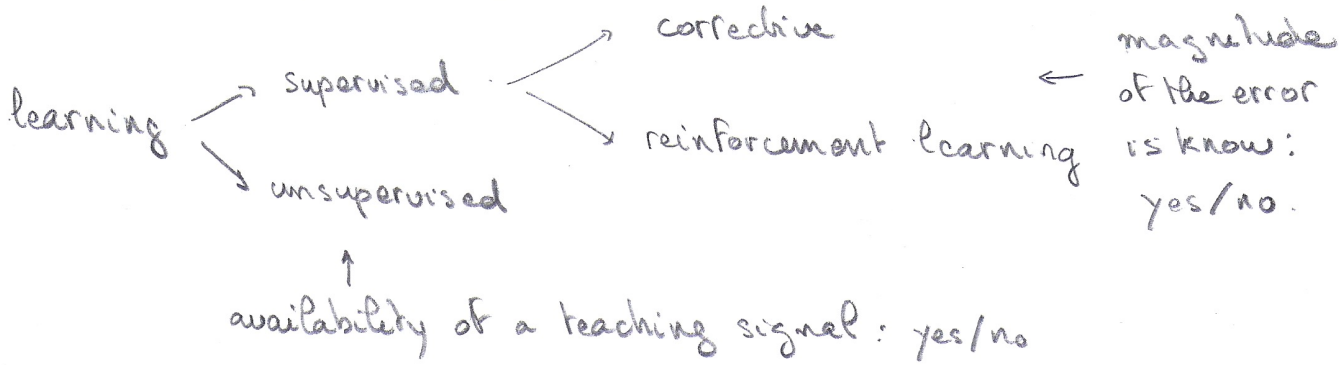
Classes of learning algorithms:



Perceptron algorithm (Rosenblatt 1957)

↳ supervised learning with reinforcement performing binary classification (task)

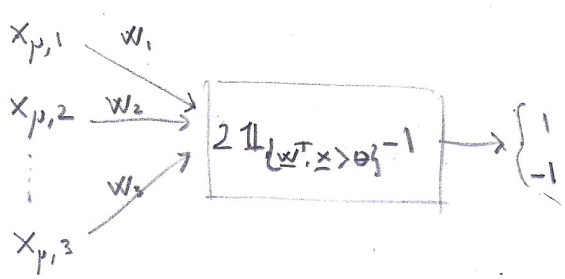
input: "pattern = collection of points $\underline{x}_\mu$, $\mu = 1, \dots, P$ config." in a N-dimensional space.

output: "boolean assignments" = $\underline{x}_\mu \mapsto y_\mu \in \{-1, 1\}$ (false, true)

Perceptron architecture

↳ single-layer perceptron = linear threshold function specified by a N-dimensional weight vector $\underline{w}$ and a threshold θ.

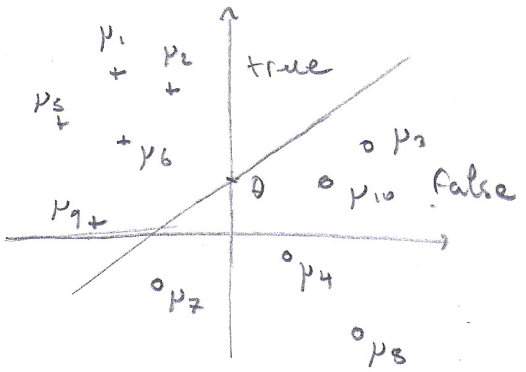
$$y_\mu = 1 \iff \underline{w}^T \cdot \underline{x}_\mu > \theta$$



Inspiration from neural network: first instance of artificial neural networks.

Linear separability

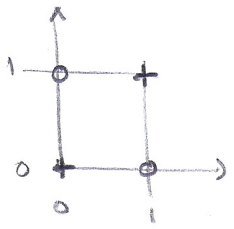
↳ Geometrically, a perceptron defined an hyperplane as a decision surface separating patterns into true instances and false instances



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

Are all configurations of points linearly separable? — NO!

↳ XOR :



: In 2-D, there is no line separating $\{(1,0), (0,1)\}$ from $\{(0,0), (1,1)\}$

↳ inherent limitation of the linear setting!

The complexity of the perceptron model is defined by the number of true false assignment, i.e. the number of dichotomies, that can be realized via linear threshold functions given generic input configurations.

↳ concept of general position

General position

↳ extension of the concept of linear dependence to the case of a family of points whose size P may exceed the space dimension N.

