Lecture 2: Bayesian inference and Hidden Markov Models

Scribe: *Audrow Nash*

1 Probability

Chain rule

$$
\begin{aligned}
P(X_1, \ldots, X_n) &= P(X_1|X_2, \ldots, X_n)P(X_2, \ldots, X_n) \\
&= P(X_1|X_2, \ldots, X_n)P(X_2|X_3, \ldots, X_n)P(X_3, \ldots, X_n) \\
&\vdots \\
&= \Pi_{i=1}^{n} P(X_i|parents(X_i))
\end{aligned}
$$

A note on notation: $P(A|B, C) = P(A|B \cap C)$.

2 Bayesian Networks

1. Each *node* corresponds to a random variable, which is either discrete or continuous.
2. Directed links connect pairs of nodes; $X \rightarrow Y$ means X is a *parent* of Y.
3. Each node has a conditional probability distribution: $P(X|parents(X))$.
4. Each node is conditionally independent of its non-descendents given its parents.

2.1 Example

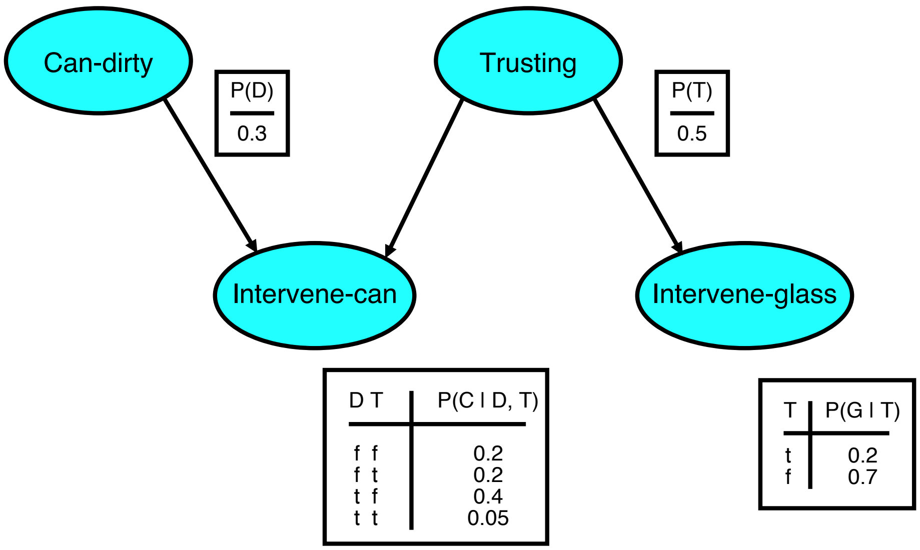

Figure 1: The Bayesian Network for the example, which uses the example of a robot reaching for a glass or a can where a human can intervine.

$$P(D, C, T, G) = P(C|D, T)P(G|T)P(D)P(G)$$

$$\begin{aligned} P(G) =& P(G|T)P(T) + P(G|\neg T)P(\neg T) \\ =& 0.2 \cdot 0.5 + 0.7 \cdot 0.5 \\ =& \boxed{0.45} \end{aligned}$$

$$\begin{aligned} P(C) =& P(C|D, T)P(D, T) \\ &+ P(C|\neg D, T)P(\neg D, T) \\ &+ P(C|D, \neg T)P(D, \neg T) \\ =& P(C|D, T)P(D)P(T) \\ &+ P(C|\neg D, T)P(\neg D)P(T) \\ &+ P(C|\neg D, \neg T)P(\neg D, \neg T) \\ &+ P(C|D, \neg T)P(D, \neg T) \\ =& 0.05 \cdot 0.3 \cdot 0.5 + 0.2 \cdot 0.7 \cdot 0.5 + 0.9 \cdot 0.7 \cdot 0.5 + 0.4 \cdot 0.7 \cdot 0.5 \\ =& \boxed{0.3825} \end{aligned}$$

$$P(T|G) = \frac{P(G|T)P(T)}{P(G)} = \frac{0.2 \cdot 0.5}{0.45} = \boxed{0.22}$$

$$\begin{aligned} P(C|G) =& P(C|D,T,G)P(D,T,G) \\ &+ P(C|\neg D,\neg T,G)P(\neg D,\neg T,G) \\ &+ P(C|D,\neg T,G)P(D,\neg T,G) \\ &+ P(C|\neg D,T,G)P(\neg D,T,G) \\ &= P(C|D,T)P(T|G)P(D) \\ &+ P(C|\neg D,\neg T)P(\neg D)P(\neg T|G) \\ &+ P(C|D,\neg T)P(D)P(\neg T|G)P(C|\neg D,T)P(T|G) \\ &= \boxed{0.51} \end{aligned}$$

