Lecture 5: Uncertainty
CS486/686 Intro to Artificial Intelligence

2023-5-25

Pascal Poupart
David R. Cheriton School of Computer Science

UNIVERSITY OF WATERLOO

Outline

- Probability theory
- Uncertainty via probabilities
- Probabilistic inference

Terminology

- **Probability distribution:**
  - A specification of a probability for each event in our sample space
  - Probabilities must sum to 1
- Assume the world is described by two (or more) random variables
  - **Joint probability distribution**
    - Specification of probabilities for all combinations of events

Joint distribution

- Given two random variables A and B:
- Joint distribution:

  $\Pr(A = a \wedge B = b)$ for all a, b

- **Marginalisation (sumout rule):**

  $\Pr(A = a) = \Sigma_b \Pr(A = a \wedge B = b)$

  $\Pr(B = b) = \Sigma_a \Pr(A = a \wedge B = b)$

Example: Joint Distribution

sunny

	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny

	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

P(headache Λ sunny Λ cold) =

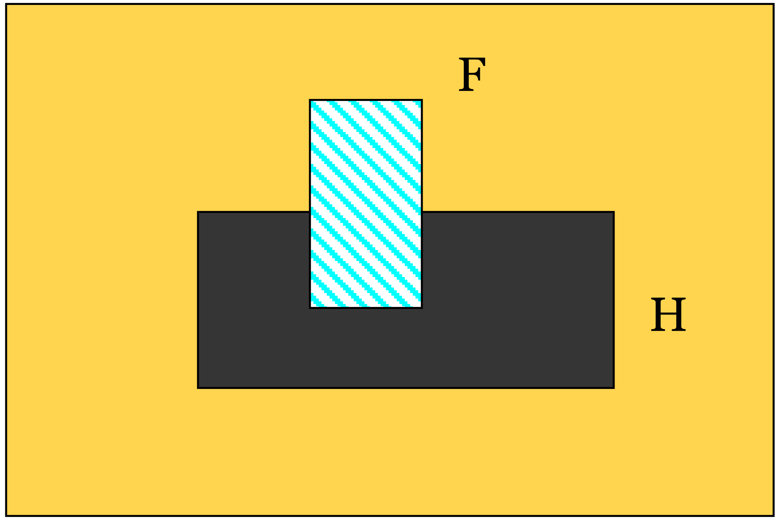
P(~headache Λ sunny Λ ~cold) =

P(headache) =

marginalization

Conditional Probability

- $\Pr(A|B)$: fraction of worlds in which B is true that also have A true

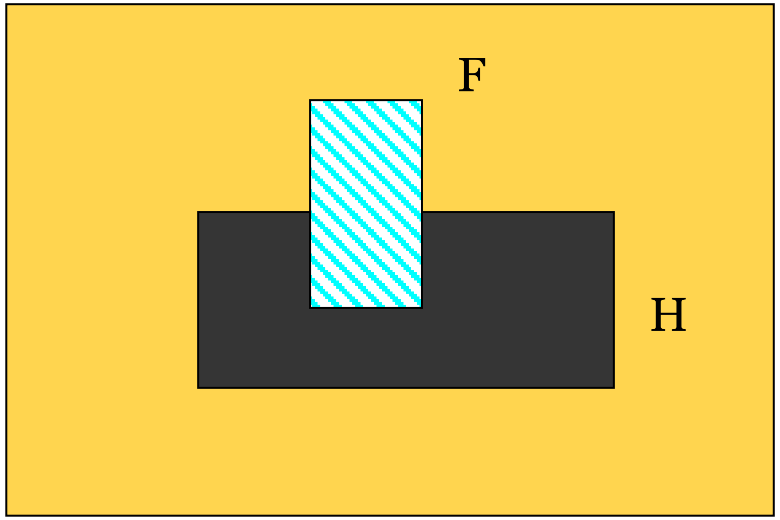


H = "Have headache"
F = "Have Flu"

$$\Pr(H) = 1/10$$
$$\Pr(F) = 1/40$$
$$\Pr(H|F) = 1/2$$

Headaches are rare and flu is rarer, but if you have the flu, then there is a 50-50 chance you will have a headache

