22.03.2019

Statistical Methods in AI (CSE/ECE 471)

Lecture-19: ML for Sequential Data - Hidden Markov Models

Ravi Kiran

Center for Visual Information Technology (CVIT), IIIT Hyderabad

Announcements

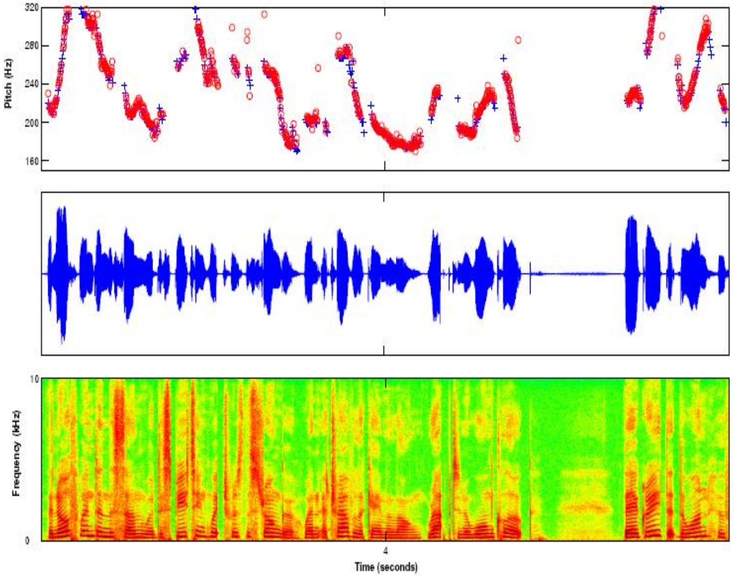
- For final exam
  - Any answer written with a pencil will AUTOMATICALLY get 0 marks !

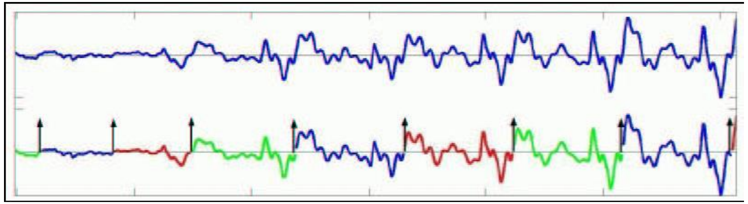
Analysis of Sequential Data

- Sequential structure arises in a huge range of applications
  - Repeated measurements of a temporal process
  - Online decision making & control
  - Text, biological sequences etc.

Speech Recognition

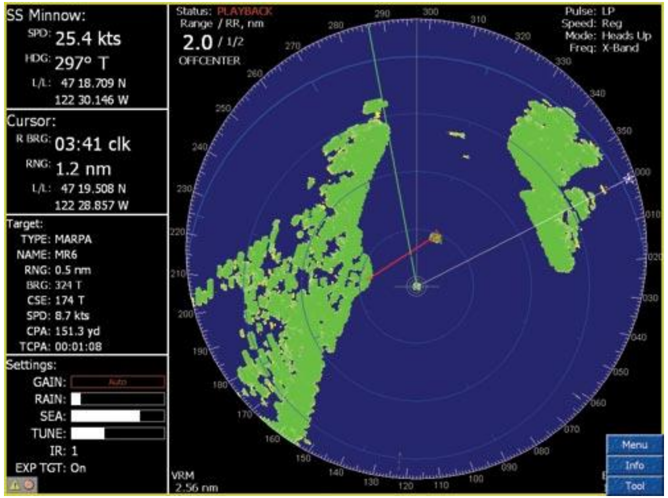
- Given an audio waveform, robustly extract & recognize any spoken words
- Statistical models can be used to
  – Provide greater robustness to noise
  – Adapt to accent of different speakers
  – Learn from training



