Sequence labeling problems

- Many problems in NLP can be formulated as sequence labeling problems
  - POS tagging:
    - The_DT man_NN who_WP whi _ _VBZ
  - Named
    - T _ _ _ Microsoft_B-ORG co _ _ _ _O venture_O capitalist_O Andre _B- _ R
  - Time expression detection
    - Bedford_O police_O said_O they_O received_O a_O call_O about_O 3:45_B-TIMEX p.m._I-TIMEX Monday_B-TIMEX
  - Spoken language understanding
    - Which_O flights_FLIGHT arrive_ARRIVE in_O Burbank_CITY from_O Denver_CITY on_ON Saturday_Day
  - .....

Search and Learning

Recall most natural lan
mathematically as optimization:

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} \Psi(\boldsymbol{x}, \boldsymbol{y})$$

There are two modu

- Search, the module that is responsible for
  of the score function Ψ
- Learning, the module that is responsible f
  optimal parameters $\boldsymbol{\theta}$

For simple text classification problems, the search module is fairly straightforward, and most of the work goes to learning. For sequence labeling and more complicated NLP problems, the search module is getting more complicated.

Sequence labeling: first idea

- Classify the se

Sequence labeling example: POS tagging

- Let's use POS tagging as an example
- The most common used data set for training P
  the Penn TreeBank

DT					NNS
The					pianos

- DT: Determiner
- NN: uncountable noun or noun in singula
- WP: Wh-pronoun
- VBZ: 3rd person singular verb
- NNS: plural noun

How do we extract features from sequences in a linear model?

- We take as input a sequence of word tokens $\boldsymbol{x}$, their corresponding POS tags $\boldsymbol{y}$, as well as a po , and return a set of features
- Typically we matters for classifying the word at positi re its surrounding words. We define a window t position m of size k, and only extract contextual information from this window.

Extracting features from a window size of 1

- Assuming a window of 1, the features we will be extracting from the example sentence will be:

$\boldsymbol{f}((\boldsymbol{w} = \text{t}, m = 1), DT)$

$= (w$

$\boldsymbol{f}((\boldsymbol{w} = \text{the man who whistles tunes pi}, m = 2), NN)$

$= (w_0 = man, NN)$

.....

How many features will we extract if we use a window of size 1?

Weights

We can then train a classifier using these features and get a weight for each feature:

	DT	NN			NS
w_0=the	-0.05	-3.9	-		4.6
w_0=ma				4	-3.5
w_0=wh				6	-4.6
w_0=whistles	-4.6	-4.6			-0.63
w_0=tunes	-4.6	-4.6			-0.6
w_0=pianos	-4.6	-4.6	-4.6	-3.0	-0.08

For example, the weight $\theta_1 = -0.05$ for the feature $f_1(w_0 = the, DT)$

Using these weights we can classify each word in the sequence

$$\psi((\boldsymbol{w} = \text{the man who whistles tunes pianos}, m = 1), DT)$$
$$= \sum_i f_i \theta_i = -0.05$$

$$\psi((\boldsymbol{w} = \text{the man who whistles tunes pianos}, m = 1), NN)$$
$$= \sum_i$$

$$\psi((\boldsymbol{w} = \quad \text{tunes pianos} \quad = 1), WP)$$
$$= \sum_i f_i \theta_i = -4.6$$

$$\psi((\boldsymbol{w} = \text{the man who whistles tunes pianos}, m = 1), VBZ)$$
$$= \sum_i f_i \theta_i = -4.6$$

$$\psi((\boldsymbol{w} = \text{the man who whistles tunes pianos}, m = 1), NNS)$$
$$= \sum_i f_i \theta_i = -4.6$$

Predicting the tag for each word in the sequence

After finding the argmax at all positions of the sentence, we get:

DT	NN	WP	NNS		
The	man	who	whistles		s

				Z	NNS
w_0=the				6	-4.6
w_0=man	-4.6	-0.35			-3.5
w_0=who	-4.6	-4.6			-4.6
w_0=whistles	-4.6	-4.6	-4.6	-0.8	-0.63
w_0=tunes	-4.6	-4.6	-4.6	-0.8	-0.6
w_0=pianos	-4.6	-4.6	-4.6	-3.0	-0.08

