INTRODUCTION TO MACHINE LEARNING COMPSCI 4ML3

Lecture 15

Hassan Ashtiani

LINEARLY SEPARABLE DATA

- A binary classification data set $Z = \{(x^i, y^i)\}_{i=1}^{n}$ is linearly separable if
  - There exists W^* such that
    - For every $i \in [n]$ we have $\mathrm{sgn}(< x^i, W^* >) = y^i$
    - Or equivalently, for every $i \in [n]$ we have $({W^*}^T x^i) y^i > 0$
- In other words, the classification error on Z is 0
- Can we find W^* efficiently for linearly separable data?

LP FOR LINEAR CLASSIFICATION

- DEFINE $A = \left[x_j^i y^i\right]_{n\times d}$
- THEN FINDING THE OPTIMAL W IS EQUIVALENT TO

$$\max_{w\in\mathbb{R}^d} < \vec{0}, w >$$

$$s.t.\ \ Aw \geq \vec{1}$$

WE CAN USE OFF-THE-SHELF LP SOLVERS!

APPROACH 2: PERCEPTRON

- Proposed in 50's by Rosenblatt
- Predecessor of Neural Networks
  - Multi-layer Perceptron!

ROSENBLATT'S PERCEPTRON

Batch Perceptron

input: A training set $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
initialize: $\mathbf{w}^{(1)} = (0, \ldots, 0)$
for $t = 1, 2, \ldots$
  if $(\exists\, i \text{ s.t. } y_i \langle \mathbf{w}^{(t)}, \mathbf{x}_i \rangle \leq 0)$ **then**
    $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i \mathbf{x}_i$
  else
    output $\mathbf{w}^{(t)}$

- In each update, W becomes "more correct" on x^i
- https://phiresky.github.io/kogsys-demos/neural-network-demo/?preset=Rosenblatt+Perceptron

THE GREEDY UPDATE

- In each update, W becomes "more correct" on x^i:

- What about other x^j's?

NOVIKOFF, 1962

UNCLASSIFIED

AD 298 258

Reproduced by the

ARMED SERVICES TECHNICAL INFORMATION AGENCY
ARLINGTON HALL STATION
ARLINGTON 12, VIRGINIA

Technical Report

ON CONVERGENCE PROOFS FOR PERCEPTRONS

Prepared for:

OFFICE OF NAVAL RESEARCH
WASHINGTON, D.C.

CONTRACT Nonr 3438(00)

By: *Albert B. J. Novikoff*

CATALOGED BY
AS AD NO.

