Course 16:198:440: Introduction To Artificial Intelligence
Lecture 12

Temporal Models

Abdeslam Boularias

Wednesday, April 9, 2025

RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

Outline

In the previous lectures on Bayesian and Markov networks, we were concerned about problems where the values of variables are **static**. We now consider problems wherein variables **change over time**.

Example 1: gait modeling

- The joint angles of a person are random variables that change over time.
- Probabilistic modeling can be used to predict the values of the joint angles in a sequence, or to recognize a person from her gait.

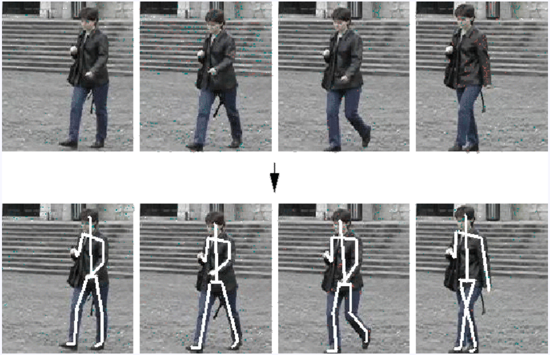
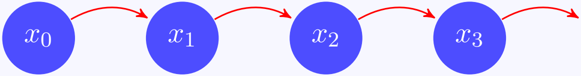
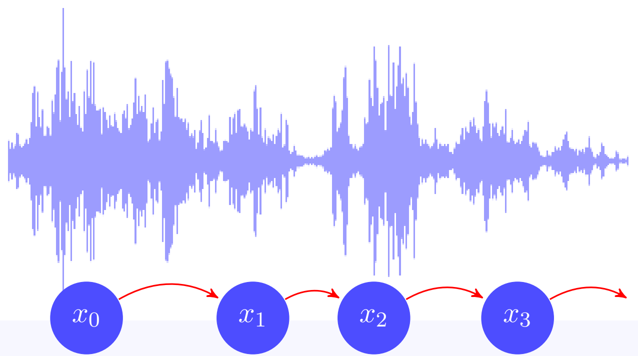


image from inf.ed.ac.uk



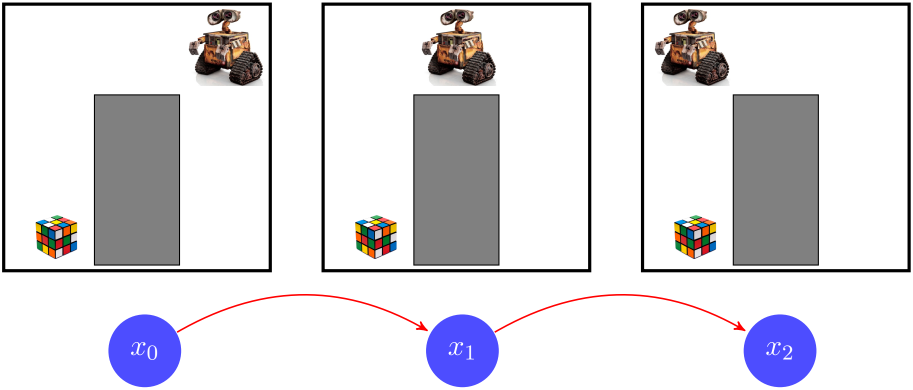
Example 2: natural language processing

- A speech is a sequence of utterances take different values over time.
- Probabilistic modeling is used for speech processing and recognition.



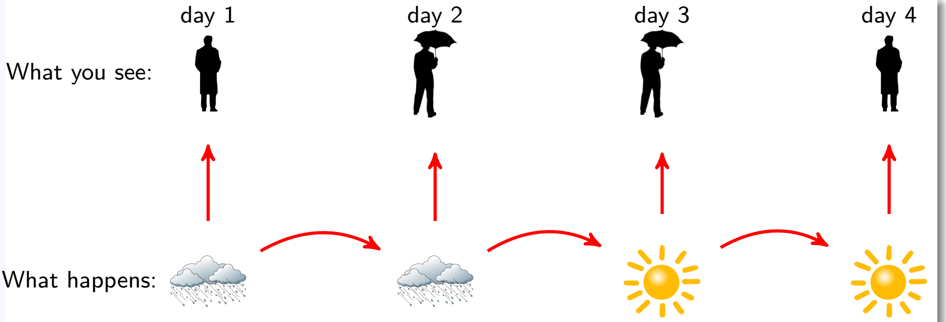
Example 3: robot navigation

The robot's exact position is a random variable that changes over time.



