Review of Probability

CPS 570

Ron Parr

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

Probabilities

- Natural way to represent uncertainty
- People have intuitive notions about probabilities
- Many of these are wrong or inconsistent
- Most people *don't get* what probabilities mean
- Finer details of this question still debated

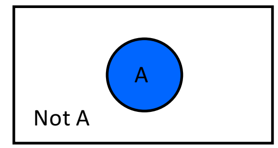
Bogus Probabilistic Reasoning

- Is the sequence 123456 any less likely than any other sequence of lottery numbers?
- Is it good to bet on rare events because they are "due" to come up?
- Cancer clusters

Relative Frequencies

- Probabilities defined over events
- Space of all possible events is the "event space"

Event space:



- Think: Playing blindfolded darts with the Venn diagram…
- P(A) ≅ percentage of dart throws that hit A (assuming a uniform distribution of dart hits over the area of the box)

Understanding Probabilities

- Initially, probabilities are "relative frequencies"
- This works well for dice and coin flips
- For more complicated events, this is problematic
- Probability Trump winning election in 2017?
  - 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

Probabilities and Beliefs

- Suppose I have flipped a coin and hidden the outcome
- What is P(Heads)?
- Note that this is a statement about a *belief*, not a statement about the world
- The world is in exactly one state (at the macro level) and it is in that state with probability 1.
- Assigning truth values to probability statements is very tricky business
- Must reference speakers state of knowledge

Frequentism and Subjectivism

- Frequentists: Probabilities = relative frequencies
  - Purist viewpoint
  - But, relative frequencies often unobtainable
  - Often requires complicated and convoluted assumptions to come up with probabilities
- Subjectivists: Probabilities = degrees of belief
  - Taints purity of probabilities
  - Often more practical

The Middle Ground

- No two events are ever identical, but
- No two events are ever totally unique either
- Probability that Trump will win the election in 2017?
  - We now how states have leaned in the past
  - Performance in debates informs our expectations
- In reality, we use probabilities as beliefs, but we allow data (relative frequencies) to influence these beliefs
- More precisely: We can use Bayes rule to combine our prior beliefs with new data

Why probabilities are good

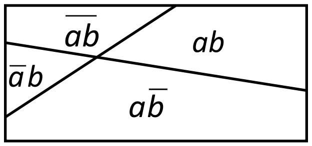
- Subjectivists: probabilities are degrees of belief
- Are all degrees of belief probability?
  - AI has used many notions of belief:
    - Certainty Factors
    - Fuzzy Logic
- Can prove that a person who holds a system of beliefs inconsistent with probability theory can be tricked into accepted a sequence of bets that is guaranteed to lose (Dutch book) in expecation

What are probabilities mathematically?

- Probabilities are defined over random variables
- Random variables represented with capitals: X,Y,Z
- RVs take on values from a finite domain: d(X), d(Y), d(Z)
- We use lower case letters for values from domains
  - X=x asserts: RV X has taken on value x
  - P(x) is shorthand for P(X=x)

Event spaces for binary, discrete RVs

- 2 variable case



- Important: Event space grows exponentially in number of random variables
- Components of event space = atomic events

Domains

- In the simplest case, domains are Boolean
- In general may include many different values
- Most general case: domains may be continuous
- Continuous domains introduce complications

Kolmogorov’s axioms of probability

- $0 \leq P(a) \leq 1$
- P(true) = 1; P(false)=0
- P(a or b) = P(a) + P(b) – P(a and b)
- Subtract to correct for double counting

- Sufficient to completely specify probability theory for discrete variables
- Continuous variables need *density functions*

Atomic Events

- When several variables are involved, it is useful to think about **atomic events**
- Complete assignment to variables in the domain
  - Atomic events are mutually exclusive
  - Exhaust space of all possible events
  - Atomic events = states
- For n binary variables, how many unique atomic events are there?

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 ∧ B) = P(A)P(B)

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

- Are cold and headache independent?

