Probability

Sanja Lazarova-Molnar

Uncertainty

- Let action A_t = leave for airport t minutes before flight
  - Will A_t get me there on time?
- Problems:
  - Partial observability (road state, other drivers' plans, etc.)
  - Uncertainty in action outcomes (flat tire, etc.)
  - Complexity of modeling and predicting traffic
- Hence a purely logical approach either
  - Risks falsehood: "A_{25} will get me there on time," or
  - Leads to conclusions that are too weak for decision making:
    - 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., etc.
    - A_{1440} might reasonably be said to get me there on time but I'd have to stay overnight in the airport

Probability

Probabilistic assertions summarize effects of

- Laziness: failure to enumerate exceptions, qualifications, etc.
- Ignorance: lack of explicit theories, relevant facts, initial conditions, etc.
- Intrinsically random behavior

Making decisions under uncertainty

- Suppose the agent believes the following:

  $P(A_{25}$ gets me there on time$) = 0.04$

  $P(A_{90}$ gets me there on time$) = 0.70$

  $P(A_{120}$ gets me there on time$) = 0.95$

  $P(A_{1440}$ gets me there on time$) = 0.9999$

- Which action should the agent choose?

Making decisions under uncertainty

- Suppose the agent believes the following:

  $P(A_{25}$ gets me there on time$) = 0.04$

  $P(A_{90}$ gets me there on time$) = 0.70$

  $P(A_{120}$ gets me there on time$) = 0.95$

  $P(A_{1440}$ gets me there on time$) = 0.9999$

- Which action should the agent choose?
  - Depends on preferences for missing flight vs. time spent waiting
  - Encapsulated by a *utility function*
- The agent should choose the action that maximizes the *expected utility*:

  $P(A_t$ succeeds$) * U(A_t$ succeeds$) + P(A_t$ fails$) * U(A_t$ fails$)$

- Utility theory is used to represent and infer preferences
- Decision theory = probability theory + utility theory

Monty Hall problem

- You're a contestant on a game show. You see three closed doors, and behind one of them is a prize. You choose one door, and the host opens one of the other doors and reveals that there is no prize behind it. Then he offers you a chance to switch to the remaining door. Should you take it?

Monty Hall problem

- With probability 1/3, you picked the correct door, and with probability 2/3, picked the wrong door. If you picked the correct door and then you switch, you lose. If you picked the wrong door and then you switch, you win the prize.

- Expected payoff of switching:

(1/3) * 0 + (2/3) * Prize

- Expected payoff of not switching:

(1/3) * Prize + (2/3) * 0

http://www.shodor.org/interactivate/activities/SimpleMontyHall/

Where do probabilities come from?

- **Frequentism**
  - Probabilities are relative frequencies
  - For example, if we toss a coin many times, P(heads) is the proportion of the time the coin will come up heads
- **Subjectivism**
  - Probabilities are degrees of belief
  - But then, how do we assign belief values to statements?

Random variables

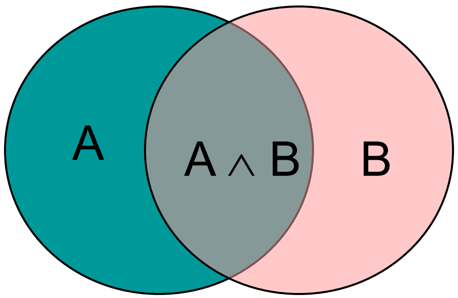
- We describe the (uncertain) state of the world using *random variables*
  - Denoted by capital letters
  - **R**: *It will rain tomorrow*
  - **W**: *Weather condition*
  - **D**: *Outcome of rolling two dice*
  - **S**: *Speed of my car (in KPH)*
- Just like variables in CSP’s, random variables take on values in a *domain*
  - Domain values must be mutually exclusive and exhaustive
  - **R** in {True, False}
  - **W** in {Sunny, Cloudy, Rainy, Snow}
  - **D** in {(1,1), (1,2), … (6,6)}
  - **S** in [0, 260]

