Learning with Neural Networks

Reiner Schulz[1]

rschulz@cs.umd.edu

[1] And many others whose copyright I shamelessly infringed by including their illustrations in this presentation.

Outline

Biology: Neurons and the Brain

Computational Model of the Neuron

Logical Neuron

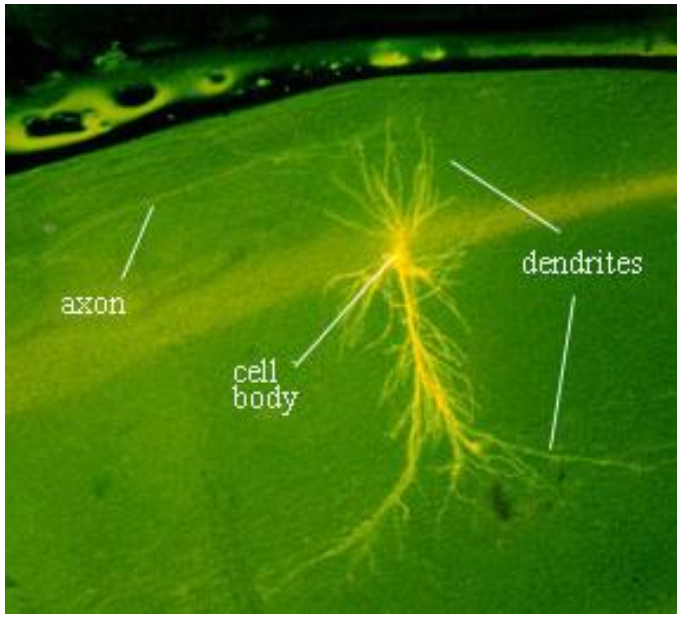
Perceptron

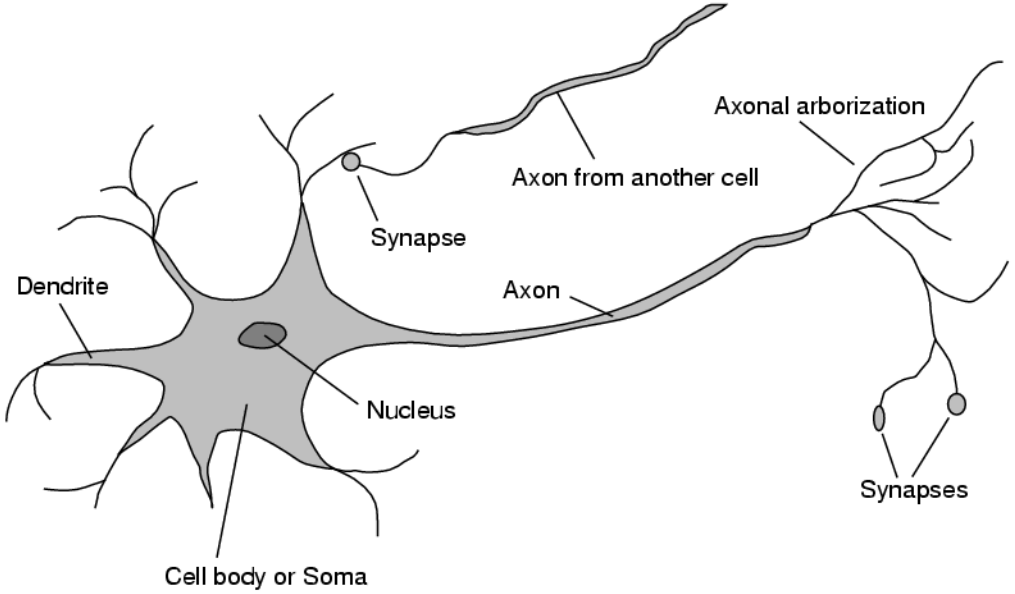
ADAptive LINear Element (ADALINE)

