Computational cognitive modeling

Bayesian approach

Karla Štěpánová

ČVUT v Praze
Fakulta Elektrotechnická
Katedra kybernetiky
Výzkumná skupina BioDat
http://bio.felk.cvut.cz
Výzkumná skupina Incognite
http://incognite.felk.cvut.cz

April 13, 2015

Obsah

Computational cognitive modeling

Computational cognitive modeling

= simulations of complex mental processes in different areas of cognition, the goal - to understand, describe, model and predict observed human behavior

Cognition

=mental process of knowing, including aspects such as awareness, perception, reasoning and judgement

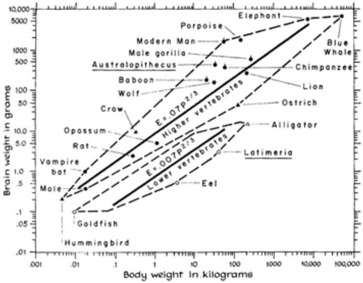
Latin word cognitio: **-co** (intensive) + **nosecere** (to learn)

Modeling

Data never speak for themselves, require a model to be understood and explained

Several alternative models $->$ compare – quantitative evaluation and intellectual judgement

Motivation



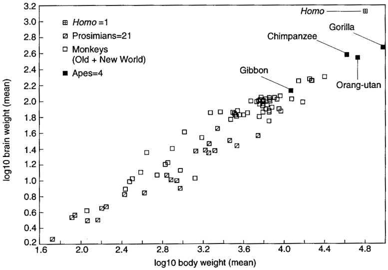


Figure: Encephalisation quotient

Motivation

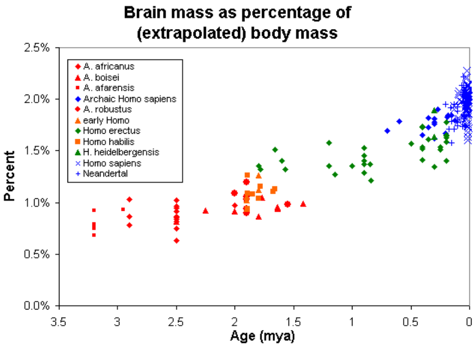


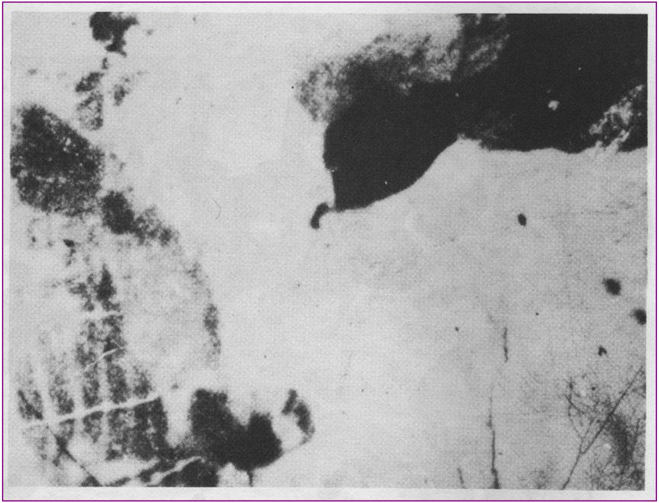
Figure: Brain mass: Chart by Nick Matzke

Language

Technology

Art, culture, high tech

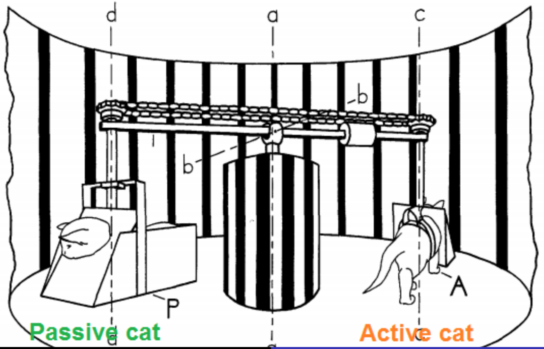
Motivation



Motivation

Motivation

Movement is essential for perceptual learning (visually-guided behavior - depth perception, paw placement, visual cliff, blink to an approaching object etc.)- brain doesn't consist of separated neurons



Motivation

John Langford:

"A human brain has about 10^{15} synapses which operate at about 10^{2} per second implying about 10^{17} bit ops per second"