Toy example

- You are the security guard stationed at a secret underground installation.
- You want to know whether it's raining today,
- but your only access to the outside world occurs each morning when you see the director coming in with, or without, an umbrella.



States and observations

- In a temporal model, the world is seen as a sequence of snapshots, or *time slices*.
- A **state** is defined as a set of random variables $\mathbf{X_t}$.
- The variables in $\mathbf{X_t}$ take new values at each time t.
- The variables in $\mathbf{X_t}$ are **unobservable** (or unknown).
- The set of observed variables (**evidence**) at time t is denoted by $\mathbf{E_t}$.
- We assume that time t is discrete, $t \in \{0, 1, 2, 3, \dots\}$.
- We use $X_{a:b}$ to denote the variable X at times $t = a$ to $t = b$. $X_{a:b} = [X_a, X_{a+1}, \dots, X_b]$ and $E_{a:b} = [E_a, E_{a+1}, \dots, E_b]$

In our previous example:

- Each time step corresponds to a day.
- There is one state variable $X_t = \{\text{rain}, \text{no rain}\}$.
- There is one evidence variable $E_t = \{\text{umbrella}, \text{no umbrella}\}$.

Transition model

- The transition model specifies the probability distribution over current state values, given all the previous states: $P(X_t \mid X_{0:t-1})$.
- Problem: $X_{0:t-1}$ is unbounded in size as t increases.
- **Markov assumption**: the last state X_{t-1} has all the information from the past. This defines a first-order Markov chain.

$$P(X_t \mid X_{0:t-1}) = P(X_t \mid X_{t-1}).$$

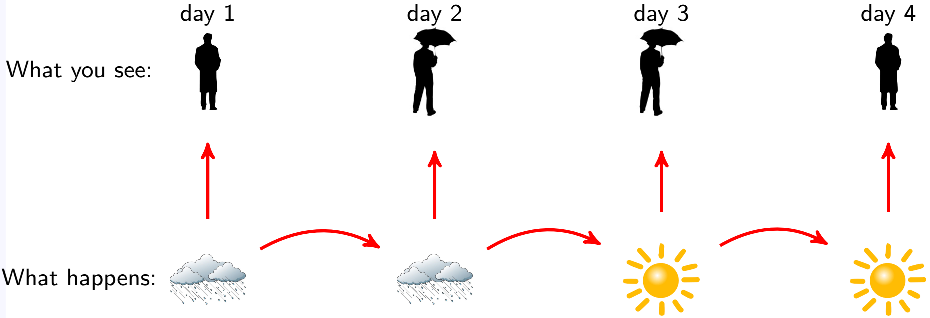
- One can also define a second-order Markov chain by assuming

$$P(X_t \mid X_{0:t-1}) = P(X_t \mid X_{t-1}, X_{t-2}).$$

- Any n-order Markov chain can be reduced to a first-order Markov chain by redefining the states.

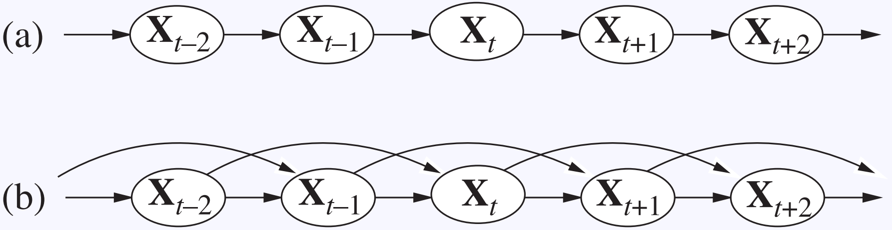
Transition model

The transition model is the probabilities:
P(it rains today | it rained yesterday),
P(it rains today | it didn't rain yesterday).



