PERCEPTRON ALGORITHM

- $\mathcal{C} = \{\text{HALFSPACES}\}$

$$f(x) = \text{SIGN}\left(\sum_{i=1}^{n} w_i x_i - \theta\right)$$

threshold

W IS UNKNOWN

$f(x) \in \{-1, +1\}$

$x_i \in \mathbb{R}$

A "COMPLICATED" ALGORITHM FOR LEARNING A HALFSPACE

$(x^1, \underline{+1}) \rightsquigarrow w_1 x_1^1 + \cdots + w_n x_n^1 \geq \underline{\theta}$

$(x^2, \underline{-1}) \rightsquigarrow w_1 x_1^2 + \cdots + w_n x_n^2 < \underline{\theta}$

$\vdots \qquad\qquad \vdots$

m training SET $\to$ m inequalities

WE CAN USE LP $\rightsquigarrow$ FIND A WEIGHT VECTOR $\underline{w}$ CONSISTENT W/ TRAINING SET.

LINEAR PROGRAMMING SOLVERS

PERCEPTRON ALGORITHM.

MISTAKE-BOUNDED MODEL.

INITIALLY $w^0 = (0, \dots, 0)$ $w^0 = \left(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\right)$

LEARNER w AS ITS STATE

$$\xleftarrow{\text{TEACHER}} \quad x \in \mathbb{R}^n$$

LEARNER RESPONDS WITH

$$\xrightarrow{\text{SIGN}(w \cdot x)}$$

IF A MISTAKE IS MADE

CASE 1: x WAS TRULY A NEGATIVE EXAMPLE

$$w_{NEW} = w_{OLD} - x$$

2: x WAS TRULY A POSITIVE EXAMPLE

$$w_{NEW} = w_{OLD} + x$$

EQUIVALENT WAY TO VIEW UPDATE RULE:

EVERY TIME WE MAKE A MISTAKE

$$w_{NEW} = w_{OLD} + \underbrace{y}_{\text{LABEL}} \cdot x$$

ASSUMPTIONS:

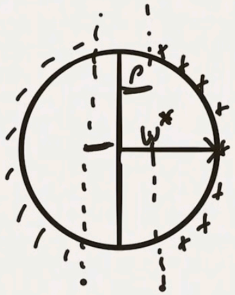
- ASSUME $\exists\, w^*$, TRUE UNKNOWN WEIGHT VECTOR

  $$\|w^*\|_2 = 1$$

- ASSUME X HAS NORM 1 $\|X\|_2 = 1$
- $\theta = 0$

<u>MAIN ASSUMPTION</u>: THERE EXISTS A <u>MARGIN</u> ρ

ALL POINTS ARE AT LEAST DISTANCE ρ FROM w^*



ALL POSITIVE/NEGATIVE POINTS HAVE DISTANCE $\geq \rho$ FROM HALFSPACE.

(BECAUSE $\|x\| = \|w^*\| = 1$)

EQUIVALENTLY $|\langle x, w^* \rangle| \geq \rho$

"MARGIN ASSUMPTION"

MAIN THEOREM "PERCEPTRON CONVERGENCE THEOREM" IS THAT THE MISTAKE BOUND OF PERCEPTRON ALGORITHM IS $O\left(\frac{1}{\rho^2}\right)$.

PROOF NOT TOO COMPLICATED!

UPDATE STEP:

$$W_{NEW} = W_{OLD} + y \cdot x$$

Let's say W is current state of learner
W^* is true normal to halfspace.

CLAIM 1: on every mistake $W \cdot W^*$ increases by at least ρ

CLAIM 2 $\|W\|^2$ increases after every mistake by at most 1

QUESTION: How to obtain the $O(\frac{1}{\rho^2})$ mistake bound given claims 1 and 2?

Let t be # mistakes we've made at some point during execution

$$t \cdot \rho \leq W \cdot W^* \leq \underbrace{\|W\|}_{\sqrt{t}} \cdot \cancel{\|W^*\|}^{1}$$

$$t \cdot \rho \leq \sqrt{t} \implies t \leq \frac{1}{\rho^2}$$

<u>CLAIM 1</u>

$w \cdot w^*$ INCREASES ON EVERY MISTAKE BY AT LEAST ρ

$$w_{NEW} = w_{OLD} + y \cdot x$$

$$w_{NEW} \cdot w^* = (w_{OLD} + y \cdot x) \cdot w^* = \underline{w_{OLD} \cdot w^*} + \underbrace{y \cdot x \cdot w^*}_{\geq \rho} \quad \blacksquare$$

CLAIM 2 $\|w\|^2$ INCREASES BY AT MOST 1 ON EVERY MISTAKE.

$$\underline{W_{NEW}} = \|W_{OLD} + y\cdot x\|^2 =$$

$$\underline{\|W_{OLD}\|^2} + \underbrace{2\cdot y\langle \underline{x, W_{OLD}}\rangle}_{\text{NEGATIVE}} + \underbrace{\|x\|^2}_{1}$$

ENDS PROOF OF CLAIM 2. ■

CLAIM 2 $\|w\|^2$ INCREASES BY AT MOST 1 ON EVERY MISTAKE.

$$W_{NEW} = \|W_{OLD} + y \cdot x\|^2 =$$

$$\|W_{OLD}\|^2 + \underbrace{2 \cdot y \langle x, W_{OLD} \rangle}_{\text{NEGATIVE}} + \underbrace{\|x\|^2}_{1}$$

ENDS PROOF OF CLAIM 2.

