Artificial Intelligence

CS4365 --- Fall 2022

Probabilistic Reasoning and Methods

Instructor: Yunhui Guo

Uncertainty

- General situation:
  - Observed variables (evidence): Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
  - Unobserved variables: Agent needs to reason about other aspects (e.g. where an object is or what disease is present)
  - Model: Agent knows something about how the known variables relate to the unknown variables

Probability

- By far the most universally accepted and used formalism for uncertainty.
- Well-developed **semantics** and **theory** and **proofs** and **examples** and **software** and **education**.
- But how good is it for representing large amounts of general purpose knowledge about an uncertain world?
- What are the computational issues?
- We'll come back to this at the end. But for now let's steam ahead with the world of..... ....probability.

Sample Space

- Sample space:
  - The set of all possible outcomes Ω

- Examples:
  - Tossing a coin: {Head, Tail}
  - Tosing a dice: {1, 2, 3, 4, 5, 6}

Events

- Events:
  - A set of outcomes based on the sample space

- We say the event A occurs if the outcome of the experiment is in the set A

- Examples:
  - Tossing a coin:
    - Events: { {}, {Head}, {Tail}, {Head, Tail}}
  - Tosing a dice: {1, 2, 3, 4, 5, 6}
    - Events: { {}, {1}, {1,2}, {1,2,3}..., {1,2,3,4,5,6} }

The fundamental rules of probability

- Let A be an event, the occurrence of which we are uncertain about.

- The axioms of probability:

  - $0 <= P(A)$
  - $P(A) <= 1$ $P(\Omega) = 1$
  - $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \varnothing$

Another fact about probability theory

- $P(\sim A) = P(\Omega - A) = 1 - P(A)$

- This can be deduced from the axioms we just saw:
  - $0 <= P(A)$
  - $P(A) <= 1 \qquad P(\Omega) = 1$
  - $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \varnothing$

  - How?

Another fact

- $P(B) = P(B \cap A) + P(B \cap \sim A)$
- Why?

- $B = \{B \cap A\} \cup \{B \cap \sim A\}$
- $P(B) = P(\{B \cap A\} \cup \{B \cap \sim A\}) = P(B \cap A) + P(B \cap \sim A)$
- Since $B \cap A$ and $B \cap \sim A$ are disjoint.

Another fact

- $P(B) = P(B \cap A) + P(B \cap \sim A)$

- More generally (prove by induction):

  If $P(A_1 \cup A_2 \cup \ldots A_n) = 1$, and for all i, j unequal, $P(A_i \cap A_j) = 0$) THEN we know:

  $P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n)$

Another fact

- $P(A \cup B) = P(A) + P(B) - P(B \cap A)$ (inclusion–exclusion principle)

- $A \cup B = A \cup (B \cap \sim A)$

- $P(A \cup B) = P(A) + P(B \cap \sim A) = P(A) + P(B) - P(B \cap A)$

Conditional probability

- P(A | B) denotes the probability of A given that B's occurrence is known.

- P(Cavity) = 0.04

  "There's a 4% chance at any time that you have a cavity"

- P(Cavity | Toothache) = 0.8

  "If you have a toothache, there's an 80% chance you have a cavity"

P(B|A) = 1 is equivalent to A => B

P(B|A) = 0.95 is a bit like a "soft fuzzy" version of A=>B.

Conditional Probability

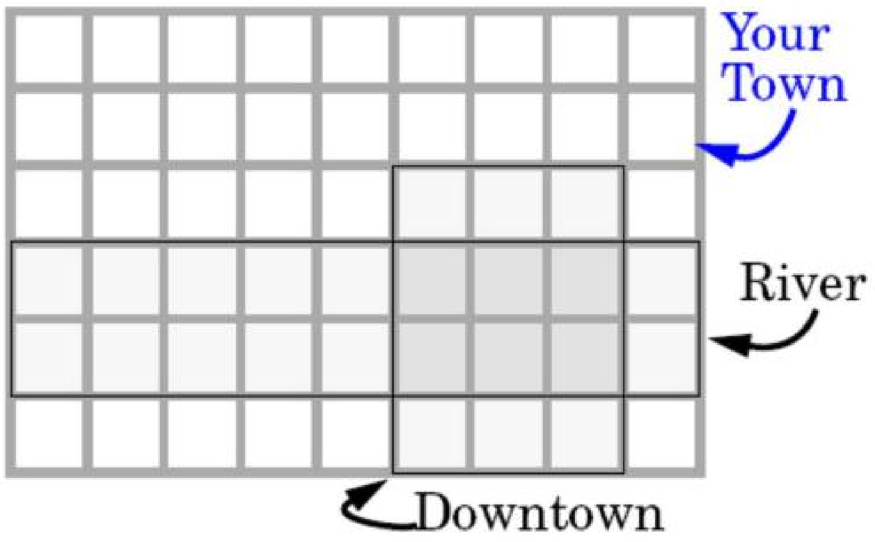
- Prior probablity: the degrees of belief in propositions in the absence of any other information
  - P(cavity)

