Lecture 13: Sequence Modeling, HMMs CS486/686 Intro to Artificial Intelligence

2026-2-24

Pascal Poupart
David R. Cheriton School of Computer Science

UNIVERSITY OF WATERLOO

Outline

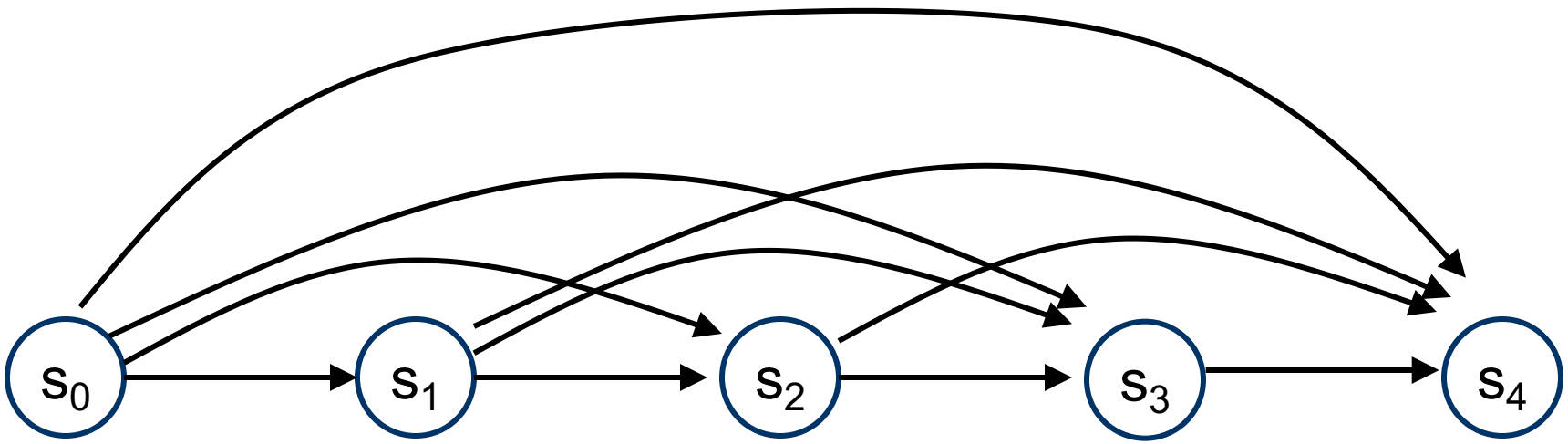
- Reasoning over time
  - Time series, speech, text
  - Dynamic inference
- Sequence modeling
  - **Markov Processes**
  - **Hidden Markov Models (HMMs)**
  - Dynamic Bayesian Networks (DBNs)
  - Recurrent Neural Networks (next lecture)
  - Transformers (next lecture)

Static Inference

- So far…
  - Assume the world doesn't change
  - Static probability distribution
  - Ex: when repairing a car, whatever is broken remains broken during the diagnosis
- But the world evolves over time…
  - How can we use probabilistic inference for weather predictions, stock market predictions, patient monitoring, etc?

Stochastic Process

- Definition
  - Set of States: S
  - Stochastic dynamics: $Pr(s_t|s_{t-1}, \ldots, s_0)$



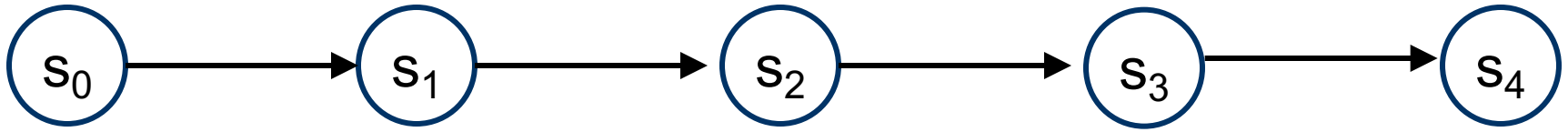
  - Can be viewed as a Bayes net with one random variable per time slice

Stochastic Process

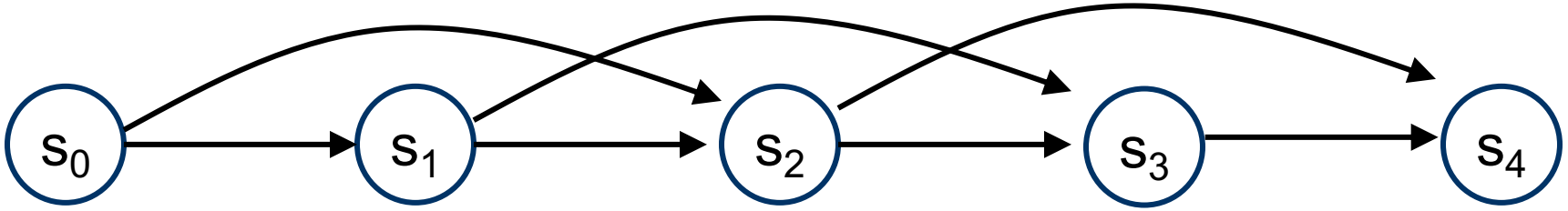
- Problems:
  - Infinitely many variables
  - Infinitely large conditional probability tables
- Solutions:
  - Stationary process: dynamics do not change over time
  - Markov assumption: current state depends only on a finite history of past states