STANFORD RESEARCH INSTITUTE

MENLO PARK, CALIFORNIA

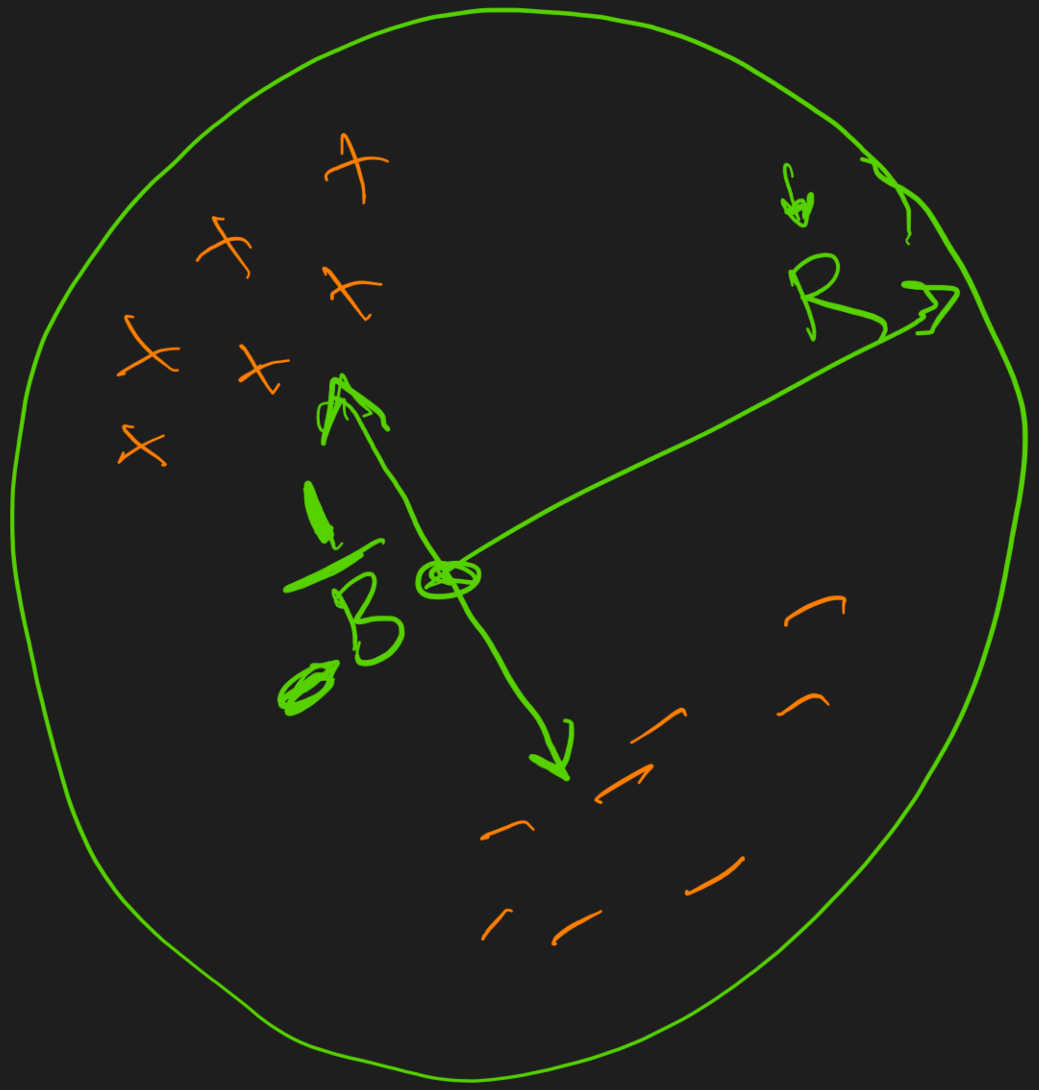
SRI

Novikoff, A. B. J. (1962). On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, Volume 12, pp. 615–622. Polytechnic Institute of Brooklyn.

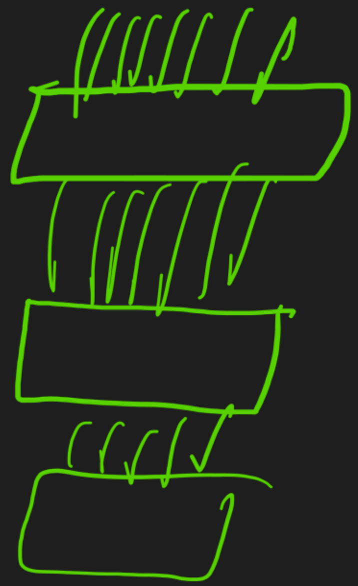
CONVERGENCE OF PERCEPTRON

THEOREM 9.1 *Assume that* $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$ *is separable, let* $B = \min\{\|\mathbf{w}\| : \forall i \in [m],\ \ y_i\langle \mathbf{w}, \mathbf{x}_i\rangle \geq 1\}$, *and let* $R = \max_i \|\mathbf{x}_i\|$. *Then, the Perceptron algorithm stops after at most* $(RB)^2$ *iterations, and when it stops it holds that* $\forall i \in [m],\ \ y_i\langle \mathbf{w}^{(t)}, \mathbf{x}_i\rangle > 0$.

- #STEPS DOES NOT EXPLICITLY DEPEND ON $\boxed{d}$
- YOU CAN FIND MORE DETAILS ABOUT THIS LECTURE IN
  - UNDERSTANDING MACHINE LEARNING, CHAPTER 9
  - HTTPS://WWW.CS.HUJI.AC.IL/~SHAIS/UNDERSTANDINGMACHINELEARNING/UNDERSTANDING-MACHINE-LEARNING-THEORY-ALGORITHMS.PDF

