Lecture 14: Hidden Markov Models
CS480/680 Intro to Machine Learning

2023-3-2

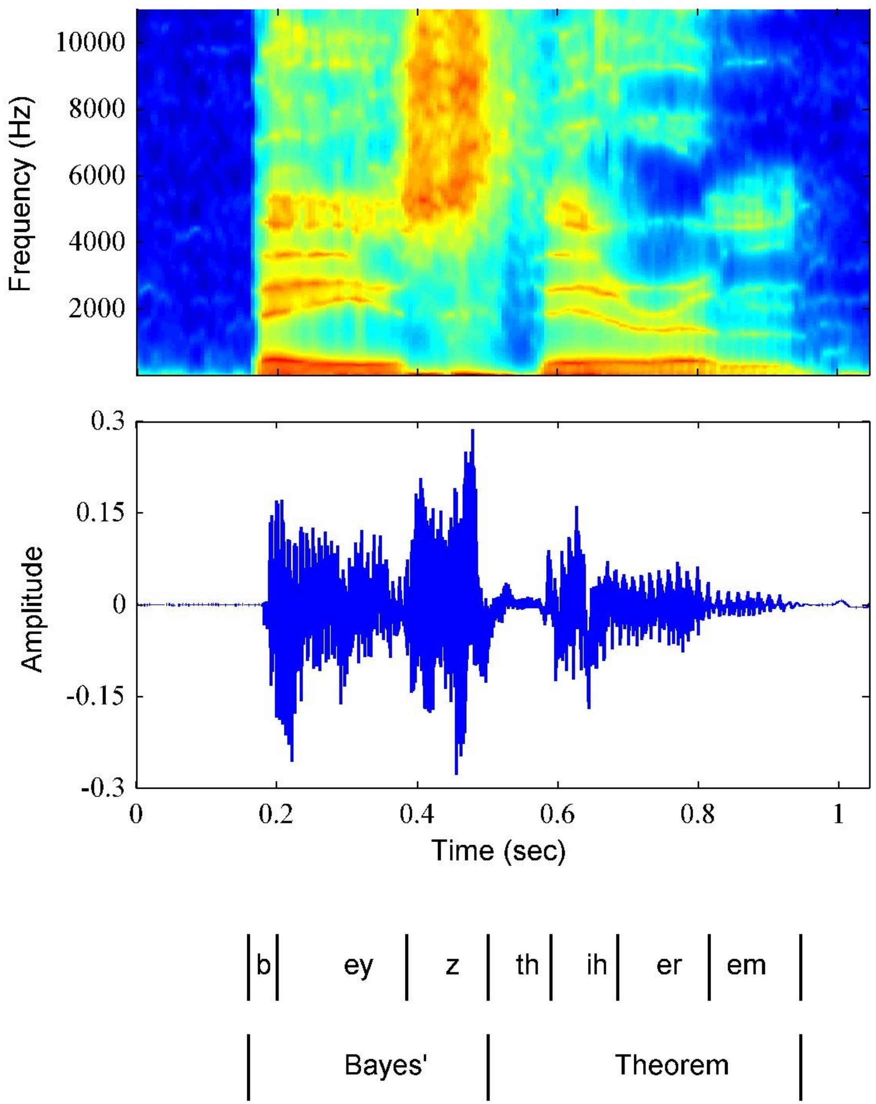
Pascal Poupart
David R. Cheriton School of Computer Science

UNIVERSITY OF WATERLOO

Sequence Data

- So far, we assumed that the data instances are classified independently
  - More precisely, we assumed that the data is iid (identically and independently distributed)
    - E.g., text categorization, digit recognition in separate images, etc.
- In many applications, the data arrives sequentially and the classes are correlated
  - E.g., weather prediction, robot localization, speech recognition, activity recognition

Speech Recognition



Classification

- Extension of some classification models for sequence data

	Independent classification	Correlated classification
Generative models	Mixture of Gaussians	**Hidden Markov Model**
Discriminative models	Logistic Regression	**Conditional Random Field**
Discriminative models	Feed Forward Neural Network	**Recurrent Neural Network**

Hidden Markov Models (HMMs)

