Linear model for classification:

Reading: (UML ch. 9, PRML ch. 4)

$\mathcal{X}$: domain set

$\mathcal{X} = \mathbb{R}^D$

$\underline{x}_i \in \mathbb{R}^D$ $\quad i = 1, \ldots, N \rightarrow$ training examples

$X \in \mathbb{R}^{N \times D}$

$\mathcal{Y} = \{+1, -1\}$

Correct labelling function : $y_i \in \mathcal{Y}$
$:= f(\underline{x}_i)$

Training set : $\mathcal{D} = \{(\underline{x}_1, y_1), \ldots (\underline{x}_N, y_N)\}$

Goal: To assign each $\underline{x} \in \mathbb{R}^D$ (or $\mathcal{X}$) to one of the two classes C_1 or C

Hypothesis: $h : \mathcal{X} \to \mathcal{Y} \equiv \mathbb{R}^D \to \{+1, -1\}$

ERM:

$$\min_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left[h(\underline{x}_i) \neq y_i\right]$$

Indicator function

$$\mathbb{1}(a) = \begin{cases} 1 & a \text{ is true} \\ 0 & \text{o.w.} \end{cases}$$

Assumptions: $\{\underline{x}_i\}$ are linearly separable

Discriminant functions: A function that assigns $\underline{x}$ to a class C_k.

Affine functions of $\underline{x}$:

