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 \( P(\text{Color} | \text{Distance}) \)

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 \((1,1)\)

Sensor readings

— On the ghost \((1 \text{ location})\): red
— 1 or 2 away \((5 \text{ locations})\): orange
— 3 or 4 away \((3 \text{ locations})\): yellow

Sensors noisy

| \( P(\text{red} | 0) \) | \( P(\text{orange} | 1 \text{ or } 2) \) | \( P(\text{yellow} | 3 \text{ or } 4) \) |
|---|---|---|
| 0.7 | 0.2 | 0.1 |
| \( P(\text{red} | 1 \text{ or } 2) \) | \( P(\text{orange} | 1 \text{ or } 2) \) | \( P(\text{yellow} | 1 \text{ or } 2) \) |
| 0.15 | 0.7 | 0.15 |
| \( P(\text{red} | 3 \text{ or } 4) \) | \( P(\text{orange} | 3 \text{ or } 4) \) | \( P(\text{yellow} | 3 \text{ or } 4) \) |
| 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)

<table>
<thead>
<tr>
<th></th>
<th>0.11</th>
<th>0.11</th>
<th>0.11</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Intractability of Probabilistic Inference

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

Probabilistic Independence

• 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?

• 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 ⊥ 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

\[ 2^3 = 8 \text{ entries in the full joint probability distribution} \]

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.

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
- Each sensor depends only on where the ghost is
- Sensors are conditionally independent given the ghost
- Givens:
  - \( P(g) = 0.5 \)
  - \( P(t | g) = 0.8 \)
  - \( P(t | \neg g) = 0.4 \)
  - \( P(b | g) = 0.4 \)
  - \( P(b | \neg g) = 0.8 \)

\[ \begin{array}{ccc}
T & B & G \\
\hline
 t & b & g & 0.16 \\
 t & b & \neg g & 0.16 \\
 t & \neg b & g & 0.24 \\
 t & \neg b & \neg g & 0.04 \\
 \neg t & b & g & 0.04 \\
 \neg t & b & \neg g & 0.24 \\
 \neg t & \neg b & g & 0.06 \\
 \neg t & \neg b & \neg g & 0.06 \\
 \end{array} \]

\[ P(T,B,G) = P(T|G)P(B|G)P(G) \]

Conditional Independence: Notation

- \( X \) and \( Y \) are conditionally independent given \( Z \)

\[ X \perp Y | Z \]

Model for Ghostbusters cont’d

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

Bayesian Network
**Ghostbusters Example**

[Diagram showing a network with nodes labeled G, T, and B, and probabilities associated with each node and edge]

**Space and Time**

- Bayesian networks are generally much more compact than the full joint probability distribution
  - Joint distribution: $O(2^n)$
  - Bayes net: $O(n^2k)$, 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

[CS 188 Berkeley]

**Joint Distribution of a Markov Model**

- Joint distribution:
  \[ P(x_1, x_2, \ldots, x_n) = P(x_n|x_{n-1})P(x_{n-1}|x_{n-2})\ldots P(x_1|x_0)P(x_0) \]

- But can we really call this a joint distribution?
  \[ P(x_1, x_2, \ldots, x_n) = P(x_n|x_1, \ldots, x_{n-1})P(x_{n-1}|x_1, \ldots, x_{n-2})\ldots P(x_1|x_0)P(x_0) \]

**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: \( W = \{ \text{rain}, \text{sun} \} \)

\[
\begin{array}{c|cc}
W_t & W_{t-1} & \text{P}(W_t | W_{t-1}) \\
\hline
\text{Rain} & \text{Rain} & 0.9 \\
\text{Rain} & \text{Sun} & 0.1 \\
\text{Sun} & \text{Rain} & 0.1 \\
\text{Sun} & \text{Sun} & 0.9 \\
\end{array}
\]

Example

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

\[
\begin{array}{c|c|c}
W_t & \text{P}(W_t) \\
\hline
\text{Rain} & 1.0 \\
\text{Sun} & 0.0 \\
\end{array}
\]

\[
\begin{array}{c|cc}
W_t & W_{t-1} & \text{P}(W_t | W_{t-1}) \\
\hline
\text{Rain} & \text{Rain} & 0.9 \\
\text{Rain} & \text{Sun} & 0.1 \\
\text{Sun} & \text{Rain} & 0.1 \\
\text{Sun} & \text{Sun} & 0.9 \\
\end{array}
\]

\[
P(W_2 = \text{Sun}) = P(W_1 = \text{Sun} | W_2 = \text{Sun}) P(W_1 = \text{Sun}) + P(W_1 = \text{Rain} | W_2 = \text{Rain}) P(W_1 = \text{Rain}) = 0.1
\]

Example cont’d

- From initial observation of sun:

\[
\begin{array}{c|c|c|c|c|c|c}
& \text{Sun} & \text{Rain} & \multicolumn{4}{c}{\text{other days}} \\
\hline
\text{Sun} & 1.0 & 0.9 & 0.82 & \ldots & 0.5 \\
\text{Rain} & 0.0 & 0.1 & 0.18 & \ldots & 0.5 \\
\end{array}
\]

If we simulate the chain long enough, uncertainty accumulates