CS6784
Primer on Hidden Markov Models

Spring 2014

Thorsten Joachims

Cornell University
Department of Computer Science

Reading:
Koller, Friedman, Getoor, Taskar, “Graphical Models in a Nutshell”
http://www.seas.upenn.edu/~taskar/pubs/gms-srl07.pdf

Warm-Up Assignment

- Submission
  - Deadline today, Thursday 1/30, by 11:59pm
  - Make sure to not include your name in PDF → double-blind reviewing
- Reviewing
  - Double-blind → academic integrity
    - You do not know who you reviewed. Authors do not know who reviewed them.
    - Do not talk about who you reviewed.
    - Assignments done at random. Let us know if you feel conflicted with some assignment.
  - Answer review questions
  - Text should justify and your scores as convincingly as possible.

Part-of-Speech Tagging

- Predict sequence of POS tags for sequence of words:

sentence	POS
$\mathbf{x}_1 = (The, bear, chased, the, cat)$	$\mathbf{y}_1 = (DET, N, V, DET, N)$
$\mathbf{x}_2 = (Students, bear, a, burden)$	$\mathbf{y}_2 = (N, V, DET, N)$

- Ambiguity
  - He will race/V the car.
  - When will the race/NOUN end?
  - I bank/V at CFCU.
  - Go to the bank/NOUN!
- Average of ~2 parts of speech for each word
- 20 – 400 different tags (i.e. word classes)

Predicting Sequences

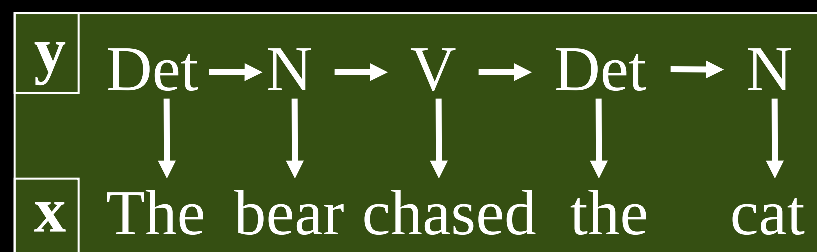
- Bayes rule:
  - Generative model
- Design decisions:
  - Representation
    - Linear chain Hidden Markov Model
  - Prediction (i.e. inference)
    - Viterbi algorithm
  - Learning
    - Maximum likelihood

Representation: Hidden Markov Model

- Bayes rule: $h(x) = \underset{y \in Y}{\operatorname{argmax}}[P(X = x|Y = y)P(Y = y)]$
- Independence assumptions for compact representation

$$P(Y = (y^1, \dots, y^1)) = \prod_{i=1}^{l} P\left(Y^i = y^i \middle| Y^{i-1} = y^{i-1}\right)$$

$$P\left(X = \left(x^1, \dots, x^l\right) \middle| Y = \left(y^1, \dots, y^l\right)\right) = \prod_{i=1}^{l} P\left(X^i = x^i \middle| Y^i = y^i\right)$$

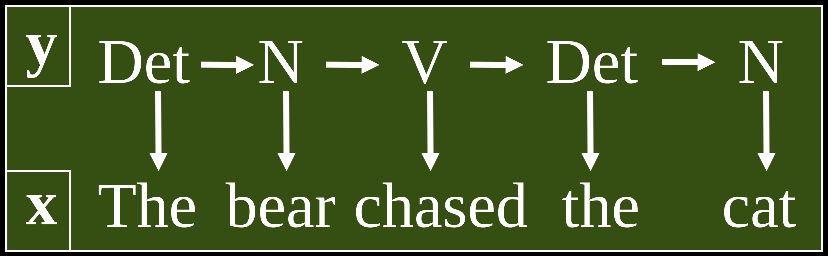


- Each sequence pair has probability:

$$P(X = x, Y = y) = \left[\prod_{i=1}^{l} P\left(Y^i = y^i \middle| Y^{i-1} = y^{i-1}\right) P\left(X^i = x^i \middle| Y^i = y^i\right)\right]$$

Representation: Hidden Markov Model

- States: $y \in \{s_1,\dots,s_k\}$
  - Special starting state s_0
- Outputs symbols: $x \in \{o_1,\dots,o_m\}$
- Transition probability $P(Y^i = s | Y^{i-1} = s')$
  - Probability that one states succeeds another
- Output/Emission probability $P(X^i = o | Y^i = s)$
  - Probability that word is generated in this state



Learning: Estimating HMM Probabilities

- Maximum Likelihood: Given $(x_1, y_1), \ldots, (x_n, y_n)$, find

$$\hat{w} = \underset{w \in W}{\operatorname{argmax}} \prod_{i=1}^{n} [P(Y_i = y_i, X_i = x_i | w)]$$

- Closed-form solutions
  - Estimating transition probabilities

$$P\left(Y^j = a \middle| Y^{j-1} = b\right) = \frac{\#of\ Times\ State\ a\ Follows\ State\ b}{\#\ of\ Times\ State\ b\ Occurs}$$

  - Estimating mission probabilities

$$P\left(X^j = o \middle| Y^j = b\right) = \frac{\#of\ Times\ Output\ o\ is\ Observed\ in\ State\ b}{\#\ of\ Times\ State\ b\ Occurs}$$

- Need for smoothing the estimates (e.g. Laplace)

