Page 1

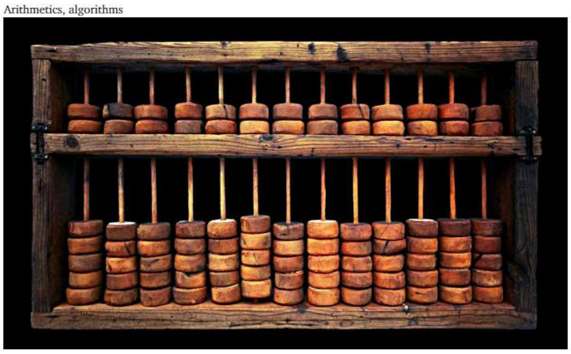
ML1 – Introduction

Stefano Rovetta

A.y. 2023-2024

Page 2

Arithmetics, algorithms



Page 3

Logic and logical reasoning

Aristotle

Page 4

Automata and programmable machines

A writing automaton and the Jaquard programmable loom

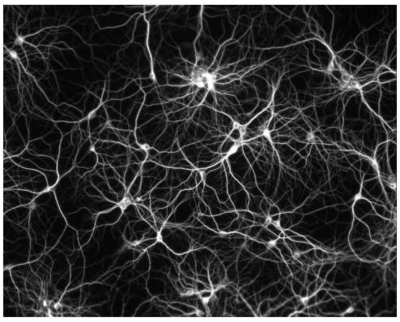
Page 5

Turing and the science of computation

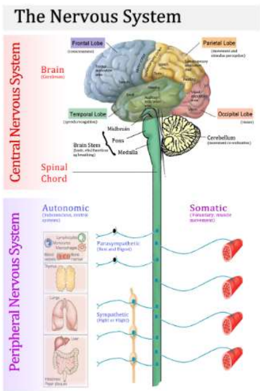
Alan M. Turing (1912 - 1954)

Page 6

What's in our brain?



Page 7



Page 8

More on the brain later

Page 9

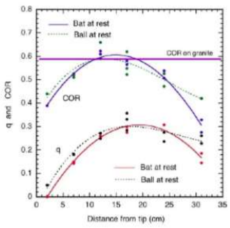
Computation meets the brain

Simulate the **final result** of brain processing → Artificial **intelligence**

Simulate the **inner mechanics** of brain processing → Artificial **neural networks**

Modern machine learning was mostly developed for neural networks

Page 10



Page 11

Page 12

For these problems we do have algorithms

- sorting → insert sort, bubble sort, Shell sort, radix sort, heapsort, bogosort…
- spectral analysis of periodic signals → FFT
- database filtering → SQL SELECT

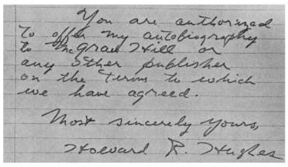
Page 13

We have no algorithms for…

Recognizing a face

Page 14

We have no algorithms for…

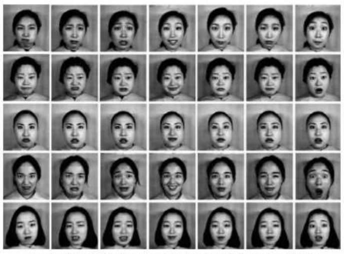


You are authorized
to offer my autobiography
to McGraw Hill or
any other publisher
on the terms to which
we have agreed.
Most sincerely yours,
Howard R. Hughes

Distinguishing between genuine works by a given author and fakes

Page 15

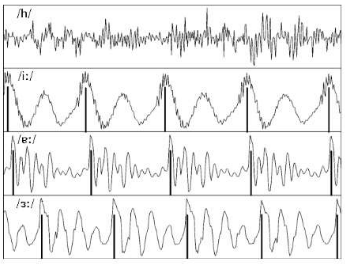
We have no algorithms for…



Recognizing emotions from gestures or facial expressions

Page 16

We have no algorithms for…



Understanding speech

Page 17

We have no algorithms for...

Controlling a soft robotic tentacle

Page 18

We have no algorithms for...

