CSE 573: Artificial Intelligence

Hanna Hajishirzi
Hidden Markov Models

slides adapted from
Dan Klein, Pieter Abbeel ai.berkeley.edu
And Dan Weld, Luke Zettelmoyer
Reasoning over Time or Space

- Often, we want to reason about a sequence of observations
  - Speech recognition
  - Robot localization
  - User attention
  - Medical monitoring

- Need to introduce time (or space) into our models
Markov Models

- Value of $X$ at a given time is called the **state**

\[ P(X_2|X_1) = P(X_3|X_2) \ldots \]

\[ P(X_1) \quad P(X_t|X_{t-1}) \]

- Parameters: called **transition probabilities** or dynamics, specify how the state evolves over time (also, initial state probabilities)
- Stationarity assumption: transition probabilities the same at all times
- Same as MDP transition model, but no choice of action
- A (growable) BN: We can always use generic BN reasoning on it if we truncate the chain at a fixed length

\[ P(X_t) = ? \]
Markov Assumption: Conditional Independence

- Basic conditional independence:
  - Past and future independent given the present
  - Each time step only depends on the previous
  - This is called the (first order) Markov property
Example Markov Chain: Weather

- States: \( X = \{ \text{rain, sun} \} \)

- Initial distribution: 1.0 sun

- CPT \( P(X_t | X_{t-1}) \):

\[
P(X_{t-1} = \text{sun}) = 1
\]

\[
P(X_{t-1} = \text{sun} | X_t = \text{sun}) = 0.9
\]

\[
P(X_{t-1} = \text{sun} | X_t = \text{rain}) = 0.1
\]

\[
P(X_{t-1} = \text{rain} | X_t = \text{sun}) = 0.3
\]

\[
P(X_{t-1} = \text{rain} | X_t = \text{rain}) = 0.7
\]
Bayes Nets -- Independence

- **Bayes Net**
  \[ P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \text{parents}(X_i)) \]

- **Chain Rule**
  \[ P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | x_1 \ldots x_{i-1}) \]
Markov Models (Markov Chains)

A Markov model defines a joint probability distribution:

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

More generally:

\[
P(X_1, X_2, \ldots, X_T) = P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1})
\]

\[
P(X_1, \ldots, X_n) = P(X_1)\prod_{t=2}^{N} P(X_t|X_{t-1})
\]

One common inference problem:

- Compute marginals \(P(X_t)\) for all time steps \(t\)

Why?
- Chain Rule, Indep. Assumption?
Example Markov Chain: Weather

- Initial distribution: 1.0 sun
  \[ P(X_1 = \text{sun}) = 1 \]
  \[ P(X_1 = \text{rain}) = 0 \]

- What is the probability distribution after one step?

\[
P(X_2 = \text{sun}) = \sum_{x_1} P(x_1, X_2 = \text{sun}) = \sum_{x_1} P(X_2 = \text{sun} | x_1)P(x_1)
\]

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

\[
0.9 \cdot 1.0 + 0.3 \cdot 0.0 = 0.9
\]
Mini-Forward Algorithm

- Question: What’s $P(X)$ on some day $t$?

$P(x_1) = \text{known}$

$P(x_t) = \sum_{x_{t-1}} P(x_{t-1}, x_t)$

$= \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1})$

Forward simulation
Example Run of Mini-Forward Algorithm

- From initial observation of sun
  \[
  \begin{pmatrix}
  1.0 \\
  0.0
  \end{pmatrix}
  \begin{pmatrix}
  0.9 \\
  0.1
  \end{pmatrix}
  \begin{pmatrix}
  0.84 \\
  0.16
  \end{pmatrix}
  \begin{pmatrix}
  0.804 \\
  0.196
  \end{pmatrix}
  \begin{pmatrix}
  0.75 \\
  0.25
  \end{pmatrix}
  \]
  \[P(X_1) \quad P(X_2) \quad P(X_3) \quad P(X_4) \quad P(X_{\infty})\]

- From initial observation of rain
  \[
  \begin{pmatrix}
  0.0 \\
  1.0
  \end{pmatrix}
  \begin{pmatrix}
  0.3 \\
  0.7
  \end{pmatrix}
  \begin{pmatrix}
  0.48 \\
  0.52
  \end{pmatrix}
  \begin{pmatrix}
  0.588 \\
  0.412
  \end{pmatrix}
  \begin{pmatrix}
  0.75 \\
  0.25
  \end{pmatrix}
  \]
  \[P(X_1) \quad P(X_2) \quad P(X_3) \quad P(X_4) \quad P(X_{\infty})\]

- From yet another initial distribution \(P(X_1)\):
  \[
  \begin{pmatrix}
  p \\
  1 - p
  \end{pmatrix}
  \begin{pmatrix}
  \vdots
  \end{pmatrix}
  \begin{pmatrix}
  0.75 \\
  0.25
  \end{pmatrix}
  \]
  \[P(X_1) \quad \ldots \quad P(X_{\infty})\]
Pac-man knows the ghost’s initial position, but gets no observations!
Video of Demo Ghostbusters Circular Dynamics
Stationary Distributions

- For most chains:
  - Influence of the initial distribution gets less and less over time.
  - The distribution we end up in is independent of the initial distribution

- **Stationary distribution:**
  - The distribution we end up with is called the stationary distribution $P_\infty$ of the chain.
  - It satisfies
    \[ P_\infty(X) = P_{\infty+1}(X) = \sum_x P(X|x)P_\infty(x) \]
Example: Stationary Distributions

