Hidden Markov Models

Andrea Danyluk

April 3, 2017

With thanks to CS188 slides, as well as content from University of Washington CSE515, Penn State Stats, Yale University Stats, and others.

Announcements

- Filtering assignment is posted
- Start thinking about final projects
- Will return midterm exams no later than Friday

Today's Lecture

- Quick review/reminders
- Hidden Markov Models

Probabilistic Reasoning

- General situation:
  - **Observed variables** (evidence): Agent knows certain things about the state of the world (e.g., sensor readings or symptoms)
  - **Unobserved variables**: Agent needs to reason about other aspects (e.g., where an object is or what disease is present)
  - **Model**: Agent knows something about how the known variables related to the unknown variables
- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge

Joint Distributions

A joint distribution over a set of random variables X_1, X_2,… X_n specifies a probability for each possible outcome (i.e., assignment).

	female	male
kate	0.04	0.0
kim	0.02	0.01
michael	0.01	0.1
tom	0.0	0.05
other	0.43	.34

P(kim ∧ male) = 0.01

P(kim, male) = 0.01

If these are all the random variables in the "world", then this table gives the **full joint** probability distribution. All entries sum to 1.

Events

- From a joint probability distribution, can calculate the probability of any event
  - P(kate)? P(male)?
  - P(michael ∧ female)?
  - P(kate ∨ kim)?

	female	male
kate	0.04	0.0
kim	0.02	0.01
michael	0.01	0.1
tom	0.0	0.05
Other	0.43	.34

Conditional (Posterior) Probabilities

A simple relationship between joint and conditional probabilities

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

P(female|michael)

	female	male
kate	0.04	0.0
kim	0.02	0.01
michael	0.01	0.1
tom	0.0	0.05
Other	0.43	.34

= (female, michael) / P(michael)

= 0.01 / 0.11

= 0.09

"given that *michael* is all I know, what is the probability that you're female"

Conditional Distributions

Conditional distributions are probability distributions over some variables given fixed values of others.

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

P(Name|Sex)

	female	male
kate	0.04	0.0
kim	0.02	0.01
michael	0.01	0.1
tom	0.0	0.05
Other	0.43	.34

	S=f
kate	0.08
kim	0.04
michael	0.02
tom	0.0
Other	0.86

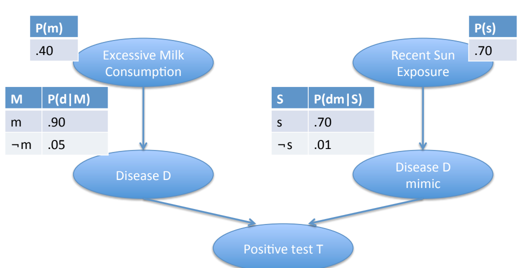
	S=m
kate	0.0
kim	0.02
michael	0.2
tom	0.1
Other	.68

Conditional Independence

- Say we have three random variables: R = rash; T = test for a particular disease; D = the disease, which sometimes causes a rash.
- We can say that R and T are conditionally independent, given information about D.
- P(R|T,D) = P(R|D). That is, if I have the disease, the probability that I expect a rash does not depend on how the test turns out.
  - P(T|R,D) = P(T|D)
  - P(R, T|D) = P(R|D)P(T|D)
- We say T and R are conditionally independent given D.

Bayesian Network

- Concise representation for a joint probability distribution
- Explicitly represents dependencies among random variables



P(m)
.40

P(s)
.70

M	P(d|M)
m	.90
¬m	.05

S	P(dm|S)
s	.70
¬s	.01

D	Dm	P(t|D, Dm)
d	dm	.95
d	¬dm	.80
¬d	dm	.10
¬d	¬dm	.05

A patient tests positive for Disease D. The patient also reports drinking a great deal of milk.
What's the probability that the patient has Disease D?