Build a complete self-driving car

Page 19

To sum up

Many interesting problems are too complex to admit an algorithmic solution or even a complete description

For these problems, only **data** are available

Nowadays we have **LOTS OF DATA FROM LOTS OF SOURCES**

Machine learning is about using data to solve problems

Page 20

Perceptual tasks

Perceptual tasks are tasks related to perception.
They have a typical structure, based on sets of individual measurements.
It is generally difficult to write a program (an algorithm) to solve a perceptual task
LEARNING FROM DATA

Page 21

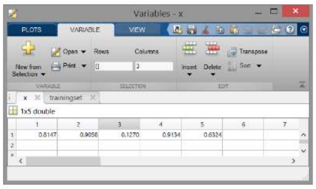
Representing perceptual tasks

NAMES:

- sensors
- inputs, variables; also features
- patterns (vectors)
- experimental observation, example, instance; sometimes "sample" (cfr. statistics)
- data set
  - training set
  - validation set
  - test set
- if data are **vectors** or d-dimensional points, then a data set of n observations is a $n \times d$ **matrix**

Page 22

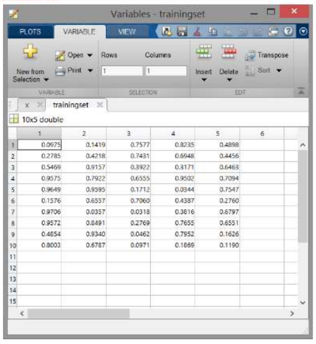
A pattern as a vector



(Matlab screenshot)

Page 23

A training set as a matrix



(Matlab screenshot)

Page 24

Operations on vectors

Two main operations are defined:

1. Vector sum: $\mathbf{u} \in \mathscr{V}$ and $\mathbf{v} \in \mathscr{V} \quad \Rightarrow \quad \mathbf{u}+\mathbf{v} \in \mathscr{V}$
2. Multiplication by a scalar: $\mathbf{v} \in \mathscr{V}$ and $a \in \mathscr{F} \quad \Rightarrow \quad a\mathbf{v} \in \mathscr{V}$

On real vectors ($\mathbf{u} \in \mathbb{R}^d$):

1. Vector sum:
   $\mathbf{u} = [u_1, u_2, \dots, u_d]$ and $\mathbf{v} = [v_1, v_2, \dots, v_d] \quad \Rightarrow \quad \mathbf{u}+\mathbf{v} = [u_1+v_1, u_2+v_2, \dots, u_d+v_d]$
2. Multiplication by a scalar:
   $\mathbf{v} \in \mathscr{V}$ and $a \in \mathscr{F} \quad \Rightarrow \quad a\mathbf{v} = [av_1, av_2, \dots, av_d]$

Page 25

Other operations are possible on real vectors.

- Scalar (inner, dot-) product between two vectors: outputs a scalar and is defined as:

$$\mathbf{u} \cdot \mathbf{v} = \sum_i u_i v_i$$

- (Euclidean) norm of a vector:

$$||\mathbf{u}|| = \sqrt{\sum_i u_i^2}$$

- (Euclidean) distance between two vectors:

$$d_E(\mathbf{u}, \mathbf{v}) = ||\mathbf{u} - \mathbf{v}|| = \sqrt{\sum_i (u_i - v_i)^2}$$

Page 26

Note that:

1. $||\mathbf{u}|| = \sqrt{\sum_i (u_i u_i)} = \sqrt{\mathbf{u} \cdot \mathbf{u}}$
2. $\mathbf{u} \cdot \mathbf{v} = ||\mathbf{u}|| \, ||\mathbf{v}|| cos\alpha$
   where α is the angle between $\mathbf{u}$ and $\mathbf{v}$.
3. Therefore $\mathbf{u} \cdot \mathbf{v} = 0$ for orthogonal vectors ($cos\alpha = 0$).
4. If $||\mathbf{u}|| = 1$ and $||\mathbf{v}|| = 1$, then $||\mathbf{u} - \mathbf{v}||^2 = 2 - 2\mathbf{v} \cdot \mathbf{u}$

