Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2016/2017

Lesson 13 24 march 2017

Reasoning with Bayesian Networks

Sources:

1. Bishop, C. "Pattern Recognition and Machine Learning (Information Science and Statistics), " *Springer, New York* (2007).
2. Koller, D., and Friedman, N., Probabilistic graphical models: principles and techniques. MIT press, 2009.
3. Kelleher, John D., Mac Namee, B, and D'Arcy A., Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press, 2015.

Naïve Bayesian Systems

Example Problem

For today's lecture we will illustrate Bayesian Reasoning with the problem of diagnosis of Lung disease.

Example Problem

A patient has been suffering from shortness of breath (called dyspnoea) and visits the doctor, worried that he has lung cancer. The doctor knows that other diseases, such as tuberculosis and bronchitis, are possible causes, as well as lung cancer. She also knows that other relevant information includes whether or not the patient is a smoker (increasing the chances of cancer and bronchitis) and what sort of air pollution he has been exposed to. A positive X-ray would indicate either Tuberculosis or lung cancer. Recent travail in Asia can increase the risk of Tuberculosis.

Probability Distribution Tables

A possible approach would be to build a histogram that tells the probability cancer given the symptoms.

For example, consider the set of Boolean random variables:
C: Cancer (Boolean) - patient has Lung Cancer,
S: Smoker: Patient smokes
X: X-Ray: Patients X-Ray is positive.
D: Dyspnoea (Shortness of breath)
B: Bronchitis - Patient has Bronchitis
P: Pollution - The atmosphere is polluted
A: Asthma - The patient has asthma
T: Tuberculosis - The patient has Tuberculosis.

Collect a sample of M patients $\{\vec{X}_m\}$ where each patient is described by the Boolean vector of variables: $\vec{X}_m = (C, S, X, D, N, P, A, T)$.

Construct a table $h(C, S, X, D, N, P, A, T)$ with $Q=2^8$ cells and count the number of times each vector (C, S, X, D, N, P, A, T) occurs in a sample population.

The table $h(C, S, X, D, N, P, A, T)$ is a probability distribution table.

A probability distribution table lists the joint outcome for the variables.

We can use the table to build a "naive Bayesian" diagnostic system.

For a probability distribution *P(A,B)*, the sum rule tells us $\sum_x P(x,B) = P(B)$

Bayes rule (Conditional probability) can be defined as

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

Consider a table constructed from C = Patient has lung Cancer, S=Patient smokes
X=Patients X-Ray is positive and D= Dyspnoea (Shortness of breath)

Thus $$P(C = T \mid S,X,D) = \frac{h(C = T,S,X,D)}{\sum_x h(x,S,X,D)} = \frac{h(C = T,S,X,D)}{h(C = T,S,X,D) + h(C = F,S,X,D)}$$

Limitations of Naïve Bayesian Systems

A probability distribution table requires one cell for each possible combination of values for all variables. For D Boolean variables, the size of the table is $Q = 2^D$.

However, random variable can take on values other than Boolean.
Variables can be:

- Ordered symbols: X ∈{low, medium, high}
- Arbitrary (unordered) Symbols: X ∈{English, French, Spanish}
- Integer Values: X ∈{1, 2, ..., 120}
- Sets of symbols: X ∈{{A,B}, {A, C}, {B, C, D}}

For real problems Q can grow very large very fast.

In this case the size of the table is the product of the number of values of the variables.

$$Q=N_1 \cdot N_2 \cdot \ldots \cdot N_D$$

Where *N* is the number of values of each random Variable.
In addition, such a table only describes correlation of observation.
It does not allow us to reason about causes and relations.
It does not allow us to explain cause and effects.

For this, we can use Bayesian Networks.

Bayesian Networks

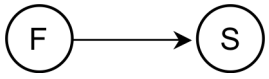
Bayesian Networks (BNs) are graphical models for reasoning under uncertainty, where the nodes represent random variables (discrete or continuous) and the arcs represent relations between variable. Arcs are often causal connections but can be other forms of association.

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. In our previous lectures these were referred to as "features".

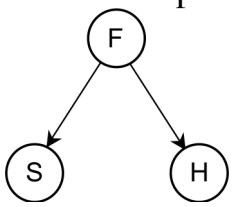
Directed arcs (or links) connect pairs of nodes, $X_1 \rightarrow X_2$, representing the direct dependencies between random variables.

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

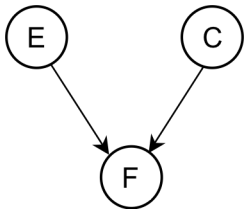


We can use graphical models to represent causal relations.

