Markov Chains and HMMs

Ronald Parr

Overview

- Prediction not always 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

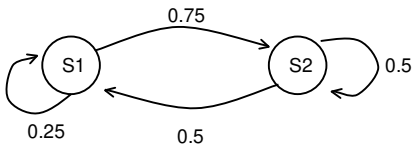
Atomic Events

- 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/features used for generalization (more advanced)

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.

Simple Properties of Markov Chains

- For distribution over states: x (row vector)
- Distribution at next time step is xP
- Distribution after t steps is xP^t
- Steady state distribution is xP = x
- Steady state existence (informal):
  - aperiodic
  - irreducible

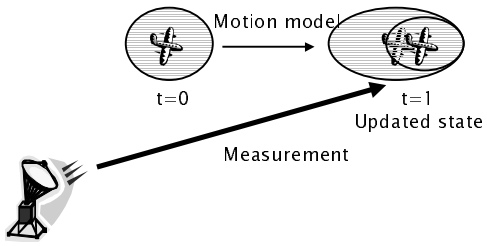
Hidden Markov Models

- Introduce E_t for the observation at time t
- Observations are like evidence/features
- 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

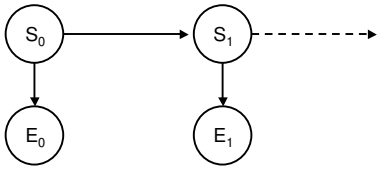
Application: Tracking Aircraft

- Example: Radar tracking
- Hidden state variable(s) = aircraft state
- Dynamic model = Markov chain for aircraft dynamics
- Noisy observations
- Problem: Infer hidden variables

Tracking Example



A Graphical Model



Note: These are random variables, not states!

HMM Uses

- 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?

Our Main Tool

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

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

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

Extending Bayes Rule

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

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

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

Recursive

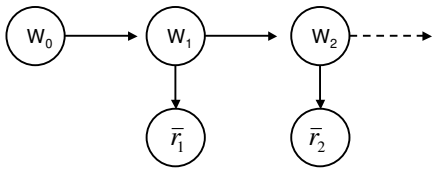
Example

- W = Student is working
- R = Student has produced results
- Advisor observes whether student has produced results
- Must infer whether student is working given observations

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

Problem

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



Let's Do The Math

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

$$P(W_2 \mid \bar{r}_2 \bar{r}_1) = \alpha_1 P(\bar{r}_2 \mid W_2) \sum_{W_1} P(W_2 \mid W_1) P(W_1 \mid \bar{r}_1)$$
$$P(W_1 \mid \bar{r}_1) = \alpha_2 P(\bar{r}_1 \mid W_1) \sum_{W_0} P(W_1 \mid W_0) P(W_0)$$
$$P(w_1 \mid \bar{r}_1) = \alpha_2 0.4(0.8*1.0 + 0.3*0.0) = \alpha_2 0.32$$
$$P(\overline{w_1} \mid \bar{r}_1) = \alpha_2 0.8(0.2*1.0 + 0.7*0.0) = \alpha_2 0.16$$
$$P(w_1 \mid \bar{r}_1) = 0.67, P(\overline{w_1} \mid \bar{r}_1) = 0.33$$

More Math

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

$$P(W_2 \mid \bar{r}_2 \bar{r}_1) = \alpha_1 P(\bar{r}_2 \mid W_2) \sum_{W_1} P(W_2 \mid W_1) P(W_1 \mid \bar{r}_1)$$
$$P(w_2 \mid \bar{r}_2 \bar{r}_1) = \alpha_1 0.4(0.8*0.67 + 0.3*0.33) = \alpha_1 0.25$$
$$P(\overline{w_2} \mid \bar{r}_2 \bar{r}_1) = \alpha_1 0.8(0.2*0.67 + 0.7*0.33) = \alpha_1 0.292$$
$$P(w_2 \mid \bar{r}_2 \bar{r}_1) = 0.46, P(\overline{w_2} \mid \bar{r}_2 \bar{r}_1) = 0.54$$

Hindsight

$$P(S_k \mid e_t \ldots e_0) = \alpha P(S_k \mid e_k \ldots e_0) P(e_t \ldots e_{k+1} \mid S_k, e_k \ldots e_0)$$
$$= \alpha P(S_k \mid e_k \ldots e_0) P(e_t \ldots e_{k+1} \mid S_k)$$
$$P(e_t \ldots e_{k+1} \mid S_k) = \sum_{S_{k+1}} P(e_t \ldots e_{k+1} \mid S_k S_{k+1}) P(S_{k+1} \mid S_k)$$
$$= \sum_{S_{k+1}} P(e_t \ldots e_{k+1} \mid S_{k+1}) P(S_{k+1} \mid S_k)$$
$$= \sum_{S_{k+1}} P(e_{k+1} \mid S_{k+1}) P(e_t \ldots e_{k+2} \mid S_{k+1}) P(S_{k+1} \mid S_k)$$