Page 27

- $\mathbf{u}$ is a linear combination of vectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots$
  when $\mathbf{u} = \sum_i a_i \mathbf{v}_i$
- $\mathbf{u}$ is a convex combination of $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots$
  when
  1. $\mathbf{u} = \sum_i a_i \mathbf{v}_i$ (a linear combination)
  2. $\sum_i a_i = 1$ and $a_i > 0 \; \forall i$

Page 28

Typical perceptual problems

- Mapping a stimulus to a response:
  supervised learning
- Describing the data:
  unsupervised learning

Page 29

- In **supervised tasks**
  the data contain both input patterns and output values

- In **unsupervised tasks**
  the data contain just input patterns

Page 30

Typical perceptual problems

- Mapping a stimulus to a response (from input to output):
  - Classification
  - Regression

- Describing the data (from input to a more compact representation of the input itself):
  - Clustering
  - Mapping in lower dimensionality

Page 31

Types of quantity we want to learn

- **Real values** (one or more)

- **Categorical values**

Page 32

examples of categorical information:

- colour = {red, green, blue, cyan, magenta, yellow, black}
- name = {Socrates, Plato}
- truth value = {true, false}

No natural ordering, only qualitative information

Page 33

		type of output	
		quantitative	nominal
supervised	YES	REGRESSION	CLASSIFICATION
	NO	LOW-DIMENSIONAL MAPPING	CLUSTERING

Page 34

More detailed examples of perceptual problems

- A wall-following robot has to make decisions as to the direction to take, depending on a circular array of ultrasound sensors
- The robot has 24 such sensors evenly spread over 360 degrees.
- Possible directions are: Sharp-Left-Turn | Slight-Left-Turn | Move-forward | Slight-Right-Turn | Sharp-Right-Turn

Page 35

The scitos G5 robot is a multipurpose, modular platform for robotic research and development.

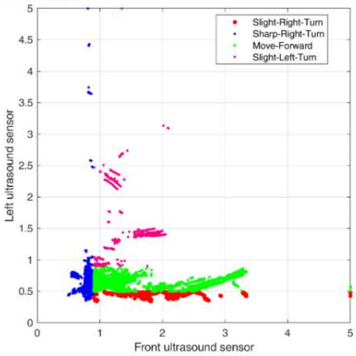
https://www.metralabs.com/en/mobile-robot-scitos-g5/

Page 36

The ultrasonic sensor's output is available as a voltage in the range $0 \ldots 5V$

Page 37

Minimum readings from two groups of sensors, on the forward and on the left.
Colors correspond to directions to take.



Page 38

Iris (flower) recognition

- The Iris dataset has been in use since 1936
- Collected by botanist Edgar Anderson in 1935
- Used by statistician Sir Ronald A. Fisher in 1936

Sources:

Edgar Anderson (1935). "The irises of the Gaspe Peninsula". Bulletin of the American Iris Society *59: 25.*

Fisher, R.A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". Annals of Eugenics *7: 179-188.*

Download it from the University of California - Irvine repository at:
`http://archive.ics.uci.edu/ml/datasets/Iris`