Mixture of Gaussians

HMMs

Assumptions

- **Stationary Process**: transition and emission distributions are identical at each time step

$$\Pr(x_t|y_t) = \Pr(x_{t+1}|y_{t+1}) \quad \forall t$$

$$\Pr(y_t|y_{t-1}) = \Pr(y_{t+1}|y_t) \quad \forall t$$

- **Markovian Process**: next state is independent of previous states given the current state

$$\Pr(y_{t+1}|y_t, y_{t-1}, \dots, y_1) = \Pr(y_{t+1}|y_t) \quad \forall t$$

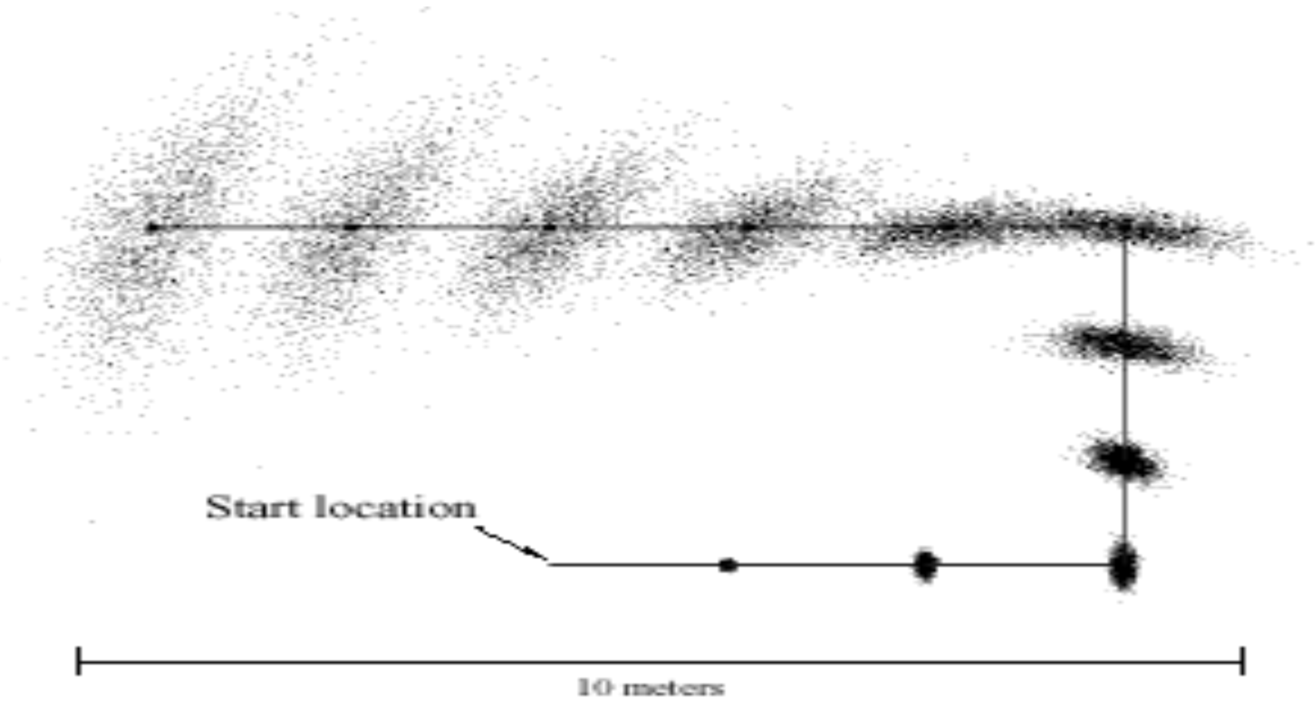
Hidden Markov Model

- Graphical Model

- Parameterization
  - Initial distribution:
  - Transition distribution:
  - Emission distribution:
- Joint distribution:

Mobile Robot Localisation

- Example of a Markov process
- Problem: uncertainty grows over time...



Mobile Robot Localisation

- Hidden Markov Model:
  $\boldsymbol{y}$: coordinates of the robot on a map
  $\boldsymbol{x}$: distances to obstacles (measured by laser range finders or sonars)
  $\mathbf{Pr}(\boldsymbol{y}_t|\boldsymbol{y}_{t-1})$: movement of the robot with uncertainty
  $\mathbf{Pr}(\boldsymbol{x}_t|\boldsymbol{y}_t)$: uncertainty in measurements by laser range finders and sonars

- **Localisation:** $\Pr(y_t|x_t, \ldots, x_1)$?

Inference in temporal models

- Four common tasks:
  - **Monitoring:** $\Pr(y_t|x_{1..t})$
  - **Prediction:** $\Pr(y_{t+k}|x_{1..t})$
  - **Hindsight:** $\Pr(y_k|x_{1..t})$ where $k < t$
  - **Most likely explanation:** $argmax_{y_1,\ldots,y_t} \Pr(y_{1..t}|x_{1..t})$

- What algorithms should we use?

