Machine Learning 10-701

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

March 22, 2011

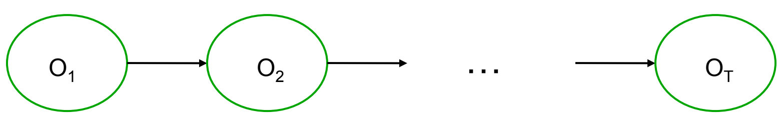
Today:
- Time series data
- Markov Models
- Hidden Markov Models
- Dynamic Bayes Nets

Reading:
- Bishop: Chapter 13 (very thorough)

thanks to Professors Venu Govindaraju, Carlos Guestrin, Aarti Singh, and Eric Xing for access to slides on which some of these are based

Sequential Data

- stock market prediction
- speech recognition
- gene data analysis



how shall we represent and learn $P(O_1, O_2 \ldots O_T)$?

Markov Model

$$O_1 \rightarrow O_2 \rightarrow \cdots \rightarrow O_T$$

how shall we represent and learn $P(O_1, O_2 \ldots O_T)$?

Use a Bayes net: $\boxed{P(O_1 \ldots O_T) = \prod_{t=1}^{T} P(O_t | Pa(O_t))}$

$\boxed{\text{Markov model: } Pa(O_t) \equiv O_{t-1}}$

$$O_1 \rightarrow O_2 \rightarrow O_3 \rightarrow \cdots \rightarrow O_T$$

$\boxed{\text{nth order Markov model: } Pa(O_t) \equiv O_{t-1}, O_{t-2}, \ldots O_{t-n}}$

Markov Model

$$O_1 \rightarrow O_2 \rightarrow \cdots \rightarrow O_T$$

how shall we represent and learn $P(O_1, O_2 \ldots O_T)$?

Use a Bayes net: $P(O_1 \ldots O_T) = \prod_{t=1}^{T} P(O_t | Pa(O_t))$

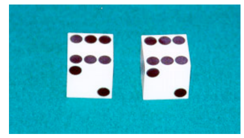
Markov model: $Pa(O_t) \equiv O_{t-1}$

nth order Markov model: $Pa(O_t) \equiv O_{t-1}, O_{t-2}, \ldots O_{t-n}$

if O_t real valued, and assume $P(O_t) \sim N(f(O_{t-1}, O_{t-2} \ldots O_{t-n}), \sigma)$, where f is some linear function, called nth order autoregressive (AR) model

Hidden Markov Models: Example

An experience in a casino



<u>Game:</u>

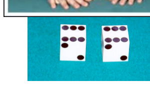
1. You bet $1
2. You roll (always with a fair die)
3. Casino player rolls (sometimes with fair die, sometimes with loaded die)
4. Highest number wins $2

Here is **his** sequence of die rolls:

1245526462146146136136661664661636
616366163616515615115146123562344

Which die is being used in each play?

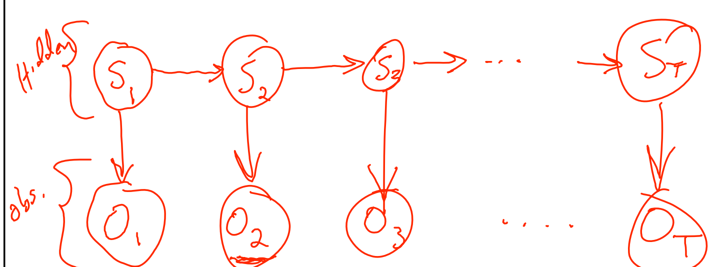
Question:



1245526462146146136136661664661636
616366163616515615115146123562344

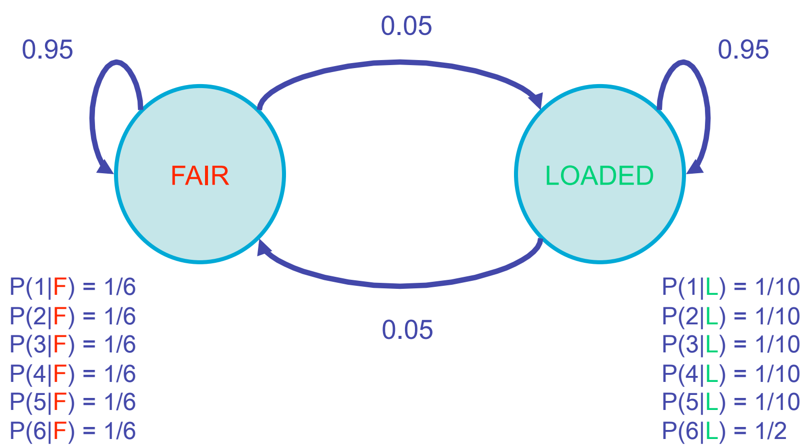
Which die is being used in each play?