So.. A transcription of 1 second of brain activity at the neural spike level would fill up about 40,000 ordinary 300Gb hard drives

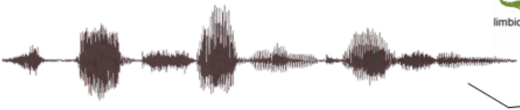
...and consumes 20% of body's oxygen (approx 1.3 kg)

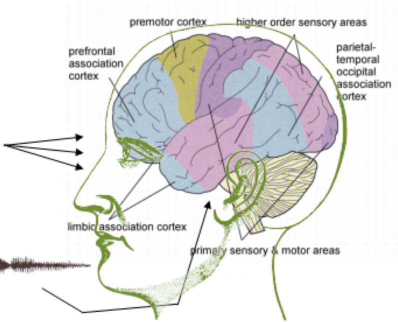
Is it worth?

Kandel (1995)

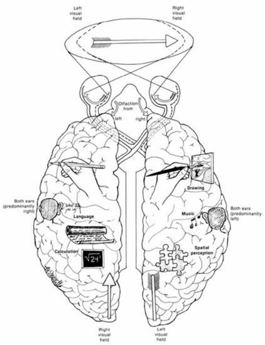
Processing information

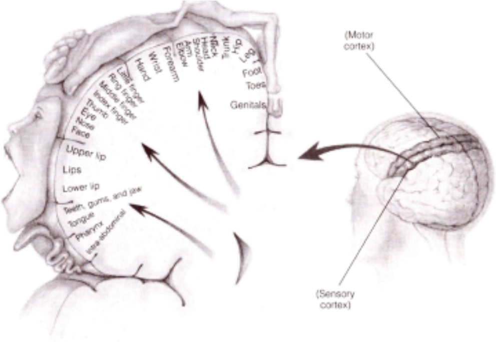
Single units → Networks, Unimodal areas → Association → Evolution in time, reasoning, induction

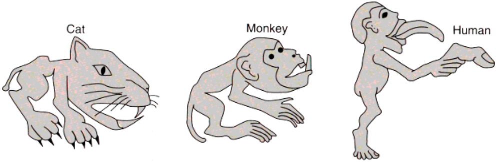




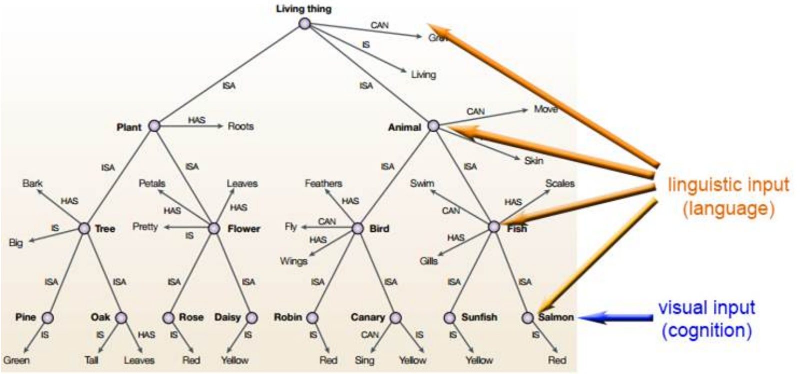
Internalized representations of world







Multimodal association - creating internal representations



- **Traditional models of cognition:**
  - „connectionism" adaptivity
  - „rule-based" (Minsky 1968, a priori rules) apriori knowledge
  - „parametric model-based" adaptivity + apriori knowledge

Combinatorical explosion or computational complexity

- Neural and biological plausability
- Parametric X nonparametric methods

- **Parametric model-based models** - Parameters can capture variablities and uncertainities in the data (pdf)

- **Physical theory of mind:** apriori knowledge + adaptivity + **ability of computation in the real time**

Cognitive architectures - Marr's levels of abstraction

Marr's levels of abstraction

Computational: What are the abstract inference problems that the mind needs to solve, and what are the solutions? *Bayesian parametric modeling*

Algorithmic: What information and processing steps are followed to arrive at the solutions? *Connectionism*

Implementation: How does the brain carry out these operations?

Marr, D (1982). Vision. *A Computational Investigation into the Human Representation and Processing of Visual*

Cognitive architectures - Marr's levels of abstraction

Marr's levels of abstraction

Computational: What are the abstract inference problems that the mind needs to solve, and what are the solutions? *Bayesian parametric modeling*

Algorithmic: What information and processing steps are followed to arrive at the solutions? *Connectionism*

Implementation: How does the brain carry out these operations?