Prediction/Inference: Viterbi Algorithm

Prediction: Find most likely state sequence

- Given x and fully specified HMM:
  - transition probabilities
  - emission probabilities
- Find the most likely state (i.e tag) sequence $(y^1, \dots, y^l)$ for a given sequence of observed output symbols (i.e. words) $(x^1, \dots, x^l)$

$$h(x) = \underset{(y^1,\dots,y^l)\in Y}{\operatorname{argmax}} \left[\prod_{i=1}^{l} P\left(Y^i = y^i \middle| Y^{i-1} = y^{i-1}\right) P\left(X^i = x^i \middle| Y^i = y^i\right) \right]$$

- Viterbi algorithm uses dynamic programming
  - Construct trellis graph for HMM
  - Shortest path in this graph is most likely state sequence
- Viterbi algorithm has runtime linear in length of sequence

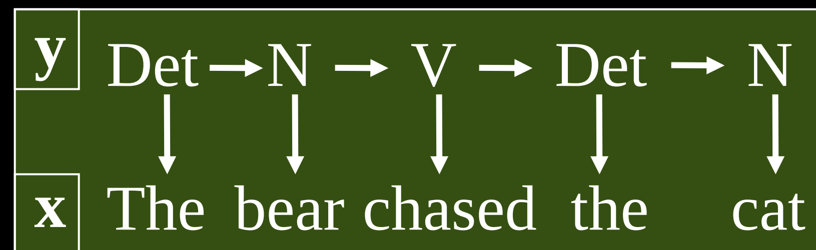
Viterbi Example

$P(X^i \mid Y^i)$	I	bank	at	CFCU	go	to	the
DET	0.01	0.01	0.01	0.01	0.01	0.01	0.94
PRP	0.94	0.01	0.01	0.01	0.01	0.01	0.01
N	0.01	0.4	0.01	0.4	0.16	0.01	0.01
PREP	0.01	0.01	0.48	0.01	0.01	0.47	0.01
V	0.01	0.4	0.01	0.01	0.55	0.01	0.01

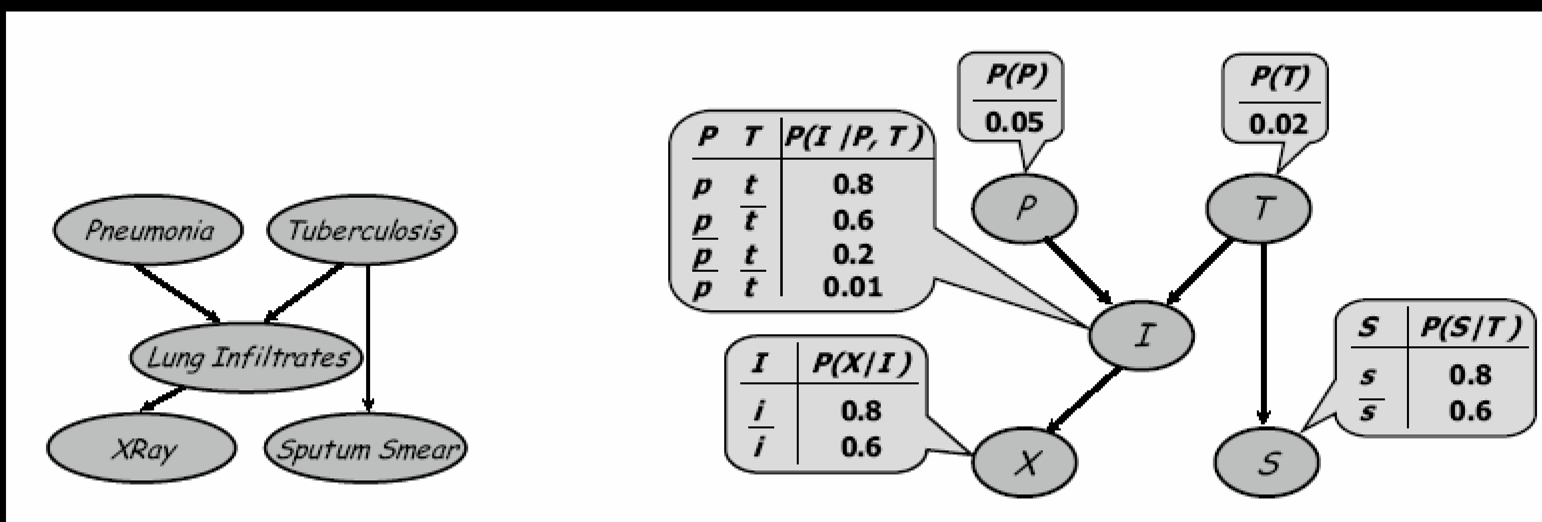
$P(Y^i \mid Y^{i-1})$	DET	PRP	N	PREP	V
START	0.3	0.3	0.1	0.1	0.2
DET	0.01	0.01	0.96	0.01	0.01
PRP	0.01	0.01	0.01	0.2	0.77
N	0.01	0.2	0.3	0.3	0.19
PREP	0.3	0.2	0.3	0.19	0.01
V	0.2	0.19	0.3	0.3	0.01

Directed Graphical Models

- Representation of joint distribution
  - Exploit conditional independence between random variables
- Example
  - Joint distribution



$$P(P,T,I,X,S) = P(P)P(T)P(I|P,T)P(X|I)P(S|T)$$



from [Koller/etal/07]

Undirected Graphical Models

- Markov Networks / Markov Random Fields
  - More flexible representation of joint distribution
- Example
  - Joint distribution $P_H(X_1, \ldots, X_n) = \frac{1}{Z} P'(X_1, \ldots, X_n)$
  - $P'_H(X_1, \ldots, X_n) = \pi_1[D_1] \times \cdots \times \pi_m[D_m]$
  - $Z = \sum_{X_1,\ldots,X_n} P'_H(X_1, \ldots, X_n)$

from [Koller/etal/07]