Artificial Intelligence Programming

Probability

Cindi Thompson

Department of Computer Science

University of San Francisco

Uncertainty

- In many interesting agent environments, *uncertainty* plays a central role.
- Actions may have nondeterministic effects.
  - Shooting an arrow at a target, retrieving a web page, moving
- Agents may not know the true state of the world.
  - Incomplete sensors, dynamic environment
- Relations between facts may not be deterministic.
  - Sometimes it rains when it's cloudy.
  - Sometimes I play tennis when it's humid.
- Rational agents will need to deal with uncertainty.

Logic and Uncertainty

- We've already seen how to use logic to deal with uncertainty.
  - $Studies(Bart) \vee WatchesTV(Bart)$
  - $Hungry(Homer) \Rightarrow Eats(Homer, HotDog) \vee Eats(Homer, Pie)$
  - $\exists x Hungry(x)$
- Unfortunately, the logical approach has some drawbacks.

Weaknesses with logic

- Qualifying all possible outcomes.
  - "If I leave now, I'll be on time, unless there's an earthquake, or I run out of gas, or there's an accident ..."
- We may not know all possible outcomes.
  - "If a patient has a toothache, she may have a cavity, or may have gum disease, or maybe something else we don't know about."
- We have no way to talk about the likelihood of events.
  - "It's possible that I'll get hit by lightning today."

Qualitative vs. Quantitative

- Logic gives us a *qualitative* approach to uncertainty.
  - We can say that one event is more common than another, or that something is a possibility.
  - Useful in cases where we don't have statistics, or we want to reason more abstractly.
- Probability allows us to reason *quantitatively*
  - We assign concrete values to the chance of an event occurring and derive new concrete values based on observations.

Uncertainty and Rationality

- Recall our definition of rationality:
  - A rational agent is one that acts to maximize its performance measure.
- How do we define this in an uncertain world?
- We will say that an agent has a *utility* for different outcomes, and that those outcomes have a *probability* of occurring.
- An agent can then consider each of the possible outcomes, their utility, and the probability of that outcome occurring, and choose the action that produces the highest *expected* (or average) utility.
- The theory of combining preferences over outcomes with the probability of an outcome's occurrence is called *decision theory*.

Basic Probability

- A probability signifies a *belief* that a proposition is true.
  - P(BartStudied) = 0.01
  - P(Hungry(Homer)) = 0.99
- The proposition itself is true or false - we just don't know which.
- This is different than saying the sentence is partially true.
  - "Bart is short" - this is *sort of* true, since "short" is a vague term.
- An agent's *belief state* is a representation of the probability of the value of each proposition of interest.

Terminology

- A *Random Variable* (or just variable) is a variable whose value can be described using probabilities
  - Use Upper Case for variables: X, Y, Z, etc.
- Random Variables can have discrete or continuous values (for now, we will assume discrete values)
  - use lower case for values of variables: x, y, x1, x2, etc.
- P(X = x) is the probability that variable X has the value x
  - Can also be written as P(x)

Terminology & Notation

- If variable X can have the values $x_1, x_2, \ldots, x_n$, then the expression P(X) stands for a vector which contains $P(X = x_k)$, for all values x_k of X
  - $P(X) = [P(X = x_1), P(X = x_2), \ldots, P(X = x_n)]$
- For example, If D is a variable that represents the value of a fair die, then
  - P(D) = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6]

Another Example

- Variable W, represents Weather, which can have values sunny, cloudy, rain, or snow.
  - P(W = sunny) = 0.7
  - P(W = cloudy) = 0.2
  - P(W = rain) = 0.08
  - P(W = snow) = 0.02
- P(W) = [0.7, 0.2, 0.08, 0.02]

Notation: AND

$$P(x, y) = P(X = x \wedge Y = y)$$

- Given two fair dice D1 and D2: P(D1 = 3, D2 = 4) = 1/36
- P(X, Y) represents the set of P(x, y) for all values x of X and y of Y.
- Thus, P(D1, D2) represents 36 different values.