Sun's levels

Sociological level - inter-agent processes, collective behavior of agents

Psychological level - individual behavior of agents

Componential level - intra-agent processes, modular construction of agents

Physiological level - biological implementation

Marr, D (1982). Vision. *A Computational Investigation into the Human Representation and Processing of Visual*

Disiderata - Cognitive architectures

Newell (1990). Unified theories of cognition

Flexibility

Adaptivity

Autonomy

Self-awarness

Operation in real-time and in complex environment

Usage of symbol and abstractions

Usage of language

Learning from environment

Acquiring capabilities through development,

Be realizable as a neural system

Be constructable by an embryological growth process

Arise through evolution

Disiderata - Computational cognitive neuroscience model

The neuroscience ideal

A CCN model should not make any assumptions that are known to contradict the current neuroscience literature.

The simplicity heuristic

No extra neuroscientific detail should be added to the model unless there are data to test this component of the model or the model cannot function without this detail.

The Set-in-Stone Ideal

Once set, the architecture of the network and the models of each individual unit should remain fixed throughout all applications.

The Goodness-of-Fit Ideal

A CCN model should provide good accounts of behavioral and at least some neuroscience data.

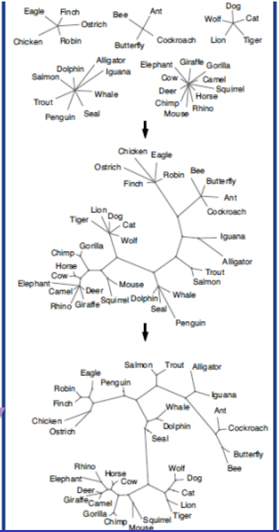
G. F. Ashby and S. Helie(2011). A tutorial on computational cognitive neuroscience: Modeling the neurodynamics of cognition

Functionalism

Two ideas

Two different ways of thinking about cognition:

– **Functionalism**: the mind is an information system, so we're interested in what inferences are licenced by data



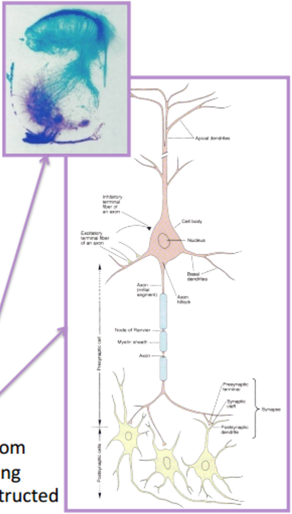
A sequence of theories about animals licensed by the data presented to a child (Kemp & Tenenbaum, 2008)

Connectionism

Two ideas

Two different ways of thinking about cognition:

- **Connectionism**: the mind is built from the brain, a physical system built out of massively parallel networks of simple processors (neurons)
- What kind of behaviours does such a network produce?



The basic components from which the concept learning system needs to be constructed

Similarities and differences

Similarities and differences

Connectionists and functionalists agree on lots of things

- Form of the mental representation is critical
- The nature of human induction is central
- Learning is a cool topic

We differ on one very big question

- Are we more interested in the kind of statistical inference performed by the mind (a question of **why**), or what the brain does to implement the inferences (a question of **how**)?
- Connectionists operate at the **algorithmic** level, while functionalists operate at the **computational** level

The challange

How do we generalize successfully from very limited data?

Bayes rule

For any hypothesis *h* and data *d*,

Posterior probability

Likelihood

Prior probability

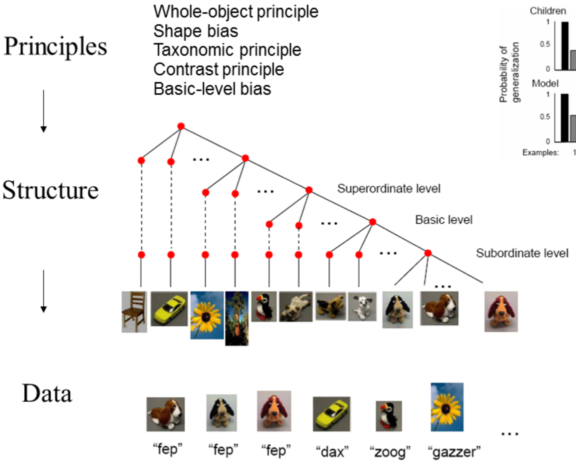
$$p(h \mid d) = \frac{p(d \mid h)\,p(h)}{\sum_{h' \in H} p(d \mid h')\,p(h')}$$

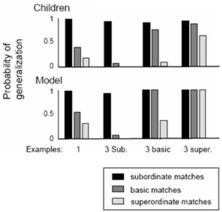
Sum over space of alternative hypotheses

Why Bayes?