↳ A configuration $(\underline{x}_1, \dots, \underline{x}_P)$ is in general position if there is no subset $J \subset \{1, \dots, P\}$, $|J| \leq N$, that is linearly dependent.

↳ This property is satisfied with probability one whenever $(\underline{x}_1, \dots, \underline{x}_P)$ are sampled from distribution with smooth density.

Cover's function counting theorem:

IF $(x_1, \dots, x_P)$ is in general position the number of linearly separable patterns (i.e. of dichotomies) is

$$C(P,N) = 2 \sum_{k=0}^{N-1} \binom{P-1}{k} \quad \text{where} \quad \binom{n}{m} = \frac{n!}{m!(n-m)!}$$

Consequences: * $P \leq N$, $C(P,N) = 2(1+1)^{P-1} = 2^P$ ← all combinations are possible

* $P = 2N$, $C(P,N) = 2 \sum_{k=0}^{N-1} \binom{2N-1}{k} = 2 \frac{1}{2} 2^{2N-1} = 2^{P-1}$ ↑ half of the possible combinations

* $P \gg 2N$, $C(P,N) \propto P^N$ ← polynomial growth

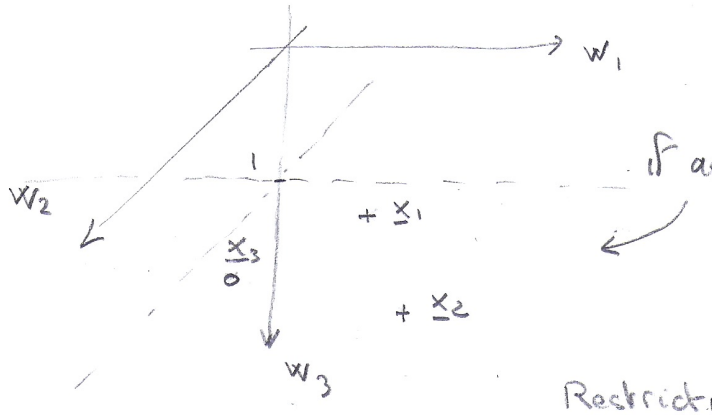
Proof: next class!

Reduction of the problem

↳ Finding (w, θ) realizing a dichotomy assumed to be realizable?

↳ Threshold plays no particular role: extended vectors

$$\underline{x} = (x_1, \dots, x_n, 1), \quad \underline{w} = (w_1, \dots, w_n, -\theta)$$



not necessary!

↑

if assumed in general position:
in 2-D: no three points lie on the same line

Restriction to zero threshold Function

Perceptron algorithm

$\underline{x}_\mu$, $\mu = 1, \dots, N$: input, $y_\mu = \{-1, 1, 1, -1, \dots\}$: teaching signal

initial state: random weight vector. $\underline{w}_0$

The algorithm cycles through all the inputs, possibly many times: at each step $n \geq 1$, the μ_n-th input is presented and two outcomes are possible:

* either: $y_n \cdot (\underline{w}_{n-1}^T \cdot \underline{x}_n) > 0$ with $y_n = y_{\mu_n}$

  ↑ current weights $\quad \underline{x}_n = \underline{x}_{\mu_n}$

  $\Rightarrow$ the perceptron properly classifies $\underline{x}_n$

* either: $y_n \cdot (\underline{w}_{n-1}^T \cdot \underline{x}_n) < 0$

  $\Rightarrow$ the perceptron is wrong!

  update rule: $\underline{w}_n = \underline{w}_{n-1} + \eta\, y_n\, \underline{x}_n$

  ↑ learning rate

Perceptron convergence theorem

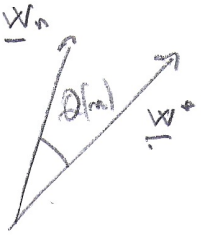
For any finite set of linearly separable labeled configurations, the perceptron will halt after a finite number of iterations.

↳ the number of iterations required for the algorithm to converge is typically larger than the training set, as as points need to be presented multiple times.

Proof: linear separability ⇒ there exists $\underline{w}^*$ such that

$$y_\mu (\underline{w}^{*T} \underline{x}_\mu) > 0 \quad \text{for all } \mu.$$

idea: consider the angle between $\underline{w}^*$ and $\underline{w}_n$, the current perceptron weight vector.

