Artificial Neural Networks

Machine Learning for Data Science 1

CENTRAL IDEA:

- extract derived features as linear combinations of inputs
- model the internal target features through non-linear transformations
- combine and ~~repeat~~

1, >2 ≡ deep model

1943 : models of the true NN

1940s : plasticity ≡ learning

1953 : perceptron ; failed idea, looked very promising

1965 : Idea of many layers

1973 : backpropagation ||

1975 : →

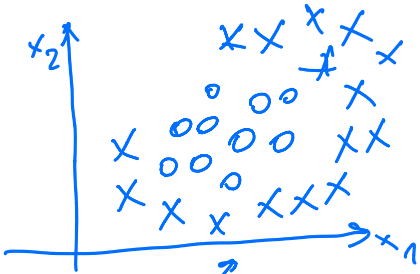
1989 : LeCun : hand-written digits } deep networks
convolution

1992 : max pooling , 3D

2006 : Hinton ; Boltzmann NN

2009 - 2012 : ANN major competition

2012 : deep learning , images , text , ... more data
comp. power



$x_1 \cdot x_2$

MOTIVATION

- can we learn hard concepts?
- Interactions

1000 features

→ $\frac{1000 \cdot 999}{2}$ two-interactions

$\frac{1000 \cdot 999 \cdot 998}{3 \cdot 2}$ three interactions

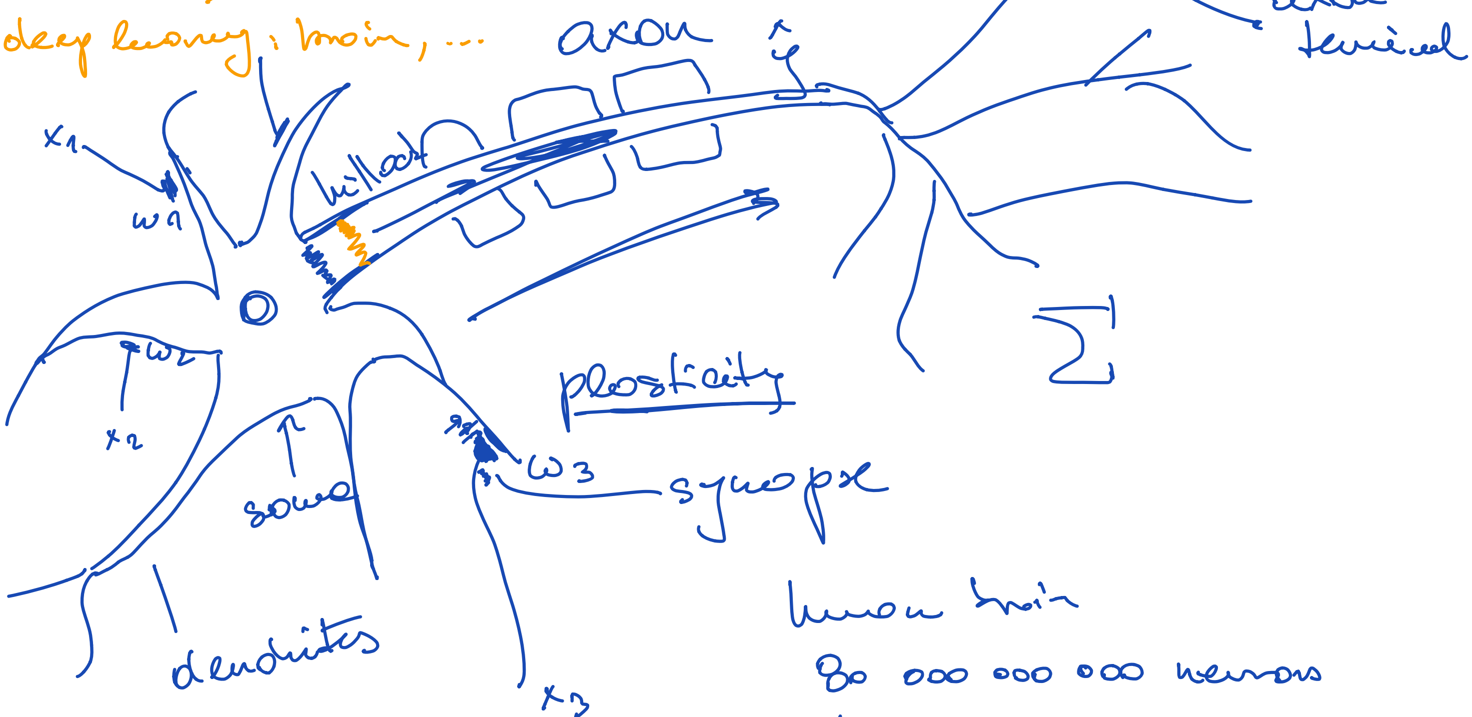
a modelling framework that incorporates all these interactions

