TDDC17

Seminar 8
Reasoning with Uncertainty
Bayesian Networks

Artificial Intelligence & Integrated Computer Systems Division
Department of Computer and Information Science
Linköping University, Sweden

Seminar Outline

- Basic Probability Theory from a logical perspective.
- Bayesian Networks
  - An "efficient" means of doing probabilistic reasoning.

Propositional Logic and Models

Table 1: Propositional Truth Tables and Models

	Cavity (Cav)	Toothache (Too)	Catch (Cat)	$Cav \vee Too$	$Cav \rightarrow Too$	$\neg Too$
1	T	T	T	T	T	F
2	T	T	F	T	T	F
3	T	F	T	T	F	T
4	T	F	F	T	F	T
5	F	T	T	T	T	F
6	F	T	F	T	T	F
7	F	F	T	F	T	T
8	F	F	F	F	T	T

DNF Characterization of Models

Any propositional formula can be equivalently represented in DNF form based on its truth table characterization:

For example:

$$Cav \vee Too \equiv 1 \vee 2 \vee 3 \vee 4 \vee 5 \vee 6$$

$$\begin{aligned} Cav \vee Too \quad \equiv \quad & \\ & (Cav \wedge Too \wedge Cat) \vee (Cav \wedge Too \wedge \neg Cat) \vee (Cav \wedge \neg Too \wedge Cat) \vee \\ & (Cav \wedge \neg Too \wedge \neg Cat) \vee (\neg Cav \wedge Too \wedge Cat) \vee (\neg Cav \wedge Too \wedge \neg Cat) \vee \\ & (\neg Cav \wedge \neg Too \wedge Cat) \vee (\neg Cav \wedge \neg Too \wedge \neg Cat) \end{aligned}$$

Observe that:

$$True \equiv 1 \vee 2 \vee 3 \vee 4 \vee 5 \vee 6 \vee 7 \vee 8$$

$$False \equiv \neg True$$

Degrees of Truth/Belief

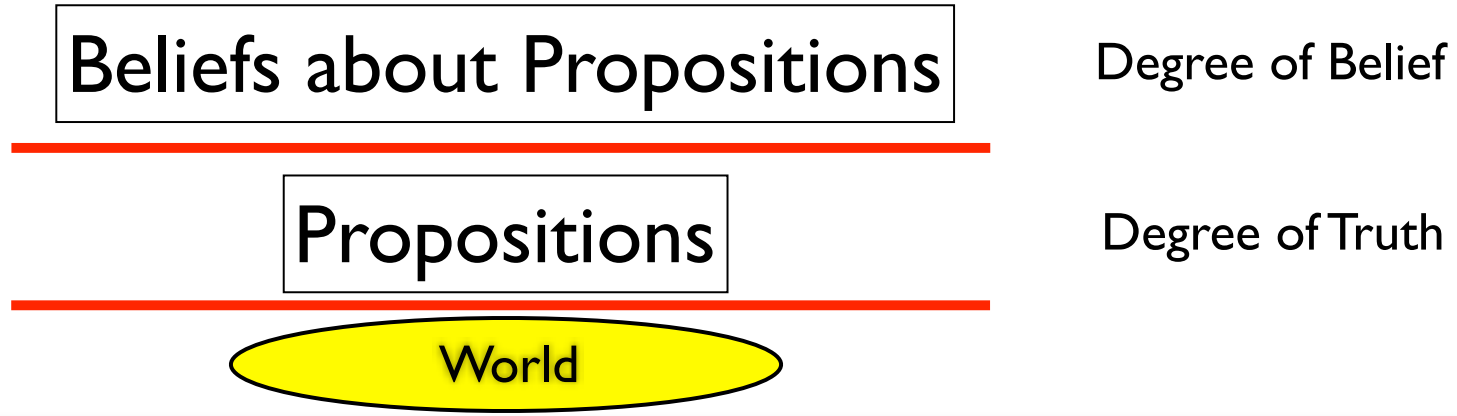
The truth table method can be used to evaluate the truth or falsity of any formula, although it is very inefficient. One requires a table with 2^n rows, where n is the number of propositional variables in the language.

Propositional logic allows the representation of propositions about the world which are true or false. In this case, a proposition has a degree of truth, either true or false.

Suppose our knowledge about the truth or falsity of a proposition is uncertain. In this case, we might want to attach a degree of belief in its truth status.

Observe that this degree of belief is subjective, in the sense that the proposition in question is still considered to be true or false in the world, we simply do not have enough information to determine this.

In this case, it is important to distinguish between degrees of truth and degrees of belief.



A Language of Probability

Just as propositional atoms provide the primitive vocabulary for propositions in propositional logic, random variables will provide the primitive vocabulary for our probabilistic language.

Random Variables:

Boolean Cavity: {true, false}

Discrete Weather: {sunny, rainy, cloudy, snow}

Continuous Temperature: $\{x \mid -43.0 \leq x \leq 100.0\}$

A random variable may be viewed as an aspect of the world that is initially unknown. A degree of belief may then be attached to a variable/value pair. Complex formulas may be formed using Boolean combinations of variable/value pairs.

