Lecture 11: Uncertainty - 2

Victor Lesser

CMPSCI 683
Fall 2004

Review of Key Issues *with respect to* Probability Theory

- Uncertainty arises because of both laziness and ignorance.
  - **inescapable in complex, dynamic, or inaccessible worlds.**
- Uncertainty means that many of the simplifications that are possible with deductive inference are no longer valid
  - **lack of modularity.**
- Probabilities express the agent's inability to reach a definite decision regarding the truth of a sentence,
  - **summarize the agent's degree of belief.**

Review of Key Issues *with respect to* Probability Theory

- Basic probability statements include prior probabilities and conditional probabilities over simple and complex propositions.
  - **Product rule, Marginalization(summing out) and conditioning**
- The axioms of probability specify constraints on reasonable assignments of probabilities to propositions.
  - **An agent that violates the axioms will behave irrationally in some circumstances.**
- The joint probability distribution specifies the probability of each complete assignment of values to random variables
  - **It is usually far too large to create or use.**

Bayes' Rule

P(A,B,C,D,..) = P(A|B,C, D,..) P(B,C, D,..) ; product rule
P(A,B) = P(A|B)P(B) = P(B|A)P(A)

Thus, Bayes' Rule: $P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$

This allows us to compute a conditional probability from its inverse.

E.g., $P(disease \mid symptom) = \frac{P(symptom \mid disease)P(disease)}{P(symptom)}$

Bayes' rule is typically written as: P(***B*** **|** ***A***)=αP(***A*** **|** ***B***)P(***B***)

(α **is the n**ormalization constant **needed to make the** P(***B*** **|** ***A***) **entries sum to 1, it eliminates the need to know** P(***A***) **)**

Why is Bayes' Rule Useful?

- **P(object | image) proportional to:**
  P(image | object) P(object)
- **P(sentence | audio) proportional to:**
  P(audio | sentence) P(sentence)
- **P(fault | symptoms) ...**
  P(symptoms | fault) P(fault)

Basis of Abductive Inference -- From Casual Knowledge to Diagnostic Knowledge!!

Combining evidence

- Consider a diagnosis problem with multiple symptoms:
  $P(d|s_i,s_j) = P(d)P(s_i,s_j|d)/P(s_i,s_j)$
- For each pair of symptoms, we need to know $P(s_i,s_j|d)$ and $P(s_i,s_j)$. Large amount of data is needed.
- Need to make **independence assumptions**:
  $P(s_i|s_j) = P(s_i) \text{ -> } P(s_i,s_j)= P(s_i)P(s_j)$;

Or **conditional independence assumptions**:
  $P(s_i|s_j,d) = P(s_i|d) \quad P(s_i,s_j|d) = P(s_i|d)\, P(s_j|d)$
  implicitly d causes s_i and s_j
- With conditional independence, Bayes' rule becomes:
  $\mathbf{P}(Z|X,Y) = \alpha\, \mathbf{P}(Z)\, \mathbf{P}(X|Z)\, \mathbf{P}(Y|Z)$

Causal vs. Diagnostic Knowledge

S =patient has a stiff neck
M =patient has meningitis

P(S/M) = .5
P(M) = 1/50,000
P(S) = 1/20

$$P(M/S) = \frac{P(S/M)P(M)}{P(S)} = \frac{.5 \times 1/50{,}000}{1/20} = .0002$$

Suppose given only P(M/S) based on actual observation of data… what happens if there is a sudden outbreak of meningitis:

⇒ P(M) goes up significantly
P(S/M) not affected

"Diagnostic knowledge is often more tenuous than Causal knowledge."

Combining evidence

- Consider a diagnosis problem with multiple symptoms:
  $P(d|s_i,s_j) = P(d)\, P(s_i,s_j|d)/P(s_i,s_j)$
- For each pair of symptoms, we need to know $P(s_i,s_j|d)$ and $P(s_i,s_j)$. Large amount of data is needed.
- Suppose we make **independence assumptions**:
  $P(s_i|s_j) = P(s_i); \quad P(s_i,s_j)= P(s_i|s_j)P(s_j)= P(s_i)P(s_j)$
- Or **conditional independence assumptions**:
  $P(s_i|s_j,d) = P(s_i|d); \quad P(s_i,s_j|d) = P(s_i|\, s_j,d)\, P(s_j|d)= P(s_i|d)\, P(s_j|d)$
- With conditional independence, Bayes' rule becomes:
  $P(d|s_i,s_j) = \alpha\, P(d)\, P(s_i|d)\, P(s_j|d)$

Bayes' Rule: Incremental Evidence Accumulation

Probabilistic inference involves computing probabilities that are not explicitly stored by the reasoning system.

