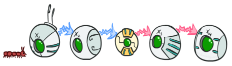
CS5100: Foundations of Artificial Intelligence

Markov Models
Hidden Markov Models

Dr. Rutu Mulkar-Mehta
Lecture 10



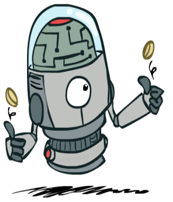
[These slides were created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley. All CS188 materials are available at http://ai.berkeley.edu.]

Independence

- Two variables are *independent* in a joint distribution if:

$$P(X,Y) = P(X)P(Y) \qquad X \perp\!\!\!\perp Y$$
$$\forall x, y\; P(x,y) = P(x)P(y)$$

  - Says the joint distribution *factors* into a product of two simple ones
  - Usually variables aren't independent!
- Can use independence as a *modeling assumption*
  - Independence can be a simplifying assumption
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity}?



Example: Independence?

$P_1(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

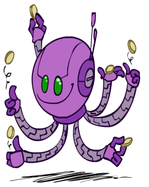
W	P
sun	0.6
rain	0.4

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

T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

Example: Independence

- N fair, independent coin flips:



$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

H	0.5
T	0.5

. . .

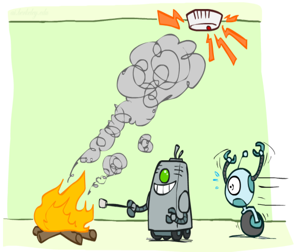
$P(X_n)$

H	0.5
T	0.5

$P(X_1, X_2, \ldots X_n)$

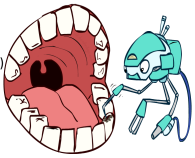
2^n

Conditional Independence



Conditional Independence

- P(Toothache, Cavity, Catch)
- If I have a cavity, the probability that the probe catches it doesn't depend on whether I have a toothache:
  - P(+catch | +toothache, +cavity) = P(+catch | +cavity)
- The same independence holds if I don't have a cavity:
  - P(+catch | +toothache, -cavity) = P(+catch| -cavity)
- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)



- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other easily

Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z $X \perp\!\!\!\perp Y | Z$

if and only if:

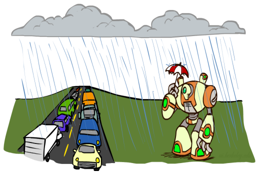
$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

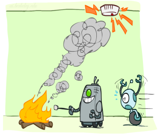
Conditional Independence

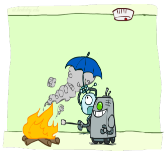
- What about this domain:
  - Traffic
  - Umbrella
  - Raining



Conditional Independence

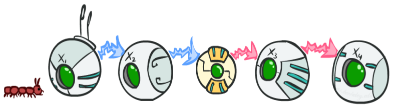
- What about this domain:
  - Fire
  - Smoke
  - Alarm





Probability Recap

- Conditional probability $P(x|y) = \frac{P(x, y)}{P(y)}$
- Product rule $P(x, y) = P(x|y)P(y)$
- Chain rule $$P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})$$
- X, Y independent if and only if: $\forall x, y : P(x, y) = P(x)P(y)$
- X and Y are conditionally independent given Z if and only if: $\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$ $X \perp\!\!\!\perp Y | Z$



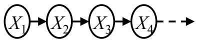
MARKOV MODELS

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

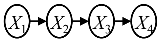


$P(X_1)$ $P(X_t|X_{t-1})$

– 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

Joint Distribution of a Markov Model



$P(X_1)$ $P(X_t|X_{t-1})$

– Joint distribution:

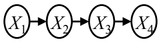
$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

– More generally:

$$P(X_1, X_2, \ldots, X_T) = P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1})$$
$$= P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})$$

– Questions to be resolved:
- Does this indeed define a joint distribution?
- Can every joint distribution be factored this way, or are we making some assumptions about the joint distribution by using this factorization?

Chain Rule and Markov Models



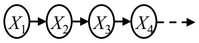
- From the chain rule, every joint distribution over X_1, X_2, X_3, X_4 can be written as:

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

- Assuming that $X_3 \perp\!\!\!\perp X_1 \mid X_2$ and $X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$ results in the expression posited on the previous slide:

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

Chain Rule and Markov Models



- From the chain rule, every joint distribution over $X_1, X_2, \ldots, X_T$ can be written as:

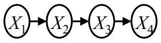
$$P(X_1, X_2, \ldots, X_T) = P(X_1)\prod_{t=2}^{T} P(X_t|X_1, X_2, \ldots, X_{t-1})$$