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

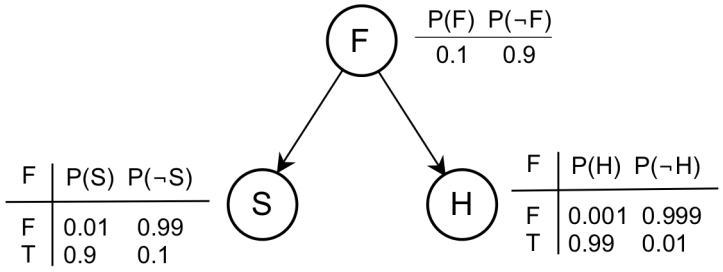


Graphical models can also express multiple possible causes for diagnostic reasoning. For example, Fire can be caused by an Electrical problem (E) or by a Cigarette (C)



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

Conditional Probability Tables.



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.

Independent Random Variables
Two random variables are Independent if

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

Formally, this is written: $A \perp B$

Independence 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

When A and B are conditionally independent given C,

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

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)*.

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

Bayesian networks explicitly express conditional independencies in probability distributions and allows computation of probabilities distributions using the chain rule.

Conditional Independence Properties.

We can identify several useful properties for conditional independence:

Symmetry: $(A \perp B \mid C) = (B \perp A \mid C)$
Decomposition: $(A \perp B, C \mid D) \Rightarrow (A \perp B \mid D)$
Weak Union: $(A \perp B, C \mid D) \Rightarrow (A \perp B \mid D, C)$

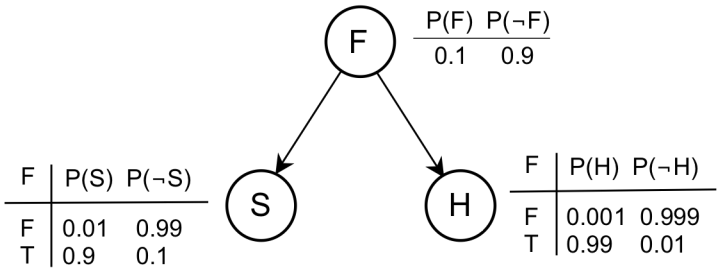
Computing with Conditional Probability Tables

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.

$$P(F, S, H) = P(S \mid F)\, P(H \mid F)\, P(F)$$

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). $F \rightarrow S$ 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)*
We get this as the product of the nodes:

$$P(F,S) = \sum_H P(F,S,H) = \sum_H P(H \mid F)P(S \mid F)P(F)$$

$$P(F,S) = P(H \mid F)P(S \mid F)P(F) + P(\neg H \mid F)P(S \mid F)P(F)$$

$$= 0.9 \text{ x } 0.99 \text{ x } 0.1 + 0.90 + 0.01 + 0.1 = 0.09$$

What is the probability of seeing Smoke?

$$P(S) = \sum_F \sum_H P(F,S,H) = \sum_F \sum_H P(H \mid F)P(S \mid F)P(F)$$

$$\begin{aligned} P(S) = {} & P(H \mid F)P(S \mid F)P(F) \\ & + P(\neg H \mid F)P(S \mid F)P(F) \\ & + P(H \mid \neg F)P(S \mid \neg F)P(\neg F) \\ & + P(\neg H \mid \neg F)P(S \mid \neg F)P(\neg F) \end{aligned}$$

$$\begin{aligned} P(S) = {} & (0.90 \cdot 0.99 \cdot 0.10) + (0.001 \cdot 0.0001 \cdot 0.10) \\ & + (0.001 \cdot 0.0001 \cdot 0.90) + (0.001 \cdot 0.9999 \cdot 0.90) \\ & = 0.0909 \end{aligned}$$

From which we have:

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

A Joint Distribution in Structured form

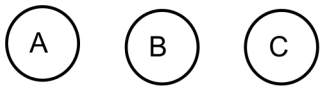
A Bayesian Network is a Joint Distribution in Structured form. The network is an Acyclic Directed Graph.

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

Node = random Variable
Directed Edge = Conditional Dependence
Absence of an Edge = Conditional Independence.

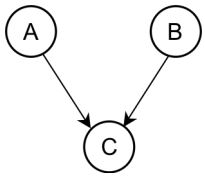
The graph shows conditional (and causal) relationships
The tables provide data for computing the probability distribution.

Marginal Independence:



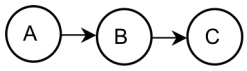
$$P(A, B, C) = P(A)\cdot P(B)\cdot P(C)$$

Independence Causes: (Common Effect)



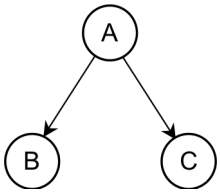
$$P(A, B, C)=P(C \mid A, B)\cdot P(A)\cdot P(B)$$

Markov Dependence (Causal Chain)



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

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 $\text{Parent}(B) = A$

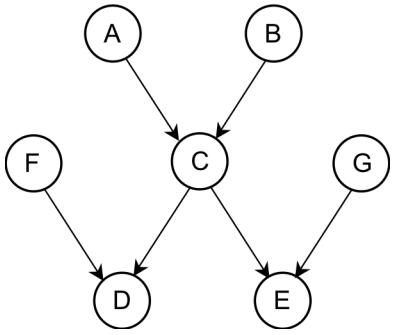
The set of all parents of a node x is the function Parents(x).