$S \in \{\overset{0}{\text{Fair}}, \overset{1}{\text{Loaded}}\}$



$O_t \in \{1, 2, 3, 4, 5, 6\}$

The Dishonest Casino Model



Puzzles Regarding the Dishonest Casino

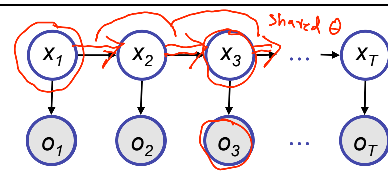
GIVEN: A sequence of rolls by the casino player

1245526462146146136136661664661636616366163616515615115146123562344

QUESTION

- How likely is this sequence, given our model of how the casino works?
  - This is the **EVALUATION** problem
- What portion of the sequence was generated with the fair die, and what portion with the loaded die?
  - This is the **DECODING** question
- 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

Definition of HMM

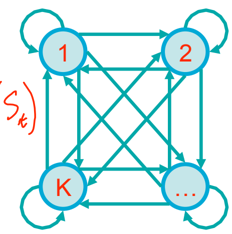


Graphical model

1. N: number of hidden states X_t can take on
   $X = \{1, 2, \ldots N\}$
2. O: set of values O_t can take on
3. Initial state distribution: $P(X_1=k)$ for $k=1, 2, \ldots N$
4. State transition distribution: $P(X_{t+1}=k \mid X_t=i)$, for $k,i =1, 2, \ldots N$
5. Emission distribution: $P(O_t \mid X_t)$

CPD's shared parameters

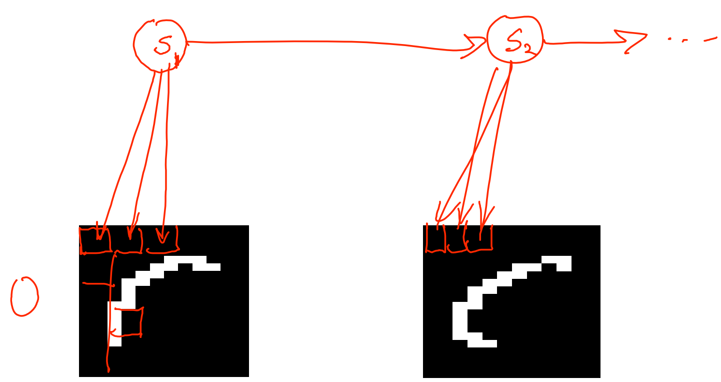
$P(S_{t+1} \mid S_t)$



State automata view

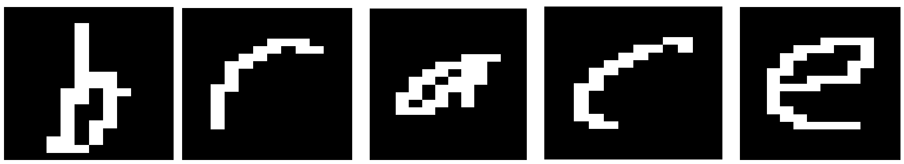
Handwriting recognition

$S \in \{a, b, c, \ldots z\}$



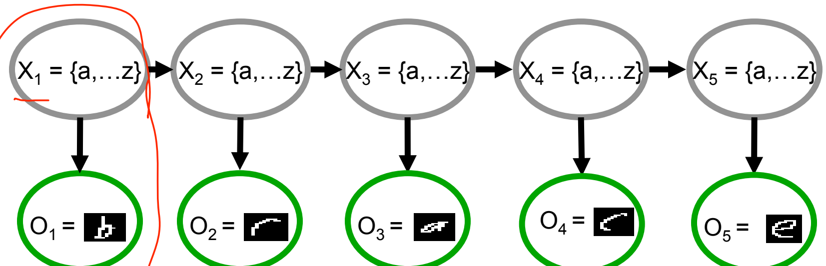
Character recognition, e.g., logistic regression, Naïve Bayes

Example of a hidden Markov model (HMM)



$P(X_i = \text{'i'})$

Understanding HMM Semantics



vector of pixel intensities?