Page 39

Iris Setosa

Iris Virginica

Iris Versicolor

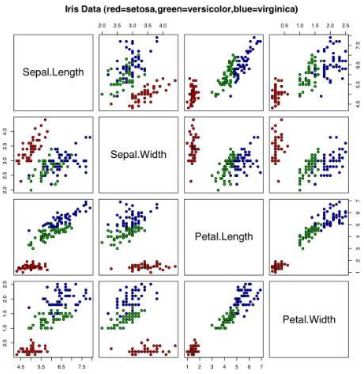
Page 40

180 MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

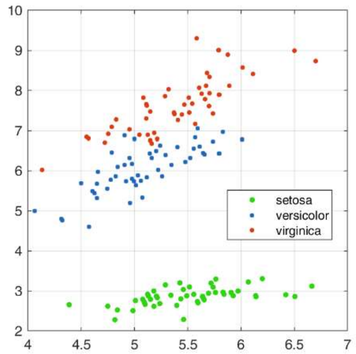
Table I

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5·1	3·5	1·4	0·2	7·0	3·2	4·7	1·4	6·3	3·3	6·0	2·5
4·9	3·0	1·4	0·2	6·4	3·2	4·5	1·5	5·8	2·7	5·1	1·9
4·7	3·2	1·3	0·2	6·9	3·1	4·9	1·5	7·1	3·0	5·9	2·1
4·6	3·1	1·5	0·2	5·5	2·3	4·0	1·3	6·3	2·9	5·6	1·8
5·0	3·6	1·4	0·2	6·5	2·8	4·6	1·5	6·5	3·0	5·8	2·2
5·4	3·9	1·7	0·4	5·7	2·8	4·5	1·3	7·6	3·0	6·6	2·1
4·6	3·4	1·4	0·3	6·3	3·3	4·7	1·6	4·9	2·5	4·5	1·7
5·0	3·4	1·5	0·2	4·9	2·4	3·3	1·0	7·3	2·9	6·3	1·8
4·4	2·9	1·4	0·2	6·6	2·9	4·6	1·3	6·7	2·5	5·8	1·8
4·9	3·1	1·5	0·1	5·2	2·7	3·9	1·4	7·2	3·6	6·1	2·5
5·4	3·7	1·5	0·2	5·0	2·0	3·5	1·0	6·5	3·2	5·1	2·0
4·8	3·4	1·6	0·2	5·9	3·0	4·2	1·5	6·4	2·7	5·3	1·9
4·8	3·0	1·4	0·1	6·0	2·2	4·0	1·0	6·8	3·0	5·5	2·1
4·3	3·0	1·1	0·1	6·1	2·9	4·7	1·4	5·7	2·5	5·0	2·0
5·8	4·0	1·2	0·2	5·6	2·9	3·6	1·3	5·8	2·8	5·1	2·4
5·7	4·4	1·5	0·4	6·7	3·1	4·4	1·4	6·4	3·2	5·3	2·3
5·4	3·9	1·3	0·4	5·6	3·0	4·5	1·5	6·5	3·0	5·5	1·8
5·1	3·5	1·4	0·3	5·8	2·7	4·1	1·0	7·7	3·8	6·7	2·2
5·7	3·8	1·7	0·3	6·2	2·2	4·5	1·5	7·7	2·6	6·9	2·3
5·1	3·8	1·5	0·3	5·6	2·5	3·9	1·1	6·0	2·2	5·0	1·5
5·4	3·4	1·7	0·2	5·9	3·2	4·8	1·8	6·9	3·2	5·7	2·3

Page 41



Page 42



Page 43

Recognizing handwritten digits

Images of size 28x28 pixels, grayscale

Page 44

When stored as rows of a matrix: 784 real values between 0 (black) and 1 (white)

Variables - mnist

mnist

10000x784 double

	177	178	179	180	181	182	183	184	185	186	187
5	0	0	0	0	0	0.0588	0.0863	0.0353	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0
7	0	0.0863	0.6392	0.6353	0.6353	0.2000	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0.0235	0.5216	0.9922	0.9922	0.9922	0.9922	0.9922	0.2078	0
10	0	0	0	0.0314	0.2824	0.2941	0.0353	0	0	0	0
11	0	0	0	0	0	0	0	0.0431	0.3843	0.3843	0.3843
12	0	0	0	0	0	0	0	0	0	0	0.1020
13	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0.0392	0.0392	0.0235	0	0	0
18	0.9961	0.5294	0	0	0	0	0	0	0	0	0
19	0.9922	0.9922	0.9922	0.9922	0.9922	0.3608	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0.0118	0.0627	0.0353	0	0	0
26	0	0	0.5333	0.9922	0.9922	0.9922	0.9922	0.9922	0.9922	0.9922	0.9922
27	0.0824	0.5569	0.5961	0.9922	0.4000	0	0	0	0	0	0