P(hypothesis* | *evidence) **is a common value we want, and we want to compute this incrementally as evidence accumulates.**

possible with conditional independence

$\mathrm{P}(H \mid E_1,E_2) = \alpha \mathrm{P}(E_2 \mid H)\mathrm{P}(E_1 \mid H)\mathrm{P}(H)$

[$\mathrm{P}(E_1 \mid H)\mathrm{P}(H)$ **is just the belief based on** E_1]

Review of Key Issues *with respect to* Baye' Rule

- Bayes' rule allows unknown probabilities to be computed from known, stable ones.
- In the general case, combining many pieces of evidence may require assessing a large number of conditional probabilities.
- Conditional independence brought about by direct causal relationships in the domain allows Bayesian updating to work effectively even with multiple pieces of evidence.

Probabilistic reasoning

- **Can be performed using the joint probability distribution:**

$$\mathbf{P}(X|\mathbf{e}) = \alpha\,\mathbf{P}(X,\mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(X,\mathbf{e},\mathbf{y})$$

- **Problem: How to represent the joint probability distribution compactly to facilitate inference.**
- **We will use a belief network as a data structure to represent the conditional independence relationships between the variables in a given domain.**


```
function ENUMERATE-JOINT-ASK(X, e, P) returns a distribution over X
  inputs: X, the query variable
          e, observed values for variables E
          P, a joint distribution on the variables {X} ∪ E ∪ Y

  Q(X) ← a distribution over X, initially empty
  for each value x_i of X do
     Q(x_i) ← ENUMERATEJOINT(x_i, e, Y, [], P)
  return NORMALIZE(Q(X))
```

```
function ENUMERATE-JOINT(x, e, vars, values, P) returns a real number
  if EMPTY?(vars) then return P(x, e, values)
  Y ← FIRST(vars)
  return Σ_y ENUMERATE-JOINT(x, e, REST(vars), [y|values], P)
```

Figure 13.4 An algorithm for probabilistic inference by enumeration of the entries in a full joint distribution.

Belief Networks

A major advance in making probabilistic reasoning systems practical for AI has been the development of belief networks **(also called** Bayesian/probabilistic networks**).**

The main purpose of the belief network is to encode the *conditional independence relations* in a domain.

- real domains have a lot of structure

This makes it possible to specify a complete probabilistic model using far fewer (and more natural/available) probabilities while keeping probabilistic interference tractable.

- **Considered one of the major advances in AI**
  - puts diagnostic and classification reasoning on a firm theoretical foundation
  - makes possible large applications

Joint vs. Conditional Probabilities

Traditionally, probabilistic models are defined using the joint.

Conditional probabilities are then defined in terms of the joint:

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

Note that specifying the joint can require a huge number of probabilities:

2^n for n Boolean random variables.

The Bayesian/subjectivist movement in AI views the conditional probabilities as more basic (and more compatible with human knowledge).

Conditional Independence

In addition, in most domains there are independence relations that make it possible to specify the joint more compactly with conditional probabilities:

$$\mathbf{P}(A \mid B, C) = P(A \mid C)$$

A is conditionally independent of B given C

The product rule**:**

$$P(A, B) = P(A \mid B)\, P(B) \quad (or\ \mathbf{P}(A, B) = \mathbf{P}(B \mid A)\, \mathbf{P}(A))$$

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

conditional independence $\Rightarrow$ $P(A \mid C)$ $\Rightarrow$ reduces tables

Belief (or Bayesian) networks

- **Set of nodes, one per variable**
- **Directed acyclic graph (DAG): link represents “direct” influence**
- **Conditional probability tables (CPTs): P(Child | Parent$_1$, ..., Parent$_n$)**

Earthquake example (Pearl)

Suppose that 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.