Conditional Probability



H = "Have headache"
F = "Have Flu"

$$Pr(H) = 1/10$$
$$Pr(F) = 1/40$$
$$Pr(H|F) = 1/2$$

$Pr(H|F)$ = Fraction of flu inflicted worlds in which you have a headache
= (# worlds with flu and headache)/(# worlds with flu)
= (Area of "H and F" region)/(Area of "F" region)
= $Pr(H \land F) / Pr(F)$

UNIVERSITY OF WATERLOO

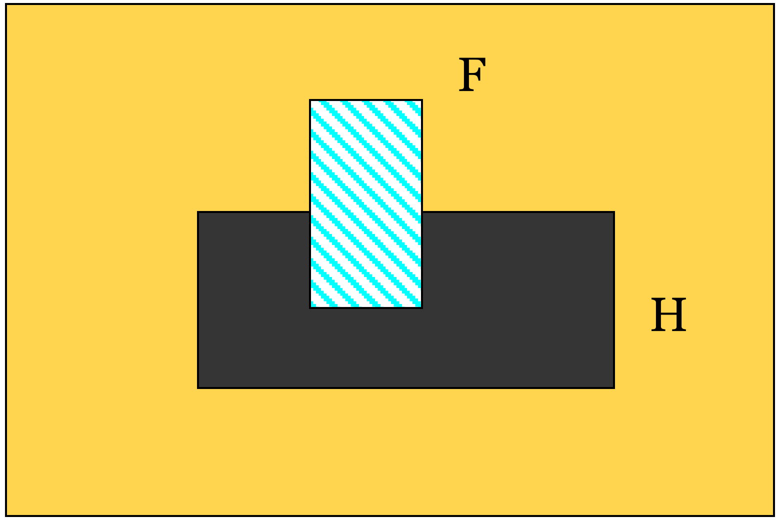
Conditional Probability

- Definition: $\Pr(A|B) = \Pr(A \wedge B) / \Pr(B)$

- Chain rule: $\Pr(A \wedge B) = \Pr(A|B)\Pr(B)$

Memorize these rules!

Inference



H = “Have headache”
F = “Have Flu”

$\Pr(H) = 1/10$
$\Pr(F) = 1/40$
$\Pr(H|F) = 1/2$

One day you wake up with a headache. You think “Drat! 50% of flues are associated with headaches so I must have a 50-50 chance of coming down with the flu”

Is your reasoning correct?

$\Pr(F \wedge H) =$

$\Pr(F|H) =$

Example: Conditional Distribution

sunny

	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny

	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

$$\Pr(headache \wedge cold \mid sunny) =$$

$$\Pr(headache \wedge cold \mid \sim sunny) =$$

Bayes Rule

- Note: $\Pr(A|B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \wedge A) = \Pr(B|A)Pr(A)$

- Bayes Rule: $\Pr(B|A) = \frac{\Pr(\mathrm{A}|\mathrm{B})\Pr(B)}{\Pr(A)}$

Memorize this!

Using Bayes' Rule for inference

- Often, we want to form a hypothesis about the world based on what we have observed
- Bayes' rule allows us to compute a belief about hypothesis H, given evidence e

Likelihood

Prior probability

$$P(H|e) = \frac{P(e|H)P(H)}{P(e)}$$

Posterior probability

Normalizing constant

More General Forms of Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}$$

$$P(A|B \land X) = \frac{P(B|A \land X)P(A|X)}{P(B|X)}$$

$$P(A = v_i|B) = \frac{P(B|A = v_i)P(A = v_i)}{\sum_{k=1}^{n} P(B|A = v_k)P(A = v_k)}$$

Probabilistic Inference

- By probabilistic inference, we mean
  - given a *prior* distribution $\Pr(\boldsymbol{X})$ over variables $\boldsymbol{X}$ of interest, representing degrees of belief
  - and given new evidence $E = e$ for some variable E
  - Revise your degrees of belief: *posterior* $\Pr(\boldsymbol{X}|E = e)$
- Applications:
  - Medicine: $\Pr(disease|symptom1, symptom2, \ldots, symptomN)$
  - Troubleshooting: $\Pr(cause|test1, test2, \ldots, testN)$

