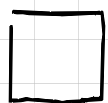
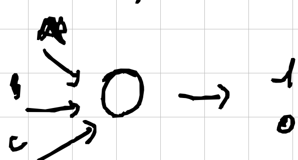
Statistical Physics of Learning
(Michael Biehl)

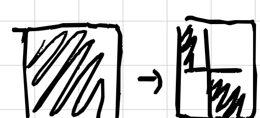
A bit of history...

IN →  → OUT

FF
DT
SVM
RF
DNN
⋮

- McCulloch & Pitts (1943) : neuron



- Hebb : synaptic plasticity (1949) 

  " Fire together, wire together "

- John Hopfield → NN
  (Amari, Little)

AUTO - ASSOCIATIVE MEMORY

Energy driven dynamics, analogy with magnet materials (spin-glass)

The space of interactions in NN (Gardner)

UNIVERSALLY VALID PRINCIPLE / ERROR BOUNDS

STATISTICAL PHYSICS OF LEARNING

OBSERVATIONS / CONCRETE APPLICATIONS

Ex. Vapnik - Chervonenkis theory

[Books: - Introduction to the Theory of Neural Computation
- Statistical Mechanics of Learning (Engels)
- Information, Physics and Computation]

The Perceptron (1957) [Rosenblatt]

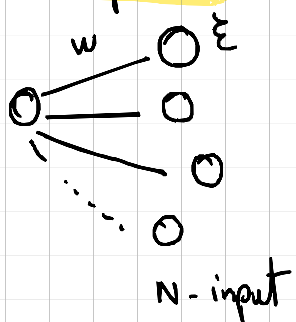
[Perceptrons: An introduction to computational geometry]

Adaptive Linear Neuron - Adaline

Basic concepts of SPL

- Linear separability
  - Rosenblit's perception algorithm
  - Learning in version space
- Student / teacher scenario
  - HD data, large Networks, generalization error
- typical learning curves
  - ONLINE / OFFLINE LEARNING

The Perceptron



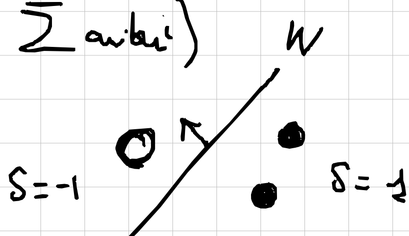
$$S = \text{sign}(W \cdot \xi - \theta) = \pm 1$$

- Perceptron Convergence Theorem [Rosenblatt]
- Capacity [Cover Theorem]

Linearly separable Function

$$S(\xi) = \text{sign}(w \xi) \quad \forall\, w \in \mathbb{R}^N$$

$$\left(a \cdot b = a^T b = \sum a_i b_i\right)$$



$$D = \{ \xi^\mu, S^\mu = \text{sign}(w \cdot \xi^\mu) \}_{\mu=1}^{P}$$

↳ linearly separable algorithm

Given D, initialize $w(0) = \emptyset$,

present $\{ \xi^t, S^t \}$ in order

$$t = \underbrace{1, \dots P}_{\text{epoch}}, 1, \dots P, 1 \dots P$$

$$w(t) = \begin{cases} w(t-1) & \text{CORRECT} \\ w(t-1) + \frac{\eta}{N} \xi^t S^t & \text{ELSE} \end{cases}$$

Hebbian → INPUT * OUTPUT
(Fire together, wire together)

Perceptron Convergence Theorem

[if $D = \{ \xi^\mu, S^\mu \}$ is linearly separable

Rosenblatt's perceptron converges after a finite number of epochs with $\text{sign}(w \cdot \xi^\mu) = S^\mu$ $\forall \mu = 1, \dots P$.]

↳ How long will it take?

Consider a rule $S_R(\xi) = \text{sign}(w^* \cdot \xi)$ defined by a specific vector w^* TEACHER

Consider a Training set $D = \{\xi^\mu\}_{\mu=1}^{P}$,

Find a $\bar{w}$ (student) with a consistent hypothesis S_H with

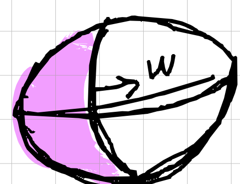
$$S_H(\xi^\mu) = S_H^\mu = \text{sign}(w \cdot \xi^\mu) = S_R^\mu \quad \mu = 1 \ldots P$$