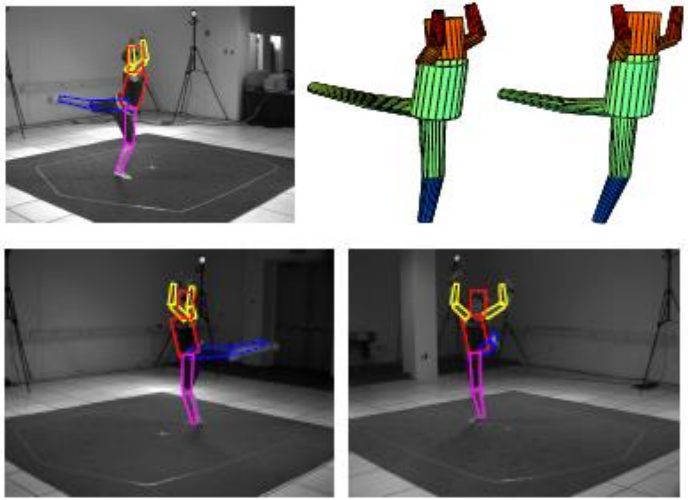


S. Roweis, 2004

Target Tracking



Radar-based tracking of multiple targets

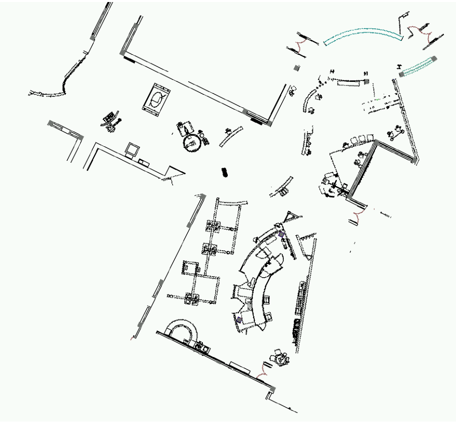


Visual tracking of articulated objects
(L. Sigal et. al., 2006)

- Estimate motion of targets in 3D world from indirect, potentially noisy measurements

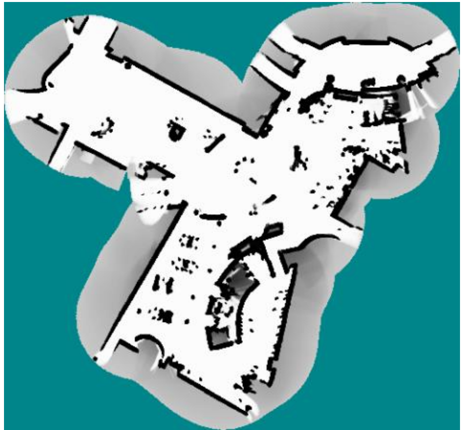
Robot Navigation: *SLAM*

Simultaneous Localization and Mapping



CAD Map

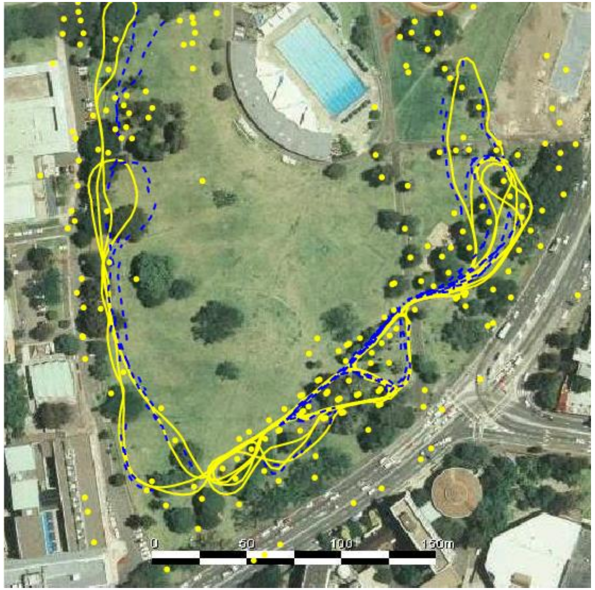
(S. Thrun, San Jose Tech Museum)



