

Certainty in AI

- The main problem with techniques up to this point is the inability to deal with uncertainty
 - searching**: moves may or may not lead to the planned state
 - game theory**: opponent has a certain probability of making one move or another, based on unknown heuristic
 - reasoning**: axioms are only correct most of the time
 - planning**: actions do not always lead to subsequent state
 - language**: sentences do not always break down into part with equal frequency
- Example: **PCFG** (probabilistic context-free grammars)
 - apply weights to decomposition rules and lexical entries
 - $P(\text{"time" is } N) = 0.7$; $P(\text{"time" is } V) = 0.25$;
 - $P(\text{"time" is } A_{adj}) = 0.05 \rightarrow$ influences likelihood of parse

Let's Make a Deal

- Classic probability problem:
 - Assume you're on the game show "Let's Make a Deal"
 - You are presented with three doors, behind one of which is a brand new car. The other doors have joke prizes.
 - Monty Hall (the host) asks you to pick a door
 - Once you pick a door, he reveals one of the other doors, with a joke prize behind it
 - He then offers you a chance to switch the door that you picked.
 - What do you do? Stay, or switch?



I'll Take Door #1, Monty



- Classic mistake: assuming that prize has equal likelihood of being behind either remaining door
- Actual probability of winning:
 - if you stay: 33%
 - if you switch: 67%

Probability Theory

- Assuming discrete events to start, probability theory deals with the prediction of the likelihood of an event
- Example: flipping 3 coins
 - sample space** (Ω) = all possible outcomes of an experiment
 - $\Omega = \{HHH, HHT, HTH, \dots, TTT\}$
 - event** (A) = a subset of the sample space
 - $A_1 = \text{"2 heads flipped"} = \{HHT, HTH, THH\}$
 - $|A_i| = \#$ of elements in set $i = 3$
 - probability distribution** (probability function) = assignment of probability value to events, such that $P(\emptyset) = 0$, $P(\Omega) = 1$, and $0 \leq P(A_i) \leq 1$
 - for **disjoint** sets $A_j \in F$ (i.e. $A_j \cap A_k = \emptyset$ for $j \neq k$)

$$P\left(\bigcup_{j=1}^n A_j\right) = \sum_{j=1}^n P(A_j)$$
 - probability space** = collection of sample space Ω , field of events F and probability function P .
 - $P(A_i) = |A_i|/|\Omega| = 3/8$

Axioms of Probability

- Probability range**:
 - For all events A : $0 \leq P(A) \leq 1$
 - Probability limits**:
 - $P(\text{true}) = 1$, $P(\text{false}) = 0$
 - Disjunctions**:
 - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
 - Marginalization**:
 - if $\bigcup_j A_j = \Omega$, then $\sum_j P(A_j) = 1$
- **Note**: Probability is like a degree of belief, not a function like its appearance might make it seem

Joint & Conditional Probability

- Joint probability** $P(A, B)$ = the probability of two simultaneous events A and B
 - given disjoint events A_i and A_j , $P(A_i, A_j) = 0$
- Conditional probability** $P(A|B)$ = the probability of one event occurring, given knowledge about another event
 - probability of A alone \rightarrow the **prior** probability of A
 - probability of A given B knowledge \rightarrow the **posterior** probability of A
 - conditional probability can also be written as:

$$P(A|B) = \frac{P(A, B)}{P(B)} = \frac{P(A, B)}{\sum_x P(A, B)}$$

Monty Hall Revisited

- Assume that prizes are cars (C) and goats (G)
 - $F_1: \{CGG, GCG, GGC\}$
 - $A_1: \{CGG\}, A_2: \{GCG\}, A_3: \{GGC\}$;
 - $P(A_1)=P(A_2)=P(A_3)=1/3$
 - $P(\neg A_1)=P(\neg A_2)=P(\neg A_3)=2/3$



- Assume that Monty opens Door #1, #2, #3
 - $F_2 = \{\text{"Door #1"}, \text{"Door #2"}, \text{"Door #3"}\}$
 - $B_1: \text{"Door #1"}, B_2: \text{"Door #2"}, B_3: \text{"Door #3"}$
 - $P(B_1|A_1)=0, P(B_1|A_2)=1/2, P(B_1|A_3)=1/2$
- Probability of winning if you pick Door #1 and stay after Door #2 is opened (WLOG):
 - $P(A_1|B_2) = P(B_2|A_1)P(A_1)/P(B_2)$
 $= [(1/2)(1/3)]/(1/2) = 1/3$
- Therefore, 2/3 probability of winning if you switch

Random Variables

- Instead of referring to values like "stay" and "switch", random variables act as a function between $X: \Omega \rightarrow \mathfrak{R}$, where \mathfrak{R} is the set of real numbers
 - discrete random variable = a function $X: \Omega \rightarrow S$ where S is a countable subset of \mathfrak{R} . If $X: \Omega \rightarrow \{0,1\}$, then X is called an indicator random variable, or a Bernoulli trial
- Example: rolling a 6-sided die
 - $P(X)$ = general expression for the probability of some value stored in X
 - For a "fair" die: $P(X=1) = 1/6$
 - For a weighted, "unfair" die: $P(X=1) = 1/2$

Expectation & Variance

- The **expected value** (μ) for a random variable X is the average value of X
 - not the same as the average of X 's possible values.

$$E(X) = \sum_x xP(x)$$

- The **variance** (σ) of a random variable X is a measure of whether the values of the random variable are consistent over several trials, or whether they tend to vary a lot from the expected value

$$V(X) = E((X-E(X))^2) = E(X^2) - E^2(X)$$

E(X) and V(X) Examples

- Expectation for rolling of a single die:

$$E(X) = \sum_{x=1}^6 xP(x) = 1/6 \sum_{x=1}^6 x = 21/6 = 3\frac{1}{2}$$

- Expectation for rolling of two dice:

$$E(Y) = E(X) + E(X) = 3\frac{1}{2} + 3\frac{1}{2} = 7$$

- Variance for rolling of two dice:

$$V(Y) = E((Y - E(Y))^2) = 5\frac{5}{6}$$

- How do we get the values for these calculations?