Recursive

Problem II

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

We initially thought:

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

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

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

$$P(W_1 \mid \bar{r}_2 \bar{r}_1) = \alpha P(W_1 \mid \bar{r}_1) P(\bar{r}_2 \mid W_1)$$
$$P(\bar{r}_2 \mid w_1) = \sum_{W_2} P(\bar{r}_2 \mid W_2) P(W_2 \mid w_1)$$
$$P(\bar{r}_2 \mid w_1) = (0.4 * 0.8 + 0.8 * 0.2) = 0.48$$
$$P(\bar{r}_2 \mid \overline{w}_1) = (0.4 * 0.3 + 0.8 * 0.7) = 0.68$$
$$P(\overline{w}_1 \mid \bar{r}_2 \bar{r}_1) = \alpha 0.33 * 0.68 = 0.4556\alpha = 0.5862$$
$$P(w_1 \mid \bar{r}_2 \bar{r}_1) = \alpha 0.67 * 0.48 = 0.3216\alpha = 0.4138$$

What Happened?

- After one observation, we initially think it is somewhat less likely that the student is working. However, not all working students have results all of the time.
- After two observations, we conclude that the student was much less likely to have been working in the first time step.
- Moral: Never go two meetings without having some results for your advisor.

Checkpoint

- Done: Forward Monitor 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.

Most Likely Path

- A similar problem to filtering/hindsight problem is one of finding the most likely path through the state space
- Useful for reconstructing events:
  - What caused the factory to explode?
  - Why is the most likely signal given our measurements?
  - What did a person say?

Viterbi Paths

- We can use the Markov property to decompose the problem
- Main observation: The most likely path from $S_0 \ldots S_k \ldots S_t$ can be decomposed as the most likely path from $S_0 \ldots S_k$ combined with the most like path from $S_k \ldots S_t$.
- Idea: Very similar to forward/backward, but we replace sums with maxes in the filtering equations (see Rabiner)

Learning

- Fully observable case
- Inputs: Time stamped record of states, observations
- Output: Transition and emission probs.
- Compute conditional probabilities by maximum likelihood

Learning II

- Partial observability
  - Observe emissions, but not underlying states
- Solution: EM
  - Start with initial guess about probs
  - Use forward/backward to compute distribution over states
  - Compute new transition probs based on state occupancy probs.

Estimating Transition Probabilities

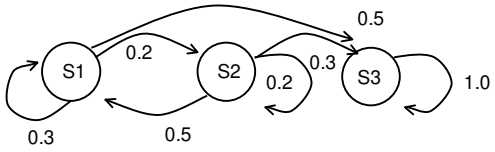
- Need:

$$P(Si_k, Sj_{k+1} \mid e_0 \ldots e_t) = \alpha P(Si_k \mid e_0 \ldots e_k) P(e_{k+2} \ldots e_t \mid Sj_{k+1}) P(Sj_{k+1} \mid Si_k) P(e_{k+1} \mid Sj_{k+1})$$

Forward Backward

- (Same as forward backward w/o marginalizing out one time step)
- Estimate new transition probabilities by using the observed transition probs.

Example



P(A|S1) = 0.6, P(B|S1) = 0.4,P(x|S1 = 0)
P(A|S2) = 0.2, P(B|S2) = 0.8,P(x|S2 = 0)
P(x|S3) = 1.0

Inputs:
Known start in S1
ABxxx, ABABx

EM Iteration 1

- ABxxx state probs:

1.00	0.51	0.00	0.00	0.00
0.00	0.48	0.00	0.00	0.00
0.00	0.00	1.00	1.00	1.00

- ABABx state probs

1.00	0.30	0.79	0.54	0.00
0.00	0.70	0.21	0.46	0.00
0.00	0.00	0.00	0.00	1.00

- New Transition Probs

0.35	0.40	0.25
0.38	0.11	0.51
0	0	1.00

After 3rd EM Iteration

- Transition matrix:

0.00	1.00	0.00
0.33	0.00	0.67
0.00	0.00	1.00

Observation probs:

1.00	0.00	0.00
0.00	1.00	0.00
0.00	0.00	1.00

Effects of Initial Conditions

- Suppose we assume to start in S2:
  - Converges to same model with S2 and S1 roles switched
  - Takes 5 iterations instead of 3
- In general
  - May get stuck in local optima
  - May be sensitive to initial conditions
  - Problem if number of states not known?

HMM summary

- Markov property allows use of dynamic programming to compute:
  - Forward probabilities (tracking)
  - Backward probabilities (prob. future observations)
  - Occupancy probs, transition probs
- Use EM to iteratively update model
- Same old issues about regularization, local minima, initial parameters, etc.