Multi-Layer Perceptron

Error Back-Propagation

Biological Neuron

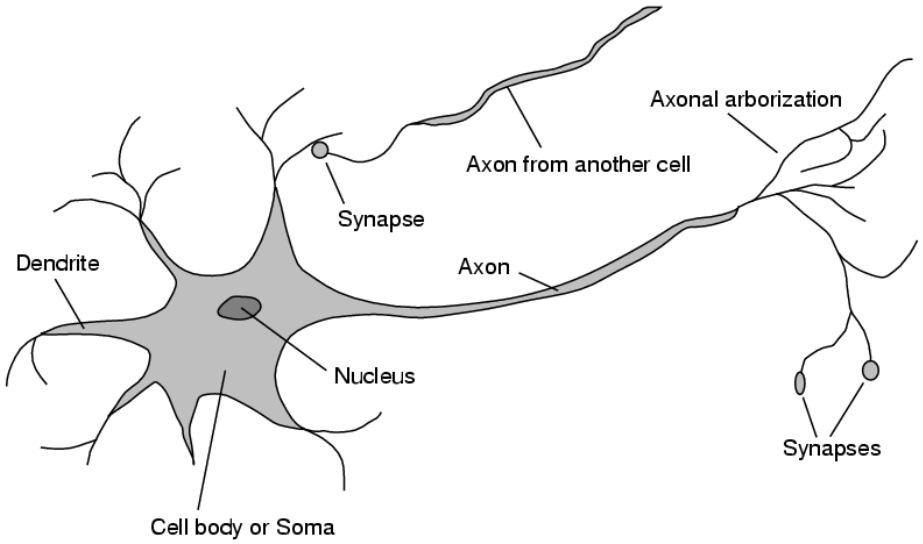


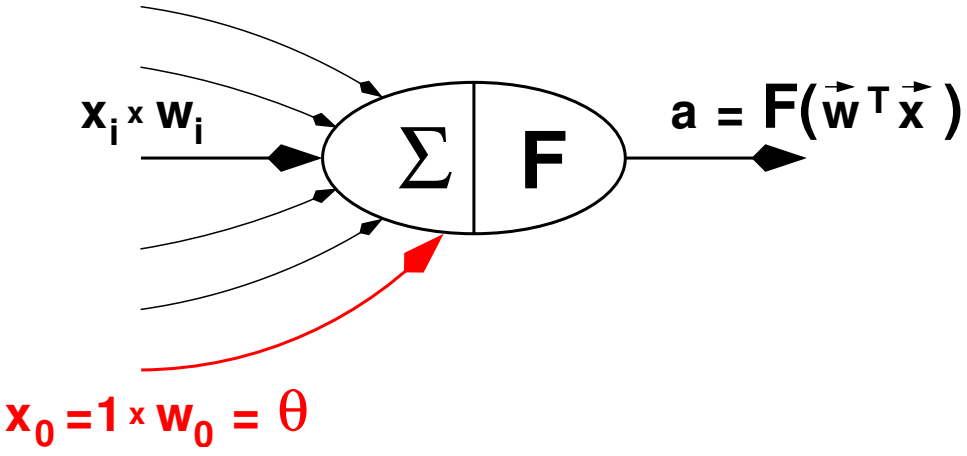


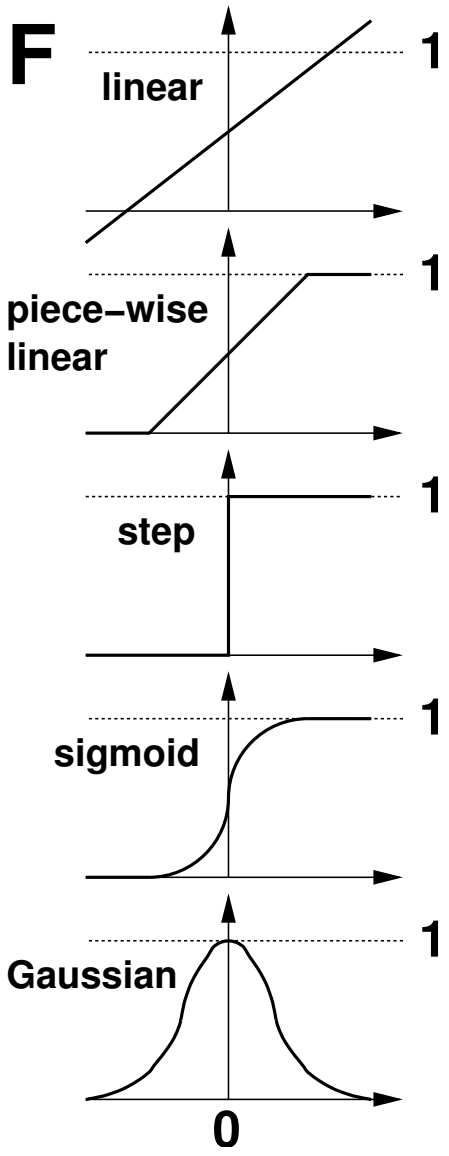
Axon $\xrightarrow{\text{AP}}$ Synapse (LEARNING: LTP or LTD) $\xrightarrow{\text{excitatory or inhibitory Neurotransmitter}}$ Dendrite $\xrightarrow{\text{EPSP or IPSP}}$ Soma $\xrightarrow{\text{AP}}$ Axon

Slow ($\approx$10ms), but MASSIVELY REDUPLICATED (10^{11} neurons, 10^{14} synapses)