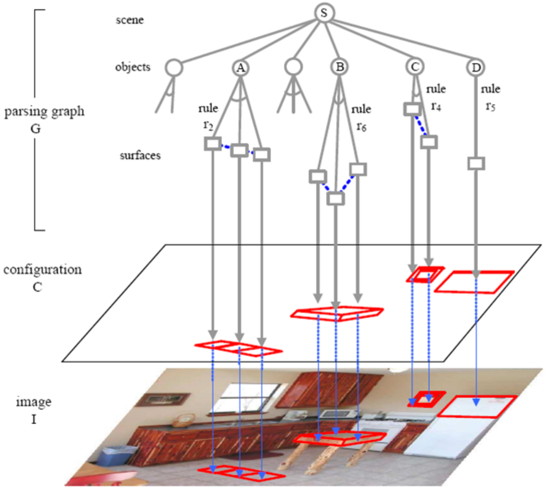
- The problem of induction
  - How does the mind form inferences, generalizations, models or theories about the world from impoverished data?
- Induction is ubiquitous in cognition
  - Vision (+ audition, touch, or other perceptual modalities)
  - Language (understanding, production)
  - Concepts (semantic knowledge, "common sense")
  - Causal learning and reasoning
  - Decision-making and action (production, understanding)
- A unifying framework for explaining cognition.
  - How people can learn so much from such limited data.
  - Strong quantitative models with minimal ad hoc assumptions.
  - Why algorithmic-level models work the way they do.
- A framework for understanding how structured knowledge and statistical inference interact.
  - How structured knowledge guides statistical inference, and may itself be acquired through statistical means.
  - What forms knowledge takes, at multiple levels of abstraction.
  - What knowledge must be innate, and what can be learned.
  - How flexible knowledge structures may grow as required by the data, with complexity controlled by Occam's razor.

Examples - Learning word meanings





Examples - Vision as probabilistic parsing



Examples - Vision as probabilistic parsing

Examples - Grammar

Universal Grammar

P(grammar | UG)

Grammar

P(phrase structure | grammar)

Phrase structure

P(utterance | phrase structure)

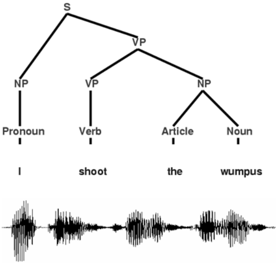
Utterance

P(speech | utterance)

Speech signal

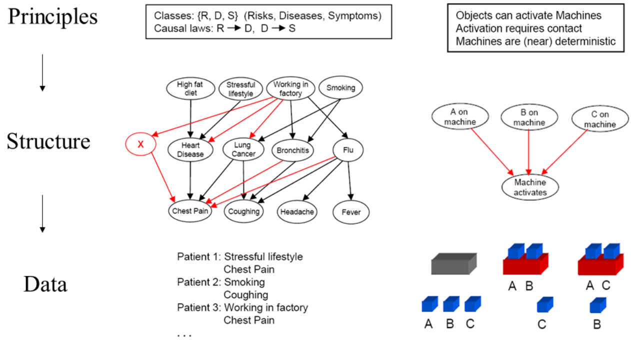
Hierarchical phrase structure grammars (e.g., CFG, HPSG, TAG)

$$S \rightarrow NP\ VP$$
$$NP \rightarrow Det\,[Adj]\,Noun\,[RelClause]$$
$$RelClause \rightarrow [Rel]\,NP\ V$$
$$VP \rightarrow VP\ NP$$
$$VP \rightarrow Verb$$