Estimated Map

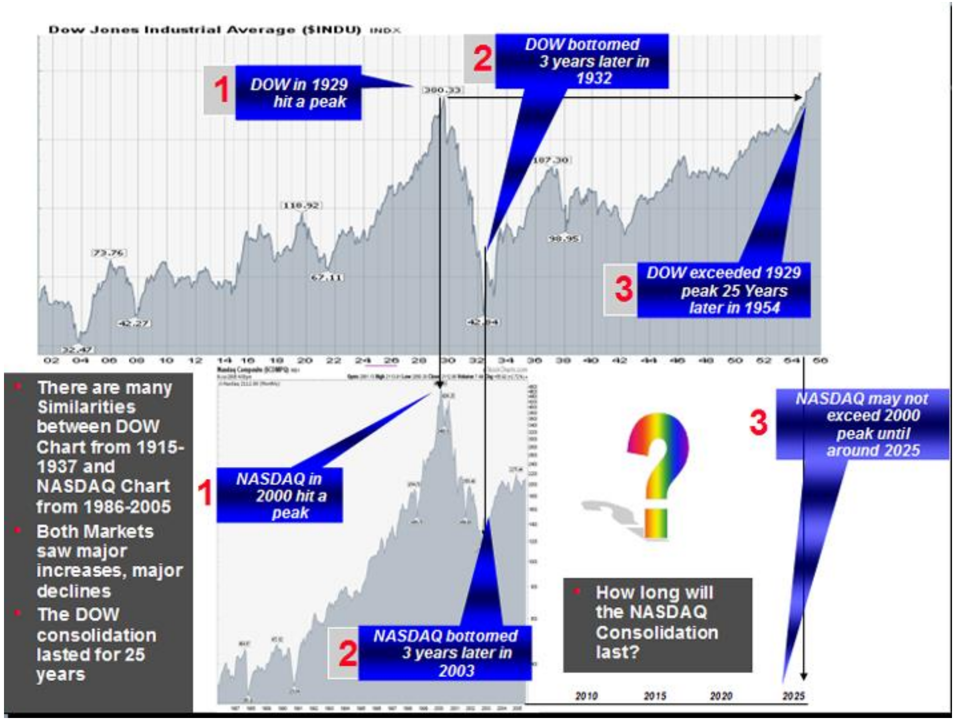
Landmark SLAM

(E. Nebot, Victoria Park)



- As robot moves, estimate its pose & world geometry

Financial Forecasting



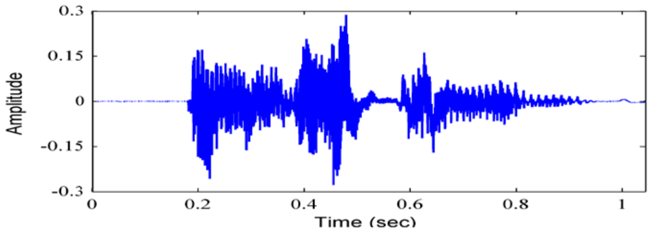
http://www.steadfastinvestor.com/

- Predict future market behavior from historical data, news reports, expert opinions, …

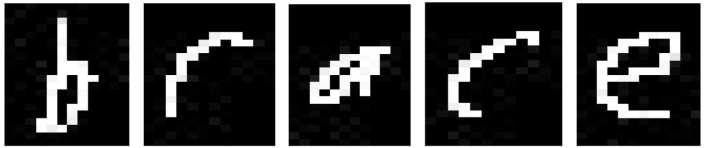
i.i.d to sequential data

- So far we assumed independent, identically distributed data $\{X_i\}_{i=1}^n \overset{iid}{\sim} p(\mathbf{X})$
- Sequential (non i.i.d.) data
  - Time-series data

    E.g. Speech



  - Characters in a sentence



  - Base pairs along a DNA strand

Sequential Processes

- Consider a system which can occupy one of *N* discrete *states* or *categories*

$$x_t \in \{1, 2, \ldots, N\} \longrightarrow \text{state at time } t$$

- We are interested in *stochastic* systems, in which state evolution is random
- Any *joint* distribution can be factored into a series of *conditional* distributions:

$$p(x_0, x_1, \ldots, x_T) = p(x_0) \prod_{t=1}^{T} p(x_t \mid x_0, \ldots, x_{t-1})$$

$p(A,B) = p(A)\,p(B|A)$

$p(A,B,C) = p(A)\,p(B|A)$
$p(C|A,B)$

Markov Processes

- For a *Markov* process, the next state depends only on the current state:

$$p(x_{t+1} \mid x_0, \ldots, x_t) = p(x_{t+1} \mid x_t)$$

- This property in turn implies that

$$p(x_0, \ldots, x_{t-1}, x_{t+1}, \ldots, x_T \mid x_t)$$
$$= p(x_0, \ldots, x_{t-1} \mid x_t) p(x_{t+1}, \ldots, x_T \mid x_t)$$

"Conditioned on the present,
the past & future are independent"

X_1 X_2 X_3 N N $x_1 = a_3$ $x_2 = a_5$ $x_3 = a_2$ .....

State Transition Matrices

- A *stationary* Markov chain with *N* states is described by an *NxN* *transition matrix:*

$X_1 = a_1$ ⋮ a_N a_3 a_5 .....

$$Q = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix}$$

$$q_{ij} \triangleq p(x_{t+1} = i \mid x_t = j)$$

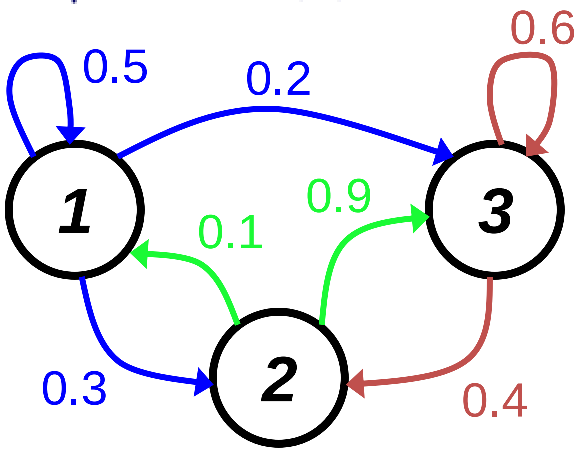
- Constraints on valid transition matrices:

$$q_{ij} \geq 0 \qquad \sum_{i=1}^{N} q_{ij} = 1 \quad \text{for all } j$$

State Transition Diagrams

$$q_{ij} \triangleq p(x_{t+1} = i \mid x_t = j)$$

$$Q = \begin{bmatrix} 0.5 & 0.1 & 0.0 \\ 0.3 & 0.0 & 0.4 \\ 0.2 & 0.9 & 0.6 \end{bmatrix}$$



