



Why Language is Hard: Structure and Predictions (Description)

Advanced Machine Learning for NLP

Jordan Boyd-Graber

SLIDES ADAPTED FROM LIANG HUANG

How do we set the feature weights?

- Goal is to minimize errors
- Want to reward features that lead to right answers
- Penalize features that lead to wrong answers
- Problem: predictions are correlated

Perceptron Algorithm

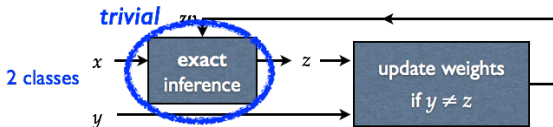
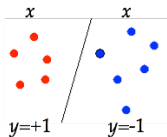
- Rather than just counting up how often we see events?
- We'll use this for intuition in 2D case

Perceptron Algorithm

```
1:  $\vec{w}_1 \leftarrow \vec{0}$ 
2: for  $t \leftarrow 1 \dots T$  do
3:   Receive  $x_t$ 
4:    $\hat{y}_t \leftarrow \text{sgn}(\vec{w}_t \cdot \vec{x}_t)$ 
5:   Receive  $y_t$ 
6:   if  $\hat{y}_t \neq y_t$  then
7:      $\vec{w}_{t+1} \leftarrow \vec{w}_t + y_t \vec{x}_t$ 
8:   else
9:      $\vec{w}_{t+1} \leftarrow w_t$ 
return  $w_{T+1}$ 
```

Binary to Structure

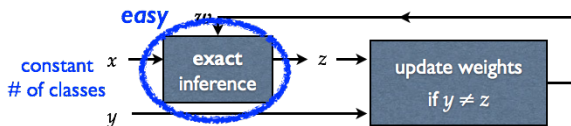
binary perceptron
(Rosenblatt, 1959)



Binary to Structure

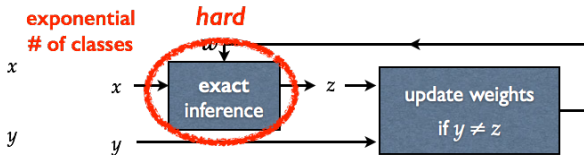
multiclass perceptron
(Freund/Schapire, 1999)

0 1 2 3 4 5 6 7 8 9



Binary to Structure

structured perceptron
(Collins, 2002)



Generic Perceptron

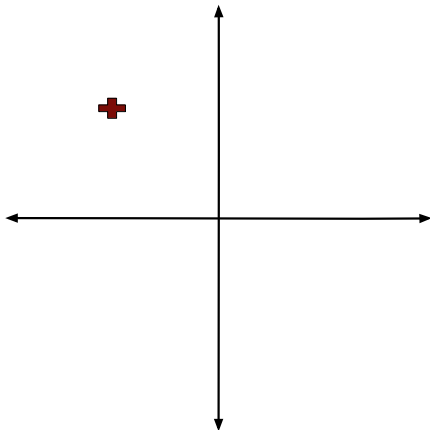
- perceptron is the simplest machine learning algorithm
- online-learning: one example at a time
- learning by doing
 - find the best output under the current weights
 - update weights at mistakes

2D Example

Initially, weight vector is zero:

$$\vec{w}_1 = \langle 0, 0 \rangle \quad (1)$$

Observation 1



$$x_1 = \langle -2, 2 \rangle \quad (2)$$

$$\hat{y}_1 = 0 \quad (3)$$

$$y_1 = +1 \quad (4)$$

Update 1

$$\vec{w}_{t+1} \leftarrow \vec{w}_t + y_t \vec{x}_t \quad (5)$$

$$\vec{w}_2 \leftarrow \quad (6)$$

Update 1

$$\vec{w}_{t+1} \leftarrow \vec{w}_t + y_t \vec{x}_t \quad (5)$$

$$\vec{w}_2 \leftarrow \langle 0, 0 \rangle + \langle -2, 2 \rangle \quad (6)$$

$$(7)$$

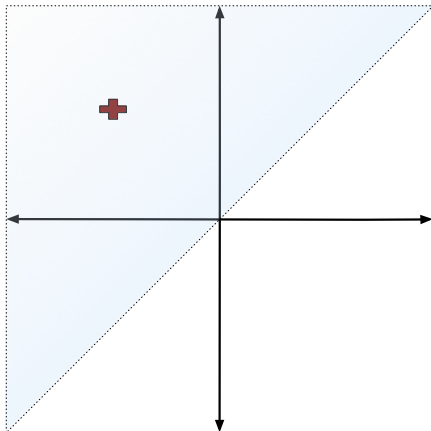
Update 1

$$\vec{w}_{t+1} \leftarrow \vec{w}_t + y_t \vec{x}_t \quad (5)$$

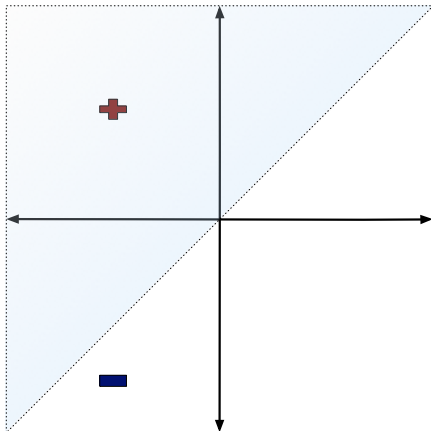
$$\vec{w}_2 \leftarrow \langle 0, 0 \rangle + \langle -2, 2 \rangle \quad (6)$$

