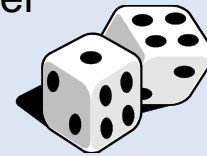


CSCI 400 Artificial Intelligence
David Keil, Framingham State University

Topic 4: Uncertainty and probabilistic reasoning

1. Acting under uncertainty
2. Probability theory and belief
3. Bayes' Theorem
4. Markov models



Inquiry

- Is *probabilistic reasoning* part of intelligence?
- What is the form of knowledge in partially observable and non-deterministic environments?

Objectives

- 4a. Describe ways to operate under conditions of *uncertain knowledge*
- 4b. Use some concepts of probability theory
- 4c. Demonstrate Markov models or Bayesian reasoning

1. Acting under uncertainty

- *Uncertainty*: a property of environments that are
 - Partially observable, or
 - Stochastic (probabilistic)
- *Belief* (quantified in 0..1) replaces *knowledge* (T/F)
- *Inference* under uncertainty is probabilistic
- Uncertainty affects *planning*
- *Related*: fuzzy logic and sets; Bayesian diagnostic reasoning

Conditional planning

- In fully observable environments, plans can include actions conditional on percepts
- In partially observable environments, *belief state* about the environment exists, represented by a *state set*
- Knowledge propositions may describe the agent's knowledge state, using closed-world assumption

Acting under uncertainty

- *Rational decisions* under uncertain information depend on
 - Relative importance of multiple goals
 - Probabilities of achieving goals by alternative actions
- *Diagnosis*: knowledge only provides a *degree of belief* in $[0..1)$
- Degree of belief is expressed using probability theory

Planning under uncertainty

- In partially observable and non-deterministic environments, an agent may *interact* with its environment, obtaining *percepts* to verify or correct planned actions
- For *bounded* uncertainty, sensorless planning may be used to coerce the environment, or contingency planning may be used
- For *unbounded* uncertainty, agent may use execution monitoring and replanning, or continuous planning

Execution monitoring

- *Action monitoring* verifies that the environment is ready for the next action to work
- *Plan monitoring* verifies that the remaining part of plan should work
- *Replanning* entails responding to the unexpected by creating a new plan
- *Continuous planning agents* include planning in their activities, continuously monitoring their environments; similar to partial-order planning

Application of stochastic methods

Some applications:

- *Diagnostic reasoning*, because cause-effect relationship is not always obvious
- *Natural language processing*, because semantics are fuzzy or ambiguous
- *Planning*, because of uncertainty of future events and cause-effect relationships
- *Learning*, because conclusions to draw from experience are ambiguous and probabilistic

Nonmonotonic reasoning

- *Monotone* functions have nondecreasing values as arguments rise; graph never slopes downward
- Mathematical logic is monotonic, in that adding facts makes the set of true assertions larger
- Beliefs, in contrast, change over time
- *Nonmonotonic reasoning* allows for subtracting beliefs and consequences (defeasibility)

Truth maintenance

- When inference is uncertain and contrary evidence arises, *belief revision* must occur
- *Justification based* truth maintenance annotates each sentence in KB with justification, enabling efficient retraction
- Truth-maintenance systems can generate *explanations* for sentences in KB
- TM is NP-hard

2. Probability theory and belief

- In probabilistic reasoning, *belief* is quantified
- *Random process*: one whose outcome is from a set of possibilities that are uncertainly predictable
- *Examples*: tossing a coin, playing lottery, or rolling dice are random processes
- *Sample space*: the set of possible outcomes in a random process
- *Event*: a subset of a sample space
- *Atomic events*: mutually exclusive and exhaustive

Uniform probability space

- A probability space S is a set of possible outcomes of an experiment
- *Example:* S for a die throw is $\{1, 2, 3, 4, 5, 6\}$
- Let $|S| = n$ for probability space S
- Uniform probability function $P : S \rightarrow R$ is defined $P(x) = (1/n)$ for any x in S
- *Example:* Using fair die, $P(3) = 1/6$, because there are 6 possible events, all equally likely

