Machine Intelligence 2

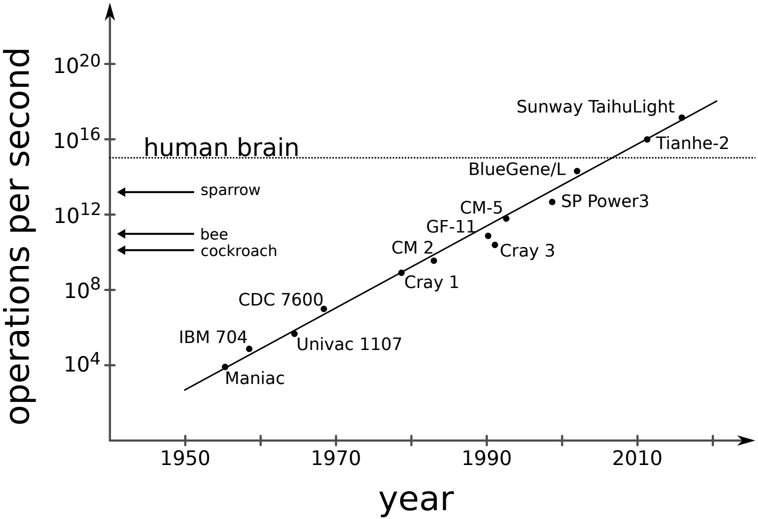
0.1 Introduction

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

SS 2018

Moore's law



operations per second
10^{20}
10^{16}
10^{12}
10^{8}
10^{4}
human brain
sparrow
bee
cockroach
Sunway TaihuLight
Tianhe-2
BlueGene/L
SP Power3
CM-5
GF-11
Cray 3
CM 2
Cray 1
CDC 7600
IBM 704
Univac 1107
Maniac
1950
1970
1990
2010
year

Brains vs. machines

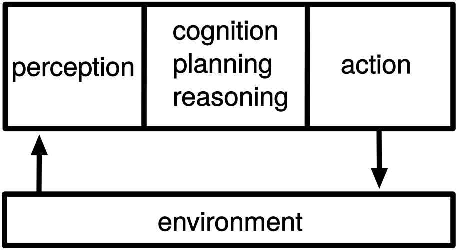
Brains are good where machines are bad:

- pattern recognition (images,audio, touch, multimodal data, but also abstract patterns)
- communication (language, speech)
- categorization and classification
- model building, inference and prediction
- control (robots, plants, software agents)

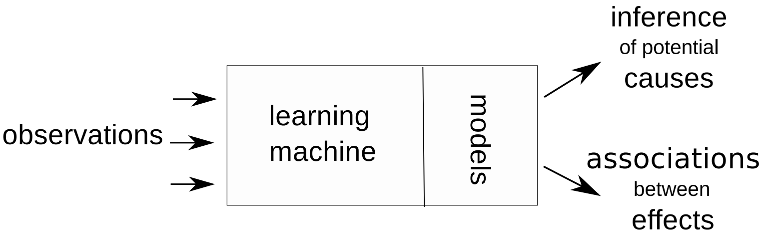
Machines are good where brains are bad:

- calculus
- chess
- manipulating symbols/strings

Machine intelligence: embedded agents

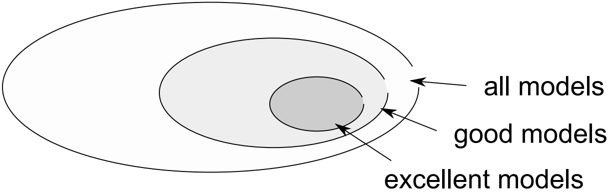


Learning to predict



- inductive learning: learning from examples

Learning as model selection



Learning as model selection

data representation

↓

model class

↓

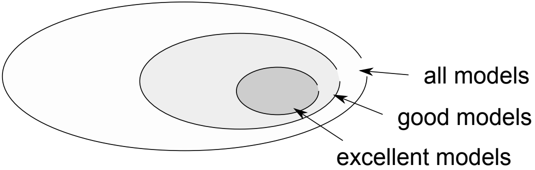
performance measure

↓

optimization

↓

validation



Supervised learning: "learning with a teacher"