- Think of a particle randomly following an arrow at each discrete time step
- Most useful when *N* small, and *Q* *sparse*

Markov Models

parameters in
stationary model
K-ary variables

- Markov Assumption

1st order $\quad p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n | X_{n-1})$ $\quad O(K^2)$

mth order $\quad p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n | X_{n-1}, \ldots, X_{n-m})$ $\quad O(K^{m+1})$

n-1th order $\quad p(\mathbf{X}) = \prod_{i=1}^{n} p(X_n | X_{n-1}, \ldots, X_1)$ $\quad O(K^n)$

≡ no assumptions – complete (but directed) graph

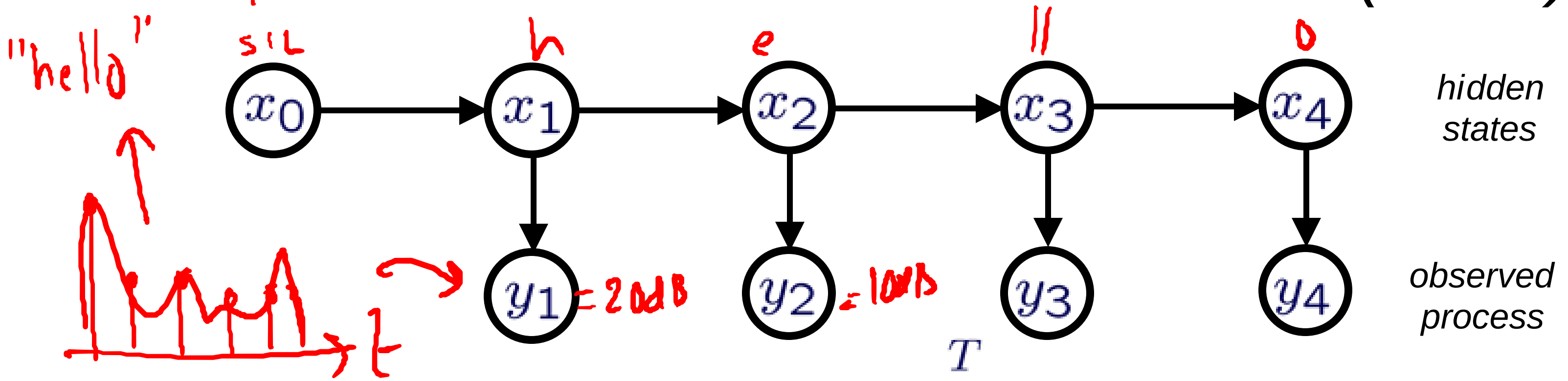
Homogeneous/stationary Markov model (probabilities don't depend on n)

Hidden Markov Models

- Few realistic time series directly satisfy the assumptions of Markov processes:

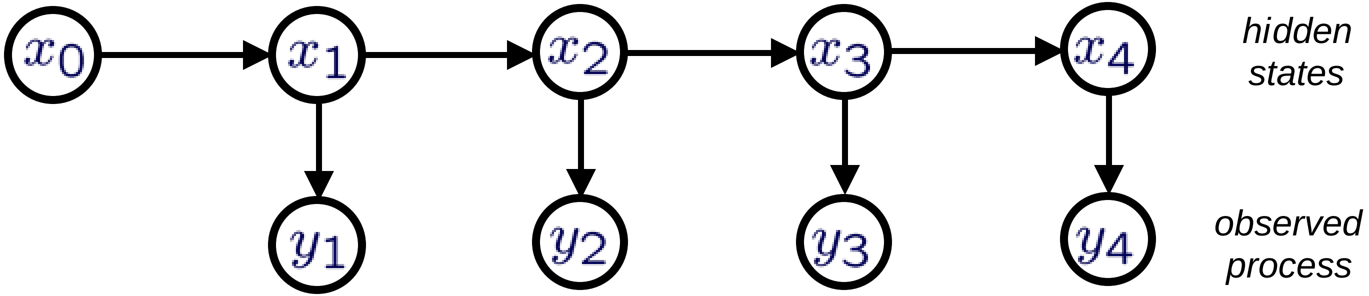
"Conditioned on the present,
the past & future are independent"

