Probability Overview (very brief)

CSCI 2951-F
Brown University
Ronald Parr

Goals Of These Slides

- Revisit a topic most of you have seen already to
  - Refresh your memories
  - Synchronize notation
- Provide context – for AI, RL, etc.

Why does AI need uncertainty?

- Reason: Sh*t happens
- Actions don't have deterministic outcomes

- Can logic be the "language" of AI???
- Problem: General logical statements are almost always false

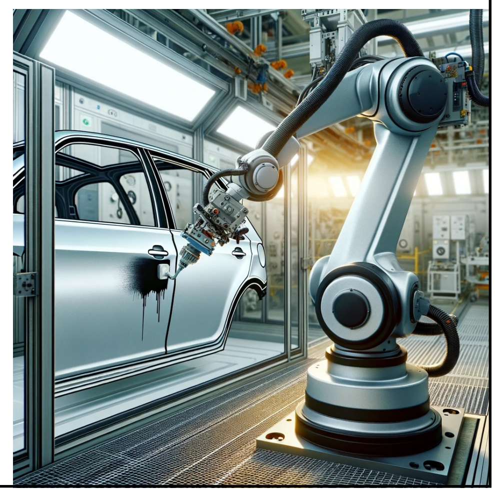
- Truthful and accurate statements about the world would seem to require an endless list of *qualifications*
- How do you start a car?
- Call this "The Qualification Problem"

The Qualification Problem

- Is this a real concern?
- YES!
- Systems that try to avoid dealing with uncertainty tend to be brittle.
- Plans fail
- Finding shortest path to goal isn't that great if the path doesn't really get you to the goal

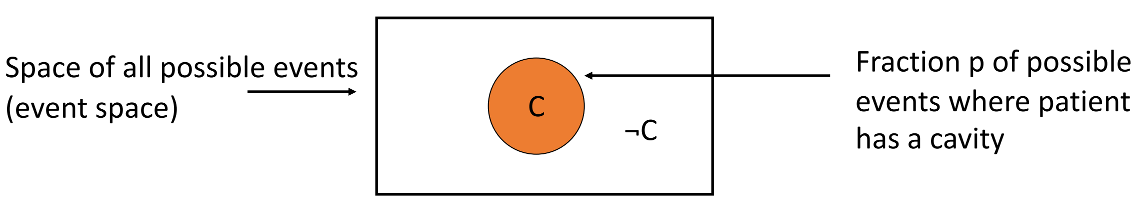
When can we (mostly) ignore qualifications?

- When environment is highly engineered/controlled
- Objects moving in free space
- Carefully controlled factories
- When replanning is relatively cheap



Relative Frequencies (simplest view of probs)

- Consider a world where a dentist agent D meets a new patient
- D is interested in **only one thing**: whether patient has a cavity (C)
- Before making any observation, D's belief state is:



- This means that D believes that a fraction p of patients have cavities

Notation

- P(XY) = P(X,Y) = a joint probability distribution over all settings of X and Y (potentially a **table** with a large number of entries)
- P(xy) = P(x,y) = P(x AND y) = P(x ^ y) = P(X=x,Y=y),P(X=x AND Y=y)=P(X=x ^ Y=x) = **a single number** corresponding the probability that both X=x and Y=y
- P(Xy) = a table with one entry for each value of X when y is true – not a distribution
- P(X=false)=P($\overline{x}$)=P(¬x) = P(~x) = a single number for case where X is a binary variable takes value false (or zero)

Why Probabilities Are Messy

- Probabilities are not truth-functional
- Computing P(a and b) requires the joint distribution
  - It is not, in general, a function of P(a) and P(b)
  - It is not, in general, a function of P(a) and P(b)
  - It is not, in general, a function of P(a) and P(b)
- This fact led to many approximations methods such as certainty factors and fuzzy logic (Why?)

Working With Joint Distributions

- A joint distribution is an assignment of probabilities to every possible atomic event
- We can define all other probabilities in terms of the joint probabilities by *marginalization*:

$$P(a) = P(a \wedge b) + P(a \wedge \neg b)$$

$$P(a) = \sum_{e_i \in e(a)} P(e_i)$$

Example

- P(cold ⋀ headache) = 0.4
- P(¬cold ⋀ headache) = 0.2
- P(cold ⋀ ¬ headache) = 0.3
- P(¬ cold ⋀ ¬ headache) = 0.1

- What are P(cold) and P(headache)?