Observations:

- series of observations: $\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{x}}^{(2)}, \ldots, \underline{\mathbf{x}}^{(p)}$
- corresponding labels/targets: $y^{(1)}, y^{(2)}, \ldots, y^{(p)}$

Goal:

- predict the label/target of new (previously unseen) observation

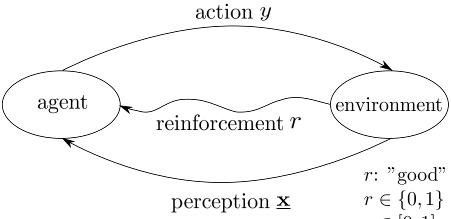
Application:

- classification
- regression

Reinforcement learning: "learning behavior"

Observations:

- series of visited states: $\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{x}}^{(2)}, \ldots, \underline{\mathbf{x}}^{(p)}$
- series of executed actions: $y^{(1)}, y^{(2)}, \ldots, y^{(p)}$
- series of experienced rewards: $r^{(1)}, r^{(2)}, \ldots, r^{(p)}$



r: "good" or "bad"
$r \in \{0, 1\}$
$r \in [0, 1]$

Goal:

- find for every state $\underline{\mathbf{x}}$ the action y that maximizes future reward

Unsupervised learning: "self-organization"

Observations:

- series of observations: $\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{x}}^{(2)}, \ldots, \underline{\mathbf{x}}^{(p)}$

Goal:

- build a *useful* representation of observations $\underline{\mathbf{x}}$
- extract relevant structure of observations $\underline{\mathbf{x}}$

Application:

- dimensionality reduction
- clustering
- categorization
- source separation

Learning paradigms

- phenomenological characterization of learning paradigms
- not at all based on mathematical principles (e.g. same inductive learning approaches for "supervised" and "unsupervised" problems)

Overview over MI2: unsupervised methods

1. Principal Component Analysis
   - 1.1 Principal Component Analysis
   - 1.2 Hebbian Learning for Linear Neurons
   - 1.3 Kernel Principal Component Analysis
   - 1.4 Novelty Filter
2. Independent Component Analysis
   - 2.1 Independent Component Analysis
   - 2.2 Model-based Independent Component Analysis
   - 2.3 Second Order Source Separation
   - 2.4 Fast ICA
3. Stochastic Optimization
   - 3.1 Simulated Annealing
   - 3.2 The Gibbs Distribution
   - 3.3 Mean-Field Annealing

Overview over MI2: unsupervised methods

4 Clustering and Embedding
   4.1 K-means Clustering
   4.2 Pairwise Custering
   4.3 Self-Organizing Maps
   4.4 Locally Linear Embedding

5 Probability Density Estimation
   5.1 Density Estimation: Kernel-based/Parametric
   5.2 Maximum Likelihood & Estimation Theory
   5.3 Mixture Models & EM Algorithm

6 Hidden-Markov Models
   6.1 Modeling sequential data
   6.2 Forward-Backward Algorithm
   6.3 Viterbi Algorithm

Overview over MI1: supervised methods

Artificial neural networks

- Connectionist neurons
- Multilayer perceptrons & radial basis function networks
- Learning, generalisation, regularisation

Learning theory and support vector machines

- Statistical learning theory
- Support vector machines (SVMs) & the kernel trick

Probabilistic methods

- Uncertainty and inference
- Bayesian networks, Bayesian inference and neural networks

Reinforcement Learning

Textbooks for MI2

- Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, 2006.
- Cichocki & Amari, Adaptive Blind Signal and Image Processing, Wiley, 2002.
- Duda, Hart & Stock, Pattern Classification, Wiley, 2000.
- Haykin, Neural Networks, Prentice Hall, 1998
- Hyvärinen, Karhunen & Oja, Independent Component Analysis, Wiley, 2001
- Kohonen, Self-Organizing Maps, Springer-Verlag, 1997.
- Schölkopf & Smola, Learning with Kernels, MIT Press 2002.

