

- Once the agent has obtained some evidence concerning the previously unknown random variables making up the domain, we have to switch to using *conditional* (posterior) probabilities
- P(a | b) is the probability of proposition a, given that all we know is h

- P(cavity) = P(cavity |)
- We can express conditional probabilities in terms of unconditional probabilities:

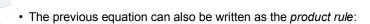
$$P(a \mid b) = \frac{P(a \land b)}{P(b)}$$

whenever P(b) > 0



OHJ-2556 Artificial Intelligence, Spring 2011

10.3.201



• We can, of course, have the rule the other way around

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

 $P(a \land b) = P(a \mid b) P(b)$

· Conditional distributions:

$$\underline{P}(X \mid Y) \equiv P(X = x_i \mid Y = y_i) \forall i, j$$

· By the product rule

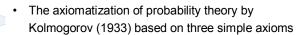
$$\underline{P}(X,Y) = \underline{P}(X \mid Y) \underline{P}(Y)$$

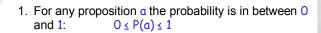
(entry-by-entry, not a matrix multiplication)

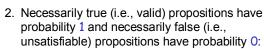


OHJ-2556 Artificial Intelligence, Spring 2011









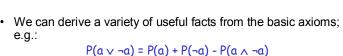
$$P(true) = 1$$
 $P(false) = 0$

3. The probability of a disjunction is given by the *inclusion-exclusion principle*

$$P(a \lor b) = P(a) + P(b) - P(a \land b)$$



OHJ-2556 Artificial Intelligence, Spring 2011



 The fact of the third line can be extended for a discrete variable b with the domain d_{1,...},d_n:

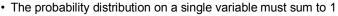
$$\sum_{i=1,...,n} P(D = d_i) = 1$$

 For a continuous variable X the summation is replaced by an integral:

$$\int_{-\infty}^{\infty} P(X=x) dx = 1$$

TAMPERE UNIVERSITY OF TECHNOLOGY Department of Software Systems

OHJ-2556 Artificial Intelligence, Spring 2011



- It is also true that any joint probability distribution on any set of variables must sum to 1
- Recall that any proposition a is equivalent to the disjunction of all the atomic events in which a holds
- Call this set of events e(a)
- Atomic events are mutually exclusive, so the probability of any conjunction of atomic events is zero, by axiom 2
- · Hence, from axiom 3

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

 Given a full joint distribution that specifies the probabilities of all atomic events, this equation provides a simple method for computing the probability of any proposition



OHJ-2556 Artificial Intelligence, Spring 2011

10.3.201

13.3 Inference Using Full Joint Distribution

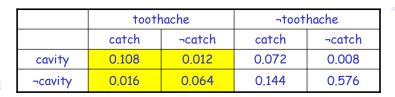
	toothache		¬toothache	
	catch	-catch	catch	-catch
cavity	0.108	0.012	0.072	0.008
¬cavity	0.016	0.064	0.144	0.576

- E.g., there are six atomic events for cavity v toothache:
 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28
- Extracting the distribution over a variable (or some subset of variables), marginal probability, is attained by adding the entries in the corresponding rows or columns
- E.g., P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2
- We can write the following general marginalization (summing out) rule for any sets of variables Y and Z:

$$\underline{P}(Y) = \sum_{z \in Z} \underline{P}(Y, z)$$

TAMPERE UNIVERSITY OF TECHNOLOGY Department of Software Systems

OHJ-2556 Artificial Intelligence, Spring 2011



Computing a conditional probability

```
P(cavity | toothache) =
    P(cavity \( \) toothache)/P(toothache) =
    (0.108 + 0.012)/(0.108 + 0.012 + 0.016 + 0.064) =
    0.12/0.2 = 0.6
```

Respectively

```
P(¬cavity | toothache) = (0.016 + 0.064)/0.2 = 0.4
```

· The two probabilities sum up to one, as they should



OHJ-2556 Artificial Intelligence, Spring 2011

10.3.201



- 1/P(toothache) = 1/0.2 = 5 is a normalization constant ensuring that the distribution P(Cavity | toothache) adds up to 1
- \bullet Let α denote the normalization constant

 In other words, we can calculate the conditional probability distribution without knowing P(toothache) using normalization



OHJ-2556 Artificial Intelligence, Spring 2011



- · More generally:
- we need to find out the distribution of the query variable X (Cavity),
- evidence variables E (Toothache) have observed values e, and
- the remaining unobserved variables are Y (Catch)
- Evaluation of a query:

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

where the summation is over all possible ys; i.e., all possible combinations of values of the unobserved variables Y



TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Software Systems

OHJ-2556 Artificial Intelligence, Spring 2011



- $\underline{P}(X, e, y)$ is simply a subset of the joint probability distribution of variables X, E, and Y
- · X, E, and Y together constitute the complete set of variables for the domain
- · Given the full joint distribution to work with, the equation in the previous slide can answer probabilistic queries for discrete variables
- · It does not scale well
- For a domain described by n Boolean variables, it requires an input table of size $O(2^n)$ and takes $O(2^n)$ time to process the table
- · In realistic problems the approach is completely impractical



OHJ-2556 Artificial Intelligence, Spring 2011



- If we expand the previous example with a fourth random variable Weather, which has four possible values, we have to copy the table of joint probabilities four times to have 32 entries together
- Dental problems have no influence on the weather, hence:

```
P(Weather = cloudy | toothache, catch, cavity) = P(Weather = cloudy)
```