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

$$= \langle \underline{w}, \underline{x} \rangle + b$$

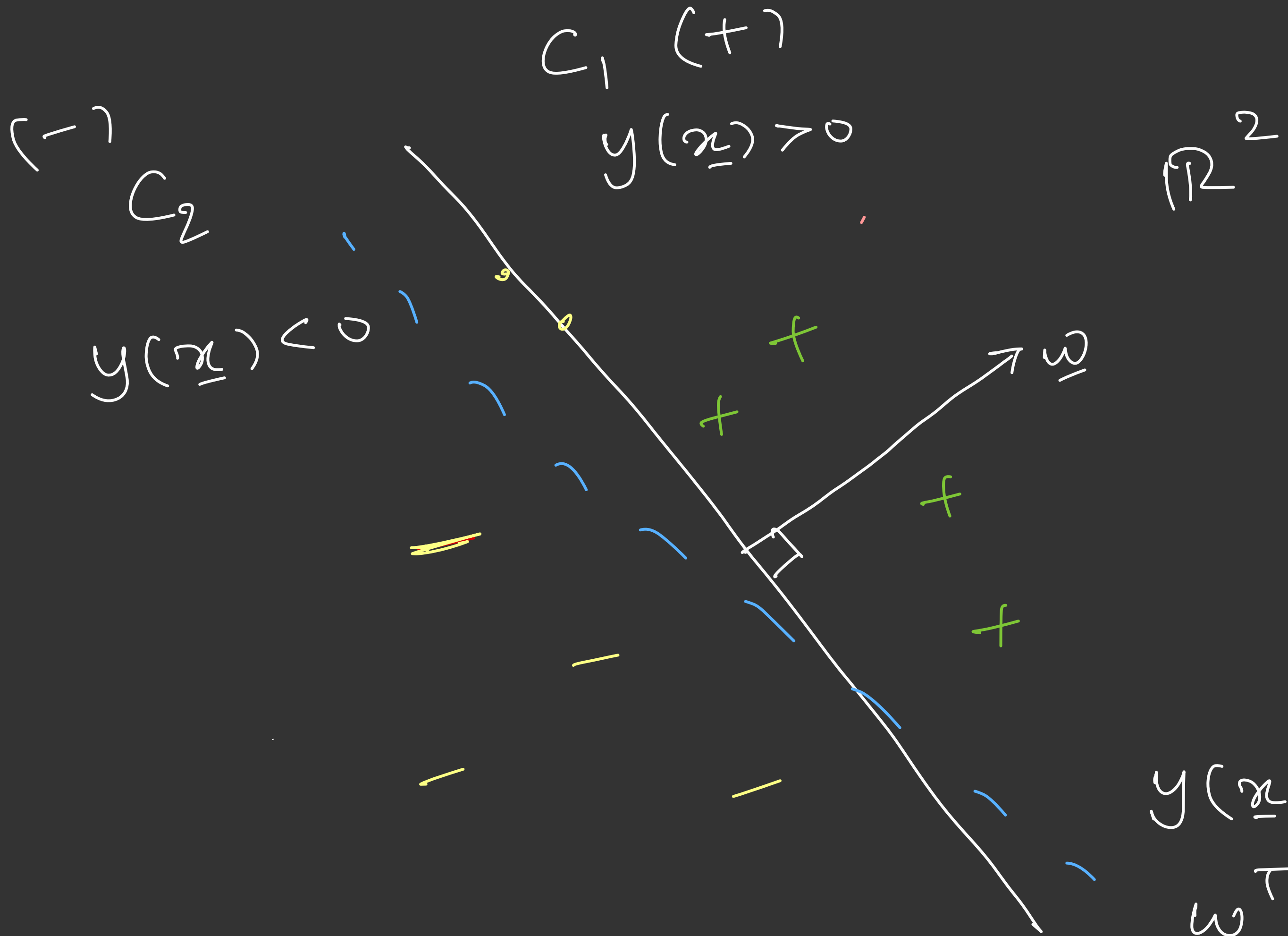
$$= \sum_{i=1}^{D} w_i x_i + b$$

weight

bias

Parameters: $\underline{w} \in \mathbb{R}^D$ and $b \in \mathbb{R}$

Half space:



$y(\underline{x}) = 0$

$$\underline{w}^T \underline{x} = \underbrace{-b}_{\text{"decision threshold"}}$$

$$\underline{w}' = \begin{bmatrix} \underline{w} \\ b \end{bmatrix} \in \mathbb{R}^{D+1} \qquad \underline{x}' = \begin{bmatrix} x \\ 1 \end{bmatrix}$$

$$y(\underline{x}) = (\underline{w}')^T \underline{x}$$

<u>ER</u> :

$$L_D(\underline{w}, b) = \frac{1}{N} \sum_{i=1}^{D} \mathbb{1}\left[\operatorname{sign}(\underline{w}^T \underline{x} + b) \neq y_i\right]$$

<u>Hypothesis class</u> of <u>Half Spaces</u>:

$$\mathcal{H} := \left\{ \underline{x} \longmapsto \operatorname{sign}(\underline{w}^T \underline{x} + b) \;\middle|\; \underline{w} \in \mathbb{R}^D,\ b \in \mathbb{R} \right\}$$

$$\text{sign}(a) = \begin{cases} +1 & \text{if } a > 0 \\ -1 & \text{o.w.} \end{cases}$$

ERM.

$$\underset{\substack{\underline{w} \in \mathbb{R}^D \\ b \in \mathbb{R}}}{\text{minimize}} \quad \mathcal{L}_D(\underline{w}, b)$$

Realizable case: w.l.o.g. $b = 0$

$\exists \, \underline{w}^*$ such that $\underline{y} = \text{sign}\,\cancel{(}\underline{w}^*, \underline{x})$

I. Linear program (LP)

Seek $\underline{w} \in \mathbb{R}^D$ s.t.

$$\text{sign}(\underline{w}^T \underline{x}_i) = y_i$$

$$i = 1, \ldots, N$$

$$\equiv \quad y_i(\underline{w}^T \underline{x}_i) > 0$$

Define

$$\gamma = \min_i \left(y_i (\underline{w}^T \underline{x}_i) \right)$$

$$\underline{\bar{w}} = \frac{\underline{w}^*}{\gamma}$$

Therefore, $$y_i\left(\underline{\bar{w}}^T \underline{x}_i\right) = \frac{1}{\gamma}\langle \underline{w}^*, x_i \rangle$$
$$\geq 1$$

Define. $$\underline{u} = \underline{0}$$
$$A_{ij} = y_i x_{ij}$$

LP: $$\max_{\underline{w}} \; \underline{0}^T \underline{w} \quad \text{s.t.} \quad A\underline{w} \geq \underline{1}$$

Perceptron for Half space:

Frank Rosenblat 1958

Iterative: $\underline{w}^{(1)}, \underline{w}^{(2)}, \ldots,$

At iteration t, find example i

s.t. $\operatorname{sign}\langle \underline{w}^{(t)}, \underline{x}_i \rangle \neq y_i$

update: $\underline{w}^{(t+1)} = \underline{w}^{(t)} + y_i \underline{x}_i$

$$y_i \langle \underline{w}^{(t+1)}, \underline{x}_i \rangle = y_i \langle (\underline{w}^{(t)} + y_i x_i), x_i \rangle$$

$$= \underbrace{y_i \langle \underline{w}^{(t)}, x_i \rangle}_{\leq 0} + \|x_i\|^2$$

