Quantifying Uncertainty

Melih Kandemir

University of Southern Denmark
Department of Mathematics and Computer Science (IMADA)

Outline

- Acting under uncertainty
- Basic probability notation
- Inference using full joint distributions
- Independence
- Bayes' rule and its use
- Naive Bayes models
- The Wumpus World revisited

Acting under uncertainty

Real-world problems contain **uncertainties** due to:

- Partial observability
- Nondeterminism
- Adversaries

Example of dental diagnosis using propositional logic:

$\mathit{toothache} \Rightarrow \mathit{cavity}$

However, this rule is inaccurate since not all patients with toothaches have cavities:

$\mathit{toothache} \Rightarrow \mathit{cavity} \lor \mathit{gum\ problem} \lor \mathit{abscess}...$

To make the rule true, we would need an exhaustive list of all possible problems.

Acting under uncertainty

An agent strives to choose the right thing to do—the rational decision— depends on both the relative importance of various goals and the likelihood that, and degree to which, they will be achieved.

- Large domains such as medical diagnosis fail for three main reasons:
    - **Laziness:** It is too much work to list the complete set of antecedents or consequents needed to ensure an exceptionless rule.
    - **Theoretical ignorance:** Medical science has no complete theory for the domain.
    - **Practical ignorance:** Even if we know all the rules, we might be uncertain about a particular patient because not all the necessary tests have been or can be run.
- An agent only has a degree of belief in the relevant sentences.

Acting under uncertainty

Probability Theory:

- Tool to deal with degrees of belief of relevant sentences.
- Summarizes the uncertainty that comes from our laziness and ignorance.

Uncertainty and Rational Decisions:

- An agent requires **preference** among different possible outcomes of various plans.
- **Utility Theory:** The quality of the outcome being useful.
  - Every state has a degree of usefulness/utility.
  - Higher utility is preferred.
- **Decision Theory:** Preferences (utility theory) + probabilities.
  - *Decision theory = probability theory + utility theory.*
  - An agent is **rational** if and only if it chooses the action that **yields the highest expected utility**, averaged over all **possible outcomes**.
  - Principle of Maximum Expected Utility (MEU).

Decision-theoretic agent

```
Algorithm DT-AGENT( percept) returns an action
Input: belief state: probabilistic beliefs about the current state of the world,
  action: the agent's action
Update belief state based on action and percept
Calculate outcome probabilities for actions given belief state
Select action with highest expected utility, given probabilities of outcomes and
  utility information
return action
```

Basic probability notation

- For our agent to represent and use probabilistic information, we need a formal language.
- A **sample space** Ω is the set of all possible worlds ω. The possible worlds are mutually exclusive and exhaustive.
- A probability function assigns a numerical score $P(\omega)$ between 0 and 1 such that:
    - The sum of the probabilities of all possible outcomes is one: $\sum_{\omega \in \Omega} P(\omega) = 1$
    - The probability of the sure event is one: $P(\Omega) = 1$
    - The probability of the impossible event is zero: $P(\emptyset) = 0$
- **Unconditional probability**: degrees of belief in propositions in the absence of any other information.

Basic probability notation

- **Conditional probability**: given evidence that has happened, degree of belief in a new event: Make use of unconditional probabilities.
- Probability of a given b:

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

- Can also be written as:

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

- Example of rolling fair dice, rolling doubles when the first die is 5:

$$P(\text{doubles}|\text{Die}_1 = 5) = \frac{P(\text{doubles} \wedge \text{Die}_1 = 5)}{P(\text{Die}_1 = 5)}$$

Basic probability notation

- **Factored representation**: A possible world is represented by a set of variable/value pairs. Variables in probability theory are called **random variables**, and their names begin with an uppercase letter (e.g., Total and Die$_1$).
- Sometimes we will want to talk about the probabilities of all the possible values of a random variable. We could write:

$$P(\text{Weather} = \text{sun}) = 0.6$$

$$P(\text{Weather} = \text{rain}) = 0.1$$

$$P(\text{Weather} = \text{cloud}) = 0.29$$

$$P(\text{Weather} = \text{snow}) = 0.01$$

- Abbreviation of this will be:

$$P(\text{Weather}) = (0.6, 0.1, 0.29, 0.01)$$

- The P statement defines a **probability distribution** for the random variable *Weather*.