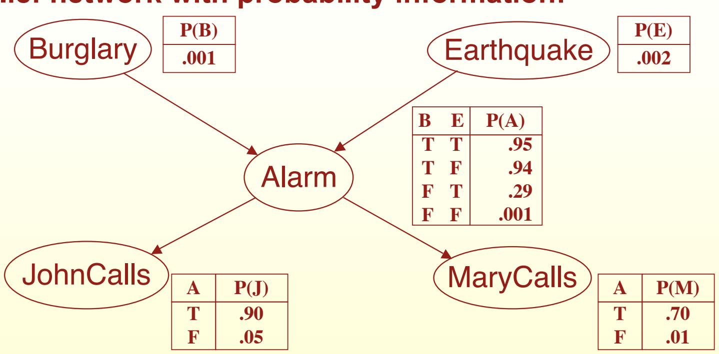
Conditional probability tables

Burglary	Earthquake	P(A=True | B,E)	P(A=False | B,E)
True	True	0.950	0.050
True	False	0.940	0.060
False	True	0.290	0.710
False	False	0.001	0.999

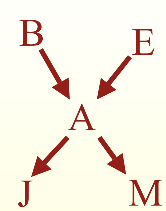
How much data is needed to represent a particular problem? How can we minimize it?

Earthquake Example, *Cont'd*

Belief network with probability information:



Earthquake example cont.



Priors: P(B), P(E)

CPTs: P(A|B,E),

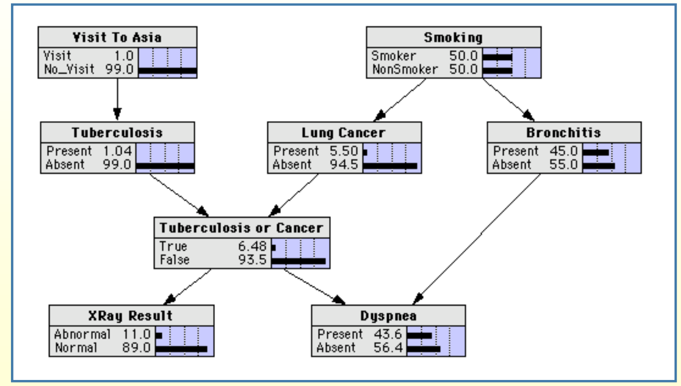
P(J|A), P(M|A)

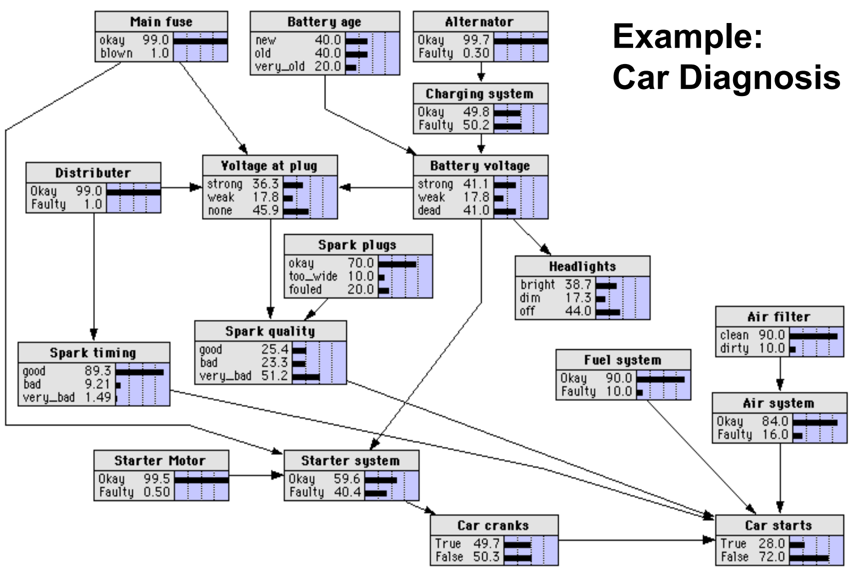
10 parameters in Belief Network but 31 parameters in the 5-variable Joint Distribution

Ignorance /Laziness in Example

- **Not included**
  - Mary is currently listening to music
  - telephone ringing and confusing John
- **Factor summarized in**
  - Alarm → John calls
  - Alarm → Mary calls
- **Approximating Situation**
  - eliminating hard-to-get information
  - reducing computational complexity