Algorithm:

I/p : D , $\underline{w}^{(1)} = \underline{0}$

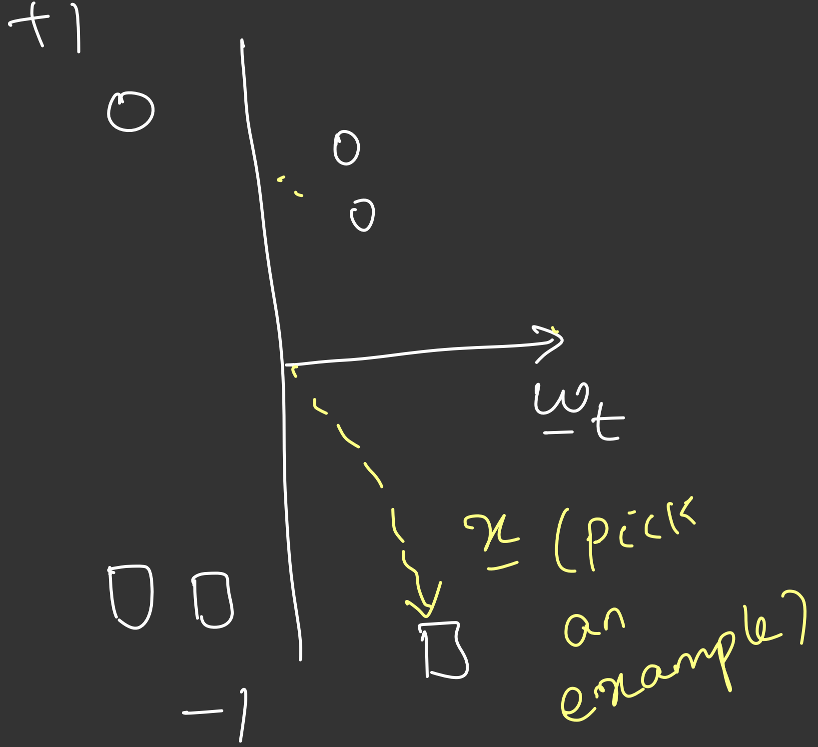
for $t = 1, 2, \ldots$

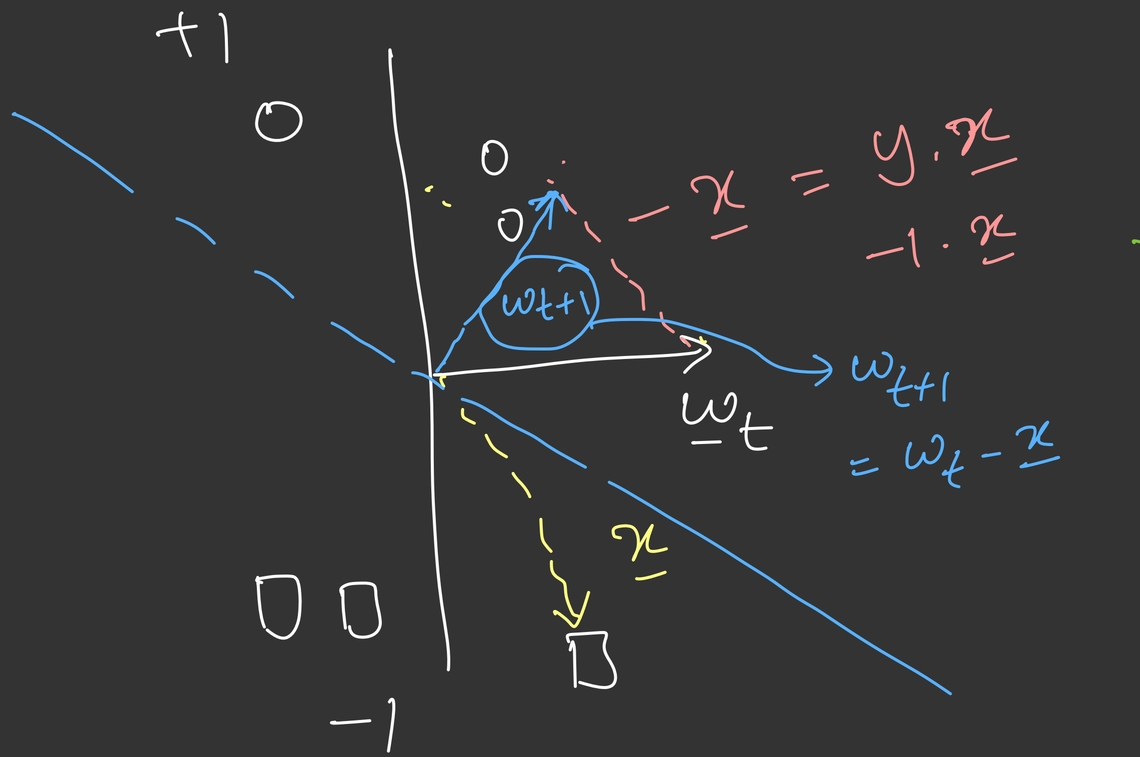
if [$\exists\, i$ s.t. $y_i \langle \underline{w}^{(t)}, \underline{x}_i \rangle \leq 0$]

