The Perceptron

CMSC 422

MARINE CARPUAT
marine@cs.umd.edu

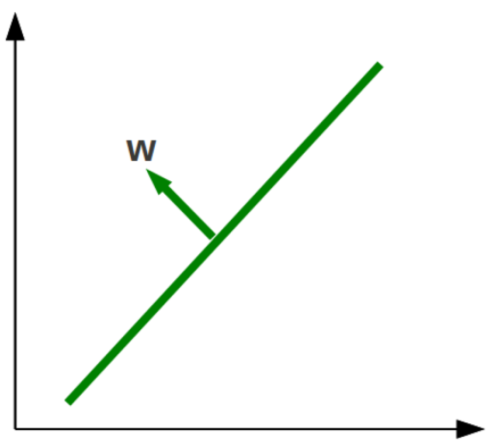
Credit: figures by Piyush Rai and Hal Daume III

This week

- A new model/algorithm
  - the perceptron
  - and its variants: voted, averaged
- Fundamental Machine Learning Concepts
  - Online vs. batch learning
  - Error-driven learning

- Project 1 coming soon!

Geometry concept: **Hyperplane**

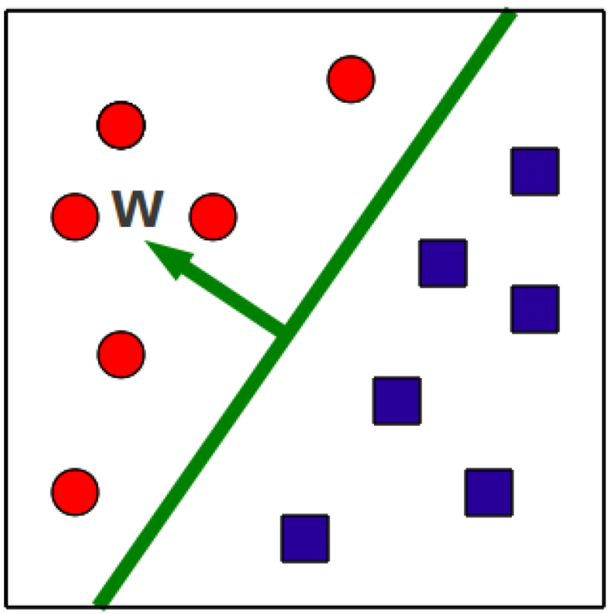


- Separates a D-dimensional space into two half-spaces

- Defined by an outward pointing normal vector $w \in \mathbb{R}^D$
  - w is **orthogonal** to any vector lying on the hyperplane

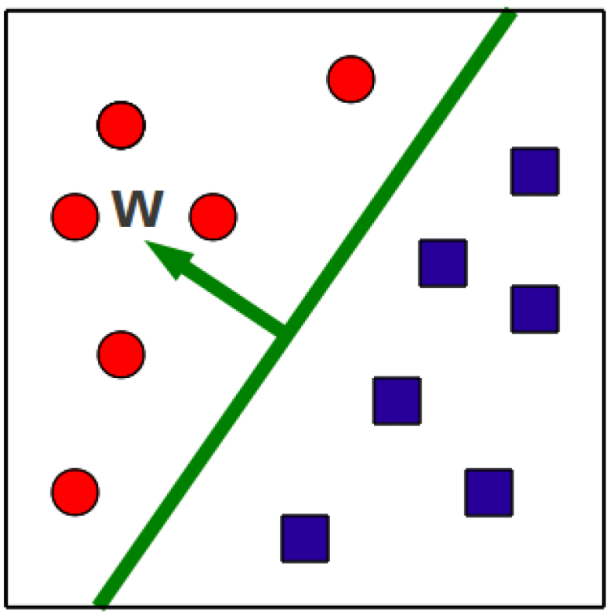
- Hyperplane passes through the origin, unless we also define a **bias** term b

Binary classification via hyperplanes



- Let's assume that the decision boundary is a hyperplane
- Then, training consists in finding a hyperplane w that separates positive from negative examples

Binary classification via hyperplanes



- At test time, we check on what side of the hyperplane examples fall

$$\hat{y} = sign(w^T x + b)$$

Function Approximation with Perceptron

Problem setting

- Set of possible instances X
  - Each instance $x \in X$ is a feature vector $x = [x_1, \ldots, x_D]$
- Unknown target function $f: X \rightarrow Y$
  - Y is binary valued {-1; +1}
- Set of function hypotheses $H = \{h \mid h: X \rightarrow Y\}$
  - Each hypothesis h is a hyperplane in D-dimensional space

Input

- Training examples $\{(x^{(1)}, y^{(1)}), \ldots (x^{(N)}, y^{(N)})\}$ of unknown target function f

