HMMs

CPS 271

Ronald Parr

Overview

- Bayes nets are (mostly) atemporal
- Need a way to talk about a world that changes over time
- Necessary for planning
- Many important applications
  - Target tracking
  - Patient/factory monitoring
  - Speech recognition

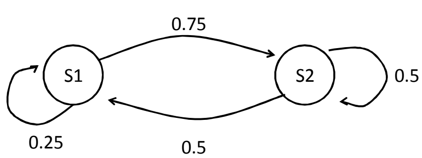
Back to Atomic Events

- We began talking about probabilities from the perspective of atomic events
- An atomic event is an assignment to every random variable in the domain
- For n random variables, there are 2^n possible atomic events
- State variables return later (briefly)

States

- When reasoning about time, we often call atomic events states
- States, like atomic events, form a mutually exclusive and jointly exhaustive partition of the space of possible events
- We can describe how a system behaves with a state-transition diagram

State Transition Diagram



P(S2|S1)=0.75
P(S1|S1)=0.25
P(S2|S2)=0.50
P(S1|S2)=0.50

Don't confuse states with state variables!
Don't confuse states with state variables!
Don't confuse states with state variables!

State Transition Diagrams

- Make a lot of assumptions
  - Transition probabilities don't change over time (*stationarity*)
  - The event space does not change over time
  - Probability distribution over next states depends only on the current state (*Markov assumption*)
  - Time moves in uniform, discrete increments

The Markov Assumption

- Let S_t be a random variable for the state at time t
- $P(S_t|S_{t-1},\ldots,S_0) = P(S_t|S_{t-1})$
- (Use subscripts for time; S0 is different from S_0)
- Markov is special kind of conditional independence
- Future is independent of past given current state

Markov Models

- A system with states that obey the Markov assumption is called a *Markov Model*
- A sequence of states resulting from such a model is called a *Markov Chain*
- The mathematical properties of Markov chains are studied heavily in mathematics, statistics, computer science, electrical engineering, etc.

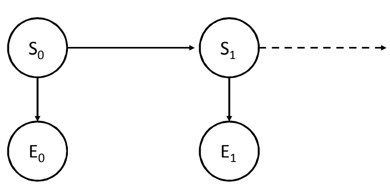
What's The Big Deal?

- A system that obeys the Markov property can be described succinctly with a transition matrix, where the i,jth entry of the matrix is P(Sj|Si)
- The Markov property ensures that we can maintain this succinct description over a potentially infinite time sequence
- Properties of the system can be analyzed in terms of properties of the transition matrix
  - Steady-state probabilities
  - Convergence rate, etc.

Observations

- Introduce E_t for the observation at time t
- Observations are like evidence
- Define the probability distribution over observations as function of current state: P(E|S)
- Assume observations are conditionally independent of other variables given current state
- Assume observation probabilities are stationary

A Graphical Model



Note: These are random variables, not states!

Applications

- Monitoring/Filtering
  - S is the current status of the patient/factory
  - E is the current measurement
- Prediction
  - S is the current/future position of an object
  - E are our past observations
  - Project S into the future

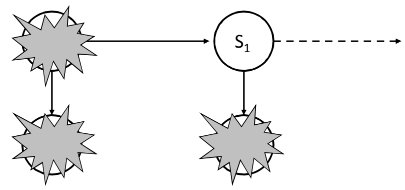
Applications

- Smoothing/hindsight
  - Update view of the past based upon future
  - Diagnosis: Factory exploded at time t=20, what happened at t=5 to cause this?
- Most likely explanation
  - What is the most likely sequence of events (from start to finish) to explain what we have seen?