↳ Learning a VERSION SPACE

$$V_D = \{ w \mid \text{sign}(w \cdot \xi^\mu) = S_R^\mu \}$$

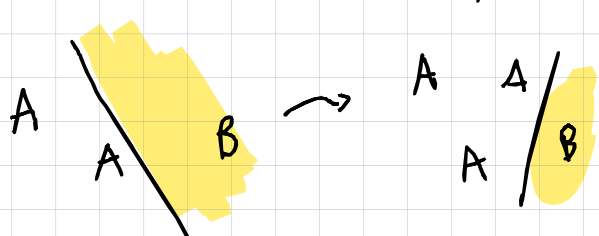
Each pair $\{\xi^\mu, S^\mu\}$ defines an HYPERPLANE in $\mathbb{R}^N$ which separates correct from incorrect

$$\text{sign}(w \cdot \xi^\mu) = S$$
$$\text{sign}(w \cdot \xi^\mu) = -S$$



TEACHER "somewhere" in VERSION SPACE

↳ Smaller Version space



So w STUDENT $\longrightarrow w^*$ Teacher

want a degree of agreement between w & w^*

- Consider random Test INPUT, drawn $P(\xi)$

$$\epsilon_g = \text{Prob}\left(\sigma_H(\xi) \neq \sigma_R(\xi)\right)$$

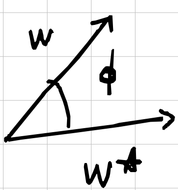
- $P(\xi)$ spherically symmetric $\langle \xi_i \rangle = 0$

$\langle \xi_j \xi_k \rangle = \delta_{jk}$

GENERALIZATION ERROR

- $$\epsilon_g = \frac{2\phi}{2\pi} = \frac{1}{\pi} \arccos\left[\frac{w \cdot w^*}{\|w\| \|w^*\|}\right]$$

$$= \begin{cases} 0 & \text{perfect } w \parallel w^* \\ 1/2 & w \perp w^* \end{cases}$$



CENTRAL LIMIT THEOREM

Consider iid random numbers x_i $\langle x_i \rangle = \mu_i$

$$Y = \sum x_i \sim \mathcal{N}(n\mu, n\sigma^2) \quad n \to \infty \qquad \langle x_i^2 \rangle - \mu^2 = \sigma_i^2$$

independently of the details of x_i

(can be extended of course)

Thermodynamic limit

HD spaces $N \to \infty$

$w, w^*, \xi \in \mathbb{R}^N$

$|w|^2 = o(N)$

$|w^*|^2 = o(N)$

Let's draw INPUT vectors $\langle \xi_j \rangle = \emptyset$

$\xi^2 = \sum \xi_j^2 = N$

$\langle \xi_i \xi_j \rangle = \delta_{jk} = \begin{cases} 1 & j=k \\ 0 & j \neq k \end{cases}$

UNCORRELATED

Ex. $P(\xi) = \mathcal{N}(0,1) \Rightarrow \quad \xi_j = \begin{cases} 1 & P = 1/2 \\ -1 & P = 1/2 \end{cases}$

$h = \dfrac{w \cdot \xi}{\sqrt{N}} = \dfrac{1}{\sqrt{N}} \sum_i w_j \xi_j$

joint normal density

$b = \dfrac{w^* \cdot \xi}{\sqrt{N}} = \dfrac{1}{\sqrt{N}} \sum_i w_j^* \xi_j$

$\langle b \rangle = \emptyset = \dfrac{1}{\sqrt{N}} \sum w_j \langle \xi_j \rangle = \emptyset$

ASSUMED FIXED

$\langle h \rangle = \emptyset$

$\boxed{}$ ORDER PARAMETERS

$\langle b^2 \rangle = \quad \langle \xi_j \xi_k \rangle = \dfrac{1}{N} \sum (w_j^*)^2 = T = 1$

$\langle h^2 \rangle = \quad = \dfrac{1}{N} \sum w_j^2 = Q$

$\langle h b \rangle = \quad = \dfrac{1}{N} \sum w_j w_k^* = R$

$$P(h,b) = \frac{1}{2\pi\sqrt{\det C}} \exp\left[-\frac{1}{2} \begin{pmatrix} h - \langle h \rangle \\ b - \langle b \rangle \end{pmatrix}^T C^{-1} \begin{pmatrix} h \\ b \end{pmatrix} \right]$$

$$C = \begin{pmatrix} Q & R \\ R & 1 \end{pmatrix}$$

$$= \frac{1}{2\pi\sqrt{Q - R^2}} \exp\left[-\frac{1}{2} \frac{h^2 + Qb^2 - 2Rhb}{Q - R^2} \right]$$

$$\epsilon_g = \mathrm{Prob}(hb < 0) = \langle \Theta[-hb] \rangle = \int_0^{\infty} dh \int_{-\infty}^{0} dh + \int_{-\infty}^{0} db \int_0^{\infty} P(h,b)$$

$$= \frac{1}{\pi} \arccos\left[R/\sqrt{Q} \right] = \frac{1}{\pi} \arccos\left[\frac{w \cdot w^*}{\|w\| \|w^*\|} \right]$$

We find the same result in HD

↳ HD space describet with same order parameter

Remarks with learning in a version space

- linear separability of the rule ??
- " " of the data ??
- Is data reliable ??

We should

→ Tollerate errors
→ aim at small # of errors
→ apply heuristic algorithms
→ SVM : nonlinear