K-order Markov Process

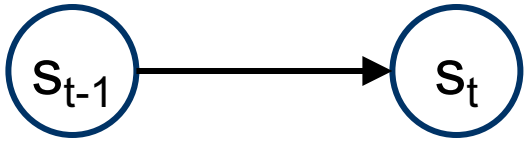
- Assumption: last k states sufficient
- First-order Markov Process
  - $Pr(s_t|s_{t-1}, ..., s_o) = Pr(s_t|s_{t-1})$



- Second-order Markov Process
  - $Pr(s_t|s_{t-1}, ..., s_o) = Pr(s_t|s_{t-1}, s_{t-2})$



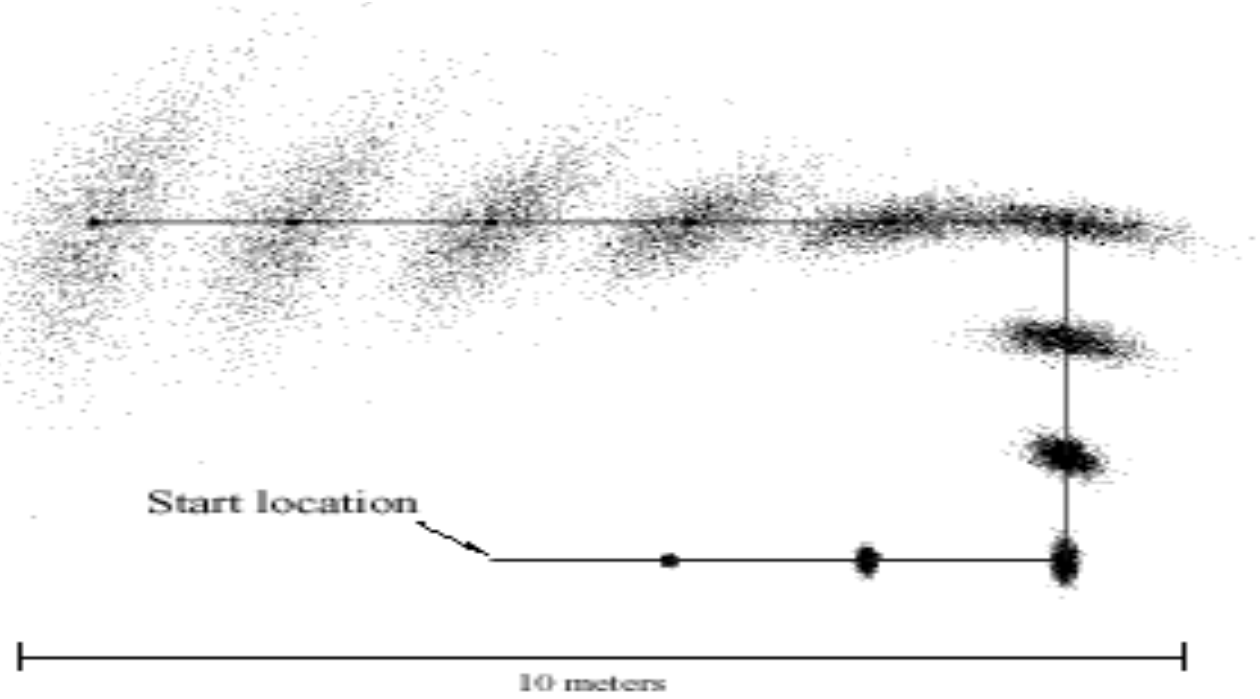
K-order Markov Process

- Advantage:
  - Can specify entire process with finitely many time slices
- Two slices sufficient for a first-order Markov process…
  - Graph: 