- Evidence: has already been revealed P(toothache)

- Posterior probability: P(cavity |toothache=True)

Lunar Lander Example

A lunar lander crashes somewhere in your town (one of the cells at random on the grid). The crash point is **uniformly random**. R is the event that it crashes in the river. D is the event that it crashes downtown



- What are P(R), P(D), P(D ∩ R) ?
- What is P(D | R) ?
- What is P(R | D) ?
- What is P(R ∩ D) / P(D) ?

Conditional Probability

- Useful to remember "the chain rule":

  $P(A \cap B) = P(A \mid B)\, P(B)$ (product rule).

Exercise: Prove that $P(A \cap B) <= P(A)$ for any events A and B.

$P(W) = 0.001$

$P(S \mid W) = 0.5$

$P(L \mid S \cap W) = 0.1$

What is $P(W \cap S \cap L)$?

Combining what we know

If $P(A_1 \cup A_2 \cup \ldots A_n) = 1$, and for all i, j unequal, $P(A_i \cap A_j) = 0$
THEN we know:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \ldots + P(B \cap A_n)$$

$$P(B \cap A) = P(B|A)\ P(A)$$

- So ...

- $P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_n)\ P(A_n)$

Example

My mood can take one of two values: Happy, Sad.
The weather can take one of three values: Rainy, Sunny, Cloudy.

My knowledge base says.

- P(Mood=Happy and Weather=Rainy) = 0.2
- P(Mood=Happy and Weather=Sunny) = 0.1
- P(Mood=Happy and Weather=Cloudy) = 0.4

- Can I compute P(Mood=Happy)?
- Can I compute P(Mood=Sad)?
- Can I compute P(Weather=Rainy) ?

Random Variable

- A random variable is some aspect of the world about which we (may) have uncertainty
  - R = Is it raining?
  - T = Is it hot or cold?
  - D = How long will it take to drive to work?
- We denote random variables with capital letters X
- Like variables in a CSP, random variables have domains
  - R in {true, false}
  - T in {hot, cold}
  - D in $[0, \infty)$

Random Variable

- A random variable X is a function from the sample space Ω into the real numbers

- **Probability function:** $P_X(X = x_i) = P(\{s_j \in \Omega : X(s_j) = x_i\})$
- Example:
  - Toss two coins:
    - Define a random variable X to be the number of heads obtained

    - [H, H], [H,T], [T, T,] [T, H]

    - $P(X = 1) = P(\{[H, T], [T, H]\}) = P([H, T]) + P([H, T]) = 1/2$

Probability Distributions

- Three coin tosses
- Define random variable X as the number of heads

s:	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
X:	3	2	2	2	1	1	1	0

x	0	1	2	3
P(X=x)	1/8	3/8	3/8	1/8

Probability Distributions

- Associate a probability with each value

- Temperature:

$$P(T)$$

T	P
hot	0.5
cold	0.5

Weather:

$$P(W)$$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Probability Mass Function

- The **probability mass function** of a discrete random variable is defined as,

$$f_X(x) = P(X=x) \text{ for all } x$$

Bernoulli distribution:

Takes the value 1 with probability p and the value 0 with probability q=1-p

$$f_X(x;p) = p \text{ if } x = 1 \text{ or } q \text{ if } x = 0$$

Bayes Rule

- Named after **Thomas Bayes**, describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

Prior + Likelihood -> Posterior

Bayes Rule

- Simplest form:

Prior distribution

Likelihood

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

Posterior distribution

Marginalization

- General form:

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

Generalizing Bayes Rule

- If we know that exactly one of A_1, A2, ..., A_n are true, then:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + ... + P(B|A_n)\,P(A_n)$$

and in general

$$P(B|X) = P(B|A_1,X)P(A_1|X) + ... + P(B|A_n,X)\,P(A_n|X)$$

- so

$$\mathrm{P}(A_k|B,X) = \frac{\mathrm{P}(A_k|X)\mathrm{P}(B|A_k,X)}{\sum_i \mathrm{P}(A_i|X)P(B|A_i,X)}$$

Medical Diagnosis

A doctor knows that meningitis causes a stiff neck 50% of the time.