Computational Neuron Model (Node)



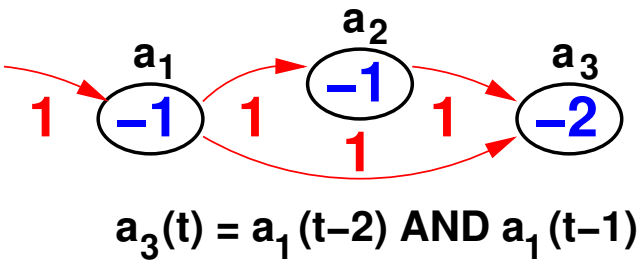


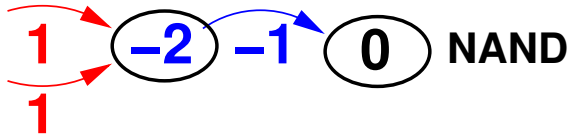


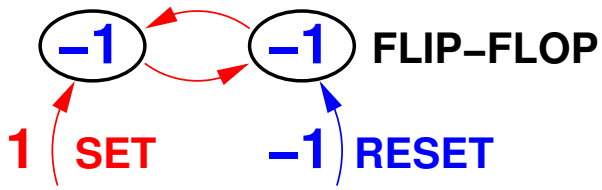
Logical Neuron

McCulloch & Pitts, 1943

- $w_{i>0} \in \{-\Phi, 1\}$ (fixed; NO learning), $w_0 = \theta \in \{0\} \cup Z^-$, F = step function
- syncronous updates: $a(t+1) = \begin{cases} 1 \text{ if } \vec{w}^T \vec{x}(t) \geq 0 \\ 0 \text{ otherwise} \end{cases}$







