



Artificial Intelligence Programming

Introduction to Probability

Chris Brooks

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.
 - $P(\text{BartStudied}) = 0.01$
 - $P(\text{Hungry}(\text{Homer})) = 0.99$
- 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.

Random Values

- A random variable is a variable or proposition whose value is unknown.
- It has a domain of values that it can take on.
- These variables can be:
 - Boolean (true, false) - Hungry(Homer), isRaining
 - Discrete - values taken from a countable domain.
 - Temperature: <hot, cool, mild>, Outlook: <sunny, overcast, rain>
 - Continuous - values can be drawn from an interval such as $[0, 1]$
 - Velocity, time, position
- Most of our focus will be on the discrete case.

Atomic Events

- We can combine propositions using standard logical connectives and talk about conjunction and disjunction
 - $P(\text{Hungry}(\text{Homer}) \wedge \neg \text{Study}(\text{Bart}))$
 - $P(\text{Brother}(\text{Lisa}, \text{Bart}) \vee \text{Sister}(\text{Lisa}, \text{Bart}))$
- 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.
- Propositions that are unsatisfiable have probability 0.
- The probability of $(A \vee B)$ is $P(A) + P(B) - P(A \wedge B)$

Prior Probability

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

Inference with Joint Probability Distr

- The simplest way of doing probabilistic inference is to keep a table representing the joint probability distribution.
- Observe each of the independent variables, then look up the probability of the dependent variable.

	Hum = High Sky = Overcast	Hum = High Sky = Sunny	Hum = Normal Sky = Overcast	Hum = Normal Sky = Sunny
Rain	0.1	0.05	0.15	0.05
\neg Rain	0.2	0.15	0.1	0.2

Inference with Joint Probability Distributions

- We can also use the joint probability distribution to determine the *marginal probability* of the dependent variable by summing all the ways the dependent variable can be true.
 - $P(\text{Rain}) = 0.1 + 0.05 + 0.15 + 0.05 = 0.35$
- What is the problem with using the joint probability distribution to do inference?

Inference with Joint Probability Distributions

- We can also use the joint probability distribution to determine the *marginal probability* of the dependent variable by summing all the ways the dependent variable can be true.
 - $P(\text{Rain}) = 0.1 + 0.05 + 0.15 + 0.05 = 0.35$
- What is the problem with using the joint probability distribution to do inference?
- A problem with n independent Boolean variables requires a table of size 2^n

Conditional Probability

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

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 $P(\text{rain}|\text{Cloudy}, \text{Monday}) = P(\text{rain}|\text{cloudy}, \text{Tuesday})\dots = P(\text{rain}|\text{cloudy})$
- 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?
- $$P(meningitis|stiffNeck) = \frac{P(stiffNeck|meningitis)P(meningitis)}{P(stiffNeck)} = \frac{0.5 \times 0.00002}{0.05} = 0.0002$$

Another example

- Suppose a lab test comes back saying that a patient has cancer. Should we believe it?
- $P(\text{cancer}) = 0.008$
- $P(\text{positiveTest}|\text{cancer}) = 0.98$
- $P(\text{positiveTest}|\neg\text{cancer}) = 0.03$
- We want to know whether *cancer* or $\neg\text{cancer}$ is more likely. $P(\text{cancer}|\text{positive})$

Another example

- We can ignore the denominator for the moment.
- $P(\text{positive}|\text{cancer})P(\text{cancer}) = 0.98 * 0.008 = 0.0078$
- $P(\text{positive}|\neg\text{cancer})P(\neg\text{cancer}) = 0.03 * 0.992 = 0.0298$
- We see that it's more likely that the patient does not have cancer.
- we can get the exact probabilities by normalizing these values.
- $P(\text{cancer}|\text{positive}) = \frac{0.0078}{0.0078+0.0298} = 0.21$

Why is this useful?

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

Combining Evidence with Bayes' Theorem

- We can extend this to work with multiple observed variables.
- $P(a|b \wedge c) = \alpha P(a \wedge b|c)P(c)$
- where α is a normalization parameter representing $\frac{1}{b \wedge c}$
- 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: $P(\textit{toothache} \wedge \textit{catch} | \textit{cavity}) = P(\textit{toothache} | \textit{cavity})P(\textit{catch} | \textit{cavity})$.
- We then use Bayes' theorem to rewrite this as:
- $P(\textit{cavity} | \textit{toothache} \wedge \textit{catch}) = \alpha P(\textit{toothache} | \textit{cavity})P(\textit{catch} | \textit{cavity})P(\textit{cavity})$

Applications of probabilistic inference

- 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.