Probability Recap

- Conditional probability $P(x|y) = \frac{P(x,y)}{P(y)}$
- Product rule $P(x,y) = P(x|y)P(y)$
- Chain rule $$P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1,X_2)\ldots = \prod_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1})$$
- X, Y independent if and only if: $\forall x,y : P(x,y) = P(x)P(y)$
- X and Y are conditionally independent given Z if and only if: $X \perp\!\!\!\perp Y|Z$
  $\forall x,y,z : P(x,y|z) = P(x|z)P(y|z)$

Space and Time

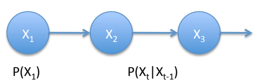
- Bayesian networks are generally much more compact than the full joint probability distribution
  - Joint distribution: $O(2^n)$
  - Bayes net: $O(n2^k)$, where k is the max # parents a node can have
- But the complexity of inference is still exponential in the number of random variables in the worst case

Reasoning over Time

- Often, we want to reason about a sequence of observations
  - Speech recognition
  - Robot localization
  - Medical monitoring
- Need to introduce time into our models
- Basic approach: Hidden Markov Models (HMMs)
- More general: dynamic Bayesian networks

Markov Models

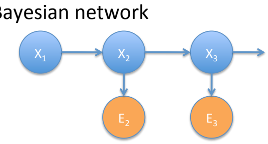
- A Markov Model is a chain-structured Bayesian network



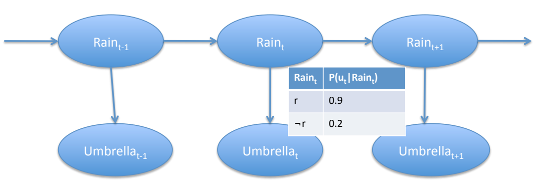
- Value of X at a given time is called the state
- Parameters:
  - Initial probabilities
  - transition probabilities specify how the state evolves over time

Hidden Markov Models

- Markov chains not terribly useful for most agents
  - At some point, stop knowing anything
  - Need observations to update beliefs
- Hidden Markov Models (HMMs)
  - Underlying Markov chain over states S
  - You observe outputs (effects) at each time step
  - As a Bayesian network



$Rain_{t-1}$	$P(r_t \mid Rain_{t-1})$
r	0.7
¬r	0.3

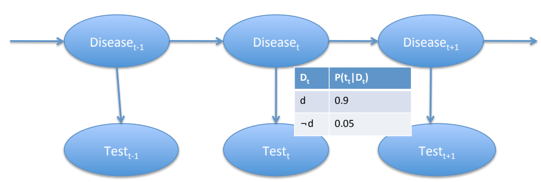


$Rain_t$	$P(u_t \mid Rain_t)$
r	0.9
¬r	0.2

An HMM is defined by:

Initial distribution: $P(X_1)$
Transitions: $P(X|X_{-1})$
Emissions: $P(E|X)$

D_{t-1}	$P(d_t \mid D_{t-1})$
d	.99
¬d	0.2



D_t	$P(t_t \mid D_t)$
d	0.9
¬d	0.05

An HMM is defined by:

Initial distribution: $P(X_1)$
Transitions: $P(X|X_{-1})$
Emissions: $P(E|X)$

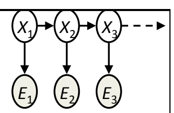
Conditional Independence

- HMMs have two important independence properties
  - *Markov* hidden process: Future depends on the past via the present
  - Current observation (emission) is independent of all else given the current state

Filtering = State Estimation

- Process of computing the belief state (posterior distribution over the most recent state), given evidence to date
- Begin with P(X) in an initial setting, usually uniform
- As time passes/get observations update belief state

Chain Rule and HMMs



- From the chain rule, *every* joint distribution over $X_1, E_1, X_2, E_2, X_3, E_3$ can be written as:

$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1, E_1)P(E_2|X_1, E_1, X_2)$$
$$P(X_3|X_1, E_1, X_2, E_2)P(E_3|X_1, E_1, X_2, E_2, X_3)$$

