CS 378 Lecture 10
HMMs, Viterbi

Announcements

- A2 due today
- Bias response due today
- A3 out today
- Midterm: in 2.5 weeks
- Final proj: if independent project, start thinking

Recap POS tagging

NNP	VBZ	NN	NNS
Fed	raises	interest	rates

How to do with a classifier?

$$P(y_i = \underset{\substack{\uparrow \\ \text{prediction}}}{\hat{y}} \mid \bar{x}) = \frac{e^{\bar{w}_{\hat{y}}^T f(\bar{x}, i)}}{\sum_{y \in \mathcal{Y}} e^{\bar{w}_y^T f(\bar{x}, i)}}$$

index these + treat as your features

$$f(\bar{x}, i) = \begin{cases} \text{Prevword} = \ldots \\ \text{Currword} = \ldots \\ \text{Nextword} = \ldots \end{cases}$$

Problem: all y_i are predicted independently

How do we model the sequence all at once?

Today Hidden Markov Models

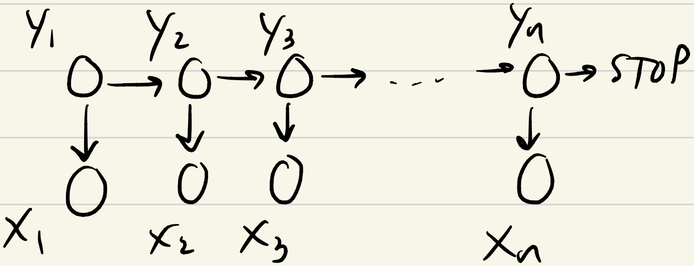
- Definition
- Training
- Inference (Viterbi)

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

There are discriminative seq. models, Conditional Random Fields $P(\bar{y} | \bar{x})$

HMMs Tags $y_i \in \overset{\text{tag set}}{\mathcal{T}}$, words $x_i \in \mathcal{V}$

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



Generative story: draw $\overset{\text{NNP}}{y_1}$ to start tag sequence

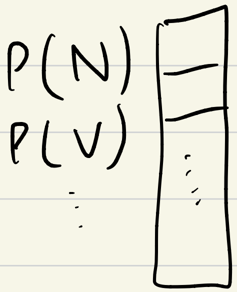
draw $\overset{\text{Fed}}{x_1} | y_1$ as the first word