- Assuming that for all *t*: $X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$

gives us the expression posited on the earlier slide:

$$P(X_1, X_2, \ldots, X_T) = P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})$$

Implied Conditional Independencies



- We assumed: $X_3 \perp\!\!\!\perp X_1 \mid X_2$ and

$X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$

- Do we also have $X_1 \perp\!\!\!\perp X_3, X_4 \mid X_2$?
  - Yes!
  - Proof:

$$\begin{aligned} P(X_1 \mid X_2, X_3, X_4) &= \frac{P(X_1, X_2, X_3, X_4)}{P(X_2, X_3, X_4)} \\ &= \frac{P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_3)}{\sum_{x_1} P(x_1)P(X_2 \mid x_1)P(X_3 \mid X_2)P(X_4 \mid X_3)} \\ &= \frac{P(X_1, X_2)}{P(X_2)} \\ &= P(X_1 \mid X_2) \end{aligned}$$

Markov Models Recap

- Explicit assumption for all *t*: $X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$
- Consequence, joint distribution can be written as:

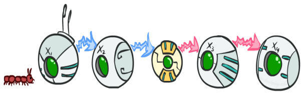
$$P(X_1, X_2, \ldots, X_T) = P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1})$$
$$= P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})$$

- Implied conditional independencies: (try to prove this!)
  - Past variables independent of future variables given the present

i.e., if $t_1 < t_2 < t_3$ or $t_1 > t_2 > t_3$ then: $X_{t_1} \perp\!\!\!\perp X_{t_3} \mid X_{t_2}$

- Additional explicit assumption: $P(X_t \mid X_{t-1})$ is the same for all *t*

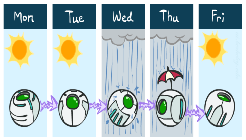
Conditional Independence



- Basic conditional independence:
  - – Past and future independent of the present
  - – Each time step only depends on the previous
  - – This is called the (first order) Markov property
- Note that the chain is just a (growable) BN
  - – We can always use generic BN

Example Markov Chain: Weather

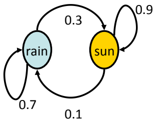
- States: X = {rain, sun}

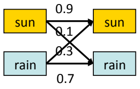


- Initial distribution: 1.0 sun
- CPT P(X_t | X_{t-1}):

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

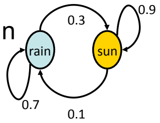
Two new ways of representing the same CPT





Example Markov Chain: Weather

- Initial distribution: 1.0 sun



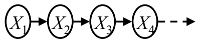
- What is the probability distribution after one step?

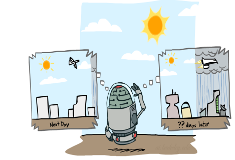
$$P(X_2 = \text{sun}) = P(X_2 = \text{sun}|X_1 = \text{sun})P(X_1 = \text{sun}) + P(X_2 = \text{sun}|X_1 = \text{rain})P(X_1 = \text{rain})$$

$$0.9 \cdot 1.0 + 0.3 \cdot 0.0 = 0.9$$

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

$$\underset{P(X_1)}{\left\langle \begin{matrix} 1.0 \\ 0.0 \end{matrix} \right\rangle} \underset{P(X_2)}{\left\langle \begin{matrix} 0.9 \\ 0.1 \end{matrix} \right\rangle} \underset{P(X_3)}{\left\langle \begin{matrix} 0.84 \\ 0.16 \end{matrix} \right\rangle} \underset{P(X_4)}{\left\langle \begin{matrix} 0.804 \\ 0.196 \end{matrix} \right\rangle} \Rightarrow \underset{P(X_\infty)}{\left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle}$$

- From initial observation of rain

$$\underset{P(X_1)}{\left\langle \begin{matrix} 0.0 \\ 1.0 \end{matrix} \right\rangle} \underset{P(X_2)}{\left\langle \begin{matrix} 0.3 \\ 0.7 \end{matrix} \right\rangle} \underset{P(X_3)}{\left\langle \begin{matrix} 0.48 \\ 0.52 \end{matrix} \right\rangle} \underset{P(X_4)}{\left\langle \begin{matrix} 0.588 \\ 0.412 \end{matrix} \right\rangle} \Rightarrow \underset{P(X_\infty)}{\left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle}$$

- From yet another initial distribution P(X_1):

$$\underset{P(X_1)}{\left\langle \begin{matrix} p \\ 1-p \end{matrix} \right\rangle} \quad \cdots \quad \Rightarrow \underset{P(X_\infty)}{\left\langle \begin{matrix} 0.75 \\ 0.25 \end{matrix} \right\rangle}$$