Transition model

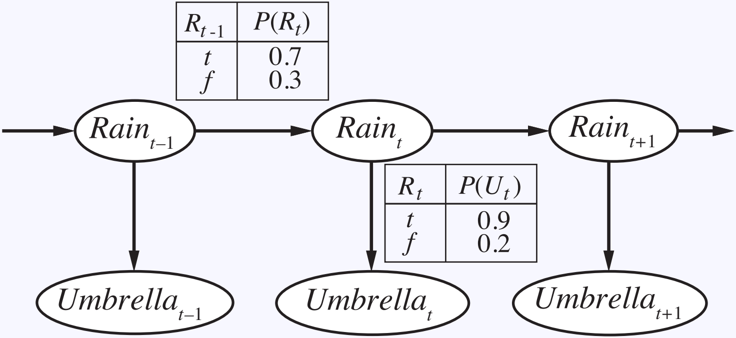
The transition model is a *dynamic* Bayesian network. Previous states are the cause of the current state.



(a) first-order Markov chain, (b) second-order Markov chain

Observation model

- The observation model specifies the probability distribution over current observations, given the current state: $P(E_t \mid X_t)$.
- $P(E_t \mid X_t)$ is also called the **sensor model**.
- The transition and observation models form together a *dynamic* Bayesian network.



Inference problems in temporal models

Inference

The joint probability of a sequence of states $X_{0:t}$ and observations $E_{1:t}$ is given by

$$P(X_{0:t}, E_{1:t}) = P(X_0) \prod_{i=1}^{t} P(X_i \mid X_{i-1}) P(E_i \mid X_i).$$

Inference problems in temporal models

Inference

The joint probability of a sequence of states $X_{0:t}$ and observations $E_{1:t}$ is given by

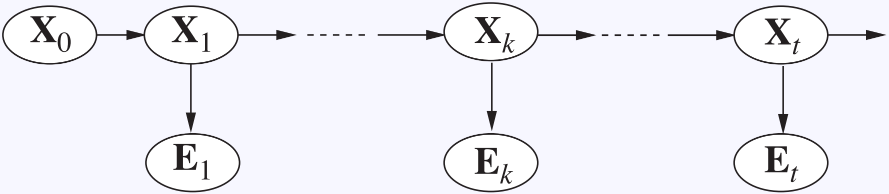
$$P(X_{0:t}, E_{1:t}) = P(X_0) \prod_{i=1}^{t} P(X_i \mid X_{i-1}) P(E_i \mid X_i).$$

Inference problems

- **Filtering** (a.k.a state estimation): Find $P(X_t \mid e_{1:t})$.
- **Prediction**: Find $P(E_{t+1:T} \mid e_{1:t})$.
- **Smoothing**: Find $P(X_{0:t} \mid e_{1:t})$.
- **Most Likely Explanation**: Find $\arg\max_{X_{0:t}} P(X_{0:t} \mid e_{1:t})$.

Filtering (state estimation)

Given all the past and the current observations $e_{1:t} = [e_1, e_2, \ldots e_t]$, we want to compute a distribution on the current state X_t (without knowing the previous states $X_{0:t-1}$).
$P(X_t \mid e_{1:t})$ is know as the **belief state**.



Filtering (state estimation)

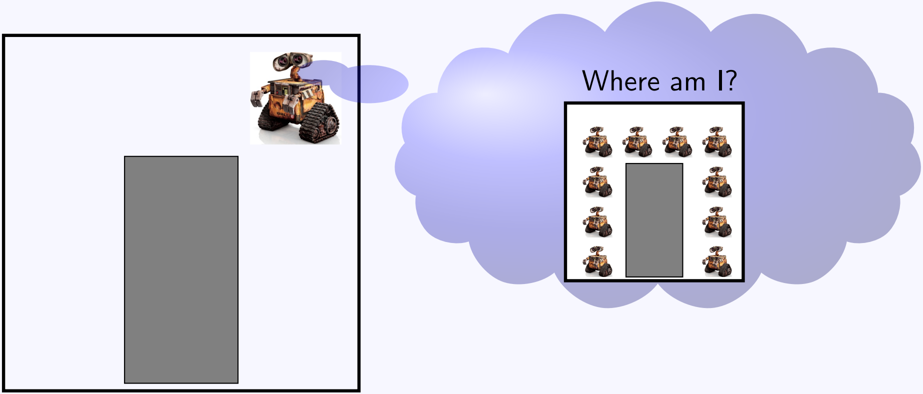


Figure: Initial belief state.