$$\underline{w}^{(t+1)} = \underline{w}^{(t)} + y_i \underline{x}_i$$

else

o/p $\underline{w}^{(t)}$





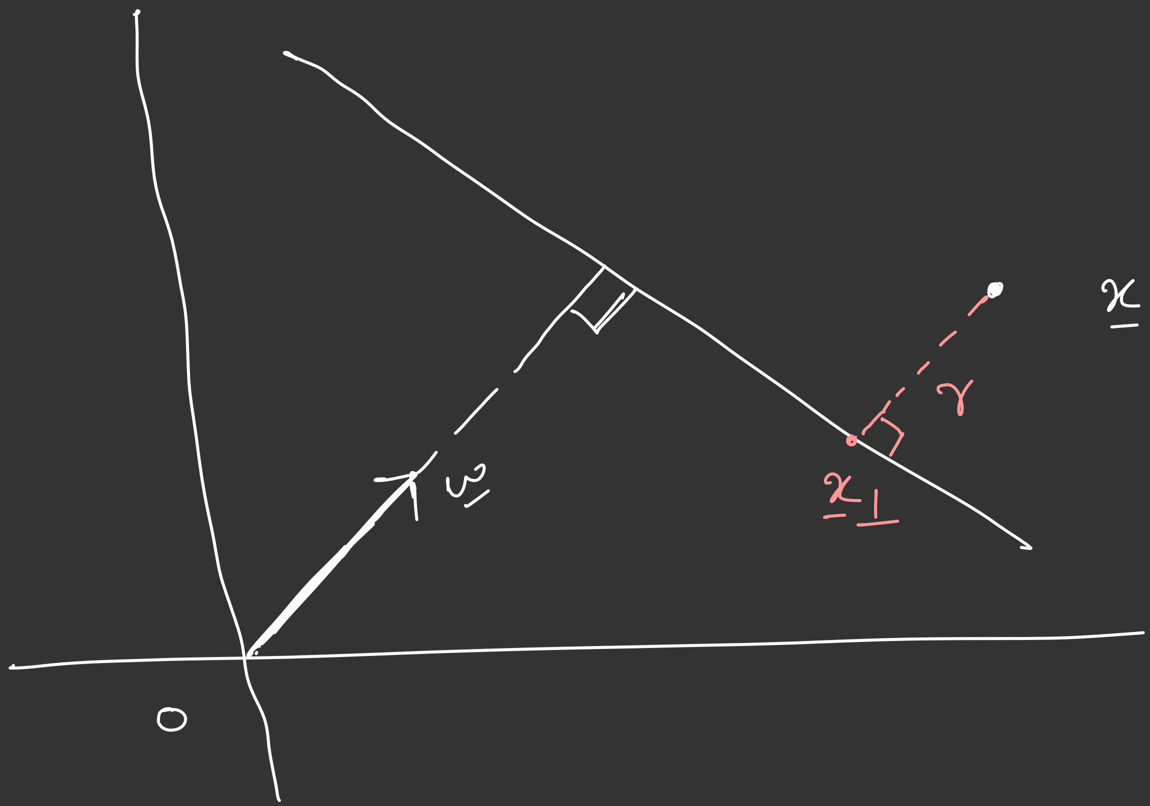
Side note: $y_i(\underline{w}^T \underline{x}_i) > 0 \; , \; \forall i$

Surrogate loss to minimize : $-\underbrace{\sum_{i=1}^{N} y_i(\underline{w}^T \underline{x}_i)}_{\text{make this small } (\leq 0)}$

Gradient descent : $w_{t+1} = \underline{w}_t + y_i \underline{x}_i$

[Perceptron also.]