Factored representations like that used with CSPs

Probability Distributions

$P(Cavity = true) = P(cavity) = 0.4$

$P(Cavity = false) = P(\neg cavity) = 0.6$

$\mathbf{P}(Cavity) = \langle 0.4, 0.6 \rangle$

P Notation

$P(Weather = sunny) = 0.7$

$P(Weather = rainy) = 0.2$

$P(Weather = cloudy) = 0.08$

$P(Weather = snow) = 0.02$

$\mathbf{P}(Weather) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$

$\mathbf{P}(X)$ is the *Probability Distribution (Unconditional or Prior Probability)* of the random variable X.

Joint Probability Distributions

Joint Probability Distributions:

$$\mathbf{P}(Cavity, Weather) = \langle 0.1, 0.1, 0.1, 0.1, 0.2, 0.1, 0.2, 0.1 \rangle \quad (2 \times 4)$$

$$\mathbf{P}(cavity, Weather) = \langle 0.1, 0.1, 0.1, 0.1 \rangle \quad (1 \times 4)$$

$$\mathbf{P}(Cavity, weather = rainy) = \langle 0.1, 0.2 \rangle \quad (2 \times 1)$$

Full Joint Probability Distribution: $\{X_1, \ldots, X_n\}$

Assume a domain of random variables, $\{X_1, \ldots, X_n\}$.
A *full joint probability distribution* $\mathbf{P}(X_1, \ldots, X_n)$ assigns a probability to each of the possible combinations of variable/value pairs.

Joint Probability Distributions

Joint Probability Distribution:

$$\mathbf{P}(Cavity, Weather) = \langle 0.28, 0.08, 0.032, 0.008, 0.42, 0.12, 0.048, 0.12 \rangle \quad (2 \times 4)$$

Marginalizing: use of **P** notation:

$$\mathbf{P}(cavity, Weather) = \langle 0.28, 0.08, 0.032, 0.008 \rangle \quad (1 \times 4)$$

$$\mathbf{P}(Cavity, Weather = rainy) = \langle 0.08, 0.12 \rangle \quad (2 \times 1)$$

Assume a domain of variables : $\{X_1, \ldots, X_n\}$

A *full joint probability distribution*, $\mathbf{P}(X_1, \ldots, X_n)$,
assigns a probability to each of the the possible combinations
Of variable/value pairs

An Example

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

Table 2: Full joint probability distribution

	Cavity (Cav)	Toothache (Too)	Catch (Cat)	
1	T	T	T	0.108
2	T	T	F	0.012
3	T	F	T	0.072
4	T	F	F	0.008
5	F	T	T	0.016
6	F	T	F	0.064
7	F	F	T	0.144
8	F	F	F	0.576

The probabilities for each atomic event (an interpretation) must sum to 1.

Full Joint Probability Distribution

Given a full joint probability distribution, arbitrary Boolean combinations of random variable/value pairs can be interpreted by taking the sum of the degree of beliefs attached to each interpretation (atomic event) which satisfies the formula.

	Cavity (Cav)	Toothache (Too)	Catch (Cat)		$Cav \vee Too$	$Cav \rightarrow Too$	$\neg Too$
1	T	T	T	0.108	T	T	F
2	T	T	F	0.012	T	T	F
3	T	F	T	0.072	T	F	T
4	T	F	F	0.008	T	F	T
5	F	T	T	0.016	T	T	F
6	F	T	F	0.064	T	T	F
7	F	F	T	0.144	F	T	T
8	F	F	F	0.576	F	T	T

$$P(cav \vee too) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

$$P(cav \rightarrow too) = 0.108 + 0.012 + 0.016 + 0.064 + 0.144 + 0.576$$

$$P(\neg too) = 1 - P(Too) = 1 - (0.108 + 0.012 + 0.016 + 0.064) = 0.8$$

$$P(\neg too) = 0.072 + 0.008 + 0.144 + 0.576 = 0.8$$

Conditional Probability

Prior probabilities are not adequate once additional evidence concerning previously unknown random variables is introduced. One must condition any random variable of interest relative to the new information. This conditioning is represented using conditional or posterior probabilities.

The probability of $X = x_i$ given $Y = y_j$ is denoted: $P(X = x_i \mid Y = y_j)$

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i \wedge Y = y_j)}{P(Y = y_j))}$$

Another way to write this is in the form of the *product rule*:

$$P(X = x_i \wedge Y = y_j) = P(X = x_i \mid Y = y_j) * P(Y = y_j)$$

$$P(X = x_i \wedge Y = y_j) = P(Y = y_j \mid X = x_i) * P(X = x_j)$$

From the product rule, we may derive Bayes Rule!

Some (P) Notation

$\mathbf{P}(X \mid Y)$ denotes the set of equations $P(X = x_i \mid Y = y_j)$, for each possible i, j.

For example:

$$\mathbf{P}(X \wedge Y) = \mathbf{P}(X, Y) = \mathbf{P}(X \mid Y) * \mathbf{P}(Y)$$

denotes

$$\begin{aligned}
P(X = x_1 \wedge Y = y_1) &= P(X = x_1 \mid Y = y_1) * P(Y = y_1) \\
P(X = x_1 \wedge Y = y_2) &= P(X = x_1 \mid Y = y_2) * P(Y = y_2) \\
&\vdots \\
P(X = x_i \wedge Y = y_j) &= P(X = x_i \mid Y = y_j) * P(Y = y_j) \\
&\vdots
\end{aligned}$$

Kolmogorov's Axioms

Probability theory can be built up from three axioms:

1. All probabilities are between 0 and 1.
   - For any proposition a, $0 \leq P(a) \leq 1$
2. Necessarily true (i.e. valid) propositions have probability 1, and necessarily false propositions have probability 0.
   - $P(True) = 1$ and $P(False) = 0$
3. The probability of a disjunction is given by
   - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

Some Useful Properties

A probability distribution on a random variable X with domain $\langle x_1, \ldots, x_n \rangle$ must add up to 1:

$$\sum_{i=1}^{n} P(X = x_i) = 1$$

The probability of a proposition is equal to the sum of the probabilities of the atomic events (interpretations) in which it holds:

$$P(a) = \sum_{e_i \in \mathbf{e}(a)} P(e_i)$$

where $\mathbf{e}(a)$ denotes the set of atomic events where a is true.