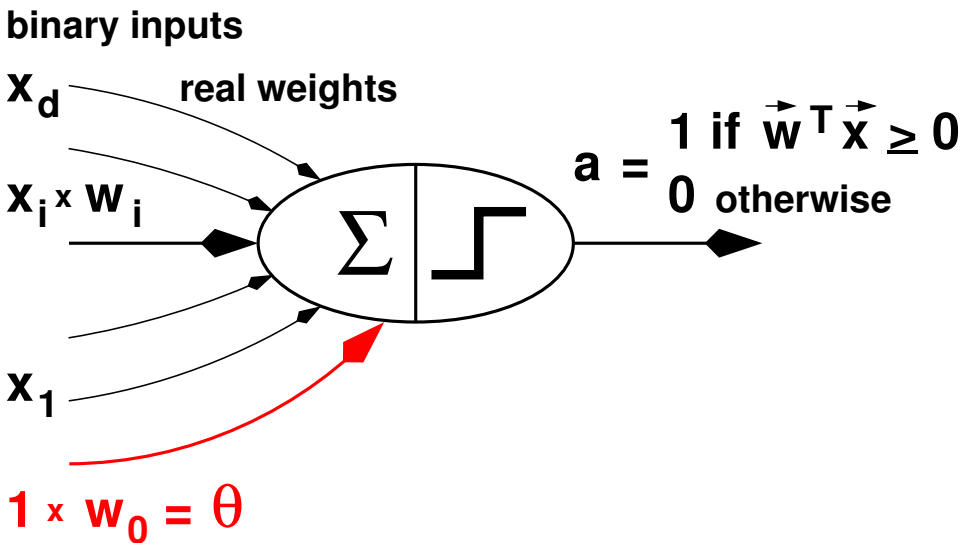
Model of the Brain?

- N nodes, $p(0) \in [0, 1]$ fraction of initially active nodes
- NO external inputs, i.e., $\vec{x} = \vec{a}$
- $n = n_e$(excitatory) $+ n_i$(inhibitory) incoming connections per node from random (other) nodes
- $n_e \gg n_i, \Phi \gg 0, \theta \approx -1 \Leftrightarrow$ limit cycles $\ldots \vec{a}(t), \vec{a}(t+1), \ldots, \vec{a}(t+k) = \vec{a}(t) \ldots$

Perceptron

Rosenblatt, 1958 vs. Minsky & Papert, 1969

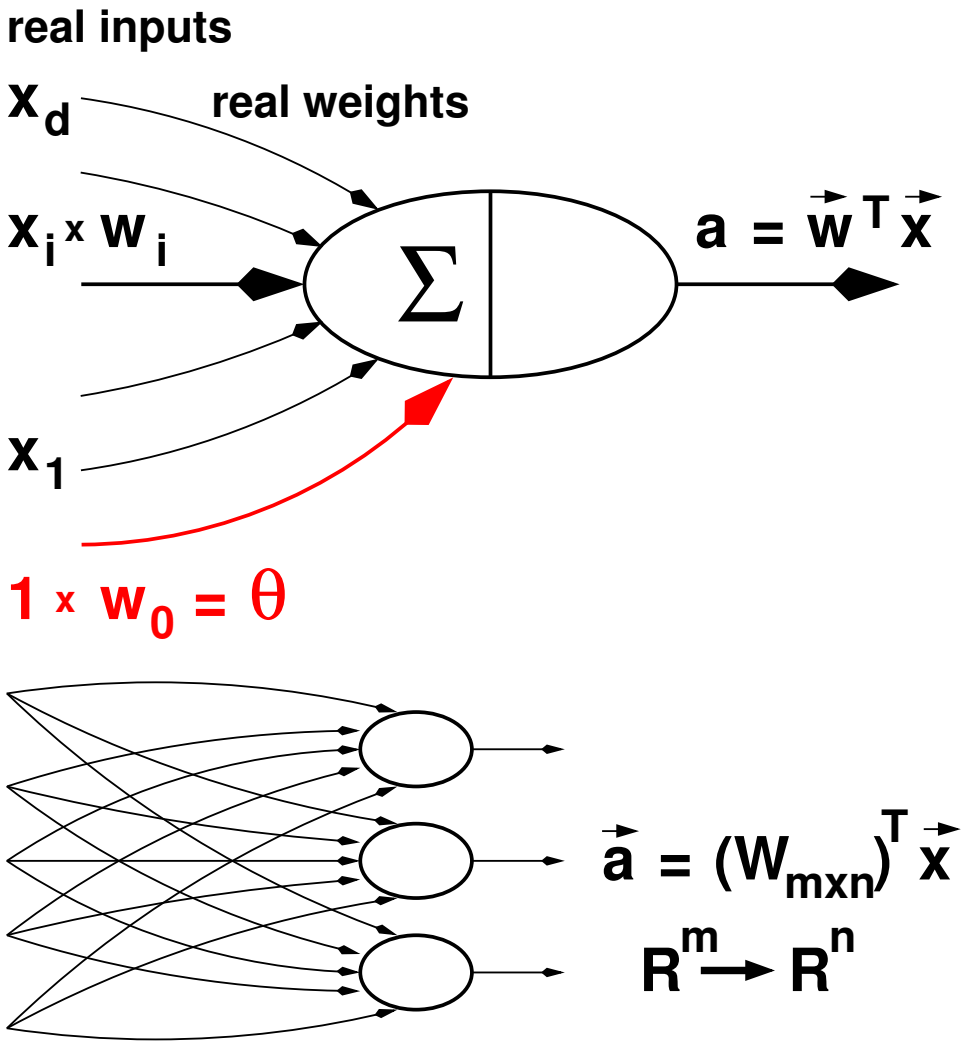
- more complex, but subsumed by the following 'modern' perceptron