Binary random variables

If X has two values (false and true), we can represent:

- $P(X = false)$ as $P(\neg x)$ and
- $P(X = true)$ as $P(x)$

Distributions

- The assignment of probabilities to different outcomes is known as a *distribution*.
- This lets us evaluate the relative frequency of different outcomes.
- We might have a closed-form description of a distribution:
  - Uniform distribution
  - Normal distribution
  - Binomial distribution
- Or we might simply have an enumeration of events and probabilities.

What are Probabilities?

Assertions about possible worlds!

- How probable each world is
- The set of all possible worlds is the *sample space*
- A *probability model* associates a numerical probability $P(\omega)$ with each possible world.
- Sets of possible worlds are called *events*
  - For example, raining and windy, or two die adding to 11

Atomic Events

- We can combine propositions using standard logical connectives and talk about conjunction and disjunction
  - $P(Hungry(Homer) \wedge \neg Study(Bart))$
  - $P(Brother(Lisa, Bart) \vee Sister(Lisa, Bart)$
- A sentence that specifies a possible value for every uncertain variable is called an *atomic event*.
  - Atomic events are mutually exclusive
  - The set of all atomic events is exhaustive
  - An atomic event predicts the truth or falsity of every proposition
- Atomic events will be useful in determining truth in cases with multiple uncertain variables.

Axioms of Probability

- All probabilities are between 0 and 1. $0 \leq P(a) \leq 1$
- Propositions that are necessarily true have probability 1. $P(true) = 1$
- Propositions that are unsatisfiable have probability 0. $P(false) = 0$
- The probability of $(a \vee b)$ is $P(a) + P(b) - P(a \wedge b)$

Everything follows from these axioms

For example, prove $P(x) = 1 - P(\neg x)$

Axioms of Probability

- $0 \leq P(a) \leq 1$
- $P(true) = 1$
- $P(false) = 0$
- $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

Prove $P(x) = 1 - P(\neg x)$

$$P(x \vee \neg x) = P(x) + P(\neg x) - P(x \wedge \neg x)$$
$$1 = P(x) + P(\neg x) - 0$$
$$1 - P(\neg x) = P(x)$$

Prior Probability

- The *prior probability* of a proposition is its probability of taking on a value *in the absence of any other information*.
  - P(Rain) = 0.1, P(Overcast) = 0.4, P(Sunny) = 0.5
- We can also list the probabilities of combinations of variables
  - $P(Rain \wedge Humid) = 0.1, P(Rain \wedge \neg Humid) = 0.1, P(Overcast \wedge Humid) = 0.2, P(Overcast \wedge \neg Humid) = 0.2, P(Sunny \wedge Humid) = 0.15, P(Sunny \wedge \neg Humid) = 0.25$
- This is called a *joint probability distribution*

Continuous Variables

- For continuous variables, we can't enumerate values
- Instead, we use a parameterized function.
  - $P(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{-z^2}{2}} dz$ (Normal distribution)

Bootstrapping

- Where do distributions of priors come from? How can we determine what they are?
- We might estimate based on past observations.
  - Do we have enough data?
  - Is the past always like the present? What is the probability of (for example) the sun going out?
- We might choose an existing model, such as a normal distribution.
  - Which model is best?

Joint Probability

- Probability for all possible values of all possible variables

cavity	toothache	0.04
cavity	$\neg$toothache	0.06
$\neg$cavity	toothache	0.01
$\neg$cavity	$\neg$toothache	0.89

- From the joint, we can calculate anything
- Also called the Joint Probability Distribution (JPD)

Joint Probability

cavity	toothache	0.04
cavity	$\neg$toothache	0.06
$\neg$cavity	toothache	0.01
$\neg$cavity	$\neg$toothache	0.89

- From the joint, we can calculate anything
  - P(cavity) = 0.04 + 0.06 = 0.1
  - P(cavity $\vee$ toothache) = 0.04 + 0.06 + 0.01 = 0.11
  - P(cavity|toothache) = P(c, t)/P(t) = 0.04 / (0.04 + 0.01) = 0.80