Marginalization

Marginalization, or summing out, extracts the distribution over a single variable of some subset of variables from an existing joint distribution.

Table 4: default

	toothach		¬*toothache*	
	catch	¬*catch*	*catch*	¬*catch*
cavity	0.108	0.012	0.072	0.008
¬*cavity*	0.016	0.064	0.144	0.576

catch

For example, with no knowledge about toothaches, the marginal probability of $Cavity = true$ would be:

$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008$ = 0.2

General Marginalization Rule

For any sets of variables $\mathbf{Y}$ and $\mathbf{Z}$,

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{z})$$

where $\mathbf{z}$ is the exhaustive set of sequences of variable/value pairs from the variable set $\mathbf{Z}$.

Some Examples

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{z})$$

Let $\mathbf{Y} = \{Cavity, Catch\}, \mathbf{Z} = \{Toothache\}$

$$\mathbf{P}(\mathbf{Y}) = \mathbf{P}(\mathbf{Y}, toothache) + \mathbf{P}(\mathbf{Y}, \neg toothache)$$

$$\begin{aligned} P(cavity, catch) &= P(cavity, catch, toothache) + P(cavity, catch, \neg toothache) \\ &= 0.108 + 0.072 = 0.18 \end{aligned}$$

Let $\mathbf{Y} = \{Cavity\}, \mathbf{Z} = \{Catch, Toothache\}$

$$\begin{aligned} \mathbf{P}(\mathbf{Y}) &= \mathbf{P}(\mathbf{Y}, catch, toothache) + \mathbf{P}(\mathbf{Y}, \neg catch, toothache) + \\ &\quad \mathbf{P}(\mathbf{Y}, catch, \neg toothache) + \mathbf{P}(\mathbf{Y}, \neg catch, \neg toothache) \end{aligned}$$

$$\begin{aligned} P(cavity) &= P(cavity, catch, toothache) + P(cavity, \neg catch, toothache) + \\ &\quad P(cavity, catch, \neg toothache) + P(cavity, \neg catch, \neg toothache) \\ &= 0.108 + 0.012 + 0.072 + 0.008 = 0.2 \end{aligned}$$

Conditionalization

Conditionalization, generates the weighted sum over the beliefs of all the distinct ways (**z**), a variable or set of variables (**Y**) might be realized.

For any sets of variables $\mathbf{Y}$ and $\mathbf{Z}$,

Application of The product rule

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y} \mid \mathbf{z}) * \mathbf{P}(\mathbf{z})$$

where $\mathbf{z}$ is the exhaustive set of sequences of variable/value pairs from the variable set $\mathbf{Z}$.

Some Examples:

Let $\mathbf{Y} = \{Cavity\}, \mathbf{Z} = \{Toothache\}$

$$\mathbf{P}(\mathbf{Y}) = \mathbf{P}(\mathbf{Y} \mid toothache) * \mathbf{P}(toothache) + \mathbf{P}(\mathbf{Y} \mid \neg toothache) * \mathbf{P}(\neg toothache)$$

$$\begin{aligned} P(cavity) \quad = \quad & P(cavity \mid toothache) * P(toothache) + \\ & P(cavity \mid \neg toothache) * P(\neg toothache) \end{aligned}$$

Computing Conditionals

The main form of inference with probabilities is to compute the conditional probabilities of some variables, given evidence about others.

What is the probability that I have a cavity given I have evidence of a toothache?

$$
\begin{aligned}
P(cavity|toothache) &= \frac{P(cavity \wedge toothache)}{P(toothache)} \\
&= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} \\
&= 0.6
\end{aligned}
$$

$$
\begin{aligned}
P(\neg cavity|toothache) &= \frac{P(\neg cavity \wedge toothache)}{P(toothache)} \\
&= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} \\
&= 0.4
\end{aligned}
$$

Note that computing a conditional is in some sense dependent on having a full joint distribution for all random variables in the domain in question.

Conditional Probability Distribution

Given the conditional distribution: $\mathbf{P}(Cavity \mid toothache)$

$$\alpha = \frac{1}{P(toothache)}$$

can be viewed as a normalization constant for the distribution ensuring that it adds up to 1.

The two equations can then be written as one:

$$\begin{aligned}
\mathbf{P}(Cavity \mid toothache) &= \alpha * \mathbf{P}(Cavity, toothache) \\
&= \alpha * [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
&= \alpha * [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle] \\
&= \alpha * \langle 0.12 , 0.08 \rangle = \langle 0.6, 0.4 \rangle
\end{aligned}$$

$$\alpha = \frac{1}{0.12 + 0.08}$$

A General Inference Procedure

Let X be the *query variable*, $\mathbf{E}$ be the *evidence variables*, $\mathbf{e}$ be the observed values for them, $\mathbf{Y}$ be the remaining *unobserved (hidden) variables* and $\mathbf{y}$ be the exhaustive set of sequences of distinct variable/value pairs of the unobserved variables Y.