  - Dynamics: $Pr(s_t|s_{t-1})$
  - Prior: $Pr(s_o)$

Mobile Robot Localisation

- Example of a first-order Markov process



- Problem: uncertainty grows over time…

Hidden Markov Models

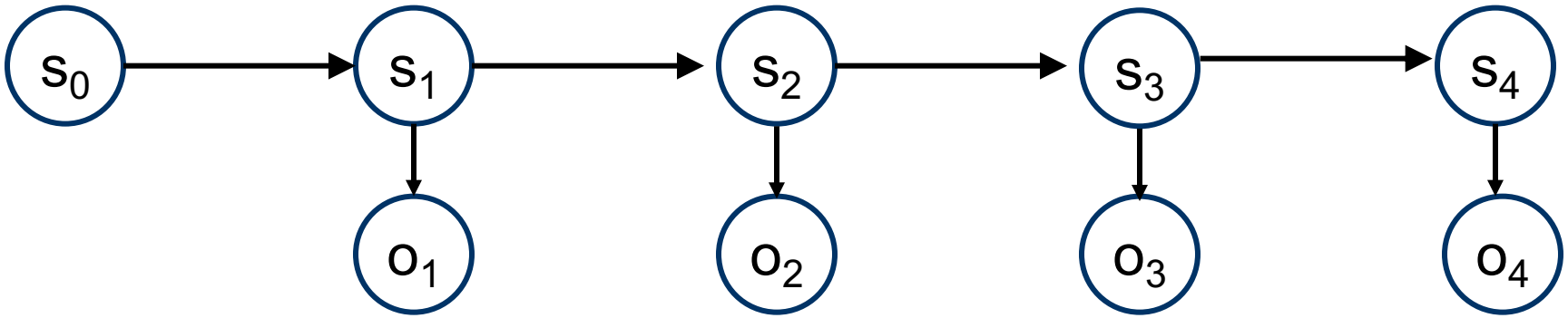
- Robot could use sensors to reduce location uncertainty…

- In general:
  - States not directly observable, hence uncertainty captured by a distribution
  - Uncertain dynamics increase state uncertainty
  - Observations made via sensors reduce state uncertainty

- Solution: Hidden Markov Model

First-order Hidden Markov Model

- Definition:
  - Set of states: S
  - Set of observations: O
  - Transition model: $Pr(s_t|s_{t-1})$
  - Observation model: $Pr(o_t|s_t)$
  - Prior: $Pr(s_o)$



Mobile Robot Localisation

- (First-order) Hidden Markov Model:
  - S: (x,y) coordinates of the robot on a map
  - O: distances to surrounding obstacles (measured by laser range finders or sonars)
  - $Pr(s_t|s_{t-1})$: movement of the robot with uncertainty
  - $Pr(o_t|s_t)$: uncertainty in the measurements provided by laser range finders and sonars

- Localisation corresponds to the query: $Pr(s_t|o_t, ..., o_1)$?

Inference in temporal models

- Four common tasks:
  - Monitoring: $Pr(s_t|o_t, ..., o_1)$
  - Prediction: $Pr(s_{t+k}|o_t, ..., o_1)$
  - Hindsight: $Pr(s_k|o_t, ..., o_1)$ where $k < t$
  - Most likely explanation: $\text{argmax}_{st,...,s1} Pr(s_t, ..., s_1|o_t, ..., o_1)$

- What algorithms should we use?
  - First 3 tasks can be done with variable elimination and 4^{th} task with a variant of variable elimination

Monitoring

- $Pr(s_t|o_t, ..., o_1)$: distribution over current state given observations

- Examples: robot localisation, patient monitoring

- Forward algorithm: corresponds to variable elimination
  - Factors: $Pr(s_o)$, $Pr(s_i|s_{i-1})$, $Pr(o_i|s_i)$, $1 \le i \le t$
  - Restrict $o_1, ..., o_t$ to the observations made
  - Summout $s_o, ..., s_{t-1}$
  - $\sum_{s_0 ... s_{t-1}} Pr(s_0) \prod_{1 \le i \le t} Pr(s_i|s_{i-1}) Pr(o_i|s_i)$

Prediction

- $Pr(s_{t+k}|o_t, ..., o_1)$: distribution over future state given observations

- Examples: weather prediction, stock market prediction

- Forward algorithm: corresponds to variable elimination
  - Factors: $Pr(s_o)$, $Pr(s_i|s_{i-1})$ $1 \le i \le t+k$, $Pr(o_j|s_j)$ $1 \le j \le t$
  - Restrict $o_1, ..., o_t$ to the observations made
  - Summout $s_o, ..., s_{t+k-1}$
  - $\sum_{s_0 ... s_{t+k-1}} Pr(s_0) \prod_{1 \le i \le t+k} Pr(s_i|s_{i-1}) \prod_{1 \le j \le t} Pr(o_j|s_j)$

Hindsight

- $Pr(s_k|o_t, ..., o_1)$ for k<t: distribution over a past state given observations

- Example: delayed activity/speech recognition

- Forward-backward algorithm: corresponds to variable elimination
  - Factors: $Pr(s_o)$, $Pr(s_i|s_{i-1})$, $Pr(o_i|s_i)$, $1 \le i \le t$
  - Restrict $o_1, ..., o_t$ to the observations made
  - Summout $s_0, s_1, ..., s_{k-1}, s_t, s_{t-1}, ..., s_{k+1}$
  - $\sum_{s_0,s_1,...,s_{k-1},s_t,s_{t-1},...,s_{k+1}} \Pr(s_0) \prod_{1 \le i \le t} \Pr(s_i|s_{i-1}) \Pr(o_i|s_i)$