The doctor knows that if a person is randomly selected from the US population, there's a 1/50,000 chance the person will have meningitis.

The doctor knows that if a person is randomly selected from the US population, there's a 5% chance the person will have a stiff neck.

You walk into the doctor complaining of the **symptom** of a stiff neck. What's the probability that the **underlying cause** is meningitis?

Example

- I have three identical boxes labeled H1, H2 and H3.
    - **Into H1 I place 1 black bead, 3 white beads.**
    - **Into H2 I place 2 black beads, 2 white beads.**
    - **Into H3 I place 4 black beads, no white beads.**

I draw a box at random. I remove a bead at random from that box.

What can I deduce from the color of the bead as to which box I drew?

If I replace the bead, then redraw another bead at random from the same box, how well can I predict its color before drawing it?

- These two questions are the foundations of reasoning with **uncertainty** and **machine learning**.

Bayesian Rule

A nice way to look at this

H1 , H2 and H3 were my prior models of the world. The fact that P(H1) = 1/3, P(H2) = 1/3, P(H3) = 1/3 was my **prior distribution**.

The color of the bead was a piece of **evidence** about the true model of the world.

The use of bayes' rule was a piece of probabilistic inference, giving me a **posterior distribution** on possible worlds.

Learning is prior + evidence ---> posterior

- A piece of evidence decreases my ignorance about the world.
- Distributions are good ways of describing your **state of knowledge**. Knowledge that includes uncertainty measure can mean much better decision-making.

Joint Probability Distribution

- Given **two random variables**, the joint probability distribution is the corresponding probability distribution on all possible pairs of output

- E.g., tossing two coins

  $P(A) = 1/2 \quad$ for $A \in \{0, 1\}$

  $P(B) = 1/2 \quad$ for $B \in \{0, 1\}$

  $P(A=1, B=1) = 1/4$

Another example

- Suppose we will wish to reason about flying ability, birdhood and youth in animals
- We can set up or knowledge base as a probability distribution before we receive any information about an animal

Bird	Flier	Young	Prob
T	T	T	0
T	T	F	0.2
T	F	T	0.04
T	F	F	0.01
F	T	T	0.01
F	T	F	0.01
F	F	T	0.23
F	F	F	0.5

- Each row is a prior hypothesis about the state of the animal.

Marginalizing

- Joint Probability Distribution:
  - Its probabilities must add up to 1. It defines all basic conjunctive probabilities, e.g: P(Bird = T, Flier=F, Young=T) = 0.04

- We can compute marginal probabilities (probabilities for subsets of variables taking specified values) easily... e.g. P(Bird = T, Young=F)

- Marginalization,

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

Marginalizing

- **Handy tip**: For P(expression), just find the rows that match the expression, and add up the associated probabilities.
- We can compute conditional probabilities with ease. e,g,.: P(Y |B and ~F) =
- **Handy tip**: For P(this|that), just do two marginals: P(this and that) and P(that). Then compute their ratio!

Conditional probabilities from the joint distribution

Bird	Flier	Young	Prob
T	T	T	0
T	T	F	0.2
T	F	T	0.04
T	F	F	0.01
F	T	T	0.01
F	T	F	0.01
F	F	T	0.23
F	F	F	0.5

- Let x_1, x_2 ... be values True or False

- $P(X_1=x_1,X_2=x_2,...,X_n=x_n, |$

  $X_{n+1}=x_{n+1},...,X_{n+k}=x_{n+k})=$

  sum of all entries for which $X_1=x_1,X_2=x_2,...,X_{n+k}=x_{n+k}$ /

entries for which $X_{n+1}=x_{n+1},...,X_{n+k}=x_{n+k}$

- P(B=False | F=True,Y=False) =

  P(B=False, F=True, Y=False) /

P(B=True,F=True,Y=False)+P(B=False,F=True,Y=False) = 0.1/0.21 = 1/21

Another Example

- Three Boolean variables
  - Toothache: the patient has a toothache
  - Cavity: the patient has a cavity
  - Catch: the dentist catches in the patients tooth with his nasty steel probe

