Markov Models and Hidden Markov Models

Robert Platt
Northeastern University

Some images and slides are used from:
1. CS188 UC Berkeley
2. RN, AIMA
Markov Models

We have already seen that an MDP provides a useful framework for modeling stochastic control problems.

**Markov Models**: model any kind of temporally dynamic system.
Probability again: Independence

Two random variables, $x$ and $y$, are independent when:

$$\forall (x, y), \ P(x, y) = P(x)P(y) \quad \iff \quad x \perp y$$

The outcomes of two different coin flips are usually independent of each other.

Image: Berkeley CS188 course notes (downloaded Summer 2015)
Probability again: Independence

If: \[ P(x, y) = P(x)P(y) \]

Then: \[ P(x) = P(x|y) \]
\[ P(y) = P(y|x) \]

Why?
Are $T$ and $W$ independent?

$$P_1(T, W)$$

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Are T and W independent?

\[ P_1(T, W) \]

\[
\begin{array}{ccc}
T & W & P \\
\text{hot} & \text{sun} & 0.4 \\
\text{hot} & \text{rain} & 0.1 \\
\text{cold} & \text{sun} & 0.2 \\
\text{cold} & \text{rain} & 0.3 \\
\end{array}
\]

\[
P(T)\]

\[
\begin{array}{cc}
T & P \\
\text{hot} & 0.5 \\
\text{cold} & 0.5 \\
\end{array}
\]

\[
P(W)\]

\[
\begin{array}{cc}
W & P \\
\text{sun} & 0.6 \\
\text{rain} & 0.4 \\
\end{array}
\]

Slide: Berkeley CS188 course notes (downloaded Summer 2015)
Are $T$ and $W$ independent?

$P_1(T, W)$

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.4</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.1</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.3</td>
</tr>
</tbody>
</table>

$P(T)$

<table>
<thead>
<tr>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>0.5</td>
</tr>
<tr>
<td>cold</td>
<td>0.5</td>
</tr>
</tbody>
</table>

$P_2(T, W) = P(T)P(W)$

<table>
<thead>
<tr>
<th>T</th>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>hot</td>
<td>sun</td>
<td>0.3</td>
</tr>
<tr>
<td>hot</td>
<td>rain</td>
<td>0.2</td>
</tr>
<tr>
<td>cold</td>
<td>sun</td>
<td>0.3</td>
</tr>
<tr>
<td>cold</td>
<td>rain</td>
<td>0.2</td>
</tr>
</tbody>
</table>

$P(W)$

<table>
<thead>
<tr>
<th>W</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>sun</td>
<td>0.6</td>
</tr>
<tr>
<td>rain</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Slide: Berkeley CS188 course notes (downloaded Summer 2015)
Conditional independence

Independence: \( \forall (x, y), P(x, y) = P(x)P(y) \)
\[
x \perp y
\]

Conditional independence: \( \forall (x, y, z), P(x, y|z) = P(x|z)P(y|z) \)
\[
x \perp y|z
\]

Equivalent statements of conditional independence:
\[
P(x|z) = P(x|z, y)
\]
\[
P(y|z) = P(y|z, x)
\]
Conditional independence: example

P(toothache, catch | cavity) = P(toothache | cavity) = P(catch | cavity)

or...

P(toothache | cavity) = P(toothache | cavity, catch)
P(catch | cavity) = P(catch | cavity, toothache)
Conditional independence: example

- What about this domain:
  - Traffic
  - Umbrella
  - Raining
Conditional independence: example

- What about this domain:
  - Fire
  - Smoke
  - Alarm
Markov Processes

State at time=1

State at time=2

transitions
Since this is a Markov process, we assume transitions are Markov:

Process model: \[ P(X_t|X_{t-1}) = P(X_t|X_{t-1}, \ldots, X_1) \]

Markov assumption: \[ X_t \perp X_{t-2} | X_{t-1} \]
How do we calculate: \( P(X_1, X_2, X_3, X_4) =? \)

\[
P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)P(X_4|X_3, X_2, X_1)
\]
How do we calculate:  \( P(X_1, X_2, X_3, X_4) = ? \)

\[
P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)P(X_4|X_3, X_2, X_1)
\]

\[
\underbrace{P(X_2, X_1)}
\]
Markov Processes

How do we calculate: \[ P(X_1, X_2, X_3, X_4) = ? \]

\[
P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)P(X_4|X_3, X_2, X_1)
\]

\[
P(X_3, X_2, X_1)
\]
How do we calculate: \( P(X_1, X_2, X_3, X_4) =? \)

\[
P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)P(X_4|X_3, X_2, X_1)
\]

Can we simplify this expression?
Markov Processes

How do we calculate: \( P(X_1, X_2, X_3, X_4) = ? \)

\[
P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)P(X_4|X_3, X_2, X_1)
\]

\[
P(X_3|X_2) \quad P(X_4|X_3)
\]
Markov Processes

How do we calculate: $P(X_1, X_2, X_3, X_4) =$?

