CS 378 Lecture 9
Sequence Tagging: Classifiers, HMMs

Today
- Using classifiers for sequence labeling
- Hidden Markov Models
- Parameter estimation
- (if time) inference

Announcements
- A1
- A2
- A3

Recap Sequence labeling

Input = $\bar{x} = (x_1, \cdots, x_n)$

Output = $\bar{y} = (y_1, \cdots, y_n)$ $y_i \in$

x_1 Fed	x_2 raises	x_3 interest	rates	0.5	percent
y_1 N	y_2 V	y_3 N	N	CD	N
(V)	(N)	(V)	(V)		

Tagging with Classifiers

$P(y_i = y \mid \bar{x})$ with Logistic Regression?

Run n times — ^ MC

$P(y_3 = N \mid \bar{x})$? Classify "interest" as N or V or ...

① Using BoW features

$f(\bar{x}) = [0\ 0\ 1\ 0\ 0\ 1\ 0\ 0\ 0\ 1\ \ldots]$

raises Fed interest

$y = N$ or V

$$P(y_3 = N \mid \bar{x}) = \frac{e^{\bar{w}_N^T f(\bar{x})}}{\cdots}$$

$$P(y_2 = N \mid \bar{x}) = \frac{e^{\bar{w}_N^T f(\bar{x})}}{\cdots}$$

Same

New feature vector: $f(\bar{x}, i)$ depends on posn

$f(\bar{x}, i)$ → skip-gram

Fed raises interest rates ...

i = 1 2 3

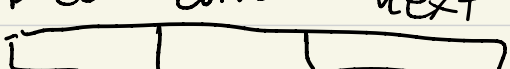
$f(\bar{x}, i=3) = [0\,0\,0\,0 \;\; 1 \;\; 0\,0\,0\,0 \ldots]$

interest

BoW only including current word

Look at context words

prev current next

$f(\bar{x}, i=3) =$ 

raises interest rates

3 skip-gram concatenated

$\overline{w}_N$ $+1.2$

$\overline{w}_{PREP}$ -10

$f(\bar{x}, i=3) = [\ 0 \quad 1 \quad \cdots \quad 1 \quad \cdots$

Curr word = interest

Prev word = raises

$1 \quad \cdots \quad 0 \quad \cdots]$

next word = rates

curr word = rates

"Feature" = descriptor of the context the POS tag occurs in

$3|V|$ features

"vector length

Treat each POS tag as a training ex

N	V	N	N	$\overline{Y}$
Fed	raises	interest	rates	$\overline{X}$

4 examples $y_1 = N \mid \overline{X}$

$y_2 = V \mid \overline{X}$

Problem w/ POS as classification

VBZ ≠ VBP present tense verbs

Fed raises interest rates

We never have these two tags in a sequence like this

If our classifier makes a mistake, sequence of tags makes no sense

Constraint: no VBZ–VBP
Build this into model

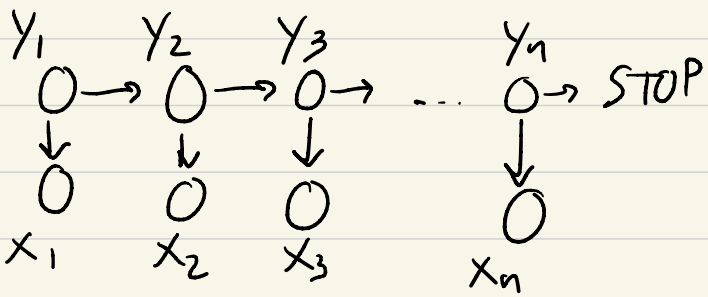
Hidden Markov Models: model the entire sequence of tags

Generative: HMMs
Discriminative seq. models: conditional random fields (CRFs)
"HMMs + LR"

Hidden Markov Models

Tags $y_i \in \underset{\text{tags}}{\mathcal{T}}$, $x_i \in \mathcal{V}$ words

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



"Generative story": draw a tag y_1
to start the sentence
draw tag y_i based on y_{i-1}
word x_i based on tag y_i

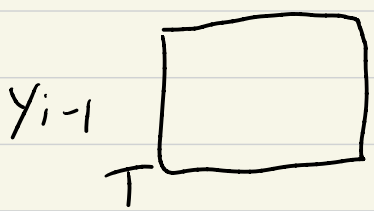
Params : $P(y_1)$

initial distribution

$|\tau|$-len vector that sums to 1

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

transitions

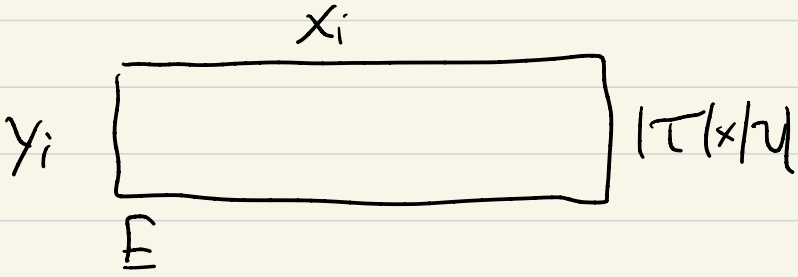


$|\tau| \times |\tau|$ matrix

each row sum to 1

$P(x_i | y_i)$

emissions



Why called HMM? y's form a Markov process: y_i is independent of $y_1 \dots y_{i-2}$ conditioned on y_{i-1}

Goals

(1) Inference: $\underset{\bar{y}}{\text{argmax}}\ P(\bar{y}|\bar{x})$

(2) Estimation: labeled data $(\bar{x}^{(i)}, \bar{y}^{(i)})$

$\Rightarrow$ params

Ex $T = \{N, V, STOP\}$ $V = \{they, can, fish\}$

$P(y_1)$ "we start sentences w/ N w/ probability 1"

S

N	1.0
V	0.0

T

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

"when I'm at a N, the next tag sampled is V w/ prob 3/5"

E

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

"if I'm a V, we see word can w/ prob 1/2 and fish w/ prob 1/2"

Poll

N V V sent +tags

they can fish

V|V STOP|V

① Prob: initial $P(N)=1$ $P(V|N)=3/5$ $1/5$ $3/5$

$P(\text{they}|N)=1$ $P(\text{can}|V)=1/2$ $1/2$

$= 1 \cdot 1 \cdot 3/5 \cdot 1/2 \cdot 1/5 \cdot 1/2 \cdot 3/5$

② other seqs

they they can they fish they fish can ...

③ other tags: no, all zero prob

Estimating Parameters

Maximize log likelihood of observed data

Data: N V two labeled sents

they can

N V

they fish

What params make these sents look likely?

// Maximum likelihood estimation

MLE by counting + normalizing

When you have categorical distributions (coin with n sides) and you see m trials, prob of each outcome is just $\frac{\text{count (outcome)}}{\text{Count (trials)}}$

coin {H,T}

HHHT ⇒ prob of H = 3/4

$P(x_i \mid V) =$ Can: count 1, fish: count 1, they: count 0 ⇒

th	0
can	1/2
fish	1/2

⇒ exactly the emissions we saw before

Transitions were smoothed: pretended we saw every event w/ count 1

N → N 0
N → V 2
N → STOP 0

⇒ 1, 3, 1 ⇒ 1/5, 3/5, 1/5