Artificial Intelligence (CS303)

Lecture 8: Representing Uncertainty

Hints for this lecture

- **An agent can seldom precisely knows the state, knowledge should be represented such that wise decisions/actions can still be made.**

Outline of this lecture

- **Uncertainty and Rational Decisions**
- **Basic Probability Theory and Its Use**
- **Bayesian Network**

I. Uncertainty and Rational Decisions

The World is Uncertain

- We never know what "state" we are in exactly, because the world is only partially observable (to us).

- Agents may encounter similar situations in a real-world AI task.

- Maintaining the belief states is a possible approach, but impractical
  - Combinatorial explosion
  - An agent may not do multiple (possibly conflict actions) in one state.

Alternative

- Utility theory: Assign utility to each state/actions
- Probability theory: Summarize the uncertainty associated with each state
- Rational Decisions: Maximize the expected utility (Probability + Utility)
- Thus we need to represent states in a language of probability.

II. Basic Probability Theory and Its Use

Basic Probability Theory and Its Use

- **Joint probability distribution specifies probability of every atomic event.**
- **Queries can be answered by summing over atomic events.**

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

Basic Probability Theory and Its Use

Prior probability

Prior or unconditional probabilities of propositions
e.g., $P(Cavity = true) = 0.1$ and $P(Weather = sunny) = 0.72$
correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:
$\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)

Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)
$\mathbf{P}(Weather, Cavity) =$ a 4×2 matrix of values:

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

Every question about a domain can be answered by the joint distribution because every event is a sum of sample points

Basic Probability Theory and Its Use

Conditional probability

Conditional or posterior probabilities
e.g., $P(cavity|toothache) = 0.8$
i.e., **given that** *toothache* **is all I know**
NOT "if *toothache* then 80% chance of *cavity*"

(Notation for conditional distributions:
$\mathbf{P}(Cavity|Toothache)$ = 2-element vector of 2-element vectors)

If we know more, e.g., *cavity* is also given, then we have
$P(cavity|toothache, cavity) = 1$
Note: the less specific belief **remains valid** after more evidence arrives, but is not always **useful**

New evidence may be irrelevant, allowing simplification, e.g.,
$P(cavity|toothache, 49ersWin) = P(cavity|toothache) = 0.8$
This kind of inference, sanctioned by domain knowledge, is crucial

Basic Probability Theory and Its Use

Conditional probability

Definition of conditional probability:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0$$

Product rule gives an alternative formulation:

$$P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$$

A general version holds for whole distributions, e.g.,

$$\mathbf{P}(Weather, Cavity) = \mathbf{P}(Weather|Cavity)\mathbf{P}(Cavity)$$

(View as a 4×2 set of equations, **not** matrix mult.)

Chain rule is derived by successive application of product rule:

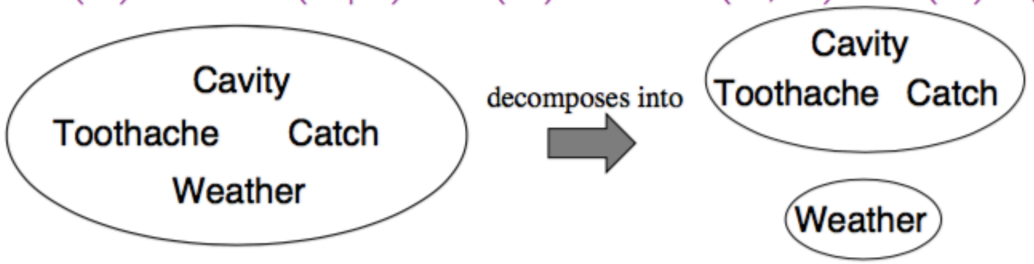
$$\begin{aligned}\mathbf{P}(X_1, \ldots, X_n) &= \mathbf{P}(X_1, \ldots, X_{n-1})\, \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\ &= \mathbf{P}(X_1, \ldots, X_{n-2})\, \mathbf{P}(X_{n_1}|X_1, \ldots, X_{n-2})\, \mathbf{P}(X_n|X_1, \ldots, X_{n-1}) \\ &= \ldots \\ &= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1})\end{aligned}$$