potentially

complexity
needs a lot of data
overfitting
explanation

Biology – motivation

- important at the start
- similarity was lost
- deep learning: brain, ...



human brain

80 000 000 000 neurons

10 000 synapses

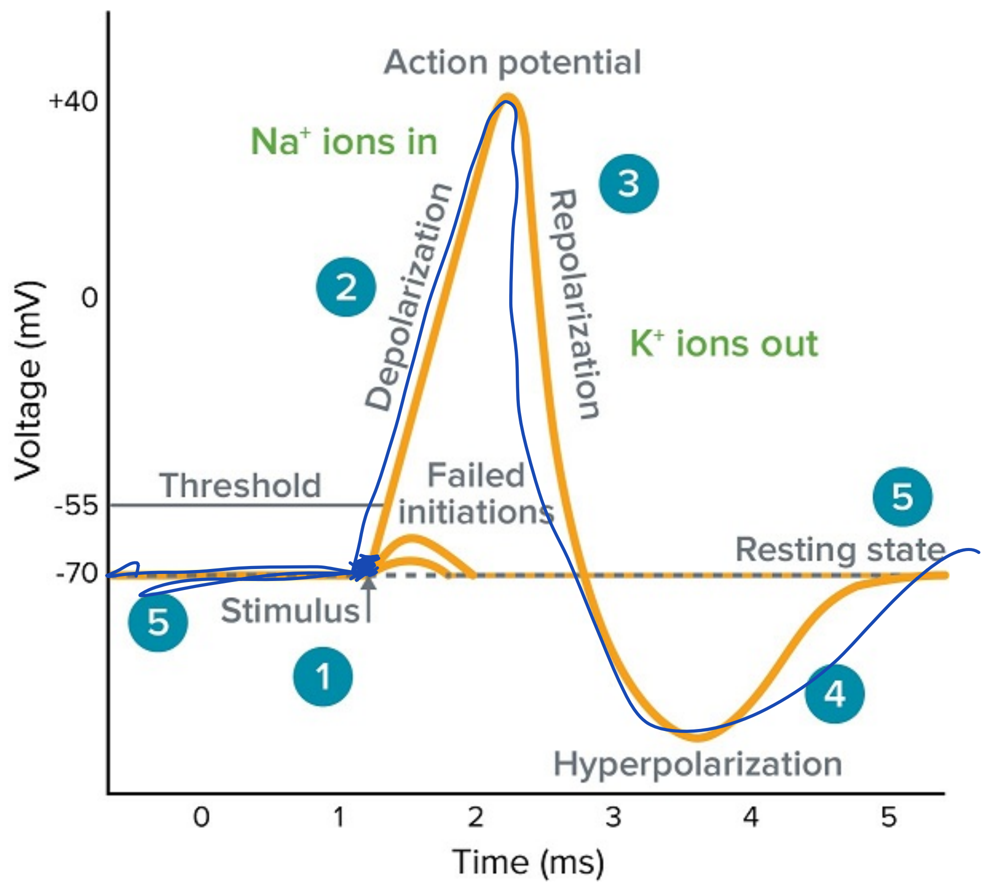
10^{15} connections

synapses are slow

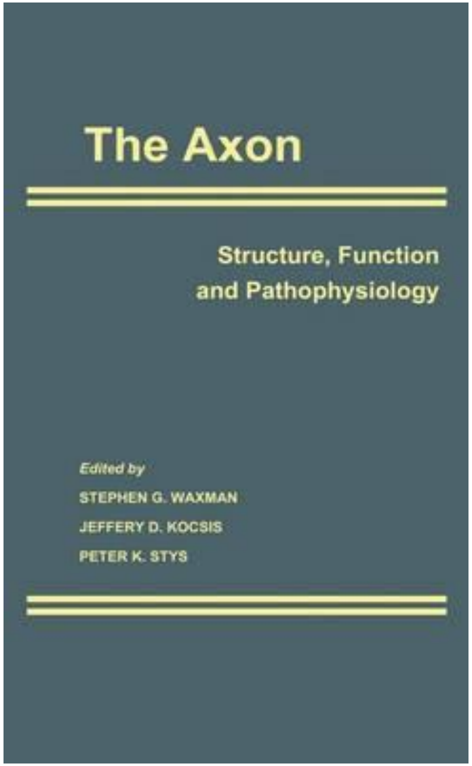
everything is ||

Modularity of the Brain

- different regions = different tasks
- anatomically similar
- damage in one region
  $\Rightarrow$ a different region can take over



Action potential
+40
Na+ ions in
Depolarization
Repolarization
K+ ions out
0
Voltage (mV)
Threshold
-55
Failed initiations
-70
Resting state
Stimulus
Hyperpolarization
1
2
3
4
5
5
0
1
2
3
4
5
Time (ms)



1 Electrical activity of nerve: The background up to 1952

SIR ANDREW HUXLEY

My interest in physiology, and in the physiology of nerve in particular, dates from the autumn of 1935, when I went up to Trinity College, Cambridge, as an undergraduate. I was expecting to specialize in physics, in which I had been very well taught at school, but the rules of the Natural Sciences Tripos ("tripos" is a Cambridge word for courses leading to a first degree) required me to take a third experimental science as well as the physics, chemistry, and mathematics that were the obvious choices. I picked physiology on the advice of a friend a few years older than myself who told me that it was a lively subject in which even the initial courses included material recently discovered or even still controversial, unlike the courses in physics, which included nothing that had not been cut and dried for decades. I was inspired to switch to physiology as my final-year specialty subject largely by my teachers W. A. H. (William) Rushton and F. J. W. (Jack) Roughton and by personal contacts with Glenn Millikan (son of R. A. Millikan of the oil-drop experiment; too little known on account of his death in 1946 in a climbing accident) and Alan Hodgkin, all Fellows of Trinity College working in the physiology laboratory. E. D. Adrian (later Lord