Textbooks for MI2

- Kohonen, Self-Organization and Associative Memory, Springer, 1989
- Kay, Fundamentals of Statistical Signal Processing - Vol.I: Estimation Theory, Prentice Hall, 1993 *.
- Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, 1996 *

* advanced readings

on-line review and tutorial:

- Hyvärinen, Survey on Independent Component Analysis, on-line via: http://www.cis.hut.fi/aapo/ps/NCA99.pdf
- Hyvärinen, Independent Component Analysis: Algorithms and Applications, on-line via: http://www.cs.helsinki.fi/u/ahyvarin/papers/NN00new.pdf

Recommended readings for MI2 chapters

1. Principal Component Analysis
   - 1.1 Principal Component Analysis Haykin, Ch. 8.3; Bishop, Ch. 12.1
   - 1.2 Kernel Principal Component Analysis Haykin, Ch. 8.10; Schölkopf & Smola, Chs. 14.1-14.3
   - 1.3 Hebbian Learning for Linear Neurons Haykin, Chs. 8.4, 8.5
   - 1.4 Novelty Filter Kohonen 1989, Chs. 4.3, 4.4
2. Independent Component Analysis
   - 2.2 Model-based Independent Component Analysis Haykin, Ch. 10.11; Hyvärinen et al., Ch. 9; Cichochi & Amari, Chs. 5.5, 6.1, 6.12
   - 2.3 Second Order Source Separation Hyvärinen 2001, Ch 18
   - 2.4 Fast ICA Hyvärinen 2001, Ch 8
3. Stochastic Optimization
   - 3.1 Simulated Annealing: Haykin, Ch. 11.1-11.7
   - 3.2 The Gibbs Distribution: Haykin, Ch. 11.1-11.7
   - 3.3 Mean-Field Annealing: original publications

Recommended readings for MI2 chapters

4. Clustering and Embedding
   - 4.1 K-means Clustering Haykin, Ch. 11.1-11.7
   - 4.2 Pairwise Clustering original publications
   - 4.3 Self-Organizing Maps Haykin, Chs. 9.1-9.6, 9.11; Vertiefung: Kohonen 1997
   - 4.4 Locally Linear Embedding Saul & Roweis tutorial; Roweis & Saul 2000
5. Probability Density Estimation
   - 5.1 Density Estimation: Kernel-based/Parametric Bishop, Ch. 2.5
   - 5.2 Maximum Likelihood & Estimation Theory Kay, Chs. 3, 7
   - 5.3 Mixture Models & EM algorithm Bishop, Chs. 9.1, 9.2
6. Hidden-Markov Models

   all: Bishop, Chs. 13.1, 13.2

End of Section 0.1

the following slides contain

OPTIONAL MATERIAL

learning 'without a teacher'

The problem with labeled data

- expensive
- contains relatively little information (binary)

$\Rightarrow$ often not enough to estimate the parameters of complex models.

Methods subsumed under the term unsupervised learning deal with finding structure or regularities in a set of observations $\underline{\mathbf{x}}^{(1)}, \underline{\mathbf{x}}^{(2)}, \ldots, \underline{\mathbf{x}}^{(p)}$

Applications: Knowledge discovery, explorat. data analysis, data mining

Frequent Item Sets for Market basket analysis at WalMart,Google Searches, Youtube tags, Twitter hashtags, administrative claims (eHealth), targeted advertising (e-commerce), density estimation (Starbucks) ...

what are we looking for?

Many datasets ...

... are grouped or clustered – identifying groups / categories, construction of taxonomies, generalisation, preprocessing for prediction

... are high-dimensional – dimension reduction, compression, visualisation: Humans are very good at discovering nonlinear structure.