- two sets of input patterns
  $X^+ = \{\vec{x}_1^+, \vec{x}_2^+, \ldots\}$,
  $X^- = \{\vec{x}_1^-, \vec{x}_2^-, \ldots\}$, $X^+ \cap X^- = \emptyset$
- GOAL: $\vec{w}$ (initially random) so that
  $$a = \begin{cases} 1 \text{ if } \vec{x} \in X^+ \\ 0 \text{ if } \vec{x} \in X^- \end{cases}$$
- Solution: $w_i = w_i + \mu x_i (t - a)$ where t (a) is the target (actual) output for $\vec{x}$ and $\mu \in]0, 1]$ is the learning rate
- guaranteed to find a 'correct' $\vec{w}$, provided X^+ and X^- are linearly separable
- very rare for $|X| = |X^+ \cup X^-| \gg d + 1$: $P(|X|, d) = \frac{1}{2^{|X|-1}} \sum_{i=0}^{d} \binom{|X| - 1}{i}$
- $P(|X| = 4, d = 2) = 14/16$ (XOR and EQU not linearly separable)

ADAptive LINear Element (ADALINE)

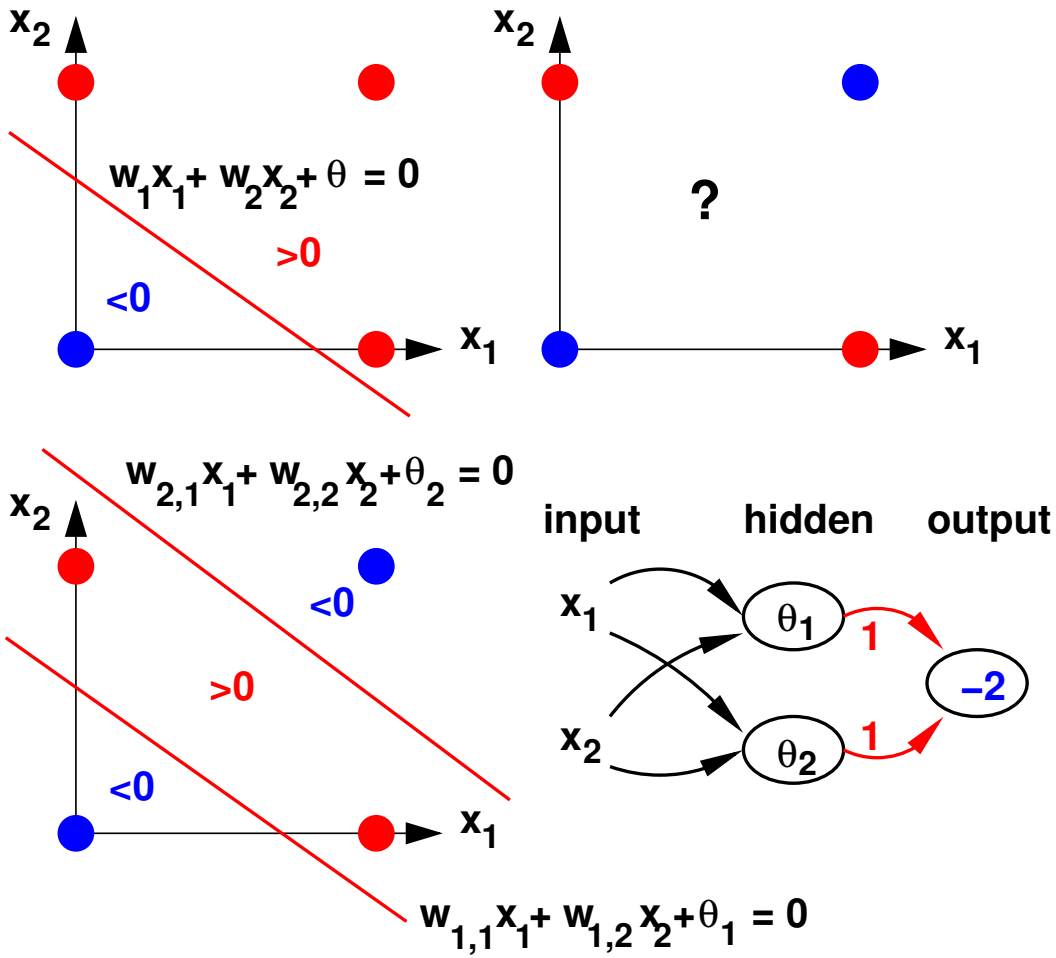
Widrow & Hoff, 1960



- paired input and target output patterns $(\vec{x} \in \mathcal{R}^m, \vec{t} \in \mathcal{R}^n)$
- GOAL: weight matrix W so that $\vec{a} = W^T \vec{x} = \vec{t}$ for each pair $(\vec{x}, \vec{t})$
