CS 161 Fundamentals of Artificial Intelligence
Lecture 11

Uncertainty Quantification

Quanquan Gu

Department of Computer Science
UCLA

March 8th, 2022

Teaching/Course Evaluation

- The online evaluations has opened last week and will close 8:00 AM Saturday, March 12.
- Your feedback is valuable and more than welcome!
- Students who have done the online evaluation will receive 1 extra point in the final grade:)

Outline

- Uncertainty
- Probability
- Syntax and Semantics
- Inference
- Independence and Bayes' Rule

Uncertainty

Let action A_t = leave for airport t minutes before flight
Will A_t get me there on time?
Hence a purely logical approach:

- "A_{25} will get me there on time", may be wrong!
- "A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact, etc.", there exists uncertainty!

How to handle uncertainty?

Methods for handling uncertainty

Making **assumptions**:

- Assume my car does not have a flat tire
- Assume A_{25} works unless contradicted by evidence

Issues: What assumptions are reasonable? How to handle contradiction?

Probability can help us!

- Given the available evidence, A_{25} will get me there on time with probability 0.04

What is probability?

Probability

- Probabilistic assertions **summarize** effects of
  laziness: failure to enumerate exceptions, qualifications, etc.
  ignorance: lack of relevant facts, initial conditions, etc.
- **Subjective** or **Bayesian** probability:

Probabilities relate propositions to one's own state of knowledge

e.g., $P(A_{25}|\text{no reported accidents}) = 0.06$

- Probabilities of propositions change with new evidence:

e.g., $P(A_{25}|\text{no reported accidents, 5 a.m.}) = 0.15$

Making decisions under uncertainty

- Suppose I believe the following:

$$\begin{aligned}
P(A_{25} \text{ gets me there on time}|\ldots) &= 0.04 \\
P(A_{90} \text{ gets me there on time}|\ldots) &= 0.70 \\
P(A_{120} \text{ gets me there on time}|\ldots) &= 0.95 \\
P(A_{1440} \text{ gets me there on time}|\ldots) &= 0.9999
\end{aligned}$$

Which action to choose?

- Depends on my **preferences** for missing flight vs. airport cuisine, etc.
- **Utility theory** says that every state has a degree of usefulness, or utility, to an agent and that the agent will prefer states with higher utility.
- **Decision theory** = utility theory + probability theory

Probability basics

- Begin with a set Ω—the **sample space**
  e.g., 6 possible rolls of a die.
  $\omega \in \Omega$ is a **sample point/possible world/atomic event**
- A **probability space** or **probability model** is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ s.t.
  $0 \leq P(\omega) \leq 1$
  $\sum_{\omega} P(\omega) = 1$
  e.g., $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$.
- An **event** A is any subset of Ω

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

E.g.,
$P(\text{die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

Random variables

- A **random variable** is a function from sample points to some range, e.g., the reals or Booleans
  e.g., $Odd(1) = true$.
- P induces a **probability distribution** for any r.v. X:

$$P(X = x_i) = \sum_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

e.g.,
$P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$

Propositions

- We call events in probability as **propositions** in AI.
  - Events: given Boolean random variables A and B:
    - Event a = set of sample points where $A(\omega) = true$
    - Event $\neg a$ = set of sample points where $A(\omega) = false$
    - Event $a \wedge b$ = points where $A(\omega) = true$ and $B(\omega) = true$
- With Boolean variables, sample point = propositional logic model
  e.g., $A = true$, $B = false$, or $a \wedge \neg b$.
- Proposition = disjunction of events in which it is true
  e.g., $(a \vee b) \equiv (\neg a \wedge b) \vee (a \wedge \neg b) \vee (a \wedge b)$
  $\Rightarrow P(a \vee b) = P(\neg a \wedge b) + P(a \wedge \neg b) + P(a \wedge b)$

Syntax for propositions

- **Propositional** or **Boolean** random variables
  e.g., $Cavity$ (do I have a cavity?)
  $Cavity = true$ is a proposition, also written $cavity$
- **Discrete** random variables (finite or infinite)
  e.g., $Weather$ is one of $\langle sunny, rain, cloudy, snow \rangle$
  $Weather = rain$ is a proposition
  Values must be exhaustive and mutually exclusive
- **Continuous** random variables (bounded or unbounded)
  e.g., $Temp = 21.6$; also allow, e.g., $Temp < 22.0$.
- Arbitrary Boolean combinations of basic propositions