Chest clinic example





Example: Car Diagnosis

The semantics of belief networks

- Any joint can be decomposed into a product of conditionals:
  $P(X_1, X_2, ..., X_n) = P(X_n|X_{n-1}, ...,X_1)P(X_{n-1}, ...,X_1) = \Pi\ P(X_i|X_{i-1}, ..., X_1)$
- Value of belief networks is in "exposing" conditional independence relations that make this product simpler:
  $P(X_1, X_2, ..., X_n) = \Pi\ P(X_i \mid Parents(X_i))$

Conditional independence in BNs

- Each node is conditionally independent of its non-descendants, given its parents.
  - **Says nothing about other dependencies**
- Causality is intricately related to conditional independence.
- Conditional independence is one type of knowledge that we use.

d-separation: Direction-Dependent Separation

- **Network construction**
  - Conditional independence of a node and its predecessors, given its parents
  - The absence of a link between two variables does not guarantee their independence
- **Effective inference needs to exploit all available conditional independences**
  - Which set of nodes X are conditionally independent of another set Y, given a set of evidence nodes E
    - P(X,Y/E) = P(X/E) · P(Y/E)
  - Limits propagation of information
  - Comes directly from structure of network

d-separation

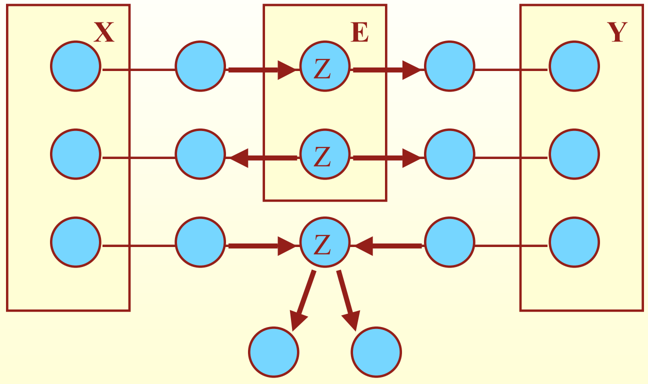
Definition: If X, Y and E are three disjoint subsets of nodes in a DAG, then E is said to d-separate X from Y if every undirected path from X to Y is blocked by E. A path is blocked if it contains a node Z such that:

(1) Z has one incoming and one outgoing arrow; or

(2) Z has two outgoing arrows; or

(3) Z has two incoming arrows and neither Z nor any of its descendants is in E.

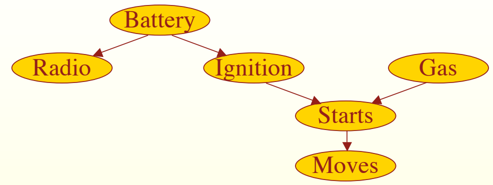
d-separation cont.



d-separation cont.

- **Property of belief networks**: if X and Y are d-separated by E, then X and Y are conditionally independent given E.
- An "if-and-only-if" relationship between the graph and the probabilistic model cannot always be achieved.

d-separation example



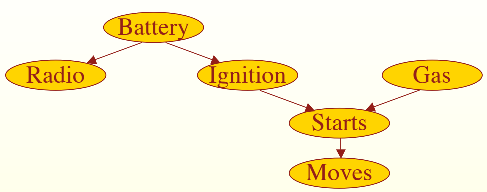
Whether there is *Gas* in the car and whether the car *Radio* plays are independent given evidence about whether the *SparkPlugs* fire (case 1).

P(R,G/I) = P(R/I) · P(G/I)
P(G/I,R) = P(G/I)

Gas and *Radio* are independent if it is known if the battery works (case2).

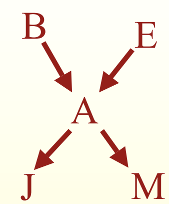
P(R/B,G) = P(R/B)

d-separation example:



. *Gas* and *Radio* are independent given no evidence at all. But they are dependent given evidence about whether the car *Starts*. For example, if the car does not start, then the radio playing is increased evidence that we are out of gas. *Gas* and *Radio* are also dependent given evidence about whether the car *Moves*, because that is enabled by the car starting.