Adrian, Master of Trinity College and President of the Royal Society) was also a Fellow of Trinity College, but I hardly came across him until my final undergraduate year because he was a research professor of the Royal Society, taking little part in undergraduate teaching, until 1937; in that year he became head of the Department and in my final year he lectured to us on the central nervous system.

I hope that my account of the ideas then current about nerve conduction and of developments up to 1952 is not too heavily biased by my Cambridge background.

EXCITATION OF NERVE

Our first-year lectures on nerve were given by William Rushton. We were, of course, taught the elementary facts about excitation of nerve, mostly established in the mid-19th century in Germany by experiments on the sciatic nerve of the frog with the gastrocnemius muscle attached to indicate by its contraction whether the motor nerve fibers had been activated: the impulse arises at the point where a stimulating current of short duration leaves the nerve (the cathode) and it travels in both directions. If a direct current of fairly long duration is used, an impulse may be set up both at the cathode at the start of the current and at the anode when the current is terminated (anode break excitation). There is no response if the strength of the stimulus is below a well-defined critical level (the "threshold"), and the response increases with stimulus strength up to a maximum; the impulse is accompanied by a wave of electrical negativity passing along the surface of the nerve. A second stimulus is ineffective if it follows a maximal stimulus within a certain time interval (the "absolute refractory period," roughly equal to the duration of the propagated electric change), and this is followed by a "relative refractory period" in which the threshold is higher than when the nerve is fully rested. The threshold value of current strength varies inversely with its duration, the product of these quantities approaching a finite limit as the duration is reduced toward zero.