Geometric margin:



$$\underline{x} = \underline{x}_\perp + \gamma \cdot \frac{\underline{w}}{\|\underline{w}\|}$$

$$\underline{w}^T \underline{x} = \underbrace{\underline{w}^T \underline{x}_\perp}_{=0} + \gamma \|w\|$$

$$\tilde{\gamma} = \frac{\underline{w}^T \underline{x}}{\|\underline{w}\|}$$

Define:

$$\gamma_i := y_i \frac{(\underline{w}^T x)}{\|\underline{w}\|}$$

Let

$$\gamma = \min_i \gamma_i \quad \text{and} \quad \tilde{\omega} = \frac{w^*}{\gamma}$$

$$y_i < \underline{\tilde{w}}, x_i > = \frac{y_i}{\gamma} < \underline{w}^*, x_i > \geq \| \underline{w}^* \|$$

Perceptron convergence:

Assume $\{ \underline{x}_i, y_i \}_{i=1}^{N}$ is linearly separable.

Let $B = \min \{ \| \underline{w} \| \mid y_i \langle \underline{w}, x_i \rangle \geq 1, \ i=1, \ldots, N \}$

and $R = \max_i \| \underline{x}_i \|$. Then, the

Perceptron method stops after at most

$(RB)^2$ iterations, and when it stops

$y_i \langle w^{(t)}, \underline{x}_i \rangle > 0 \quad \forall i$.

Proof:

Let $\underline{w}^*$ be the optimal vector

s.t. $y_i \langle \underline{w}^*, \underline{x}_i \rangle \geq 1 \quad \forall i$

i.e., $B = \| \underline{w}^* \|$

To show:

$$\underbrace{\frac{\langle \underline{w}^*, \underline{w}^{(T+1)} \rangle}{\| \underline{w}^* \| \, \| \underline{w}^{(T+1)} \|}} \geq \frac{\sqrt{T}}{RB}$$

Init: $\underline{w}^{(0)} = \underline{0}$

$$1 \geq \frac{\sqrt{T}}{RB}$$

(due to Cauchy-Schwarz)

$$\Rightarrow (RB)^2 \geq T$$

$$\langle \underline{w}^*, \underline{w}^{(0)} \rangle = 0$$

$$\langle \underline{w}^*, \underline{w}^{(t+1)} \rangle - \langle \underline{w}^*, \underline{w}^{(t)} \rangle$$

$$= \langle \underline{w}^*, \underline{w}^{(t+1)} - \underline{w}^{(t)} \rangle$$

$$= \langle \underline{w}^*, y_i \underline{x}_i \rangle = y_i \langle \underline{w}^*, x_i \rangle$$

$$\geq 1$$

Telescopic sum:

$$\langle \underline{w}^*, \underline{w}^{(T+1)} \rangle = \sum_{t=0}^{T} \langle \underline{w}^*, \underline{w}^{(t+1)} \rangle - \langle \underline{w}^*, \underline{w}^{(t)} \rangle$$

$$\geq T$$

$$\|\underline{w}^{(t+1)}\|^2 = \|w^{(t)} + y_i \underline{x}_i\|^2$$

$$= \|\underline{w}^{(t)}\|^2 + 2\, y_i \langle \underline{w}^{(t)}, \underline{x}_i \rangle + y_i^2 \|x_i\|^2$$

$$\leq \|\underline{w}^{(t)}\|^2 + R^2$$

$$\sum_{t=0}^{T} \left(\|\underline{w}^{(t+1)}\|^2 - \|\underline{w}^{(t)}\|^2 \right) \leq T R^2$$

$$\|\underline{w}^{(T+1)}\| \leq \sqrt{T}\, R$$

$$\Rightarrow \quad \frac{\langle \underline{w}^{(T+1)}, \underline{w}^{*} \rangle}{\|\underline{w}^{(T+1)}\| \cdot \|\underline{w}^{*}\|} \geq \frac{T}{B\sqrt{T}R} = \frac{\sqrt{T}}{RB}$$

$\blacksquare$