Most likely explanation

- $\text{Argmax}_{s0...st}\ \Pr(s_o,...,s_t|o_t, ..., o_1)$: most likely state sequence given observations

- Example: speech recognition

- Viterbi algorithm: corresponds to a variant of variable elimination
  - Factors: $\Pr(s_o)$, $\Pr(s_i|s_{i-1})$, $\Pr(o_i|s_i)$, $1 \le i \le t$
  - Restrict $o_1, ..., o_t$ to the observations made
  - Maxout $s_o, ..., s_t$
  - $\max_{s_0 ... s_t} \Pr(s_0) \prod_{1 \le i \le t} \Pr(s_i|s_{i-1}) \Pr(o_i|s_i)$

Complexity of temporal inference

- Hidden Markov Models are Bayes nets with a polytree structure

- Hence, variable elimination is
  - Linear with respect to # of time slices
  - Linear with respect to largest conditional probability table ($Pr(s_t|s_{t-1})$ or $Pr(o_t|s_t)$)

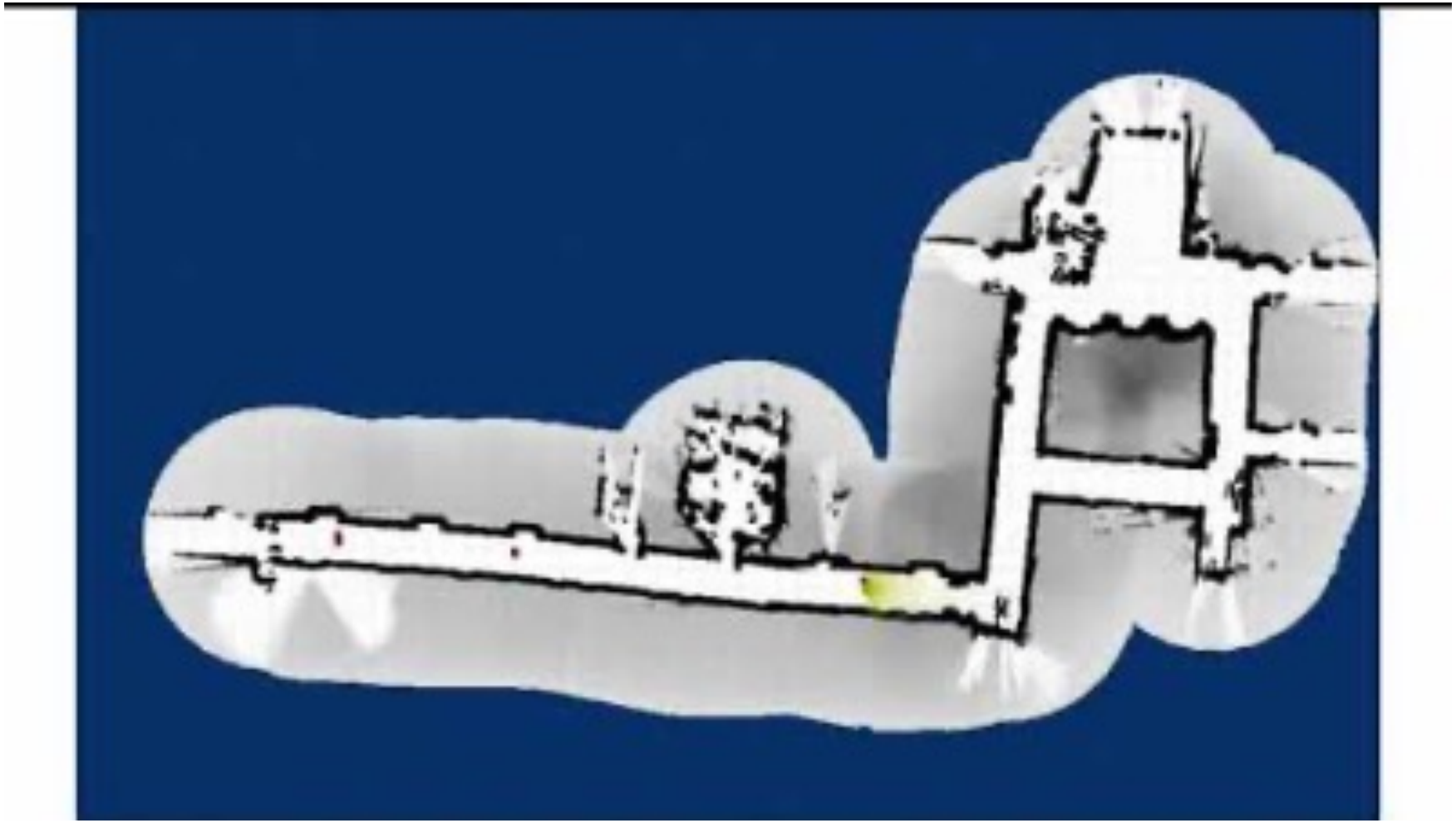
Probabilistic Inference

- Applications of static and temporal inference are virtually limitless

- Some examples:
  - mobile robot navigation
  - vacuum cleaners
  - speech recognition
  - patient monitoring
  - weather prediction
  - fault diagnosis in Mars rovers
  - etc.