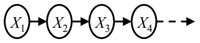
Stationary Distributions

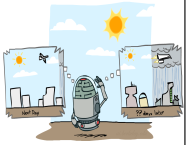
- For most chains:
  - – Influence of the initial distribution gets less and less over time.
  - – The distribution we end up in is independent of the initial distribution
- Stationary distribution:
  - The distribution we end up with is called the stationary distribution P_∞ of the chain
  - It satisfies

$$P_\infty(X) = P_{\infty+1}(X) = \sum_x P(X|x)P_\infty(x)$$

Example: Stationary Distributions

Question: What's P(X) at time t = infinity?





$$P_\infty(sun) = P(sun|sun)P_\infty(sun) + P(sun|rain)P_\infty(rain)$$
$$P_\infty(rain) = P(rain|sun)P_\infty(sun) + P(rain|rain)P_\infty(rain)$$

$$P_\infty(sun) = 0.9P_\infty(sun) + 0.3P_\infty(rain)$$
$$P_\infty(rain) = 0.1P_\infty(sun) + 0.7P_\infty(rain)$$

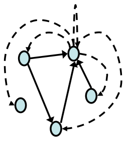
$$P_\infty(sun) = 3P_\infty(rain)$$
$$P_\infty(rain) = 1/3P_\infty(sun)$$

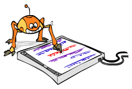
Also: $P_\infty(sun) + P_\infty(rain) = 1$ ⇨ $P_\infty(sun) = 3/4$ $P_\infty(rain) = 1/4$

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

Application of Stationary Distribution: Web Link Analysis

- PageRank over a web graph
  - Each web page is 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)



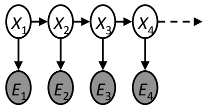


- Stationary distribution
  - Will spend more time on highly reachable pages
  - E.g. many ways to get to the Acrobat Reader download page
  - Somewhat robust to link spam
  - Google 1.0 returned the set of pages containing all your keywords in decreasing rank, now all search engines use link analysis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

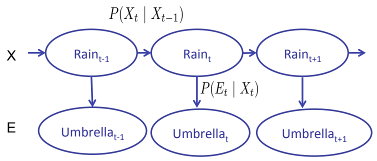
Hidden Markov Models

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

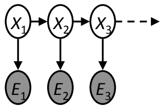


- An HMM is defined by:
  - Initial distribution: $P(X_1)$
  - Transitions: $P(X_t \mid X_{t-1})$
  - Emissions: $P(E_t \mid X_t)$

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

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

Joint Distribution of an HMM



- Joint distribution:

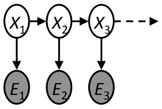
$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1)P(E_2|X_2)P(X_3|X_2)P(E_3|X_3)$$

- More generally:

$$P(X_1, E_1, \ldots, X_T, E_T) = P(X_1)P(E_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(E_t|X_t)$$

- Questions to be resolved:
  - Does this indeed define a joint distribution?
  - Can every joint distribution be factored this way, or are we making some assumptions about the joint distribution by using this factorization?

Chain Rule and HMMs



- From the chain rule, *every* joint distribution over $X_1, E_1, X_2, E_2, X_3, E_3$ can be written as:

$$P(X_1,E_1,X_2,E_2,X_3,E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1,E_1)P(E_2|X_1,E_1,X_2)$$
$$P(X_3|X_1,E_1,X_2,E_2)P(E_3|X_1,E_1,X_2,E_2,X_3)$$

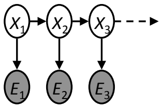
- *Assuming* that

$$X_2 \perp\!\!\!\perp E_1 \mid X_1, \quad E_2 \perp\!\!\!\perp X_1, E_1 \mid X_2, \quad X_3 \perp\!\!\!\perp X_1, E_1, E_2 \mid X_2, \quad E_3 \perp\!\!\!\perp X_1, E_1, X_2, E_2 \mid X_3$$

gives us the expression posited on the previous slide:

$$P(X_1,E_1,X_2,E_2,X_3,E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1)P(E_2|X_2)P(X_3|X_2)P(E_3|X_3)$$

Chain Rule and HMMs



- From the chain rule, *every* joint distribution over $X_1, E_1, \ldots, X_T, E_T$ can be written as:

$$P(X_1,E_1,\ldots,X_T,E_T) = P(X_1)P(E_1|X_1)\prod_{t=2}^{T} P(X_t|X_1,E_1,\ldots,X_{t-1},E_{t-1})P(E_t|X_1,E_1,\ldots,X_{t-1},E_{t-1},X_t)$$