$$\vec{w}_2 = \langle -2, 2 \rangle \quad (7)$$

Observation 2



Observation 2



$$x_2 = \langle -2, -3 \rangle \quad (8)$$

$$\hat{y}_2 = +4 + -6 = -2 \quad (9)$$

$$y_2 = -1 \quad (10)$$

Update 2

$$\vec{w}_{t+1} \leftarrow \vec{w}_t \quad (11)$$

$$\vec{w}_2 \leftarrow \quad (12)$$

Update 2

$$\vec{w}_{t+1} \leftarrow \vec{w}_t \quad (11)$$

$$\vec{w}_2 \leftarrow \langle -2, 2 \rangle \quad (12)$$

$$(13)$$

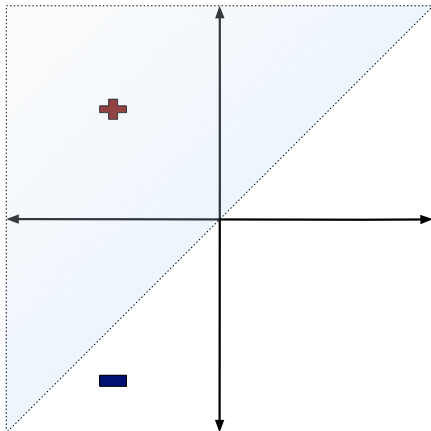
Update 2

$$\vec{w}_{t+1} \leftarrow \vec{w}_t \quad (11)$$

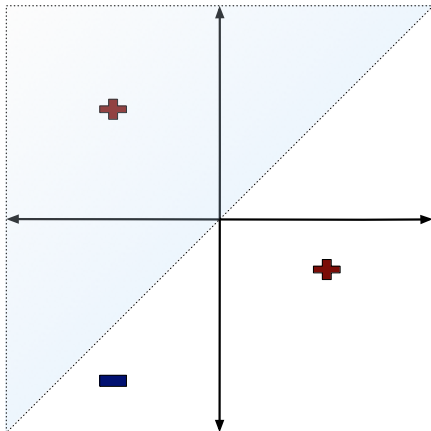
$$\vec{w}_2 \leftarrow \langle -2, 2 \rangle \quad (12)$$

$$\vec{w}_2 = \langle -2, 2 \rangle \quad (13)$$

Observation 3



Observation 3



$$x_3 = \langle 2, -1 \rangle \quad (14)$$

$$\hat{y}_3 = -4 + -2 = -6 \quad (15)$$

$$y_3 = +1 \quad (16)$$

Update 3

$$\vec{w}_{t+1} \leftarrow \vec{w}_t + y_t \vec{x}_t \quad (17)$$

$$\vec{w}_3 \leftarrow \quad (18)$$

Update 3

$$\vec{w}_{t+1} \leftarrow \vec{w}_t + y_t \vec{x}_t \quad (17)$$

$$\vec{w}_3 \leftarrow \langle -2, 2 \rangle + \langle 2, -1 \rangle \quad (18)$$

$$(19)$$

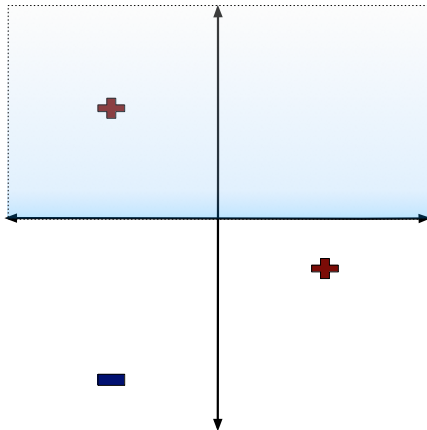
Update 3

$$\vec{w}_{t+1} \leftarrow \vec{w}_t + y_t \vec{x}_t \quad (17)$$

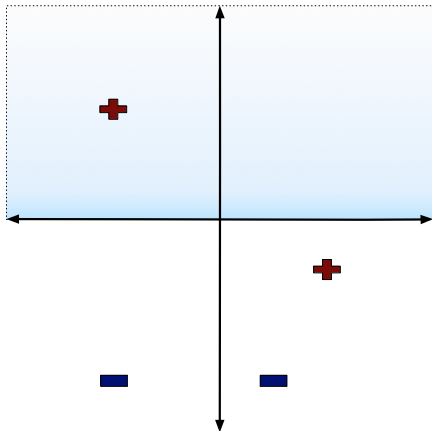
$$\vec{w}_3 \leftarrow \langle -2, 2 \rangle + \langle 2, -1 \rangle \quad (18)$$