Robot localisation

- Demo at 1:15



Localization and Mapping in Robotic Vacuums

Neato Robotics

Uses particle filtering (approximate inference technique based on sampling) for simultaneous localisation and mapping

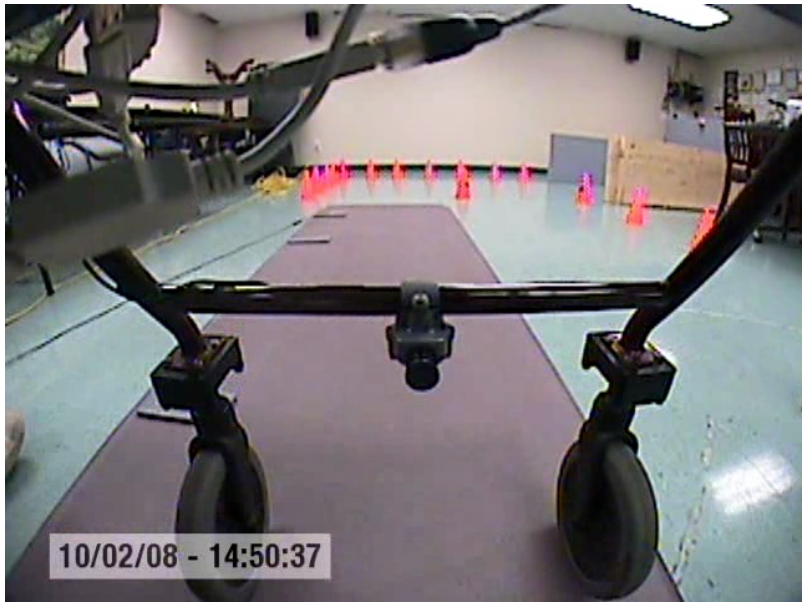
See patent: http://www.faqs.org/patents/assignee/neato-robotics-inc/

Case Study: Activity Recognition

- Task: infer activities performed by a user of a smart walker
  - Inputs: sensor measurements
  - Output: activity

Backward view

Forward view



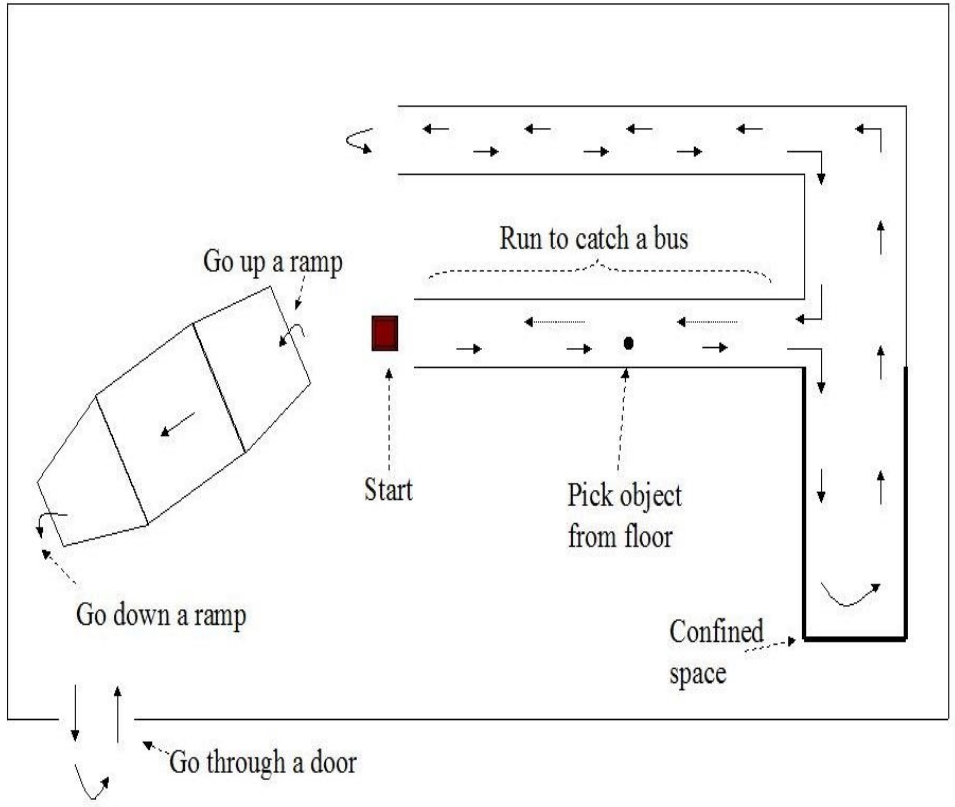
Inputs: Raw Sensor Data

- 8 channels:
  - Forward acceleration
  - Lateral acceleration
  - Vertical acceleration
  - Load on left rear wheel
  - Load on right rear wheel
  - Load on left front wheel
  - Load on right front wheel
  - Wheel rotation counts (speed)