Note that $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$ is the set of all variables in the full joint distribution.

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

An Example

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

$X = \{Cavity\}, \mathbf{E} = \{Toothache\}, \mathbf{e} = \{toothache\}, \mathbf{Y} = \{Catch\}, \mathbf{y} = \{\{catch\}, \{\neg catch\}\}$

$$\begin{aligned}
\mathbf{P}(Cavity \mid toothache) &= \alpha * \mathbf{P}(Cavity, toothache) \\
&= \alpha * \sum_{\mathbf{y}} \mathbf{P}(Cavity, toothache, \mathbf{y}) \\
&= \alpha * [\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
&= \alpha * [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle] \\
&= \alpha * \langle 0.12 , 0.08 \rangle \\
&= \langle 0.6, 0.4 \rangle
\end{aligned}$$

Note: $P(X = x_i \mid e)$ is not related to $e \vDash X = x_1$

Comments

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

- The equation above can serve as the basis for an implementation of an inference procedure. Unfortunately it is not efficient
  - It requires an input table for the full joint distribution. Assuming n variables, this would require a table size of $O(2^n)$ and $O(2^n)$ time to run the algorithm.
- This should be viewed as the theoretical foundation for development of more efficient inferencing techniques.

Independence (1)

Suppose we add a new variable $Weather$ with four potential values, to the full joint distribution for $Toothache, Catch, Cavity$:

$\mathbf{P}(Toothache, Catch, Cavity, Weather)$ $(2 * 2 * 2 * 4)$

Previously, the table contained 8 entries. With the new addition, it would contain 32 entries.

Given any values of the four variables,the product rule tells us the following:

$P(toothache, catch, cavity, Weather = cloudy)$

$= P(Weather = cloudy \mid toothache, catch, cavity) * P(toothache, catch, cavity)$

However, it is intuitively correct to assume that weather has nothing to do with cavities, etc.

Independence (2)

However, it is intuitively correct to assume that weather has nothing to do with cavities, etc.
So it should be possibile to assert

$$P(Weather = cloudy \mid toothache, catch, cavity) = P(Weather = cloudy)$$

From this, we can deduce:

$$P(toothache, catch, cavity, Weather = cloudy)$$

$$= P(Weather = cloudy) * P(toothache, catch, cavity)$$

More generally,

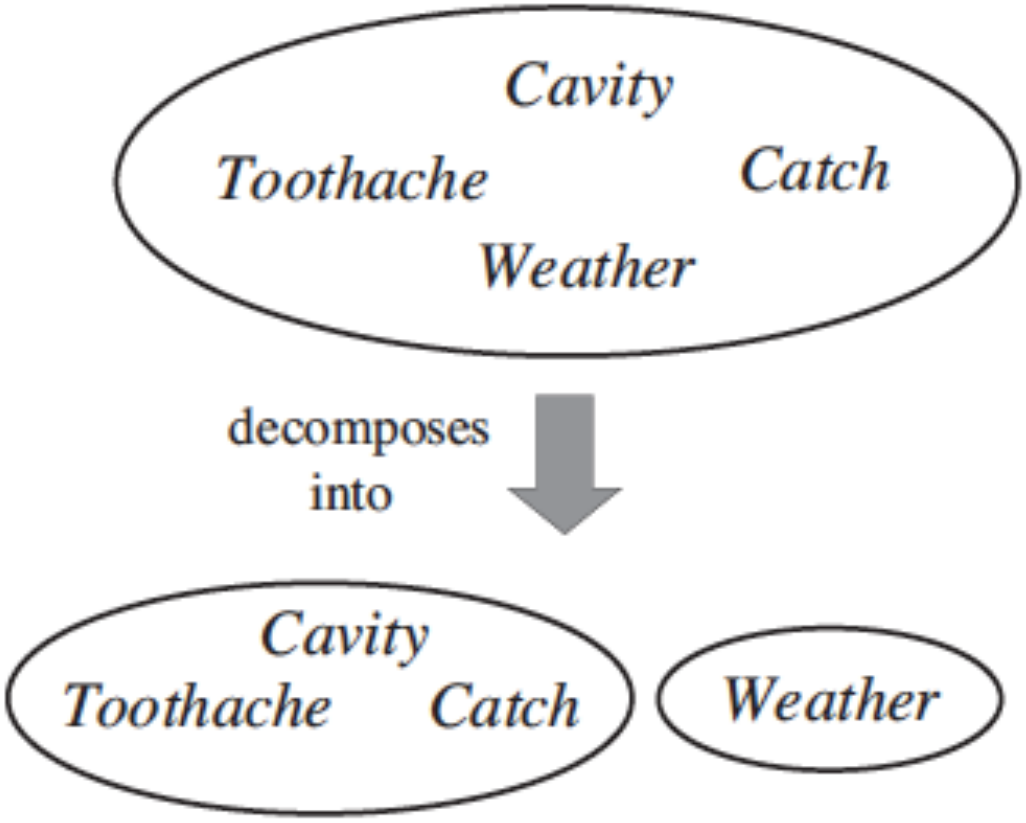
$$\mathbf{P}(Toothache, Catch, Cavity, Weather)$$

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

The full joint distribution can now be defined in terms of an 8 element table (the original) and a new four element table instead of a 32 element table!