Filtering (state estimation)

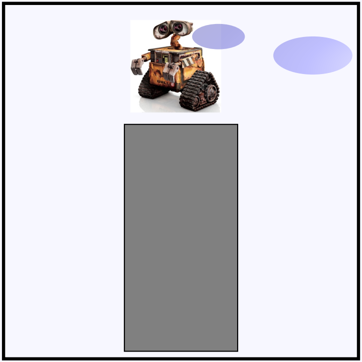


Figure: Belief state after moving left.

Filtering (state estimation)

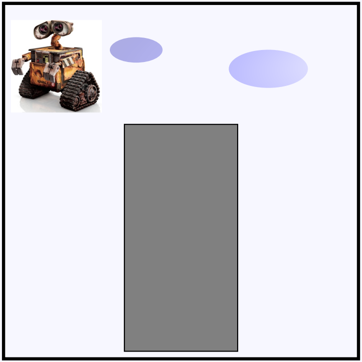


Figure: Belief state after moving left twice.

Filtering (state estimation)

$$P(X_t \mid e_{1:t}) \quad = \quad P(X_t \mid e_{1:t-1}, e_t)$$

Filtering (state estimation)

$$
\begin{aligned}
P(X_t \mid e_{1:t}) &= P(X_t \mid e_{1:t-1}, e_t) \\
&= \frac{P(e_t \mid X_t, e_{1:t-1}) P(X_t \mid e_{1:t-1})}{P(e_t | e_{1:t-1})} \text{ (Bayes' Rule)}
\end{aligned}
$$

Filtering (state estimation)

$$
\begin{aligned}
P(X_t \mid e_{1:t}) &= P(X_t \mid e_{1:t-1}, e_t) \\
&= \frac{P(e_t \mid X_t, e_{1:t-1}) P(X_t \mid e_{1:t-1})}{P(e_t | e_{1:t-1})} \text{ (Bayes' Rule)} \\
&= \frac{P(e_t \mid X_t) P(X_t \mid e_{1:t-1})}{P(e_t | e_{1:t-1})}
\end{aligned}
$$

Filtering (state estimation)

$$
\begin{aligned}
P(X_t \mid e_{1:t}) &= P(X_t \mid e_{1:t-1}, e_t) \\
&= \frac{P(e_t \mid X_t, e_{1:t-1}) P(X_t \mid e_{1:t-1})}{P(e_t|e_{1:t-1})} \text{ (Bayes' Rule)} \\
&= \frac{P(e_t \mid X_t) P(X_t \mid e_{1:t-1})}{P(e_t|e_{1:t-1})} \\
&= \frac{P(e_t \mid X_t) \sum_{x_{t-1}} P(X_t \mid x_{t-1}, e_{1:t-1}) P(x_{t-1} \mid e_{1:t-1})}{P(e_t|e_{1:t-1})}
\end{aligned}
$$

Filtering (state estimation)

$$
\begin{aligned}
P(X_t \mid e_{1:t}) &= P(X_t \mid e_{1:t-1}, e_t) \\
&= \frac{P(e_t \mid X_t, e_{1:t-1}) P(X_t \mid e_{1:t-1})}{P(e_t | e_{1:t-1})} \text{ (Bayes' Rule)} \\
&= \frac{P(e_t \mid X_t) P(X_t \mid e_{1:t-1})}{P(e_t | e_{1:t-1})} \\
&= \frac{P(e_t \mid X_t) \sum_{x_{t-1}} P(X_t \mid x_{t-1}, e_{1:t-1}) P(x_{t-1} \mid e_{1:t-1})}{P(e_t | e_{1:t-1})} \\
&= \frac{P(e_t \mid X_t) \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid e_{1:t-1})}{P(e_t | e_{1:t-1})} \text{ (Markov)}
\end{aligned}
$$

Filtering (state estimation)

$$
\begin{aligned}
P(X_t \mid e_{1:t}) &= P(X_t \mid e_{1:t-1}, e_t) \\
&= \frac{P(e_t \mid X_t, e_{1:t-1}) P(X_t \mid e_{1:t-1})}{P(e_t | e_{1:t-1})} \text{ (Bayes' Rule)} \\
&= \frac{P(e_t \mid X_t) P(X_t \mid e_{1:t-1})}{P(e_t | e_{1:t-1})} \\
&= \frac{P(e_t \mid X_t) \sum_{x_{t-1}} P(X_t \mid x_{t-1}, e_{1:t-1}) P(x_{t-1} \mid e_{1:t-1})}{P(e_t | e_{1:t-1})} \\
&= \frac{P(e_t \mid X_t) \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid e_{1:t-1})}{P(e_t | e_{1:t-1})} \text{ (Markov)} \\
&= \frac{P(e_t \mid X_t) \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid e_{1:t-1})}{\sum_{x_t} P(e_t \mid x_t) \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1} \mid e_{1:t-1})}
\end{aligned}
$$

