CS480/680: Intro to ML

Lecture 01: Perceptron

UNIVERSITY OF WATERLOO

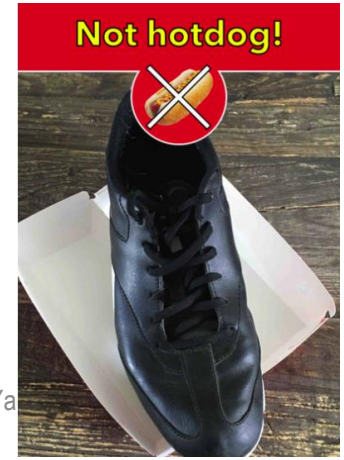
Announcements

- Assignment 1 to be posted on Thursday 1/14

- Projects
  - NeurIPS: https://papers.nips.cc/
  - ICML: http://proceedings.mlr.press/v80/
  - COLT: http://proceedings.mlr.press/v75/
  - AISTATS: http://proceedings.mlr.press/v84/
  - ICLR: https://iclr.cc/Conferences/2018/Schedule?type=Poster
  - JMLR: http://www.jmlr.org/papers/v18/

Perceptron

Supervised learning



UNIVERSITY OF WATERLOO

Spam filtering example

	and	viagra	the	of	nigeria	y
$\mathbf{x}_1$	1	1	0	1	1	+1
$\mathbf{x}_2$	0	0	1	1	0	-1
$\mathbf{x}_3$	0	1	1	0	0	+1
$\mathbf{x}_4$	1	0	0	1	0	-1
$\mathbf{x}_5$	1	0	1	0	1	+1
$\mathbf{x}_6$	1	0	1	1	0	-1

- Training set ($X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$, $\mathbf{y}=[y_1, y_2, \ldots, y_n]$)
  - $\mathbf{x}_i$ in $X = R^d$: instance i with d dimensional features
  - y_i in $Y = \{-1, 1\}$: instance i is spam or ham?
- Good feature representation is of uttermost importance

Batch vs. Online

- Batch learning
  - Interested in performance on test set X'
  - Training set (X, **y**) is just a means
  - Statistical assumption on X and X'

- Online learning
  - Data comes one by one (streaming)
  - Need to predict y before knowing its true value
  - Interested in making as few mistakes as possible
  - Compare against some baseline

Thought Experiment

- Repeat the following game
  - Observe instance $\mathbf{x}_i$
  - Predict its label $\hat{y}_i$ (in whatever way you like)
  - Reveal the true label y_i
  - Suffer a mistake if $\hat{y}_i \neq y_i$
- How many mistakes in the worst-case?
- Predict first, reveal next: no peek into the future!

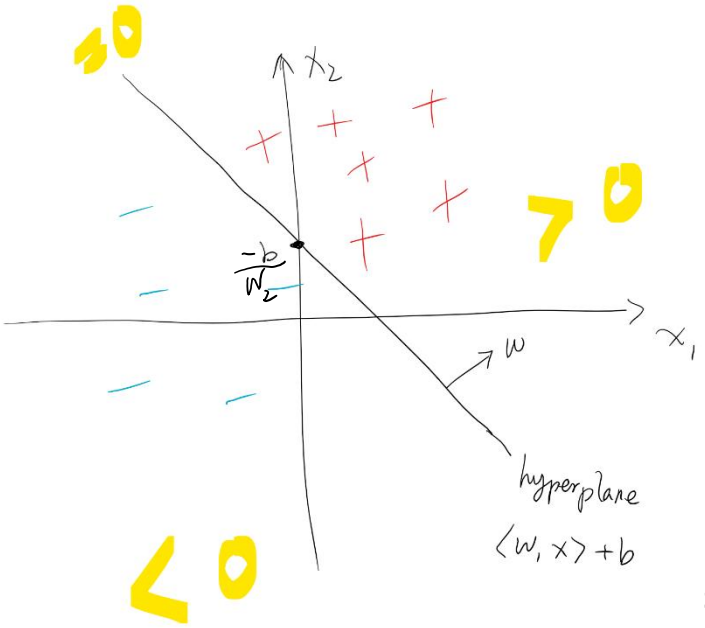
Linear threshold function

- Find (**w**, b) such that for all i:

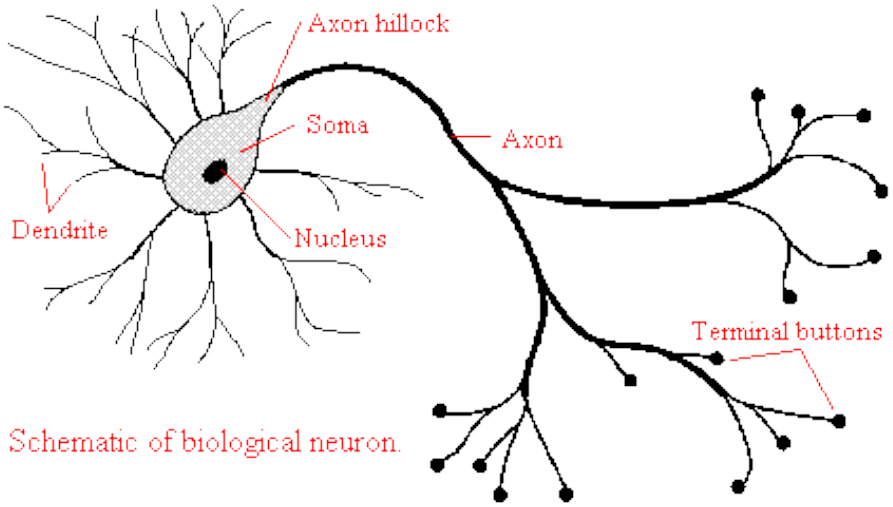
$$y_i = \text{sign}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)$$

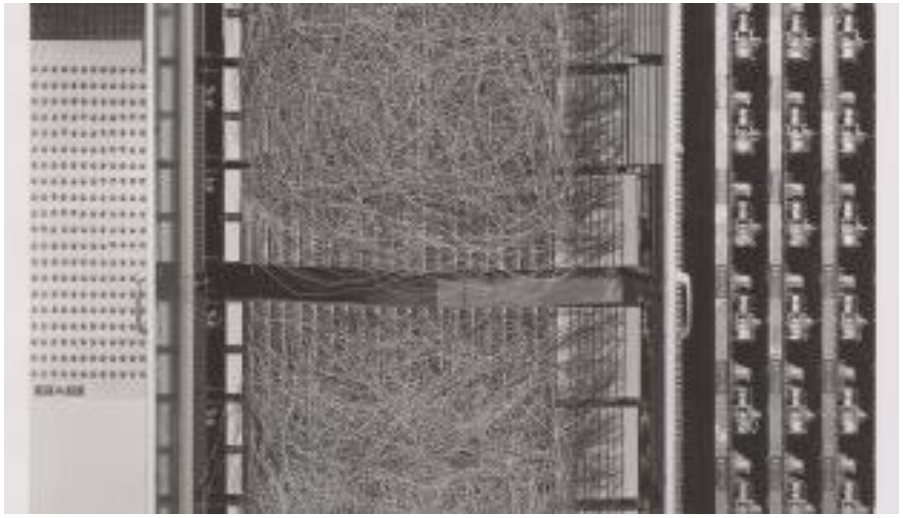
  - **w** in R^d: weight vector for the separating hyperplane
  - b in R: offset (threshold, bias) of the separating hyperplane
  - sign: thresholding function

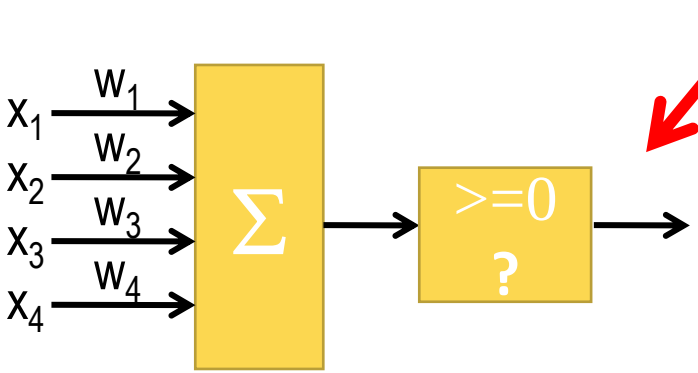
$$\text{sign}(t) = \begin{cases} 1, & t > 0 \\ -1, & t \leq 0 \end{cases}$$