Inference using full joint distributions

- Start with the 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

- For any proposition $A \subseteq \Omega$, sum the atomic events where it is true:

$$P(A) = \sum_{\omega \in A} P(\omega)$$

$$\begin{aligned} P(\text{toothache}) &= 0.108 + 0.012 + 0.016 + 0.064 = 0.2 \\ P(\text{cavity} \vee \text{toothache}) &= 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 \\ &= 0.28 \end{aligned}$$

Inference using full joint distributions

- Start with the 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

- One can also compute conditional probabilities:

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

Inference using full joint distributions

- Start with the 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

- Denominator can be viewed as a normalization constant α:

$$
\begin{aligned}
&P(\text{cavity}|\text{toothache}) = \alpha P(\text{cavity}, \text{toothache}) \\
&= \alpha \Big[P(\text{cavity}, \text{toothache}, \text{catch}) + P(\text{cavity}, \text{toothache}, \neg\text{catch}) \Big] \\
&= \alpha \left[(0.108, 0.016) + (0.012, 0.064) \right] \\
&= \alpha(0.12, 0.08) = (0.6, 0.4)
\end{aligned}
$$

- **General idea:** Compute the distribution on the query variable by fixing **evidence variables** and summing over **hidden variables**.

Inference using full joint distributions

- Let X be all the variables. Typically, we want the joint distribution of the **query variables** Y given specific values e for the **evidence variables** E.
- Let the **hidden variables** be $H = X \backslash (Y \cup E)$.
- Then the required summation of joint entries is done by summing out the hidden variables:

$$P(Y|E=e) = \alpha P(Y, E=e) = \alpha \sum_h P(Y, E=e, H=h)$$

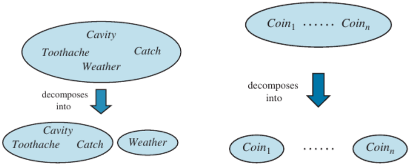
- The terms in the summation are joint entries because Y, E, and H together exhaust the set of random variables.
- Obvious problems:
  1. Worst-case time complexity $O(d^n)$, where d is the largest arity and n the number of variables.
  2. Space complexity $O(d^n)$ to store the joint distribution.
  3. How to find the numbers for $O(d^n)$ entries?

Independence

- Two examples of factoring a large joint distribution into smaller distributions, using absolute independence: (a) Weather and dental problems are independent. (b) Coin flips are independent.
- $P(a|b) = P(a)$ or $P(b|a) = P(b)$ or $P(a \wedge b) = P(a)P(b)$
- One's dental problems influence the weather thus:

$$P(a|b) = P(a) \quad \text{or} \quad P(b|a) = P(b) \quad \text{or} \quad (a \wedge b) = P(a)P(b)$$

$$P(\text{toothache, catch, cavity, cloud})$$
$$= P(\text{cloud}|\text{toothache, catch, cavity})P(\text{toothache, catch, cavity})$$
$$P(\text{cloud}|\text{toothache, catch, cavity}) = P(\text{cloud})$$
$$P(\text{toothache, catch, cavity, cloud}) = P(\text{cloud})P(\text{toothache, catch, cavity})$$



Bayes' rule and its use

- Bayes' rule is derived from the product rule:

$$P(a \wedge b) = P(a|b)P(b) \quad \text{and} \quad P(a \wedge b) = P(b|a)P(a)$$

- Equating the two right-hand sides and dividing by $P(a)$, we get:

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

- Often, we perceive as evidence the effect of some unknown cause and we would like to determine that cause. In that case, Bayes' rule becomes:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- The conditional probability $P(\text{effect}|\text{cause})$ quantifies the relationship in the **causal** direction, whereas $P(\text{cause}|\text{effect})$ describes the **diagnostic** direction.