Output

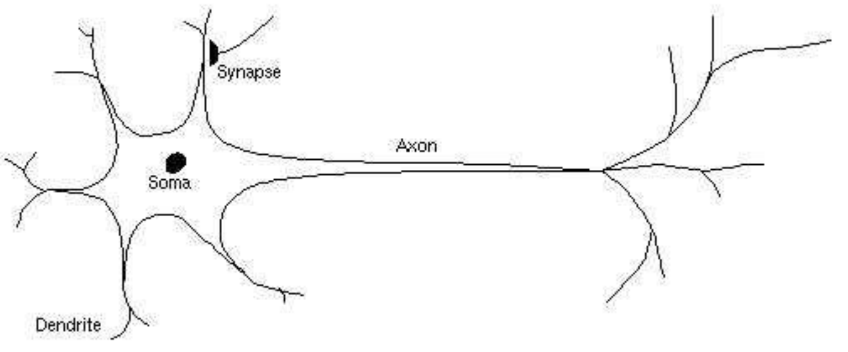
- Hypothesis $h \in H$ that best approximates target function f

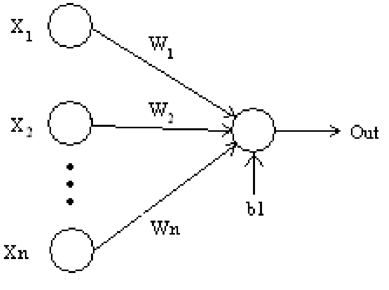
Perception: Prediction Algorithm

Algorithm 6 $\textsc{PerceptronTest}(w_0, w_1, \ldots, w_D, b, \hat{\boldsymbol{x}})$

1: $a \leftarrow \sum_{d=1}^{D} w_d \, \hat{x}_d + b$ // compute activation for the test example
2: **return** $\textsc{sign}(a)$

Aside: biological inspiration





Analogy: the perceptron as a neuron

Perceptron Training Algorithm

Algorithm 5 $\textsc{PerceptronTrain}(\mathbf{D}, \mathit{MaxIter})$

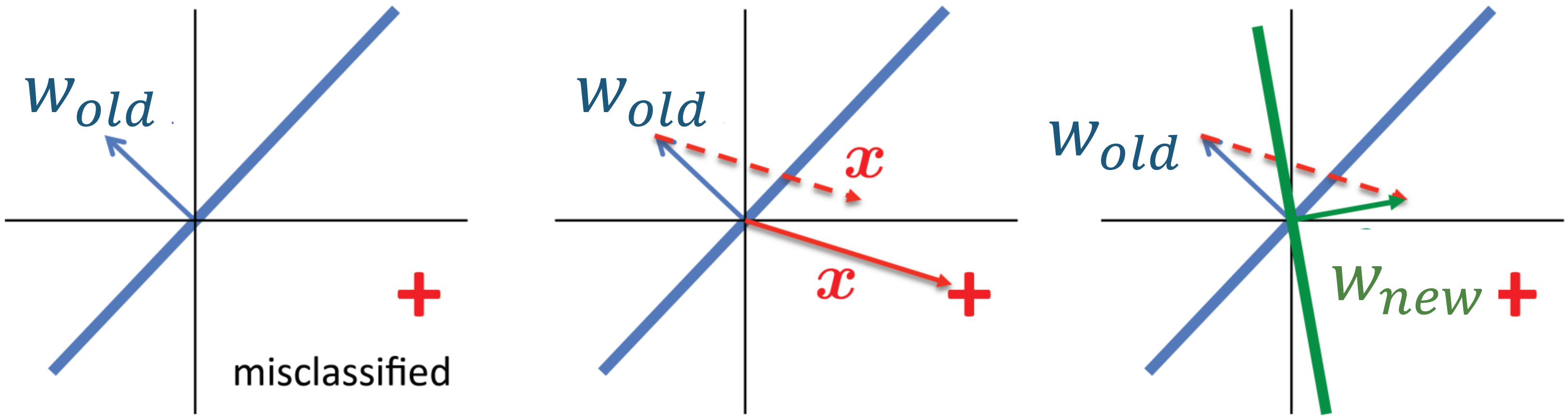
$w_d \leftarrow 0$, for all $d = 1 \dots D$ // initialize weights
$b \leftarrow 0$ // initialize bias
for $\mathit{iter} = 1 \dots \mathit{MaxIter}$ **do**
    for all $(\boldsymbol{x}, y) \in \mathbf{D}$ **do**
        $a \leftarrow \sum_{d=1}^{D} w_d \, x_d + b$ // compute activation for this example
        if $ya \leq 0$ **then**
            $w_d \leftarrow w_d + y x_d$, for all $d = 1 \dots D$ // update weights
            $b \leftarrow b + y$ // update bias
        end if
    end for
end for
return $w_0, w_1, \dots, w_D, b$

Properties of the Perceptron training algorithm

- Online
  - We look at one example at a time, and update the model as soon as we make an error
  - As opposed to batch algorithms that update parameters after seeing the entire training set

- Error-driven
  - We only update parameters/model if we make an error

Perceptron update: geometric interpretation



Practical considerations

- The order of training examples matters!
  - Random is better
- Early stopping
  - Good strategy to avoid overfitting
- Simple modifications dramatically improve performance
  - voting or averaging