Monitoring/Prediction

We want: $P(S_t | e_t \ldots e_0)$

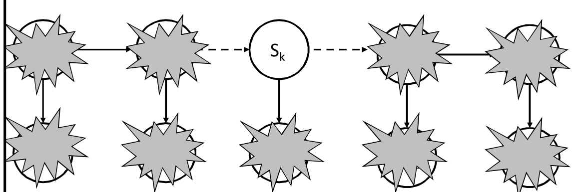
By variable elimination:



Smoothing/Hindsight

We want: $P(S_k | e_t \ldots e_0)$, $0<k<t$

By variable elimination:



Viterbi Path

From definition of Bayes net (or HMM):

$$P(S_0 E_0 \ldots S_t E_t) = P(S_0) P(E_0 | S_0) \prod_{i=1}^{t} P(S_i | S_{i-1}) P(E_i | S_i)$$

Suppose we want max probability sequence of states:

$$\begin{aligned} \max_{S_0 \ldots S_t} P(S_0 E_0 \ldots S_t E_t) &= \max_{S_0 \ldots S_t} P(S_0) P(E_0 | S_0) \prod_{i=1}^{t} P(S_i | S_{i-1}) P(E_i | S_i) \\ &= \max_{S_1 \ldots S_t} \prod_{i=1}^{t} P(S_i | S_{i-1}) P(E_i | S_i) \max_{S_0} P(S_1 | S_0) P(S_0) P(E_0 | S_0) \\ &= \max_{S_2 \ldots S_t} \prod_{i=2}^{t} P(S_i | S_{i-1}) P(E_i | S_i) \max_{S_0} P(S_2 | S_1) P(S_1 | E_1) \max_{S_0} P(S_1 | S_0) P(S_0) P(E_0 | S_0) \end{aligned}$$

Keep distributing max over product!

Algebraic View: Our Main Tool

$$P(A \wedge B) = P(B \wedge A)$$

$$P(A | B) P(B) = P(B | A) P(A)$$

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

Extending Bayes Rule

$$P(A | BC) = \frac{P(B | AC) P(A | C)}{P(B | C)}$$

How to think about this: The C is like “extra” evidence.
This forces us into one corner of the event space.
Given that we are in this corner, everything behaves the same.

Monitoring

We want: $P(S_t | e_t \ldots e_0)$

$$P(S_t | e_t \ldots e_0) = \frac{P(e_t | S_t, e_{t-1} \ldots e_0) P(S_t | e_{t-1} \ldots e_0)}{P(e_t | e_{t-1} \ldots e_0)}$$

$$= \alpha P(e_t | S_t e_{t-1} \ldots e_0) P(S_t | e_{t-1} \ldots e_0)$$

$$= \alpha P(e_t | S_t) P(S_t | e_{t-1} \ldots e_0)$$

$$= \alpha P(e_t | S_t) \sum_{S_{t-1}} P(S_t | S_{t-1}) P(S_{t-1} | e_{t-1} \ldots e_0)$$

Recursive

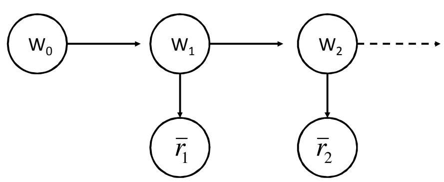
Example

- W = student is working
- R = student has produced results
- adviser observed whether student has produced results
- Must infer whether student is working given observations

$$P(W_{t+1} | W_t) = 0.8$$
$$P(W_{t+1} | \overline{W_t}) = 0.3$$
$$P(R | W) = 0.6$$
$$P(R | \overline{W}) = 0.2$$

Problem

Assume student starts school in a productive (working) state.
adviser has observed two consecutive months without results.
What is probability that student was working in the second month?



Let's Do The Math

$P(W_{t+1} | W_t) = 0.8$
$P(W_{t+1} | \overline{W_t}) = 0.3$
$P(R | W) = 0.6$
$P(R | \overline{W}) = 0.2$

$$P(W_2 | \bar{r}_2 \bar{r}_1) = \alpha_1 P(\bar{r}_2 | W_2) \sum_{W_1} P(W_2 | W_1) P(W_1 | \bar{r}_1)$$

$$P(W_1 | \bar{r}_1) = \alpha_2 P(\bar{r}_1 | W_1) \sum_{W_0} P(W_1 | W_0) P(W_0)$$