UNIVERSITY OF WATERLOO

Issues

- How do we specify the full joint distribution over a set of random variables $X_1, X_2, \ldots, X_n$?
  - Exponential number of possible worlds
  - e.g., if X_i is Boolean, then 2^n numbers (or $2^n - 1$ parameters, since they sum to 1)
  - These numbers are not robust/stable
- Inference is frightfully slow
  - Must sum over exponential number of worlds to answer queries
    - $\Pr(X_i) = \sum_{X_1} \ldots \sum_{X_{i-1}} \sum_{X_{i+1}} \ldots \sum_{X_n} \Pr(X_1, X_2, \ldots, X_n)$
    - $Pr(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n | X_i) = \frac{P(X_1, \ldots, X_n)}{P(X_i)} = \frac{P(x_1, \ldots, X_n)}{\sum_{X_1} \ldots \sum_{X_{i-1}} \sum_{X_{i+1}} \ldots \sum_{X_n} \Pr(X_1, \ldots, X_n)}$

Small Example: 3 Variables

sunny

	cold	~cold
headache	0.108	0.012
~headache	0.016	0.064

~sunny

	cold	~cold
headache	0.072	0.008
~headache	0.144	0.576

$$\Pr(headache) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

$$\Pr(headache \wedge cold | sunny) = \Pr(headache \wedge cold \wedge sunny) \,/\, \Pr(sunny)$$

$$= 0.108/(0.108 + 0.012 + 0.016 + 0.064) = 0.54$$

$$\Pr(headache \wedge cold | {\sim}sunny) = \Pr(headache \wedge cold \wedge {\sim}sunny) \,/\, \Pr({\sim}sunny)$$

$$= 0.072/(0.072 + 0.008 + 0.144 + 0.576) = 0.09$$

Intractable Inference

- How do we avoid the exponential blow up of joint distribution and probabilistic inference?
  - no solution in general
  - but in practice there is structure we can exploit

- We'll use conditional independence

UNIVERSITY OF WATERLOO

Independence

- Recall that X and Y are *independent* iff:

$$\Pr(X = x) = \Pr(X = x|Y = y)$$
$$\Leftrightarrow \Pr(Y = y) = \Pr(Y = y|X = x)$$
$$\Leftrightarrow \Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$$
$$\forall x \in dom(X), y \in dom(Y)$$

- Intuitively, learning the value of Y doesn't influence our beliefs about X and vice versa.

- Example: $\Pr(Sunny|ToothCavity) = \Pr(Sunny)$
  $\Pr(ToothCavity|Sunny) = \Pr(ToothCavity)$

Conditional Independence

- Two *variables* X and Y are conditionally independent given variable Z

$$\Pr(X = x|Z = z) = \Pr(X = x|Y = y, Z = z)$$
$$\Leftrightarrow \Pr(Y = y|Z = z) = \Pr(Y = y|X = x, Z = z)$$
$$\Leftrightarrow \Pr(X = x, Y = y|Z = z) = \Pr(X = x|Z = z)\Pr(Y = y|Z = z)$$
$$\forall x \in dom(X), y \in dom(Y), z \in dom(Z)$$

- If you know the value of Z (*whatever* it is), nothing you learn about Y will influence your beliefs about X

- Example: $\Pr(ToothAche|ToothCavity, ToothCatch) = \Pr(ToothAche|ToothCavity)$

$$\Pr(ToothCatch|ToothCavity, ToothAche) = \Pr(ToothCatch|ToothCavity)$$

UNIVERSITY OF WATERLOO

What good is independence?