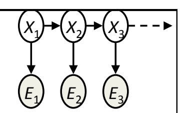
- *Assuming* that

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

gives us:

$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1)P(E_2|X_2)P(X_3|X_2)P(E_3|X_3)$$

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:

$$P(X_1, E_1, \ldots, X_T, E_T) = P(X_1)P(E_1|X_1)\prod_{t=2}^{T} P(X_t|X_1, E_1, \ldots, X_{t-1}, E_{t-1})P(E_t|X_1, E_1, \ldots, X_{t-1}, E_{t-1}, X_t)$$

- *Assuming* that for all *t*:
  - State independent of all past states and all past evidence given the previous state, i.e.:

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

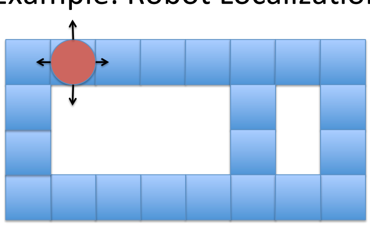
  - Evidence is independent of all past states and all past evidence given the current state, i.e.:

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

gives us the expression posited on the earlier slide:

$$P(X_1, E_1, \ldots, X_T, E_T) = P(X_1)P(E_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(E_t|X_t)$$

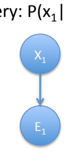
Example: Robot Localization



X_t is the location of the robot. Domain is the set of empty squares
Don't know where robot starts; assume uniform distribution over all squares
Sensor model: 4 bits (whether a wall in each direction); each sensor's error rate is ε
Neighbors(s) is a set of empty squares adjacent to s
Equally likely to move in any valid direction

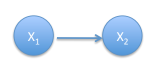
Inference: Base Cases

- Observation
  - Given: $P(X_1)$, $P(e_1|X_1)$
  - Query: $P(x_1|e_1)$ for all x_1



$P(x_1|e_1) = P(e_1,x_1)$ / $P(e_1)$
[Normalization step: do at the end.]
Focus on:
$P(e_1|x_1)\ P(x_1)$

- Passage of Time
  - Given: $P(X_1)$, $P(X_2|X_1)$
  - Query: $P(x_2)$ for all x_2



$P(x_2) = \Sigma_{\text{all } x1} P(x_2|x_1)\ P(x_1)$

Passage of Time

- Assume we have current belief P(X | evidence to date)

$B(X_t) = P(X_t|e_{1:t})$

$X_1 \rightarrow X_2$

- Then, after one time step passes:

$$P(X_{t+1}|e_{1:t}) = \sum_{x_t} P(X_{t+1}, x_t|e_{1:t})$$
$$= \sum_{x_t} P(X_{t+1}|x_t, e_{1:t})P(x_t|e_{1:t})$$
$$= \sum_{x_t} P(X_{t+1}|x_t)P(x_t|e_{1:t})$$

- Or compactly:

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

Basic idea: beliefs get "pushed" through the transitions

With the "B" notation, we have to be careful about what time step t the belief is about, and what evidence it includes

Observation

- Assume we have current belief P(X | previous evidence):

$B'(X_{t+1}) = P(X_{t+1}|e_{1:t})$

- Then, after evidence comes in:

$$P(X_{t+1}|e_{1:t+1}) = P(X_{t+1}, e_{t+1}|e_{1:t})/P(e_{t+1}|e_{1:t})$$
$$\propto_{X_{t+1}} P(X_{t+1}, e_{t+1}|e_{1:t})$$
$$= P(e_{t+1}|e_{1:t}, X_{t+1})P(X_{t+1}|e_{1:t})$$
$$= P(e_{t+1}|X_{t+1})P(X_{t+1}|e_{1:t})$$

- Basic idea: beliefs "reweighted" by likelihood of evidence
- Unlike passage of time, we have to renormalize

- Or, compactly:

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