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)
  - Tips: we use P to denote probability for specific event, **P** to denote the distribution over all events– a vector!
- **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

Conditional probability

- **Conditional** or **posterior probabilities**
  e.g., $P(cavity|toothache) = 0.8$
  - "if $toothache$ and **no further information we have**, then 80% chance of $cavity$"
  - **NOT** "if $toothache$ then 80% chance of $cavity$"!
    - If we know more, e.g., $cavity$ is also given, then we have $P(cavity|toothache, cavity) = 1$
- New evidence may be irrelevant, allowing simplification, e.g.,
  $P(cavity|toothache, LakersWin) = P(cavity|toothache) = 0.8$

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:
  $\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1, \ldots, X_{n-1})\ \mathbf{P}(X_n|X_1, \ldots, X_{n-1})$
  $=$
$\mathbf{P}(X_1, \ldots, X_{n-2})\ \mathbf{P}(X_{n-1}|X_1, \ldots, X_{n-2})\ \mathbf{P}(X_n|X_1, \ldots, X_{n-1})$
  $= \ldots$
  $= \prod_{i=1}^{n} \mathbf{P}(X_i|X_1, \ldots, X_{i-1})$

Inference by enumeration

- A naive way of doing probablistic inference is inference by enumeration.
- Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

Here catch: the dentist's nasty steel probe catches in tooth.

- For any proposition ϕ, sum the atomic events where it is true:

$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$

Inference by enumeration

- Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

- For any proposition ϕ, sum the atomic events where it is true:

$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$

$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

Inference by enumeration

- Start with the joint distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

- For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \sum_{\omega:\omega\models\phi} P(\omega)$$

$$P(cavity \vee toothache) =$$
$$0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

Inference by enumeration

- Start with the joint distribution:

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

- Can also compute conditional probabilities:

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

Normalization

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
$\neg$ *cavity*	**.016**	**.064**	**.144**	**.576**

- Denominator can be viewed as a normalization constant α

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

- General idea: compute distribution on query variable

by fixing **evidence variables** and summing over **hidden variables**

Inference by enumeration, contd.

- Let $\mathbf{X}$ be all the variables. Typically, we want the posterior joint distribution of the **query variables** $\mathbf{Y}$ given specific values $\mathbf{e}$ for the **evidence variables** $\mathbf{E}$
- Let the **hidden variables** be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

Then the required summation of joint entries is done by summing out the hidden variables:

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

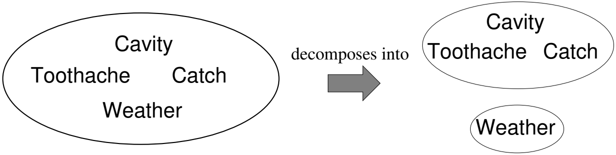
- The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$, and $\mathbf{H}$ together exhaust the set of random variables
- Obvious problems:

1) Worst-case time complexity $O(d^n)$ where d is the largest domain size
2) Space complexity $O(d^n)$ to store the joint distribution
3) How to find the numbers for $O(d^n)$ entries???

Independence

- Those problems can be solved by exploring independencies between variables.

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 is powerful but rare

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

Conditional independence

- $\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

  (1) $P(catch|toothache, cavity) = P(catch|cavity)$
- The same independence holds if I haven't got a cavity:

  (2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

$Catch$ is **conditionally independent** of $Toothache$ given $Cavity$:

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

- Equivalent statements:

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

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

Conditional independence contd.

- Write out full joint distribution using chain rule:

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

i.e., $2 + 2 + 1 = 5$ independent numbers

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.

Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Bayes' Rule

- Bayes' Rule can be used in probability inference when we have $P(b|a)$ but not $P(a|b)$.
- Product rule $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

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

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:

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

e.g., let C be COVID-19, M be cough:

$$P(c|m) = \frac{P(m|c)P(c)}{P(m)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of COVID-19 still very small!

Bayes' Rule and conditional independence

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

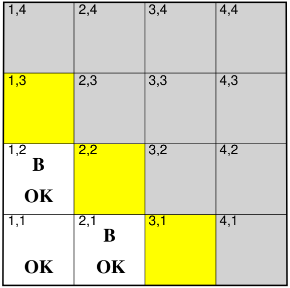
This is an example of a **naive Bayes** model:

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

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

Example: Wumpus World

• Suppose we have the information of $B_{1,1}, B_{1,2}, B_{2,1}$ and we want to know the probability distributions of the unknown squares.