Perceptron [Rosenblatt'58]







Nonlinear activation function

Linear superposition

UNIVERSITY OF WATERLOO

History

1958 New York Times…

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI) —The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of $100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

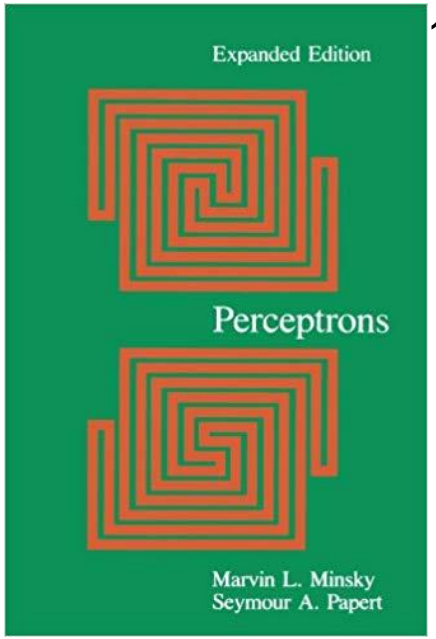
In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

Learns by Doing

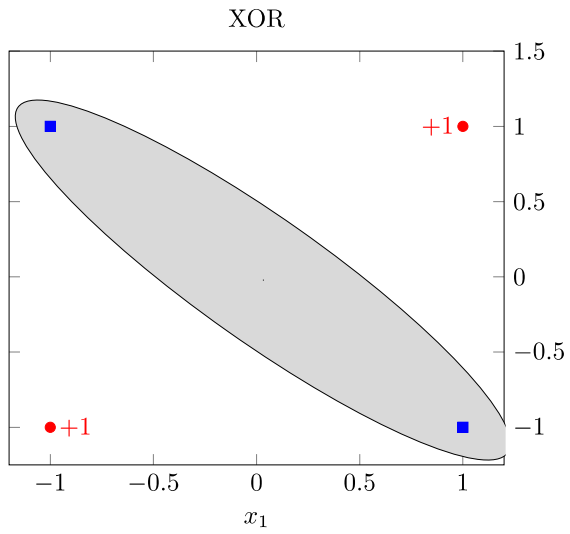
In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical terms. But he said the computer had undergone a "self-induced change in the wiring diagram."

The first Perceptron will have about 1,000 electronic "association cells" receiving electrical impulses from an eye-like scanning device with 400 photo-cells. The human brain has 10,000,000,000 responsive cells, including 100,000,000 connections with the eyes.



1969



UNIVERSITY OF WATERLOO

The perceptron algorithm

Algorithm: The Perceptron (Rosenblatt 1958)

Input: Dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{\pm 1\} : i = 1, \ldots, n\}$, initialization $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, threshold $\delta \geq 0$

Output: approximate solution $\mathbf{w}$ and b

```
for t = 1, 2, ... do
    receive training example index I_t ∈ {1, ..., n}        // the index I_t can be random
    if y_{I_t}(w^T x_{I_t} + b) ≤ δ then
        w ← w + y_{I_t} x_{I_t}                              // update only after making a "mistake"
        b ← b + y_{I_t}
```

- Typically $\delta = 0$, $\mathbf{w}_0 = \mathbf{0}$, $b_0 = 0$ (truth, prediction)
  - ✓ $y(\langle x, w\rangle + b) > 0$ implies $y = \text{sign}(\langle x, w\rangle + b)$
  - x $y(\langle x, w\rangle + b) < 0$ implies $y \neq \text{sign}(\langle x, w\rangle + b)$
  - ? $y(\langle x, w\rangle + b) = 0$ implies sitting on the hyperplane
- Lazy update: if it ain't broke, don't fix it

Does it work?