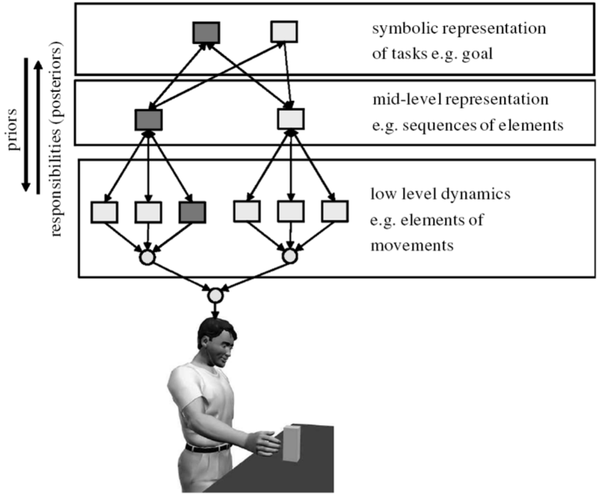


Examples - Causal learning and reasoning

Causal learning and reasoning



Examples - Motor control



Bayes rule

For any hypothesis *h* and data *d*,

Posterior probability

Likelihood

Prior probability

$$p(h \mid d) = \frac{p(d \mid h)\, p(h)}{\sum_{h' \in H} p(d \mid h')\, p(h')}$$

Sum over space of alternative hypotheses

Bayes rule - Priors

Prior knowledge about the world $->$ interpret data in the case of the uncertainity

Prediction - the more uncertain the data, the more the prior should influence the interpretation

Priors should reflect the statistics of the sensory world

Coin flipping

HHHHH

HHTHT

What process produced these sequences?

Coin flipping

Contrast simple hypotheses:

h1: "fair coin", P(H) = 0.5

h2: "always heads", P(H) = 1.0

Bayes' rule:

$$\frac{P(h|d) = P(h)P(d|h)}{\sum_{h_i} P(h_i)P(d|h_i)}$$

With two hypotheses, use odds form

Coin flipping

Comparing two simple hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D:	HHTHT		
H_1, H_2:	"fair coin", "always heads"		
$P(D\|H_1) =$	$1/2^5$	$P(H_1) =$	999/1000
$P(D\|H_2) =$	0	$P(H_2) =$	1/1000

$$P(H_1|D) \,/\, P(H_2|D) = \text{infinity}$$

Coin flipping

Comparing two simple hypotheses

$$\frac{P(H_1|D)}{P(H_2|D)} = \frac{P(D|H_1)}{P(D|H_2)} \times \frac{P(H_1)}{P(H_2)}$$

D: HHHHH

H_1, H_2: "fair coin", "always heads"

$P(D|H_1) =$ $1/2^5$ $P(H_1) =$ 999/1000

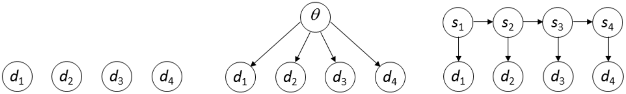
$P(D|H_2) =$ 1 $P(H_2) =$ 1/1000

$$P(H_1|D) \,/\, P(H_2|D) \approx 30$$

Coin flipping

Model selection

- Assume hypothesis space of possible models:



Fair coin: $P(\mathrm{H}) = 0.5$

$P(\mathrm{H}) = \theta$

Hidden Markov model: $S_i \in$ {Fair coin, Trick coin}

- Which model generated the data?
  - requires summing out hidden variables
  - requires some form of Occam's razor to trade off complexity with fit to the data.

Coin flipping

Parameter estimation *vs.* Model selection across learning and development

- *Causality:* learning the strength of a relation *vs.* learning the existence and form of a relation
- *Language acquisition:* learning a speaker's accent, or frequencies of different words *vs.* learning a new tense or syntactic rule (or learning a new language, or the existence of different languages)
- *Concepts:* learning what horses look like *vs.* learning that there is a new species (or learning that there *are* species)
- *Intuitive physics:* learning the mass of an object *vs.* learning about gravity or angular momentum
- *Intuitive psychology:* learning a person's beliefs or goals *vs.* learning that there can be false beliefs, or that visual access is valuable for establishing true beliefs

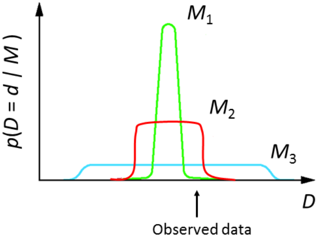
Coin flipping

Comparing simple and complex hypotheses

- $P(\mathrm{H}) = \theta$ is more complex than $P(\mathrm{H}) = 0.5$ in two ways:
  - $P(\mathrm{H}) = 0.5$ is a special case of $P(\mathrm{H}) = \theta$
  - for any observed sequence X, we can choose θ such that X is more probable than if $P(\mathrm{H}) = 0.5$
- How can we deal with this?
  - Some version of Occam's razor?
  - Bayes: automatic version of Occam's razor follows from the "law of conservation of belief".