Filtering (state estimation)

$$
\begin{aligned}
P(X_t \mid e_{1:t}) &= P(X_t \mid e_{1:t-1}, e_t) \\
&= \frac{P(e_t \mid X_t, e_{1:t-1}) P(X_t \mid e_{1:t-1})}{P(e_t|e_{1:t-1})} \text{ (Bayes' Rule)} \\
&= \frac{P(e_t \mid X_t) P(X_t \mid e_{1:t-1})}{P(e_t|e_{1:t-1})} \\
&= \frac{P(e_t \mid X_t) \sum_{x_{t-1}} P(X_t \mid x_{t-1}, e_{1:t-1}) P(x_{t-1} \mid e_{1:t-1})}{P(e_t|e_{1:t-1})} \\
&= \frac{P(e_t \mid X_t) \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid e_{1:t-1})}{P(e_t|e_{1:t-1})} \text{ (Markov)} \\
&= \frac{P(e_t \mid X_t) \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid e_{1:t-1})}{\sum_{x_t} P(e_t \mid x_t) \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1} \mid e_{1:t-1})}
\end{aligned}
$$

$P(X_t \mid e_{1:t})$ can be computed recursively by starting with the prior $P(X_0)$.

Filtering (state estimation)

$$P(X_t \mid e_{1:t}) = \frac{P(e_t \mid X_t) \sum_{x_{t-1}} P(X_t \mid x_{t-1}) P(x_{t-1} \mid e_{1:t-1})}{\sum_{x_t} P(e_t \mid x_t) \sum_{x_{t-1}} P(x_t \mid x_{t-1}) P(x_{t-1} \mid e_{1:t-1})}$$

$P(X_t \mid e_{1:t})$ can be computed recursively by starting with the prior $P(X_0)$.

$$P(X_t \mid e_{1:t}) = \text{FORWARD}\Big(P(X_{t-1} \mid e_{1:t-1}), e_t\Big).$$

Prediction

Let's say we have received a sequence of observations $e_{1:t} = (e_1, e_2, \ldots, e_t)$ and we want to compute $P(E_{t+1:T} \mid e_{1:t})$, the probability distribution over future observations. We have

$$P(E_{t+1:T} \mid e_{1:t}) = \sum_{x_t} P(x_t \mid e_{1:t}) P(E_{t+1:T} \mid x_t),$$

where computing $P(x_t \mid e_{1:t})$ is a filtering problem.
We need then to find how to compute $P(E_{t+1:T} \mid x_t)$.

Prediction

$$\begin{aligned}
P(E_{t+1:T} \mid x_t) &= \sum_{x_{t+1}} P(x_{t+1} \mid x_t) P(E_{t+1:T} \mid x_{t+1}) \text{ (Markov property)} \\
&= \sum_{x_{t+1}} P(x_{t+1} \mid x_t) P(E_{t+1}, E_{t+2:T} \mid x_{t+1}) \\
&= \sum_{x_{t+1}} P(x_{t+1} \mid x_t) P(E_{t+1} \mid x_{t+1}) P(E_{t+2:T} \mid x_{t+1})
\end{aligned}$$

Prediction

$$
\begin{aligned}
P(E_{t+1:T} \mid x_t) &= \sum_{x_{t+1}} P(x_{t+1} \mid x_t) P(E_{t+1:T} \mid x_{t+1}) \text{ (Markov property)} \\
&= \sum_{x_{t+1}} P(x_{t+1} \mid x_t) P(E_{t+1}, E_{t+2:T} \mid x_{t+1}) \\
&= \sum_{x_{t+1}} P(x_{t+1} \mid x_t) P(E_{t+1} \mid x_{t+1}) P(E_{t+2:T} \mid x_{t+1})
\end{aligned}
$$

$P(E_{t+1:T} \mid x_t)$ can be computed recursively, starting at $t = T - 1$ and setting $P(E_{T+1:T} \mid x_T) = 1$.