Basic Probability Theory and Its Use

Independence

A and B are independent iff

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



$$\mathbf{P}(Toothache, Catch, Cavity, Weather)$$
$$= \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$$

32 entries reduced to 12; for n independent biased coins, $2^n \to n$

Absolute independence powerful but rare

Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

Basic Probability Theory and Its Use

Bayes' Rule

Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$\Rightarrow$ Bayes' rule $P(a|b) = \dfrac{P(b|a)P(a)}{P(b)}$

or in distribution form

$$\mathbf{P}(Y|X) = \frac{\mathbf{P}(X|Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\mathbf{P}(X|Y)\mathbf{P}(Y)$$

Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = \frac{P(Effect|Cause)P(Cause)}{P(Effect)}$$

Basic Probability Theory and Its Use

Bayes' Rule and conditional independence

$$\begin{aligned}&\mathbf{P}(Cavity|toothache \wedge catch)\\ &= \alpha\,\mathbf{P}(toothache \wedge catch|Cavity)\mathbf{P}(Cavity)\\ &= \alpha\,\mathbf{P}(toothache|Cavity)\mathbf{P}(catch|Cavity)\mathbf{P}(Cavity)\end{aligned}$$

This is an example of a naive Bayes model:

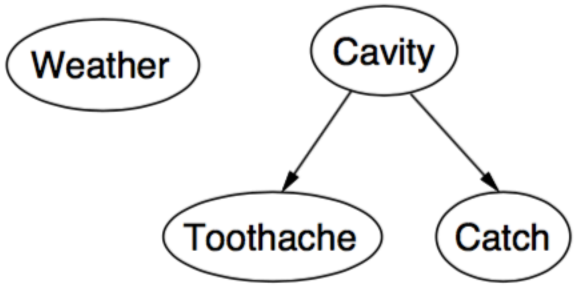
$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause)\Pi_i\mathbf{P}(Effect_i|Cause)$$

Total number of parameters is **linear** in n

III. Bayesian Networks

What is a BN?

