Fundamentals of Artificial Intelligence

Bayesian Network

Shyamanta M Hazarika
Mechanical Engineering
Indian Institute of Technology Guwahati
s.m.hazarika@iitg.ac.in

http://www.iitg.ac.in/s.m.hazarika/

Representing Knowledge in an Uncertain Domain

- Joint probability distribution can answer any question about the domain.
  - Can **become intractably large** as the **number of variables grows.**
  - Specifying **probabilities for atomic events** may be very difficult; large amount of data is required from which statistical estimates are gathered.
- Bayes' rule allows unknown probabilities to be computed from known, stable ones.
  - **Conditional independence relationships** among variables can simplify the computation of query results.
  - Greatly **reduce the number of conditional probabilities** that need to be specified.

Belief Network

Definition: A belief network is a graph in which the following holds:

Represent the dependence between variables and give a concise specification of the joint probability distribution

1. A set of random variables makes up the nodes of the network.
2. A set of directed links or arrows connects pairs of nodes. The intuitive meaning of an arrow from node X to node Y is that X has a direct influence on Y.
3. Each node has a conditional probability table that quantifies the effects that the parents have on the node. The parents are all those nodes that have arrows pointing to it.
4. The graph has no directed cycles (hence is a directed, acyclic graph, or DAG).

Example – Lung Cancer Diagnosis

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 TB or lung cancer.

Kevin B. Korb and Ann E. Nicholson; Bayesian Artificial Intelligence, Second Edition Chapter 2, Pages 30-31.

Nodes and Values

What are the nodes to represent and what values can they take, or what state can they be in? For now we will consider only nodes that take discrete values. The values should be both mutually exclusive and exhaustive.

Nodes can be discrete or continuous

- **Boolean nodes –** represent propositions taking binary values
  Example: *Cancer* node represents proposition "*the patient has cancer*"
- **Ordered values**
  Example: *Pollution* node with values *low, medium, high*
- **Integral values**
  Example: *Age* with possible values 1-120

Preliminary choices: Nodes and Values

Node	Type	Values
Pollution	Binary	{*low,high*}
Smoker	Boolean	{T,F}
Cancer	Boolean	{T,F}
Dyspnoea	Boolean	{T,F}
Xray	Binary	{*pos,neg*}

Modeling choices are to be made. For example, **an alternative to representing a patient's exact age might be to clump patients into different age groups**, such as {baby, child, adolescent, young, old}.
The trick is to **choose values that represent the domain efficiently**, but with enough detail to perform the reasoning required.

Preliminary choices: Nodes and Values

Node	Type	Values
Pollution	Binary	{*low*,*high*}
Smoker	Boolean	{T,F}
Cancer	Boolean	{T,F}
Dyspnoea	Boolean	{T,F}
Xray	Binary	{*pos*,*neg*}

Choices limit what can be represented in the network. For instance, 1. There is no representation of other diseases. 2. Lack of differentiation, for example between a heavy or a light smoker. **Note that all these nodes have only two values, which keeps the model simple**, but **in general there is no limit to the number of discrete value.**

Bayesian Network Structure

The **structure, or topology, of the network** should **capture qualitative relationships** between variables.

In particular, **two nodes should be connected directly if one affects or causes the other**, with the arc indicating the direction of the effect.

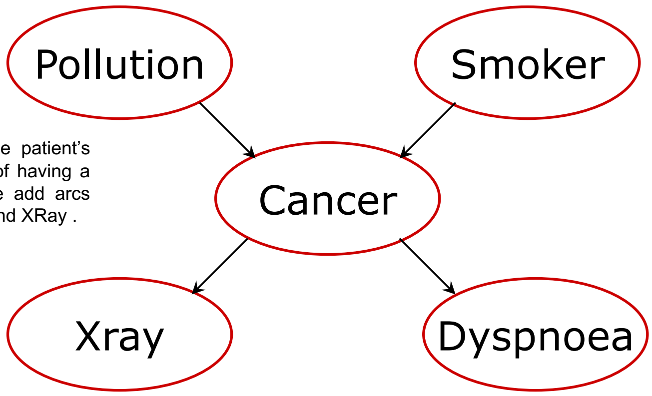
<u>For Example</u>

In the Example being discussed; What factors affect a patient's chance of having cancer? If the answer is "Pollution and smoking," then we should add arcs from Pollution and Smoker to Cancer.