Obtaining Probabilities

- When the cases are not as obvious as coins and dice (e.g. occurrence of words in English), P is often found through estimation
- $C(u)$ = the number of times that u occurs over N trials, with $u \leq N$, obviously.
 - unreliable in some cases, since occurrence of certain data values (i.e. words) are rare or non-existent in the data, especially if dataset is relatively small
 - practically, give token probability to all unseen data, to ensure that probability calculations are not zero

Some Probability Rules

- If A and B are **independent events**,
 - $P(A,B) = P(A)P(B)$
 - $P(A|B) = P(A)$
- Multiplication Rule:
 - $P(A,B) = P(B)P(A|B) = P(A)P(B|A)$
- Chain Rule:
 - $P(A_1, A_2, \dots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \dots P(A_n|A_1, A_2, \dots, A_{n-1})$
- Bayes' Theorem:

$$P(B|A) = \frac{P(A,B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$$

- extremely important theorem in the field of probability, where B cannot be observed directly from data

Importance of Bayes' Theorem

- Example #1: Medical diagnosis
 - if $P(\text{cough} | \text{lung cancer}) = 0.95$, is that useful information to somebody who has a cough?
 - more interesting would be $P(\text{lung cancer} | \text{cough})$, but how does one find that out?
 - also need $P(\text{cough}) = 0.3$ and $P(\text{lung cancer}) = 0.001$
 - $P(\text{LC} | \text{C}) = P(\text{C} | \text{LC})P(\text{LC})/P(\text{C}) = (0.95)(0.001)/(0.3) = \mathbf{0.3\%}$
 - On the other hand, if the probability of a cold is about 1 in 4, then $P(\text{cold} | \text{cough}) = \mathbf{79\%}$
 - Note:
 - $P(\text{C})$ cannot be 0 in these cases
 - It is not possible for $P(\text{C} | \text{LC})P(\text{LC})$ to be greater than $P(\text{C})$

Importance of Bayes' Theorem

- Example #2: Courtroom verdicts
 - Given that an evidence test E is valid whenever a suspect is guilty (G) of committing a crime ($P(E|G)=1$), can we infer that finding such evidence implies that the suspect is guilty?
 - It all depends on the reliability of the test, and how guilty people are in general
 - Let's say that 1% of people are generally guilty of this crime, and that the evidence occurs 5% of the time naturally
 - $P(G) = 0.01$
 - $P(E) = 0.05$
 - $P(G|E) = P(E|G)P(G)/P(E) = (1)(0.01)/(0.05) = \mathbf{20\% \text{ guilty}}$
 - What about if 0.05% of people are guilty of this crime, but the evidence occurs rarely (DNA evidence $\sim 10^{-6}\%$)
 - $P(G|E) = (1)(0.0005)/[(0.0005)(1) + (0.95)(10^{-6})] = \mathbf{99.8\%}$

Standard Distributions

- Binomial distribution
 - results when an event has two possible outcomes (i.e. Bernoulli trials), where each trial is independent of all the others

$$b(r; n, p) = \binom{n}{r} p^r (1-p)^{n-r}, \quad \text{where } \binom{n}{r} = n! / ((n-r)! r!) \quad 0 \leq r \leq n$$

- the term $\binom{n}{r}$ counts the number of different possibilities for choosing r objects out of n, not considering the order in which they are chosen.

Standard Distributions (cont'd)

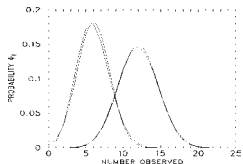
- Normal distributions
 - for continuous cases, sums become integrals and random variable values become ranges
 - the normal distribution is the most commonly-occurring phenomenon in nature, and what occurs in most random situations
 - defined in terms of the mean μ and variance σ :

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$$

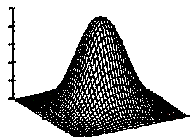
- in addition to "normal distribution", this can also be called a "bell curve", but the AI community always refers to these functions as "Gaussians", after Carl Friedrich Gauss

Distributions Illustrated

- Binomial:



- Gaussian (2-d):



Applied Probability

- Game playing:
 - Use past performance to determine the probability of particular moves, given the present situation
 - Changes heuristic values of certain states, since the opponent might not necessarily choose the "best" move available. This means that some potential dangerous moves can be explored, in the hopes that a more beneficial position can be reached
 - Can compensate for differing heuristics, because moves that the AI agent might not anticipate are still possible, and are not completely discounted in the calculation

Applied Probability

- Reasoning under uncertainty
 - As mentioned before, some propositions might not be absolute
 - Instead of stating absolute truths, reasoning systems would indicate a degree of belief in a particular fact, based on the cumulative probabilities of the steps leading up to that fact.
 - Probabilities are derived from empirical observations
- Planning under uncertainty
 - Plan steps do not lead to definite outcomes, any more than reasoning steps lead to definite facts
 - Path explosion potentially very expensive
 - Multiple paths to goal are possible; choose the path with the highest likelihood of success