- Data recorded at 50 Hz and digitized (16 bits)

Data Collection

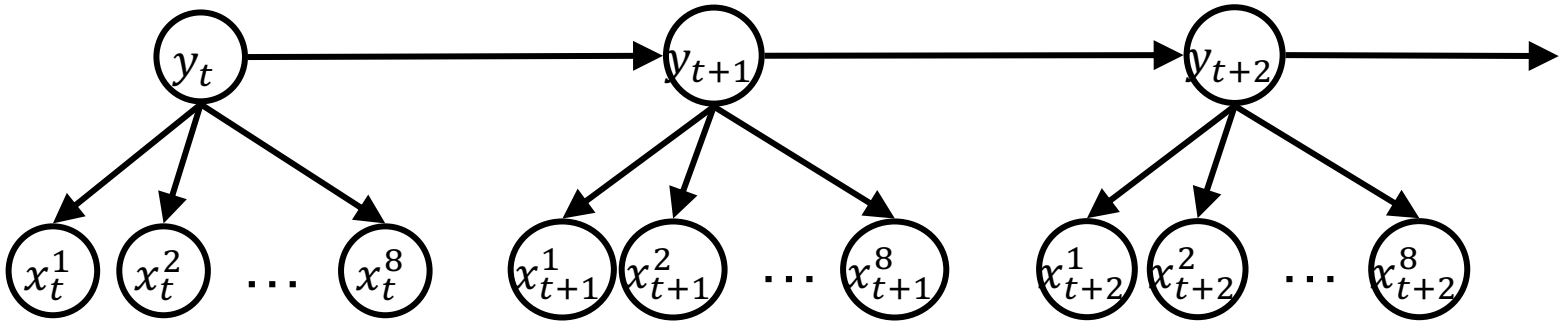
- 8 walker users at Winston Park (84-97 years old)
- 12 older adults (80-89 years old) in the KW area who do not use walkers



Output: Activities

- Not Touching Walker (NTW)
- Standing (ST)
- Walking Forward (WF)
- Turning Left (TL)
- Turning Right (TR)
- Walking Backwards (WB)
- Sitting on the Walker (SW)
- Reaching Tasks (RT)
- Up Ramp/Curb (UR/UC)
- Down Ramp/Curb (DR/DC)

