0. HW4A

1. Log-Linear Models (A) Class. (B) Sequence

2. Perceptron

3. Structured Perceptron

CS585
UMass Amherst
10/18/16

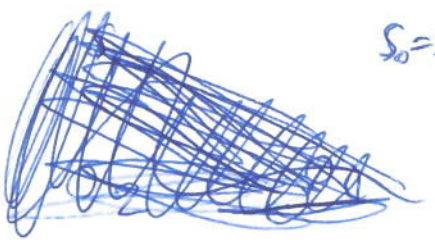
HW4 A

(1) 1-3: Can answer just for whether labels is general for most HMM params
4-5: Numbers

(2) Implicit $S_0 = \text{START}$

(3) If you know the sequence length: ~~implies~~ is n:
implies $S_{n+1} = \text{END}$

$$P(O_2 | O_1) \iff P(O_2 | O_1, S_0 = \text{START})$$



$S_0 = \text{START} \rightarrow S_1 \rightarrow S_2 \rightarrow \dots$
$\downarrow \quad \downarrow$
$O_1 \quad O_2 \quad \dots$

Homework 4 Part A: HMM, Viterbi

CS 585, UMass Amherst, Fall 2016

Overview

Due **Friday Oct 21**.

Get starter code from the course website's schedule page. You should submit a zipped directory (please don't use other compression formats like .rar) named hw4a_YOUR-USERNAME that contains:

- your vit.py file
- writeup

Our course's collaboration policy is specified on the website.

1 HMM

[15 total points]

Answer the following questions using the transition matrix T and emission probabilities E below. Below, Δ and $\square$ are two output variables, A and B are two hidden states; s_n refers to the n^{th} hidden state in the sequence and o_n refers to the n^{th} observation.

$T =$

	A	B	END
START	0.5	0.5	0.0
A	0.2	0.3	0.5
B	0.4	0.4	0.2

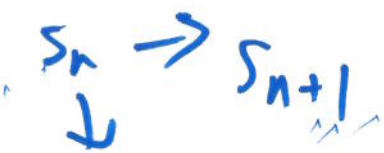
$E =$

	Δ	$\square$
A	0.5	0.5
B	0.3	0.7

1. *[2 points]* Does $P(o_2 = \Delta | s_1 = B) = P(o_2 = \Delta | o_1 = \square)$?
2. *[2 points]* Does $P(s_2 = B | s_1 = A) = P(s_2 = B | s_1 = A, o_1 = \Delta)$?
3. *[3 points]* Does $P(o_2 = \Delta | s_1 = A) = P(o_2 = \square | s_1 = A, s_3 = A)$?
4. *[3 points]* Compute the probability of observing $\square$ as the first emission of a sequence generated by an HMM with transition matrix T and emission probabilities E.
5. *[5 points]* Compute the probability of the first state being A given that the last token in an observed sequence of length 2 was the token Δ.

$$P(s_2 \mid s_1, s_3) = \frac{P(s_1, s_2, s_3)}{P(s_1, s_3)}$$

$$= P(s_2 \mid s_1, s_3, o_3) = \frac{P(s_1\ s_2\ s_3)}{\sum_{s_2'} P(s_1, s_2', s_3)}$$



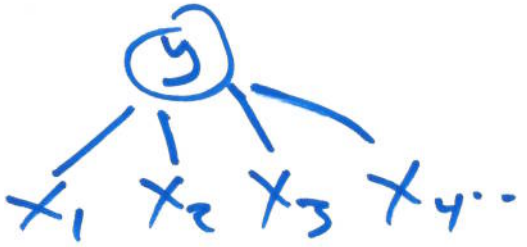
$$P(s_n, o_n, s_{n+1}) = P(s_n)\, P(o_n \mid s_n)\, P(s_{n+1} \mid s_n)$$

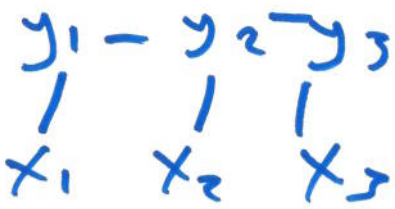
Log-Linear (Discriminative) Models

$P(y|x)$

y: doc class .... sequence of POS tags

x: ~~text~~





HMM

$P(y,x) = \prod_t P(y_t|y_{t-1}) P(x_t|y_t)$

$$\log P(y|x) = C + \theta^T f(x,y)$$

θ: Model Params

$f(x,y)$: Feature Vector for (x,y)

Classif: "Multi-class Log. Reg." $x =$ "I ♡ cats"

	POS_I	POS_♡	pos_dog	...	NEG_I
$f(x, POS) =$	1	1	0	...	0
$f(x, NEG) =$	0	0	0	...	1
$\theta =$	+2.1	3.9	0.3	...	...