Discrete probability

- *Discrete probability* assumes finite sample space
- *Probability of an event x :* ratio of the number of outcomes in the event to the size of the sample space; $0 \leq P(x) \leq 1$
- For event E in sample space S , $P(E) = |E| \div |S|$

Possibility trees

- A series of events that each has a finite number n of alternative outcomes may be diagrammed by a *possibility tree*, which is n -ary
- *Theorem* (instance of the Multiplication Rule): a series of k events, each with n possible outcomes, has n^k distinct paths from root to leaf of its possibility tree
- Note similarity to state spaces
- *Example*: four throws of a die have 6^4 possible outcomes [pic]

Permutations and combinations

- *Set*: A non-duplicating collection of items, not defined by ordering
- *Sequence*: An aggregate defined by ordering; possibly with duplication
- *Permutations*: The possible orderings of elements of a set
- *Combinations*: The set of subsets of a set, not defined by order
- Our interest is to *count* permutations and combinations in order to determine probabilities

Permutations

- *Definition:* Orderings of objects, without repetition
- There are $(n! = n (n - 1) (n - 2) \times \dots \times 2)$ permutations of n objects
- *Example:* There are $5! = 120$ ways to order the letters A, B, C, D, E
- *k-permutations* ($P(n,k)$): Orderings of n objects taken k at a time; there are $(n! / (n - k)!)$ k -permutations of n objects
- *Example:* there are $P(6, 3) = 120$ different ways to throw a die such that only 1, 2, or 3 show

Combinations

- *Definition:* the number of ways to select from k objects at a time, taken from a set of n objects, without order or repetition
- $C(n, k) = n! / ((n - k)! k!)$
- *Example:* There are $C(36, 6)$ ways to play the lottery where 6 numbers are chosen out of 36
- $C(n, k)$ is also written $\binom{n}{k}$ (“ n choose k ”)

Combinations vs. permutations

- $\text{Combinations}(n, k) = \text{Permutations}(n, k) / k!$
- This is because with combinations, order is not significant, but with permutations, it is
- Hence for every (unordered) combination or selection of k items from a set of n items, there are $(k!)$ (ordered) permutations
- This explains the divisor $k!$ in the ratio between $P(n, k)$ and $C(n, k)$

Example: poker

- **Problem:** how many five-card poker hands are there?
- Note that order is not significant, so we are *selecting* five cards from a possible 52
- **Solution:** there are $C(52, 5)$ hands
 $= (52! / ((52 - 5)! 5!)) = \underline{\hspace{2cm}}$
- **Problem:** what is the probability of each hand?
- **Solution:** $C(52, 5) = \dots$

Kolmogorov's axioms

For sample space S and events $A, B \subseteq S$,

1. $(\forall A) 0 \leq P(A) \leq 1$
 2. $P(S) = 1, P(\emptyset) = 0$
 3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- *Usefulness*: it is irrational (not rewarding) to have beliefs that violate the axioms because they will result in poor bets
 - *Theorems* that follow: $P(A \cup A^c) = 1$;
 $P(A^c) = 1 - P(A)$
 - **Show Venn diagram**

Probability of events

- For event E , $P(E) =$
(number of ways E can occur) \div
(# possible outcomes)
- *Example*: Probability of rolling 9 with two dice is

$$\frac{\text{Count}((3, 6), (4, 5), (5, 4), (6, 3))}{\text{Count}((1, 1), (1, 2), \dots, (6, 6))}$$

$$= 4 / 36 = 1/9$$

Expected values

- For n different possible outcomes of a random process, where a_k is the value of the k^{th} outcome, the expected value of the process is

$$\sum_{k=1..n} a_k P_k$$

- *Examples:*
 - In coin toss, expected value is (0.5)
 - In die throw, expected outcome is $(1 + 2 + 3 + 4 + 5 + 6) / 6 = 3.5$
 - Expected time for linear search is $(n / 2)$

Independent events

- *Intuition:* Independent events can have no effect on each other or overlap with each other
- *Formally:* Events A and B are independent iff $P(A \cap B) = P(A) P(B)$
- Single coin tosses and die rolls are *independent*
- *Example:* For draw of cards, $P(\heartsuit)$ is independent of $P(\text{J or Q or K})$
- For *non-independent* events, notion of *conditional probability* is used, i.e., probability of E_1 given E_2