... may display interesting (or uninteresting) directions – find informative features (projection methods) for characterisation & denoising

... may be determined by different causes – unmixing a mixture of sources, definition of components, infering causes

many datasets are large $\Rightarrow$ automatic data analysis

Artificial **Intelligence** (from Russel & Norvig)

Motivating question

"How is it possible for a slow, tiny brain, whether biological or electronic, to perceive, understand, predict, and manipulate a world far larger and more complicated than itself?"

General Tasks: Perception, logical reasoning, navigation
Specific Tasks: Chess, search, theorem proving, disease diagnosis, speech recognition, translation, etc.

Intelligence

- strategies/activities we would call "intelligent" if done by a person to make decisions, solve problems, learn ($\rightarrow$ homo sapiens).
- *operational definition*: Turing test (imitation game, A. Turing, 1950)

$\rightarrow$ **Extraction** and **representation** of knowledge; **reasoning** (inference)

AI: From Hanoi to ELIZA, ACT-R and ANNs

GOAL: build agents that think/act rational/like (successful) humans.
OR: build systems that efficiently optimize cost functions ($\rightarrow$ heuristics)

A list of Problems / Challenges

- simple decision making (sensor fusion)
- simple problem solving (Tower of Hanoi, means-ends analysis)
- chatterbots (ELIZA, Help systems), intelligent tutoring ($\rightarrow$ CogSci)
- chess (Deep Blue), video game adversaries
- shift to technology industry: smartphones, SIRI, google, translation
- DARPA Challenge, Google driverless car, Jeopardy! (Watson)

leading-edge definition of AI

- ”AI research is that which computing scientists do not know how to do cost-effectively today”. (Wikipedia)

Machine Learning

Definition: Learning

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." (Tom Mitchell)

- important part of many AI systems ($\rightarrow$ extraction of knowledge)
- data driven, adaptive systems
- programming & deduction vs. learning & self-organization
- symbolic approaches vs. sub-symbolic approaches
  biologically inspired learning rules
  (knowledge bases & reasoning vs. easy & robust learning rules, Hebb)
- Many problems involve uncertainty $\rightarrow$ Probability & Decision theory

Machine Intelligence

Focus: statistical approaches and learning algorithms

→ Extract, analyse, and use principles of neural information processing to build intelligent "machines".

MI has large overlap with: Methods from AI, Statistics, Pattern Recognition, Machine Learning, Learning theory, Data mining.

Main Topics

- Learning / prediction & generalization of statistical relationships
- Statistical inference in graphical models
- Finding structure in high dimensional data sets

Influences from biology

2 Perspectives: engineering $\leftrightarrow$ reverse engineering of biological systems

- Biological systems are amazingly good at certain perceptual tasks
- Intelligent machines can be much better at many tasks

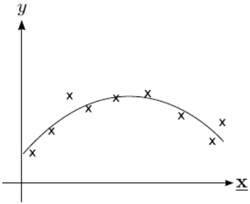
$\rightarrow$ Learn from Biology to build better machines (biomorphic engineering)
$\rightarrow$ Understanding the statistical structure of sensory environments helps to understand principles of adapted perceptual system

Design Principles from Biology

- simple, but highly optimized hardware (echolocation in bats, sound localization in barn owls, ultra fast face recognition)
- Plasticity (synaptic strength, lifelong renewal of cells in the olfactory system) drifting environments,
- Adaptation in sensory systems
- Graceful degradation

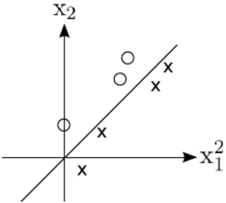
Illustration of MI1 paradigms

Regression



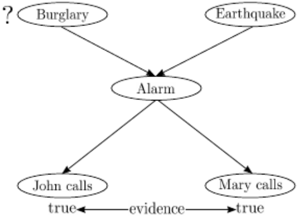
predict value y
given $\underline{\mathbf{x}}$

Classification



predict label y
given $\underline{\mathbf{x}}$

Inference



predict p(Burglary)
given evidence

Artificial Neural Networks (ANNs)

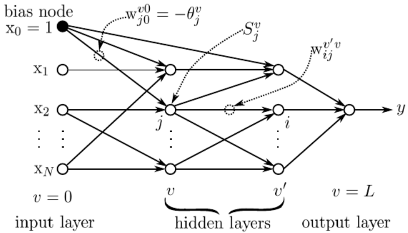
ANNs are ...