Independence assumptions imply partitioning and local computation of sorts.
Perhaps this insight can be used to develop more efficient inferencing techniques!

Factoring



Independence assertions can both reduce the size of the domain representation and make the inferencing problem more efficient.

Absolute Independence

Independence between variables can be written as follows:

$$\mathbf{P}(X \mid Y) = \mathbf{P}(X) \text{ or } \mathbf{P}(Y \mid X) = \mathbf{P}(Y) \text{ or } \mathbf{P}(X, Y) = \mathbf{P}(X) * \mathbf{P}(Y)$$

Independence assumptions are domain dependent, but if the set of variables can be divided in independent subsets, then the full joint probability distribution can be factored into separate joint distributions on those subsets.

This implies a reduction in the size of the domain representation and in the complexity of the the inference problem.

Baye's Rule (Simple Case)

The product rule states that,

$$\mathbf{P}(X, Y) = \mathbf{P}(X \mid Y) * \mathbf{P}(Y) = \mathbf{P}(Y \mid X) * \mathbf{P}(X)$$

From this we can derive,

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

Bayes' Rule has widespread applications:

$$\mathbf{P}(Hypothesis \mid Evidence) = \frac{\mathbf{P}(Evidence|Hypothesis)*\mathbf{P}(Hypothesis)}{\mathbf{P}(Evidence)}$$

$$\mathbf{P}(Cause \mid Effects) = \frac{\mathbf{P}(Effects|Cause)*\mathbf{P}(Cause)}{\mathbf{P}(Effects)}$$

$$\mathbf{P}(Disease \mid Symptoms) = \frac{\mathbf{P}(Symptons|Disease)*\mathbf{P}(Disease)}{\mathbf{P}(Symptoms)}$$

Intuitions

$$\mathbf{P}(Hypothesis \mid Evidence) = \frac{\mathbf{P}(Evidence|Hypothesis)*\mathbf{P}(Hypothesis)}{\mathbf{P}(Evidence)}$$

Given a prior probability for a Hypothesis, $\mathbf{P}(Hypothesis)$, upon receiving new evidence, where its prior probability is given, $\mathbf{P}(Evidence)$, what is my revised belief for the hypothesis in the context of the new evidence: $\mathbf{P}(Hypothesis \mid Evidence)$?

$\mathbf{P}(Hypothesis)$ is called the *prior probability* and $\mathbf{P}(Hypothesis \mid Evidence)$ is called the *posterior probability*.

An Example

Doctors often know how many patients with a given disease exhibit various symptoms:

$$P(StiffNeck \mid Meningitis) = 0.5$$

Doctors generally also know some unconditional facts:

$$P(Meningitis) = \frac{1}{50{,}000}, P(StiffNeck) = \frac{1}{20}$$

What is the probability a patient has Meningitis given evidence of a stiff neck?

$$\mathbf{P}(Disease \mid Symptoms) = \frac{\mathbf{P}(Symptons \mid Disease) * \mathbf{P}(Disease)}{\mathbf{P}(Symptoms)}$$

$$\begin{aligned} \mathbf{P}(Meningitis \mid Stiffneck) &= \frac{\mathbf{P}(StiffNeck \mid Meningitis) * \mathbf{P}(Meningitis)}{\mathbf{P}(StiffNeck)} \\ &= \frac{0.5 * \frac{1}{50{,}000}}{\frac{1}{20}} \\ &= 0.0002 = \frac{1}{5000} \end{aligned}$$

A marked increase!

Generalizations: Normalized Form

P(X)

In order to avoid assessing the probability of the evidence , one can instead compute the posterior probability for each value of the query variable and then normalize the results. The normalized form of Bayes' rule is,

$$\mathbf{P}(Y \mid X) = \alpha * \mathbf{P}(X \mid Y) * \mathbf{P}(Y)$$

where α is the normalization constant needed to make the entries in $\mathbf{P}(Y \mid X)$ sum to 1. α can be computed using the numerator of Bayes' rule and the conditionalization rule.

$$\alpha = \frac{1}{\mathbf{P}(X)} = \frac{1}{\sum_{\mathbf{y}} \mathbf{P}(X \mid \mathbf{y}) * P(\mathbf{y})}$$

The denominator of α is the conditionalization of X relative to all values of Y.

Many Pieces of Evidence

In the case if more than one piece of evidence, Bayes' rule can be generalized:

$$\mathbf{P}(Y \mid X, \mathbf{e}) = \frac{\mathbf{P}(X \mid Y, \mathbf{e}) * \mathbf{P}(Y \mid \mathbf{e})}{\mathbf{P}(X \mid \mathbf{e})}$$

$$\mathbf{P}(Meningitis \mid Stiffneck, SwollenBrain) =$$

$$\frac{\mathbf{P}(StiffNeck \mid Meningitis, SwollenBrain) * \mathbf{P}(Meningitis \mid SwollenBrain)}{\mathbf{P}(StiffNeck \mid SwollenBrain)}$$

The Chain Rule

A joint distribution can be factored into a product of conditional probabilities using the chain rule:

$$\mathbf{P}(X_1, X_2, \ldots, X_n) = \mathbf{P}(X_1 \mid X_2, \ldots, X_n) * \mathbf{P}(X_2 \mid X_3, \ldots, X_n) *, \ldots, * \mathbf{P}(X_{n-1} \mid X_n) * \mathbf{P}(X_n)$$

In general form,

$$\mathsf{P}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \mathsf{P}(X_i \mid X_{i+1}, \ldots, X_n)$$

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

This formulation allows us to compute any value in the table representing the full joint distribution provided we have the conditional probabilities. Consequently, we can then compute the value of any formula.