Prediction

$$P(E_{t+1:T} \mid X_t) = \text{BACKWARD}\Big(P(E_{t+2:T} \mid X_{t+1}), E_{t+1}\Big)$$

$P(E_{t+1:T} \mid x_t)$ can be computed recursively, starting at $t = T - 1$ and setting $P(E_{T+1:T} \mid x_T) = 1$.

Smoothing

Let's say we have received a sequence of observations $e_{1:t} = (e_1, e_2, \ldots, e_t)$ and we want to compute $P(X_k \mid e_{1:t})$, the probability distribution over past states at time $k < t$.

$$
\begin{aligned}
P(X_k \mid e_{1:t}) &= P(X_k \mid e_{1:k}, e_{k+1:t}) \\
&= \frac{P(e_{k+1:t} \mid e_{1:k}, X_k) P(X_k \mid e_{1:k})}{P(e_{k+1:t} \mid e_{1:k})} \text{ (Bayes' Rule)} \\
&= \frac{P(e_{k+1:t} \mid X_k) P(X_k \mid e_{1:k})}{P(e_{k+1:t} \mid e_{1:k})} \\
&= \frac{P(e_{k+1:t} \mid X_k) P(X_k \mid e_{1:k})}{\sum_{x_k} P(e_{k+1:t} \mid x_k) P(x_k \mid e_{1:k})}
\end{aligned}
$$

Smoothing

$$
\begin{aligned}
P(X_k \mid e_{1:t}) &= P(X_k \mid e_{1:k}, e_{k+1:t}) \\
&= \frac{P(e_{k+1:t} \mid e_{1:k}, X_k) P(X_k \mid e_{1:k})}{P(e_{k+1:t} \mid e_{1:k})} \text{ (Bayes' Rule)} \\
&= \frac{P(e_{k+1:t} \mid X_k) P(X_k \mid e_{1:k})}{P(e_{k+1:t} \mid e_{1:k})} \\
&= \frac{P(e_{k+1:t} \mid X_k) P(X_k \mid e_{1:k})}{\sum_{x_k} P(e_{k+1:t} \mid x_k) P(x_k \mid e_{1:k})}
\end{aligned}
$$

Notice that computing

- $P(X_k \mid e_{1:k})$ is a filtering problem (state estimation), which can be done **forward**,
- $P(e_{k+1:t} \mid X_k)$ is a prediction problem, which can be done **backward**.

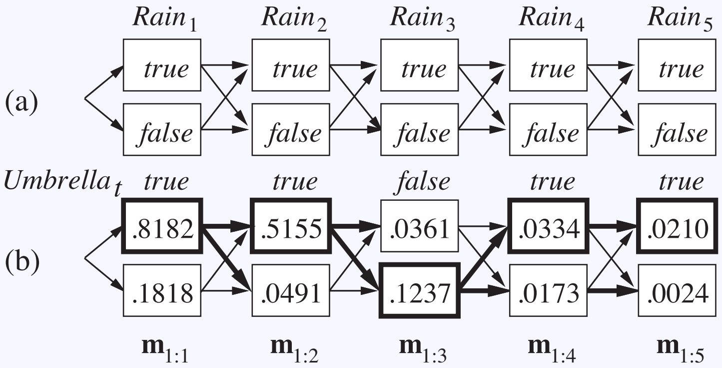
The forward-backward algorithm for smoothing

function FORWARD-BACKWARD(**ev**, $prior$) **returns** a vector of probability distributions
inputs: **ev**, a vector of evidence values for steps $1, \ldots, t$
$prior$, the prior distribution on the initial state, $\mathbf{P}(\mathbf{X}_0)$
local variables: **fv**, a vector of forward messages for steps $0, \ldots, t$
b, a representation of the backward message, initially all 1s
sv, a vector of smoothed estimates for steps $1, \ldots, t$

```
fv[0] ← prior
for i = 1 to t do
    fv[i] ← FORWARD(fv[i − 1], ev[i])
for i = t downto 1 do
    sv[i] ← NORMALIZE(fv[i] × b)
    b ← BACKWARD(b, ev[i])
return sv
```

Most Likely Explanation

Suppose that [true, true, false, true, true] is the umbrella sequence in the previous example. What is the weather sequence most likely to explain this?