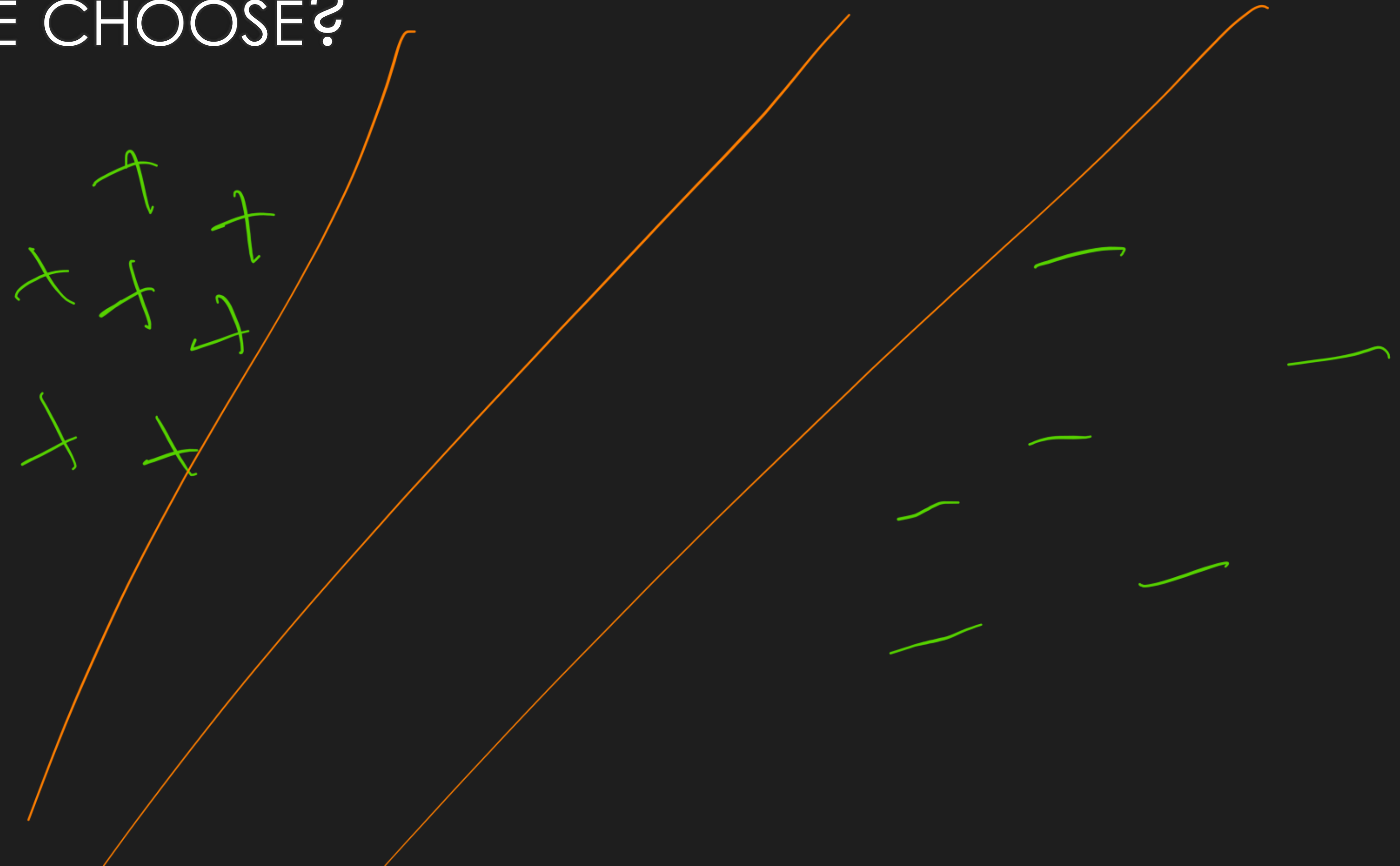

R
B

- In 1969, Marvin Minsky and Seymour Papert argued that it is impossible to learn XOR function using multilayer Perceptron…
  - Only good for linearly separable data
- Stacking perceptrons?
- 70's: AI (connectionism) winter