- Solution: Gradient Descent
- Error: $E = \frac{1}{2} \sum_l (t_l - a_l)^2$
- Gradient: $\frac{\delta E}{\delta w_{i,j}} = (t_i - a_i)(-\frac{\delta a_i}{\delta w_{i,j}})$
- $\frac{\delta a_i}{\delta w_{i,j}} = \frac{\delta \vec{w}_i^T \vec{x}}{\delta w_{i,j}} = \frac{\delta \sum_l w_{i,l} x_l}{\delta w_{i,j}} = x_j$
- Least Mean Square (LMS) or Δ-rule: $w_{i,j} = w_{i,j} - \mu \frac{\delta E}{\delta w_{i,j}} = \ldots$ $\ldots = w_{i,j} + \mu \Delta_i x_j$ where $\Delta_i = t_i - a_i$
- finds (global) minimum of E

Multi-Layer Perceptron

Rosenblatt, 1958 (original Perceptron had multiple layers)



- input $\rightarrow$ hidden $\rightarrow$ output
- computes the function $\vec{a}_O = F_O(W^T F_H(V^T \vec{x}))$ where V (W) contains the incoming weights of the hidden (output) nodes
- used for classification and function approximation
- typically, $\mathrm{F}(x) = \sigma(x) = \frac{1}{1+e^{-sx}}$ (sigmoid function)
- if trained via Gradient Descend, F needs to have a derivative F'
- $\sigma'(x) = s\sigma(x)(1-\sigma(x))$