- *Assuming* that for all *t*:
  - State independent of all past states and all past evidence given the previous state, i.e.:

$$X_t \perp\!\!\!\perp X_1, E_1, \ldots, X_{t-2}, E_{t-2}, E_{t-1} \mid X_{t-1}$$

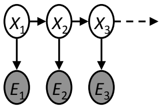
  - Evidence is independent of all past states and all past evidence given the current state, i.e.:

$$E_t \perp\!\!\!\perp X_1, E_1, \ldots, X_{t-2}, E_{t-2}, X_{t-1}, E_{t-1} \mid X_t$$

gives us the expression posited on the earlier slide:

$$P(X_1,E_1,\ldots,X_T,E_T) = P(X_1)P(E_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(E_t|X_t)$$

Implied Conditional Independencies



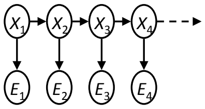
- Many implied conditional independencies, e.g.,

$$E_1 \perp\!\!\!\perp X_2, E_2, X_3, E_3 \mid X_1$$

- To prove them
  - Approach 1: follow similar (algebraic) approach to what we did in the Markov models section
  - Approach 2: directly from the graph structure
    - Intuition: If path between U and V goes through W, then $U \perp\!\!\!\perp V \mid W$

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

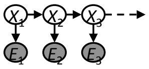


- Quiz: does this mean that evidence variables are guaranteed to be independent?
  - [No, they tend to correlated by the hidden state]

Real HMM Examples

- 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)

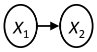
Passage of Time



- Assume we have current belief P(X | evidence to date)

$$B(X_t) = P(X_t|e_{1:t})$$

- Then, after one time step passes:



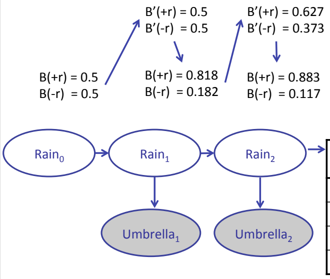
$$P(X_{t+1}|e_{1:t}) = \sum_{x_t} P(X_{t+1}, x_t|e_{1:t})$$
$$= \sum_{x_t} P(X_{t+1}|x_t, e_{1:t})P(x_t|e_{1:t})$$
$$= \sum_{x_t} P(X_{t+1}|x_t)P(x_t|e_{1:t})$$

- Or compactly:

$$B'(X_{t+1}) = \sum_{x_t} P(X'|x_t)B(x_t)$$

- Basic idea: beliefs get "pushed" through the transitions
  - With the "B" notation, we have to be careful about what time step t the belief is about, and what evidence it includes

Example: Weather HMM



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

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

The Forward Algorithm

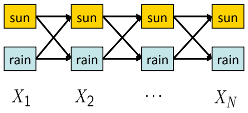
- We are given evidence at each time and want to know $B_t(X) = P(X_t|e_{1:t})$

- We can derive the following updates

$$P(x_t|e_{1:t}) \propto_X P(x_t, e_{1:t})$$
$$= \sum_{x_{t-1}} P(x_{t-1}, x_t, e_{1:t})$$
$$= \sum_{x_{t-1}} P(x_{t-1}, e_{1:t-1})P(x_t|x_{t-1})P(e_t|x_t)$$
$$= P(e_t|x_t) \sum_{x_{t-1}} P(x_t|x_{t-1})P(x_{t-1}, e_{1:t-1})$$

We can normalize as we go if we want to have P(x|e) at each time step, or just once at the end…

Forward / Viterbi Algorithms



Forward Algorithm (Sum)

$$f_t[x_t] = P(x_t, e_{1:t})$$

$$= P(e_t|x_t) \sum_{x_{t-1}} P(x_t|x_{t-1})f_{t-1}[x_{t-1}]$$

Viterbi Algorithm (Max)

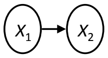
$$m_t[x_t] = \max_{x_{1:t-1}} P(x_{1:t-1}, x_t, e_{1:t})$$

$$= P(e_t|x_t) \max_{x_{t-1}} P(x_t|x_{t-1})m_{t-1}[x_{t-1}]$$

Online Belief Updates

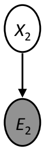
- Every time step, we start with current P(X | evidence)
- We update for time:

$$P(x_t|e_{1:t-1}) = \sum_{x_{t-1}} P(x_{t-1}|e_{1:t-1}) \cdot P(x_t|x_{t-1})$$



- We update for evidence:

$$P(x_t|e_{1:t}) \propto_X P(x_t|e_{1:t-1}) \cdot P(e_t|x_t)$$



- The forward algorithm does both at once (and doesn't normalize)

Now: In Class Assignment