Standard Perceptron: predict based on final parameters

Algorithm 5 PerceptronTrain($\mathbf{D}$, *MaxIter*)

```
w_d ← 0, for all d = 1 ... D                         // initialize weights
b ← 0                                                // initialize bias
for iter = 1 ... MaxIter do
    for all (x,y) ∈ D do
        a ← Σ_{d=1}^{D} w_d x_d + b                  // compute activation for this example
        if ya ≤ 0 then
            w_d ← w_d + y x_d, for all d = 1 ... D   // update weights
            b ← b + y                                // update bias
        end if
    end for
end for
return w_0, w_1, ..., w_D, b
```

Predict based on final + intermediate parameters

- The voted perceptron

$$\hat{y} = \text{sign}\left(\sum_{k=1}^{K} c^{(\mathsf{k})}\text{sign}\left(\boldsymbol{w}^{(\mathsf{k})} \cdot \hat{\boldsymbol{x}} + b^{(\mathsf{k})}\right)\right)$$

- The averaged perceptron

$$\hat{y} = \text{sign}\left(\sum_{k=1}^{K} c^{(\mathsf{k})}\left(\boldsymbol{w}^{(\mathsf{k})} \cdot \hat{\boldsymbol{x}} + b^{(\mathsf{k})}\right)\right)$$

- Require keeping track of "survival time" of weight vectors $c^{(1)}, \ldots, c^{(\mathsf{K})}$

How would you modify this algorithm for voted perceptron?

Algorithm 5 PerceptronTrain($\mathbf{D}$, $MaxIter$)

$w_d \leftarrow 0$, for all $d = 1 \dots D$ // initialize weights
$b \leftarrow 0$ // initialize bias
for $iter = 1 \dots MaxIter$ **do**
  for all $(x,y) \in \mathbf{D}$ **do**
    $a \leftarrow \sum_{d=1}^{D} w_d \, x_d + b$ // compute activation for this example
    if $ya \leq 0$ **then**
      $w_d \leftarrow w_d + y x_d$, for all $d = 1 \dots D$ // update weights
      $b \leftarrow b + y$ // update bias
    end if
  end for
end for
return $w_0, w_1, \dots, w_D, b$

How would you modify this algorithm for averaged perceptron?

Algorithm 5 PerceptronTrain($\mathbf{D}$, $MaxIter$)

$w_d \leftarrow 0$, for all $d = 1 \ldots D$ // initialize weights
$b \leftarrow 0$ // initialize bias
for $iter = 1 \ldots MaxIter$ **do**
  for all $(x,y) \in \mathbf{D}$ **do**
    $a \leftarrow \sum_{d=1}^{D} w_d \, x_d + b$ // compute activation for this example
    if $ya \leq 0$ **then**
      $w_d \leftarrow w_d + y x_d$, for all $d = 1 \ldots D$ // update weights
      $b \leftarrow b + y$ // update bias
    end if
  end for
end for
return $w_0, w_1, \ldots, w_D, b$

Averaged perceptron decision rule

$$\hat{y} = \text{sign}\left(\sum_{k=1}^{K} c^{(\mathsf{k})}\left(\boldsymbol{w}^{(\mathsf{k})} \cdot \hat{\boldsymbol{x}} + b^{(\mathsf{k})}\right)\right)$$

can be rewritten as

$$\hat{y} = \text{sign}\left(\left(\sum_{k=1}^{K} c^{(\mathsf{k})}\boldsymbol{w}^{(\mathsf{k})}\right) \cdot \hat{\boldsymbol{x}} + \sum_{k=1}^{K} c^{(\mathsf{k})} b^{(\mathsf{k})}\right)$$

Averaged Perceptron Training

Algorithm 7 AveragedPerceptronTrain($\mathbf{D}$, $MaxIter$)

$\boldsymbol{w} \leftarrow \langle 0, 0, \ldots 0 \rangle$, $b \leftarrow 0$ // initialize weights and bias
$\boldsymbol{u} \leftarrow \langle 0, 0, \ldots 0 \rangle$, $\beta \leftarrow 0$ // initialize cached weights and bias
$c \leftarrow 1$ // initialize example counter to one
for $iter = 1 \ldots MaxIter$ **do**
  for all $(\boldsymbol{x}, y) \in \mathbf{D}$ **do**
    if $y(\boldsymbol{w} \cdot \boldsymbol{x} + b) \leq 0$ **then**
      $\boldsymbol{w} \leftarrow \boldsymbol{w} + y\ \boldsymbol{x}$ // update weights
      $b \leftarrow b + y$ // update bias
      $\boldsymbol{u} \leftarrow \boldsymbol{u} + y\ c\ \boldsymbol{x}$ // update cached weights
      $\beta \leftarrow \beta + y\ c$ // update cached bias
    end if
    $c \leftarrow c + 1$ // increment counter regardless of update
  end for
end for
return $\boldsymbol{w} - \frac{1}{c}\ \boldsymbol{u}$, $b - \frac{1}{c}\ \beta$ // return averaged weights and bias

This week

- A new model/algorithm
  - the perceptron
  - and its variants: voted, averaged
- Fundamental Machine Learning Concepts
  - Online vs. batch learning
  - Error-driven learning

- Project 1 coming soon!