Random variables

- These are probability functions
- *Boolean*: domain is {T, F}
- *Discrete*: countable domain whose values are mutually exclusive and exhaustive
- *Continuous*: domain is subset of \mathbf{R}

Random variables

- *Definition*: A random variable is a function $f: S \rightarrow \mathbf{R}$ where a probability is assigned to each outcome in the sample space
- A random variable is a distribution that describes the likelihood of outcomes
- *Kinds*: Boolean, discrete, continuous
- *Example*: random variable for throw of two dice:

1	2	3	4	5	6	7	8	9	
		10	11	12					
0	1	2	3	4	5	6	5	4	3
	2	1							

Discrete random variables

- *Definition:* A function from a finite sample space to a finite set of outcomes
- *Example:*
 - Let random variable χ (“Chi”) be the sum of scores for two dice.
 - Then χ takes the value 1 in no case, 2 in 1 case, 3 in 2 cases $\{(1,2), (2,1)\}$, etc.

Random distributions

- Probability that a random variable takes a given value is the probability of the set of outcomes where that holds,
$$P(\chi = k) = P(\{ s \in S \mid \chi(s) = k \})$$
- Probability distribution function, $f_\chi(x)$, maps from outcomes to their probabilities
- Examples:
 - Uniform distribution (flat graph)
 - Gaussian distribution (“normal curve”)

Probability in predicate logic

- A probabilistic knowledge base should give probabilities of all models in predicate logic
- For sentence ϕ , where μ gives probability of a model, $P(\phi) = \sum_{M \text{ s.t. } \phi \text{ holds}} \mu(M)$
- Causal dependencies can be denoted by parent relationships, similar to semantic networks
- Inference may occur if network representation is finite and has fixed structure

Prior probability

- *Prior (unconditional) probability* $P(\alpha)$: degree of belief in the absence of other information
- *Probability distribution*: sequence of probabilities of possible event outcomes
- *Joint probability distribution*: grid of probabilities of all combinations chosen from sets of random variables, e.g., weather, traffic
- *Probability density function*: probability distribution of a continuous variable

Conditional probability

- *Definition:* $P(A | B) = P(A \cap B) \div P(B)$
- *Interpretation:* The probability of event A , given event B , is the probability that both will occur, divided by the probability of B
- *Example:* Given that the first of two coin tosses is heads, what's the chance of two heads?

$$P(c_1 = c_2 = H | c_1 = H)$$

$$= P(\text{both H and } c_1 = H) / P(c_1 = H) = 1/4 / 1/2 = 1/2$$
- It follows from the definition that
 - $P(A \cap B) = P(A | B) P(B)$
 - $P(B) = P(A \cap B) \div P(A | B)$

Independent events and conditional probability

- *Definition:* A and B are *independent* if $P(A \cap B) = P(A)P(B)$, or the probability of each is the probability of itself given the other
- That is, with independent events, knowing that B is true gives us no hint as to whether A is true, and conversely
- In planning, we can predict events better using independent events or conditional probabilities

Conditionally independent events

- A, B are *conditionally independent given event C* iff $P((A \cap B) | C) = P(A | C) P(B | C)$
- This means that if C occurs, then knowledge of B gives no information on $P(A)$
- *Example:* Slow traffic on Rt. 9 is independent of slow traffic in LA but may be causally related to slow traffic on Rt. 128
- Conditional probability of slow traffic on Rt. 9, given construction, is higher than if we knew nothing about the construction situation

Pigeonhole principle

- (Intuition) If n pigeons enter m pigeon holes, and if $n > m$, then at least one hole must have at least two pigeons
- (Formal) *Theorem:* If $|A| > |B|$ then $f: A \rightarrow B$ cannot be injective; i.e., $(\exists a, b \in A, a \neq b) f(a) = f(b)$
- *Example:* at least two people in Framingham have the same last-four, because there are 10K last-4s and more than 10K persons in Framingham
- *Corollary:* Any function from an infinite set to a finite one is non-injective [show non-inj surjection]