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

Example

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

- P(cavity) = $\Sigma_{z \in \{Catch, Toothache\}}$ P(cavity, z)
- P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2 (marginalization)
- P(cavity $\vee$ toochaches) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28

Example

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

- P(cavity | toochache) = P(cavity ∧ toochache) / P(toothache)

  = (0.108 + 0.012) / (0.108 + 0.012 + 0.016 + 0.064) = 0.6

- P(¬cavity | toochache) = P(¬cavity ∧ toochache) / P(toothache)

  = (0.016 + 0.064) / (0.108 + 0.012 + 0.016 + 0.064) = 0.4

Example

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

- Normalization constant: 1/P(toothache) (denoted as α)
- 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>.

A general inference precedure

- E: the list of evidence variables
- e: the list of observed values for them,
- Y be the remaining unobserved variables

- The query can be computed as:

$$P(X|e) = \alpha\, P(X, e) = \alpha\, \Sigma_y P(X, e, y)$$

The joint probability distribution

- Contains all the domain knowledge you'll need for any conditional probability.
- What's the probability that it's young given that it either flies and is a bird?
- I took an animal. I don't know what they were, but it is definitely young. What's the chance it is a bird?"
- Joint probability distributions are a great way of describing knowledge.
- Question: What's the big problem with this?

Using fewer numbers

Suppose there are two events:

M : Mr. M teaches algebra

S : It is sunny

The joint p.d.f for these events contain four entries.

If we want to build the joint p.d.f we'll have to invent those four numbers.
We don't have to specify with bottom level conjunctive events such as P(not M and S) IF instead it may sometimes be more convenient for us to specify things like: P(M) , P(S).
But just P(M) and P(S) don't derive the joint distribution. So you can't answer all questions.

What **extra assumption** can you make?

Independence

- "The sunshine levels do not depend on and do not influence who is teaching."

This can be specified very simply: $P(S \mid M) = P(S)$

This is a powerful statement! It required extra domain knowledge. A different kind of knowledge than p.d.f.s.

From $P(S \mid M) = P(S)$, the rules of probability imply:

- $P(\sim S \mid M) = P(\sim S)$
- $P(M \mid S) = P(M)$
- $P(M \text{ and } S) = P(M)\, P(S)$
- $P(\sim M \text{ and } S) = P(\sim M)P(S)$, $P(M \text{ and } \sim S) = P(M)P(\sim S)$
- $P(\sim M \text{ and } \sim S) = P(\sim M)P(\sim S)$

Independence

- We've stated

  P(M) = 0.6, P(S) = 0.3, P(S | M) = P(S)

  From these two numbers, and an independence assumption, we can derive the full joint pdf

M	S	Prob
T	T	
T	F	
F	T	
F	F	

And since we now have the joint pdf, we can make any queries we like.

A more interesting case

Suppose there are three events:
M: Mr. M teaches algebra (otherwise it's Mr. B)
S : It is sunny
L : The lecturer arrives slightly late

Assume both lecturers are sometimes delayed by bad weather. And Mr. B is more likely to arrive late than Mr. M.
Let's begin with writing down knowledge we're happy about:

$$P(S \mid M) = P(S),\ P(S) = 0.3,\ P(M) = 0.6$$

Now, lateness is not independent of the weather and is not independent of the lecturer. We must choose what else we need to write down.

We know the joint pdf of S and M, so let's just write down $P(L \mid S=x, M=y)$ in the 4 cases.

A more interesting case

M: Mr. M teaches algebra (otherwise it's Mr. B)
S : It is sunny
L : The lecturer arrives slightly late

Assume both lecturers are sometimes delayed by bad weather. And Mr. B is more likely to arrive late than Mr. M.

$P(S \mid M) = P(S) \quad P(S) = 0.3 \quad P(M) = 0.6$

$P(L \mid M \wedge S) = 0.05 \quad P(L \mid M \wedge \sim S) = 0.1$

$P(L \mid \sim M \wedge S) = 0.1 \quad P(L \mid \sim M \wedge \sim S) = 0.2$

Now we can derive a full joint p.d.f with six numbers instead of eight. (Savings are larger for larger numbers of variables).

Question: Express $P(L=x \wedge M=y \wedge S=z)$ in terms that only need the above expressions, where x, y and z may each be True or False