Example – Lung Cancer Diagnosis

Two nodes should be connected directly if one affects or causes the other, with the arc indicating the direction of the effect

Having cancer will affect the patient's breathing and the chances of having a positive X-ray result. So we add arcs from Cancer to Dyspnoea and XRay .



It is important to note that this NETWORK is just one possible structure for the problem.

Structure Terminology

- Node is a parent of a child, if there is an arc from the former to the latter.
  - For a directed chain of nodes, one node is an ancestor of another if it appears earlier in the chain, whereas a node is a descendant of another node if it comes later in the chain.
    - Cancer node has two parents, Pollution and Smoker, while Smoker is an ancestor of both X-ray and Dyspnoea.
    - Xray is a child of Cancer and descendant of Smoker and Pollution.
- **Markov blanket** of a node consists of the node's parents, its children, and its children's parents.
- Given a **causal understanding of the structure**, the root nodes represent original causes, while leaf nodes represent final effects.
  - Causes Pollution and Smoker are root nodes, while the effects X-ray and Dyspnoea are leaf nodes.

Conditional Probability Tables (CPTs)

After specifying topology, must **specify the CPT for each discrete node**

- Each row contains the conditional probability of each node value for each possible combination of values in its parent nodes.
- Each row must sum to 1.
- A CPT for a Boolean variable with n Boolean parents contains 2^{n+1} probabilities.
- A node with no parents has one row (its prior probabilities).

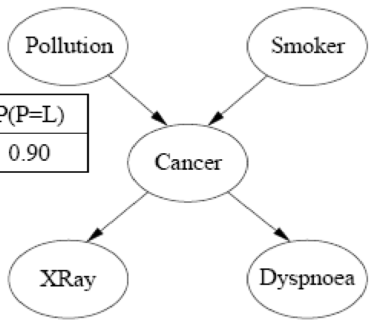
Once the topology of the BN is specified, the next step is to quantify the relationships between connected nodes. Done by specifying a conditional probability distribution for each node. As we are only considering discrete variables at this stage, this takes the form of a conditional probability table

Example – Lung Cancer Diagnosis

Root nodes also have an associated CPT, although it is degenerate, containing only one row representing its prior probabilities

For each node we need to look at all the possible combinations of values of those parent nodes. Each such combination is called an **instantiation** of the parent set.

For **each distinct instantiation of parent node values**, we need to **specify the probability** that the child will take each of its values



P(S=T)
0.30

P(P=L)
0.90

P	S	P(C=T|P,S)
H	T	0.05
H	F	0.02
L	T	0.03
L	F	0.001

C	P(X=pos|C)
T	0.90
F	0.20

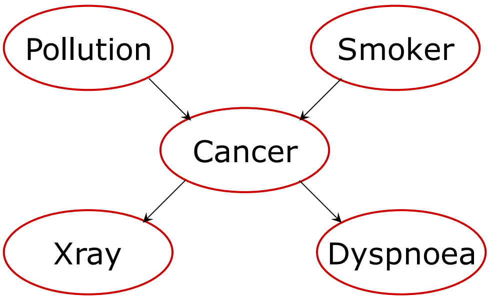
C	P(D=T|C)
T	0.65
F	0.30

If a node has many parents or if the parents can take a large number of values, the CPT can get very large!