	and	viagra	the	of	nigeria	y
$\mathbf{x}_1$	1	1	0	1	1	+1
$\mathbf{x}_2$	0	0	1	1	0	-1
$\mathbf{x}_3$	0	1	1	0	0	+1
$\mathbf{x}_4$	1	0	0	1	0	-1
$\mathbf{x}_5$	1	0	1	0	1	+1
$\mathbf{x}_6$	1	0	1	1	0	-1

$$\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$$
$$b \leftarrow b + y$$

- $\mathbf{w}_0 = [0, 0, 0, 0, 0]$, $b_0 = 0$, pred -1 on $\mathbf{x}_1$, wrong
- $\mathbf{w}_1 = [1, 1, 0, 1, 1]$, $b_1 = 1$, pred 1 on $\mathbf{x}_2$, wrong
- $\mathbf{w}_2 = [1, 1, -1, 0, 1]$, $b_2 = 0$, pred -1 on $\mathbf{x}_3$, wrong
- $\mathbf{w}_3 = [1, 2, 0, 0, 1]$, $b_3 = 1$, pred 1 on $\mathbf{x}_4$, wrong
- $\mathbf{w}_4 = [0, 2, 0, -1, 1]$, $b_4 = 0$, pred 1 on $\mathbf{x}_5$, correct
- $\mathbf{w}_4 = [0, 2, 0, -1, 1]$, $b_4 = 0$, pred -1 on $\mathbf{x}_6$, correct

Simplification: Linear Feasibility

- Padding constant 1 to the end of each **x**

$$\langle \mathbf{w}, \mathbf{x} \rangle + b = \left\langle \begin{pmatrix} \mathbf{w} \\ b \end{pmatrix}, \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \right\rangle$$

denote as **z**

- Pre-multiply **x** with its label y

$$y(\langle \mathbf{w}, \mathbf{x} \rangle + b) = \langle \mathbf{z}, y \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \rangle$$

denote as **a**

- Find **z** such that $A^T\mathbf{z} > \mathbf{0}$, where $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$

Perceptron Convergence Theorem

Theorem (Block'62; Novikoff'62). Assume there exists some $\mathbf{z}$ such that $A^T\mathbf{z} > 0$, then the perceptron algorithm converges to some $\mathbf{z}^*$. If each column of A is selected indefinitely often, then $A^T\mathbf{z}^* > \delta$.

Corollary. Let $\delta = 0$ and $\mathbf{z}_0 = \mathbf{0}$. Then perceptron converges after at most $(R/\gamma)^2$ steps, where

$$R = \|A\|_{2,\infty} := \max_i \|\mathbf{a}_i\|_2 \qquad \gamma = \max_{\mathbf{z}:\|\mathbf{z}\|_2 \le 1} \min_i \langle \mathbf{z}, \mathbf{a}_i \rangle$$

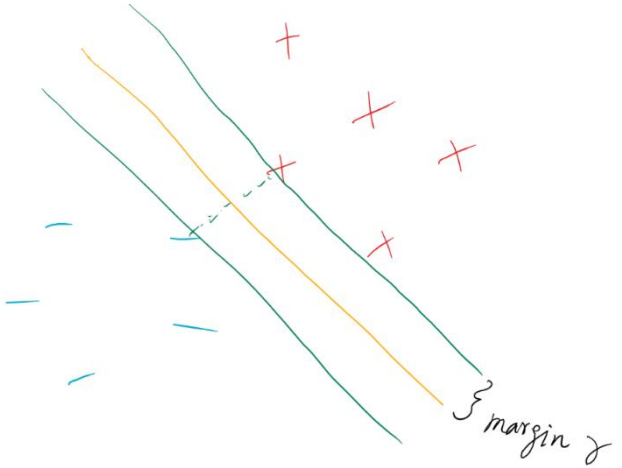
The Margin

- $\exists \mathbf{z}$ s.t. $A^T\mathbf{z} > \mathbf{0}$ iff for some hence all $s>\mathbf{0}$ $\exists \mathbf{z}$ s.t. $A^T\mathbf{z} > s\mathbf{1}$