JPD Inference

sunny	windy	playTennis	0.1
sunny	windy	$\neg$playTennis	0.1
sunny	$\neg$windy	playTennis	0.3
sunny	$\neg$windy	$\neg$playTennis	0.05
$\neg$sunny	windy	playTennis	0.05
$\neg$sunny	windy	$\neg$playTennis	0.2
$\neg$sunny	$\neg$windy	playTennis	0.1
$\neg$sunny	$\neg$windy	$\neg$playTennis	0.1

- P(sunny)?
- P(playTennis|$\neg$windy)?
- P(playTennis)?
- P(playTennis|sunny $\wedge \neg$windy)?
  - (also written P(playTennis|sunny, $\neg$windy))

JPD Inference

- Joint can tell us everything
- Calculate the joint, read off what you want to know

JPD Inference

- Joint can tell us everything
- Calculate the joint, read off what you want to know
- This will not work!
  - x different variables, each of which has v values
  - Size of joint = v^x
  - For example, 50 variables, each has 7 values, $1.8 * 10^{42}$ table entries

Conditional Probability

- Working with the joint is impractical
- Work with conditional probabilities instead:
- Once we begin to make observations about the value of certain variables, our belief in other variables changes.
  - Once we notice that it's cloudy, $P(Rain)$ goes up.
- this is called *conditional probability*
- Written as: $P(Rain|Cloudy)$
- $P(a|b) = \frac{P(a \wedge b)}{P(b)}$
- or $P(a \wedge b) = P(a|b)P(b)$
  - This is called the *product rule*.

Conditional Probability

- Example: $P(cloudy) = 0.3$
- $P(rain) = 0.2$
- $P(cloudy \wedge rain) = 0.15$
- $P(cloudy \wedge \neg rain) = 0.1$
- $P(\neg cloudy \wedge rain) = 0.1$
- $P(\neg cloudy \wedge \neg rain) = 0.65$
  - Initially, $P(rain) = 0.2$. Once we see that it's cloudy, $P(rain|cloudy) = P\frac{(rain \wedge cloudy)}{P(cloudy)} = \frac{0.15}{0.3} = 0.5$

Independence

- In some cases, we can simplify matters by noticing that one variable has no effect on another.
- For example, what if we add a fourth variable $DayOfWeek$ to our Rain calculation?
- Since the day of the week will not affect the probability of rain, we can assert $P(Rain|Cloudy, monday) = P(Rain|Cloudy, tuesday) ... = P(Rain|Cloudy)$
- We say that $DayOfWeek$ and $Rain$ are independent.
- We can then split the larger joint probability distribution into separate subtables.
- Independence will help us divide the domain into separate pieces.

Bayes' Theorem

- Often, we want to know how a probability changes as a result of an observation.
- Recall the Product Rule:
  - $P(a \wedge b) = P(a|b)P(b)$
  - $P(a \wedge b) = P(b|a)P(a)$
- We can set these equal to each other
  - $P(a|b)P(b) = P(b|a)P(a)$
- And then divide by $P(a)$
  - $P(b|a) = \frac{P(a|b)P(b)}{P(a)}$
- This equality is known as Bayes' theorem (or rule or law).

Bayes' Theorem

- we can generalize this to the case with more than two variables:
  - $P(Y|X,e) = \frac{P(X|Y,e)P(Y|e)}{P(X|e)}$
- We can then recursively solve for the conditional probabilities on the right-hand side.
- In practice, Bayes' rule is useful for transforming the question we want to ask into one for which we have data.

Bayes' theorem example

- Say we know:
  - Meningitis causes a stiff neck in 50% of patients.
  - $P(stiffNeck|Meningitis) = 0.5$
  - Prior probability of meningitis is 1/50000.
  - $P(meningitis) = 0.00002$
  - Prior probability of a stiff neck is 1/20
  - $P(stiffNeck) = 0.05$
- A patient comes to use with a stiff neck. What is the probability she has meningitis?
- $P(meningitis|stiffNeck) = \frac{P(stiffNeck|meningitis)P(meningitis)}{P(stiffNeck)} = \frac{0.5 \times 0.00002}{0.05} = 0.0002$

Why is this useful?