3 Bayesian Inference

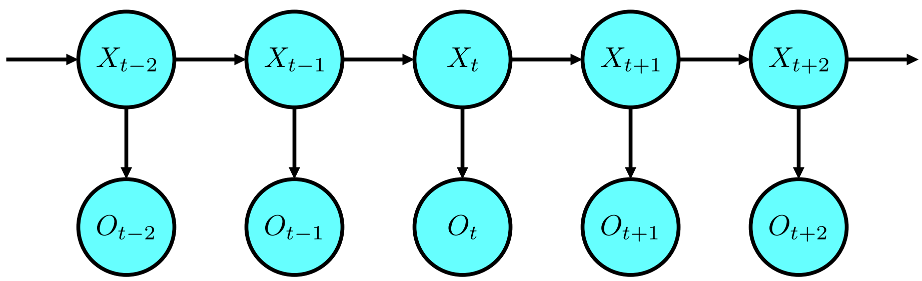

Figure 2: A Hidden Markov Model, where X_i and O_i denotes the state and observation at time step i.

We'll be focusing on Discrete Random Variables.

Markov assumption: The next state depends only on the current state - not past states.

$$P(X_t|X_0,\ldots,X_{t-1}) = P(X_t|X_{0:t-1}) = P(X_t|X_{t-1})$$

Components of a Hidden Markov Model:

1. χ: set of states
2. Ω: set of observations
3. $T : X \rightarrow \Pi(X)$: state transition probabilities
4. $M : X \rightarrow \Pi(\Omega)$: observation probabilities

3.1 Example

3.1.1 Setup

$$X = \{Desk, Whiteboard, Door\}$$

Dynamics model:

$$T = \left(\begin{array}{llccc} & & & X_t = & \\ & & \text{Desk} & \text{Whiteboard} & \text{Door} \\ X_{t-1} = & \text{Desk} & 0.4 & 0.4 & 0.2 \\ & \text{Whiteboard} & 0.4 & 0.4 & 0.2 \\ & \text{Door} & 0.0 & 0.0 & 1.0 \end{array} \right)$$

Sensor model:

$$M = \left(\begin{array}{llccc} & & & O_t = & \\ & & \text{Desk} & \text{Whiteboard} & \text{Door} \\ X_t = & \text{Desk} & 0.8 & 0.1 & 0.1 \\ & \text{Whiteboard} & 0.1 & 0.8 & 0.1 \\ & \text{Door} & 0.1 & 0.1 & 0.8 \end{array} \right)$$

3.1.2 Given

$$observations = \{desk, door, desk\}$$

$$P(X_0) = \{1/3,\ 1/3,\ 1/3\}$$

3.1.3 Exercises

$$
\begin{aligned}
P(X_t|O_{1:t}) =& P(X_t|O_{1:t-1}, O_t) \\
=& \frac{P(O_t|O_{1:t-1}, X_t)P(X_t|O_{1:t-1})}{P(O_t|O_{1:t-1})} \\
=& \eta P(O_t|O_{1:t-1}, X_t)P(X_t|O_{1:t-1}) \\
=& \eta P(O_t|X_t)P(X_t|O_{1:t-1}) \\
=& \eta P(O_t|X_t) \sum_{X_{t-1}} P(X_t|X_{t-1}, O_{1:t-1})P(X_{t-1}|O_{1:t-1})
\end{aligned}
$$

$$
P(X_1|O_1) = \eta\ \mu(O_1|X_1) \sum_{X_0} T(X_1|X_0)P(X_0)
$$

$$
\begin{aligned}
P(X_1 = desk|O_1 = desk) =& \eta\ M(O_1 = desk|X_0 = desk) \\
& (T(X_1 = desk|X_0 = desk)P(X_0 = desk) \\
& + T(X_1 = desk|X_0 = whiteboard)P(X_0 = whiteboard) \\
& + T(X_1 = desk|X_0 = door)P(X_0 = door)) \\
& = \eta \cdot 0.8 \cdot (0.4 \cdot 1/3 + 0.4 \cdot 1/3 + 0 \cdot 1/3) \\
& = \eta \cdot 0.213
\end{aligned}
$$