$$P(w_1 | \bar{r}_1) = \alpha_2 0.4(0.8 * 1.0 + 0.3 * 0.0) = \alpha_2 0.32$$

$$P(\overline{w}_1 | \bar{r}_1) = \alpha_2 0.8(0.2 * 1.0 + 0.7 * 0.0) = \alpha_2 0.16$$

$$P(w_1 | \bar{r}_1) = 0.67, P(\overline{w}_1 | \bar{r}_1) = 0.33$$

More Math

$P(W_{t+1} | W_t) = 0.8$
$P(W_{t+1} | \overline{W_t}) = 0.3$
$P(R | W) = 0.6$
$P(R | \overline{W}) = 0.2$
$P(w_1 | \bar{r}_1) = 0.67$
$P(\overline{w}_1 | \bar{r}_1) = 0.33$

$$P(W_2 | \bar{r}_2 \bar{r}_1) = \alpha_1 P(\bar{r}_2 | W_2) \sum_{W_1} P(W_2 | W_1) P(W_1 | \bar{r}_1)$$

$$P(w_2 | \bar{r}_2 \bar{r}_1) = \alpha_1 0.4(0.8 * 0.67 + 0.3 * 0.33) = \alpha_1 0.25$$

$$P(\overline{w}_2 | \bar{r}_2 \bar{r}_1) = \alpha_1 0.8(0.2 * 0.67 + 0.7 * 0.33) = \alpha_1 0.292$$

$$P(w_2 | \bar{r}_2 \bar{r}_1) = 0.46, P(\overline{w}_2 | \bar{r}_2 \bar{r}_1) = 0.54$$

Hindsight

$$P(S_k | e_t \ldots e_0) = \alpha P(e_t \ldots e_{k+1} | S_k, e_k \ldots e_0) P(S_k | e_k \ldots e_0)$$

$$= \alpha P(e_t \ldots e_{k+1} | S_k) \boxed{P(S_k | e_k \ldots e_0)}$$ Monitoring!

$$P(e_t \ldots e_{k+1} | S_k) = \sum_{S_{k+1}} P(e_t \ldots e_{k+1} | S_k S_{k+1}) P(S_{k+1} | S_k)$$

$$= \sum_{S_{k+1}} P(e_t \ldots e_{k+1} | S_{k+1}) P(S_{k+1} | S_k)$$

$$= \sum_{S_{k+1}} P(e_{k+1} | S_{k+1}) P(e_t \ldots e_{k+2} | S_{k+1}) P(S_{k+1} | S_k)$$

Recursive

Hindsight Summary

- Forward: Compute k state distribution given
  - Forward distribution up to k
  - Observations up to k
  - Equivalent to monitoring up to k
  - Equivalent to eliminating variables <k
- Backward: Compute conditional evidence distribution after k
  - Work backward from t to k
  - Equivalent to to eliminating variables >k
- Smoothed state distribution is proportional to product of forward and backward components

Problem II

Can we revise our estimate of the probability that the student worked at step 1?

We initially thought:

$$P(w_1 \mid \overline{r}_1) = 0.67, P(\overline{w}_1 \mid \overline{r}_1) = 0.33$$

Since the student didn't have results at time 2, is it now less likely that he was working at time 1?

Let's Do More Math

$P(W_{t+1} \mid W_t) = 0.8$
$P(W_{t+1} \mid \overline{W}_t) = 0.3$
$P(R \mid W) = 0.6$
$P(R \mid \overline{W}) = 0.2$
$P(w_1 \mid \overline{r}_1) = 0.67$
$P(\overline{w}_1 \mid \overline{r}_1) = 0.33$

$$P(W_1 \mid \overline{r}_2\overline{r}_1) = \alpha P(W_1 \mid \overline{r}_1)P(\overline{r}_2 \mid W_1)$$
$$P(\overline{r}_2 \mid w_1) = \sum_{W_2} P(\overline{r}_2 \mid W_2)P(W_2 \mid w_1)$$
$$P(\overline{r}_2 \mid w_1) = (0.4*0.8+0.8*0.2) = 0.48$$
$$P(\overline{r}_2 \mid \overline{w}_1) = (0.4*0.3+0.8*0.7) = 0.68$$
$$P(w_1 \mid \overline{r}_1) = \alpha 0.33*0.48 = 0.1584$$
$$P(\overline{w}_1 \mid \overline{r}_2\overline{r}_1) = \alpha 0.67*0.68 = 0.4556$$
$$P(w_1 \mid \overline{r}_2\overline{r}_1) = 0.258, P(\overline{w}_1 \mid \overline{r}_1) = 0.742$$