Page 45

Preparing the data

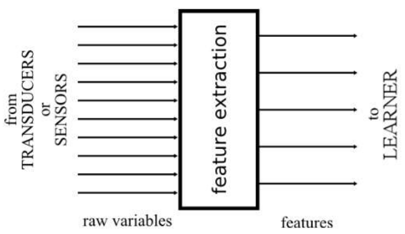
Page 46

Data cleaning

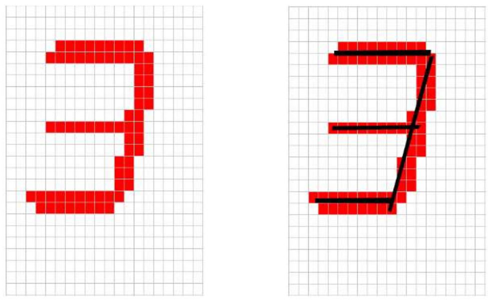
- Change data types to make them suitable for your software (es. change strings into numerical codes)
- Remove data with out-of-range values
- Deal with **missing** data – Several possible strategies:
  - Removing observations (rows)
  - Removing input variables (columns)
  - **Imputation** of missing values
- Align timestamped data

Page 47

Feature extraction



Page 48



Page 49

Learning problems

- Representation: learn to reproduce what is in your data
- Generalization: learn to understand what your data represent

Solving the **representation** problem finds the best solution for the training set

Solving the **generalization** problem finds the best solution for any data from the same source that generated the training set

Page 50

More names

- **Learning machine** – also "learner".
  Not necessarily a real machine! Maybe a program
- **Task** – a problem to be solved.
  We don't have a description of the problem, but data
- **Learning** – adjusting quantities "inside the machine"
  (e.g., algorithm parameters) to do a certain task

Page 51

Even more names

- Hypothesis – a specific learning machine that implements a certain task
  e.g., a neural network that reads images and recognizes whether there is a known person (biometric recognition)
- Hypothesis space – the set of all tasks that can be learned by a specific learning machine
  e.g., the set of all classifiers that can be implemented by a specific neural network by setting its internal parameters

To make it more intuitive:

Hypothesis space = a learning machine before learning
Hypothesis = a learning machine after learning a task

Page 52

Possible scenarios

1 (useful mostly for reasoning): not very usefull
- Learner = hypothesis space = $\mathscr{H}$ is fixed in advance
- Data are fixed (population)

→ Find correct learners in $\mathscr{H}$ (a *representation* task)

2 (not realizable): 2
- No data necessary: Probabilities are assumed to be known!

→ Find best hypotesis space $\mathscr{H}$ and optimal learner

3 (your usual situation): 1
- Data will be **stochastic** but probabilities are **not known**
- Hypothesis space $\mathscr{H}$ fixed, chosen in advance

→ Find learners in $\mathscr{H}$ which are correct **for any possible realization of the data** (a *generalization* task)

1

In this scenario, you can expect other data that is not training data.

2

In security risk, concepts like "risk", are introduced, where you talk about the probability of that event.

Page 53

Scenario 1:
We have the whole population

Page 54

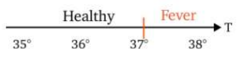
Linear threshold classifiers

Page 55

Intuition

What do you do when you want to categorize something according to one measurement?

You mark a change **point** and categorize values either to the left or to the right on the axis of values



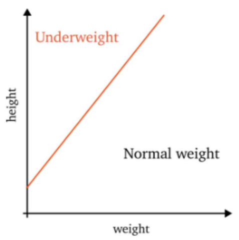
Page 56

Intuition

What do you do when you want to categorize something according to two measurements?

You draw a **line** and categorize values either to one side or to the other side of the line in the plane of values (data space)

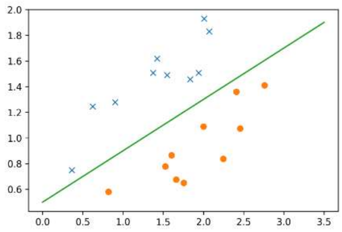
In a three dimensions, the equivalent of the line is a plane.