SUPPORT VECTOR MACHINES

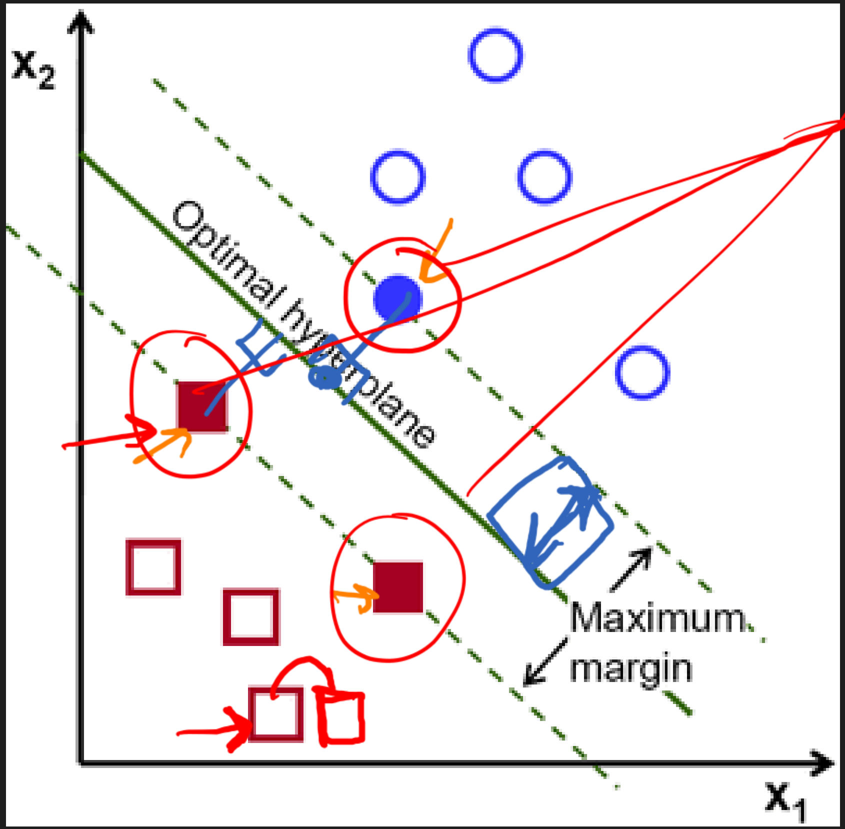
- AMONG PERFECT LINEAR SEPARATORS, WHICH ONE SHOULD WE CHOOSE?



SUPPORT VECTOR MACHINES

support vectors

- Pick the linear separator that maximizes the "margin"
- More robust to "perturbation"
- Less prone to overfitting
  - Works well for high-dimensional data (?)
  - More on that later!



DISTANCE OF A POINT TO A HYPERPLANE

THE EUCLIDEAN DISTANCE BETWEEN A POINT x AND THE HYPERPLANE PARAMETRIZED BY W IS (WHY?)