Extracting features from a larger window

- If we increase the window size to 2 and also include the previous word in the context

$\boldsymbol{f}((\boldsymbol{w} = \text{the man who whistles tunes pianos}, m = 1), DT)$
$= \{(w \qquad\qquad\qquad\qquad\qquad\qquad T)\}$

$\boldsymbol{f}((\boldsymbol{w} = \text{th} \qquad\qquad\qquad\qquad\qquad\qquad m = 2), NN)$
$= \{(w_0 = man, NN), (w_{-1}$

$\ldots\ldots$

$\boldsymbol{f}((\boldsymbol{w} = \text{the man who whistles tunes pianos}, m = 4), VBZ)$
$= \{(w_0 = whistles, VBZ), (w_{-1} = who, VBZ)\}$

Include weights for the new features

	DT	NN	WP	VBZ	NNS
w_0=the	-0.05	-3.9	-4.6	-4.6	-4.6
w_0=man	-4.6	-0.35	-		.5
w_0=who	-3.6	-4.6	-		.6
w_0=wh				8	-0.63
w_0=tu				8	-0.6
w_0=pia				0	-0.08
w_{-1}=START	-0.92	-3.9			-0.92
w_{-1}=the	-4.6	-0.7			-0.75
w_{-1}=man	-1.6	-2.3			-2.3
w_{-1}=who	-1.8	-4.6	-4.6	-0.2	-4.6
w_{-1}=whistles	-2.3	-4.6	-4.6	-1.6	-0.4
w_{-1}=tunes	-1.6	-4.6	-4.6	-4.6	-0.26

Classification with the new weights

$$\psi((\boldsymbol{w} = \text{th} \quad \text{nes pianos} \quad), DT)$$
$$= \sum_i f_i\theta_i = -4.6 - 1.8 = -6.4$$

$$\psi((\boldsymbol{w} = \text{the man who whi} \quad \text{es pianos} \quad N \quad)$$
$$= \sum_i f_i\theta_i = -4.6 - 4.6 = -9.2$$

$$\psi((\boldsymbol{w} = \quad \text{tunes pianos} \quad = 4), WP)$$
$$= \sum_i f_i\theta_i = -4.6 - 4.6 = -9.2$$

$$\psi((\boldsymbol{w} = \text{the man who w} \quad \text{nes pianos} \quad), VBZ)$$
$$= \sum_i f_i\theta_i = -0.8 - 0.2 = -1$$

$$\psi((\boldsymbol{w} = \text{the man who whistles tunes pianos}, m = 4), NNS)$$
$$= \sum_i f_i\theta_i = -0.63 + -4.6 = -5.23$$

So VBZ receives the highest score when classifying position 4.

Updated classification results

DT	NN	WP	VBZ	NNS	NNS
The	man	who	whistles	tunes	pianes

	DT	NN			NS
w_0=the	-0.05	-3.9	-4.6	-4.6	-4.6
w_0=man	-4.6	-0.35	-4.6	-1.4	-3.5
w_0=wh				6	-4.6
w_0=wh				8	-0.63
w_0=tunes	-4.6	-4.6			-0.6
w_0=pianes	-4.6	-4.6			-0.08
w_{-1}=START	-0.92	-3.9			-0.92
w_{-1}=the	-4.6	-0.7	-4.6	-4.6	-0.75
w_{-1}=man	-1.6	-2.3	-0.9	-1.6	-2.3
w_{-1}=who	-1.8	-4.6	-4.6	-0.2	-4.6
w_{-1}=whistles	-2.3	-4.6	-4.6	-1.6	-0.4
w_{-1}=tunes	-1.6	-4.6	-4.6	-4.6	-0.26

Sequence labeling as structured prediction

- Enlarging the window to include more context helps, to a degree
- To further improve the classifier requires als
  sequences of t
  NNS" should
  Incorporating such information in the m
  improve tagging accuracy
- The tags are of course not observable in the data, and they need to be predicted together.

