Markov Models

Andrea Danyluk
March 17, 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 after the break
- Start thinking about final projects

Today's Lecture

- Finish up a bit of "intro to probability"
- Markov Models

Inference in Ghostbusters

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
- Sensors noisy, but we know P(Color|Distance)

Ghostbusters

- Say we have two distributions
  - P(G): say it's uniform
  - Sensor reading model P(R|G)
- Say we get a reading at (1,1)
- Can calculate the posterior distribution P(G|r) over all locations given the reading at (1,1)

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

- Sensor readings
  - On the ghost (1 location): red
  - 1 or 2 away (5 locations): orange
  - 3 or 4 away (3 locations): yellow
- Sensors noisy

P(red | 0)	P(orange | 0)	P(yellow | 0)
0.7	0.2	0.1

P(red | 1 or 2)	P(orange | 1 or 2)	P(yellow | 1 or 2)
0.15	0.7	0.15

P(red | 3 or 4)	P(orange | 3 or 4)	P(yellow | 3 or 4)
0.1	0.2	0.7

[Adapted from CS 188 Berkeley]

P(g|yellow)
P(0 away|yellow)
P(1-2 away|yellow)
P(3-4 away|yellow)
P(yellow)

Ghostbusters

- Say we have two distributions
  - P(G): say it's uniform
  - Sensor reading model P(R|G), where R=reading at (1,1)
- Can calculate the posterior distribution P(G|r) over ghost locations given a reading at (1,1)

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

0.05	0.24	0.24
0.05	0.05	0.24
0.03	0.05	0.05

Intractability of Probabilistic Inference

- Size of full joint probability distribution over n (Boolean) random variables?
  - $O(2^n)$

- Say we add a new random variable to the Ghostbusters problem: Is the number of students attending AI today > 15?
- 3 random variables:
  - Ghost location
  - Sensor reading
  - Attendance > 15?

- Say we add a new random variable to the Ghostbusters problem: Is the number of students attending AI today > 15?
-

But what does attendance in AI have to do with Ghostbusters?

Probabilistic Independence

- It seems reasonable to assert that the number of students attending AI on any given day is unrelated to ghosts or sensor readings.
- If P(X|Y)=P(X), we say X is independent of Y: $X \perp\!\!\!\perp Y$
  - Similarly, Y is independent of X.
  - P(Y|X) = P(Y), P(X, Y) = P(X)P(Y)
- This means the joint distribution factors into a product of two simpler distributions.

- Say we add a random variable to a "test and disease" problem domain: Does the patient have a rash? [And say that when a person has the disease, they tend to get a rash.]
- 3 Boolean variables:
  - T: Test positive or negative
  - D: Disease positive or negative
  - R: Rash positive or negative
- $2^3 = 8$ entries in the full joint probability distribution

Conditional Independence

- This time we can't reasonably assert that R is independent of T or D.
- But 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(D)
  - P(R, T|D) = P(R|D)P(T|D)
- We say T and R are conditionally independent given D.

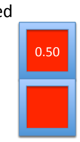
Conditional Independence: Notation

X and Y are conditionally independent given Z

$$X \perp\!\!\!\perp Y | Z$$

Model for Ghostbusters

- Reminder: ghost is hidden, sensors are noisy
- T: Top sensor is red
  B: Bottom sensor is red
  G: Ghost is in the top
- Queries:
  P(g) = ?
  P(g | t) = ?
  P(g | t, ¬b) = ?
- What happens to the joint distribution when the game gets bigger than two squares?



Joint Distribution

T	B	G	P(T,B,G)
t	b	g	0.16
t	b	¬g	0.16
t	¬b	g	0.24
t	¬b	¬g	0.04
¬t	b	g	0.04
¬t	b	¬g	0.24
¬t	¬b	g	0.06
¬t	¬b	¬g	0.06

[CS 188 Berkeley]

Joint distribution too large/complex!