Question: What’s \( P(X) \) at time \( t = \infty \)?

\[
P_{\infty}(\text{sun}) = P(\text{sun}|\text{sun})P_{\infty}(\text{sun}) + P(\text{sun}|\text{rain})P_{\infty}(\text{rain})
\]
\[
P_{\infty}(\text{rain}) = P(\text{rain}|\text{sun})P_{\infty}(\text{sun}) + P(\text{rain}|\text{rain})P_{\infty}(\text{rain})
\]

\[
P_{\infty}(\text{sun}) = 0.9P_{\infty}(\text{sun}) + 0.3P_{\infty}(\text{rain})
\]
\[
P_{\infty}(\text{rain}) = 0.1P_{\infty}(\text{sun}) + 0.7P_{\infty}(\text{rain})
\]

\[
P_{\infty}(\text{sun}) = 3P_{\infty}(\text{rain})
\]
\[
P_{\infty}(\text{rain}) = 1/3P_{\infty}(\text{sun})
\]

Also:
\[
P_{\infty}(\text{sun}) + P_{\infty}(\text{rain}) = 1
\]

\[
P_{\infty}(\text{sun}) = 3/4
\]
\[
P_{\infty}(\text{rain}) = 1/4
\]
Application of Stationary Distribution: Web Link Analysis

- **PageRank over a web graph**
  - Each web page is a possible value of a state
  - Initial distribution: uniform over pages
  - Transitions:
    - With prob. $c$, uniform jump to a random page (dotted lines, not all shown)
    - With prob. $1-c$, follow a random outlink (solid lines)

- **Stationary distribution**
  - Will spend more time on highly reachable pages
  - E.g., many ways to get to the Acrobat Reader download page
  - Google 1.0 returned the set of pages containing all your keywords in decreasing rank, now all search engines use link analysis along with many other factors (rank actually getting less important over time)
Hidden Markov Models
Pacman – Sonar
Hidden Markov Models

- Markov chains not so useful for most agents
  - Need observations to update your beliefs

- Hidden Markov models (HMMs)
  - Underlying Markov chain over states $X$
  - You observe outputs (effects) at each time step
Example: Weather HMM

An HMM is defined by:
- Initial distribution: \( P(X_1) \)
- Transitions: \( P(X_t | X_{t-1}) \)
- Emissions: \( P(E_t | X_t) \)

| \( R_{t-1} \) | \( R_t \) | \( P(R_t | R_{t-1}) \) |
|---|---|---|
| \(+r\) | \(+r\) | 0.7 |
| \(+r\) | \(-r\) | 0.3 |
| \(-r\) | \(+r\) | 0.3 |
| \(-r\) | \(-r\) | 0.7 |

| \( R_t \) | \( U_t \) | \( P(U_t | R_t) \) |
|---|---|---|
| \(+r\) | \(+u\) | 0.9 |
| \(+r\) | \(-u\) | 0.1 |
| \(-r\) | \(+u\) | 0.2 |
| \(-r\) | \(-u\) | 0.8 |
Example: Ghostbusters HMM

- \( P(X_1) = \text{uniform} \)
- \( P(X|X') = \) usually move clockwise, but sometimes move in a random direction or stay in place
- \( P(R_{ij}|X) = \) same sensor model as before: red means close, green means far away.

\[
\begin{array}{c|c|c|c}
X_1 & X_2 & X_3 & X_4 \\
\hline
R_{i,j} & R_{i,j} & R_{i,j} & R_{i,j}
\end{array}
\]

\[
\begin{array}{c|c|c|c}
1/9 & 1/9 & 1/9 \\
1/9 & 1/9 & 1/9 \\
1/9 & 1/9 & 1/9 \\
\end{array}
\]

\[
\begin{array}{c|c|c|c}
1/6 & 1/6 & 1/2 \\
0 & 1/6 & 0 \\
0 & 0 & 0 \\
\end{array}
\]
Video of Demo Ghostbusters – Circular Dynamics -- HMM
Conditional Independence

- HMMs have two important independence properties:
  - Markov hidden process: future depends on past via the present
  - Current observation independent of all else given current state

- Does this mean that evidence variables are guaranteed to be independent?
  - [No, they tend to correlated by the hidden state]
Real HMM Examples

- **Robot tracking:**
  - Observations are range readings (continuous)
  - States are positions on a map (continuous)

- **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
Filtering / Monitoring

- Filtering, or monitoring, is the task of tracking the distribution \( B_t(X) = \Pr(X_t \mid e_1, \ldots, e_t) \) (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.
Example: Robot Localization

Sensor model: can read in which directions there is a wall, never more than 1 mistake.
Motion model: may not execute action with small prob.
Example: Robot Localization

Lighter grey: was possible to get the reading, but less likely b/c required 1 mistake

$t=1$
Example: Robot Localization

t=2

Prob

0 1
Example: Robot Localization

\[ t=3 \]
Example: Robot Localization

![Robot Localization Diagram]
Example: Robot Localization

\[ t=5 \]
Inference: Find State Given Evidence

- We are given evidence at each time and want to know

\[ B_t(X) = P(X_t|e_1:t) \]

- Idea: start with \( P(X_1) \) and derive \( B_t \) in terms of \( B_{t-1} \)
  - equivalently, derive \( B_{t+1} \) in terms of \( B_t \)