The Markov Property

- Modeling with Bayesian Networks requires the assumption of the **Markov Property**:
  - *There are no direct dependencies in the system being modelled which are not already explicitly shown via arcs.*

Example: smoking can influence dyspnoea only through causing cancer



The Markov Property

- Bayesian networks which have the Markov property are also called **Independence-maps** (or, I-maps for short), since every independence suggested by the lack of an arc is real in the system.
- It is not generally required that the arcs in a BN correspond to real dependencies in the system.
  - The CPTs may be parameterized in such a way as to nullify any dependence.
  - Every fully-connected Bayesian network can represent, any joint probability distribution over the variables being modeled

The Markov Property

- We shall prefer minimal models and, in particular, minimal I-maps, which are I-maps such that the deletion of any arc violates I-mapness by implying a non-existent independence in the system.
- If, in fact, **every arc in a Bayesian network happens to correspond to a direct dependence** in the system, then the Bayesian Network is said to be a **Dependence-map** (or, D-map for short).
- A Bayesian network which is both an I-map and a D-map is said to be a **perfect map**.

Reasoning with Bayesian Networks

We have discussed how a domain and its uncertainty may be represented in a Bayesian network; NEXT we look at how to use the Bayesian network to reason about the domain.

- The basic task for any probabilistic inference system:

Compute the posterior probability distribution for a set of ***query variables***, given new information about some ***evidence variables.***

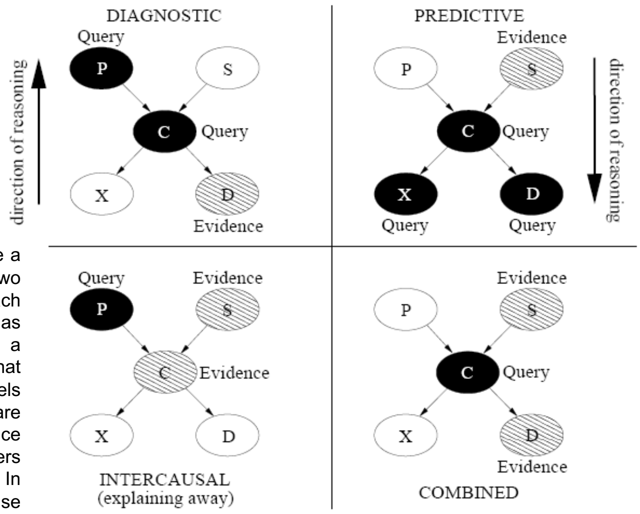
When we observe the value of some variable, we would like to condition upon the new information. The process of conditioning (also called probability propagation or inference or belief updating) is performed via a “flow of information” through the network

- Also called **conditioning** or **belief updating** or **inference**.

Types of Reasoning

Reasoning from symptoms to cause, such as when a doctor observes Dyspnoea and then updates his belief about Cancer and whether the patient is a Smoker. This reasoning occurs in the opposite direction to the network arcs.

Smoker and Pollution which have a common effect, Cancer These two causes are independent of each other; We learn that Mr. Smith has cancer. We discover that he is a smoker. Lowers the probability that he has been exposed to high levels of pollution. The two causes are initially independent; the presence of one explanatory cause renders an alternative cause less likely. In other words, the alternative cause has been explained away.



Reasoning from new information about causes to new beliefs about effects, following the directions of the network arcs. For example, the patient may tell his physician that he is a smoker; the physician knows this will increase the chances of the patient having cancer

Since any nodes may be query nodes and any may be evidence nodes, sometimes the reasoning does not fit neatly into one of the types described above. Indeed, we can combine the above types of reasoning in any way.

Types of Evidence

- **Evidence**

  Bayesian networks can be used to calculate new beliefs when new information – which we call evidence – is available.

- **Specific Evidence**

  Evidence as a definite finding that a node X has a particular value, x, which we write as X = x.

  Example Smoker = True.

- **Negative Evidence**

  Evidence might be that Y is not in state y_1 (but may take any of its other values).

- **Likelihood Evidence**

  New information - new probability distribution over Y

Understanding Bayesian Networks

We now consider how to interpret the information encoded in a BN — the probabilistic semantics of Bayesian networks

- A (more compact) representation of the joint probability distribution
  - There is a useful underlying structure to the problem being modeled that can be captured with a BN.
- Bayesian networks which satisfy the Markov property explicitly encode conditional independence statements
  - understand how to design inference procedures

    via *Markov property:*
    Each conditional independence implied by the graph is present in the probability distribution.