$$
\begin{aligned}
P(X_1 = whiteboard|O_1 = desk) =& \eta \cdot 0.1 \cdot (0.4 \cdot 1/3 + 0.4 \cdot 1/3 + 0) \\
=& \eta \cdot 0.027
\end{aligned}
$$

$$
\begin{aligned}
P(X_1 = door|O_1 = desk) =& \eta \cdot 0.1 \cdot (0.2 \cdot 1/3 + 0.2 \cdot 1/3 + 1 \cdot 1/3) \\
=& \eta \cdot 0.097
\end{aligned}
$$

$$
\begin{aligned}
1 =& P(X_1 = desk|O_1 = desk) \\
& + P(X_1 = whiteboard|O_1 = desk) \\
& + P(X_1 = door|O_1 = desk) \\
=& \eta \cdot (0.213 + 0.027 + 0.097) \Rightarrow \boxed{\eta = 1/0.287}
\end{aligned}
$$

$$P(X_1) = \{\text{Desk: } 0.744,\ \text{Whiteboard: } 0.093,\ \text{Door: } 0.163\}$$

Repeat the same process using the updated probabilities to solve for $P(X_2 = desk | O_1 = desk,\ O_2 = door)$ and beyond.

t	$P(desk)$	$P(whiteboard)$	$P(door)$
0	1/3	1/3	1/3
1	0.74	0.09	0.16
2	0.10	0.10	0.80
3	0.41	0.05	0.54

Can we predict the future? Condition without observations:

$$\begin{aligned} P(X_{t+k} | O_{1:t}) &= \sum_{x_t+k-1} P(X_{t+k} | X_{t+k-1}, O_{1:t}) P(X_{t+k-1} | O_{1:t}) \\ &= \sum_{x_t+k-1} P(X_{t+k} | X_{t+k-1}) P(X_{t+k-1} | O_{1:t}) \end{aligned}$$

Prediction:

t	$P(desk)$	$P(whiteboard)$	$P(door)$
4	0.19	0.19	0.62
5	0.15	0.15	0.70
6	0.12	0.12	0.76
7	0.10	0.10	0.80
8	0.08	0.08	0.84

Note: The door is an absorbing state, so when predicting without evidence the belief converges to the door.

3.2 Discussion of examples

What is the time complexity of predictions?

- $\mathcal{O}(c)$
- $\mathcal{O}(\log t)$
- $\boldsymbol{\mathcal{O}(t)}$ $\Leftarrow$ (answer; aka, linear)
- $\mathcal{O}(t^2)$

3.3 Misc. Notes

Forward pass:

$$
\begin{aligned}
P(X_k|O_{1:t}) =& \eta P(X_k|O_{1:k}, O_{k+1:t}) \\
=& \eta P(O_{k+1:t}|X_k, O_{1:k}) P(X_k|O_{1:k}) \\
=& \eta P(O_{k+1:t}|X_k, O_{1:k}) P(X_k|O_{1:k}) \\
=& \eta P(X_k|O_{1:k}) P(O_{k+1:t}|X_k) \\
=& \eta \cdot f_{1:k} \cdot b_{k+1:t}
\end{aligned}
$$

Backwards pass:

$$
\begin{aligned}
P(O_{k+1:t}|X_k) &= \sum_{X_k+1} P(O_{k+1:t}|X_k, X_{k+1}) P(X_{k+1}|X_k) \\
&= \sum_{X_k+1} P(O_{k+1:t}|X_{k+1}) P(X_{k+1}|X_k) \\
&= \sum_{X_k+1} P(O_{k+1}, O_{k+2:t}|X_{k+1}) P(X_{k+1}|X_k) \\
&= \sum_{X_k+1} P(O_{k+1}|X_{k+1}) P(O_{k+2:t}|X_{k+1}) P(X_{k+1}|X_k) \\
&= \sum_{X_k+1} M(O_{k+1}|X_{k+1}) P(O_{k+2:t}|X_{k+1}) T(X_{k+1}|X_k)
\end{aligned}
$$