Monitoring

- $\Pr(y_t|x_{1..t})$: distribution over current state given observations
- Examples: robot localisation, patient monitoring
- Recursive computation:

$\Pr(y_t|x_{1..t}) \propto \Pr(x_t|y_t, x_{1..t-1})\Pr(y_t|x_{1..t-1})$ by Bayes' theorem

$= \Pr(x_t|y_t)\Pr(y_t|x_{1..t-1})$ by conditional independence

$= \Pr(x_t|y_t)\sum_{y_{t-1}}\Pr(y_t, y_{t-1}|x_{1..t-1})$ by marginalization

$= \Pr(x_t|y_t)\sum_{y_{t-1}}\Pr(y_t|y_{t-1}, x_{1..t-1})\Pr(y_{t-1}|x_{1..t-1})$ by chain rule

$= \Pr(x_t|y_t)\sum_{y_{t-1}}\Pr(y_t|y_{t-1})\Pr(y_{t-1}|x_{1..t-1})$ by conditional independence

Forward Algorithm

- Compute $\Pr(y_t|x_{1..t})$ by forward computation

$\Pr(y_1|x_1) \propto \Pr(x_1|y_1)\Pr(y_1)$
For $i = 2$ to t do
$\quad \Pr(y_i|x_{1..i}) \propto \Pr(x_i|y_i) \sum_{y_{i-1}} \Pr(y_i|y_{i-1}) \Pr(y_{i-1}|x_{1..i-1})$
End

- Linear complexity in t

Prediction

- $\Pr(y_{t+k}|x_{1..t})$: distribution over future state given observations
- Examples: weather prediction, stock market prediction

- Recursive computation

$\Pr(y_{t+k}|x_{1..t}) = \sum_{y_{t+k-1}} \Pr(y_{t+k}, y_{t+k-1}|x_{1..t})$ by marginalization

$= \sum_{y_{t+k-1}} \Pr(y_{t+k}|y_{t+k-1}, x_{1..t}) \Pr(y_{t+k-1}|x_{1..t})$ by chain rule

$= \sum_{y_{t+k-1}} \Pr(y_{t+k}|y_{t+k-1}) \Pr(y_{t+k-1}|x_{1..t})$ by conditional independence

Forward Algorithm

1. Compute $\Pr(y_t|x_{1..t})$ by forward computation

   $\Pr(y_1|x_1) \propto \Pr(x_1|y_1)\Pr(y_1)$
   For $i = 1$ to t do

   $$\Pr(y_i|x_{1..i}) \propto \Pr(x_i|y_i)\sum_{y_{i-1}}\Pr(y_i|y_{i-1})\Pr(y_{i-1}|x_{1..i-1})$$

2. Compute $\Pr(y_{t+k}|x_{1..t})$ by forward computation

   For $j = 1$ to k do

   $$\Pr\left(y_{t+j}\middle|x_{1..t}\right) = \sum_{y_{t+j-1}}\Pr\left(y_{t+j}\middle|y_{t+j-1}\right)\Pr(y_{t+j-1}|x_{1..t})$$

- Linear complexity in $t + k$

Hindsight

- $\Pr(y_k|x_{1..t})$ for $k < t$: distribution over a past state given observations
- Example: delayed activity/speech recognition
- Computation:

$$\Pr(y_k|x_{1..t}) \propto \Pr(y_k, x_{k+1..t}|x_{1..k}) \text{ by conditioning}$$
$$= \Pr(y_k|x_{1..k}) \Pr(x_{k+1..t}|y_k) \text{ by chain rule}$$

- Recursive computation

$$\Pr(x_{k+1..t}|y_k) = \sum_{y_{k+1}} \Pr(y_{k+1}, x_{k+1..t}|y_k) \text{ by marginalization}$$
$$= \sum_{y_{k+1}} \Pr(y_{k+1}|y_k) \Pr(x_{k+1..t}|y_{k+1}) \text{ by chain rule}$$
$$= \sum_{y_{k+1}} \Pr(y_{k+1}|y_k) \Pr(x_{k+1}|y_{k+1}) \Pr(x_{k+2..t}|y_{k+1}) \text{ by conditional independence}$$