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)P(X_4|X_3, X_2, X_1)$$

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$
How do we calculate: \( P(X_1, X_2, X_3, X_4) = ? \)

\[
P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)P(X_4|X_3, X_2, X_1)
\]

\[
P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)
\]

In general: \( P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=1}^{T-1} P(X_{t+1}|X_t) \)
Markov Processes

How do we calculate: \( P(X_1, X_2, X_3, X_4) = ? \)

\[
P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)
\]

In general: \( P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=1}^{T-1} P(X_{t+1}|X_t) \)
Markov Processes: example

- States: $X = \{\text{rain, sun}\}$
- Initial distribution: $1.0$ sun
- Process model: $P(X_t | X_{t-1})$:

| $X_{t-1}$ | $X_t$ | $P(X_t|X_{t-1})$ |
|-----------|-------|------------------|
| sun       | sun   | 0.9              |
| sun       | rain  | 0.1              |
| rain      | sun   | 0.3              |
| rain      | rain  | 0.7              |

Two new ways of representing the same CPT

Slide: Berkeley CS188 course notes (downloaded Summer 2015)
Simulating dynamics forward

Joint distribution: \( P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=1}^{T-1} P(X_{t+1} | X_t) \)

But, suppose we want to predict the state at time \( T \), given a prior distribution at time 1?

\[
P(X_2) = \sum_{X_1} P(X_1)P(X_2 | X_1)
\]

\[
P(X_3) = \sum_{X_2} P(X_2)P(X_3 | X_2)
\]

\[
\vdots
\]

\[
P(X_T) = \sum_{X_{T-1}} P(X_{T-1})P(X_T | X_{T-1})
\]
Markov Processes: example

- Initial distribution: 1.0 sun

- What is the probability distribution after one step?

\[ P(X_2 = \text{sun}) = P(X_2 = \text{sun}|X_1 = \text{sun})P(X_1 = \text{sun}) + P(X_2 = \text{sun}|X_1 = \text{rain})P(X_1 = \text{rain}) \]

\[ 0.9 \cdot 1.0 + 0.3 \cdot 0.0 = 0.9 \]
Simulating dynamics forward

- **From initial observation of sun**

<table>
<thead>
<tr>
<th>P(X₁)</th>
<th>P(X₂)</th>
<th>P(X₃)</th>
<th>P(X₄)</th>
<th>P(X_∞)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.9</td>
<td>0.84</td>
<td>0.804</td>
<td>0.75</td>
</tr>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.16</td>
<td>0.196</td>
<td>0.25</td>
</tr>
</tbody>
</table>

- **From initial observation of rain**

<table>
<thead>
<tr>
<th>P(X₁)</th>
<th>P(X₂)</th>
<th>P(X₃)</th>
<th>P(X₄)</th>
<th>P(X_∞)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.3</td>
<td>0.48</td>
<td>0.588</td>
<td>0.75</td>
</tr>
<tr>
<td>1.0</td>
<td>0.7</td>
<td>0.52</td>
<td>0.412</td>
<td>0.25</td>
</tr>
</tbody>
</table>

- **From yet another initial distribution P(X₁):**

<table>
<thead>
<tr>
<th>P(X₁)</th>
<th>P(X_∞)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>0.75</td>
</tr>
<tr>
<td>1 − p</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Slide: Berkeley CS188 course notes (downloaded Summer 2015)
Simulating dynamics forward

- From initial observation of sun:
  
  \[
  \begin{pmatrix}
  1.0 \\
  0.0
  \end{pmatrix}
  \begin{pmatrix}
  0.9 \\
  0.1
  \end{pmatrix}
  \begin{pmatrix}
  0.84 \\
  0.16
  \end{pmatrix}
  \begin{pmatrix}
  0.804 \\
  0.196
  \end{pmatrix}
  \begin{pmatrix}
  0.75 \\
  0.25
  \end{pmatrix}
  \]
  
