# CS 378 Lecture 10: HMMs

## Announcements:

- Poll on Piazza (OHs)
- Instapoll survey (end of class)
- A2 due today / A3 out
- Bias response due today

## Recap

Sequence tagging

Map $\bar{X} = (x_1, \ldots, x_n)$ to

$\bar{Y} = (y_1, \ldots, y_n)$

Ex: POS tagging

| NNP | VBZ | NN |
|---|---|---|
| Fed | raises | interest |

$$f(\bar{x}, i) = \begin{cases} \text{Prev Word} \\ \text{Curr Word} \\ \text{Next Word} \end{cases}$$

Can classify $y_i$ independent of others

$P(y_i \mid \bar{x})$ loop over $i$ need feats $f(\bar{x}, i)$

Alternative: Hidden Markov Models

<u>Today</u>
- HMMs
- Example of computing probs
- Parameter estimation
- Start Viterbi

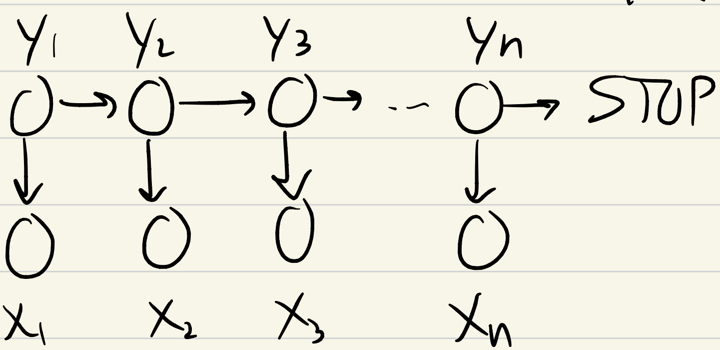
HMMs generative model $P(\bar{y}, \bar{x})$

(discriminative seq. models $P(\bar{y} | \bar{x})$

Conditional random fields CRFs)

Tags $y_i \in \overset{\text{tagset}}{\tilde{\mathcal{T}}}$ words $x_i \in \mathcal{V}$

$$P(\bar{y}, \bar{x}) = P(y_1)\, P(x_1 | y_1)\, P(y_2 | y_1)\, P(x_2 | y_2)\, P(y_3 | y_2)\, P(x_3 | y_3) \ldots P(\text{STOP} | y_n)$$



Generative "story":

sample $y_1$

draw $x_1 | y_1$ as first word

draw $y_2 | y_1$ as next tag

etc.

Two big assumptions:

(1) Each $x_i$ is indep. of the other vars. given $y_i$

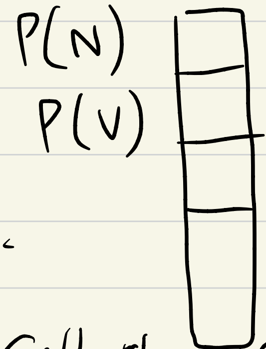
(2) Each $y_i$ is indep. of $y_1 \ldots y_{i-2}$ given $y_{i-1}$ (Markov)

## Parameters

Three types of params

$P(y_1)$

initial dist.



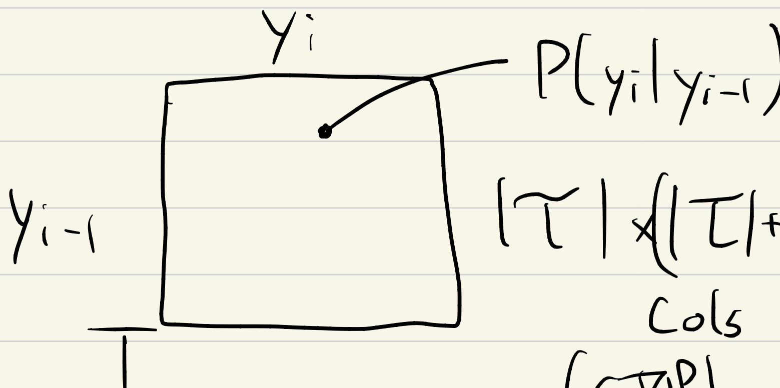

$|T|$-len vector

sums to 1

Call this $S$

$P(y_i | y_{i-1})$

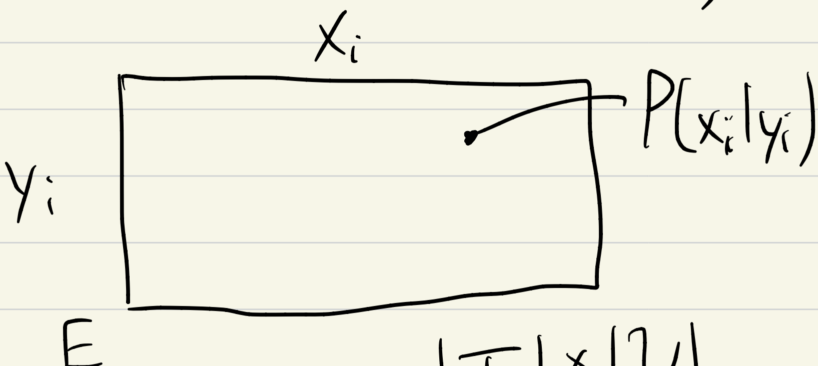
transition probs.