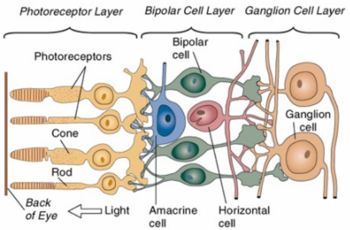
- ... useful for Regression and Classification
- ... brain *inspired* model architectures (McCulloch & Pitts, 1943)
- ... built from simple elements ($\rightarrow$ connectionist neurons)
- ...... with low precision & robustness ($\rightarrow$ binary, noisy)
- ... massively parallel systems ($\rightarrow$ "networks")

Consequences

- Distributed representation of information
- No clear separation between "data" and "program"

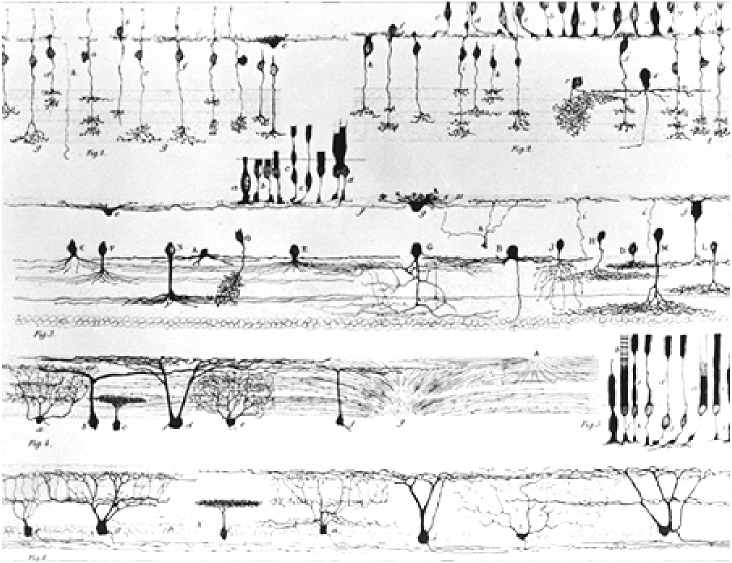
ANNs & brain style computation



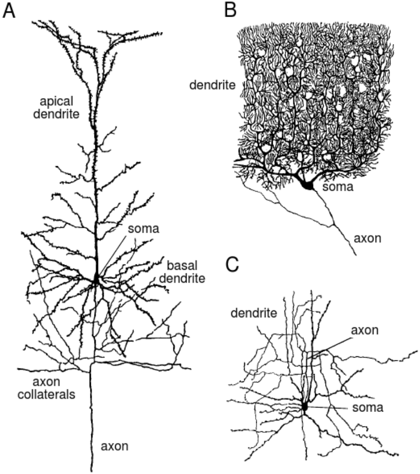


Source: Adapted from Dowling, J. E., and Boycott, B. B. *Proceedings of the Royal Society of London, B.*, 1966, *166*, 80–11

Retinal cells (Ramon y Cajal)



Excitatory cortical cells



from Dayan & Abbott (2001)

Unsupervised Learning / Exploratory Statistics

When we're learning to see, nobody's telling us what the right answers are - we just look. Every so often, your mother says 'that's a dog', but that's very little information. You'd be lucky if you got a few bits of information - even one bit per second - that way. The brain's visual system has 10^{14} *neural connections. And you only live for* 10^{9} *seconds. So it's no use learning one bit per second. You need more like* 10^{5} *bits per second. And there's only one place you can get that much information: from the input itself. (Geoffrey Hinton, 1996)*