Events

- Probabilistic statements are defined over *events*, or sets of world states
  - *"It will rain tomorrow"*
  - *"The weather is either cloudy or snowy"*
  - *"The sum of the two dice rolls is 11"*
  - *"My car is going between 50 and 90 kilometers per hour"*
- Events are described using propositions:
  - R = True
  - W = "Cloudy" $\vee$ W = "Snowy"
  - $D \in \{(5,6), (6,5)\}$
  - $50 \leq S \leq 90$
- Notation: P(A) is the probability of the set of world states in which proposition A holds
  - P(X = x), or P(x) for short, is the probability that random variable X has taken on the value x

Kolmogorov's axioms of probability

- For any propositions (events) A, B
  - $0 \le P(A) \le 1$
  - $P(True) = 1$ and $P(False) = 0$
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
    - Subtraction accounts for double-counting



- Based on these axioms, what is $P(\neg A)$?

- These axioms are sufficient to completely specify probability theory for *discrete* random variables
  - For continuous variables, need *density functions*

Atomic events

- Atomic event: a complete specification of the state of the world, or a complete assignment of domain values to all random variables
  - Atomic events are mutually exclusive and exhaustive

- E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

$$Cavity = false \wedge Toothache = false$$
$$Cavity = false \wedge Toothache = true$$
$$Cavity = true \wedge Toothache = false$$
$$Cavity = true \wedge Toothache = true$$

Joint probability distributions

- A **joint distribution** is an assignment of probabilities to every possible atomic event

Atomic event	**P**
Cavity = false ∧ *Toothache = false*	0.8
Cavity = false ∧ *Toothache = true*	0.1
Cavity = true ∧ *Toothache = false*	0.05
Cavity = true ∧ *Toothache = true*	0.05

  – From the axioms of probability it follows that the probabilities of all possible atomic events must sum to 1.

Joint probability distributions

- Suppose we have a joint *distribution* $P(X_1, X_2, \ldots, X_n)$ of n random variables with domain sizes d
  - What is the size of the probability table?
  - Impossible to write out completely for all but the smallest distributions

- Notation:
  - $P(X = x)$ is the probability that random variable X takes on value x
  - $P(X)$ is the *distribution* of probabilities for all possible values of X

Marginal probability distributions

- Suppose we have the joint distribution P(X,Y) and we want to find the *marginal distribution* P(Y)

P(Cavity, Toothache)	
Cavity = false $\wedge$ *Toothache = false*	0.8
Cavity = false $\wedge$ *Toothache = true*	0.1
Cavity = true $\wedge$ *Toothache = false*	0.05
Cavity = true $\wedge$ *Toothache = true*	0.05

P(Cavity)	
Cavity = false	?
Cavity = true	?

P(Toothache)	
Toothache = false	?
Toothache = true	?

Marginal probability distributions

- Suppose we have the joint distribution P(X,Y) and we want to find the *marginal distribution* P(X)

$$P(X = x) = P\big((X = x \wedge Y = y_1) \vee \ldots \vee (X = x \wedge Y = y_n)\big)$$

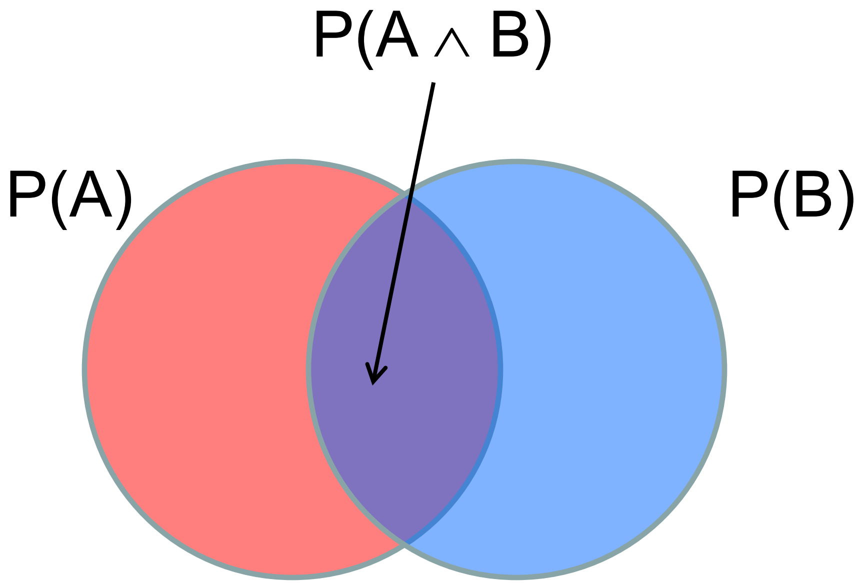
$$= P\big((x, y_1) \vee \ldots \vee (x, y_n)\big) = \sum_{i=1}^{n} P(x, y_i)$$