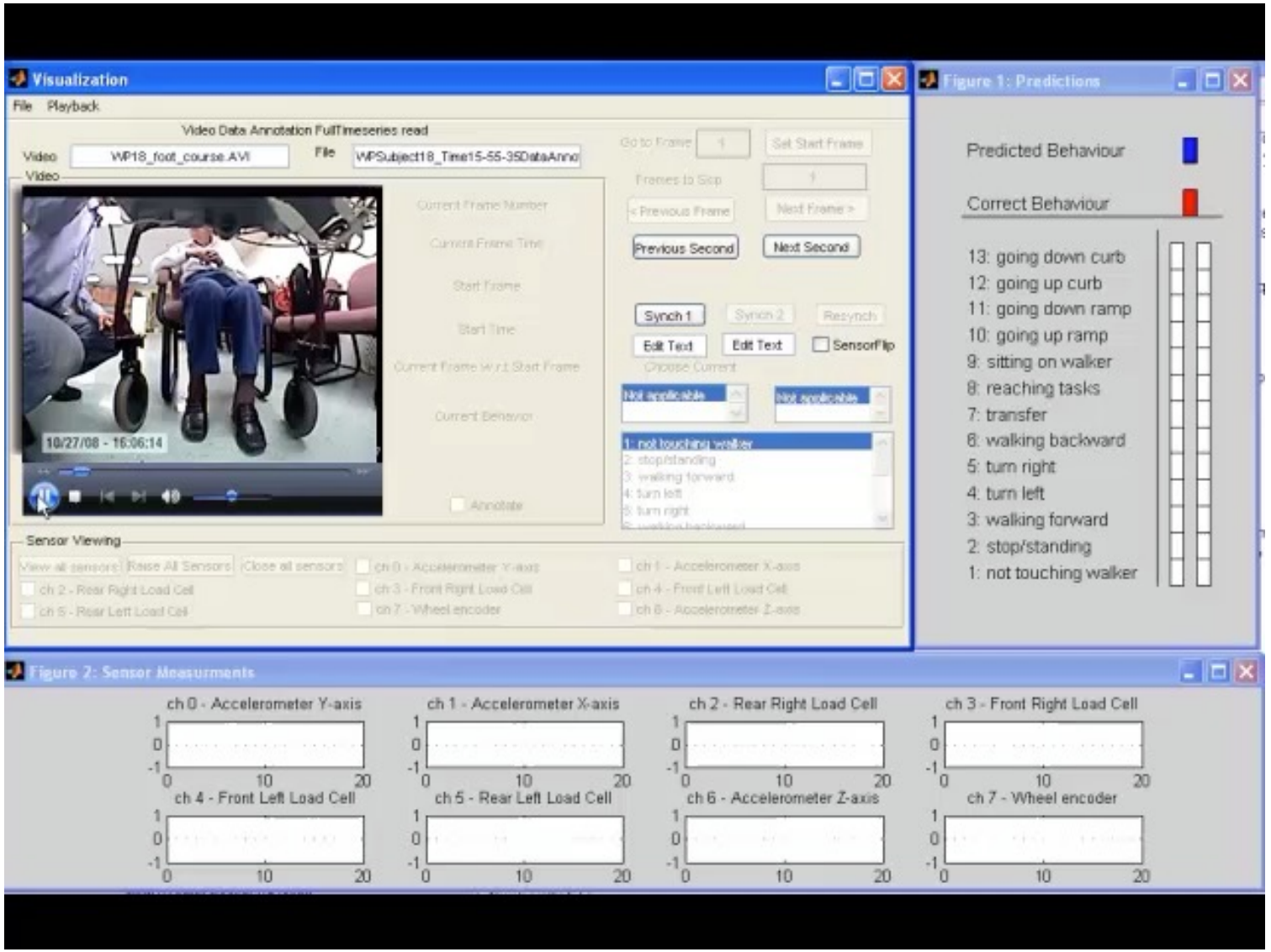
Hidden Markov Model (HMM)



- Parameters
  - Initial state distribution: $\pi_{class} = \Pr(y_1 = class)$
  - Transition probabilities: $\theta_{class'|class} = \Pr(y_{t+1} = class' | y_t = class)$
  - Emission probabilities: $\phi^i_{val|class} = \Pr(x_t^i = val | y_t = class)$

    or $N(val|\mu^i_{class}, \sigma^i_{class}) = \Pr(x_t^i = val | y_t = class)$
- Maximum likelihood:
  - Supervised: $\pi^*, \theta^*, \phi^* = argmax_{\pi,\theta,\phi} \Pr(y_{1:T}, x_{1:T} | \pi, \theta, \phi)$
  - Unsupervised: $\pi^*, \theta^*, \phi^* = argmax_{\pi,\theta,\phi} \Pr(x_{1:T} | \pi, \theta, \phi)$

Demo


Visualization
File Playback
Video Data Annotation FullTimeseries read
Video
WP18_foot_course.AVI
File
WPSubject18_Time15-55-35DataAnno
Video
10/27/08 - 16:06:14
Previous Second
Next Second
Synch 1
Synch 2
Resynch
Edit Text
Edit Text
SensorFlip
Annotate
Sensor Viewing
Figure 1: Predictions
Predicted Behaviour
Correct Behaviour
13: going down curb
12: going up curb
11: going down ramp
10: going up ramp
9: sitting on walker
8: reaching tasks
7: transfer
6: walking backward
5: turn right
4: turn left
3: walking forward
2: stop/standing
1: not touching walker
Figure 2: Sensor Measurments
ch 0 - Accelerometer Y-axis
ch 1 - Accelerometer X-axis
ch 2 - Rear Right Load Cell
ch 3 - Front Right Load Cell
ch 4 - Front Left Load Cell
ch 5 - Rear Left Load Cell
ch 6 - Accelerometer Z-axis
ch 7 - Wheel encoder

Maximum Likelihood

- Supervised Learning: y's are known
- Objective: $argmax_{\pi,\theta,\phi} \Pr(y_{1..t}, x_{1..t} | \pi, \theta, \phi)$

- Derivation:
  - Set derivative to 0
  - Isolate parameters π, θ, ϕ

- Consider a single input x per time step
- Let $y \in \{c_1, c_2\}$ and $x \in \{v_1, v_2\}$