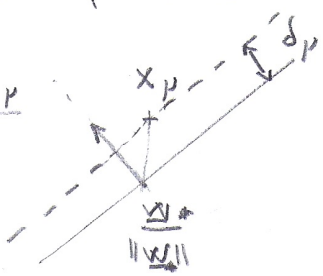


$$\cos\theta(n) = \frac{\underline{w}_n^T \cdot \underline{w}^*}{\|\underline{w}_n\| \cdot \|\underline{w}^*\|}$$

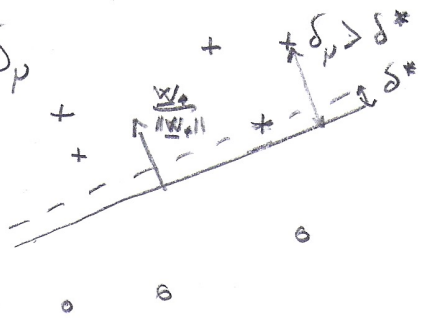
contradiction argument: we will see that if $n \to +\infty$, because $\underline{w}_n$ can be seen as some form of biased random walk toward $\underline{w}^*$, $\cos\theta(n)$ would grow strictly larger than 1, which is impossible.

important quantity: margin $\delta_\mu = \dfrac{y_\mu \underline{w}^{*T} \underline{x}_\mu}{\|\underline{w}^*\|}$

↑ Euclidean distance to the hyperplane $\underline{w}^* \cdot \underline{x} = 0$



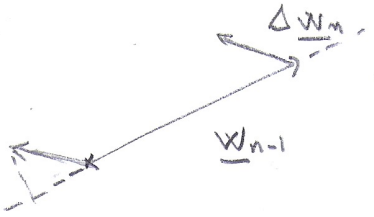
minimal margin: $\delta^* = \min_\mu \delta_\mu$



update step

$$y_n \cdot (\underline{w}_{n-1}^T \cdot \underline{x}_n) < 0 \quad \Rightarrow \quad \Delta \underline{w}_n = \underline{w}_n - \underline{w}_{n-1} = \eta\, y_n\, \underline{x}_n$$

moreover : $$\underline{w}_{n-1}^T \cdot \Delta \underline{w}_n = \eta\, y_n (\underline{w}_{n-1}^T \cdot \underline{x}_n) < 0$$



"projection of correction vector is negative"

numerator :

$$\begin{aligned} \underline{w}_n^T \cdot \underline{w}^* &= \underline{w}_{n-1}^T \cdot \underline{w}^* + \Delta \underline{w}_n^T \cdot \underline{w}^* \\ &= \underline{w}_{n-1}^T \cdot \underline{w}^* + \eta\, y_n \underline{w}^{*T} \cdot \underline{x}_n \\ &= \underline{w}_{n-1}^T \cdot \underline{w}^* + \eta\, \delta_n \|\underline{w}^*\| \geqslant \underline{w}_{n-1}^T \cdot \underline{w}^* + \eta\, \delta^* \|w^*\| \end{aligned}$$

definition of δ^* ↓

iterating on n : $$\underline{w}_n^T \cdot \underline{w}^* \geqslant \underline{w}_0^T \cdot \underline{w}^* + n \eta\, \delta^* \|w^*\|$$

denominator :

$$\begin{aligned} \|w_n\|^2 &= \|\underline{w}_{n-1} + \Delta \underline{w}_n\|^2 \\ &= \|\underline{w}_{n-1}\|^2 + 2\, \underline{w}_{n-1}^T \cdot \Delta \underline{w}_n + \|\Delta \underline{w}_n\|^2 \\ &\leqslant \|\underline{w}_{n-1}\|^2 + \eta^2 D^2 \end{aligned}$$

$\underline{w}_{n-1}^T \cdot \Delta \underline{w}_n < 0$

$\|\Delta \underline{w}_n\|^2 = \eta^2 \|x_n\|^2 \leqslant \eta^2 D^2$

iterating on n : $$\|w_n\|^2 \leqslant \|w_0\|^2 + n \eta^2 D^2$$

contradiction : if $n \longrightarrow +\infty$, $\cos \theta(n) \geqslant \dfrac{n \eta\, \delta^* \|w^*\|}{\sqrt{n}\, \eta\, D\, \|w^*\|} = \dfrac{\delta^*}{D} \sqrt{n} \longrightarrow +\infty$

convergence time : $n \leqslant D^2 / \delta_{max}^2$, $\delta_{max} = \arg \max_{w^*} \delta^*$

↑ does not play a particular role