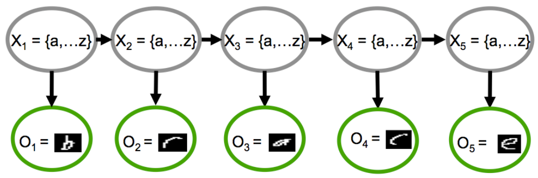
$P(O_t \mid X_t) \sim N(\mu_{X_t}, \sigma)$

$\begin{bmatrix} O_{t1} \\ O_{t2} \\ \vdots \\ O_{t900} \end{bmatrix}$

$\theta_{b,u}$ $P(X_{t+1} = \text{'b'} \mid X_t = \text{'u'})$

26×26 θ_{ij}

HMMs semantics: Details



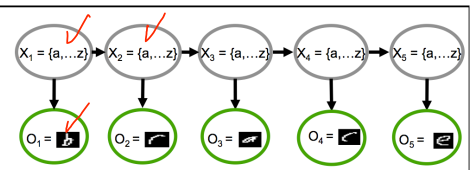
Just 3 distributions:

$P(X_1)$

$P(X_i \mid X_{i-1})$

$P(O_i \mid X_i)$

How do we generate a random output sequence following the HMM?

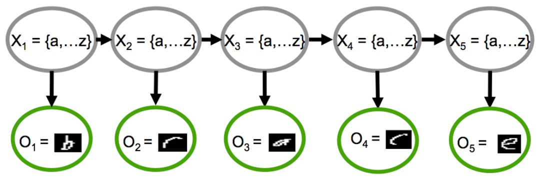


1. Randomly draw x_1 from $P(X_1)$

2. " " O_t given X_t
X_{t+1} given X_t

for $t=1$ to T

Using and Learning HMM's



$P(X_i \mid o_{1..n})$

Core HMM questions:

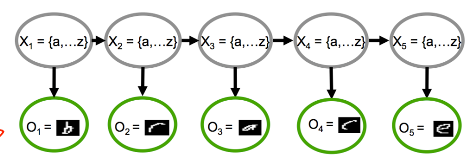
1. How do we calculate $P(o_1, o_2, \ldots o_n)$?

2. How do we calculate argmax over $x_1, x_2, \ldots x_n$
   of $P(x_1, x_2, \ldots x_n \mid o_1, o_2, \ldots o_n)$?

3. How do we train the HMM, given its structure and
   - 3a. Fully observed training examples: $< x_1, \ldots x_n, o_1, \ldots o_n >$
   - 3b. Partially observed training examples: $< o_1, \ldots o_n >$

How do we compute

$P(o_1, o_2, \ldots o_T)$

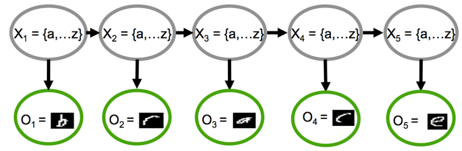


1. brute force:

$$\sum_{i='a'}^{i='z'} \sum_{j='a'}^{j='z'} \sum \sum \sum \quad P(X_1{=}i, X_2{=}j \ldots X_T{=}k, O_1 \ldots O_T)$$

$P(O_1 \ldots O_T)$

How do we compute $\boxed{P(o_1, o_2, \dots o_T)}$ ✓



1. brute force:

$$\alpha_2(k) = P(O_1=o_1, O_2=o_2, X_2=k) = \left[\sum_j \underbrace{P(o_1, X_1=j)}_{\alpha_1(j)} P(X_2=k|X_1=j)\right] P(O_2=o_2|X_2=k)$$

2. Forward algorithm (dynamic progr., variable elimination):

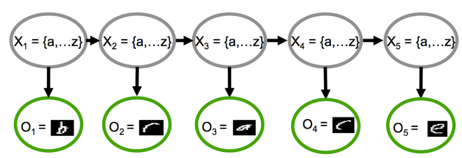
define $\boxed{\alpha_t(k) = P(o_1, o_2, \dots o_t, X_t = k)}$

$$\alpha_1(k) = P(O_1=o_1, X_1=k) = P(X_1=k)P(O_1=o_1|X=k)$$

$$\alpha_{t+1}(k) = \sum_{j=1}^{N=26} \alpha_t(j)\, P(X_{t+1}=k|X_t=j)\, P(O_{t+1}=o_{t+1}|X_{t+1}=k)$$

$$P(o_1, o_2, \dots o_T) = \sum_{k=1}^{N} \alpha_T(k)$$

How do we compute
$P(X_t = k | o_1, o_2, \dots o_T)$

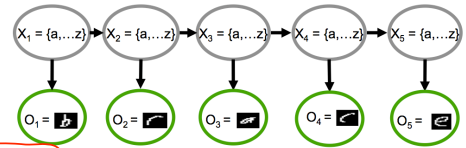


2. Backward algorithm (dynamic progr., variable elimination):

✓ $\alpha_t(k) = \boxed{P(o_1, o_2, \dots o_t, X_t = k)}$

define $\beta_t(k) = \boxed{P(o_{t+1}, o_{t+2}, \dots o_T | X_t = k)}$

✓ $$P(X_t = k | o_1, o_2, \dots o_T) = \frac{P(X_t = k, o_1, o_2, \dots o_T) = \alpha_t(k)\beta_t(k)}{P(o_1, o_2, \dots o_T) = \sum_k \alpha_T(k)}$$