Independence

- If A and B are independent: $P(A \wedge B) = P(A)P(B)$

- $P(\text{cold} \wedge \text{headache}) = 0.4$
- $P(\neg\text{cold} \wedge \text{headache}) = 0.2$
- $P(\text{cold} \wedge \neg\,\text{headache}) = 0.3$
- $P(\neg\,\text{cold} \wedge \neg\,\text{headache}) = 0.1$

- Are cold and headache independent?

Independence and Mutual Exclusivity

- Examples of independent events:
  - KC winning Superbowl, Biden winning reelection
  - Two successive, fair coin flips
  - My car starting and my iPhone working
  - etc.

- If A and B are mutually exclusive:
  $P(A \vee B) = P(A)+P(B)$

Expectation

- Most of us use expectation in some form when we compute averages
- What is the average value of a fair die roll?

- (1+2+3+4+5+6)/6 = 3.5
  (we divide by 6 because all outcomes are equally likely)

- Is it possible for all children to be above average?

Expectation in General

- Suppose we have some RV X
- Suppose we have some function f(X)
- What is the expected value of f(X)?

$$\underset{x}{E} f(x) = \sum_{x} P(X) f(X)$$

Linearity of Expectation

- Suppose we have f(X) and g(Y).
- What is the expected value of f(X)+g(Y)?

$$
\begin{aligned}
\underset{XY}{E}\, f(X)+g(Y) &= \sum_{XY} P(X \wedge Y)(f(X)+g(Y)) \\
&= \sum_{XY} P(X \wedge Y) f(X) + \sum_{XY} P(X \wedge Y) g(Y) \\
&= \sum_{X}\sum_{Y} P(X \wedge Y) f(X) + \sum_{Y}\sum_{X} P(X \wedge Y) g(Y) \\
&= \sum_{X} f(x) \sum_{Y} P(X \wedge Y) + \sum_{Y} g(Y) \sum_{X} P(X \wedge Y) \\
&= \sum_{X} f(x) P(X) + \sum_{Y} g(Y) \sum_{X} P(X \wedge Y) \\
&= \underset{X}{E}\, f(X) + \underset{Y}{E}\, g(Y)
\end{aligned}
$$

AI avoided probabilities for decades

- Reasoning about probabilities correctly requires the joint distribution
  - Exponentially large! (all truth values of all variables)
  - Very inconvenient!
- But…assuming independence (mutual exclusivity) when there is not independence (mutual exclusivity) leads to incorrect answers
- Examples:
  - ANDing symptoms by multiplying (independence)
  - ORing symptoms by adding (mutual exclusivity)
- “Dutch Book” argument shows that any system of beliefs not consistnet with probability theory can lead to bad bets

Conditional Probabilities

- Ordinary probabilities for random variables:
  unconditional or *prior* probabilities
- P(a|b) = P(a AND b)/P(b)
- This tells us the probability of a **given that we know *only* b**
- If we know c and d, we can't use P(a|b) directly
  (without additional assumptions)
- Annoying, but solves the qualification problem…

Probability **Solves** the Qualification Problem

- P(disease|symptom1)
- Probability of a disease given that we have observed ***only*** symptom1
- The conditioning bar indicates that the probability is defined with respect to a particular state of knowledge, *not as an absolute thing*

Condition with Bayes's Rule

$$P(A \wedge B) = P(B \wedge A)$$
$$P(A \mid B)P(B) = P(B \mid A)P(A)$$
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Note that we will usually call Bayes's rules "Bayes Rule"

Let's Play Doctor

- P(cold) = 0.7, P(headache) = 0.6
- P(headache|cold) = 0.57
- What is P(cold|headache) using Bayes Rule?