Error Back-Propagation

Rumelhart, 1986

- generalization of the LMS or δ-rule for Multi-Layer Perceptrons
- Error: $E = \frac{1}{2}\sum_l (t_l - a_l)^2$
- Gradient of E with respect to weight $w_{i,j}$ from j^{th} hidden to i^{th} output node:
  - $\frac{\delta E}{\delta w_{i,j}} = (t_i - a_i)(-\frac{\delta a_i}{\delta w_{i,j}}) = -\frac{\delta F(\vec{w}_i^T \vec{a}_H)}{\delta w_{i,j}}(t_i - a_i) = -F'(\vec{w}_i^T \vec{a}_H) a_{H,j}(t_i - a_i)$
    so that $w_{i,j} = w_{i,j} + \mu \Delta_i a_{H,j}$ *with* $\Delta_i = F'(\vec{w}_i^T \vec{a}_H)(t_i - a_i)$
- Gradient of E with respect to weight $v_{i,j}$ from j^{th} input to i^{th} hidden node:
  - $\frac{\delta E}{\delta v_{i,j}} = \sum_l (t_l - a_l)(-\frac{\delta a_l}{\delta v_{i,j}}) = -\sum_l \frac{\delta F(\vec{w}_l^T \vec{a}_H)}{\delta v_{i,j}}(t_l - a_l) =$
    $-\sum_l F'(\vec{w}_l^T \vec{a}_H)\frac{\delta \vec{w}_l^T \vec{a}_H}{\delta v_{i,j}}(t_l - a_l) = -\sum_l \Delta_l \frac{\delta \sum_k w_{l,k} a_{H,k}}{\delta v_{i,j}} = -\sum_l \Delta_l w_{l,i}\frac{\delta a_{H,i}}{\delta v_{i,j}} =$
    $-\sum_l \Delta_l w_{l,i}\frac{\delta F(\vec{v}_i^T \vec{x})}{\delta v_{i,j}} = -\sum_l \Delta_l w_{l,i} F'(\vec{v}_i^T \vec{x})\frac{\delta \vec{v}_i^T \vec{x}}{\delta v_{i,j}} = -\sum_l \Delta_l w_{l,i} F'(\vec{v}_i^T \vec{x}) x_j =$
    $-\Delta_i x_j$ *where* $\Delta_i = F'(\vec{v}_i^T \vec{x}) \sum_l w_{l,i}\Delta_l$ *so that* $v_{i,j} = v_{i,j} + \mu \Delta_i x_j$
- NOT guaranteed to find the global minimum of E (is NOT monotonously decreasing due to non-linearities)

Error Back-Propagation

Power

- with enough hidden units, can approximate any arbitrary function as precisely as desired

Shortcomings

- premature convergence on local minima
- very very slow
- representational 'mysticism' (what was learned?)
- many parameters to chose (how many hidden units? μ? F_H? F_O?)
- overfitting ('memorized' training examples; performs poorly on unseen inputs)

'Better' Heuristics

- EBP with Momentum: $\Delta^{(n)} w_{i,j} = \mu \Delta_i^{(n)} a_{H,j}^{(n)} + \alpha \Delta^{(n-1)} w_{i,j}$
- Resilient EBP (RPROP): $\Delta w_{i,j} = -\Delta_{i,j} \text{sign}(\frac{\delta E}{\delta w_{i,j}})$ where $\Delta_{i,j}$ ($w_{i,j}$'s learning rate) increases (decreases) if sign$(\frac{\delta E}{\delta w_{i,j}})$ remains the same (changes) from one time step to the next

References

- McCulloch, W. & Pitts, W. A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics: 5, pp. 115-137, 1943.

- Rosenblatt, F. The Perceptron: a probabilistic model for information storage & organization in the brain. Psychological Review: 65, pp. 386-408, 1958.

- Minsky, M. & Papert, S. Perceptrons. MIT Press, Cambridge, 1969.

- Widrow, B. & Hoff, M. E. Adaptive switching circuits. IRE WESCON Convention Record: 4, pp. 96-104, 1960.

- Rumelhart D. et al. Learning Representations by Back-Propagating Errors. Nature: 329, 533-536, 1986.

- Riedmiller, M. & Braun, H. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. Proceedings of the ICNN, San Francisco, pp. 123-134, 1993.

- Sejnowski, T. J. & Rosenberg C. R. NETtalk: A parallel network that learns to read aloud. Technical Report JHU/EECS-86/1, EE and CS Dept., John Hopkins University, 1986.

- Pomerleau, D. A. ALVINN: an autonomous land vehicle in a neural network. Technical Report CMU-CS-89-107, CS Dept., Carnegie Mellon University, 1989.