Forward-backward algorithm

1. Compute $\Pr(y_k|x_{1..k})$ by forward computation

   $\Pr(y_1|x_1) \propto \Pr(x_1|y_1)\Pr(y_1)$
   For $i = 2$ to k do
   $\Pr(y_i|x_{1..i}) \propto \Pr(x_i|y_i)\sum_{y_{i-1}}\Pr(y_i|y_{i-1})\Pr(y_{i-1}|x_{1..i-1})$

2. Compute $\Pr(x_{k+1..t}|y_k)$ by backward computation

   $\Pr(x_t|y_{t-1}) = \sum_{y_t}\Pr(y_t|y_{t-1})\Pr(x_t|y_t)$
   For $j = t-1$ downto $k+1$ do
   $\Pr(x_{j..t}|y_{j-1}) = \sum_{y_j}\Pr(y_j|y_{j-1})\Pr(x_j|y_j)\Pr(x_{j+1..t}|y_j)$

3. $\Pr(y_k|x_{1..t}) \propto \Pr(y_k|x_{1..k})\Pr(x_{k+1..t}|y_k)$

- Linear complexity in t

UNIVERSITY OF WATERLOO

Most likely explanation

- $argmax_{y_{1..t}} \Pr(y_{1..t}|x_{1..t})$: most likely state sequence given observations
- Example: speech recognition
- Computation:

$$\max_{y_{1..t}} \Pr(y_{1..t}|x_{1..t}) = \max_{y_t} Pr(x_t|y_t) \max_{y_{1..t-1}} Pr(y_{1..t}|x_{1..t-1})$$

- Recursive computation:

$$\max_{y_{1..i-1}} \Pr(y_{1..i}|x_{1..i-1}) \propto \max_{y_{i-1}} \Pr(y_i|y_{i-1}) \Pr(x_{i-1}|y_{i-1}) \max_{y_{1..i-2}} \Pr(y_{1..i-1}|x_{1..i-2})$$

Viterbi Algorithm

1. Compute $\max_{y_{1..t}} \Pr(y_{1..t}|x_{1..t})$ by dynamic programming

$$\max_{y_1} \Pr(y_{1..2}|x_1) \propto \max_{y_1} \Pr(y_2|y_1)\Pr(x_1|y_1)\Pr(y_1)$$

For $i = 2$ to $t-1$ do

$$\max_{y_{1..i}} \Pr(y_{1..i+1}|x_{1..i}) \propto \max_{y_i} \Pr(y_{i+1}|y_i)\Pr(x_i|y_i) \max_{y_{1..i-1}} \Pr(y_{1..i}|x_{1..i-1})$$

$$\max_{y_{1..t}} \Pr(y_{1..t}|x_{1..t}) \propto \max_{y_t} \Pr(x_t|y_t) \max_{y_{1..t-1}} \Pr(y_{1..t}|x_{1..t-1})$$