- General rule: to find P(X = x), sum the probabilities of all atomic events where X = x.

Conditional probability

- Probability of cavity given toothache:
  P(*Cavity* = *true* | *Toothache* = *true*)

- For any two events A and B, $P(A \mid B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(A, B)}{P(B)}$



Conditional probability

P(Cavity, Toothache)	
Cavity = false $\wedge$ *Toothache = false*	0.8
Cavity = false $\wedge$ *Toothache = true*	0.1
Cavity = true $\wedge$ *Toothache = false*	0.05
Cavity = true $\wedge$ *Toothache = true*	0.05

P(Cavity)	
Cavity = false	0.9
Cavity = true	0.1

P(Toothache)	
Toothache = false	0.85
Toothache = true	0.15

- What is P(*Cavity = true* | *Toothache = false*)?

  0.05 / 0.85 = 0.059

- What is P(*Cavity = false* | *Toothache = true*)?

  0.1 / 0.15 = 0.667

Conditional distributions

- A conditional distribution is a distribution over the values of one variable given fixed values of other variables

P(Cavity, Toothache)	
Cavity = false $\wedge$ *Toothache = false*	0.8
Cavity = false $\wedge$ *Toothache = true*	0.1
Cavity = true $\wedge$ *Toothache = false*	0.05
Cavity = true $\wedge$ *Toothache = true*	0.05

P(Cavity | Toothache = true)	
Cavity = false	0.667
Cavity = true	0.333

P(Cavity|Toothache = false)	
Cavity = false	0.941
Cavity = true	0.059

P(Toothache | Cavity = true)	
Toothache= false	0.5
Toothache = true	0.5

P(Toothache | Cavity = false)	
Toothache= false	0.889
Toothache = true	0.111

Normalization trick

- To get the whole conditional distribution P(X | y) at once, select all entries in the joint distribution matching Y = y and renormalize them to sum to one

P(Cavity, Toothache)	
Cavity = false $\wedge$ *Toothache = false*	0.8
Cavity = false $\wedge$ *Toothache = true*	0.1
Cavity = true $\wedge$ *Toothache = false*	0.05
Cavity = true $\wedge$ *Toothache = true*	0.05

Select

Toothache, Cavity = false	
Toothache= false	0.8
Toothache = true	0.1

Renormalize

P(Toothache | Cavity = false)	
Toothache= false	0.889
Toothache = true	0.111

Product rule

- Definition of conditional probability: $P(A \mid B) = \dfrac{P(A,B)}{P(B)}$

- Sometimes we have the conditional probability and want to obtain the joint:

$$P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

Product rule

- Definition of conditional probability: $P(A \mid B) = \dfrac{P(A,B)}{P(B)}$

- Sometimes we have the conditional probability and want to obtain the joint:

$$P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- The chain rule:

$$P(A_1,\ldots,A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1,A_2)\ldots P(A_n \mid A_1,\ldots,A_{n-1})$$

$$= \prod_{i=1}^{n} P(A_i \mid A_1,\ldots,A_{i-1})$$

Bayes Rule

Rev. Thomas Bayes
(1702-1761)

- The product rule gives us two ways to factor a joint distribution:

$$P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- Therefore, $P(A \mid B) = \dfrac{P(B \mid A)P(A)}{P(B)}$

- Why is this useful?
  - Can get *diagnostic probability* P(cavity | toothache) from *causal probability* P(toothache | cavity)
  - Can update our beliefs based on evidence
  - Important tool for probabilistic inference

Bayes Rule example

- Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year (5/365 = 0.014). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on Marie's wedding?