THE ALL-OR-NONE "LAW"

At the turn of the century, it had been debated whether the gradation of response with strength of stimulus was solely a matter of the number of fibers within the nerve trunk being activated, or whether the impulse in an individual fiber could vary with the strength of the stimulus. The former alternative was found to be correct: the invariant "all-or-none" character of the propagated response of individual motor nerve fibers, and of individual fibers of skeletal muscle, was well established in the first decade of this century by Keith Lucas (1905, 1909) (another Fellow of Trinity College, and, like Millikan, too little known on account of his early death in a flying accident during World War I), using the twitch of a muscle fiber or a motor unit as the indication of activity. It was recognized that the energy dissipated by

Format: Hardback | 708 pages

Dimensions: 225 x 289 x 39mm | 2,278g

Publication date: 30 Mar 1995

Publisher: Oxford University Press Inc

Publication City/Country: New York, United States

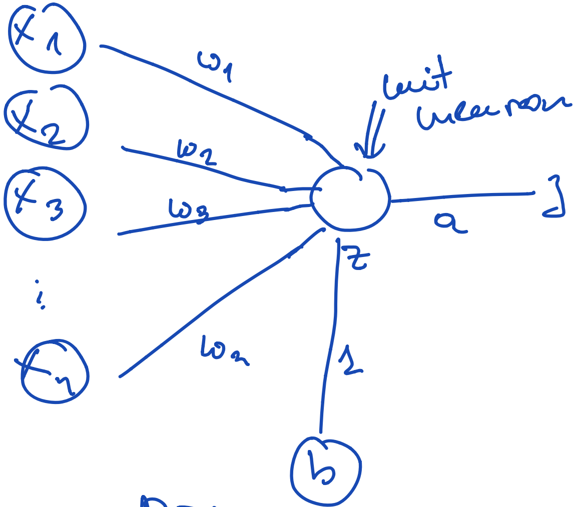
Language: English

Edition Statement: New

Illustrations note: halftones, line figures and tables

ISBN10: 0195082931

ISBN13: 9780195082937

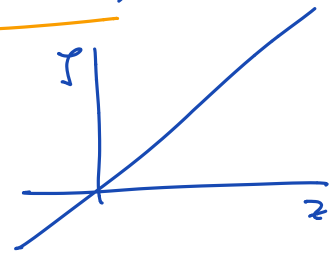


Linear neuron

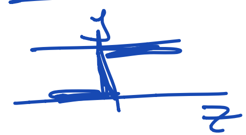
$$z = \sum w_i x_i + b$$

$$y = \sigma(z)$$

$$\sigma(z) = z$$

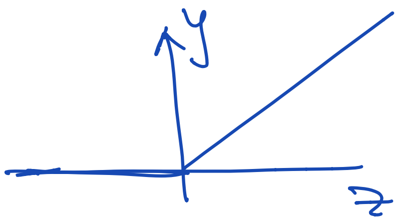


Binary Unit

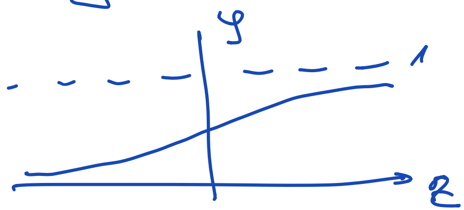


RELU

Rectified Linear



Sigmoid neurons



$$y = \frac{1}{1 + e^{-z}}$$

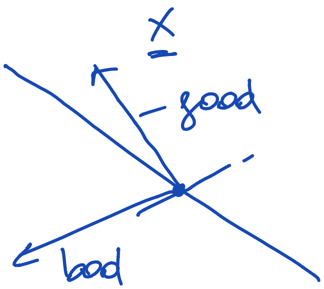
Perceptrons

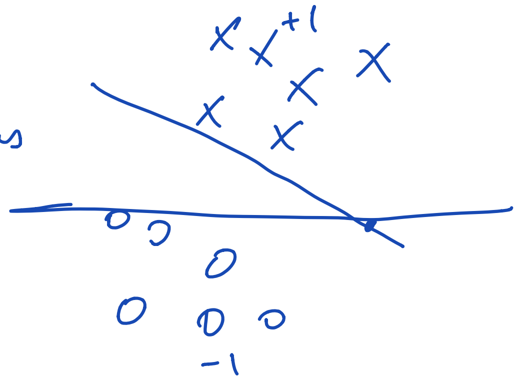
1960s : Frank Rosenblatt

how to learn weight vectors

weight-space

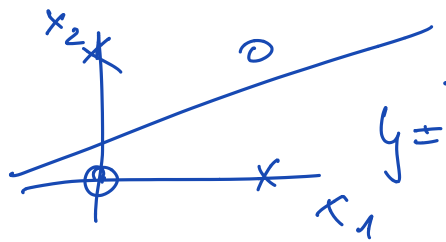
$z = w^T x$





iterations

	x_1	x_2	y
1	0	0	0
0	0	1	1
0	1	0	1
1	1	1	0



$y = x_1 \equiv x_2$

initialize w

<u>repeat</u>

randomly choose x from training set

if $h(z) \neq y$:

if $h(z) = 0$:

$w \leftarrow w + x$

else:

$w \leftarrow w - x$

Some conventions

- units of ANN, activation $\in 0..1$
- output of ANN, real values $\Rightarrow$ regression

  probabilities

cat

mouse

dog

$\vdots$

K

$\Sigma = 1$

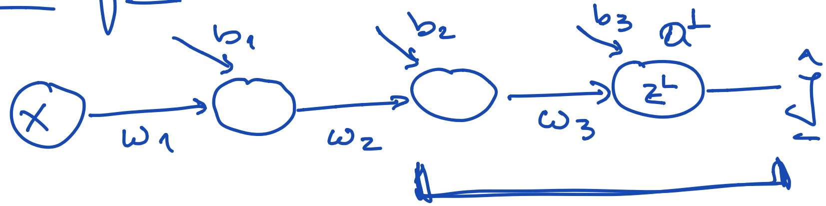
softmax

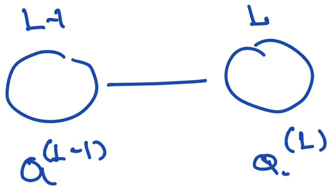
$$\hat{y}_i(z) = \frac{e^{z_i}}{\sum_{i=1}^{K} e^{z_i}}$$

$K = 2 \equiv$ logistic regression

$K > 2$

example





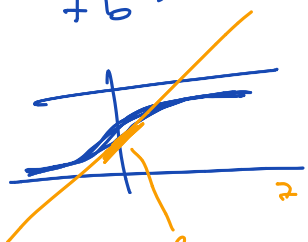
$$a^{(L)} = \sigma(z^{(L)})$$

$$z^{(L)} = \sum w_i^{(L)} x_i^{(L)} + b^{(L)}$$

$$J(w_1, b_1, \dots w_3, b_3) = (a^{(L)} - y)^2$$

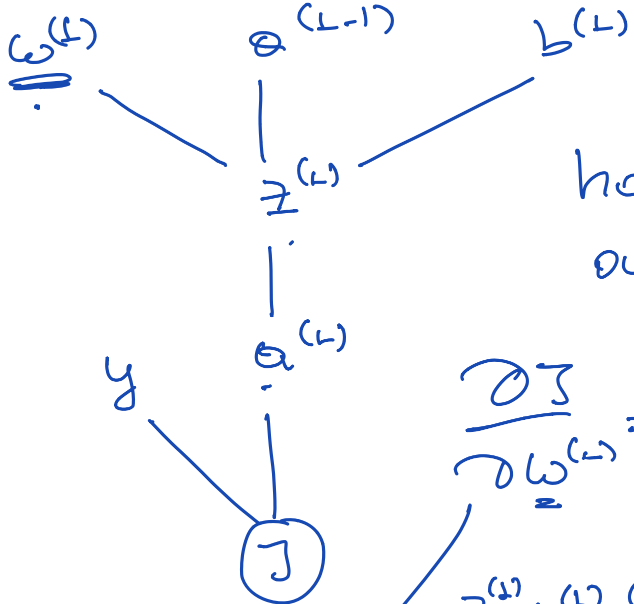
$$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$$

$$a^{(L)} = \sigma(z^{(L)})$$



$$\frac{\partial g(z)}{\partial z} = g(z)(1-g(z))$$

linear part of the sigmoid



how dos J depend on $w^{(L)}$

Σ

$J=(a^{(L)}-y)^2$

$$\frac{\partial J}{\partial w^{(L)}} = \frac{\partial z^{(L)}}{\partial w^{(L)}} \cdot \frac{\partial a^{(L)}}{\partial z^{(L)}} \cdot \frac{\partial J}{\partial a^{(L)}}$$