- Motivates *hidden Markov models (HMM):*



$$p(x_0, x_1, \ldots, x_T) = p(x_0) \prod_{t=1}^{T} p(x_t \mid x_{t-1}) p(y_t \mid x_t)$$

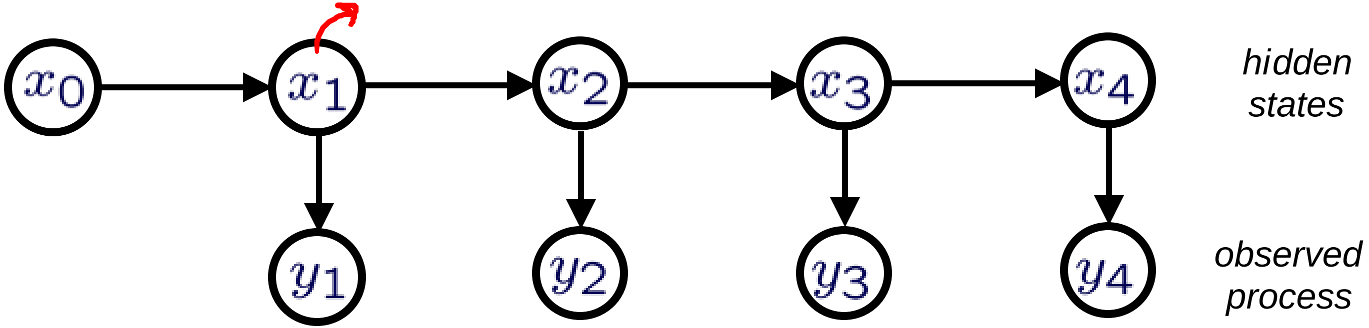
Hidden states



- Given x_t, earlier observations provide no *additional information* about the future:

$$p(y_t, y_{t+1}, \ldots \mid x_t, y_{t-1}, y_{t-2}, \ldots) = p(y_t, y_{t+1}, \ldots \mid x_t)$$

Where do states come from?

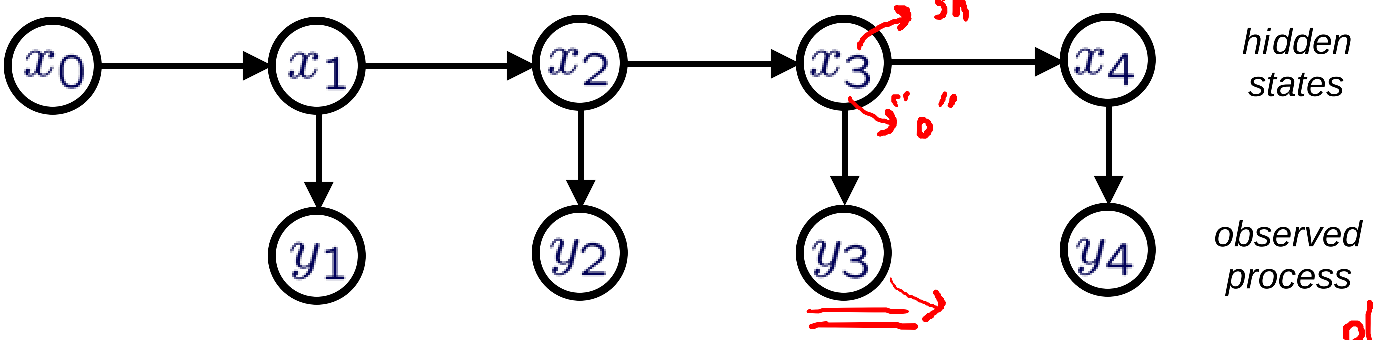


M

- Analysis of a *physical phenomenon*:
  - Dynamical models of an aircraft or robot
  - Geophysical models of climate evolution
- Discovered from *training data*:
  - Recorded examples of spoken English
  - Historic behavior of stock prices

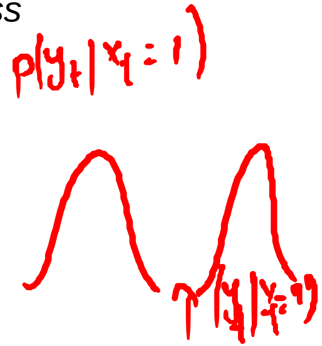
Discrete State HMMs

$$x_t \in \{1, 2, \ldots, N\}$$



- Associate each of the *N* hidden states with a different observation distribution:

$$p(y_t \mid x_t = 1) \qquad p(y_t \mid x_t = 2) \quad \cdots$$



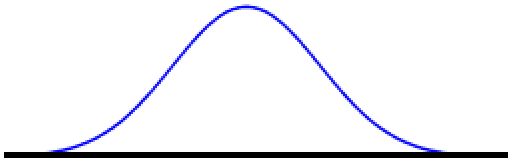
- Observation densities are typically chosen to encode domain knowledge