Page 57

Linearly separable data

In a **classification task**, data are said to be **linearly separable** if there exists a line (or a plane if d=3, or a hyperplane if d>3) such that **object of a given class are all on the same side of the line** (plane, hyperplane)



Page 58

Linearly separable data

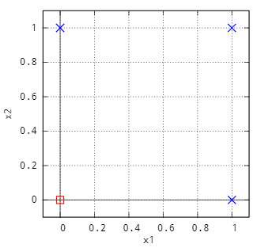
The two classes lie on opposite sides of a hyperplane.

Hyperplane:

- A d-dimensional homogeneous hyperplane: $\mathbf{x} \cdot \mathbf{w} = 0$ where x is our data and w is its weight
- homogeneous $\equiv$ through the origin
- Normal: $\mathbf{w}$ is a vector that goes out of the plane in a normal direction
- Unit-length normal: $\hat{\mathbf{w}} = \frac{\mathbf{w}}{\|\mathbf{w}\|}$
- Positive side: $\mathbf{x} \cdot \mathbf{w} > 0$ and on the negative outspace we have x w < 0
- Non-homogeneous hyperplane: $\mathbf{x} \cdot \mathbf{w} = \theta$

Page 59

Example: data



Page 60

Example: problem statement

- We have to encode classes into 0 and 1 where the values 0 and 1 are cathegorical.
- We decide to solve using a separating hyperplane
  (a line in the 2-dim plane)
- Positive side → one class, negative side → the other class

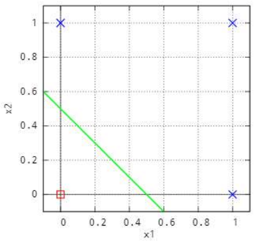
Page 61

Example: discussion

- Infinite lines solve the problem $\infty^2 \cdot \infty$ solutions
- Infinite parameter sets represent each equation (we only look at the sign!)
- Even the sign is arbitrarily assigned to the classes

Page 62

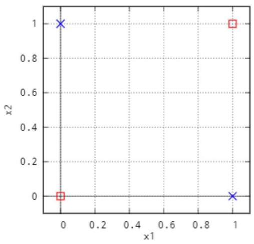
Example: one possible solution



Page 63

An example without solution

It is not possible to solve this problem and
separating the two set with a straight line



Page 64

A learner suited for linearly separable data

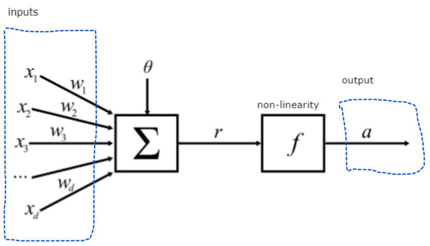
$r = \mathbf{x} \cdot \mathbf{w} - \theta$

$a = f(r)$ applying non linearity

- $\mathbf{x}$ is a d-dimensional vector of input values
- $\mathbf{w}$ is the corresponding (d-dimensional) vector of parameters
- $\cdot$ indicates scalar product
- r indicates the net, "integrated" input where r is 0 on the hyperplane
- $f()$ is a nonlinear, monotonic **activation function**
- θ is a threshold
- a indicates the output.

Page 65

Diagram of linear-threshold classifier

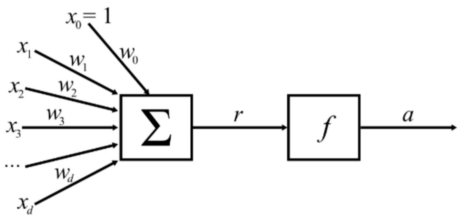


Page 66

Getting rid of the threshold

$r = \mathbf{x} \cdot \mathbf{w}$ we put θ inside w because are both numerical parameters

$a = f(r)$



Page 67

$$
\begin{aligned}
r - \theta &= \sum_{i=1}^{d} w_i x_i - \theta \\
&= w_1 x_1 + w_2 x_2 + \ldots + w_d x_d - \theta
\end{aligned}
$$

Let's call θ with another name: $\theta = -w_0 x_0$, with $x_0 \equiv 1$:

$$
\begin{aligned}
r - \theta &= \sum_{i=1}^{d} w_i x_i + w_0 \\
&= \sum_{i=0}^{d} w_i x_i
\end{aligned}
$$

θ was a **threshold** (subtracted),
w_0 is a **bias** (summed as an offset value).