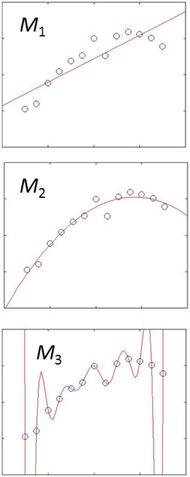
Coin flipping

$$\sum_{\text{all } d \in D} p(D = d \mid M) = 1$$



$$p(D \mid M) = p(y \mid x, M)$$
$$= \int p(y \mid x, \theta, M) p(\theta \mid M) d\theta$$

[assume Gaussian parameter priors, Gaussian likelihoods (noise)]



Concepts and categories

The fundamental problem





We easily recognise all these belonging to a category of "birds", but they aren't in any obvious sense "the same" as each other

On what basis do we decide to refer to these different things as being examples of the same kind of entity?

Concepts and categories

Concepts, Categories and Knowledge

Concepts versus categories

- A "concept" is a mental representation
- A "category" is a group of things (in the world)

The reason for having concepts

- No two things in life are ever identical. All beliefs about the present and the future are necessarily inductions.
- Concepts (and knowledge more generally) exist in order to allow us to function in spite of this.

Concepts - necessity and sufficiency

The classical theory

The theory that most people intuitively have, and that the field began with

Categories are defined by a set of individually necessary and collectively sufficient "features" (i.e., **rules**)

- **Necessity**: If any one of these features is missing, it is definitely not a member of the category
- **Sufficiency**: If all of them are present, then it definitely is a member of the category.

Concepts - necessity and sufficiency

This may work for some concepts!

... But most others are quite difficult to come up with a definition for!

sport

has a ball involved... what about:

or

involves running... what about:

or

involves exertion... what about:

or

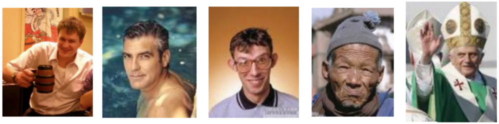
Concepts - graded membership

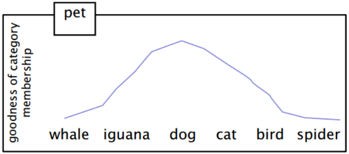
Graded membership

Graded membership: category members vary widely in terms of typicality

bachelor

typical → atypical

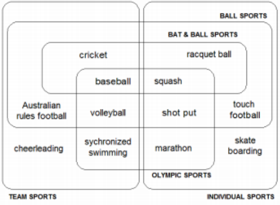




Categories

Family resemblance

A category is a statistical ensemble of features: none are necessary, and no collection is sufficient...



But items that possess more of these features are treated as better members of the category

correlations between "number of category features possessed by an item" and "how typical the item is of the category":

furniture (.88)

vehicle (.92)

fruit (.85)

vegetable (.84)

Simplest distribution=Gaussian

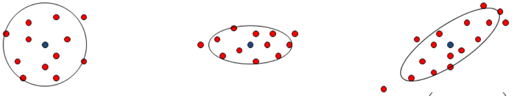
Multivariate Gaussians

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-(x-\boxed{\mu})^2 / 2\boxed{\sigma^2}\}$$

mean   variance/covariance matrix

$$p(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\{-(x-\boxed{\mu})^T \boxed{\Sigma}^{-1} (x-\mu)/2\}$$

quadratic form

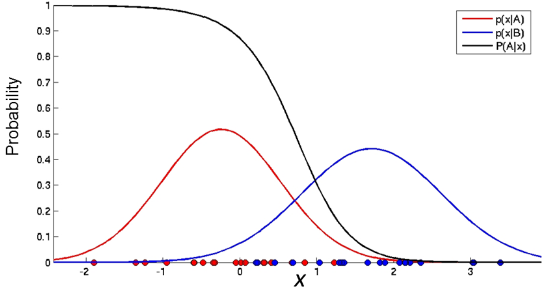


$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 0.25 \end{pmatrix}$ $\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$