  This is called the **stationary distribution**

- From initial observation of rain:
  
  \[
  \begin{pmatrix}
  0.0 \\
  1.0
  \end{pmatrix}
  \begin{pmatrix}
  0.88 \\
  0.12
  \end{pmatrix}
  \begin{pmatrix}
  0.75 \\
  0.25
  \end{pmatrix}
  \]

- From yet another initial distribution \( P(X_1) \):
  
  \[
  \begin{pmatrix}
  p \\
  1 - p
  \end{pmatrix}
  \]
  
  ...
Hidden Markov Models (HMMs)

Hidden Markov Models: markov models applied to estimation problems

- speech to text
- tracking in computer vision
- robot localization
Hidden Markov Models (HMMs)

State, $X_t$, is assumed to be unobserved.

However, you get to make one observation, $E_t$, on each timestep.
Hidden Markov Models (HMMs)

Sensor Markov Assumption: the current observation depends only on current state:

\[ P(E_t|X_t, X_{t-1}, \ldots, X_1) = P(E_t|X_t) \]

\[ E_t \perp X_{t-1}|X_t \]
An HMM is defined by:

- Initial distribution: \( P(X_1) \)
- Transitions: \( P(X_t \mid X_{t-1}) \)
- Emissions: \( P(E_t \mid X_t) \)
Real world HMM applications

- **Speech recognition HMMs:**
  - Observations are acoustic signals (continuous valued)
  - States are specific positions in specific words (so, tens of thousands)

- **Machine translation HMMs:**
  - Observations are words (tens of thousands)
  - States are translation options

- **Robot tracking:**
  - Observations are range readings (continuous)
  - States are positions on a map (continuous)
Filtering, or monitoring, is the task of tracking the distribution $B_t(X) = P_t(X_t | e_1, ..., e_t)$ (the belief state) over time.

We start with $B_1(X)$ in an initial setting, usually uniform.

As time passes, or we get observations, we update $B(X)$.

The Kalman filter was invented in the 60’s and first implemented as a method of trajectory estimation for the Apollo program.
HMM Filtering

Given a prior distribution, $P(X_1)$, and a series of observations, $E_1, \ldots, E_T$, calculate the posterior distribution: $P(X_t | E_1, \ldots, E_T)$

Two steps:

$$P(X_t | E_{1:t}) \quad \rightarrow \quad P(X_{t+1} | E_{1:t}) \quad \rightarrow \quad P(X_{t+1} | E_{1:t+1})$$

Process update  Observation update
HMM Filtering

Given a prior distribution, \( P(X_1) \), and a series of observations, \( E_1, \ldots, E_T \), calculate the posterior distribution: \( P(X_t|E_1, \ldots, E_T) \)

Two steps:

\[
\begin{align*}
B(X_t) & \quad B'(X_t) & \quad B(X_{t+1}) \\
P(X_t|E_{1:t}) & \quad P(X_{t+1}|E_{1:t}) & \quad P(X_{t+1}|E_{1:t+1})
\end{align*}
\]

Process update \quad Observation update
HMM Filtering

Given a prior distribution, $P(X_1)$, and a series of observations, $E_1, \ldots, E_T$, calculate the posterior distribution:

Two steps:

- Process update
  - $B(X_t)$
  - $P(X_t | E_{1:t})$
  - $B'(X_t)$
  - $P(X_{t+1} | E_{1:t})$

- Observation update
  - $B(X_{t+1})$
  - $P(X_{t+1} | E_{1:t+1})$

"Beliefs"
Process update

\[ B(X_t) \quad \rightarrow \quad B'(X_t) \]

\[ P(X_t|e_{1:t}) \quad \rightarrow \quad P(X_{t+1}|e_{1:t}) \]

\[
P(X_{t+1}|e_{1:t}) = \sum_{X_t} P(X_{t+1}|X_t, e_{1:t}) P(X_t|e_{1:t})
\]

\[
B'(X_{t+1}) = \sum_{X_t} P(X_{t+1}|X_t, e_{1:t}) B(X_t)
\]

This is just forward simulation of the Markov Model
Process update: example

- As time passes, uncertainty "accumulates"
  (Transition model: ghosts usually go clockwise)

\[ B'(X_{t+1}) = \sum_{X_t} P(X_{t+1} | X_t, e_{1:t}) B(X_t) \]

