Uncertainty

CSC 548, Artificial Intelligence II

Uncertainty

- General situation:
  - Observed variables (evidence): agent knows certain things about the state of the world
  - Unobserved variables: agent needs to reason about other aspects
  - Model: agent knows something about how the known variables relate to the unknown variables
- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge

Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
  - $R =$ is it raining?
  - $T =$ is it hot or cold?
  - $D =$ How long will it take to drive to work?
- We denote random variables with capital letters
- Random variables have domains
  - $R \in \{\text{true}, \text{false}\}$
  - $T \in \{\text{hot}, \text{cold}\}$
  - $D \in [0, \infty)$

Probability Distributions

- Associate a probability with each value
- Example: temperature $P(T)$

T	P
hot	0.5
cold	0.5

- Example: weather $P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3

Probability Distributions

- Unobserved random variables have distributions
- A distribution is a table of probabilities of values
- A probability is a single number

  $P(W = \text{rain} = 0.1$
- Must have:

  $\forall x P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Joint Distributions

- A joint distribution over a set of random variables $X_1, X_2, \ldots, X_n$ specifies a real number for each assignment (or outcome):

  $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$

  $P(x_1, x_2, \ldots, x_n)$

- Must obey

$$P(x_1, x_2, \ldots, x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots, x_n)} P(x_1, x_2, \ldots, x_n) = 1$$

- Size of distribution of n variables with domain sizes d?
  - Only practical to write out small distributions

Joint Distribution

- Example:

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables
- Probabilistic models:
  - (Random) variables with domains
  - Assignments are called outcomes
  - Joint distributions: say whether assignments (outcomes) are likely
  - Normalized: sum to 1.0
  - Ideally only certain variables directly interact
- Constraint satisfaction problems:
  - Variables with domains
  - Constraints: state whether assignments are possible
  - Ideally only certain variables directly interact

Events

- An event is a set E of outcomes

$$P(E) = \sum_{(x_1,\ldots,x_n)\in E} P(x_1, \ldots, x_n)$$

- From a joint distribution we can calculate the probability of any event
- Typically, the events we care about are partial assignments, for example $P(T = \text{hot})$

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): combine collapsed rows by adding
- Example:
  - $P(t) = \sum_s P(t, s) \rightarrow P(T = \text{hot}) = 0.5, P(T = \text{cold}) = 0.5$
  - $P(w) = \sum_s P(t, s) \rightarrow P(S = \text{sun}) = 0.6, P(S = \text{rain}) = 0.4$
- $P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$

Conditional Probabilities

- A simple relation between joint and conditional probabilities
- Definition:

$$P(a \mid b) = \frac{P(a, b)}{P(b)}$$

- Example:

$$P(W = s \mid T = c) = \frac{P(W = s, T = c)}{P(T = c)} = \frac{0.2}{0.5} = 0.4$$

Normalization

- Select the joint probabilities matching the evidence
- Normalize the selection
- Example:

$$
\begin{aligned}
P(W = s \mid T = c) &= \frac{P(W = s, T = c)}{P(T = c)} \\
&= \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)} \\
&= \frac{0.2}{0.2 + 0.3} = 0.4
\end{aligned}
$$

Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (for example, from joint)
- We generally compute conditional probabilities
    - These represent the agent's beliefs given the evidence
- Probabilities change with new evidence
    - Observing new evidence causes beliefs to be updated

Inference by Enumeration

- General case:
  - Evidence variables: $E_1, \ldots, E_k = e_1, \ldots, e_k$
  - Query variable: Q
  - Hidden variables: $H_1, \ldots, H_r$
- We want: $P(Q \mid e_1, \ldots, e_k)$
- Steps:
  1. Select the entries consistent with the evidence
  2. Sum out H to get joint of Query and evidence
  3. Normalize

Product Rule

- Sometimes we have conditional distributions but want the joint

$$P(y)P(x \mid y) = P(x, y) \Leftrightarrow P(x \mid y) = \frac{P(x, y)}{P(y)}$$

The Chain Rule

- More generally, we can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2 \mid x_1)P(x_3 \mid x_1, x_2)$$

- General form:

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i \mid x_1, \ldots, x_{i-1})$$

Bayes' Rule

- Two ways to factor a joint distribution over two variables:

$$P(x, y) = P(x \mid y)P(y) = P(y \mid x)P(x)$$

- Dividing, we get

$$P(x \mid y) = \frac{P(y \mid x)}{P(y)} P(x)$$

- Why is this useful?
  - We can build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems

Inference with Bayes' Rule

- Example: diagnostic probability from causal probability

$$P(\text{cause} \mid \text{effect}) = \frac{P(\text{effect} \mid \text{cause})P(\text{cause})}{P(\text{effect})}$$

Independence

- Two variables are independent, denoted $X \perp\!\!\!\perp Y$, in a joint distribution if:

$$P(X, Y) = P(X)P(Y)$$

$$\forall x, y P(x, y) = P(x)P(y)$$

  - Says the joint distribution factors into a product of two simple ones
  - Usually variables are not independent

- Can use independence as a modeling assumption
  - Independence can be a simplifying assumption
  - Empirical joint distributions: at best “close” to independent

Conditional Independence

- Example: $P(\text{Toothache}, \text{Cavity}, \text{Catch})$
- If I have a cavity, the probability that the probe catches in it does not depend on whether I have a toothache.
  - $P(+\text{catch} \mid +\text{toothache}, +\text{cavity}) = P(+\text{catch} \mid +\text{cavity})$
- The same independence holds if I do not have a cavity:
  - $P(+\text{catch} \mid +\text{toothache}, -\text{cavity}) = P(+\text{catch} \mid -\text{cavity})$
- Catch is conditionally independent of Toothache given Cavity:
  - $P(\text{Catch} \mid \text{Toothache}, \text{Cavity}) = P(\text{Catch} \mid \text{Cavity})$

Conditional Independence

- Unconditional (absolute) independence is rare
- Conditional independence is our most basic robust form of knowledge about uncertain environments.
- $X \perp\!\!\!\perp Y \mid Z$: X is conditionally independent of Y given Z
  - If and only if:

    $\forall x, y, z : P(x, y \mid z) = P(x \mid z)P(y \mid z)$
  - or, equivalently, if and only if:

    $\forall x, y, z : P(x \mid z, y) = P(x \mid z)$

Reasoning over Time or Space

- Often, we want to reason about a sequence of observations
  - Speech recognition
  - Robot localization
  - Medical monitoring
- Need to introduce time (or space) into our models

Markov Models

- Value of X at a given time is called the state
  - TODO figure
- Parameters: called transition probabilities or dynamics, specify how the state evolves over time (also, initial state probabilities)
- Stationary assumption: transition probabilities the same at all times
- Same as MDP transition model, but no choice of action

Joint Distribution of a Markov Model

- TODO figure
- Joint distribution:

  $$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_3)$$

- More generally:

$$\begin{aligned} P(X_1, X_2, \ldots, X_n) &= P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2) \ldots P(X_T \mid X_{T-1}) \\ &= P(X_1) \prod_{t=2}^{T} (P(X_t \mid X_{t-1}) \end{aligned}$$

- Questions to be resolved:
  - Does this indeed define a joint distribution?
  - Can every joint distribution be factored this way, or are we making some assumptions about the joint distribution by using this factorization?

Chain Rule and Markov Models

- From the chain rule, every joint distribution over X_1, X_2, X_3, X_4 can be written as:

  $P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1, X_2)P(X_4 \mid X_1, X_2, X_3)$

- Assuming that $X_3 \perp\!\!\!\perp X_1 \mid X_2$ and $X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$ results in the expression from the previous slide:

  $P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_3)$

Chain Rule and Markov Models

- From the chain rule, every joint distribution over $X_1, X_2, \ldots, X_T$ can be written as:

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t \mid X_1, X_2, \ldots, X_{t-1})$$

- Assuming that for all t:

  $X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$

  gives us the expression

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t \mid X_{t-1})$$

Implied Conditional Independence

- We assumed: $X_3 \perp\!\!\!\perp X_1 \mid X_2$ and $X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$
- Do we also have $X_1 \perp\!\!\!\perp X_3, X_4 \mid X_2$?
- Yes, proof:

$$
\begin{aligned}
P(X_1 \mid X_2, X_3, X_4) &= \frac{P(X_1, X_2, X_3, X_4)}{P(X_2, X_3, X_4)} \\
&= \frac{P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_3)}{\sum_{x_1} P(x_1)P(X_2 \mid x_1)P(X_3 \mid X_2)P(X_4 \mid X_3)} \\
&= \frac{P(X_1, X_2)}{P(X_2)} \\
&= P(X_1 \mid X_2)
\end{aligned}
$$

Markov Models Recap

- Explicit assumption for all t, $X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$
- Consequence: the joint distribution can be written as:

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t \mid X_{t-1})$$

- Implied conditional independences: past variables independent of future variables given the present
- Additional explicit assumption: $P(X_t \mid X_{t-1})$ is the same for all t

Stationary Distributions

- For most chains:
  - Influence of the initial distribution gets less and less over time
  - The distribution we end up in is independent of the initial distribution
- Stationary Distribution:
  - The distribution we end up with is called the stationary distribution P_∞ of the chain
  - It satisfies

$$P_\infty(X) = P_{\infty+1}(X) = \sum_x P(X \mid x) P_\infty(x)$$

Hidden Markov Models

- Markov chains not so useful for most agents
  - Need observations to update your beliefs
- Hidden Markov Models (HMMs)
  - Underlying Markov chain over states X
  - Agent observes outputs (effects) at each time step
- A HMM is defined by:
  - Initial distribution: $P(X_1)$
  - Transitions: $P(X_t \mid X_{t-1})$
  - Emissions: $P(E_t \mid X_t)$

Joint Distribution of an HMM

- TODO figure
- Joint distribution:

  $P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1 \mid X_1)P(E_2 \mid X_2)P(X_3 \mid X_2)P(E_3 \mid X_3)$
- More generally, \$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1 \mid X_1) \prod_{t=2}^{T} P(X_t \mid X_{t-1})P(E_t \mid X_t)$
- Questions to be resolved:
  - Does this indeed define a joint distribution?
  - Can every joint distribution be factored this way, or are we making some assumptions about the joint distribution by using this factorization?

Chain Rule and HMMs

- From the chain rule, every joint distribution over $X_1, E_1, \ldots, X_T, E_T$ can be written as:

  $P(X_1, E_1, \ldots, X_T, E_T) =$

  $P(X_1)P(E_1 \mid X_1)$

  $$\prod_{t=1}^{T} P(X_t \mid X_1, E_1, \ldots, X_{t-1}, E_{t-1})P(E_t \mid X_1, E_1, \ldots, X_{x-1}, E_{t-1}, X_t)$$

Chain Rule and HMMs

- Assuming that for all t:
  - State independent of all past states and all past evidence given the previous state

    $X_t \perp\!\!\!\perp X_1, E_1, \ldots, X_{t-2}, E_{t-2}, E_{t-1} \mid X_{t-1}$
  - Evidence is independent of all past states and all past evidence given the current state

    $E_t \perp\!\!\!\perp X_1, E_1, \ldots, X_{t-2}, E_{t-2}, X_{t-1}, E_{t-1} \mid X_t$

  we get the expression

$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1 \mid X_1) \prod_{t=2}^{T} P(X_t \mid X_{t-1})P(E_t \mid$$

Implied Conditional Independence

- Many implied conditional independences, for example

  $E_1 \perp\!\!\!\perp X_2, E_2, X_3, E_3 \mid X_1$

- To prove them:
  - Approach 1: follow similar (algebraic) approach to what we did for Markov models
  - Approach 2: directly from the graph structure

Real HMM Examples

- Speech recognition HMMs:
  - Observations are acoustic signals (continuous valued)
  - States are specific positions in specific words
- Machine translation HMMs:
  - Observations are words (tens of thousands)
  - States are translation options
- Robot tracking:
  - Observations are range readings (continuous)
  - States are positions on a map (continuous)

Filtering / Monitoring

- Filtering, or monitoring, is the task of tracking the distribution $B_t(X) = P_t(X_t \mid e_1, \dots, e_t)$ the belief state over time
- We start with $B_1(X)$ in an initial setting, usually uniform
- As time passes, or we get observations, we update $B(X)$
- The Kalman filter was invented in the 1960s and first implemented as a method of trajectory estimation for the Apollo program

Passage of Time

- Assume we have current belief $P(X \mid \text{evidence to date})$
- The after one time step passes:

$$
\begin{aligned}
P(X_{t+1} \mid e_{1:t}) &= \sum_{x_t} P(X_{t+1}, x_t \mid e_{1:t}) \\
&= \sum_{x_t} P(X_{t+1} \mid x_t e_{1:t}) P(x_t \mid e_{1:t}) \\
&= \sum_{x_t} P(X_{t+1} \mid x_t) P(x_t \mid e_{1:t})
\end{aligned}
$$

or compactly:

$$B'(X_{t+1} = \sum_{x_t} P(X' \mid x_t) B(x_t)$$

- Basic idea: beliefs get "pushed" through the transitions

Observation

- Assume we have current belief $P(X \mid \text{previous evidence})$
- Then after evidence comes in:

$$
\begin{aligned}
P(X_{t+1} \mid e_{1:t+1}) &= \frac{P(X_{t+1}, e_{t+1} \mid e_{1:t})}{P(e_{t+1} \mid e_{1:t}} \\
&\propto_{X_{t+1}} P(X_{t+1}, e_{t+1} \mid e_{1:t}) \\
&= P(e_{t+1} \mid e_{1:t}, X_{t+1}) P(X_{t+1} \mid e_{1:t}) \\
&= P(e_{t+1} \mid X_{t+1}) P(X_{t+1} \mid e_{1:t})
\end{aligned}
$$

or compactly:

$$B(X_{t+1})_{\propto_{X_{t+1}}} P(e_{t+1} \mid X_{t+1}) B'(X_{t+1})$$

- Basic idea: beliefs get "reweighted" by likelihood of evidence

The Forward Algorithm

- We are given evidence at each time step and want to know $B_t(X) = P(X_t \mid e_{1:t})$
- We can derive the following updates

$$\begin{aligned}
P(x_t \mid e_{1:t}) &\propto_X P(x_t, e_{1:t}) \\
&= \sum_{x_{t-1}} P(x_{t-1}, x_t, e_{1:t}) \\
&= \sum_{x_{t-1}} P(x_{t-1}, e_{1:t-1}) P(x_t \mid x_{t-1}) P(e_t \mid x_t) \\
&= P(e_t \mid x_t) \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1}, e_{1:t-1})
\end{aligned}$$

Online Belief Updates

- Every time step, we start with current $P(X \mid \text{evidence})$
- We update for time:

$$P(x_t \mid e_{1:t-1}) = \sum_{x_{t-1}} P(x_{t-1} \mid e_{e_{1:t-1}}) P(x_t \mid x_{t-1})$$

- We update for evidence:

$$P(x_t \mid e_{1:t}) \propto_X P(x_t \mid e_{1:t-1}) P(e_t \mid x_t)$$

- The forward algorithm does both at once (and does not normalize)

Particle Filtering

- Filtering: approximate solution
- Sometimes $|X|$ is too big to use exact inference
  - $|X|$ may be too big to even store $B(X)$
  - For example, X is continuous
- Solution: approximate inference
  - Track samples of X, not all values
  - Samples are called particles
  - Time per step is linear in the number of samples
  - But, the number needed may be large
  - In memory: list of particles, not states
- Particle is just a new name for sample

Representation: Particles

- Our representation of $P(X)$ is now a list of N particles (samples)
  - Generally, $N \ll |N|$
  - Storing a map from X to counts would defeat the point
- $P(X)$ approximated by number of particles with value x
  - So, many x may have $P(x) = 0$
  - More particles, more accuracy
- For now, all particles have a weight of 1

Particle Filtering: Elapse Time

- Each particle is moved by sampling its next position from the transition model

  $x' = \text{sample}(P(X' \mid x))$

  - This is like prior sampling – samples' frequencies reflect the transition probabilities
- This captures the passage of time
  - If enough samples, close to exact values before and after (consistent)

Particle Filtering: Observe

- Slightly trickier
  - Do not sample observation, fix it
  - Similar to likelihood weighting, downweight samples based on evidence

    $w(x) = P(e \mid x)$

    $B(X) \propto P(e \mid X)B'(X)$

  - As before, the probabilities do not sum to one, since all have been downweighted (in fact they now sum to (N times) an approximation of $P(e))

Particle Filtering: Resample

- Rather than tracking weighted samples, we resample
- N times, we choose from our weighted sample distribution (that is, draw with replacement)
- This is equivalent to renormalizing the distribution
- Now the update is complete for this time step, continue with the next one

Dynamic Bayes Nets (DBNs)

- We want to track multiple variables over time, using multiple sources of evidence
- Idea: repeat a fixed Bayes net structure at each time
- Variables from time t can condition on those from $t-1$
- Dynamic Bayes nets are a generalization of HMMs

DBN Particle Filters

- A particle is a complete sample for a time step
- Initialize: generate prior samples for the $t = 1$ Bayes net
  - Example particle: $G_1^a = (3, 3) G_1^b = (5, 3)$
- Elapse time: sample a successor for each particle
  - Example successor: $G_2^a = (2, 3) G_2^b = (6, 3)$
- Observe: weight each entire sample by the likelihood of the evidence conditioned on the sample
  - Likelihood: $P(E_1^a \mid G_1^a) P(E_1^b \mid G_1^b)$
- Resample: select prior samples (tuples of values) in proportion to their likelihood