CS542 (Fall 2021) Structured Perceptron Worked Toy Example

Prof. Nikhil Krishnaswamy

Structured Perceptrons

(You may find the discussion in Chapter 7.5.1 of the Eisenstein book helpful.)

Suppose we are given the following weight matrix $\boldsymbol{\Theta}$:

$\boldsymbol{\Theta}$	$y_i = \ldots$		
	NN	VB	DT
y_{i-1} = <S>	-0.3	-0.7	0.3
y_{i-1} = NN	-0.7	0.3	-0.3
y_{i-1} = VB	-0.3	-0.7	0.3
y_{i-1} = DT	0.3	-0.3	-0.7
x_i = Alice	-0.3	-0.7	0.3
x_i = admired	0.3	-0.3	-0.7
x_i = Dorothy	-0.3	0.3	-0.7
x_i = every	-0.7	-0.3	0.3
x_i = dwarf	0.3	-0.7	-0.3
x_i = cheered	-0.7	0.3	-0.3

(If you think this matrix looks like π, $\mathbf{A}$, and $\mathbf{B}$ stacked on top of each other, you are right! Note that we don't need to account for the unknown word <UNK>, and for simplicity, we will ignore the bias term.)

1 Training: Viterbi

Suppose we are given the following training sentence:

Alice/NN admired/VB Dorothy/NN

Use the Viterbi algorithm to compute the best tag sequence. As part of your answer, you should fill in the Viterbi trellis below. You should also keep track of backpointers, either using arrows or in a separate table.

	Alice	admired	Dorothy
NN	•	•	•
VB	•	•	•
DT	•	•	•

1.1 Answer

(a) We fill in the first column in our Viterbi trellis, this time using values from the weight matrix $\boldsymbol{\Theta}$, and adding rather than multiplying:

Viterbi(y_i, x_i)	Alice	admired	Dorothy
NN	(a)	(d)	(g)
VB	(b)	(e)	(h)
DT	(c)	(f)	(i)

1. $\boldsymbol{\Theta}(\texttt{<S>}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{Alice}, \texttt{NN}) = -0.3 - 0.3 = -0.6$
2. $\boldsymbol{\Theta}(\texttt{<S>}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{Alice}, \texttt{VB}) = -0.7 - 0.7 = -1.4$
3. $\boldsymbol{\Theta}(\texttt{<S>}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{Alice}, \texttt{DT}) = 0.3 + 0.3 = 0.6$

Then we can fill in the second column in our trellis:

1. $\max((\texttt{NN}, \text{Viterbi}(\texttt{NN}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{NN}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{admired}, \texttt{NN})),$
   $(\texttt{VB}, \text{Viterbi}(\texttt{VB}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{VB}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{admired}, \texttt{NN})),$
   $(\texttt{DT}, \text{Viterbi}(\texttt{DT}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{DT}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{admired}, \texttt{NN})))$
   $= \max((\texttt{NN}, -0.6 - 0.7 + 0.3), (\texttt{VB}, -1.4 - 0.3 + 0.3), (\texttt{DT}, 0.6 + 0.3 + 0.3))$
   $= \max((\texttt{NN}, -1.0), (\texttt{VB}, -1.4), (\texttt{DT}, 1.2))$
   $= (\texttt{DT}, 1.2)$
2. $\max((\texttt{NN}, \text{Viterbi}(\texttt{NN}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{NN}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{admired}, \texttt{VB})),$
   $(\texttt{VB}, \text{Viterbi}(\texttt{VB}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{VB}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{admired}, \texttt{VB})),$
   $(\texttt{DT}, \text{Viterbi}(\texttt{DT}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{DT}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{admired}, \texttt{VB})))$
   $= \max((\texttt{NN}, -0.6 + 0.3 - 0.3), (\texttt{VB}, -1.4 - 0.7 - 0.3), (\texttt{DT}, 0.6 - 0.3 - 0.3))$
   $= \max((\texttt{NN}, -0.6), (\texttt{VB}, -2.4), (\texttt{DT}, 0.0))$
   $= (\texttt{DT}, 0.0)$
3. $\max((\texttt{NN}, \text{Viterbi}(\texttt{NN}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{NN}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{admired}, \texttt{DT})),$
   $(\texttt{VB}, \text{Viterbi}(\texttt{VB}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{VB}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{admired}, \texttt{DT})),$
   $(\texttt{DT}, \text{Viterbi}(\texttt{DT}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{DT}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{admired}, \texttt{DT})))$
   $= \max((\texttt{NN}, -0.6 - 0.3 - 0.7), (\texttt{VB}, -1.4 + 0.3 - 0.7), (\texttt{DT}, 0.6 - 0.7 - 0.7))$
   $= \max((\texttt{NN}, -1.6), (\texttt{VB}, -1.8), (\texttt{DT}, -0.8))$
   $= (\texttt{DT}, -0.8)$

Finally we can fill in the third column in our trellis:

1. $\max((\texttt{NN}, \text{Viterbi}(\texttt{NN}, \texttt{admired}) + \boldsymbol{\Theta}(\texttt{NN}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{NN})),$
   $(\texttt{VB}, \text{Viterbi}(\texttt{VB}, \texttt{admired}) + \boldsymbol{\Theta}(\texttt{VB}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{NN})),$

$(\texttt{DT}, \text{Viterbi}(\texttt{DT}, \texttt{admired}) + \boldsymbol{\Theta}(\texttt{DT}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{NN})))$
$= \max((\texttt{NN}, 1.2 - 0.7 - 0.3), (\texttt{VB}, 0.0 - 0.3 - 0.3), (\texttt{DT}, -0.8 + 0.3 - 0.3))$
$= \max((\texttt{NN}, 0.2), (\texttt{VB}, -0.6), (\texttt{DT}, -0.8))$
$= (\texttt{NN}, 0.2)$

2. $\max((\texttt{NN}, \text{Viterbi}(\texttt{NN}, \texttt{admired}) + \boldsymbol{\Theta}(\texttt{NN}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{VB})),$
   $(\texttt{VB}, \text{Viterbi}(\texttt{VB}, \texttt{admired}) + \boldsymbol{\Theta}(\texttt{VB}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{VB})),$
   $(\texttt{DT}, \text{Viterbi}(\texttt{DT}, \texttt{admired}) + \boldsymbol{\Theta}(\texttt{DT}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{VB})))$
   $= \max((\texttt{NN}, 1.2 + 0.3 + 0.3), (\texttt{VB}, 0.0 - 0.7 + 0.3), (\texttt{DT}, -0.8 - 0.3 + 0.3))$
   $= \max((\texttt{NN}, 1.8), (\texttt{VB}, -0.4), (\texttt{DT}, -0.8))$
   $= (\texttt{NN}, 1.8)$

3. $\max((\texttt{NN}, \text{Viterbi}(\texttt{NN}, \texttt{admired}) + \boldsymbol{\Theta}(\texttt{NN}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{DT})),$
   $(\texttt{VB}, \text{Viterbi}(\texttt{VB}, \texttt{admired}) + \boldsymbol{\Theta}(\texttt{VB}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{DT})),$
   $(\texttt{DT}, \text{Viterbi}(\texttt{DT}, \texttt{admired}) + \boldsymbol{\Theta}(\texttt{DT}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{DT})))$
   $= \max((\texttt{NN}, 1.2 - 0.3 - 0.7), (\texttt{VB}, 0.0 + 0.3 - 0.7), (\texttt{DT}, -0.8 - 0.7 - 0.7))$
   $= \max((\texttt{NN}, 0.2), (\texttt{VB}, -0.4), (\texttt{DT}, -2.2))$
   $= (\texttt{NN}, 0.2)$

We put the max in the Viterbi trellis and the argmax in the backpointer trellis:

Viterbi(y_i, x_i)	Alice	admired	Dorothy
NN	-0.6	1.2	0.2
VB	-1.4	0.0	1.8
DT	0.6	-0.8	0.2

backpointer(y_i, x_i)	Alice	admired	Dorothy
NN	0	DT	NN
VB	0	DT	NN
DT	0	DT	NN

The best last tag is the argmax of the last column of the Viterbi trellis, i.e. VB. The best previous tag is then backpointer(VB, Dorothy) = NN. The best previous tag before that is then backpointer(NN, admired) = DT. Since we have reached the beginning of the sentence, the best tag sequence is:

```
Alice/DT admired/NN Dorothy/VB
```

2 Training: Weight Update

Update the weight matrix. Use a constant learning rate $\eta = 1$.

2.1 Answer

(b) The correct tag sequence is:

`Alice/NN admired/VB Dorothy/NN`

But the predicted tag sequence is:

`Alice/DT admired/NN Dorothy/VB`

Therefore, we increment the weights for features in the correct sequence:

$\boldsymbol{\Theta}(\texttt{<S>}, \texttt{NN}) = -0.3 + 1 = 0.7$

$\boldsymbol{\Theta}(\texttt{Alice}, \texttt{NN}) = -0.3 + 1 = 0.7$

$\boldsymbol{\Theta}(\texttt{NN}, \texttt{VB}) = 0.3 + 1 = 1.3$

$\boldsymbol{\Theta}(\texttt{admired}, \texttt{VB}) = -0.3 + 1 = 0.7$

$\boldsymbol{\Theta}(\texttt{VB}, \texttt{NN}) = -0.3 + 1 = 0.7$

$\boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{NN}) = -0.3 + 1 = 0.7$

And decrement the weights for features in the predicted sequence:

$\boldsymbol{\Theta}(\texttt{<S>}, \texttt{DT}) = 0.3 - 1 = -0.7$

$\boldsymbol{\Theta}(\texttt{Alice}, \texttt{DT}) = 0.3 - 1 = -0.7$

$\boldsymbol{\Theta}(\texttt{DT}, \texttt{NN}) = 0.3 - 1 = -0.7$

$\boldsymbol{\Theta}(\texttt{admired}, \texttt{NN}) = 0.3 - 1 = -0.7$

$\boldsymbol{\Theta}(\texttt{NN}, \texttt{VB}) = 1.3 - 1 = 0.3$

$\boldsymbol{\Theta}(\texttt{Dorothy}, \texttt{VB}) = 0.3 - 1 = -0.7$

Note that $\boldsymbol{\Theta}(\texttt{NN}, \texttt{VB})$ is a feature in both the correct sequence and the predicted sequence, and therefore was both incremented and decremented. The overall effect is that there is no change in the weight.