Decision $\underset{y'}{\text{argmax}}\ P(y'|x) = \theta^T f(x, y')$

$\theta^T f(x, \text{POS})$ vs. $\theta^T f(x, \text{NEG})$

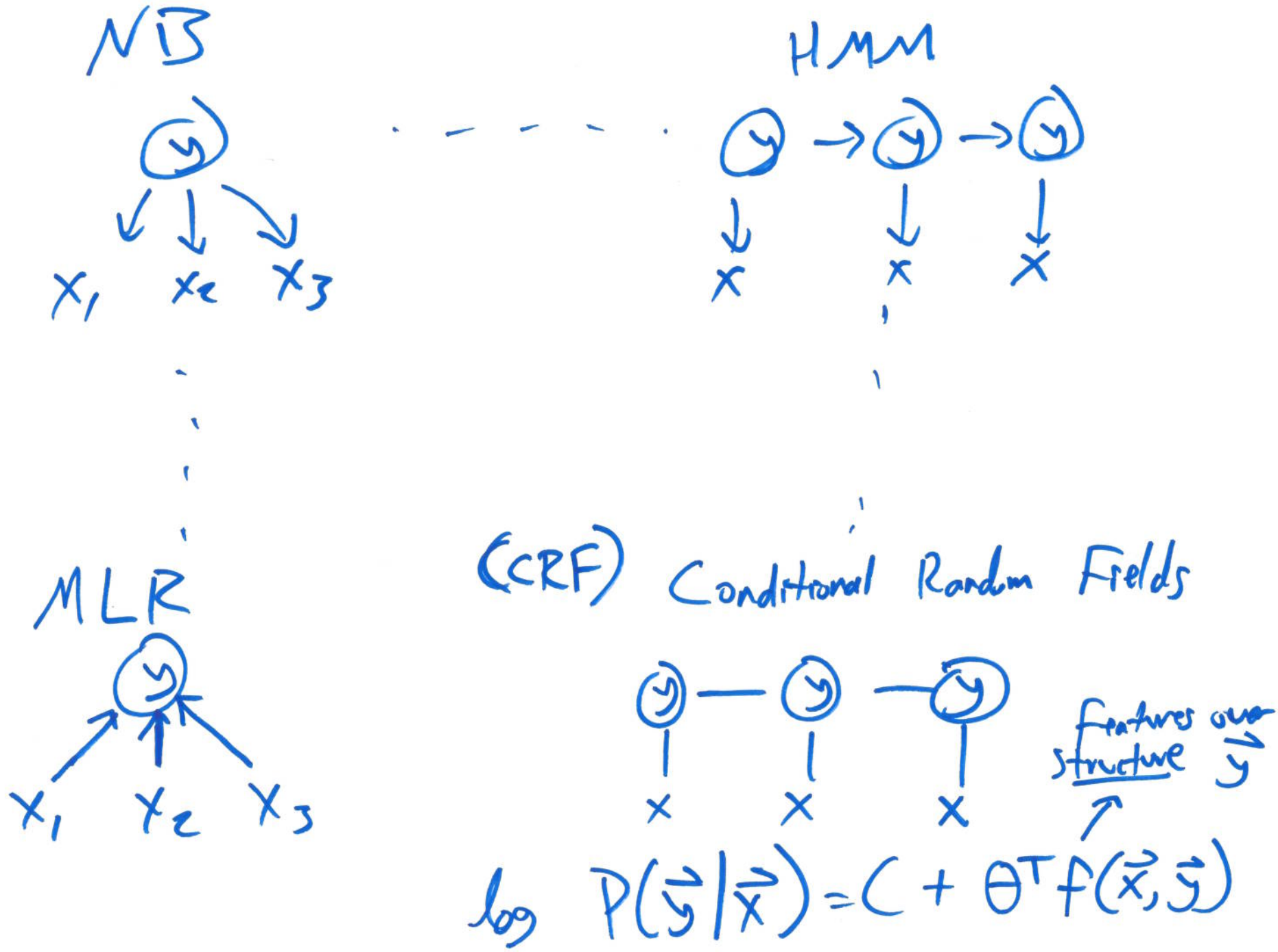
$= \sum_{j=1}^{J} \theta_j f_j(x, \text{POS})$

$\theta' f(x, \text{POS})$

NB as Log-Lin?

$f_{w,\text{POS}}(x, y) =$ Count of word w in x when $y = \text{POS}$
else 0

$\theta_{w,\text{POS}} \equiv \log P(w|\text{POS})$


NB
HMM
MLR
(CRF) Conditional Random Fields
features over structure
$\log P(\vec{y}|\vec{x}) = C + \theta^T f(\vec{x}, \vec{y})$

HMM as Log-Linear $\log P(N|N) \Rightarrow K \times K$

$y =$ A N N N $\log P(\text{Lincoln}|N)$

$x =$ happy President Abe Lincoln $\Rightarrow V \times K$

Feat Vector: for $\log P(y|x) = C \ldots + \theta^T f(x,y)$

$f(x,y) = ($ #A,N: 1, A,V: 0, N,N: 2, $\ldots$ A, happy: 1, V, happy: 0, $\ldots)$

Trans. Feats $\Rightarrow K \times K$ | obs. feats $\Rightarrow V \times K$

Params! $\downarrow$ Count(tag $\{k\}$ after $\{j\}$) Count('tag $\{y\}$ appears above $\{w\}$')

$\theta = (+2.1, -4.1, +4.4 \ldots \ldots +110.2 \ldots)$

Predict: $\underset{y'}{\text{argmax}}\ \theta^T f(x,y') \Leftrightarrow \underset{y'}{\text{argmax}}\ P(y'|x) = P(y',x)$

Viterbi

Why CRFs? $P(y|x) = \frac{1}{Z} \exp(\theta^T f(x,y))$

(1) Discrim. Training of θ

(2) More features

(HMM) Trans from j to k

(HMM) observe word w below k

- Word ends in -ing below tag k
- Word begins with a capital letter

Learning: Perceptron Algo (Structured)

Error-Driven Learning

Rosenblatt 1957

Converge?

Averaging

Algo

For each pass thru data (~30 iter)

For each $(x, y^{(gold)})$:

Pred $y^* = \arg\max_{y'} \theta^T f(x, y')$

Update $\theta := \theta + r[f(x, y^{(gold)}) - f(x, y^*)]$

$r = 1$

$\Rightarrow \theta^T f(x, y^{(gold)})$ gets bigger

What if $y^* = y^{(gold)}$?