Page 68

We like all parameters to be in one place!

NOTE: now indexes for the components of **x** and **w** start at 0, not 1!
$\mathbf{x} = [x_0, x_1, x_2, \ldots, x_d]$
$\mathbf{w} = [w_0, w_1, w_2, \ldots, w_d]$

- $w_0 = -\theta = \text{BIAS}$
- $x_0 \equiv 1$

Page 69

Examples of activation functions f

Heaviside step (defined on $[-\infty, +\infty] \rightarrow \{0, +1\}$):

$$f(r) = \mathbf{1}(r) = \begin{cases} +1 & r \geq 0 \\ 0 & r < 0 \end{cases}$$

value in zero
(ties broken arbitrarily)

Signum function (defined on $[-\infty, +\infty] \rightarrow \{-1, +1\}$):

$$f(r) = \text{sign}(r) = \begin{cases} +1 & r \geq 0 \\ -1 & r < 0 \end{cases}$$

The signum function is a symmetrization in the interval $[-1, +1]$ of the Heaviside step function: $\text{sign}(r) = 2 * \mathbf{1}(r) - 1$.

In a numerical situation where there are 2 varible that are checked at the end if they are equals

$$double\ x,\ y;$$

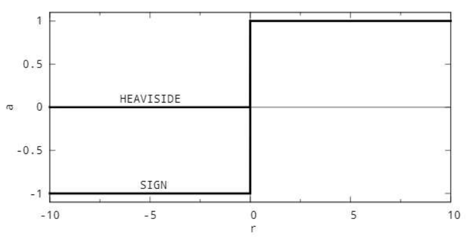
$$x\ == y$$

is better to chose a value epsilon under which the two value are considered the same

$$x - y < eps$$

Page 70

Step and signum



In some case, we want to know also the reliability of the decision and not only the decision by itself.

Page 71

Sigmoid and hyperbolic tangent

to better adapt to the classification that we are doing between the two classes compared to the linear one.

Sigmoid or *logistic function* (defined on $[-\infty, +\infty] \rightarrow [0, +1]$):

$$f(r) = \sigma(r) = \frac{1}{1 + e^{-r}}$$

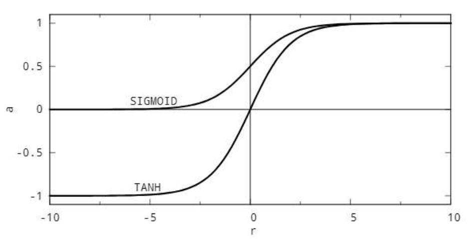
Hyperbolic tangent or tanh (defined on $[-\infty, +\infty] \rightarrow [-1, +1]$):

$$f(r) = \tanh(r) = \frac{1 - e^{-2r}}{1 + e^{-2r}}$$

The hyperbolic tangent function function is a symmetrization in the interval $[-1, +1]$ of the sigmoid function: $\tanh(r) = 2 * \sigma(2r) - 1$.

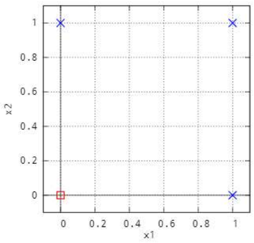
Page 72

Sigmoid and tanh



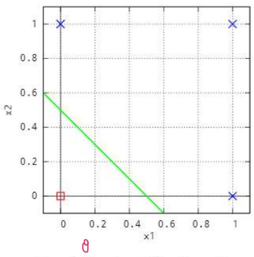
Page 73

Linearly separable example



Page 74

Linearly separable example



θ

$\mathbf{w} = [.5, -1, -1]$, red is positive