Possible state sequences for $Rain_t$ can be viewed as paths through a graph of the possible states at each time step.

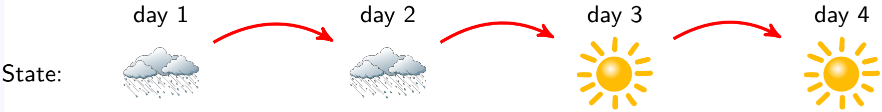
Most Likely Explanation: The Veterbi Algorithm

$$\max_{x_1 \ldots x_t} P(x_1, \ldots, x_t, X_{t+1} \mid e_{1:t+1}) =$$

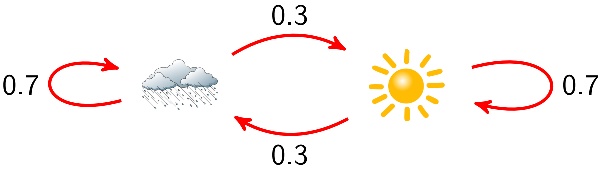
$$\frac{1}{z} P(e_{t+1} \mid X_{t+1}) \max_{x_t} \Big(P(X_{t+1} \mid x_t) max_{x_1 \ldots x_{t-1}} P(x_1, \ldots, x_{t-1}, x_t \mid e_{1:t}) \Big)$$

Markov Chain

A Markov Chain is a temporal model where the state is a single random variable that is always known.

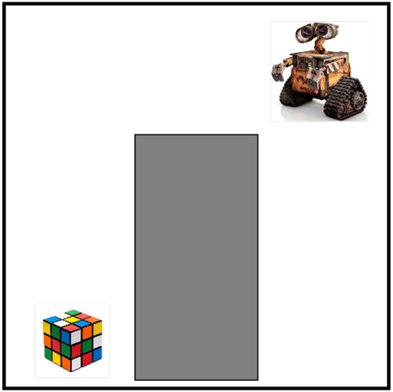


Transition Function

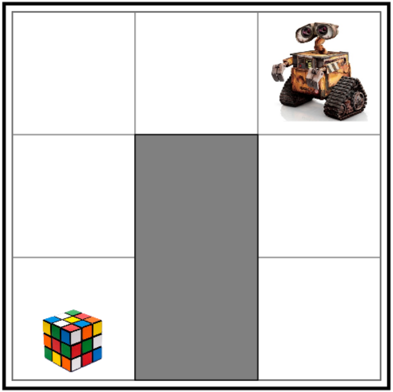


Weather remains the same with probability 0.7, and changes with probability 0.3.

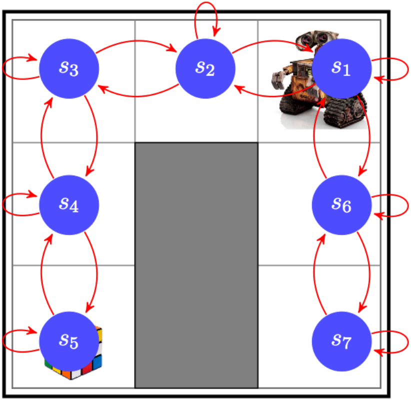
Example of a Markov Chain: Robot searching for an object



Example of a Markov Chain: Robot searching for an object



Example of a Markov Chain: Robot searching for an object

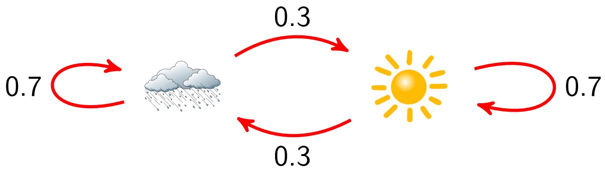


Transition Matrix

Let $\{s^i\}$ be the states of a Markov chain.
The transition function can be represented as a matrix T, where

$$T[i,j] = P(s_{t+1} = s^j \mid s_t = s^i).$$

Example



$$T = \begin{bmatrix} P(rainy \mid rainy) & P(sunny \mid rainy) \\ P(rainy \mid sunny) & P(sunny \mid sunny) \end{bmatrix} = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$$