$|T| \times (|T|+1)$ cols (STOP)

$P(x_i | y_i)$

emissions




$|T| \times |V|$

Goals What can we do?

Sample $(\bar{y}, \bar{x}) \sim P(\bar{y}, \bar{x})$
tags, words

get a random sentence

Sample $\bar{x} \sim P(\bar{x} | \bar{y})$ if we're given tags $\bar{y}$

What we really want:

Given $\bar{x}$, find $\underset{\bar{y}}{\text{argmax}}\ P(\bar{y} | \bar{x})$

most likely tags for $\bar{x}$

Ex $\mathcal{T} = \{N, V, STOP\}$

$\mathcal{V} = \{\text{they}, \text{can}, \text{fish}\}$

S $P(y_1)$

| | |
|---|---|
| N | 1.0 |
| V | 0 |
| STOP | 0 |

T

| | N | V | STOP |
|---|---|---|---|
| N | 1/5 | 3/5 | 1/5 |
| V | 1/5 | 1/5 | 3/5 |

E

| | they | fish | can |
|---|---|---|---|
| N | 1 | 0 | 0 |
| V | 0 | 1/2 | 1/2 |

① Prob. of $\begin{pmatrix} N & V & V \\ \text{they} & \text{can} & \text{fish} \end{pmatrix}$

$P_{init}(N)\, P_e(\text{they} \mid N)$

② Write another sent w/ P>0

$V|N \quad V|V \quad STOP|V$

$y_2|y_1$

1) $1.0 \cdot 3/5 \quad 1/5 \quad 3/5$

$\cdot \quad \cdot \quad \cdot$

$1.0 \quad 1/2 \quad 1/2$

multiply all these

2) N V STOP

they fish

N V V

they fish can

3) Other tags for they can fish?

No

Training Given labeled sequences

$$\left(\bar{x}^{(i)}, \bar{y}^{(i)}\right)_{i=1}^{D}$$

Data: $\begin{array}{cc} N & V \\ \text{they} & \text{can} \end{array}$ $\quad$ $\begin{array}{cc} N & V \\ \text{they} & \text{fish} \end{array}$

How to maximize likelihood of data?

$$(\log)\ P(\bar{y}, \bar{x})$$

Can compute Max. likel. params. exactly. Count + normalize

Biased coin: heads w/prob $p$

Observe: HHTH

$p = 3/4$

Likelihood: $P(H)^3 P(T)$

$= p^3(1-p)$

Take derivative $= 0$

$3p^2 - 4p^3 = 0$

$3p^2 = 4p^3 \quad p = 3/4$

E

| | they | can | fish |
|---|---|---|---|
| N | 2 | 0 | 0 |
| V | 0 | 1 | 1 |

Counts

normalize $\Rightarrow$

| | | |
|---|---|---|
| 1 | 0 | 0 |
| 0 | 1/2 | 1/2 |

0.05 0.9 0.05

T

| | N | V | STOP |
|---|---|---|---|
| N | 0 | 2 | 0 |
| V | 0 | 0 | 2 |

0.1 2.1 0.1

0.1 0.1 2.1

$\Rightarrow$

| | | |
|---|---|---|
| 0 | 1 | 0 |
| 0 | 0 | 1 |

Smoothing pretend we saw everything w/ some small count

Bulk add 0.1 to every cell

## Inference in HMMs

Goal: $P(\bar{y} | \bar{x})$ tagger

But HMM is $P(\bar{x}, \bar{y})$

$\underset{\bar{y}}{\operatorname{argmax}} P(\bar{y} | \bar{x})$ Two problems:

1. Lots of $\bar{y}$ possible
2. We don't have $P(\bar{y} | \bar{x})$

Problem 2:

$$P(\bar{y} \mid \bar{x}) = \frac{P(\bar{y} \mid \bar{x}) \cdot P(\bar{x})}{P(\bar{x})}$$

$$= \frac{P(\bar{y}, \bar{x})}{P(\bar{x})} \leftarrow \max \bar{y}$$

We care about $\underset{\bar{y}}{\operatorname{argmax}}$

~~Bottom:~~ $\cancel{P(\bar{x}) = \sum_{\bar{y}} P(\bar{y}, \bar{x})}$

$$\underset{\bar{y}}{\operatorname{argmax}} P(\bar{y} \mid \bar{x}) = \underset{\bar{y}}{\operatorname{argmax}} P(\bar{y}, \bar{x})$$

Given $\overline{x}$,

$$\underset{\overline{y}}{\operatorname{argmax}}\ P(\overline{y}, \overline{x}) = \underset{\overline{y}}{\operatorname{argmax}}\ \log P(\overline{y}, \overline{x})$$

Let $\tilde{y} = \tilde{y}_1 \ldots \tilde{y}_n$ be "pred" $y$

$$= \underset{\tilde{y}_1 \ldots \tilde{y}_n}{\operatorname{argmax}}\ \log P(\tilde{y}_1) + \log P(x_1 | \tilde{y}_1) + \log P(\tilde{y}_2 | \tilde{y}_1) + \ldots$$

Dynamic programming (Viterbi) to find $\tilde{y}$