Bayesian inference

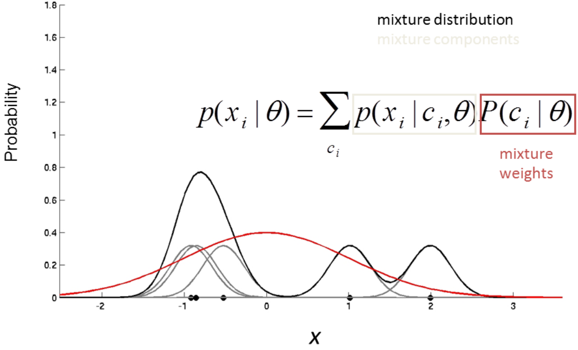
Bayesian inference

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{\sum_c P(x \mid c)P(c)}$$



Mixture of models

Mixture distributions



A chicken and egg problem

If we knew which cluster the observations were from we could find the distributions

this is just density estimation

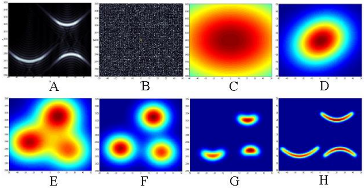
If we knew the distributions, we could infer which cluster each observation came from

this is just categorization

Modeling fields theory and Dynamic logic

Modeling fields theory (MFT): a mathematical apparatus of fuzzy adaptive logic for Aristotelian forms represented as dynamic neural fields, based on dynamic equations which maximize AZ-similarity $AZ - LL = \sum_{i=1}^{n} ll(\mathbf{x}_i) = \sum_{i=1}^{n} \log \sum_{j=1}^{K} l(\mathbf{x}_i|k_j)$

($l(\mathbf{x}_i|k_j)$ - conditional partial similarities, adequate to conditional pdf)

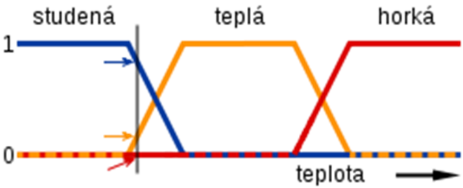


During model estimation, adaptive fuzzy membership functions $f(k_j|\mathbf{x}_i, \mathbf{\Theta}_{k_j})$ are computed from $l(\mathbf{x}_i|k_j)$:

$$f(k_j|\mathbf{x}_i, \mathbf{\Theta}_{k_j}) = l(\mathbf{x}_i|k_j)/l(\mathbf{x}_i) = r_{k_j}.l(\mathbf{x}_i|k_j)/\sum_{k_{j'} \in K} .r_{k_j}.l(\mathbf{x}_i|k_{j'}) \quad (1)$$

Fuzzy logic

- Lower computational complexity
- Adaptive class membership

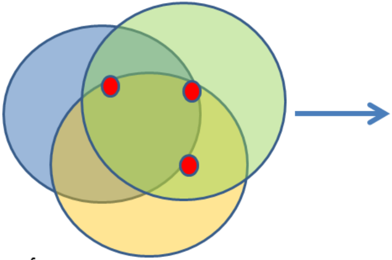


MFT-dynamics

- Dynamic creation of the relationships between internal representations and the world

MFT-dynamics

- Dynamic creation of the relationships between internal representations and the world

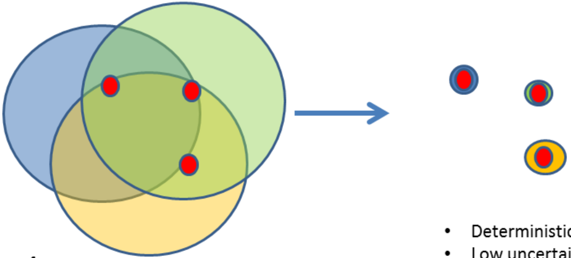


- Fuzzy forms
- Class membership with high fuzziness
- A priori models with very uncertain parameters

- Deterministic concepts
- Low uncertainity about class membership
- Models with fixed parameter values

MFT-dynamics

- Dynamic creation of the relationships between internal representations and the world



- Fuzzy forms
- Class membership with high fuzziness
- A priori models with very uncertain parameters

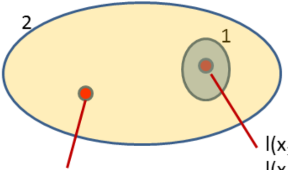
- Deterministic concepts
- Low uncertainity about class membership
- Models with fixed parameter values