Discrete HMMs: Observations

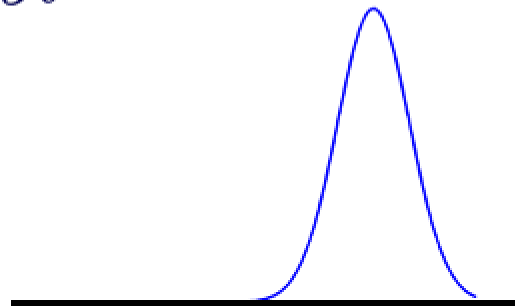
Discrete Observations $y_t \in \{1, 2, \ldots, M\}$

$$p(y_t \mid x_t = 1) = \begin{bmatrix} 0.3 \\ 0.1 \\ 0.5 \\ 0.1 \end{bmatrix} \qquad p(y_t \mid x_t = 2) = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.1 \\ 0.5 \end{bmatrix}$$

Continuous Observations $y_t \in \mathbb{R}^k$

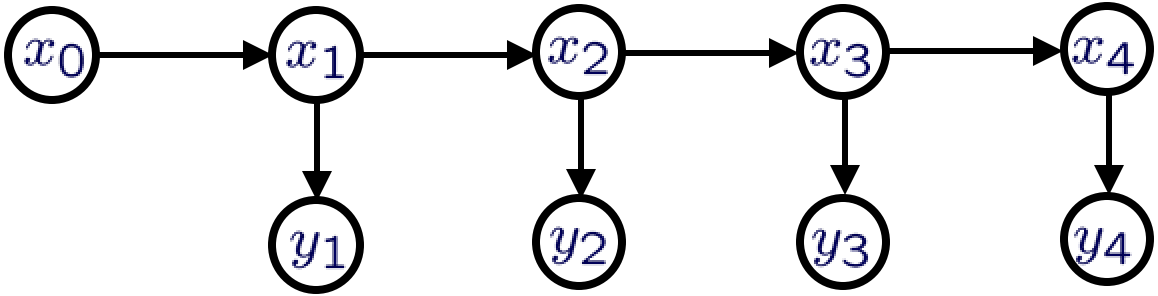


$p(y_t \mid x_t = 1)$



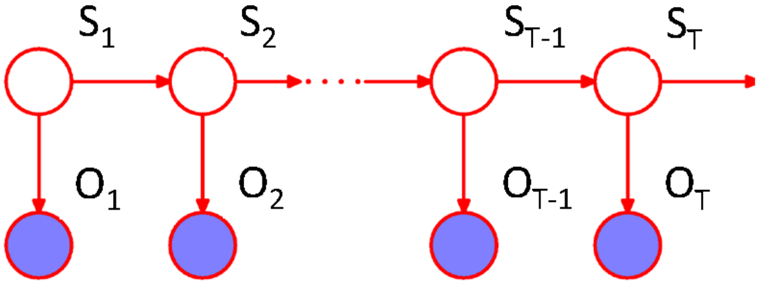
$p(y_t \mid x_t = 2)$

Specifying an HMM



- Observation model: $P(y_i|x_i)$
- Transition model: $P(x_i|x_{i-1})$
- Initial state distribution: $P(x_0)$

Hidden Markov Models



$p(A, B)$

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A,B)}{\sum_A p(A,B)}$$

$$p(S_1, \ldots, S_T, O_1, \ldots, O_T) \;=\; \prod_{t=1}^{T} p(O_t|S_t) \prod_{t=1}^{T} p(S_t|S_{t-1})$$

1st order Markov assumption on hidden states $\{S_t\}$ $t = 1, \ldots, T$

(can be extended to higher order).

Note: O_t depends on all previous observations $\{O_{t-1}, \ldots O_1\}$

Hidden Markov Models

- Parameters – stationary/homogeneous markov model (independent of time t)

Initial probabilities

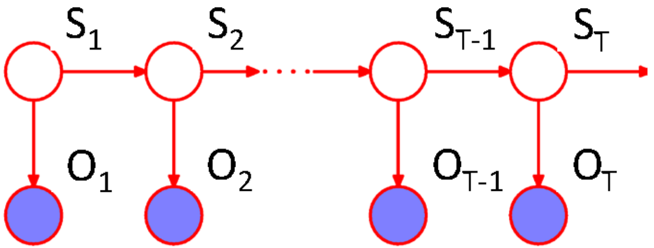
$p(S_1 = i) = \pi_i$

Transition probabilities

$p(S_t = j \mid S_{t-1} = i) = p_{ij}$

Emission probabilities

$p(O_t = y \mid S_t = i) = q_i^y$



$$p(\{S_t\}_{t=1}^T, \{O_t\}_{t=1}^T) \quad = \quad p(S_1) \prod_{t=2}^T p(S_t|S_{t-1}) \prod_{t=1}^T p(O_t|S_t)$$