- Often, a domain expert will want diagnostic information. $P(meningitis|stiffNeck)$
- We could derive this directly from statistical information.
- However, if there's a meningitis outbreak, $P(meningitis)$ will change.
- Unclear how to update a direct estimate of $P(meningitis|stiffNeck)$
- But since $P(stiffNeck|meningitis)$ hasn't changed, we can use Bayes' rule to indirectly update the diagnostic information instead.
- This makes our inference system more robust to changes in priors.

Using Bayes Rule

- Rare disease, strikes one in 10,000
- Test for the disease that is 95% accurate:
  - $P(t|d) = .95$
  - $P(\neg t|\neg d) = .95$
- Someone tests positive for the disease, what is the probability they have it?

$P(d|t) =$?

Using Bayes Rule

- $P(d) = 0.0001$
- $P(t|d) = 0.95$
- $P(\neg t|\neg d) = 0.95$

$P(d|t) = P(t|d)P(d)/P(t)$

Using Bayes Rule

- $P(d) = 0.0001$
- $P(t|d) = .95$ and hence $P(\neg t|d) = .05$
- $P(\neg t|\neg d) = .95$ and hence $P(t|\neg d) = .05$

$$\begin{aligned}
P(d|t) &= P(t|d)P(d)/P(t) \\
&= 0.95 * 0.0001/(P(t|d)P(d) + P(t|\neg d)P(\neg d)) \\
&= 0.95 * 0.0001/(0.95 * 0.0001 + 0.05 * 0.9999) \\
&= 0.0019
\end{aligned}$$

Using Bayes Rule

This is somewhat counterintuitive!

- Test is 95% accurate
- Test is positive
- Only a 0.19% chance of having the disease!
- Why?

Using Bayes Rule

- Note that for:
  - $P(a|b) = P(b|a)P(a)/P(b)$
- We needed P(b), which was a little bit of a pain to calculate
- We can often get away with not calculating it!

Using Bayes Rule

- $P(a|b) = \alpha P(b|a)P(a)$
- α is a normalizing constant
  - Calculate $P(b|a)P(a)$ and $P(b|\neg a)P(\neg a)$
  - $\alpha = \frac{1}{P(b|a)P(a)+P(b|\neg a)P(\neg a)}$ No magic here: $\alpha = \frac{1}{P(b)}$
- But you don't need it unless you want exact probabilities

Combining Evidence

- We can extend this to work with multiple observed variables.
- $P(a|b \wedge c) = \alpha P(a \wedge b|c)P(c)$
- This is still hard to work with in the general case. However, if a and b are independent of each other, then we can write:
- $P(a \wedge b) = P(a)P(b)$
- More common is the case where a and b influence each other, but are independent once the value of a third variable is known, This is called *conditional independence*.

Conditional Independence

- Suppose we want to know if the patient has a cavity. Our observed variables are $toothache$ and $catch$.
- These aren't initially independent - if the patient has a toothache, it's likely she has a cavity, which increases the probability of catch.
- Since each is caused by the having a cavity, once we know that the patient does (or does not) have a cavity, these variables become independent.
- We write this as: $P(toothache \wedge catch|cavity) = P(toothache|cavity)P(catch|cavity)$.
- We then use Bayes' theorem to rewrite this as:
- $P(cavity|toothache \wedge catch) = \alpha P(toothache|cavity)P(catch|cavity)P(cavity)$

Conditional Independence

- More generally, Variable A is conditionally independent of variable B, if $P(A|B) = P(A)$
- Examples
  - D: roll of a fair die
  - C: value of a coin flip
  - P(D|C) = P(D)
    P(C|D) = P(C)
- $P(A|B) = P(A) \iff P(B|A) = P(B)$

Conditional Independence

- If A and B are independent, then P(a, b) = P(a)P(b)
  - (Also used as a definition of conditional independence; the two definitions are equivalent)
- $P(a, b) = P(a|b)P(b) = P(a)P(b)$

Probabilities and inference

Probabilities are everywhere!

- NLP
- Machine Learning
- Bayesian Learning - classifying unseen examples based on distributions from a training set.
- Bayesian Networks. Probabilistic “rule-based” systems
  - Exploit conditional independence for tractability
  - Can perform diagnostic or causal reasoning
- Decision networks - predicting the effects of uncertain actions.