CS 371 Lecture 14
Sequence Modeling I: Part of speech

Announcements

- A3 due today
- A4 posted due in 12 days

Recap Language modeling

1. Tokenization (subword) / featurization
2. Pre-training phase: skip-gram
   language modeling
3. Fine-tuning phase: train something like in A2
   or fine-tune BERT/GPT/...

4. Inference

classify

generate:

prompting

methods: greedy, nucleus sampling

<u>Today</u> Structured prediction

- sequence modeling: part of speech tagging
- syntactic parsing

today: POS and Hidden Markov Models

<u>Part-of-speech tagging</u>

Input: sentence $x_1 \ldots x_n$

Output: POS tags $y_1 \cdots y_n$ for each word

What are POS tags?

N	N	V	N
N	V	ADJ	N
teacher	strikes	idle	kids

Predicting POS $\Leftrightarrow$ interpreting the sentence

Text-to-speech: record verb or noun

POS tags

open-class: new words with these tags are always emerging

closed-class: known set

Open-class:

(N) Nouns:
- → Proper (Google) NNP
- → common (shoe) NN
- → plural (NNS) vs. singular

(V) Verbs: features like tense, person (1st or 3rd)

In standard datasets: VBZ "3rd person present singular verb"

VBD: past tense

(J) Adjectives → yellow, idle

(RB) adverbs → swiftly

Closed-class

(DT) Determiner: articles (the, a) some, many

(CD) cardinals: numbers

(IN) prepositions : up, on, in,

(RP) particles : made <u>up</u>

Modals (could / would / should), auxiliary verbs (had)

1. What tags are possible for each word?
2. what sentences make sense?

Fed raises interest rates 0.5 percent

Fed
- NNP Federal Reserve
- VBD "They fed me"
- VBN "I was fed up"

raises
- NNS $ raises
- VBZ "she raises"

interest
- NN
- VBP "Pyramids interest me"
- VB "I want NLP to interest me"

rates
- NNS
- VBZ

0.5 CD

percent NN

<u>Sentences:</u> standard, alt 1, alt 2, alt 3

Methods for POS tagging

(later) Hidden Markov Models

(now) Classifiers

Classifier POS tags $\mathcal{Y}$

MC class. $P(y \mid \bar{x}) \in \mathcal{Y}$

For seqs: $P(y_i = t \mid \bar{x}, i)$

y_1 y_2

Fed raises

Run classifier twice

BoW $\times$ no position info

Position-sensitive BoW: "Unigram = Fed & offset = −1"

predicting y_2

$$P(\bar{y} \mid \bar{x}) = \prod_{i=1}^{n} P(y_i \mid \bar{x}, i)$$

independent classifier

this combo is bad

	N	N	
	$\boxed{V}$	$\boxed{V}$	
Fed	raises	interest	rates

(y_2, y_3) should not be (V, V)

Instead we want a <u>sequence model</u>

really model $P(\bar{y} \mid \bar{x})$

the whole sequence

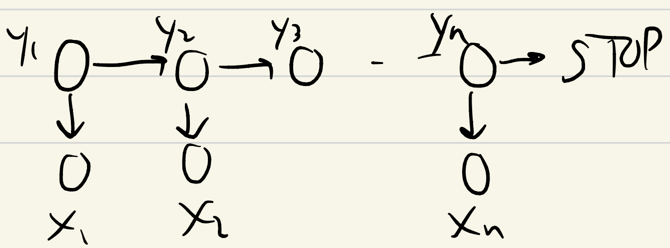
<u>HMMs</u> models of sequences, can capture

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

<u>generative</u> models $P(\bar{x}, \bar{y})$

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

$$P(y_1)\,P(x_1 \mid y_1)\,P(y_2 \mid y_1)\,P(x_2 \mid y_2)\,P(y_3 \mid y_2)\cdots P(\mathrm{STOP} \mid y_n)$$



Assumptions

① $\bar{y}$s are modeled with a "bigram LM"
(Markov property: y_i is conditionally independent of $y_1 \ldots y_{i-2}, x_1 \ldots x_{i-1}$ given y_{i-1})

② Each x_i is indep. of everything else given y_i

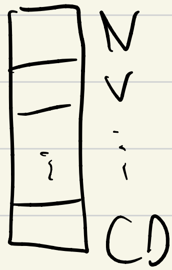
Generative story: ① Pick y_1 ② Pick $x_1 | y_1$ ③ Pick $y_2 | y_1$ ④ Pick $x_2 | y_2$ ...

Goal: model $P(\bar{x}, \bar{y})$, but ultimately we want $P(\bar{y} | \bar{x})$

$\mathcal{V}$ vocab, $\mathcal{T}$ tags

Three types of parameters:

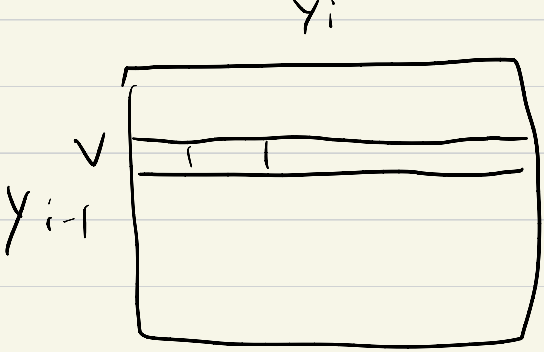
$P(y_1)$ initial distribution



$|\mathcal{T}|$-len vector, adds to 1

Transition probs

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



$\rightarrow P(y_i \mid V)$

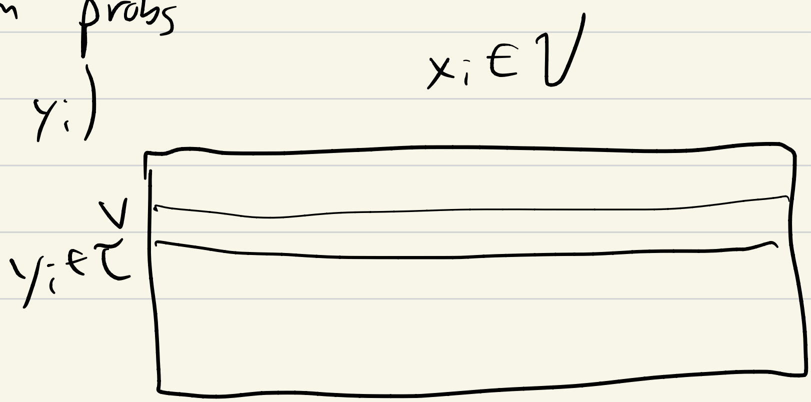
" 70% N
0% V
⋮ "

$|\mathcal{T}| \times (|\mathcal{T}|+1)$ matrix

↖ STOP

Emission probs

$P(x_i \mid y_i)$



$P(x_i \mid V)$

" 5% go
5% eat
⋮ "