Slide: Berkeley CS188 course notes (downloaded Summer 2015)
Observation update

\[ B'(X_t) \quad \xrightarrow{\text{}} \quad B(X_{t+1}) \]

\[ P(X_{t+1}|e_{1:t}) \quad \xrightarrow{\text{}} \quad P(X_{t+1}|e_{1:t+1}) \]

\[ P(X_{t+1}|e_{1:t+1}) = \eta P(e_{t+1}|X_{t+1}) P(X_{t+1}|e_{1:t}) \]

\[ B(X_{t+1}) = \eta P(e_{t+1}|X_{t+1}) B'(X_{t+1}) \]

Where \( \eta = \frac{1}{P(e_{t+1})} \) is a normalization factor
As we get observations, beliefs get reweighted, uncertainty "decreases"

$$B(X_{t+1}) = \eta P(e_{t+1} | X_{t+1}) B'(X_{t+1})$$
Observation model: can read in which directions there is a wall, never more than 1 mistake
Process model: may not execute action with small prob.
Lighter grey: was possible to get the reading, but less likely b/c required 1 mistake
Robot localization example
Robot localization example

Slide: Berkeley CS188 course notes (downloaded Summer 2015)
Robot localization example

![Diagram of robot localization example]

Prob

0

1

Slide: Berkeley CS188 course notes (downloaded Summer 2015)
Robot localization example
Weather HMM example

\[ P(R_{t+1} | R_t) \]

\begin{tabular}{|c|c|c|}
\hline
$R_t$ & $R_{t+1}$ & $P(R_{t+1} | R_t)$ \\
\hline
$+r$ & $+r$ & 0.7 \\
$+r$ & $-r$ & 0.3 \\
$-r$ & $+r$ & 0.3 \\
$-r$ & $-r$ & 0.7 \\
\hline
\end{tabular}

\[ P(U_t | R_t) \]

\begin{tabular}{|c|c|c|}
\hline
$R_t$ & $U_t$ & $P(U_t | R_t)$ \\
\hline
$+r$ & $+u$ & 0.9 \\
$+r$ & $-u$ & 0.1 \\
$-r$ & $+u$ & 0.2 \\
$-r$ & $-u$ & 0.8 \\
\hline
\end{tabular}

$B(+r) = 0.5$

$B(-r) = 0.5$

Slide: Berkeley CS188 course notes (downloaded Summer 2015)
Weather HMM example

\[ P(R_{t+1}|R_t) \]

\begin{align*}
B'(+r) &= 0.5 \\
B'(-r) &= 0.5
\end{align*}

\begin{align*}
B(+r) &= 0.5 \\
B(-r) &= 0.5
\end{align*}

\[ P(U_t|R_t) \]

\begin{tabular}{|c|c|c|}
\hline
\( R_t \) & \( R_{t+1} \) & \( P(R_{t+1}|R_t) \) \\
\hline
+\( r \) & +\( r \) & 0.7 \\
+\( r \) & -\( r \) & 0.3 \\
-\( r \) & +\( r \) & 0.3 \\
-\( r \) & -\( r \) & 0.7 \\
\hline
\end{tabular}

\begin{tabular}{|c|c|c|}
\hline
\( R_t \) & \( U_t \) & \( P(U_t|R_t) \) \\
\hline
+\( r \) & +\( u \) & 0.9 \\
+\( r \) & -\( u \) & 0.1 \\
-\( r \) & +\( u \) & 0.2 \\
-\( r \) & -\( u \) & 0.8 \\
\hline
\end{tabular}

Slide: Berkeley CS188 course notes (downloaded Summer 2015)
Weather HMM example

\[ P(R_{t+1} | R_t) \]

\[ B'(+r) = 0.5 \]
\[ B'(-r) = 0.5 \]

\[ B(+r) = 0.5 \]
\[ B(-r) = 0.5 \]

\[ B(+r) = 0.818 \]
\[ B(-r) = 0.182 \]

\[
\begin{array}{ccc}
R_t & R_{t+1} & P(R_{t+1} | R_t) \\
+ r & + r & 0.7 \\
+ r & - r & 0.3 \\
- r & + r & 0.3 \\
- r & - r & 0.7 \\
\end{array}
\]

\[
\begin{array}{ccc}
R_t & U_t & P(U_t | R_t) \\
+ r & + u & 0.9 \\
+ r & - u & 0.1 \\
- r & + u & 0.2 \\
- r & - u & 0.8 \\
\end{array}
\]
Weather HMM example