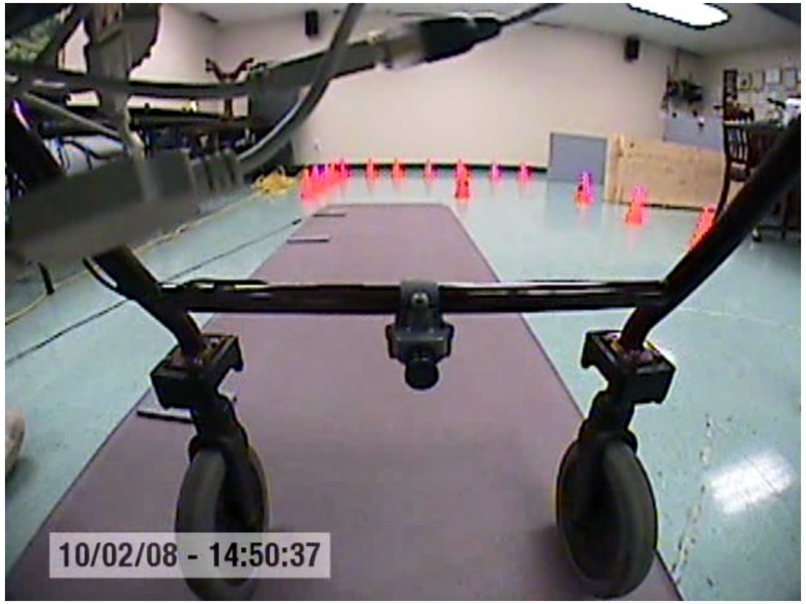
- Linear complexity in t

Case Study: Activity Recognition

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

Backward view

Forward view



UNIVERSITY OF WATERLOO

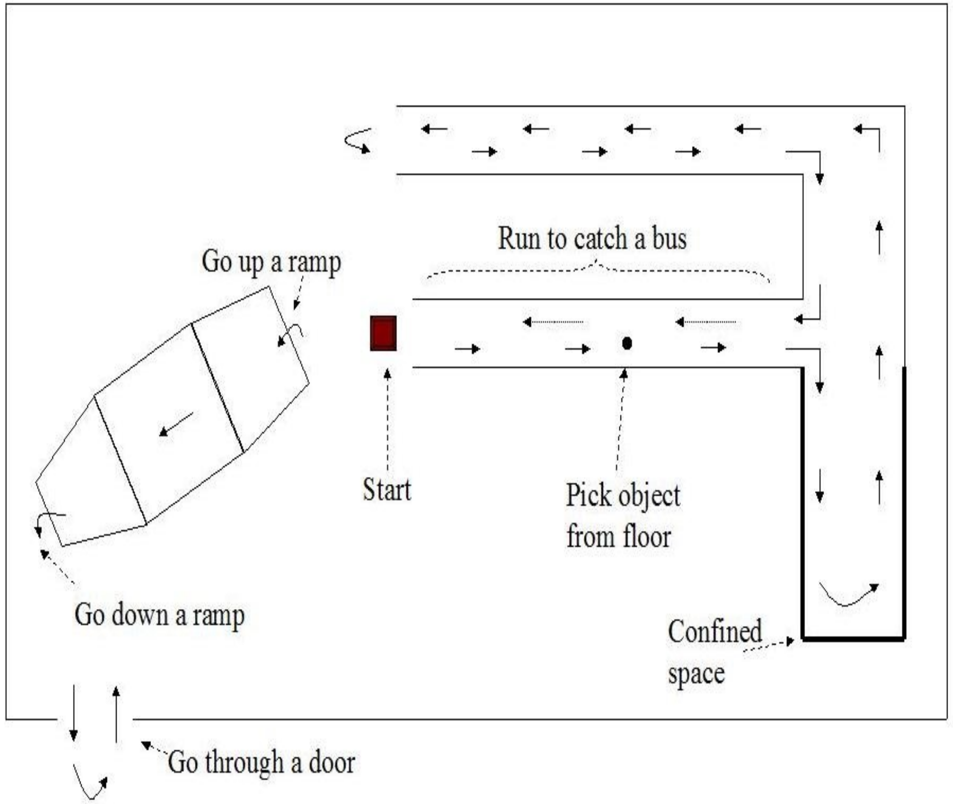
Inputs: Raw Sensor Data

- 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

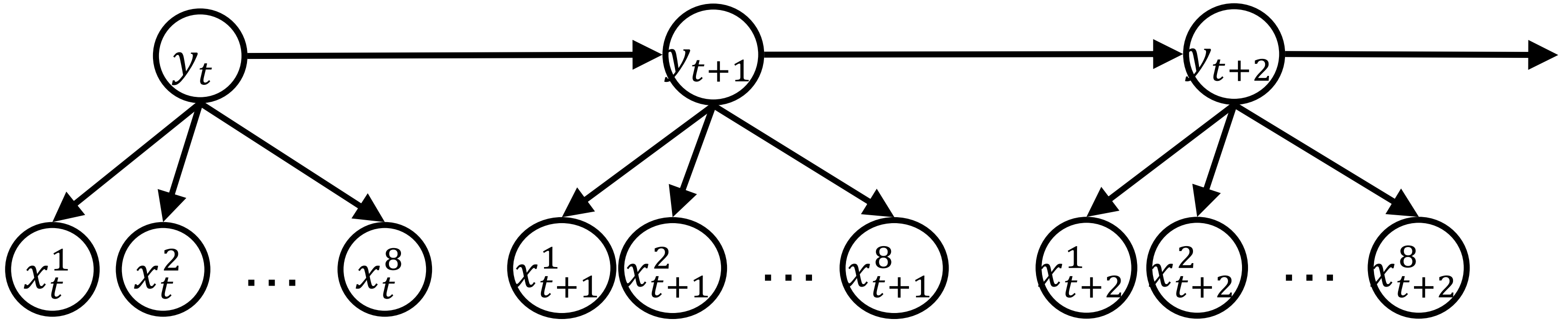
- 8 walker users at Winston Park (84-97 years old)
- 12 older adults (80-89 years old) in the KW area who do not use walkers



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)