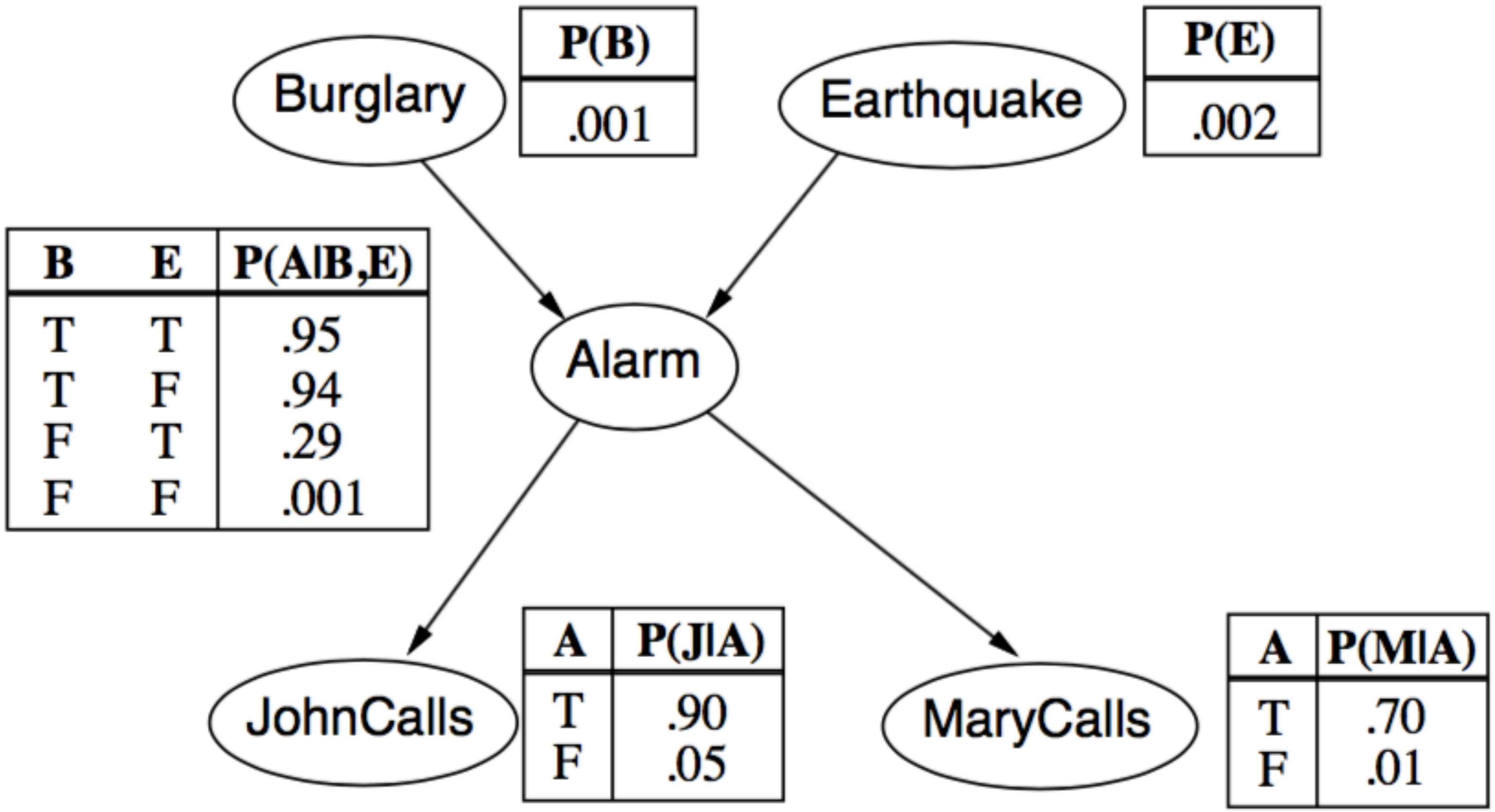
- **A Directed Acyclic Graph (DAG).**
- **Each node is a random variable, associated with conditional distribution.**
- **Each arc (link) represent direct influence of a parent node to a child node.**



$Weather$ is independent of the other variables

$Toothache$ and $Catch$ are conditionally independent given $Cavity$

What is a BN?


Burglary
P(B)
.001
Earthquake
P(E)
.002
B E P(A|B,E)
T T .95
T F .94
F T .29
F F .001
Alarm
JohnCalls
A P(J|A)
T .90
F .05
MaryCalls
A P(M|A)
T .70
F .01

Why BN?

- **More compact representation.**

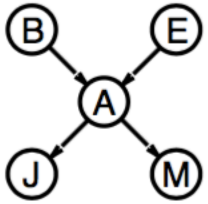
A CPT for Boolean X_i with k Boolean parents has
2^k rows for the combinations of parent values

Each row requires one number p for $X_i = true$
(the number for $X_i = false$ is just $1 - p$)

If each variable has no more than k parents,
the complete network requires $O(n \cdot 2^k)$ numbers

I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)



- **Easier to utilize independence and conditional dependence relations to define the full joint distribution.**

How to construct a BN?

1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n

   add X_i to the network

   select parents from $X_1, \ldots, X_{i-1}$ such that

   $$\mathbf{P}(X_i | Parents(X_i)) = \mathbf{P}(X_i | X_1, \ldots, X_{i-1})$$

This choice of parents guarantees the global semantics:

$$\begin{aligned} \mathbf{P}(X_1, \ldots, X_n) &= \Pi_{i=1}^{n} \mathbf{P}(X_i | X_1, \ldots, X_{i-1}) \quad \text{(chain rule)} \\ &= \Pi_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i)) \quad \text{(by construction)} \end{aligned}$$

To be continued