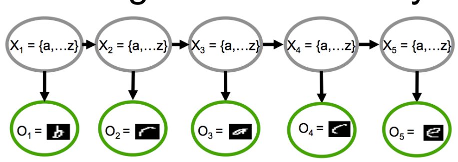
How do we compute

$$\arg\max_{x_1,\ldots x_T} P(x_1,\ldots x_T | o_1, o_2, \ldots o_T)$$

Viterbi algorithm, based on recursive computation of

$$\delta_t(k) = \max_{x_1,\ldots x_{t-1}} P(x_1,\ldots x_{t-1} X_t = k, o_1, o_2, \ldots o_t)$$

Learning HMMs from fully observable data: easy



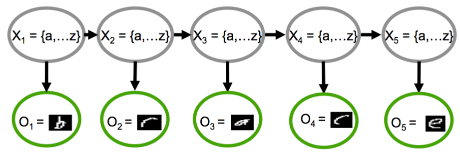
Learn 3 distributions:

$P(X_1)$

$P(O_i \mid X_i)$

$P(X_i \mid X_{i-1})$

Learning HMMs when only observe o1...oT



EM

Baum Welch

Forward-Backw

E 1 est distr $P(x_1 \cdots x_T \mid o_1 \cdots o_T)$

M choose θ to maxime $E_{P(x\ldots \mid o\ldots)} \log P(x\ldots\, o\ldots \mid \theta)$

Additional Time Series Models

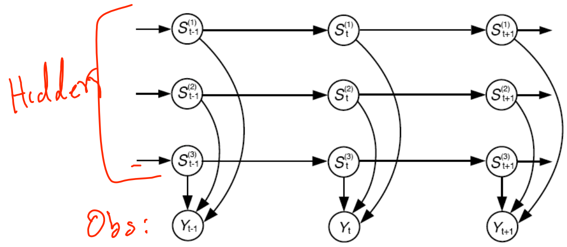


Figure 5: A Bayesian network representing the conditional independence relations in a factorial HMM with $M = 3$ underlying Markov chains. (We only show here a portion of the Bayesian network around time slice t.)

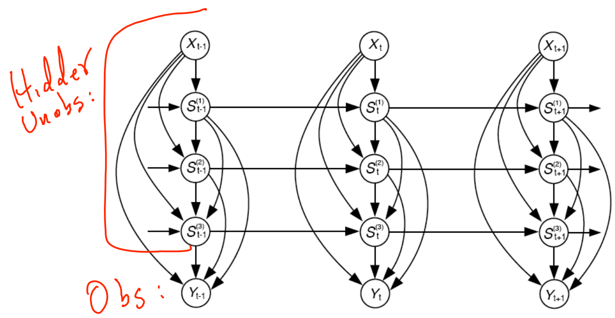


Figure 6: Tree structured hidden Markov models.

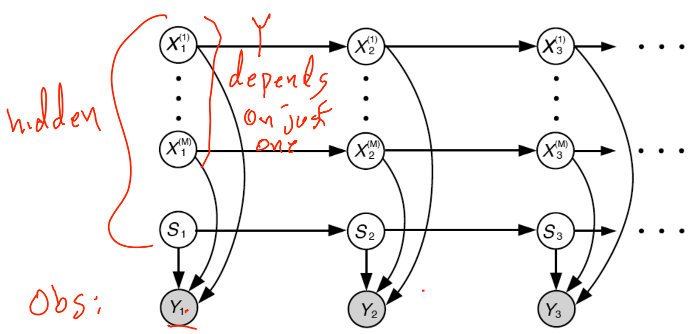


Figure 7: Bayesian network representation for switching state-space models. S_t is the discrete switch variable and $X_t^{(m)}$ are the real-valued state vectors.

What you need to know

- Hidden Markov models (HMMs)
  - Very useful, very powerful!
  - Speech, OCR, time series, …
  - Parameter sharing, only learn 3 distributions
  - Dynamic programming (variable elimination) reduces inference complexity
  - Special case of Bayes net
  - Dynamic Bayesian Networks