian Network is a Joint Distribution in Structured form. The network is an Acyclic Directed Graph.

Dependence and independence are represented as a presence or absence of edges:

Node = random Variable (equivalent to a probabilistic predicate).
Directed Edge = Conditional Dependence
Absence of an Edge = Conditional Independence.

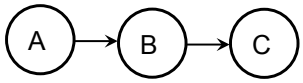
The graph shows conditional (and causal) relations. When you specify a graph, you obtain a formula. Common structures are:

Marginal Independence:



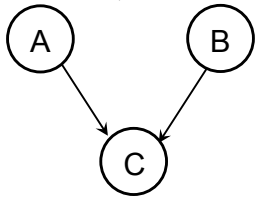
$$P(A, B, C) = P(A) \cdot P(B) \cdot P(C)$$

Markov Dependence (Causal Chain)



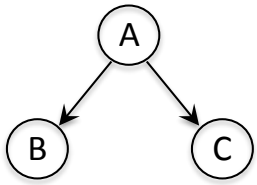
$$P(A,B,C) = P(C \mid B) \cdot P(B \mid A) \cdot P(A)$$

Independence Causes: (Common Effect)



$$P(A, B, C) = P(C \mid A, B) \cdot P(A) \cdot P(B)$$

Common Cause



$$P(A,B,C) = P(B \mid A) \cdot P(C \mid A) \cdot P(A)$$

Arcs link a "Parent node" to a "Child Node). $A \rightarrow B$
A is the Parent of B. This is written $Parent(B) = A$

The set of all parents of a node x is the function Parents(x).

In General $P(X_1, X_2, \cdots, X_D) = \prod_n P(X_n \mid \text{Parents}(X_n))$

A series of arcs list ancestors and descendents A→ B→ C

Node A is an ancestor of C. Node C is a descendent of A.

Reasoning with Bayesian networks

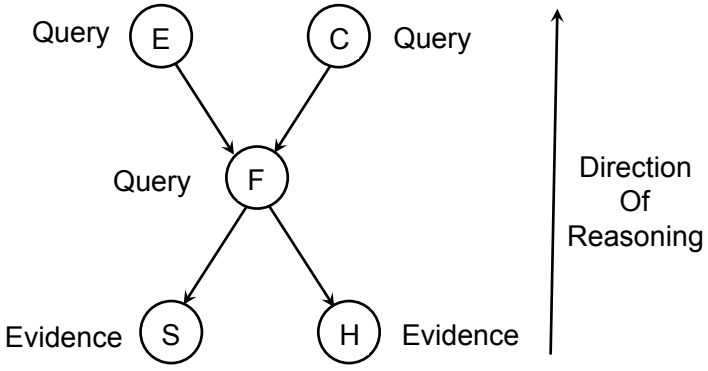
Bayesian networks support several types of reasoning.

Reasoning (inference) occurs as a flow of information through the network.
This is sometimes called propagation or belief updating or even conditioning.

Note that information flow is *not* limited to the directions of the arcs.

Diagnostic Reasoning
Diagnostic reasoning is reasoning from symptoms to cause
Diagnostic reasoning occurs in the *opposite* direction to the network arcs.

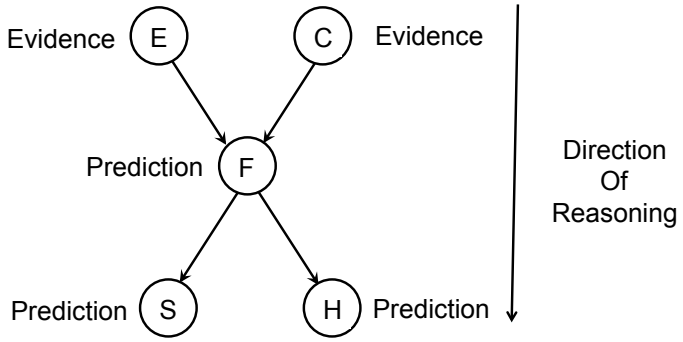
Example: A fire (F) can be caused by an electrical problem (E) or a Cigarette (C)
The fire causes smoke (S) and Heat (H).