· By this observation and product rule

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



OHJ-2556 Artificial Intelligence, Spring 2011

10.3.201

 A similar equation holds for the other values of the variable Weather, and hence

```
<u>P(</u>Toothache, Catch, Cavity, Weather) =

<u>P(</u>Toothache, Catch, Cavity) <u>P(</u>Weather)
```

- The required joint distribution tables have 8 and 4 elements
- Propositions a and b are independent if

```
P(a \mid b) = P(a) \Leftrightarrow P(b \mid a) = P(b) \Leftrightarrow P(a \land b) = P(a) P(b)
```

• Respectively variables X and Y are independent of each other if $\underline{P}(X \mid Y) = \underline{P}(X) \Leftrightarrow \underline{P}(Y \mid X) = \underline{P}(Y) \Leftrightarrow \underline{P}(X, Y) = \underline{P}(X)\underline{P}(Y)$

Independent coin flips:

 $\underline{P}(C_1,...,C_n)$ can be represented as the product of n single-variable distributions $P(C_i)$



OHJ-2556 Artificial Intelligence, Spring 2011



13.5 Bayes' Rule and Its Use

- By the product rule $P(a \land b) = P(a \mid b) P(b)$ and the commutativity of conjunction $P(a \land b) = P(b \mid a) P(a)$
- Equating the two right-hand sides and dividing by P(a), we get the Bayes' rule

$$P(b | a) = P(a | b) P(b) / P(a)$$

 The more general case of multivalued variables X and Y conditionalized on some background evidence e

$$P(Y \mid X, e) = P(X \mid Y, e) P(Y \mid e) / P(X \mid e)$$

• Using normalization Bayes' rule can be written as

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



OHJ-2556 Artificial Intelligence, Spring 2011

10.3.2011



· Half of meningitis patients have a stiff neck

$$P(s \mid m) = 0.5$$

• The prior probability of meningitis is 1 / 50 000:

$$P(m) = 1/50000$$

Every 20th patient complains about a stiff neck

$$P(s) = 1/20$$

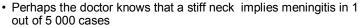
 What is the probability that a patient complaining about a stiff neck has meningitis?

$$P(m \mid s) = P(s \mid m) P(m) / P(s)$$

= 20 / (2 · 50 000) = 0.0002



OHJ-2556 Artificial Intelligence, Spring 2011

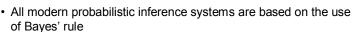


- The doctor, hence, has quantitative information in the *diagnostic* direction from symptoms to causes, and no need to use Bayes' rule
- Unfortunately, diagnostic knowledge is often more fragile than causal knowledge
- If there is a sudden epidemic of meningitis, the unconditional probability of meningitis P(m) will go up
- The conditional probability P(s | m), however, stays the same
- The doctor who derived diagnostic probability P(m | s) directly from statistical observation of patients before the epidemic will have no idea how to update the value
- The doctor who computes P(m | s) from the other three values will see P(m | s) go up proportionally with P(m)



OHJ-2556 Artificial Intelligence, Spring 2011

10.3.201



- On the surface the relatively simple rule does not seem very useful
- However, as the previous example illustrates, Bayes' rule gives a chance to apply existing knowledge
- We can avoid assessing the probability of the evidence P(s) by instead computing a posterior probability for each value of the query variable m and ¬m and then normalizing the result

 $P(M \mid s) = \alpha [P(s \mid m) P(m), P(s \mid \neg m) P(\neg m)]$

- Thus, we need to estimate $P(s \mid \neg m)$ instead of P(s)
- · Sometimes easier, sometimes harder



OHJ-2556 Artificial Intelligence, Spring 2011



- When a probabilistic query has more than one piece of evidence the approach based on full joint probability will not scale up

 P(Cavity | toothache \(\cat{catch} \)
- Neither will applying Bayes' rule scale up in general
 a P(toothache catch | Cavity) P(Cavity)
- We would need variables to be independent, but variable Toothache and Catch obviously are not: if the probe catches in the tooth, it probably has a cavity and that probably cases a toothache
- Each is directly caused by the cavity, but neither has a direct effect on the other
- · catch and toothache are conditionally independent given Cavity



OHJ-2556 Artificial Intelligence, Spring 2011

10.3.2011



- · Conditional independence:
- \underline{P} (toothache \land catch | Cavity) =

P(toothache | Cavity) P(catch | Cavity)

- · Plugging this into Bayes' rule yields
- $P(Cavity \mid toothache \land catch) =$

 $\alpha P(Cavity) P(toothache | Cavity) P(catch | Cavity)$

- Now we only need three separate distributions
- The general definition of conditional independence of variables X and Y, given a third variable Z is

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

• Equivalently, $\underline{P}(X \mid Y, Z) = \underline{P}(X \mid Z)$ and $\underline{P}(Y \mid X, Z) = \underline{P}(Y \mid Z)$



OHJ-2556 Artificial Intelligence, Spring 2011



- If all effects are conditionally independent given a single cause, the exponential size of knowledge representation is cut to linear
- A probability distribution is called a naïve Bayes model if all effects $\mathsf{E_1}$, ..., $\mathsf{E_n}$ are conditionally independent, given a single cause $\mathcal C$
- The full joint probability distribution can be written as $\underline{P}(C, E_1, ..., E_n) = \underline{P}(C) \prod_i \underline{P}(E_i \mid C)$
- It is often used as a simplifying assumption even in cases where the effect variables are not conditionally independent given the cause variable
- In practice, naïve Bayes systems can work surprisingly well, even when the independence assumption is not true



OHJ-2556 Artificial Intelligence, Spring 2011