Sequence labeling: Computing a global score for the entire sequence

- Consider all possible label sequences for the input sequence, and choose the one that has the highest score
- For a sequence of M elements wi N, there are N^M possible sequences, a very large number!
- To find the sequence with the highest score, we need to do this efficiently
- The common solution is the Viterbi Algorithm

Sequence labeling as structured prediction

- The goal of the mod
  the highest score f

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})}{\operatorname{argmax}} \Psi(\boldsymbol{w}, \boldsymbol{y})$$

- To make the co r the
  entire seque *m*

$$\Psi(\boldsymbol{w}, \boldsymbol{y}) = \sum_{m=1}^{M+1} \psi(\quad)$$

- The local score is a weighted sum of the local features at position m.

$$\psi(\boldsymbol{w}_{1:M}, y_m, y_{m-1}, m) = \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$$

Feature representation for sequences

$\boldsymbol{f}(\boldsymbol{w} = \text{the man who whistles tunes pianos}, \boldsymbol{y} = \text{DT NN WP VBZ VBZ NNS})$

$$= \sum_{m=1}^{M+1} \boldsymbol{f}(\boldsymbol{w}, y_m, y_{m-1}, m)$$

$$= \boldsymbol{f}(\boldsymbol{w}, DT, \diamond, 1) + \boldsymbol{f}(\boldsymbol{w}, NN, DT, 2) + \boldsymbol{f}(\boldsymbol{w},$$
$$+ \boldsymbol{f}(\boldsymbol{w}, VBZ, \qquad\qquad S, VBZ, 6)$$
$$+ \boldsymbol{f}(\boldsymbol{w}, \blacklozenge, NNS$$

$$= f(w_0 = the, y_0 = DT) + f(y_0 = DT, y$$

$$+ f(w_0 = man, y_0 = NN) + f(y_0 = NN$$
$$+ f(w_0 = who, y_0 = WP) + f(y_0 = W \qquad_{-1}$$
$$+ f(w_0 = whistles, y_0 = VBZ) + f(y_0 = VBZ, y_{-1} = WP)$$
$$+ f(w_0 = tunes, y_m = VBZ) + f(y_0 = VBZ, y_{-1} = VBZ)$$
$$+ f(w_0 = pianos, y_0 = NNS) + f(y_0 = NNS, y_{-1} = VBZ)$$
$$+ f(y_0 = \blacklozenge, y_{-1} = NNS)$$

Decoding for sequences: The Viterbi algorithm

- The goal is to find the sequence of tags with the highest score:

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{w})} \Psi(\boldsymbol{w}, \boldsymbol{y})$$

$$= \operatorname*{argmax}_{\boldsymbol{y}_{1:M}} \sum_{m=1}^{M+1} s(\boldsymbol{y}_{m-1}, \boldsymbol{y}_m, \boldsymbol{w}, m)$$

- Instead of finding the argmax for the entire sequence directly, we start by finding the max up to position m and keep a sequence of back pointers

Finding the max score for the sequence

$$\max_{\boldsymbol{y}_{1:M}} \Psi(\boldsymbol{x}, \boldsymbol{y}_{1:M}) = \max_{\boldsymbol{y}_{1:M}} \sum_{m=1}^{M} s_m(\boldsymbol{x}, y_m, y_{m-1})$$

$$= \max_{\boldsymbol{y}_{1:M}} \left(\max_{y_M} s_{M+1}(\blacklozenge, y_M) \right) + \left(\sum_{m=1}^{M-1} s_m(\boldsymbol{x}, y_m, y_{m-1}) \right)$$