Bayes' rule and its use

- For example, a doctor knows that the disease meningitis causes a patient to have a stiff neck, say, 70% of the time. The doctor also knows some unconditional facts: the prior probability that any patient has meningitis is $\frac{1}{50{,}000}$, and the prior probability that any patient has a stiff neck is 1%.
- Let s be the proposition that the patient has a stiff neck and m be the proposition that the patient has meningitis. We have:

$$P(s|m) = 0.7, \quad P(m) = \frac{1}{50,000}, \quad P(s) = 0.01$$

- Using Bayes' rule, we can calculate:

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.7 \times \frac{1}{50{,}000}}{0.01} = 0.0014$$

- That is, we expect only 0.14% of patients with a stiff neck to have meningitis. Notice that even though a stiff neck is quite strongly indicated by meningitis (with probability 0.7), the probability of meningitis in patients with stiff necks remains small. This is because the prior probability of stiff necks (from any cause) is much higher than the prior for meningitis.

Bayes' rule and conditional independence

- We start with the following expression:

$$P(\text{cavity}|\text{toothache} \land \text{catch}) = \alpha P(\text{toothache} \land \text{catch}|\text{cavity})P(\text{cavity})$$

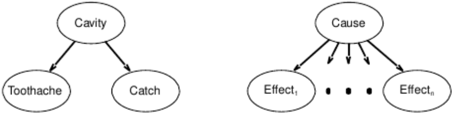
- Applying the chain rule of probability, we get:

$$= \alpha P(\text{toothache}|\text{cavity})P(\text{catch}|\text{cavity})P(\text{cavity})$$

- This is an example of a **naive Bayes** model:

$$P(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = P(\text{Cause})\prod_i P(\text{Effect}_i|\text{Cause})$$

- In this model, the total number of parameters is linear in n, the number of effects.



Naive Bayes models

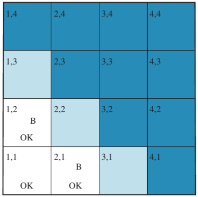
- The full joint distribution can be written as:

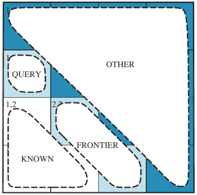
$$P(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i | \text{Cause})$$

- Such a probability distribution is called a *naive Bayes model*—“naive” because it is often used (as a simplifying assumption) in cases where the ”effect” variables are not strictly independent given the cause variable.
- Call the observed effects $E = e$, while the remaining effect variables Y are unobserved.

$$\begin{aligned} P(\text{Cause}|e) &= \alpha \sum_y P(\text{Cause}) P(y|\text{Cause}) \prod_j P(e_j|\text{Cause}) \\ &= \alpha P(\text{Cause}) \prod_j P(e_j|\text{Cause}) \sum_y P(y|\text{Cause}) \\ &= \alpha P(\text{Cause}) \prod_j P(e_j|\text{Cause}) \end{aligned}$$

The Wumpus World revisited





- Let P_{ij} be true if and only if the square $[i, j]$ contains a pit.
- Let B_{ij} be true if and only if the square $[i, j]$ is breezy.
- Apply the product rule on the full joint distribution

$$P(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1}) = \\ P(B_{1,1}, B_{1,2}, B_{2,1} | P_{1,1}, \ldots, P_{4,4}) P(P_{1,1}, \ldots, P_{4,4})$$

- First term: 1 if pits are adjacent to breezes, 0 otherwise.
- Second term: pits are placed randomly with probability 0.2 per square: $P(P_{1,1}, \ldots, P_{4,4}) = \prod_{i,j} P(P_{i,j}) = 0.2^n \times 0.8^{(16-n)}$ for n pits.

Observations and query

We know the following facts:

$$b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1} \quad \text{and} \quad \text{known} = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$$

Query is $P(p_{1,3}|\text{known}, b)$. Define "unknown" as the set of P_{ij}'s other than $P_{1,3}$ and known. Then

$$P(b|p_{1,3}, \text{known}, \text{unknown}) = P(b|p_{1,3}, \text{known}, \text{frontier})$$

For inference by enumeration, we have:

$$P(p_{1,3}|\text{known}, b) = \alpha \sum_{\text{unknown}} P(p_{1,3}, \text{unknown}, \text{known}, b)$$

This grows exponentially with the number of squares! Basic insight is that observations are conditionally independent of other hidden squares given neighboring hidden squares. Manipulate query into a form where we can use this.