\[ P(R_{t+1} | R_t) \]

\begin{align*}
B'(r) &= 0.5 \\
B'(\neg r) &= 0.5 \\
B(r) &= 0.818 \\
B(\neg r) &= 0.182
\end{align*}

\begin{align*}
R_t & | P(R_{t+1} | R_t) \\
+ r & | + r & 0.7 \\
+ r & | - r & 0.3 \\
- r & | + r & 0.3 \\
- r & | - r & 0.7
\end{align*}

\begin{align*}
R_t & | U_t & P(U_t | R_t) \\
+ r & | + u & 0.9 \\
+ r & | - u & 0.1 \\
- r & | + u & 0.2 \\
- r & | - u & 0.8
\end{align*}
Weather HMM example

\[
R_t \quad R_{t+1} \quad P(R_{t+1}|R_t)
\]

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>+r</td>
<td>+r</td>
<td>0.7</td>
</tr>
<tr>
<td>+r</td>
<td>-r</td>
<td>0.3</td>
</tr>
<tr>
<td>-r</td>
<td>+r</td>
<td>0.3</td>
</tr>
<tr>
<td>-r</td>
<td>-r</td>
<td>0.7</td>
</tr>
</tbody>
</table>

\[
B(\pm r) = 0.5 \\
B'(\pm r) = 0.5
\]

\[
B(\pm u) = 0.9 \\
B'(\pm u) = 0.1
\]

\[
B(\pm r) = 0.818 \\
B'(\pm r) = 0.182
\]

\[
B(\pm r) = 0.883 \\
B'(\pm r) = 0.117
\]

\[
B(\pm r) = 0.627 \\
B'(\pm r) = 0.373
\]

Slide: Berkeley CS188 course notes (downloaded Summer 2015)
Particle Filtering
Our representation of $P(X)$ is now a list of $N$ particles (samples)
- Generally, $N << |X|$
- Storing map from $X$ to counts would defeat the point

$P(x)$ approximated by number of particles with value $x$
- So, many $x$ may have $P(x) = 0!$
- More particles, more accuracy

For now, all particles have a weight of 1
Particle Filtering: Elapse Time

- Each particle is moved by sampling its next position from the transition model
  \[ x' = \text{sample}(P(X'|x)) \]

- This is like prior sampling – samples’ frequencies reflect the transition probabilities

- Here, most samples move clockwise, but some move in another direction or stay in place

- This captures the passage of time
  - If enough samples, close to exact values before and after (consistent)
Particle Filtering: Observe

- Slightly trickier:
  - Don’t sample observation, fix it
  - Similar to likelihood weighting, downweight samples based on the evidence
    $$w(x) = P(e|x)$$

    $$B(X) \propto P(e|X)B'(X)$$

- As before, the probabilities don’t sum to one, since all have been downweighted (in fact they now sum to (N times) an approximation of P(e))
Particle Filtering: Resample

- Rather than tracking weighted samples, we resample.

- N times, we choose from our weighted sample distribution (i.e. draw with replacement).

- This is equivalent to renormalizing the distribution.

- Now the update is complete for this time step, continue with the next one.
Recap: Particle Filtering

- Particles: track samples of states rather than an explicit distribution
In robot localization:

- We know the map, but not the robot’s position
- Observations may be vectors of range finder readings
- State space and readings are typically continuous (works basically like a very fine grid) and so we cannot store $B(X)$
- Particle filtering is a main technique
Particle Filter Localization (Sonar)
Particle Filter Localization (Laser)
Dynamic Bayes Nets
Dynamic Bayes Nets (DBNs)

- We want to track multiple variables over time, using multiple sources of evidence
- Idea: Repeat a fixed Bayes net structure at each time
- Variables from time $t$ can condition on those from $t-1$

Dynamic Bayes nets are a generalization of HMMs
A particle is a complete sample for a time step

**Initialize**: Generate prior samples for the $t=1$ Bayes net
- Example particle: $G_1^a = (3,3) \; G_1^b = (5,3)$

**Elapse time**: Sample a successor for each particle
- Example successor: $G_2^a = (2,3) \; G_2^b = (6,3)$

**Observe**: Weight each *entire* sample by the likelihood of the evidence conditioned on the sample
- Likelihood: $P(E_1^a | G_1^a) \cdot P(E_1^b | G_1^b)$

**Resample**: Select prior samples (tuples of values) in proportion to their likelihood