- Suppose (say, Boolean) variables $X_1, X_2, \ldots, X_n$ are mutually independent
  - We can specify full joint distribution using only n parameters (linear) instead of $2^n - 1$ (exponential)
- How? Simply specify $\Pr(x_1), \ldots, \Pr(x_n)$
  - From this we can recover the probability of any world or any (conjunctive) query easily
    - Recall $\Pr(x_1, \ldots, x_n) = \Pr(x_1) \ldots \Pr(x_n)$

Example

- 4 independent Boolean random vars X_1, X_2, X_3, X_4

$$\Pr(x_1) = 0.4, \Pr(x_2) = 0.2, \Pr(x_3) = 0.5, \Pr(x_4) = 0.8$$

$$\begin{aligned}\Pr(x_1, \sim x_2, x_3, x_4) &= \Pr(x_1)\,(1 - \Pr(x_2))\,\Pr(x_3)\,\Pr(x_4) \\ &= (0.4)(0.8)(0.5)(0.8) \\ &= 0.128\end{aligned}$$

$$\begin{aligned}\Pr(x_1, x_2, x_3 | x_4) &= \Pr(x_1)\,\Pr(x_2)\,\Pr(x_3)\,\mathbf{1} \\ &= (0.4)(0.2)(0.5)(1) \\ &= 0.04\end{aligned}$$

The Value of Independence

- Complete independence reduces both *representation of joint distribution* and *inference* from $O(2^n)$ to $O(n)$!!
- Unfortunately, such complete mutual independence is very rare. Most realistic domains do not exhibit this property.
- Fortunately, most domains do exhibit a fair amount of conditional independence. We can exploit conditional independence for representation and inference as well.
- **Bayesian networks** do just this

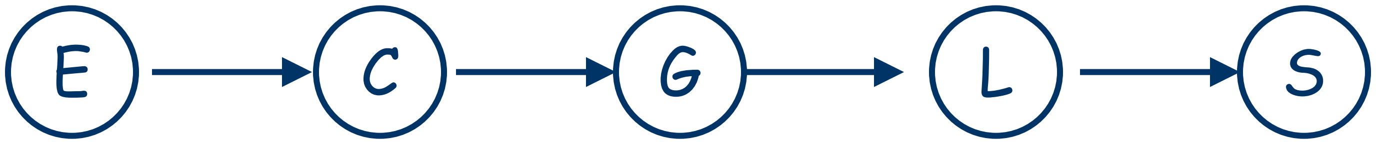
An Aside on Notation

- $\Pr(X)$ for variable X (or set of variables) refers to the *(marginal) distribution* over X. $\Pr(X|Y)$ refers to the family of conditional distributions over X, one for each $y \in Dom(Y)$.
- Distinguish between $\Pr(X)$ -- which is a distribution – and $\Pr(x)$ or $\Pr(\sim x)$ (or $\Pr(x_i)$ for non-Boolean vars) -- which are numbers. Think of $\Pr(X)$ as a function that accepts any $x_i \in Dom(X)$ as an argument and returns $\Pr(x_i)$.
- Think of $\Pr(X|Y)$ as a function that accepts any x_i and y_k and returns $\Pr(x_i|y_k)$. Note that $\Pr(X|Y)$ is not a single distribution; rather it denotes the family of distributions (over X) induced by the different $y_k \in Dom(Y)$

UNIVERSITY OF WATERLOO

Exploiting Conditional Independence

- Consider a story:
  - If Pascal woke up too early E, Pascal probably needs coffee C; if Pascal needs coffee, he's likely grumpy G. If he is grumpy then it's possible that the lecture won't go smoothly L. If the lecture does not go smoothly then the students will likely be sad S.



E – Pascal woke up too early G – Pascal is grumpy S – Students are sad
C – Pascal needs coffee L– The lecture did not go smoothly

UNIVERSITY OF WATERLOO

Conditional Independence

- If you learned any of E, C, G, or L, would your assessment of $\Pr(S)$ change?
  - If any of these are seen to be true, you would increase $\Pr(s)$ and decrease $\Pr(\sim s)$.
  - So S is *not independent* of E, or C, or G, or L.
- If you knew the value of L (true or false), would learning the value of E, C, or G influence $\Pr(S)$?
  - Influence that these factors have on S is mediated by their influence on L.
  - Students aren't sad because Pascal was grumpy, they are sad because of the lecture.
  - So S is *independent* of E, C, and G, *given* L