$P_{ij} = true$ iff square $[i, j]$ contains a pit
$B_{ij} = true$ iff square $[i, j]$ is breezy
Include only $B_{1,1}, B_{1,2}, B_{2,1}$ in the probability model

Specifying the probability model

- The full joint distribution is $\mathbf{P}(P_{1,1}, \ldots, P_{4,4}, B_{1,1}, B_{1,2}, B_{2,1})$
- Apply product rule:

$\mathbf{P}(B_{1,1}, B_{1,2}, B_{2,1} \mid P_{1,1}, \ldots, P_{4,4})\mathbf{P}(P_{1,1}, \ldots, P_{4,4})$
(Do it this way to get $P(Effect|Cause)$.)

  - First term: 1 if pits are adjacent to breezes, 0 otherwise
  - Second term: pits are placed randomly, probability 0.2 per square:

$$\mathbf{P}(P_{1,1}, \ldots, P_{4,4}) = \prod_{i,j=1,1}^{4,4} \mathbf{P}(P_{i,j}) = 0.2^n \times 0.8^{16-n}$$

for n pits.

Observations and query

- We know the following facts:
  $b = \neg b_{1,1} \wedge b_{1,2} \wedge b_{2,1}$
  $known = \neg p_{1,1} \wedge \neg p_{1,2} \wedge \neg p_{2,1}$
- Query is $\mathbf{P}(P_{1,3}|known, b)$
- Define $Unknown = P_{ij}$'s other than $P_{1,3}$ and $known$
- For inference by enumeration, we have

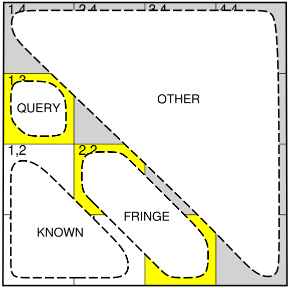
$$\mathbf{P}(P_{1,3}|known, b) = \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b)$$

Grows exponentially with number of squares!

- How to solve this problem? Use conditional independence relationships between variables.

Using conditional independence

• Basic insight: observations are conditionally independent of other hidden squares given neighbouring hidden squares

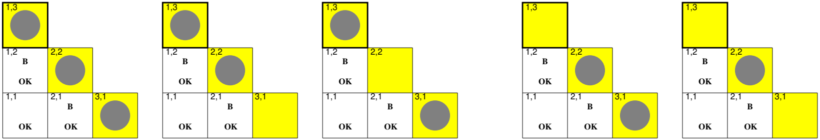


• Define $Unknown = Fringe \cup Other$
$\mathbf{P}(b|P_{1,3}, Known, Unknown) = \mathbf{P}(b|P_{1,3}, Known, Fringe)$
• Manipulate query into a form where we can use this!

Using conditional independence contd.

$$
\begin{aligned}
\mathbf{P}(P_{1,3}|known, b) &= \alpha \sum_{unknown} \mathbf{P}(P_{1,3}, unknown, known, b) \\
&= \alpha \sum_{unknown} \mathbf{P}(b|P_{1,3}, known, unknown)\mathbf{P}(P_{1,3}, known, unknown) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe, other)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \sum_{other} \mathbf{P}(b|known, P_{1,3}, fringe)\mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3}, known, fringe, other) \\
&= \alpha \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe) \sum_{other} \mathbf{P}(P_{1,3})P(known)P(fringe)P(other) \\
&= \alpha\, P(known)\mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe) \sum_{other} P(other) \\
&= \alpha'\, \mathbf{P}(P_{1,3}) \sum_{fringe} \mathbf{P}(b|known, P_{1,3}, fringe)P(fringe)
\end{aligned}
$$

Using conditional independence contd.



0.2 x 0.2 = 0.04   0.2 x 0.8 = 0.16   0.8 x 0.2 = 0.16   0.2 x 0.2 = 0.04   0.2 x 0.8 = 0.16

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

Summary

- Probability is a rigorous formalism for uncertain knowledge
- **Joint probability distribution** specifies probability of every **atomic event**
- Queries can be answered by summing over atomic events
- For nontrivial domains, we must find a way to reduce the joint size. **Independence** and **conditional independence** provide the tools to achieve this.

Acknowledgment

The slides are adapted from Stuart Russell