Independence

- If A and B are mutually exclusive:
  P(A ∨ B) = P(A)+P(B) (Why?)

- Examples of independent events:
  - Duke winning NCAA, Dem. winning white house
  - Two successive, fair coin flips
  - My car starting and my iPhone working
  - etc.

- Can independent events be mutually exclusive?

Why Probabilities Are Messy

- Probabilities are not truth-functional
- Computing P(a and b) requires the joint distribution
  - sum out all of the other variables from the distribution
  - It is not a function of P(a) and P(b)
  - It is not a function of P(a) and P(b)
  - It is not a function of P(a) and P(b)
- This fact led to many approximations methods such as certainty factors and fuzzy logic (Why?)
- Neat vs. Scruffy…

The Scruffy Trap

- Reasoning about probabilities correctly requires knowledge of the joint distribution
  - Exponentially large!
  - Very convenient!
- 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)

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)
- Defines the 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"

Conditioning and Belief Update

- Suppose we know P(ABCDE) ← Joint
- Observe B=b, update our beliefs:

$$P(ACDE \mid b) = \frac{P(ABCDE)}{P(b)} = \frac{P(ABCDE)}{\sum_{ACDE} P(AbCDE)}$$

Notation comment: This is a *very* condensed notation.
P(ACDE|b) is not a number; ***it's a distribution***

Example Revisited

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

- What is P(cold|headache)?

Let's Play Doctor

- Suppose 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

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%

Expectation

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

- (1+2+3+4+5+6)/6 = 3.5

Bias

- What if not all events are equally likely?
- Suppose weighted die makes 6 2X more likely that anything else. What is average value of outcome?

- (1 + 2 + 3 + 4 + 5 + 6 + 6)/7 = 3.86

- Probs: 1/7 for 1…5, and 2/7 for 6

- (1 + 2 + 3 + 4 + 5)*1/7 + 6 * 2/7 = 3.86

Expectation in General

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

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

Sums of Expectations

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

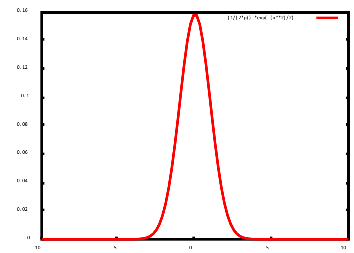
Linearity of Expectation

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

Continuous Random Variables

- Domain is some interval, region, or union of regions
- Uniform case: Simplest to visualize
  (event probability is proportional to area)
- Non-uniform case visualized with extra dimension

Gaussian (normal/bell) distribution:



Requirements on Continuous Distributions

- p(x)>1 *is* possible so long as:

$$\int_x p(x)dx = 1$$

- Don't confuse p(x) and P(X=x)
- P(X=x) for any x is 0!

$$P(x \in A) = \int_A p(x)dx$$

Cumulative Distributions

- When distribution is over numbers, we can ask:
  - P(X>=c) for some c
  - P(X<c) for some c
  - P(a<=X<=b) for some, a and b
- Solve by
  - Summation
  - Integration
- Cumulative sometimes called
  - CDF
  - Distribution function

Sloppy Comment about Continuous Distributions

- In many, many cases, you can generalize what you know about discrete distributions to continuous distributions, replacing "P" with "p" and "Σ" with "∫"

- Proper treatment of this topic requires measure theory and is beyond the scope of the class

Probability Conclusions

- Probabilistic reasoning has many advantages:
  - Solves qualification problem
  - Is better than any other system of beliefs (Dutch book argument)
- Probabilistic reasoning is tricky
  - Some things decompose nicely: linearity of expectation, conjunctions of independent events, disjunctions of disjoint events
  - Some things can be counterintuitive at first: conjunctions of arbitrary events, conditional probability
- Reasoning efficiently with probabilities poses significant data structure and algorithmic challenges for AI

(Roughly speaking, the AI community realized some time around 1990 that probabilities were **the right thing** and has spent the last 20 years grappling with this realization.)