Viterbi variable

Caching Viterbi variables as intermediate results:

$$\upsilon_m(y_m) \triangleq \max_{\boldsymbol{y}_{1:m-1}} \prod_{n=1}^{m} s_n(y_n, y_{n-1})$$

$$= \max_{y_{m-1}} \left[\max_{\boldsymbol{y}_{1:m-2}} \prod_{n=1}^{m-1} s_n(y_n, y_{n-1}) \right] s_m(y_m, y_{m-1})$$

$$= \max_{y_{m-1}} \upsilon_{m-1}(y_{m-1}) \, s_m(y_m, y_{m-1})$$

Note that $\upsilon_1(y_1) \triangleq s_1(y_1, \lozenge)$ and the maximum overall score for the sequence is the final Viterbi variable $\max_{\boldsymbol{y}_{1:M}} \Psi(\boldsymbol{w}_{1:M}, \boldsymbol{y}_{1:M}) = \upsilon_{M+1}(\blacklozenge)$

The Viterbi Algorithm

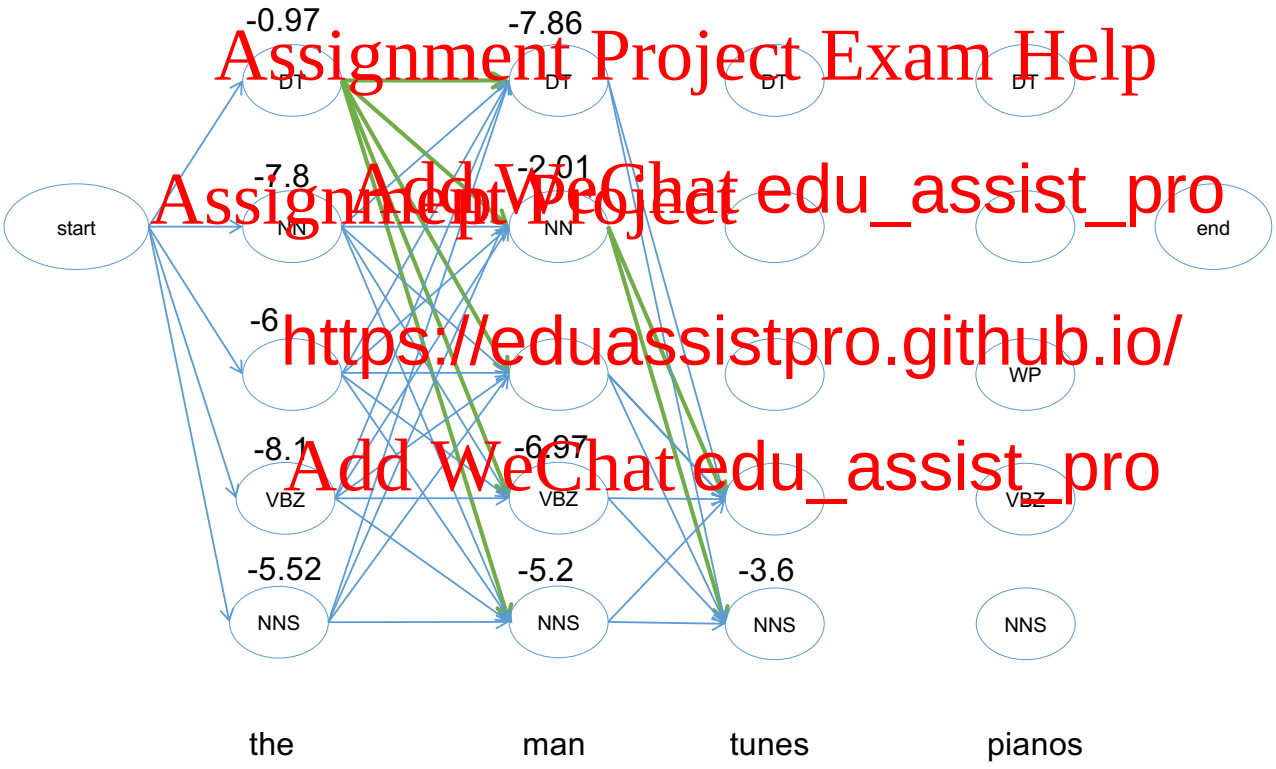
Viterbi Algorithm: Each $s_m(k, k')$ is a local score for tag $y_m = k$ and $y_{m-1} = k'$