3. Bayesian inference

- *Bayesian* reasoning allows diagnosis based on evidence and based on knowledge of statistical properties of problem domain
- *Bayesian belief networks* represent knowledge as directed acyclic graphs that reflect the likelihood of causal relationships between events

Bayesian networks

- *Advantage*: a way to exponentially reduce number of values needed to define a full joint probability distribution
- Also called *belief network* or *knowledge map*
- A BN is a directed acyclic graph with each node containing a random variable, with node X containing value of $P(X \mid \text{Parents}(X))$
- Edge denotes direct influence
- *Example*: Burglary and earthquake are causes for an alarm going off

Bayes' Theorem intuition

- Given some knowledge of an object, and some statistics about the population containing the object, what else can we surmise about the object?
- *Example:* Suppose we know $\frac{2}{3}$ of the numbered cards in a pile are red, and $\frac{1}{4}$ of the face cards are red, and $\frac{3}{4}$ of all the cards are JQK.
- If a card randomly drawn is red, then by Bayes' Theorem we can calculate the probability that it is a J, Q, or K.

Bayes' Theorem

- By Thomas Bayes, pub. 1763
- Helps relate cause and effect by showing how we can learn probability of causes by understanding an effect
- Let H be a set of hypotheses h_1, h_2, \dots , explaining evidence E
- *Theorem:* $P(h_i | E) = P(E | h_i) P(h_i) \div P(E)$

Bayesian representation

- *Full joint distribution entry:*
 $P(x_1, \dots, x_n) = \prod_{i \leq n} P(x_i | x_{i-1}, \dots, x_1)$
- Bayesian network is far more compact than full joint distribution, $n2^k$ vs. 2^n values, where k is maximum number of *local* influences
- [Clarify this]

Bayesian belief networks

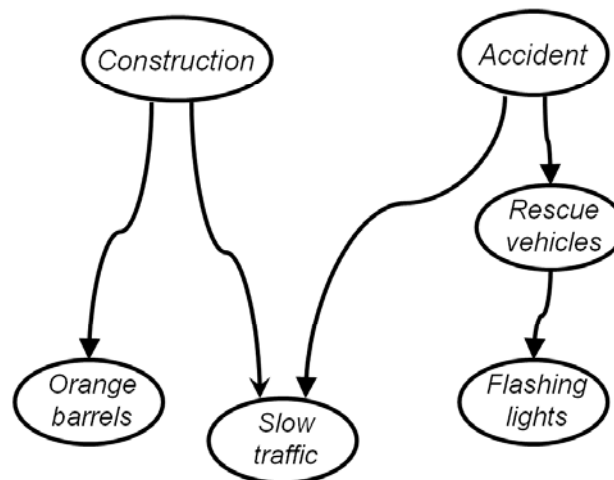
- BBNs are a selective use of Bayes' theorem, which requires a number of parameters exponential in the number of observations
- It is reasonable that some observations don't interact, such as *construction* and *accident* in traffic-jam example
- Hence nodes in belief network depend only on their parent nodes
- Because causality has direction, BBNs have directed acyclic graph (dag) form

A traffic scenario

- Bayesian networks reflect *multiple causalities*
- *Example*: Why is traffic heavy, given evidence of orange barrels or flashing lights?
- Accidents cause heavy traffic and cause emergency vehicles to arrive; these vehicles cause flashing lights
- Construction causes heavy traffic and causes orange barrels to be placed
- *Evidence* is traffic, barrels and/or flashing lights; *cause* is accident or construction

Bayesian net for traffic problem

The unlabeled BBN below reflects causal relations



Labeling a causality relation

- *Evidence* is slow traffic (T), orange barrels (B), flashing lights (L); *causes* are construction (C) or accident (A)
- Suppose we know the following *a priori*:

	<i>Construction</i>	<i>Traffic</i>	<i>Probability</i>
T	T		.3
T	F		.2
F	T		.1
F	F		.4

- Hence all data for $P(C | T)$ can be computed from the table above
- Adding orange-barrels evidence will increase likelihood of the explanation that construction is the cause