Multinomial Emissions

- Let $\#c_i^{start}$ be # times of that process **starts** in class c_i
- Let $\#c_i$ be # of times that process is in class c_i
- Let $\#(c_i, c_j)$ be # of times that c_i follows c_j
- Let $\#(v_i, c_j)$ be # of times that v_i occurs with c_j
- $\Pr(y_{0..t}, x_{1..t})$

$$= \Pr(y_0) \prod_{i=1}^{t} \Pr(y_i|y_{i-1}) \Pr(x_i|y_i)$$

$$= (\pi_{c_1})^{\#c_1^{start}} (1-\pi_{c_1})^{\#c_2^{start}} (\theta_{c_1|c_1})^{\#(c_1,c_1)} (1-\theta_{c_1|c_1})^{\#(c_2,c_1)} (\theta_{c_1|c_2})^{\#(c_1,c_2)} (1-\theta_{c_1|c_2})^{\#(c_2,c_2)}$$
$$(\phi_{v_1|c_1})^{\#(v_1,c_1)} (1-\phi_{v_1|c_1})^{\#(v_2,c_1)} (\phi_{v_1|c_2})^{\#(v_1,c_2)} (1-\phi_{v_1|c_2})^{\#(v_2,c_2)}$$

Multinomial Emissions

- $argmax_{\pi,\theta,\phi} \Pr(y_{1..t}, x_{1..t}|\pi,\theta,\phi)$

$$\Rightarrow \begin{cases} argmax_{\pi_{c_1}} (\pi_{c_1})^{\#c_1^{start}} (1-\pi_{c_1})^{\#c_2^{start}} \\ argmax_{\theta_{c_1|c_1}} (\theta_{c_1|c_1})^{\#(c_1,c_1)} (1-\theta_{c_1|c_1})^{\#(c_2,c_1)} \\ argmax_{\theta_{c_1|c_2}} (\theta_{c_1|c_2})^{\#(c_1,c_2)} (1-\theta_{c_1|c_2})^{\#(c_2,c_2)} \\ argmax_{\phi_{v_1|c_1}} (\phi_{v_1|c_1})^{\#(v_1,c_1)} (1-\phi_{v_1|c_1})^{\#(v_2,c_1)} \\ argmax_{\phi_{v_1|c_2}} (\phi_{v_1|c_2})^{\#(v_1,c_2)} (1-\phi_{v_1|c_2})^{\#(v_2,c_2)} \end{cases}$$

Multinomial Emissions

- Optimization problem:

$$\underset{\pi_{c_1}}{\text{argmax}} \left(\pi_{c_1}\right)^{\#c_1^{start}} \left(1-\pi_{c_1}\right)^{\#c_2^{start}}$$

$$= \underset{\pi_{c_1}}{\text{argmax}} \ (\#c_1^{start})\log(\pi_{c_1}) + (\#c_2^{start})\log(1-\pi_{c_1})$$

- Set derivative to 0:

$$0 = \frac{\#c_1^{start}}{\pi_{c_1}} - \frac{\#c_2^{start}}{1-\pi_{c_1}}$$

$$\Longrightarrow (1-\pi_{c_1})(\#c_1^{start}) = \left(\pi_{c_1}\right)(\#c_2^{start})$$

$$\Longrightarrow \pi_{c_1} = \frac{\#c_1^{start}}{\#c_1^{start}+\#c_2^{start}}$$

Relative Frequency Counts

- Maximum likelihood solution

$$\pi_{c_1^{start}} = \#c_1^{start} / (\#c_1^{start} + \#c_2^{start})$$

$$\theta_{c_1|c_1} = \#(c_1, c_1) / (\#(c_1, c_1) + \#(c_2, c_1))$$

$$\theta_{c_1|c_2} = \#(c_1, c_2) / (\#(c_1, c_2) + \#(c_2, c_2))$$

$$\phi_{v_1|c_1} = \#(v_1, c_1) / (\#(v_1, c_1) + \#(v_2, c_1))$$

$$\phi_{v_1|c_2} = \#(v_1, c_2) / (\#(v_1, c_2) + \#(v_2, c_2))$$

Gaussian Emissions

- Maximum likelihood solution

$$\pi_{c_1^{start}} = \#c_1^{start} / (\#c_1^{start} + \#c_2^{start})$$

$$\theta_{c_1|c_1} = \#(c_1, c_1) / (\#(c_1, c_1) + \#(c_2, c_1))$$

$$\theta_{c_1|c_2} = \#(c_1, c_2) / (\#(c_1, c_2) + \#(c_2, c_2))$$

$$\mu_{c_1} = \frac{1}{\#c_1} \sum_{\{t|y_t=c_1\}} x_t, \qquad \sigma_{c_1}^2 = \frac{1}{\#c_1} \sum_{\{t|y_t=c_1\}} (x_t - \mu_{c_1})^2$$

$$\mu_{c_2} = \frac{1}{\#c_2} \sum_{\{t|y_t=c_2\}} x_t, \qquad \sigma_{c_2}^2 = \frac{1}{\#c_2} \sum_{\{t|y_t=c_2\}} (x_t - \mu_{c_2})^2$$