Transition Matrix

The transition matrix can be used to easily compute the state distributions in the future.
Let f_t denote the state distribution at time t, i.e. $f_t[i] = P(s_t = s^i)$, then

$$\begin{aligned} f_t &= f_0 \underbrace{T \times T \times T \times \cdots \times T}_{t \text{ times}} \\ &= f_{t-1} T. \end{aligned}$$

Example

$$f_0 = \begin{bmatrix} 0.8 & 0.2 \end{bmatrix}, T = \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}$$

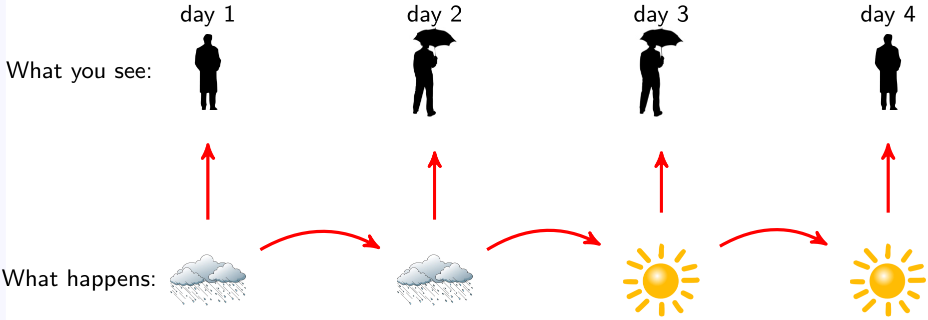
$$f_1 = f_0 T = \begin{bmatrix} 0.8 & 0.2 \end{bmatrix} \times \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix} = \begin{bmatrix} 0.62 & 0.38 \end{bmatrix},$$

$$f_2 = f_0 T T = f_1 T = \begin{bmatrix} 0.62 & 0.38 \end{bmatrix} \times \begin{bmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{bmatrix} = \begin{bmatrix} 0.548 & 0.452 \end{bmatrix}.$$

Hidden Markov Model (HMM)

A Hidden Markov Model (HMM) is a temporal model where the state is a single random variable that is unknown (hidden). An observable variable is used as evidence to infer the state.

Example

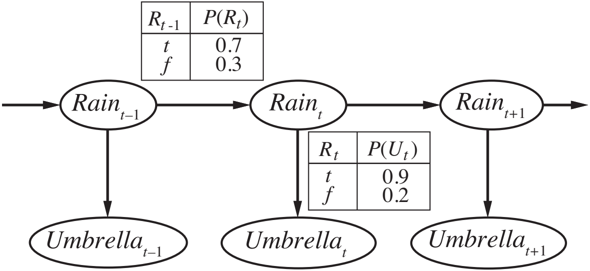


In addition to the transition matrix, we define an observation matrix O_e for each possible value e of the evidence variable E.
Observation matrix O_e has zeros everywhere except on the diagonal, where $O_e[i, i] = P(E_t = e \mid s_t = s^i)$.

$$O_e = \begin{bmatrix} P(e \mid s^0) & 0 & 0 & \dots & 0 \\ 0 & P(e \mid s^1) & 0 & \dots & 0 \\ 0 & 0 & P(e \mid s^2) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & P(e \mid s^n) \end{bmatrix}$$

Hidden Markov Model (HMM)

Example



$$O_{\text{umbrella } true} = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.2 \end{bmatrix},$$

$$O_{\text{umbrella } false} = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.8 \end{bmatrix}.$$

What is the advantage of using vector and matrix notations?

- Vector and matrix notations make the calculations simple and elegant.
- Forward: Let $f_t[i] = P(s_t = s^i)$, then

$$
\begin{aligned}
f_{t+1} &= \alpha f_t T O_{e_{t+1}}, \\
\alpha &= \left(f_t T O_{e_{t+1}} \mathbf{1}\right)^{-1} \\
\mathbf{1} &= \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}.
\end{aligned}
$$

Hidden Markov Model (HMM)

What is the advantage of using vector and matrix notations?

- Vector and matrix notations make the calculations simple and elegant.
- Backward: Let $b_t[i] = P(e_{t+1:k} \mid s_t = s^i)$, then

$$b_t = TO_{e_{t+1}} b_{t+1}.$$

where we start from the last time-step k with $b_k = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$.