HMM Example

- The Dishonest Casino

A casino has two dices:

Fair dice

P(1) = P(2) = P(3) = P(5) = P(6) = 1/6

Loaded dice

P(1) = P(2) = P(3) = P(5) = 1/10
P(6) = ½

Casino player switches back-&-forth between fair and loaded die with 5% probability

HMM Problems

GIVEN: A sequence of rolls by the casino player

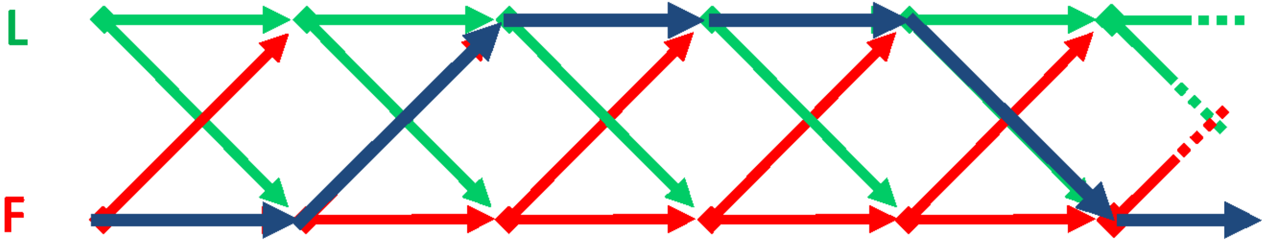
124552646214614613613666166466163661636616361651561511514612356234

QUESTION

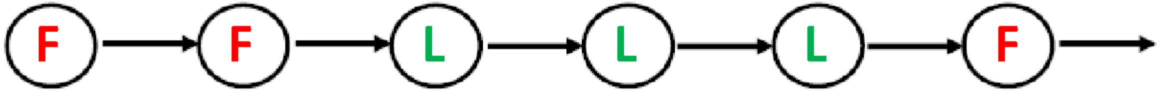
- How likely is this sequence, given our model of how the casino works?
  - This is the **EVALUATION** problem in HMMs
- What portion of the sequence was generated with the fair die, and what portion with the loaded die?
  - This is the **DECODING** question in HMMs
- How “loaded” is the loaded die? How “fair” is the fair die? How often does the casino player change from fair to loaded, and back?
  - This is the **LEARNING** question in HMMs

HMM Example

- Observed sequence: $\{O_t\}_{t=1}^T$

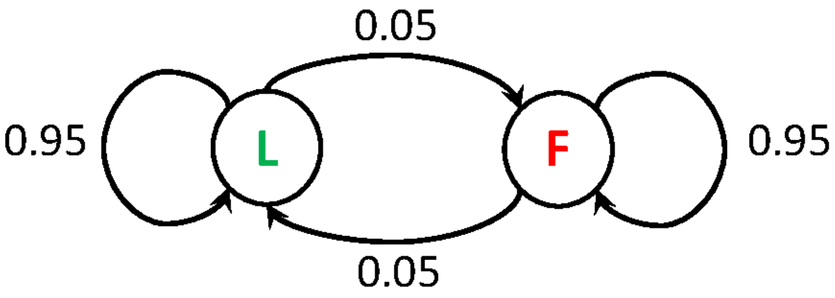


- Hidden sequence $\{S_t\}_{t=1}^T$ or segmentation):



State Space Representation

- Switch between F and L with 5% probability



- **HMM Parameters**

Initial probs	$P(S_1 = L) = 0.5 = P(S_1 = F)$
Transition probs	$P(S_t = L/F \mid S_{t-1} = L/F) = 0.95$
	$P(S_t = F/L \mid S_{t-1} = L/F) = 0.05$
Emission probabilities	$P(O_t = y \mid S_t = F) = 1/6 \qquad y = 1,2,3,4,5,6$
	$P(O_t = y \mid S_t = L) = 1/10 \qquad y = 1,2,3,4,5$
	$= 1/2 \qquad y = 6$

Three main problems in HMMs

$O_1 \rightarrow$ 1245526462146146136136661664661636616366163616515615115146123562344 $\leftarrow O_T$

- Evaluation – Given HMM parameters & observation seqn $\{O_t\}_{t=1}^T$

  find $p(\{O_t\}_{t=1}^T | \theta)$ prob of observed sequence

  - How likely is this sequence, given our model of how the casino works?