Applications of Bayes' Theorem

- By the theorem, some medical screening tests may be useful but more accurate results may be needed to diagnose a disease, because such tests may yield false positives or negatives
- *Example*: Suppose 0.5% of people have a disease, and a test has false positive rate of 3% and false negative rate of 1%
- Then under Bayes' theorem, 99.995% of negative results are correct, but only 14% of persons with positive results actually have the disease

4. Markov models

- A *Markov state machine* or chain is a system with a finite number of observable states, and with probabilistic transitions between states
- *Example*: weather at any location
- *Markov assumption*: current state depends only on finite history of previous states
- *Nth-order Markov process*: state depends only on nth-previous state
- To improve approximations, increase number of state variables or order of Markov process

Markov decision processes

- Defined by initial state s_0 , transition model $T(s, a, s')$, and reward function $R(s)$
- A solution specifies a *policy* $\pi(s)$: what agent should do given any state of environment
- Policies have *expected utilities*: utility of possible environment histories generated by it
- Optimal (maximal-utility) policy is called π^*
- *Proper policy*: one certain to reach a terminal state
- Future rewards may be discounted in deciding expected utility

Markov chains

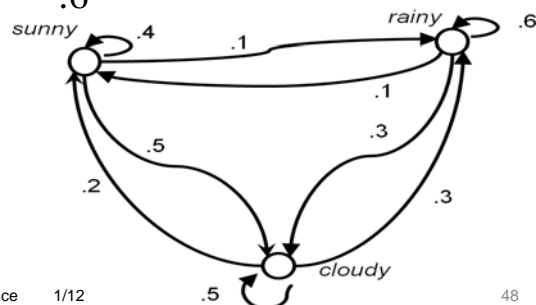
- Probability of being in a given state at a given time is dependent on state at previous times
- *First-order Markov chain* is one where probability of present state depends *only* on previous state
- *Example*: weather at any location

Example: Weather

- Let states be {sunny, cloudy, rainy}
- Let transitions be as follows:

	sunny	cloudy	rainy
sunny	.4	.5	.1
cloudy	.2	.5	.3
rainy	.1	.3	.6

- First-order Markov model:



Querying a Markov model

- *Example problem:* If it's rainy today, what is the probability that it will be rainy two days from now?
- *Solution:* Following the Markov model on previous slide, find
- $P(r, r, r) + P(r, s, r) + P(r, c, r)$
 $= (.6)(.6) + (.1)(.1) + (.3)(.3) = 0.46$

Hidden Markov models

- Let state be hidden, i.e., let observation be a probability function of current state
- *Example:* Noisy acoustic signals in speech recognition
- *Application:* Viterbi algorithm for decoding phonemes
 - Uses a table (dynamic programming) to update probability estimates
 - Algorithm starts with phoneme observations, returns most likely English spelling

Application: Speech recognition

- Bayesian probabilistic inference is used
- Where *words* is a sequence of words, *signal* is sound, we want to maximize $P(\text{words} \mid \text{signal}) = \alpha P(\text{signal} \mid \text{words}) P(\text{words})$
- *Acoustic model*: $P(\text{signal} \mid \text{words})$
- *Language model*: $P(\text{words})$
- HMM for toe-mah-toe | toe-may-toe:
[pic 9]

Concepts

action monitoring	event	plan monitoring
atomic event	expected outcome	prior probability
Bayes' theorem	hidden Markov model	probability density function
Bayesian inference	independent events	probability theory
Bayesian network	Kolmogorov axioms	random variable
belief network	Markov assumption	rational agent
belief state	Markov chains	rational decision
chain rule	Markov process	resolution proof
circumscription	minimal model	sample space
closed world assumption	modal logic	truth maintenance
combination	model	unconditional probability
conditional probability	nonmonotonic reasoning	
	permutation	

References

Susanna Epp. *Discrete Mathematics with Applications*. Brooks/Cole, 2011.

George Luger. *Artificial Intelligence*. Addison Wesley, 2005.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach, 2nd ed.* Prentice Hall, 2003.