Bayes Rule example

- Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year (5/365 = 0.014). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on Marie's wedding?

$$P(\text{Rain} \mid \text{Predict}) = \frac{P(\text{Predict} \mid \text{Rain})P(\text{Rain})}{P(\text{Predict})}$$

$$= \frac{P(\text{Predict} \mid \text{Rain})P(\text{Rain})}{P(\text{Predict} \mid \text{Rain})P(\text{Rain}) + P(\text{Predict} \mid \neg\text{Rain})P(\neg\text{Rain})}$$

$$= \frac{0.9 * 0.014}{0.9 * 0.014 + 0.1 * 0.986} = 0.111$$

Bayes rule: Another example

- 1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammographies. 9.6% of women without breast cancer will also get positive mammographies. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

Bayes rule: Another example

- 1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammographies. 9.6% of women without breast cancer will also get positive mammographies. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

$$P(\text{Cancer} \mid \text{Positive}) = \frac{P(\text{Positive} \mid \text{Cancer}) P(\text{Cancer})}{P(\text{Positive})}$$

$$= \frac{P(\text{Positive} \mid \text{Cancer}) P(\text{Cancer})}{P(\text{Positive} \mid \text{Cancer}) P(\text{Cancer}) + P(\text{Positive} \mid \neg\text{Cancer}) P(\neg\text{Cancer})}$$

$$= \frac{0.8 * 0.01}{0.8 * 0.01 + 0.096 * 0.99} = 0.0776$$

Independence

- Two events A and B are independent if and only if $P(A \wedge B) = P(A)\,P(B)$
  - In other words, $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$
  - This is an important simplifying assumption for modeling, e.g., *Toothache* and *Weather* can be assumed to be independent

- Are two *mutually exclusive* events independent?
  - No, but for mutually exclusive events we have $P(A \vee B) = P(A) + P(B)$

- **Conditional independence**: A and B are *conditionally independent* given C iff $P(A \wedge B \mid C) = P(A \mid C)\,P(B \mid C)$

Conditional independence: Example

- *Toothache*: boolean variable indicating whether the patient has a toothache
- *Cavity*: boolean variable indicating whether the patient has a cavity
- *Catch*: whether the dentist's probe catches in the cavity

- If the patient has a cavity, the probability that the probe catches in it doesn't depend on whether he/she has a toothache

  $P(Catch \mid Toothache, Cavity) = P(Catch \mid Cavity)$
- Therefore, *Catch* is conditionally independent of *Toothache* given *Cavity*
- Likewise, *Toothache* is conditionally independent of *Catch* given *Cavity*

  $P(Toothache \mid Catch, Cavity) = P(Toothache \mid Cavity)$
- Equivalent statement:

  $P(Toothache, Catch \mid Cavity) = P(Toothache \mid Cavity)\, P(Catch \mid Cavity)$

Conditional independence: Example

- How many numbers do we need to represent the joint probability table P(*Toothache, Cavity, Catch*)?

  $2^3 - 1 = 7$ independent entries, or 8 values in table

- Write out the joint distribution using chain rule:

  P(*Toothache, Catch, Cavity*)

  = P(*Cavity*) P(*Catch* | *Cavity*) P(*Toothache* | *Catch, Cavity*)

  = P(*Cavity*) P(*Catch* | *Cavity*) P(*Toothache* | *Cavity*)

- How many numbers do we need to represent these distributions?

  1 + 2 + 2 = 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*

Probabilistic inference

- In general, the agent observes the values of some random variables $X_1, X_2, \ldots, X_n$ and needs to reason about the values of some other *unobserved* random variables $Y_1, Y_2, \ldots, Y_m$
  - Figuring out a diagnosis based on symptoms and test results
  - Classifying the content type of an image or a document based on some features
- This will be the subject of the next classes