LOOK BACK AT SOME ASSUMPTIONS

- $\theta = 0$ ADD A NEW FEATURE CALL IT X_{n+1}

$(x_1, \dots, x_n) \rightarrow (x_1, \dots, x_{n+1})$ ALWAYS SET $X_{n+1} = 1$

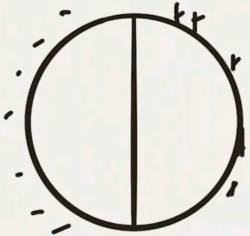
- $\|x\| = 1$ IF $\|x\| = R$ M.B. $O\left(\frac{R^2}{\rho^2}\right)$

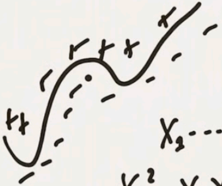
CONSIDER POLYNOMIAL THRESHOLD FUNCTIONS (PTFs)

$$f = \text{SIGN}\left(p(x)\right)$$

p IS A MULTIVARIATE POLYNOMIAL OF, LET'S SAY DEGREE d.

HOW CAN WE USE PERCEPTRON TO LEARN THIS FUNCTION CLASS?





FEATURE MAP

$(x_1, \ldots, x_n)$

$x_1 \cdots x_n \longrightarrow$

$x_1^2, x_1 x_2, \ldots x_n^2 \approx n^2$ NEW VARIABLES

$Y_1, \ldots, Y_{N=n^2}$ $\quad f = \text{SIGN}\left(\sum_{i=1}^{N} W_i Y_i\right)$

LEARNING PTFs OF DEGREE d IS EQUIVALENT TO LEARNING HALFSPACES IN n^d dimensions.

- RUN-TIME
  JUST COMPUTING THE FEATURE MAP TAKES TIME n^d.
- WHAT IS MARGIN IN THIS n^d DIMENSIONAL SPACE?
  (MAY BE COSTLY)

WE CAN SAVE ON THE RUNNING TIME

WE CAN SAVE DRAMATICALLY USING SOMETHING CALLED THE KERNEL TRICK!

KERNEL PERCEPTRON

$X \leftarrow$ INPUT $\varphi(x)$ to be the image of x IN THE FEATURE SPACE.

$x \in \mathbb{R}^n$ $\varphi(x) \in \mathbb{R}^{n^d}$

$K(\underline{x}^1, \underline{x}^2)$ AND OUTPUTS $\langle \varphi(x^1), \varphi(x^2) \rangle$

AND LET'S ASSUME $K(x^1, x^2)$ IS EASY TO COMPUTE.

KERNEL FUNCTION.

KERNEL PERCEPTRON (WANT TO WORK IN $\mathbb{R}^{n^d}$)

$W = 0^{n^d}$

LET'S ASSUME WE MAKE A MISTAKE ON POINT X^1

$$W_{NEW} = \underbrace{\cancel{W_{OLD}}}_{0} + y \cdot \varphi(x^1)$$

X^2 OR NEW POINT

WE NEED TO EVALUATE $W_{NEW} \cdot \varphi(x^2)$

$$= \langle y \cdot \varphi(x^1), \varphi(x^2) \rangle = y \cdot K(x^1, x^2)$$

$$K(x^1, x^2) = \langle \varphi(x^1), \varphi(x^2) \rangle$$

$$W_{t+1} = \sum_{i=1}^{t} y^i \cdot \varphi(x^i) \in \mathbb{R}^{n^d}$$

NEED TO COMPUTE $\langle W_{t+1}, \varphi(x^{t+1}) \rangle$

$$\sum_{i=1}^{t} y^i \cdot \underbrace{\langle \varphi(x^i), \varphi(x^{t+1}) \rangle}_{\boxed{K(x^i, x^{t+1})}}$$

EFFICIENTLY COMPUTABLE.

EXAMPLE OF A SIMPLE KERNEL FUNCTION

(CONSIDER DEG-2 POLYNOMIAL THRESHOLD FUNCTIONS)

$$\varphi(x_1 \dots x_n) = (\underbrace{x_1 x_1, x_1 x_2, \dots, x_{n-1} \cdot x_n, x_n^2}).$$

$$K(x, z) = \langle \varphi(x), \varphi(z) \rangle \quad (z_1 z_1, z_1 z_2, \dots, z_n^2)$$

$(x, z \in \mathbb{R}^n)$

$$(x_1^2 z_1^2 + x_1 x_2 z_1 z_2 + \dots x_n^2 z_n^2)$$

$$= \sum_{i,j} x_i x_j z_i z_j = \underbrace{\left(\sum_{i=1}^{n} x_i z_i\right) \cdot \left(\sum_{j=1}^{n} x_j z_j\right)}$$

$$= (x \cdot z)^2 = \underline{K(x, z)}$$