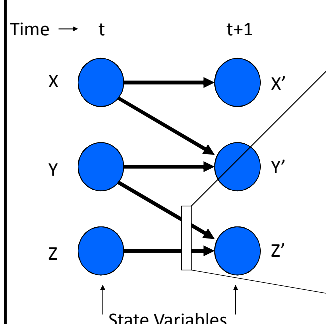
Checkpoint

- Done: Forward Monitoring and Backward Smoothing
- Monitoring is recursive from the past to the present
- Backward smoothing requires two recursive passes
- Called the forward-backward algorithm
  - Independently discovered many times throughout history
  - Was classified for many years by US Govt.
- Equivalent to doing variable elimination!

What's Left?

- We have seen that filtering and smoothing can be done efficiently, so what's the catch?
- We're still working at the level of atomic events
- There are too many atomic events!
- We need a generalization of Bayes nets to let us think about the world at the level of state variables and not states

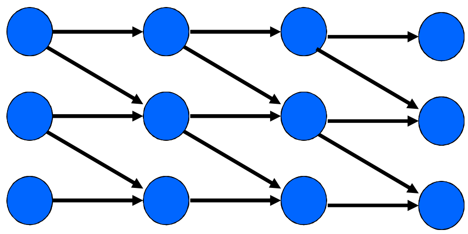
Dynamic Bayes Nets



CPT

	$P(Z')$	$P(\overline{Z}')$
$Y_t Z_t$	0.25	0.75
$Y_t \overline{Z}_t$	0.5	0.5
$\overline{Y}_t Z_t$	0.6	0.4
$\overline{Y}_t \overline{Z}_t$	0.3	0.7

Working With DBNs



Can we do variable elimination for DBNs?

Harsh Reality

- While BN inference in the static case was a very nice story, there are essentially no tractable, exact algorithms for DBNs
- Active research area:
  - Approximate inference algorithms
    - Variational methods
    - Assumed density filtering (ADF)
  - Sampling methods
    - Sequential Importance sampling
    - Sequential Importance sampling with Resampling (SISR, particle filter, condensation, etc.)

Continuous Variables

- How do we represent a probability distribution over a continuous variable?
  - Probability density function
  - Summations become integrals
- Very messy except for some special cases:
  - Distribution over variable X at time t+1 is a multivariate normal with a mean that is a linear function of the variables at the previous time step
  - This is a linear-Gaussian model

Inference in Linear Gaussian Models

- Filtering and smoothing integrals have closed form solution
- Elegant solution known as the Kalman filter
  - Used for tracking projectiles (radar)
  - State is modeled as a set of linear equations
    - S=vt
    - V=at
  - What about pilot controls?

Inference in Hybrid Networks

- Hybrid networks combine discrete and continuous variables
- Usually (but not always) a combination of discrete and Gaussian variables
- Active area of research:
  - Inference recently proven to be NP hard even for simple chains (Lerner & Parr 2001)
  - Many new approximate inference algorithms developed each year

Related Topics

- Continuous time
  - Need to model system using differential equations
- Non-stationarity
  - What if the model changes over time?
  - This touches on learning
- What about controlling the system w/actions?
  - Markov decision processes

HMM Conclusion

- Elegant algorithms for temporal reasoning over discrete atomic events, Gaussian continuous variables (many practical systems are such)
- Exact Bayes net methods don’t generalize well to state variable representation in the the temporal case: little hope for exponential savings
- Approximate inference for large systems is an active area of research