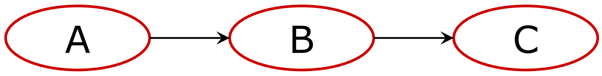
Conditional Independence

The relationship between **conditional independence** and **Bayesian network structure** is important for understanding how Bayesian networks work.

1. Causal Chains
2. Common Causes
3. Common Effects - Conditional dependence.
4. D-separation

Causal Chains

- Causal chains give rise to conditional independence



$$P(C \mid A \wedge B) = P(C \mid B)$$

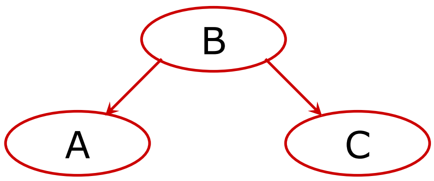
The probability of C, given B, is exactly the same as the probability of C, given both B and A. Knowing that A has occurred doesn't make any difference to our beliefs about C if we already know that B has occurred.

- Example: Smoking causes cancer, which causes dyspnoea

Probability that someone has dyspnoea depends directly only on whether they have cancer. If we don't know whether some woman has cancer, but we do find out she is a smoker, that would increase our belief both that she has cancer and that she suffers from shortness of breath. However, if we already knew she had cancer, then her smoking wouldn't make any difference to the probability of dyspnoea. That is, dyspnoea is conditionally independent of being a smoker given the patient has cancer

Common Causes

- Common Causes (or ancestors) also give rise to conditional independence



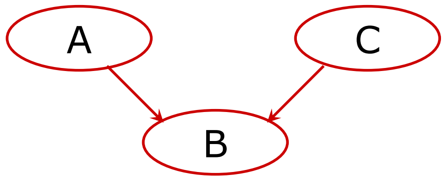
$$P(C \mid A \wedge B) = P(C \mid B) \equiv A \,\mathrm{indep}\, C \mid B$$

If there is no evidence or information about cancer, then learning that one symptom is present will increase the chances of cancer which in turn will increase the probability of the other symptom. However, if we already know about cancer, then an additional positive X-ray won't tell us anything new about the chances of dyspnoea.

- Example: Cancer is a common cause of the two symptoms: a positive X-ray and dyspnoea

Common Effects

- Common effects (or their descendants) give rise to conditional *dependence*



$$P(A \mid C \wedge B) \neq P(A)P(C) \equiv \neg(A \,\text{indep}\, C \mid B)$$

Common effects (or their descendants) produce the exact opposite conditional independence structure to that of chains and common causes. That is, the parents are marginally independent, but become dependent given information about the common effect.

- Example: Cancer is a common effect of pollution and smoking
  Given cancer, smoking “explains away” pollution

D-separation

- Graphical criterion of conditional independence.

We have seen how Bayesian networks represent conditional independencies and how these independencies affect belief change during updating.

These concepts apply not only between pairs of nodes, but also between sets of nodes.

- We can **determine whether a set of nodes X is independent of another set Y**, given a set of evidence nodes E, via the Markov property
  - If every undirected path from a node in X to a node in Y is **d-separated** by E, then X and Y are **conditionally independent** given E

Determining D-separation

<u>Definition</u>: A **Path (Undirected Path)** between two sets of nodes X and Y is any **sequence of nodes** between a member of X and a member of Y such that **every adjacent pair of nodes is connected by an arc** (regardless of direction) and **no node appears in the sequence twice**.

<u>Definition</u>: A path is a **blocked path**, given a set of nodes E, if there is a node Z on the path for which **at least one of three conditions holds**:

1. Z is in E and Z has one arc on the path leading in and one arc out (chain).
2. Z is in E and Z has both path arcs leading out (common cause).
3. Neither Z nor any descendant of Z is in E, and both path arcs lead in to Z (common effect).