draw $y_2 | y_1$ as the next tag

Parameters

$P(y_i)$

initial distribution

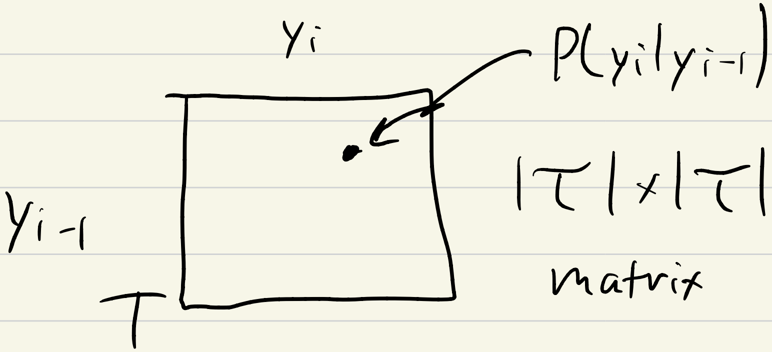


$|T|$-len vector that sums to 1

call this S

$P(y_i \mid y_{i-1})$

transitions

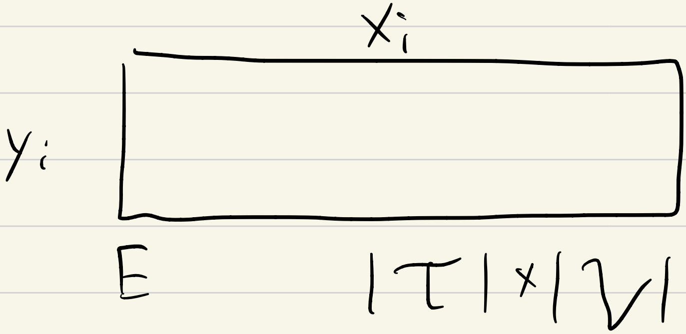


$|T| \times |T|$ matrix

each row sums to 1

$P(x_i \mid y_i)$

emissions



$|T| \times |V|$ matrix

why Markov model: $P(y_i | y_{i-1})$ y_i doesn't depend on y_{i-2}

Goals with HMMs

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

What we want: $P(\bar{y} | \bar{x}) = \frac{P(\bar{y}, \bar{x})}{P(\bar{x})}$

<u>Ex</u> $\mathcal{T} = \{N, V, STOP\}$ $\mathcal{V} = \{$ they, can, fish $\}$

S

	$P(y_1)$
N	1.0
V	0.0
STOP	0.0

T

		y_i		
		N	V	STOP
y_{i-1}	N	1/5	3/5	1/5
	V	1/5	1/5	3/5

E

		they	can	fish
	N	1	0	0
y_i	V	0	1/2	1/2

Poll

① What is the probability of

$$\left(\overset{N}{\text{they}}\ \ \overset{V}{\text{can}}\ \ \overset{V}{\text{fish}}\right)$$

$P(\bar{y}, \bar{x}) = P(N)\,P(\text{they} \mid N)$ ..

$1.0 \cdot 1.0$ ...

$1.0 \cdot 3/5 \cdot 1/5 \cdot 3/5 \leftarrow P(\text{STOP} \mid V)$

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

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

② $\overset{N}{\text{they}}\ \ \overset{V}{\text{fish}}\ \ \overset{V}{\text{can}}$ $\qquad$ $\overset{N}{\text{they}}\ \ \overset{V}{\text{can}}$

$P(y_1 = V) = 0.0$

~~V N V~~

~~can they fish~~

③ other seqs for they can fish?

No

Training Given labeled sequences

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

Data: $\overset{N}{\text{they}}\ \overset{V}{\text{can}}\quad \overset{N}{\text{they}}\ \overset{V}{\text{fish}}$

What parameters maximize the likelihood of the data?

Can find this *exactly* by counting + normalizing

Biased coin w/probability p of heads

See HHHT

$p=3/4$ is the MLE

emissions for V

$$P(x_i \mid V) = \begin{array}{l} \text{can: } 1 \\ \text{fish: } 1 \\ \text{they: } 0 \end{array} \Rightarrow \begin{array}{ll} \text{they} & 0 \\ \text{can} & 1/2 \\ \text{fish} & 1/2 \end{array}$$

This is what "training" involves

$$P(y_i \mid y_{i-1} = V) = \begin{array}{l} \text{counts} \\ N : 0 \\ V : 0 \\ \text{STOP} : 2 \end{array} \Rightarrow \begin{array}{l} \text{probs} \\ N : 0 \\ V : 0 \\ \text{STOP} : 1 \end{array}$$

<u>Smoothing</u>: pretend we saw everything with some nonzero count

Add 0.1 to every count

$$\begin{array}{ll} N & 0 \\ V & 0 \\ \text{STOP} & 2 \end{array} \Rightarrow \begin{array}{ll} N & 0.1 \\ V & 0.1 \\ \text{STOP} & 2.1 \end{array} \Rightarrow \text{probs}$$

Inference in HMMs

We defined HMMs as $P(\bar{y}, \bar{x})$

What we really want is a tagger

$P(\bar{y} | \bar{x})$ "you give me a sent, I tell you tags"

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

① Space is exponentially large $(|\mathcal{Y}|^n)$

② we don't have $P(\bar{y} | \bar{x})$ yet

Solving #2:

$$\underset{\bar{y}}{\text{argmax}}\, P(\bar{y} | \bar{x}) = \underset{\bar{y}}{\text{argmax}} \frac{P(\bar{y} | \bar{x}) \cdot P(\bar{x})}{P(\bar{x})}$$

$$= \underset{\bar{y}}{\text{argmax}} \frac{P(\bar{y}, \bar{x})}{P(\bar{x})}$$

$$P(\bar{x}) = \sum_{\bar{y}} P(\bar{y}, \bar{x})$$

We don't need this

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

Conclusion: $\underset{\bar{y}}{\operatorname{argmax}} P(\bar{y}, \bar{x})$

to build our tagger #2 ✓

Solving #1: the Viterbi algorithm

First: move to log space

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

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

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

Basic concept: dynamic programming

Best sequence of $\tilde{y}_1, \ldots \tilde{y}_n$ can be found from

best sequence from $\tilde{y}_1, \ldots, \tilde{y}_{n-1}$

+ HMM probs

optimal seq for each possible $\tilde{y}_{n-1}$