Diagnostic Reasoning

Predictive reasoning

If we discover an electrical problem, we can predict that it caused the fire.

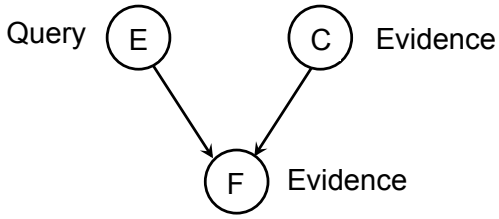


Predictive Reasoning

Note that "prediction" is not a statement about time, but about "estimation of likelihood". Predictive reasoning is reasoning from new information about causes to new beliefs about effects, following the directions of the network arcs.

Intercausal Reasoning

Intercausal reasoning involves reasoning about the mutual causes of a common effect Suppose that there are exactly two possible causes of a particular effect, represented by a v-structure in the BN.



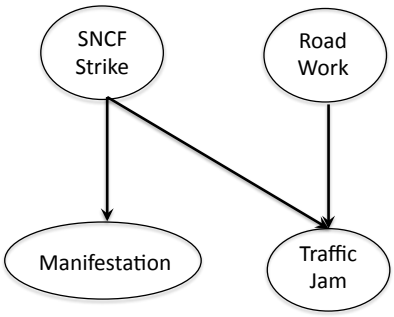
For example, a fire (F) could be caused an electrical problem (E) or a cigarette (C).

Initially these two causes are independent. Suppose that we find evidence of a smoking. This new information explains the fire, which in turn *lowers* the probability that the fire was caused by an electrical problem. Even though the two causes are initially independent, with knowledge of one cause the alternative cause is *explained away*.

The Parent nodes become dependent given information about the common effect. They are said to be conditionally dependent

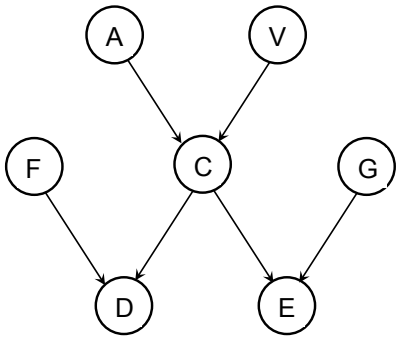
$$P(E \mid F, C) \neq P(E \mid F) \Rightarrow E \not\perp C \mid F$$

For example, suppose that you observe that there is a traffic Jam. This may be caused by a train-strike, or by roadwork. If you then discover that there is a demonstration of train workers. This confirms the trains are on strike, and explains away the possibility that roadwork has caused the traffic jam.



Markov Blanket

The Markov blanket of a node contains all the variables that shield the node from the rest of the network. This means that the Markov blanket of a node is the only knowledge needed to predict the behavior of that node. The children's parents are included, because they can be used to explain away the node in question.



MB(C)

Constructing a Bayesian Network.

Bayesian networks are generally constructed using object oriented programming. Common network patterns are coded as objects. The programmer then chains objects together. However designing a network for real problems remains somewhat of an art, and is the subject of research.

Most textbooks present only simple, pedagogically useful examples. Building real networks for practical applications is a difficult challenge.

The problems of building a complete BN for a large problems involves solving two different problems.

1) build the graph structure and

2) define the node probability tables for each node of the graph.

Building the graph structure is the hard part. This is partly because most conditional relations can be coded in several different ways. There are no obvious rules about to structure the network.

Once the network is defined, the probabilities can often be determined from statistics.

To design the network structure, researchers have assembled dictionaries of common network fragments, and expressed this using object oriented programming. These fragments are called "idioms". They represent commonly found reasoning structures.

Building a network is then reduced to assembling the objects that represent the appropriate idioms.

Five popular idioms are:

1. The Definitional/synthesis idiom
2. The Cause-Consequence Idiom
3. The Measurement idiom
4. The Induction Idiom
5. The Reconciliation Idiom