for $k \in \{0, \ldots, K\}$ **do**
    $v_1(k) \leftarrow s_1(k, \Diamond)$
for $m \in \{2 \cdot$
    for $k \in$
        $v_m(k) \leftarrow \max_{k'} s_m(k, k')$
        $b_m(k) \leftarrow \text{argmax}_{k'} s_m(k,$
$y_M \leftarrow \text{argmax}_k s_{M+1}(\blacklozenge, k) + v_M(k)$
for $m \in \{M-1, \cdots, 1\}$ **do**
    $y_m \leftarrow b_m(y_{m+1})$
return $\boldsymbol{y}_{1:M}$

Assuming these parameters

	DT	NN	WP	VBZ	NNS	♦
w_0=the	-0.05	-3.9	-4.6	-4.6	-4.6	$-\infty$
w_0=man	-4.6	-0.35	-4.6			$-\infty$
w_0=who	-4.6	-4.6	-0.25			$-\infty$
w_0=whistles	-4.6	-4.6	-4.6			$-\infty$
w_0=tunes					.6	$-\infty$
w_0=pianos					.08	$-\infty$
$t_{-1} = \Diamond$	-0.92	-3.9	-1.9			$-\infty$
t_{-1}=DT	-2.3	-0.69	-4.6			-4.6
t_{-1}=NN	-4.6	-1.6	-0.3	-0.36	-1.0	-0.7
t_{-1}=WP	-3.8	-4.6	-4.6	-0.2	-4.6	-4.6
t_{-1}=VBZ	-0.2	-1.3	-1.6	-4.6	-0.92	-2.3
w_{-1}=NNS	-4.6	-4.6	-0.1	-4.6	-3.9	-1.2

Example Viterbi computation



Additional features (and their weights) can be added

	DT	NN	WP	VBZ	NNS	♦
w_0=the	-0.05	-3.9	-4.6	-4.6	-4.6	$-\infty$
w_0=man	-4.6	-0.35	-4.6	-1.4	-3.5	$-\infty$
w_0=who	-4.6	-4.6	-0.05	-4.6		
w_0=whistles	-4.6	-4.6	-4.6	-0.8		
w_0=tunes	-4.6	-4.6	-4.6	-0.8		
w_0=pianos	-4.6	-4.6	-4.6	-3.0		∞
$t_{-1} = \Diamond$	-					∞
t_{-1}=DT	-					-6
t_{-1}=NN	-					-7
t_{-1}=WP	-3.8	-4.6	-4.6	-0.		
t_{-1}=VBZ	-0.2	-1.3	-1.6	-4.		
w_{-1}=NNS	-4.6	-4.6	-0.1	-4.		
w_{-1}=START	-0.92	-3.9	-1.9	-3.5	-0.92	$-\infty$
w_{-1}=the	-4.6	-0.7	-4.6	-4.6	-0.75	-10
w_{-1}=man	-1.6	-2.3	-0.9	-1.6	-2.3	-1
w_{-1}=who	-1.8	-4.6	-4.6	-0.2	-4.6	-9
w_{-1}=whistles	-2.3	-4.6	-4.6	-1.6	-0.4	-0.5
w_{-1}=tunes	-1.6	-4.6	-4.6	-4.6	-0.26	-0.3

Feature templates used in SoA models

State-of-the-art models tend to use richer set of features and high-order transitions

- current word, w_i

- previous words, w_{i-2}, w_{i-1}

- next words, w_{i+1}, w_{i+2}

- previous two tags, y_{-1}, y_{-2}

- for rare words:
  - first k characters, up to $K = 4$
  - last k characters, up to $k = 4$
  - whether w_m contains a number, uppercase character, or hyphen

Parameter estimation for sequence labeling

We can extend the text classification models to sequ

Text classificat	**beling**
Naïve Bayes	v Models (HMM)
Logistic Regression	dom Fields (CRF)
Perceptron	
Support Vector Machines (SVM)	achines (SVM)