Earthquake example revisited



- **Suppose you need:**
  P(J,E) = Σ P(J,m,a,b,E)
- **P(J,m,a,b,E) =**
  P(J|m,a,b,E) P(m|a,b,E) P(a|b,E) P(b|E) P(E)
- **Conditional independence saves us:**
  P(J,m,a,b,E) = P(J|a) P(m|a) P(a|b,E) P(b) P(E)

Representation of Conditional Probability Tables

- **Canonical distributions**
- **Deterministic nodes**
  - No uncertainty in decision
    - If $x_1=a$ and $x_2=b \Rightarrow x_3=c$
- **Noisy - OR**
  - Generalization of logical/OR
  - Each cause has an independent chance of causing the effect
  - All possible causes are listed
    - Add "miscellaneous cause"
  - Inhibition of causality independent among causes
  - $O(k)$ vs $O(2^k)$ parameters need to specify $P(H/C_i)$
  - $P(\sim H/C_1, \ldots C_n)$ = product of $(1-P(H/C_i))$ for all C_i=T

Example of Noisy-OR

P(Fever/Cold) = .4
P(Fever/Flu) = .8
P(Fever/Malaria) = .9

Cold	*Flu*	*Malaria*	*P(Fever)*	*P(¬Fever)*
F	F	F	0.0	1.0
F	F	T	0.9	0.1
F	T	F	0.8	0.2
F	T	T	0.98	0.02=0.2 x 0.1
T	F	F	0.4	0.6
T	F	T	0.94	0.06=0.6 x 0.1
T	T	F	0.88	0.12 = 0.6 x 0.2
T	T	T	0.988	0.012 = 0.6 x 0.2 x 0.1

Benefits of belief networks

- Individual "design" decisions are understandable: causal structure and conditional probabilities.
- BNs encode conditional independence, without which probabilistic reasoning is hopeless.
- Can do inference even in the presence of missing evidence.

Constructing belief networks

Loop:

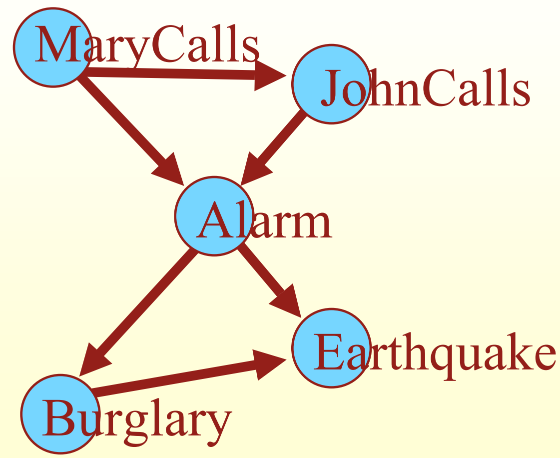
- **Pick a variable X_i to add to the graph.**
- **Find (minimal)set of parents such that $P(X_i|Parents(X_i)) = P(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)$ or $I(X_i,U_i\text{-}Parents(X_i)|Parents(X_i))$.**
- **Draw arcs from $Parents(X_i)$ to X_i.**
- **Specify the CPT: $P(X_i|Parents(X_i))$.**

Constructing belief networks cont.

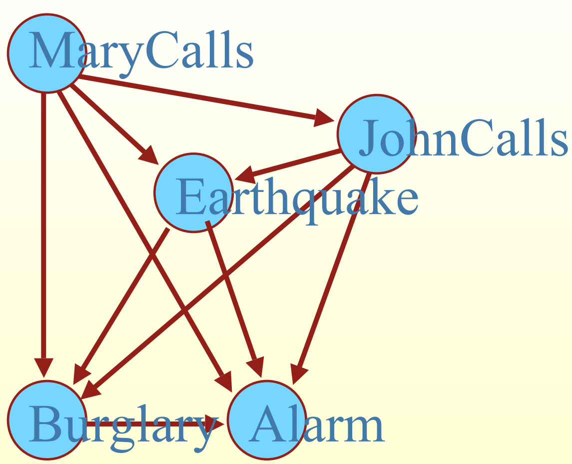
Properties of the algorithm:

- Graph is always acyclic.
- No redundant information => consistency with the axioms of probability.
- Network structure/compactness depends on the ordering of the variables.

Example: Ordering M,J,A,B,E



Example: Ordering M,J,E,B,A



Next Lecture

- Inference in Belief Networks
- Belief propagation
- Approximate inference techniques
- Alternative approaches to uncertain reasoning