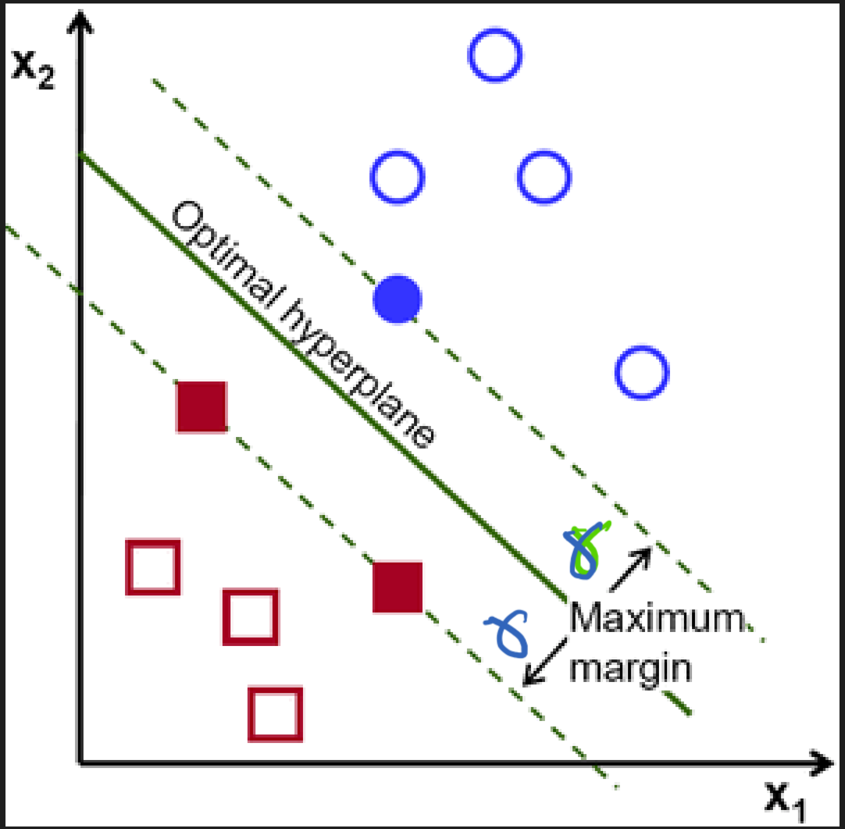
$$\frac{|W^T x + b|}{||W||_2}$$

- THE DECISION BOUNDARY OF A LINEAR CLASSIFIER IS DETERMINED BY THE DIRECTION OF W (NOT $||W||_2$)
- ASSUME $||W||_2=1$, THEN THE DISTANCE IS

$$|W^T x + b|$$

MAXIMUM MARGIN HYPERPLANE

- LET THE HYPERPLANE BE PARAMETRIZED BY W and b
- ASSUME $||W||_2 = 1$
- W HAS A γ MARGIN IF
  - $W^T x + b \geq \gamma$ FOR EVERY BLUE x, AND
  - $W^T x + b \leq -\gamma$ FOR EVERY RED x



For simplicity, assume $b=0$
$y \in \{\pm 1\}$

THE MARGIN

- $Z = \{(x^i, y^i)\}_{i=1}^n$, $y \in \{-1, +1\}$, $||W||_2 = 1$

$$\text{Margin}(Z, w) = \min_{(x,y)\in Z} w^T x y =$$

$$= \text{Max} \quad \gamma$$
$$\text{s.t.} \quad \forall (x,y) \in Z, \quad w^T x y \geq \gamma$$

$\text{Margin}(Z, w) < 0 \longrightarrow Z$ is not linearly separable.

MAXIMIZING THE MARGIN

$$\max_{w,\ \|w\|=1} \text{Margin}(Z, w) = \max \ \gamma \quad \text{s.t. } \exists w,\ \|w\|=1,\ \left\{ \forall (x,y) \in Z \right\} \left\{ w^T x y \geq \gamma \right\}$$

$$= \max \ \gamma \quad \text{s.t. } \exists w,\ \|w\|=1,\ \forall (x,y) \in Z,\ \frac{1}{\gamma} w^T x y \geq 1$$

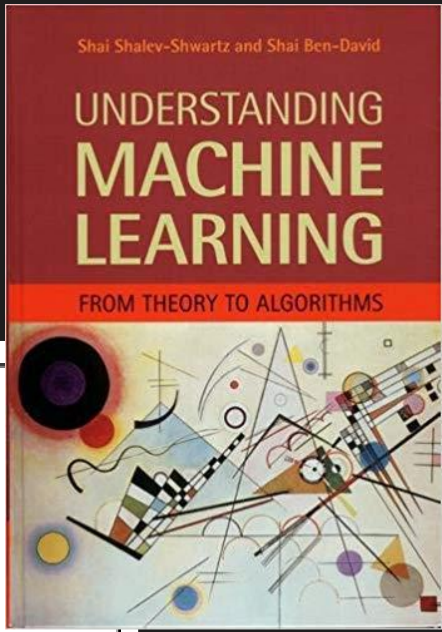
$$= \max \ \frac{1}{\|w\|} \quad \text{s.t. } \forall (x,y) \in Z,\ w^T x y \geq 1$$

so instead

$$\min_{w} \ \|w\|_2^2 \quad \text{s.t. } \forall (x,y) \in Z,\ w^T x y \geq 1$$

$w^{new} = \frac{w^{old}}{\|w\|}$

THE VERSION WITH "BIAS"



Hard-SVM

input: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
solve:

$$(\mathbf{w}_0, b_0) = \underset{(\mathbf{w},b)}{\operatorname{argmin}} \|\mathbf{w}\|^2 \text{ s.t. } \forall i,\ y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 \qquad (15.2)$$

output: $\hat{\mathbf{w}} = \frac{\mathbf{w}_0}{\|\mathbf{w}_0\|}, \quad \hat{b} = \frac{b_0}{\|\mathbf{w}_0\|}$

- WE COULD HAVE ALSO ADDED A DUMMY "1" FEATURE TO ALL POINTS SO AS TO ACCOUNT FOR THE BIAS/INTERCEPT

SENSITIVITY TO OUTLIERS