Conditional Independence

The conditional independence of two variables X and Y, given a third variable Z is,

$$\mathbf{P}(X, Y \mid Z) = \mathbf{P}(X \mid Z) * \mathbf{P}(Y \mid Z)$$

Equivalently,

$$\mathbf{P}(X \mid Y, Z) = \mathbf{P}(X \mid Z) \quad \text{and} \quad \mathbf{P}(Y \mid X, Z) = \mathbf{P}(Y \mid Z)$$

Suppose $Toothache$ and $Catch$ are independent given $Cavity$, then

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

Toothache and Catch are not absolutely independent because if a probe catches in a tooth, it probably has a cavity and that probably causes a toothache.

Toothache and catch are conditionally independent given cavity.
Each is directly caused by a cavity, but neither has a direct effect on the other.
Toothache depends on the nerves in the mouth to which catch is irrelevant.
Catch depends on the skill of the doctor to which the toothache is irrelevant.

Naive Bayes' Model (1)

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

From the previous slide, we know that $Toothache$ and $Catch$ are independent given $Cavity$, so by conditional independence,

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

Substituting above,

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

From this one can begin to discern an interesting pattern associated with causes and effects.

Naive Bayes' Model (2)

There is a common pattern in which a common cause influences a number of effects, all of which are conditionally independent relative to the cause. (The dentist example). In this case,

$$\mathbf{P}(Cause, Effect_1, Effect_2, \ldots, Effect_n) = \mathbf{P}(Cause) * \prod_{i=1}^{n} \mathbf{P}(Effect_i \mid Cause)$$

$$\mathbf{P}(Cavity, Toothache, Catch) = \mathbf{P}(Cavity) * \prod_{i=1}^{n} \mathbf{P}(Effect_i \mid Cavity)$$

$$= \mathbf{P}(Cavity) * \mathbf{P}(Toothache \mid Cavity) * \mathbf{P}(Catch \mid Cavity)$$

Naive Bayes is often used even when there are dependencies among effects due to its efficiency and relative correctness in output.

Comments

- Conditional independence assertions allow probabilistic systems to scale up by permitting implicit and compact representations of full joint distributions.
- This will be used to advantage with Bayesian Networks.
- ”The decomposition of large probabilistic domains into weakly connected subsets via conditional independence assumptions is one of the most important developments in the recent history of AI.”

Bayesian Networks

- Full joint probability distributions can answer any question about a modeled domain.
  - Intractably large as the number of variables grows
  - Specifying probabilities for all atomic events is difficult to do.
- Independence and conditional independence assumptions can greatly reduce the number of probabilities that need to be specified in order to define full joint probability distributions
- Bayesian networks are data structures that represent dependencies among variables and give precise specifications of any full joint probability distribution.

Bayesian Networks

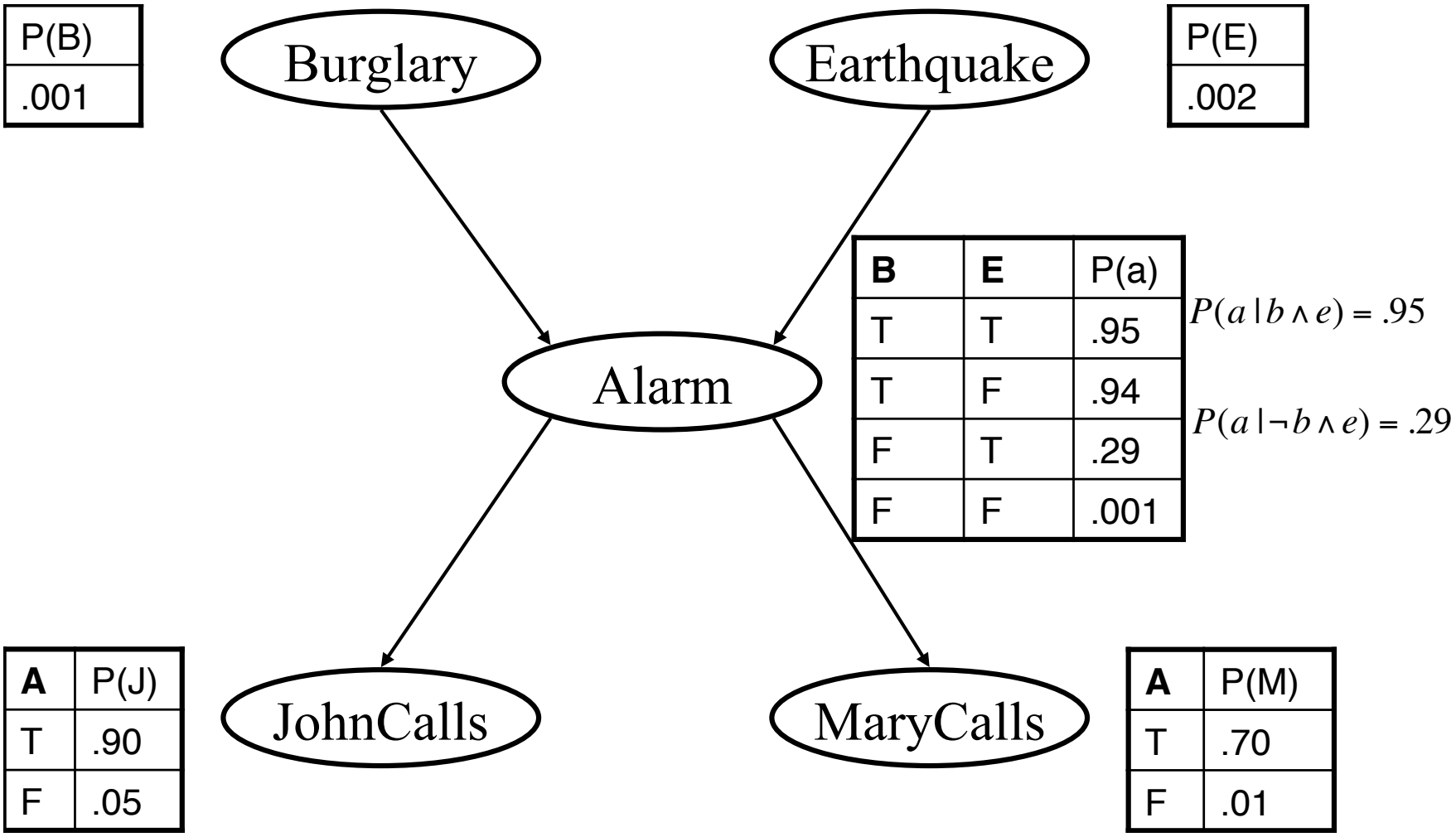
A Bayesian Network is a directed graph where each node is annotated with quantitative probability information:

1. A set of random variables makes up the nodes in the network.
2. A set of directed arrows connects pairs of nodes. If there is an arrow from X to Y, X is said to be the parent of Y.
3. Each node X_i has a conditional probability distribution $P(X_i \mid Parents(X_i))$ that quantifies the effect of the parents on the node.
4. The graph has no cycles. It is a DAG (directed, acyclic graph).

Example (J. Pearl)

- A person installs a new burglar alarm at home. It responds to burglaries, but may also respond to earthquakes on occasion.
- The person has two neighbors, John and Mary, who promise to call you at work when the alarm goes off.
  - John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm sound.
  - Mary, who likes loud music sometimes misses the alarm altogether
- Queries
  - Given evidence of who has or has not called, estimate the probability of a burglary,
  - P(burglary | john, ~mary)

Bayesian Network



Semantics of Bayesian Networks

We are interested in computing entries in the joint probability distribution:

$$P(X_1 = x_1 \wedge \ldots \wedge X_n = x_n) \quad \text{abbreviated} \quad P(x_1, \ldots, x_n)$$

This is defined as:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))$$

A bayesian network is a representation of the joint probability distribtution

where $parents(X_i)$ denotes the specific values of variables in $Parents(X_i)$

For example, what is the probability that the alarm has sounded, but neither earthquake nor burglary has occurred and both John and Mary call?

$$\begin{aligned} P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) &= \\ &= P(j \mid a) * P(m \mid a) * P(a \mid \neg b \wedge \neg e) * P(\neg b) * P(\neg e) \\ &= 0.90 * 0.70 * 0.001 * 0.999 * 0.998* = 0.00062 \end{aligned}$$

Constructing Bayesian Networks (1)

The chain rule shows that a joint distribution can be factored into a product of conditional distributions:

$$\mathbf{P}(X_1, X_2, \ldots, X_n) = \mathbf{P}(X_1 \mid X_2, \ldots, X_n) * \mathbf{P}(X_2 \mid X_3, \ldots, X_n) *, \ldots, * \mathbf{P}(X_{n-1} \mid X_n) * \mathbf{P}(X_n)$$

$$\mathsf{P}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \mathsf{P}(X_i \mid X_{i+1}, \ldots, X_n)$$

From the semantics of Bayesian Networks,

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))$$

In general,

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$

Constructing Bayesian Networks (2)

$$\text{ChainRule}: \quad \mathsf{P}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \mathsf{P}(X_i \mid X_{i+1}, \ldots, X_n)$$

$$\text{SemanticsofBN}: \quad \mathbf{P}(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$

Form the above, for every variable X_i in the network:

$$\mathsf{P}(X_i \mid X_{i+1}, \ldots, X_n) = \mathsf{P}(X_i \mid Parents(X_i)) \text{ provided } Parents(X_i) \subset \{X_{i+1}, \ldots, X_n\}$$

This condition is satisfied by labeling the nodes in any order that is consistent with the partial order implicit in the graph structure.

$$X_1 : JohnCalls, X_2 : MaryCalls, X_3 : Alarm, X_4 : Burglary, X_5 : Earthquake$$

Effects precede causes

The Bayesian Network is a correct representation of the domain only if each node is conditionally independent of other successors in the node ordering, given its parents

Exact Inference in Bayesian Networks

The main task in any probabilistic inference system is to compute the posterior probability distribution of a set of query variables, given some observed event (assignment of values to set of evidence variables)

Let X be the query variable, $\mathbf{E}$ be the evidence variables, $\mathbf{e}$ be the observed values for them, $\mathbf{Y}$ be the remaining unobserved (hidden) variables and and $\mathbf{y}$ be the exhaustive set of sequences of distinct variable/value pairs of the unobserved variables Y.

