## Conditional probability

- 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 b

  P(cavity | toothache) = 0.8
- P(cavity) = P(cavity | )
- We can express conditional probabilities in terms of unconditional probabilities:

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

  whenever $P(b) > 0$





- The previous equation can also be written as the *product rule*:

  $P(a \wedge b) = P(a \mid b) P(b)$
- We can, of course, have the rule the other way around

  $P(a \wedge b) = P(b \mid a) P(a)$
- Conditional distributions:

  $\underline{P}(X \mid Y) \equiv P(X = x_i \mid Y = y_j) \; \forall \; i, j$
- By the product rule

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

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

## 13.2.3 Probability Axioms

- The axiomatization of probability theory by Kolmogorov (1933) based on three simple axioms

1. For any proposition $a$ the probability is in between 0 and 1: $0 \leq P(a) \leq 1$
2. Necessarily true (i.e., valid) propositions have probability 1 and necessarily false (i.e., unsatisfiable) propositions have probability 0:
   $P(true) = 1 \qquad P(false) = 0$
3. The probability of a disjunction is given by the *inclusion-exclusion principle*
   $$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$





- We can derive a variety of useful facts from the basic axioms; e.g.:

$$P(a \vee \neg a) = P(a) + P(\neg a) - P(a \wedge \neg a)$$
$$P(true) = P(a) + P(\neg a) - P(false)$$
$$1 = P(a) + P(\neg a)$$
$$P(\neg a) = 1 - P(a)$$

- The fact of the third line can be extended for a discrete variable $D$ with the domain $d_1,\ldots,d_n$:

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

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

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

- The probability distribution on a single variable must sum to 1
- 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 $\mathbf{e}(a)$
- Atomic events are mutually exclusive, so the probability of any conjunction of atomic events is zero, by axiom 2
- Hence, from axiom 3

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

- 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





## 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 $cavity \vee toothache$:
  $0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$
- 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 $\mathbf{Y}$ and $\mathbf{Z}$:

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

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

- 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





- 1/P(toothache) = 1/0.2 = 5 is a normalization constant ensuring that the distribution **P**(Cavity | toothache) adds up to 1

- Let α denote the normalization constant
  **P**(Cavity | toothache) = α **P**(Cavity, toothache)
  = α(**P**(Cavity, toothache, catch) + **P**(Cavity, toothache, ¬catch))
  = α([0.108, 0.016] + [0.012, 0.064])
  = α[0.12, 0.08]
  = [0.6, 0.4]

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

- 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:

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

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





- $\underline{P}(X, \mathbf{e}, \mathbf{y})$ 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

## 13.4 Independence

- 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:

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

- By this observation and product rule

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





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

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

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

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

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

## 13.5 Bayes' Rule and Its Use

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

$$P(b \mid a) = P(a \mid b)\, P(b) \,/\, P(a)$$

- The more general case of multivalued variables $X$ and $Y$ conditionalized on some background evidence $\mathbf{e}$

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

- Using normalization Bayes' rule can be written as

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





- Half of meningitis patients have a stiff neck
  $P(s \mid m) = 0.5$
- The prior probability of meningitis is 1 / 50 000:
  $P(m) = 1/50\,000$
- 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?

$$P(m \mid s) = P(s \mid m)\, P(m) \,/\, P(s)$$
$$= 20 \,/\, (2 \cdot 50\,000) = 0.0002$$

- Perhaps the doctor knows that a stiff neck implies meningitis in 1 out of 5 000 cases
- 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 $P(s \mid m)$, however, stays the same
- The doctor who derived diagnostic probability $P(m \mid s)$ directly from statistical observation of patients before the epidemic will have no idea how to update the value
- The doctor who computes $P(m \mid s)$ from the other three values will see $P(m \mid s)$ go up proportionally with $P(m)$





- All modern probabilistic inference systems are based on the use of Bayes' rule
- 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 $\neg m$ – and then normalizing the result

$$\underline{\mathbf{P}}(M \mid s) = \alpha \, [\, P(s \mid m)\, P(m),\, P(s \mid \neg m)\, P(\neg m)\, ]$$

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

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

$$\mathbf{P}(Cavity \mid toothache \wedge catch)$$

- Neither will applying Bayes' rule scale up in general

$$\alpha\, \mathbf{P}(toothache \wedge catch \mid Cavity)\, \mathbf{P}(Cavity)$$

- 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*





- Conditional independence:

$$\mathbf{P}(toothache \wedge catch \mid Cavity) = \mathbf{P}(toothache \mid Cavity)\, \mathbf{P}(catch \mid Cavity)$$

- Plugging this into Bayes' rule yields

$$\mathbf{P}(Cavity \mid toothache \wedge catch) = \alpha\, \mathbf{P}(Cavity)\, \mathbf{P}(toothache \mid Cavity)\, \mathbf{P}(catch \mid Cavity)$$

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

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

- Equivalently, $\mathbf{P}(X \mid Y, Z) = \mathbf{P}(X \mid Z)$ and $\mathbf{P}(Y \mid X, Z) = \mathbf{P}(Y \mid Z)$

- 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 $E_1, \ldots, E_n$ are conditionally independent, given a single cause $C$
- The full joint probability distribution can be written as

$$\underline{\mathbf{P}}(C, E_1, \ldots, E_n) = \underline{\mathbf{P}}(C) \prod_i \underline{\mathbf{P}}(E_i \mid C)$$

- 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