Using conditional independence

$$P(p_{1,3}|\text{known}, b) = \alpha \sum_{\text{unknown}} P(p_{1,3}, \text{known}, b, \text{unknown})$$

$$= \alpha \sum_{\text{unknown}} P(b|p_{1,3}, \text{known}, \text{unknown}) P(p_{1,3}, \text{known}, \text{unknown})$$

$$= \alpha \sum_{\text{frontier,other}} P(b|\text{known}, p_{1,3}, \text{frontier}, \text{other}) P(p_{1,3}, \text{known}, \text{frontier}, \text{other})$$

$$= \alpha \sum_{\text{frontier,other}} P(b|\text{known}, p_{1,3}, \text{frontier}) P(p_{1,3}, \text{known}, \text{frontier}, \text{other})$$

where the final step uses the fact that b is independent of ”other” given ”known”, $p_{1,3}$, and ”frontier”. The first term does not depend on ”other”, so the summation can move inward.

$$P(p_{1,3}|\text{known}, b)$$
$$= \alpha \sum_{\text{frontier}} P(b|\text{known}, p_{1,3}, \text{frontier}) \sum_{\text{other}} P(p_{1,3}, \text{known}, \text{frontier}, \text{other})$$

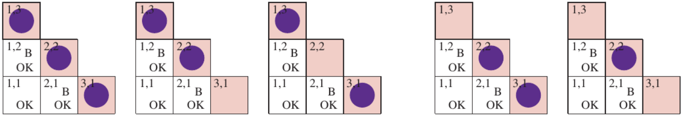
Using conditional independence

By independence, the term on the right can be factored and reordered:

$$
\begin{aligned}
&P(p_{1,3}|\text{known}, b) \\
&= \alpha \sum_{\text{frontier}} P(b|\text{known}, p_{1,3}, \text{frontier}) \sum_{\text{other}} P(p_{1,3}) P(\text{known}) P(\text{frontier}) P(\text{other}) \\
&= \alpha P(\text{known}) P(p_{1,3}) \sum_{\text{frontier}} P(b|\text{known}, p_{1,3}, \text{frontier}) P(\text{frontier}) \sum_{\text{other}} P(\text{other}) \\
&= \alpha' P(p_{1,3}) \sum_{\text{frontier}} P(b|\text{known}, p_{1,3}, \text{frontier}) P(\text{frontier})
\end{aligned}
$$

where we use $P(\text{known}) = 1$ and $\sum_{\text{other}} P(\text{other}) = 1$.

Using conditional independence



$0.2 \times 0.2 = 0.04$ $0.2 \times 0.8 = 0.16$ $0.8 \times 0.2 = 0.16$ $0.2 \times 0.2 = 0.04$ $0.2 \times 0.8 = 0.16$

The probabilities in $P(b|\text{known}, p_{1,3}, \text{frontier})$ are 1 when the breeze observations are consistent with the other variables and 0 otherwise. Thus, for each value of $p_{1,3}$, we sum over the logical models for the frontier variables that are consistent with the known facts. We have:

$$\begin{aligned} P(p_{1,3}|\text{known}, b) &= \alpha' \langle 0.2(0.04 + 0.16 + 0.16), 0.8(0.04 + 0.16) \rangle \\ &\approx \langle 0.31, 0.69 \rangle \end{aligned}$$

That is, $[1, 3]$ (and $[3, 1]$ by symmetry) contains a pit with $\approx 31\%$ probability. A similar calculation shows that $[2, 2]$ contains a pit with roughly 86% probability. The wumpus agent should definitely avoid $[2, 2]$! A logical agent cannot not know that $[2, 2]$ is worse than the other squares without probabilistic inference.

Summary

- **Probabilities** express the agent's inability to reach a definite decision regarding the truth of a sentence.
- **Decision theory** combines the agent's beliefs and desires, defining the best action as the one that maximizes expected utility.
- Basic probability statements include **prior or unconditional probabilities** and **posterior or conditional probabilities** over simple and complex propositions.
- The axioms of probability constrain the probabilities of logically related propositions.

Summary cont'd

- The **full joint probability distribution** specifies the probability of each complete assignment of values to random variables.
- **Absolute independence** between subsets of random variables factorizes the full joint distribution into smaller joint distributions, greatly reducing its complexity.
- **Bayes' rule** allows unknown probabilities to be computed from known conditional probabilities, usually in the causal direction.
- **Conditional independence** brought about by direct causal relationships in the domain allows the full joint distribution to be factored into smaller, conditional distributions.