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

We know that the terms **P**(X, **e**, **y**) in the joint distribution can be written as products of conditional probabilities from the network. So, a query is answered by computing the sums of products of conditional probabilities from the network.

An Example

$$\textbf{Query}: \quad \mathbf{P}(Burglary \mid JohnCalls = true, MaryCalls = true)$$

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

$$\begin{array}{lcl} X & = & \{Burglary\} \\ \mathbf{E} & = & \{JohnCalls, MaryCalls\} \\ \mathbf{e} & = & \{JohnCalls = true, MaryCalls = true\} \\ \mathbf{Y} & = & \{Earthquake, Alarm\} \end{array}$$

$$\mathbf{P}(Burglary \mid johncalls, marycalls) = \alpha * \mathbf{P}(Burglary, johncalls, marycalls)$$

$$= \alpha * \sum_e \sum_a \mathbf{P}(Burglary, johncalls, marycalls, Earthquake = e, Alarm = a)$$

$$= \alpha * \sum_e \sum_a \mathbf{P}(Burglary) * P(jc \mid A = a) * P(mc \mid A = a) * P(E = e) * P(A = a \mid B, E = e)$$

$$= \alpha * \langle 0.0005922, 0.001483 \rangle = \langle 28.54, 71.46 \rangle$$

The chance of a burglary given calls from both neighbors is about 28%

Example

$Burglary = True$

$$\sum_a \sum_e P(\text{jc}, \text{mc}, \text{b}, a, e) = P(\text{jc}, \text{mc}, \text{b}, \text{a}, \text{e}) + P(\text{jc}, \text{mc}, \text{b}, \text{a}, \neg\text{e}) + P(\text{jc}, \text{mc}, \text{b}, \neg\text{a}, \text{e}) + P(\text{jc}, \text{mc}, \text{b}, \neg\text{a}, \neg\text{e})$$

$$P(\text{jc}, \text{mc}, \text{b}, \text{a}, \text{e}) = P(\text{jc} \mid \text{a}) * P(\text{mc} \mid \text{a}) * P(\text{a} \mid \text{b}, \text{e}) * P(\text{b}) * P(\text{e}) = 1.197e^{-6}$$

$$P(\text{jc}, \text{mc}, \text{b}, \text{a}, \neg\text{e}) = P(\text{jc} \mid \text{a}) * P(\text{mc} \mid \text{a}) * P(\text{a} \mid \text{b}, \neg\text{e}) * P(\text{b}) * P(\neg\text{e}) = 0.000591$$

$$P(\text{jc}, \text{mc}, \text{b}, \neg\text{a}, \text{e}) = P(\text{jc} \mid \neg\text{a}) * P(\text{mc} \mid \neg\text{a}) * P(\neg\text{a} \mid \text{b}, \text{e}) * P(\text{b}) * P(\text{e}) = 5e^{-11}$$

$$P(\text{jc}, \text{mc}, \text{b}, \neg\text{a}, \neg\text{e}) = P(\text{jc} \mid \neg\text{a}) * P(\text{mc} \mid \neg\text{a}) * P(\neg\text{a} \mid \text{b}, \neg\text{e}) * P(\text{b}) * P(\neg\text{e}) = 2{,}994e^{-8}$$

$$0.0005922$$

Example

$Burglary = False$

$$\sum_a \sum_e P(\text{jc}, \text{mc}, \neg\text{b}, a, e) = P(\text{jc}, \text{mc}, \neg\text{b}, \text{a}, \text{e}) + P(\text{jc}, \text{mc}, \neg\text{b}, \text{a}, \neg\text{e}) + P(\text{jc}, \text{mc}, \neg\text{b}, \neg\text{a}, \text{e}) + P(\text{jc}, \text{mc}, \neg\text{b}, \neg\text{a}, \neg\text{e})$$

$$P(\text{jc}, \text{mc}, \neg\text{b}, \text{a}, \text{e}) = P(\text{jc} \mid \text{a}) * P(\text{mc} \mid \text{a}) * P(\text{a} \mid \neg\text{b}, \text{e}) * P(\neg\text{b}) * P(\text{e}) = 0.000365$$

$$P(\text{jc}, \text{mc}, \neg\text{b}, \text{a}, \neg\text{e}) = P(\text{jc} \mid \text{a}) * P(\text{mc} \mid \text{a}) * P(\text{a} \mid \neg\text{b}, \neg\text{e}) * P(\neg\text{b}) * P(\neg\text{e}) = 0.000628$$

$$P(\text{jc}, \text{mc}, \neg\text{b}, \neg\text{a}, \text{e}) = P(\text{jc} \mid \neg\text{a}) * P(\text{mc} \mid \neg\text{a}) * P(\neg\text{a} \mid \neg\text{b}, \text{e}) * P(\neg\text{b}) * P(\text{e}) = 7.092e^{-7}$$

$$P(\text{jc}, \text{mc}, \neg\text{b}, \neg\text{a}, \neg\text{e}) = P(\text{jc} \mid \neg\text{a}) * P(\text{mc} \mid \neg\text{a}) * P(\neg\text{a} \mid \neg\text{b}, \neg\text{e}) * P(\neg\text{b}) * P(\neg\text{e}) = 0.000189$$

$$0.001483$$

$$\alpha = \frac{1}{0.0005922 + 0.001483}$$

$$\alpha * \langle 0.0005922, 0.001483 \rangle = \langle 28.54, 71.46 \rangle$$