$$\vec{w}_3 = \langle 0, 1 \rangle \quad (19)$$

Observation 4



Observation 4



$$x_4 = \langle 1, -4 \rangle \quad (20)$$

$$\hat{y}_4 = -4 \quad (21)$$

$$y_4 = -1 \quad (22)$$

Update 4

$$\vec{w}_4 \leftarrow \quad (23)$$

Update 4

$$\vec{w}_4 \leftarrow \vec{w}_3 \quad (23)$$

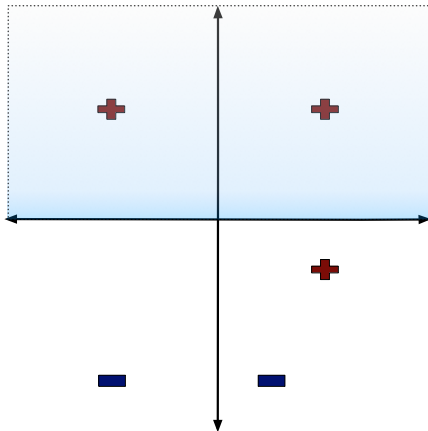
(24)

Update 4

$$\vec{w}_4 \leftarrow \vec{w}_3 \quad (23)$$

$$\vec{w}_4 = \langle 0, 1 \rangle \quad (24)$$

Observation 5



$$x_5 = \langle 2, 2 \rangle \quad (25)$$

$$\hat{y}_5 = 2 \quad (26)$$

$$y_5 = +1 \quad (27)$$

Update 5

$$\vec{w}_5 \leftarrow \quad (28)$$

Update 5

$$\vec{w}_5 \leftarrow \vec{w}_4 \quad (28)$$

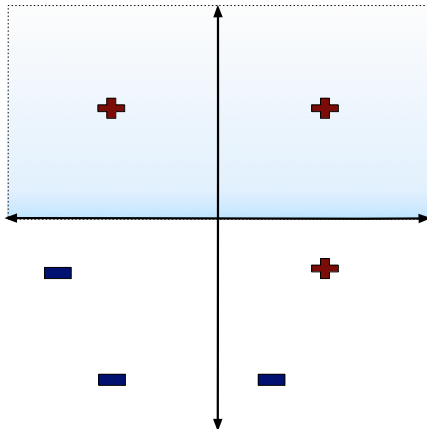
$$(29)$$

Update 5

$$\vec{w}_5 \leftarrow \vec{w}_4 \quad (28)$$

$$\vec{w}_5 = \langle 0, 1 \rangle \quad (29)$$

Observation 6



$$x_6 = \langle 2, 2 \rangle \quad (30)$$

$$\hat{y}_6 = 2 \quad (31)$$

$$y_6 = +1 \quad (32)$$

Update 6

$$\vec{w}_6 \leftarrow \quad (33)$$

Update 6

$$\vec{w}_6 \leftarrow \vec{w}_5 \quad (33)$$

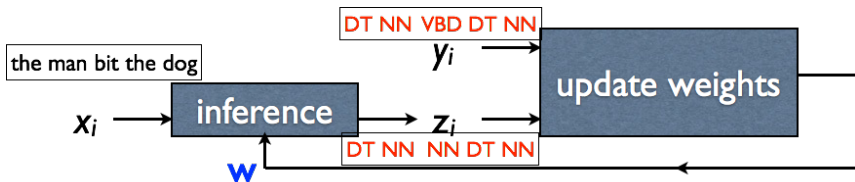
(34)

Update 6

$$\vec{w}_6 \leftarrow \vec{w}_5 \quad (33)$$

$$\vec{w}_6 = \langle 0, 1 \rangle \quad (34)$$

Structured Perceptron



Perceptron Algorithm

Inputs: Training set (x_i, y_i) for $i = 1 \dots n$

Initialization: $\mathbf{W} = 0$

Define: $F(x) = \operatorname{argmax}_{y \in \mathbf{GEN}(x)} \Phi(x, y) \cdot \mathbf{W}$

Algorithm: For $t = 1 \dots T, i = 1 \dots n$
 $z_i = F(x_i)$
 If $(z_i \neq y_i)$ $\mathbf{W} \leftarrow \mathbf{W} + \Phi(x_i, y_i) - \Phi(x_i, z_i)$

Output: Parameters \mathbf{W}

POS Example

• gold-standard: DT NN VBD DT NN y
• the man bit the dog x $\Phi(x, y)$

• current output: DT NN NN DT NN z
• the man bit the dog x $\Phi(x, z)$

• assume only two feature classes

• tag bigrams

t_{i-1} t_i

• word/tag pairs

w_i

• weights ++: (NN, VBD) (VBD, DT) (VBD \rightarrow bit)

• weights --: (NN, NN) (NN, DT) (NN \rightarrow bit)