- Decoding – Given HMM parameters & observation seqn $\{O_t\}_{t=1}^T$

  find $\arg\max_{s_1,\ldots,s_T} p(\{S_t\}_{t=1}^T | \{O_t\}_{t=1}^T, \theta)$ most probable sequence of hidden states

  - What portion of the sequence was generated with the fair die, and what portion with the loaded die?

- Learning – Given HMM with unknown parameters and $\{O_t\}_{t=1}^T$ observation sequence

  find $\arg\max_{\theta} p(\{O_t\}_{t=1}^T | \theta)$ parameters that maximize likelihood of observed data

  - How "loaded" is the loaded die? How "fair" is the fair die? How often does the casino player change from fair to loaded, and back?

HMM Algorithms

- Evaluation – What is the probability of the observed sequence? Forward Algorithm

- Decoding – What is the probability that the third roll was loaded given the observed sequence? Forward-Backward Algorithm

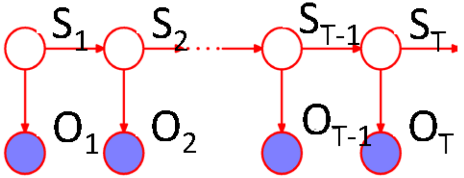
  – What is the most likely die sequence given the observed sequence? Viterbi Algorithm

- Learning – Under what parameterization is the observed sequence most probable? Baum-Welch Algorithm (EM)

Evaluation Problem

- Given HMM parameters $p(S_1), p(S_t|S_{t-1}), p(O_t|S_t)$ & observation sequence $\{O_t\}_{t=1}^T$

  find probability of observed sequence



$$\underline{p(\{O_t\}_{t=1}^T)} = \sum_{S_1,\ldots,S_T} \underline{p(\{O_t\}_{t=1}^T, \{S_t\}_{t=1}^T)}$$

$$= \sum_{S_1,\ldots,S_T} p(S_1) \prod_{t=2}^T p(S_t|S_{t-1}) \prod_{t=1}^T p(O_t|S_t)$$

$\sum_{S_1} \sum_{S_2} \cdots \sum_{S_T}$

requires summing over all possible hidden state values at all times – K^T exponential # terms!

Instead: $$p(\{O_t\}_{t=1}^T) = \sum_k \underbrace{p(\{O_t\}_{t=1}^T, S_T = k)}_{\alpha_T^k}$$

Compute recursively

Forward Probability

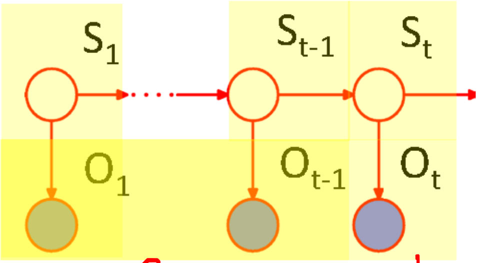
$$p(\{O_t\}_{t=1}^T) = \sum_k p(\{O_t\}_{t=1}^T, S_T = k) = \sum_k \alpha_T^k$$

Compute forward probability α_t^k recursively over t

$S_{t-1} = i$

$$\alpha_t^k \quad := \quad p(O_1, \dots, O_t, S_t = k)$$

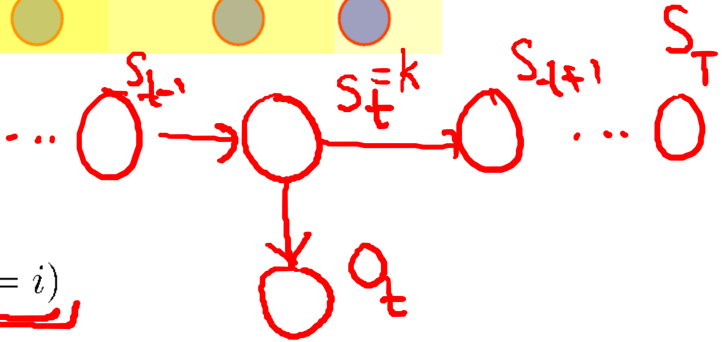
$t \leq T$



Introduce S_{t-1}

Chain rule

Markov assumption



$$= \quad p(O_t | S_t = k) \sum_{i=1}^{K} \alpha_{t-1}^i \, p(S_t = k | S_{t-1} = i)$$

Forward Algorithm

Can compute α_t^k for all k, t using dynamic programming:

- Initialize: $\alpha_1^k = p(O_1 | S_1 = k)\, p(S_1 = k)$ for all k
- Iterate: for t = 2, …, T

$$\alpha_t^k = p(O_t | S_t = k) \sum_i \alpha_{t-1}^i\, p(S_t = k | S_{t-1} = i) \qquad \text{for all k}$$

- Termination: $p(\{O_t\}_{t=1}^T) = \sum_k \alpha_T^k$