- From the proof, perceptron convergence depends on:

$$\min_{(\mathbf{z},s):A^\top \mathbf{z}\geq s\mathbf{1}} \frac{\|\mathbf{z}\|_2^2}{s^2} = \min_{(\mathbf{z},s):\|\mathbf{z}\|_2\leq 1,\ A^\top \mathbf{z}\geq s\mathbf{1}} \frac{1}{s^2} = \left[\frac{1}{\max_{(\mathbf{z},s):\|\mathbf{z}\|_2\leq 1,\ A^\top \mathbf{z}\geq s\mathbf{1}} s}\right]^2 = \Big[\underbrace{\frac{1}{\max_{\mathbf{z}:\|\mathbf{z}\|_2\leq 1} \min_i \langle \mathbf{a}_i, \mathbf{z}\rangle}}_{\text{the margin } \gamma}\Big]^2$$

- The larger the margin is, the more (linearly) separable the data is, hence the faster perceptron learns.

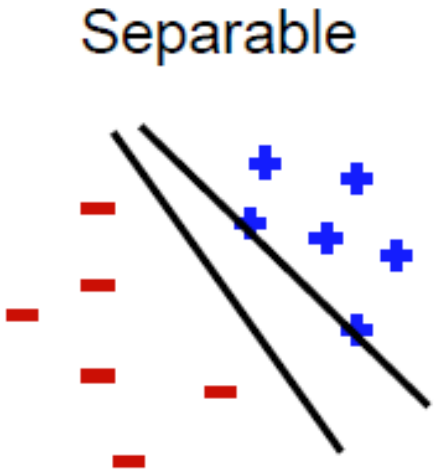


But

✓ The larger the margin, the faster perceptron converges

x But perceptron stops at an arbitrary linear separator…

- Which one do you prefer?



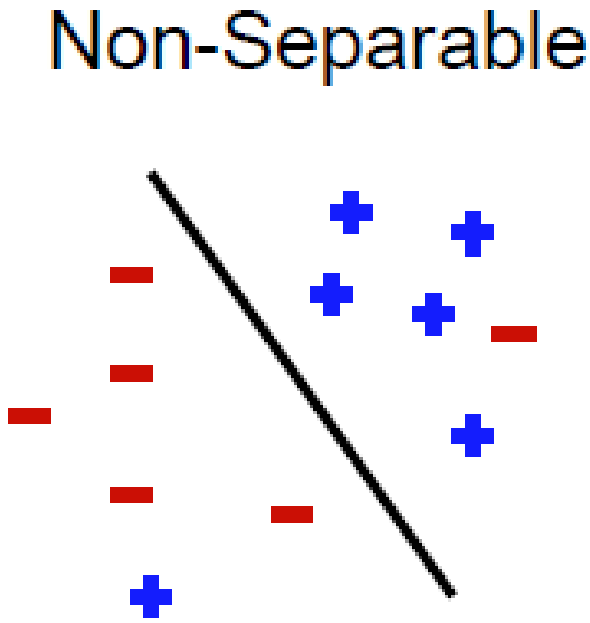
Support vector machines

$$\min_{A^\top \mathbf{w} \geq 1} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \approx \quad \min_{y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1} \frac{1}{2}\|\mathbf{w}\|_2^2$$

UNIVERSITY OF WATERLOO

What if non-separable?

- Find a better feature representation
- Use a deeper model
- Soft margin



Perceptron Boundedness Theorem

- Perceptron convergence requires the existence of a separating hyperplane
  - What if it fails? (trust me, it will)

Theorem (Minsky and Papert'67; Block and Levin'70).

The iterate $\mathbf{z} = (\mathbf{w}; b)$ of the perceptron algorithm is always bounded. In particular, if there is no separating hyperplane, then perceptron cycles.

"…proof of this theorem is complicated and obscure. So are the other proofs we have since seen..." --- Minsky and Papert, 1987

When to stop perceptron?

- Online learning: never

- Batch learning
  - Maximum number of iteration reached or run out of time
  - Training error stops changing
  - Validation error stops decreasing
  - Weights stopped changing much, if using a diminishing step size η_t, i.e., $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta_t y_i \mathbf{x}_i$

Multiclass Perceptron

- One vs. all
  - Class c as positive
  - All other classes as negative
  - Highest activation wins: pred = $\text{argmax}_c \mathbf{w}_c^T \mathbf{x}$
- One vs. one
  - Class c as positive
  - Class c' as negative
  - Voting

Questions?