Model for Ghostbusters cont'd

- T: Top sensor is red
  B: Bottom sensor is red
  G: Ghost is in the top
- Each sensor depends only on where the ghost is
- Sensors are conditionally independent given the ghost position
- Givens:
  P(g) = 0.5
  P(t | g) = 0.8
  P(t | ¬g) = 0.4
  P(b | g) = 0.4
  P(b | ¬g) = 0.8

Joint Distribution

T	B	G	P(T,B,G)
t	b	g	0.16
t	b	¬g	0.16
t	¬b	g	0.24
t	¬b	¬g	0.04
¬t	b	g	0.04
¬t	b	¬g	0.24
¬t	¬b	g	0.06
¬t	¬b	¬g	0.06

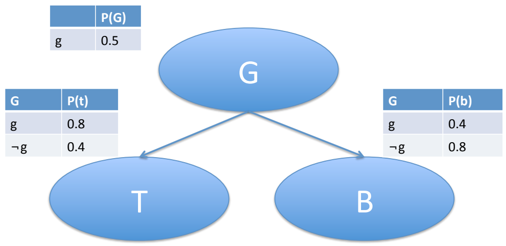
[CS 188 Berkeley]

P(T,B,G) = P(T|G)P(B|G)P(G)

Bayesian Network

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

Ghostbusters Example



	P(G)
g	0.5

G	P(t)
g	0.8
¬g	0.4

G	P(b)
g	0.4
¬g	0.8

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

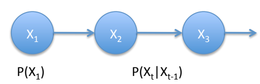
Reasoning over Time

- 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

[CS 188 Berkeley]

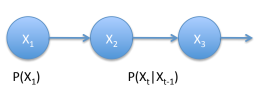
Markov Models

- 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

Joint Distribution of a Markov Model



- Joint distribution:
  $P(x_1,x_2,..x_n)=P(x_1)P(x_2|x_1)\ P(x_3|x_2)...P(x_n|x_{n-1})$
- But can we really call this a joint distribution?
  $P(x_1,x_2,..x_n)=P(x_n|x_1...x_{n-1})P(x_{n-1}|x_1...x_{n-2})...P(x_2|x_1)P(x_1)$

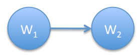
Conditional Independence

- Each time step only depends directly on the previous
  - First order Markov property
  - Past and future independent given the present
- Note that the chain is just a (growing) Bayesian net

[CS 188 Berkeley]

Example Markov Chain: Weather

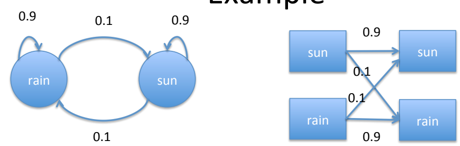
- Weather: W = {rain, sun}



W_{t-1}	W_t	$P(W_t \mid W_{t-1})$
Rain	Rain	0.9
Rain	Sun	0.1
Sun	Rain	0.1
Sun	Sun	0.9

[CS 188 Berkeley]

Example



W_{t-1}	W_t	$P(W_t \mid W_{t-1})$
Rain	Rain	0.9
Rain	Sun	0.1
Sun	Rain	0.1
Sun	Sun	0.9

[CS 188 Berkeley]

Example

- Initial distribution: 1.0 rain
- What is the probability distribution after 1 step?

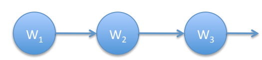
W_1	$P(W_1)$
Rain	1.0
Sun	0.0

W_{t-1}	W_t	$P(W_t \mid W_{t-1})$
Rain	Rain	0.9
Rain	Sun	0.1
Sun	Rain	0.1
Sun	Sun	0.9

- $P(W_2=Sun) = P(W_2=Sun \mid W_1=Sun)P(W_1=Sun) + P(W_2=Sun \mid W_1=Rain)P(W1=Rain) = 0.1$

[CS 188 Berkeley]

Example



W_1	$P(W_1)$
Rain	1.0
Sun	0.0

W_{t-1}	W_t	$P(W_t \mid W_{t-1})$
Rain	Rain	0.9
Rain	Sun	0.1
Sun	Rain	0.1
Sun	Sun	0.9

More generally, what's P(W) on some day t?

- $P(w_t) = \Sigma_{wt-1} P(w_t \mid w_{t-1})P(w_{t-1})$
- $P(w_1)$ known

Forward simulation

[CS 188 Berkeley]

Example cont'd

- From initial observation of **sun**:

Sun	1.0	0.9	0.82	...	0.5
Rain	0.0	0.1	0.18	...	0.5

If we simulate the chain long enough, uncertainty accumulates

[CS 188 Berkeley]