Determining D-separation

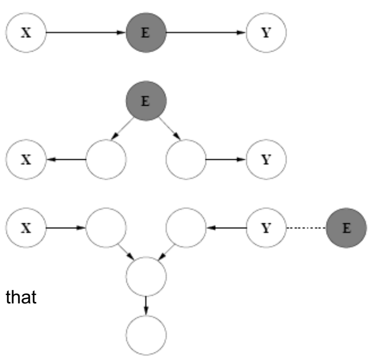
<u>Definition</u>: A set of nodes E d-separates two other sets of nodes X and Y if every path from a node in X to a node in Y is blocked given E.

If X and Y are d-separated by E, then X and Y are conditionally independent given E (given the Markov property)

Chain

Common cause

Common effect



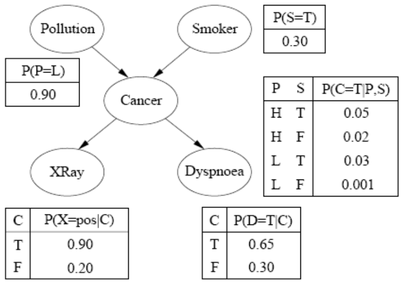
Examples of these three blocking situations are shown. Note that we have simplified by using single nodes rather than sets of nodes; also note that the evidence nodes E are shaded.

Example – Lung Cancer Diagnosis

d-separation

Evidence

Observation of the Cancer node.



1. P is d-separated from X and D. Likewise, S is d-separated from X and D (Condition 1).
2. While X is d-separated from D (Condition 2).
3. However, if C had not been observed (and also not X or D), then S would have been d-separated from P (Condition 3).

Example – Burglar Alarm

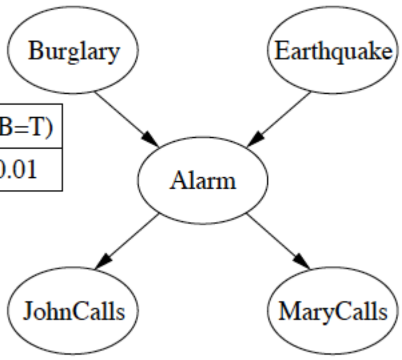
You have a new burglar alarm installed at home. It is fairly reliable at detecting a **burglary**, but also responds on occasion to minor **earthquakes**. You also have two neighbors, **John** and **Mary**, who have **promised to call** you at work when they hear the alarm. John always calls when he hears the **alarm**, but *sometimes confuses the telephone ringing with the alarm* and calls then, too. Mary, on the other hand, *likes rather loud music and sometimes misses the alarm altogether*. Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

Stuart J. Russel and Peter Norvig: Artificial Intelligence – A Modern Approach, Chapter 15, Pages 437-439.

Example – Burglar Alarm

- **All the nodes in this BN are Boolean**, representing the true/false alternatives for the corresponding propositions.
- This **BN models the assumptions**
  - John and Mary **do not perceive a burglary directly**
  - They **do not feel minor earthquakes**.
- There is **no explicit representation of loud music** preventing Mary from hearing the alarm, **nor of John's confusion of alarms and telephones**.
  - This information is summarized in the probabilities in the arcs from Alarm to JohnCalls and MaryCalls.

Example – Burglar Alarm



P(B=T)
0.01

P(E=T)
0.02

B	E	P(A=T|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J=T|A)
T	0.90
F	0.05

A	P(M=T|A)
T	0.70
F	0.01

No explicit representation of loud music preventing Mary from hearing the alarm, nor of John's confusion of alarms and telephones; this information is summarized in the probabilities in the arcs from Alarm to JohnCalls and MaryCalls .

Bayesian Networks – Summary

- Bayes rule **allows unknown probabilities to be computed** from known ones.
- **Conditional independence** (due to causal relationships) allows **efficient updating**.
- Bayesian networks are a **natural way to represent conditional independence** info
  - qualitative: links between nodes
  - quantitative: conditional probability tables (CPTs)
- Bayesian network inference
  - **computes the probability of query variables given evidence variables.**
  - is flexible – we can **enter evidence about any node and update beliefs in other nodes.**