$z^{(L)} = w^{(L)} a^{(L-1)} + b^{(L)}$

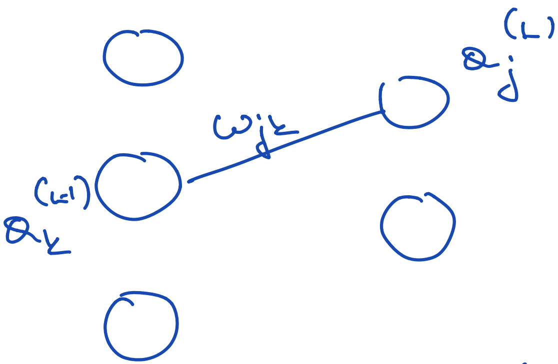
$a^{(L)}$

$\sigma(z)(1-\sigma(z))$

$\underbrace{2(a^{(L)}-y)}_{\text{error}}$

$$\frac{\partial J}{\partial w^{(L)}} = a^{(L-1)} \; \sigma(z)(1-\sigma(z)) \cdot 2(a^{(L)}-y)$$

$$J = \sum_{j=0}^{n} \left(a_j^{(L)} - y_j\right)^2$$



$$z_j^{(L)} = \sum w_{ji}^{(L)} a_i^{(L-1)} + b^{(L)}$$

$$a_j^{(L)} = \sigma\left(z_j^{(L)}\right)$$

$$J = \sum_i \left(a_j^{(L)} - y_j\right)^2$$

$$\frac{\partial J}{\partial w_{jk}^{(L)}} = \sum \frac{\partial z_j^{(L)}}{\partial w_{jk}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial J}{\partial a_j^{(L)}}$$

L-1

$$\frac{\partial J}{\partial a_k^{(L-1)}} = \sum \frac{\partial z^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial J}{\partial a_j^{(L)}}$$

$$\frac{\partial J}{\partial a_k^{(L-2)}} \quad \dots \quad \int\left(\frac{\partial J}{\partial a_k^{(L-1)}}\right) \dots$$

$$X = \underset{m}{\left[\, \equiv \,\right]}^{n_1} \quad , \quad X' = \left[\, \underset{1}{\mid} \;\; \equiv \,\right]_{m}^{n_1+1}$$

$$A^{(1)} = X'$$

$$\underset{m \times n_2}{Z^{(2)}} = \underset{m \times n_1}{A^{(1)}} \; \underset{n_1 \times n_2}{W^{(2)}}$$

$$A^{(2)} = \sigma\left(Z^{(2)}\right)$$

$$+ \frac{\alpha}{m} \sum_{l} \sum_{ij} W_{ij}^{(l)}$$

$$A^{(l)} = \sigma\left(A^{(l-1)} W^{(l)}\right)$$

$$\frac{\partial J}{\partial W^{(L)}} = \frac{\partial z^{(L)}}{\partial W^{(L)}} \cdot \underbrace{\frac{\partial a^{(L)}}{\partial z^{(L)}} \cdot \frac{\partial J}{\partial a^{(L)}}} \cdots \frac{\partial J}{\partial a^{(L-1)}} \cdot d\left(\frac{\partial J}{\partial a^{(L)}}\right)$$

$$\underset{m \times n_L}{d^{(L)}} = A^{(L)}\left(1 - A^{(L)}\right) \circ \left(A^{(L)} - \vec{Y}\right)$$

$$D^{(L)} = \frac{1}{m}\left(A^{(L-1)}\right)^{T} \times d^{(L)}$$

Learning

optimization : (stochastic) gradient descent

$\downarrow$

mini-batch , 250
adaptive learning rate
momentum

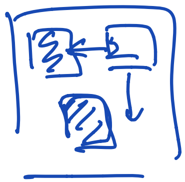
generalization : tricks

- regularization
  "weight-decay"
  0.99
- weight sharing
- early stopping | $\rightarrow$ validation set
- ensembling
- drop-out
= pre-training | transfer learning

deep learning –

hidden layers ≥ 2

– convolutional NN



max pooling

– recurrent NN

– long short-term memory units

– transfer learning

– encoders