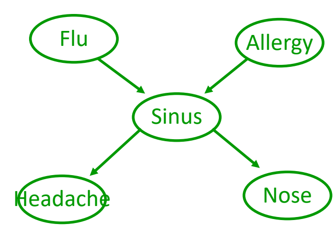
$$P(c \mid h) = \frac{P(h \mid c)P(c)}{P(h)}$$
$$= \frac{0.57 * 0.7}{0.6} = 0.66$$

- IMPORTANT: Not always symmetric

Conditional Independence

- We say that two variables, A and B, are conditionally independent given C if:
  - P(A|BC) = P(A|C)
  - P(AB|C) = P(A|C)P(B|C)
- How does this help?
- We store only a conditional probability table (CPT) of each variable given its parents
- Naïve Bayes (e.g. Spam Assassin) is a special case of this! (Words are conditionally independent given spam/ham)

What is a Bayes Net?

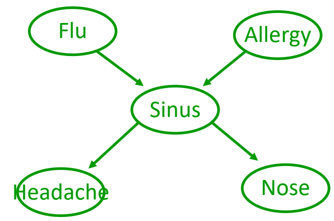


- A directed acyclic graph (DAG)
- ***Given parents,*** each variable is *conditionally independent of non-descendants,*
- Joint probability decomposes:

$$P(x_1..x_n) = \prod_i P(x_i \mid \text{parents}(x_i))$$

- For each node X_i, store $P(X_i|\text{parents}(X_i))$
- Call this a Conditional Probability Table (CPT)
- CPT size is exponential in number of parents

Space Efficiency



- Entire joint distribution as 32 (31) entries
  - P(H|S),P(N|S) have 4 (2)
  - P(S|AF) has 8 (4)
  - P(A), P(F) have 2 (1)
  - Total is 20 (10)
- This can require exponentially less space
- Space problem is solved for "most" problems

(Non)Uniqueness of Bayes Nets I

- Suppose you have two variables that are **NOT** independent
- Two possible networks:
  - A is parent of B
  - B is parent of A
- Which is right?
- There is no wrong answer!
- Each network can express arbitrary P(AB)
- Network does **NOT** encode causal or temporal dynamics

(Non)Uniqueness of Bayes Nets II

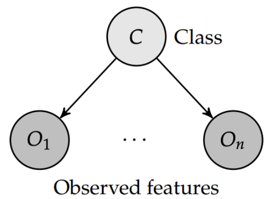
- Can construct valid Bayes net by adding variables incrementally
- For each new variable, connect all influencing variables as parents – new variables never become parents of existing variables (how does this ensure that all variables are conditionally independent of non-descendants given parents?)
- Different order → different Bayesian networks for same distribution

Working with Bayes nets

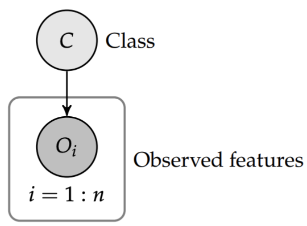
- Can give exponential reduction in storage for joint distribution
- What if we want to answer questions using joint distro, e.g., P(f|h)?
- In the worst case, answering arbitrary queries using a Bayesian network is NP-hard
- This doesn't always occur (depends upon the structure, and the query), so Bayes nets are still useful in practice
- For this class: Mostly used to show relationships between variables

Plate Notation

- A compact way of representing a Bayes net with repeated structure
- Naïve Bayes:



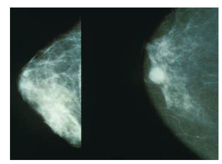
- Plate version:



Bonus material (if time permits)

Another Example

- From: http://opinionator.blogs.nytimes.com/2010/04/25/chances-are/ (attributed to Gerd Gigerenzer)
- "…The probability that one of these women has breast cancer is 0.8 percent. If a woman has breast cancer, the probability is 90 percent that she will have a positive mammogram. If a woman does ***not*** have breast cancer, the probability is 7 percent that she will still have a positive mammogram. Imagine a woman who has a positive mammogram. What is the probability that she actually has breast cancer?"
- 95/100 U.S. doctors answered ~75%



Source: Wikipedia

Understanding Probabilities More Subtly

- Initially, probabilities are "relative frequencies"
- This works well for dice and coin flips
- For more complicated events, this is problematic
- Probability Trump running and winning in 2024?
  - This event only happens once
  - We can't count frequencies
  - Still seems like a meaningful question
- In general, all events are unique
- "Reference Class" problem
- Most things are in the middle
  - Not repeatable and identical
  - Not fully unique – previous, related events may inform us