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 $A \rightarrow B \rightarrow C$

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

Markov Blanket: Parents, Children and all of Children's Parents.



MB(C)

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. (This is described below).

Bayesian Network Construction

Construction process

Networks are generally constructed by hand.
Processes for automatic construction (learning) of networks is an active research area.

To construct a Bayesian Network, a knowledge engineer must identify the relevant random variables (Features), and their possible values, Determine their dependence and causal relations, and determine the conditional probability tables. The knowledge engineer then constructs a network that captures relations between variables. She then determines the Conditional Probability Tables.

1) Identify the relevant Random Variables and their values.

The knowledge engineer must identify the relevant random variables (Features), and their possible values. The values may be Boolean, Symbolic or Discrete Numbers or even PDFs. The values of random variables must be both mutually exclusive and exhaustive. It is important to choose values that efficiently represent the domain.

2) Define the structure

The knowledge engineer then constructs a network that captures qualitative relations between variables. Two nodes should be connected directly if one affects or causes the other, with the arc indicating the direction of the effect. Causal relations are important, but other forms of correlation are possible.

The topology of the network captures qualitative relations between random variables.

Note that networks are NOT UNIQUE. Different networks can produce the same Probability Density Tables.

3) Determine the Conditional Probability Tables.

Once the network is established, we need to determine the Conditional Probability Tables (CPT)s. The tables lists the probability for each combination of values for parent nodes. Each combination of values is called an Instantiation of the parents.

For each possible instantiation, we specify (or learn) the probability of each possible value of the child.

The network can then be used for reasoning.

Reasoning with Bayesian networks

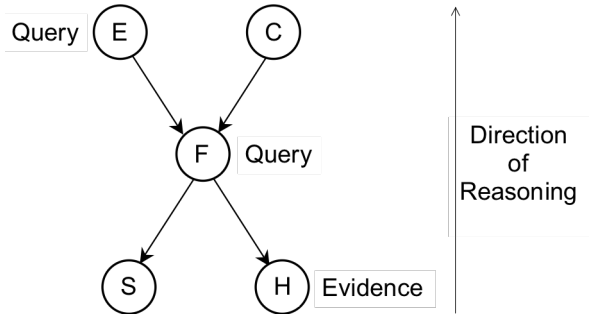
Bayesian networks provide full representations of probability distributions over their variables, and support several types of reasoning.

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.

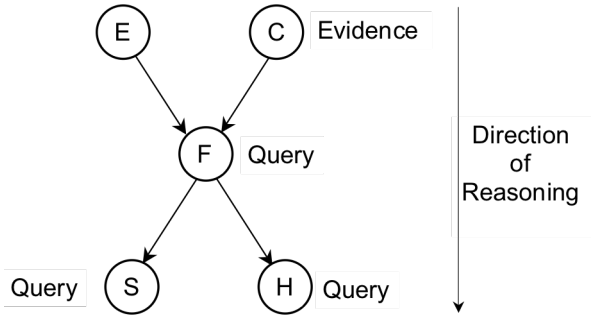
Diagnostic Reasoning
Diagnostic reasoning is reasoning from symptoms to cause
Diagnostic reasoning occurs in the *opposite* direction to the network arcs.

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.

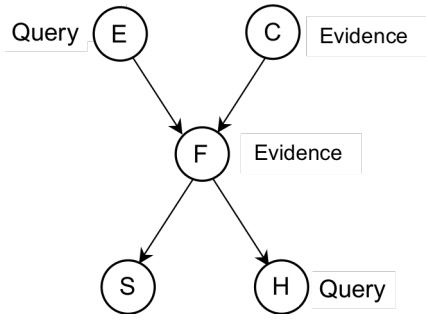


Predictive Reasoning

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

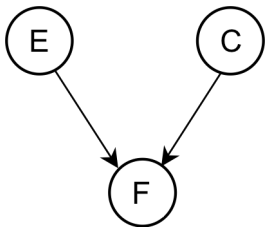
Intercausal reasoning involves reasoning about the mutual causes of a common effect



Intercausal (Explaining Away)

Explaining away

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

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

Pearl's Network Construction Algorithm

In his 1988 textbook, Judea Pearl proposed the following algorithm for constructing a Bayesian Network.

1) Choose a set of relevant variables $\{X_d\}$ that describe the problem.

2) Choose an order for the variables $[X_1, X_2, ..., X_D]$

3) For each variables X_d from d=1 to D:

a) Create a network Node for X_d.

b) determine the minimal set of previous nodes from 1 to d-1 on which X_d depends. These are the Parents of X_d : Parents(X_d).

$P(X_d \mid X_{d1},..., X_{dm}) = P(X_d \mid \text{Parents}(X_d))$

Such that $\{X_{d1},..., X_{dm}\} \subseteq \{X_1, ..., X_{d-1}\}$

c) Define the Conditional Probability Table (CPT) for X_d

Note that different node orders may result in a different network structures, with both representing the same joint probability distribution. The problem is to order the variables from Cause to Symptom so that the network representation is compact. (as few arcs as possible.)