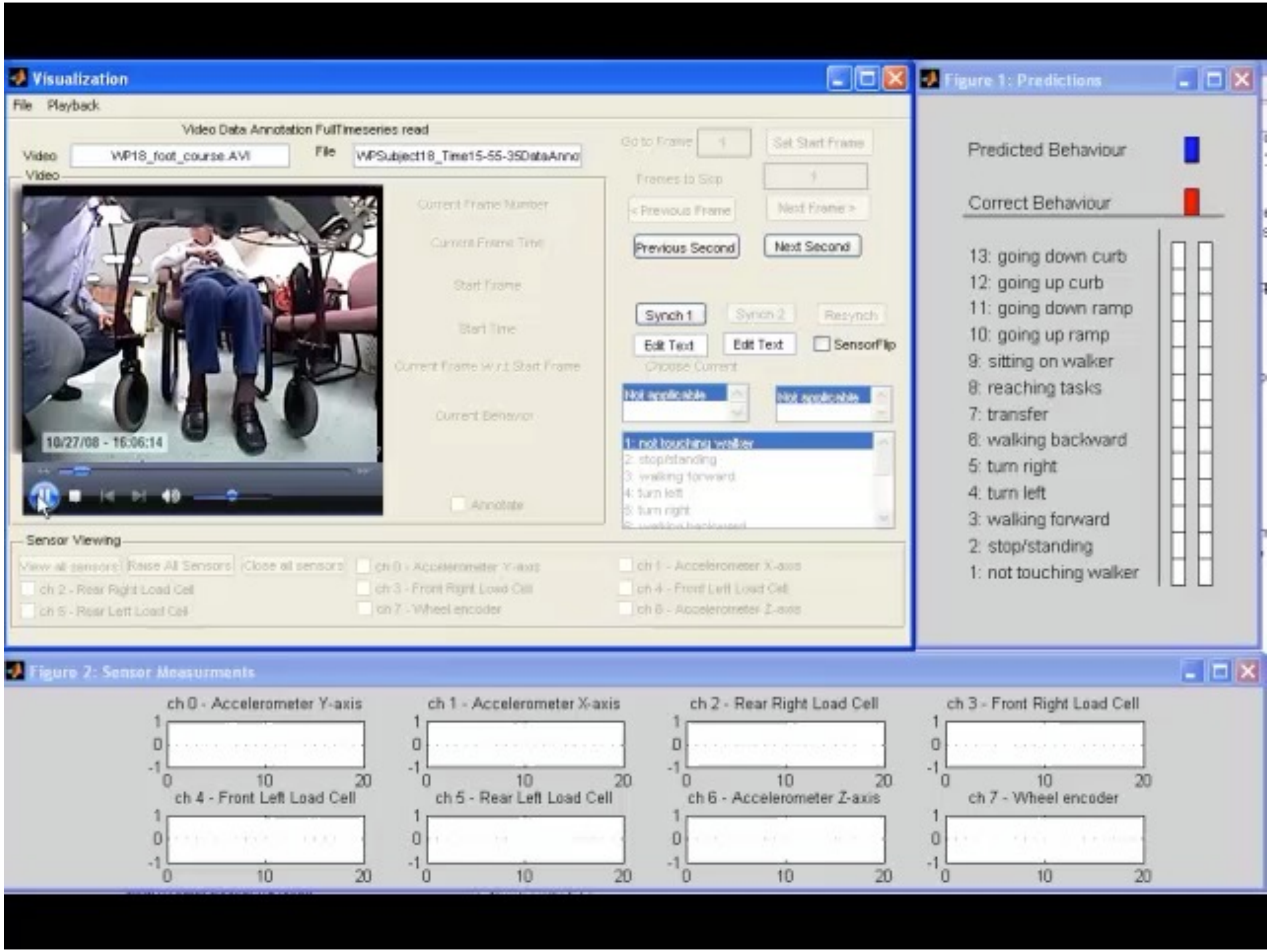
Hidden Markov Model (HMM)



- 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: $\phi^i_{val|class} = \Pr(x^i_t = val | y_t = class)$

    or $N(val | \mu^i_{class}, \sigma^i_{class}) = \Pr(x^i_t = val | y_t = class)$
- 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
10/27/08 - 16:06:14
Previous Second
Next Second
Synch 1
Edit Text
Edit Text
SensorFlip
1: not touching walker
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
- $\Pr(y_{0..t}, x_{1..t})$

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

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

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

UNIVERSITY OF WATERLOO

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:

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

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

- 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

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

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

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

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

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