3 Testing

Suppose we are given the following testing sentence:

`Alice cheered`

Use the Viterbi algorithm to compute the best tag sequence. Again, you should fill in the Viterbi trellis below, and keep track of backpointers.

	`Alice`	`cheered`
NN	•	•
VB	•	•
DT	•	•

3.1 Answer

We fill in our Viterbi trellis one more time, this time with our updated weights:

Viterbi(y_i, x_i)	`Alice`	`cheered`
NN	(a)	(d)
VB	(b)	(e)
DT	(c)	(f)

1. $\boldsymbol{\Theta}(\texttt{<S>}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{Alice}, \texttt{NN}) = 0.7 + 0.7 = 1.4$
2. $\boldsymbol{\Theta}(\texttt{<S>}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{Alice}, \texttt{VB}) = -0.7 - 0.7 = -1.4$
3. $\boldsymbol{\Theta}(\texttt{<S>}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{Alice}, \texttt{DT}) = -0.7 - 0.7 = -1.4$
4. $\max((\texttt{NN}, \text{Viterbi}(\texttt{NN}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{NN}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{cheered}, \texttt{NN})),$
   $(\texttt{VB}, \text{Viterbi}(\texttt{VB}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{VB}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{cheered}, \texttt{NN})),$
   $(\texttt{DT}, \text{Viterbi}(\texttt{DT}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{DT}, \texttt{NN}) + \boldsymbol{\Theta}(\texttt{cheered}, \texttt{NN})))$
   $= \max((\texttt{NN}, 1.4 - 0.7 - 0.7), (\texttt{VB}, -1.4 + 0.7 - 0.7), (\texttt{DT}, -1.4 - 0.7 - 0.7))$
   $= \max((\texttt{NN}, 0.0), (\texttt{VB}, -1.4), (\texttt{DT}, -2.8))$
   $= (\texttt{NN}, 0.0)$
5. $\max((\texttt{NN}, \text{Viterbi}(\texttt{NN}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{NN}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{cheered}, \texttt{VB})),$
   $(\texttt{VB}, \text{Viterbi}(\texttt{VB}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{VB}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{cheered}, \texttt{VB})),$
   $(\texttt{DT}, \text{Viterbi}(\texttt{DT}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{DT}, \texttt{VB}) + \boldsymbol{\Theta}(\texttt{cheered}, \texttt{VB})))$
   $= \max((\texttt{NN}, 1.4 + 0.3 + 0.3), (\texttt{VB}, -1.4 - 0.7 + 0.3), (\texttt{DT}, -1.4 - 0.3 + 0.3))$
   $= \max((\texttt{NN}, 2.0), (\texttt{VB}, -1.8), (\texttt{DT}, -1.4))$
   $= (\texttt{NN}, 2.0)$
6. $\max((\texttt{NN}, \text{Viterbi}(\texttt{NN}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{NN}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{cheered}, \texttt{DT})),$
   $(\texttt{VB}, \text{Viterbi}(\texttt{VB}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{VB}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{cheered}, \texttt{DT})),$
   $(\texttt{DT}, \text{Viterbi}(\texttt{DT}, \texttt{Alice}) + \boldsymbol{\Theta}(\texttt{DT}, \texttt{DT}) + \boldsymbol{\Theta}(\texttt{cheered}, \texttt{DT})))$
   $= \max((\texttt{NN}, 1.4 - 0.3 - 0.3), (\texttt{VB}, -1.4 + 0.3 - 0.3), (\texttt{DT}, -1.4 - 0.7 - 0.3))$
   $= \max((\texttt{NN}, 0.8), (\texttt{VB}, -1.4), (\texttt{DT}, -2.4))$
   $= (\texttt{NN}, 0.8)$

We put the max in the Viterbi trellis and the argmax in the backpointer trellis:

Viterbi(y_i, x_i)	Alice	cheered
NN	1.4	0.0
VB	-1.4	2.0
DT	-1.4	0.8

backpointer(y_i, x_i)	Alice	cheered
NN	0	NN
VB	0	NN
DT	0	NN

The best last tag is the argmax of the last column of the Viterbi trellis, i.e. VB. The best previous tag is then backpointer(VB, cheered) = NN. Since we have reached the beginning of the sentence, the best tag sequence is:

Alice/NN cheered/VB