Conditional Independence

- So S is *independent* of E, and C, and G, *given* L
- Similarly:
  - S is *independent* of E, and C, *given* G
  - G is *independent* of E, *given* C
- This means that:

$$\Pr(S|L,\{G,C,E\}) = \Pr(S|L)$$

$$\Pr(L|G,\{C,E\}) = \Pr(L|G)$$

$$\Pr(G|C,\{E\}) = \Pr(G|C)$$

$\Pr(C|E)$ and $\Pr(E)$ don't "simplify"

Conditional Independence

- By the chain rule (for any instantiation of $S \ldots E$):

$$\Pr(S, L, G, C, E) = \Pr(S|L, G, C, E) \Pr(L|G, C, E) \Pr(G|C, E) \Pr(C|E) \Pr(E)$$

- By our independence assumptions:

$$\Pr(S, L, G, C, E) = \Pr(S|L) \Pr(L|G) \Pr(G|C) \Pr(C|E) \Pr(E)$$

- We can specify the full joint by specifying five *local conditional distributions*:

$\Pr(S|L)$; $\Pr(L|G)$; $\Pr(G|C)$; $\Pr(C|E)$; and $\Pr(E)$

Example Quantification

$\Pr(\boldsymbol{c}|\boldsymbol{e}) = \mathbf{0.9}$
$\Pr(\sim c|e) = 0.1$
$\Pr(\boldsymbol{c}|\sim\boldsymbol{e}) = \mathbf{0.5}$
$\Pr(\sim c|\sim e) = 0.5$

$\Pr(\boldsymbol{l}|\boldsymbol{g}) = \mathbf{0.2}$
$\Pr(\sim l|g) = 0.8$
$\Pr(\boldsymbol{l}|\sim\boldsymbol{g}) = \mathbf{0.1}$
$\Pr(\sim l|\sim g) = 0.9$

$\Pr(\boldsymbol{e}) = \mathbf{0.7}$
$\Pr(\sim e) = 0.3$

$\Pr(\boldsymbol{g}|\boldsymbol{c}) = \mathbf{0.3}$
$\Pr(\sim g|c) = 0.7$
$\Pr(\boldsymbol{g}|\sim\boldsymbol{c}) = \mathbf{1.0}$
$\Pr(\sim g|\sim c) = 0.0$

$\Pr(\boldsymbol{s}|\boldsymbol{l}) = \mathbf{0.9}$
$\Pr(\sim s|l) = 0.1$
$\Pr(\boldsymbol{s}|\sim\boldsymbol{l}) = \mathbf{0.1}$
$\Pr(\sim s|\sim l) = 0.9$

- Specifying the joint requires only 9 parameters (if we note that half of these are “1 minus” the others), instead of 31 for the explicit representation
  - linear in number of variables instead of exponential!
  - linear generally if dependence has a chain structure

Inference is Easy

- Want to know $\Pr(g)$? Use sum out rule:

$$P(g) = \sum_{c_i \in Dom(C)} \Pr(g \mid c_i)\Pr(c_i)$$

$$= \sum_{c_i \in Dom(C)} \Pr(g \mid c_i) \sum_{e_i \in Dom(E)} \Pr(c_i \mid e_i)\Pr(e_i)$$

These are all terms specified in our local distributions!

Inference is Easy

- Computing $\Pr(g)$ in more concrete terms:

$$\Pr(c) = \Pr(c|e)\Pr(e) + \Pr(c|{\sim}e)\Pr({\sim}e) = 0.8 * 0.7 + 0.5 * 0.3 = 0.78$$

$$\Pr({\sim}\mathrm{c}) = \Pr({\sim}c|e)\Pr(e) + \Pr({\sim}c|{\sim}e)\Pr({\sim}e) = 0.22$$

$\Pr({\sim}c) = 1 - \Pr(c)$, as well

$$\Pr(g) = \Pr(g|c)\Pr(c) + \Pr(g|{\sim}c)\Pr({\sim}c) = 0.3 * 0.78 + 1.0 * 0.22 = 0.454$$

$$\Pr({\sim}g) = 1 - \Pr(g) = 0.546$$