- Heterohierarchical structure–many interative loops which include different levels of processing
- In each moment, many concepts (agents, objects) compete for their evidence

MFT-similarity

- Asociation(segmentation) Θ array of input data x with objects= division of inputs to subsets which are related to the given objects

l(n|k) – partial similarity of the point ***n*** with model ***k***
f(k|n) – membership of point ***n*** to model ***k***



$l(x_1|1)=0.9$
$l(x_1|2)=0.2$

$l(x_2|1)=0.01$
$l(x_2|2)=0.3$

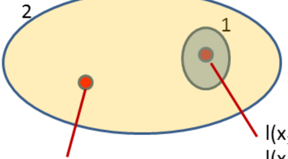
$f(1|x_1)=0.9/(0.9+0.01)=1$
$f(2|x_1)=0.3/(0.3+0.2)=0.6$

MFT-similarity

- Asociation(segmentation) Θ array of input data x with objects= division of inputs to subsets which are related to the given objects

$l(n|k)$ – partial similarity of the point ***n*** with model ***k***

$f(k|n)$ – membership of point ***n*** to model ***k***



$l(x_1|1)=0.9$
$l(x_1|2)=0.2$

$l(x_2|1)=0.01$
$l(x_2|2)=0.3$

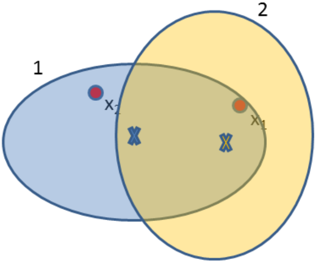
$f(1|x_1)=0.9/(0.9+0.01)=1$
$f(2|x_1)=0.3/(0.3+0.2)=0.6$

Maximalization of complete conditional log-fuzzy similarity:

$$AZ\text{-}LL=\max_{Sk}\Sigma_n \ln\,[\Sigma_k\, f(k|n)]$$

MFT-dynamic equations

1. Initialization of parameters (a priori knowledge)



MFT-dynamic equations

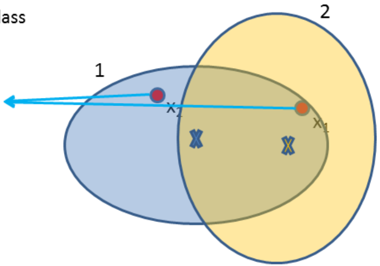
1. Initialization of parameters (a priori knowledge)

2. E – step: compute similarities $l(n|k)$ and class memberships $f(k|n)$

$l(x_2|1)=0.1$
$l(x_2|2)=0.2$
$l(x_1|1)=0.3$
$l(x_1|2)=0.3$

$f(1|x_1)$, $f(1|x_2)$,
$f(2|x_1)$,$f(2|x_2)$

$$l_j(\vec{x}_i|\vec{m}_j,\vec{S}_j) = (2\pi)^{-d/2}\vec{S}_j^{-1/2}\exp[-0.5(\vec{x}_i-\vec{m}_j)^T \vec{S}_j^{-1}(\vec{x}_i-\vec{m}_j)] \quad (5)$$



MFT-dynamic equations

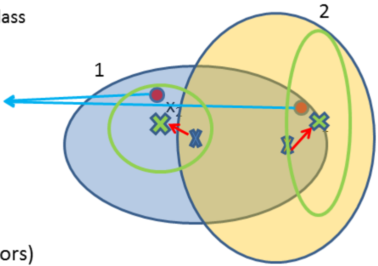
1. Initialization of parameters (a priori knowledge)

2. E – step: compute similarities l(n|k) and class memberships f(k|n)

$l(x_2|1)=0.1$
$l(x_2|2)=0.2$
$l(x_1|1)=0.3$
$l(x_1|2)=0.3$

$f(1|x_1)$, $f(1|x_2)$,
$f(2|x_1)$, $f(2|x_2)$

$$l_j(\vec{x}_i|\vec{m}_j,\vec{S}_j) = (2\pi)^{-d/2}\vec{S}_j^{-1/2}\exp[-0.5(\vec{x}_i-\vec{m}_j)^T \vec{S}_j^{-1}(\vec{x}_i-\vec{m}_j)] \quad (5)$$



3. M-step:

- $d\mathbf{S}_k/dt$ (means, covariances, priors)
- $\mathbf{S}_k(t+dt)=\mathbf{S}_k(t)+d\mathbf{S}_k/dt$

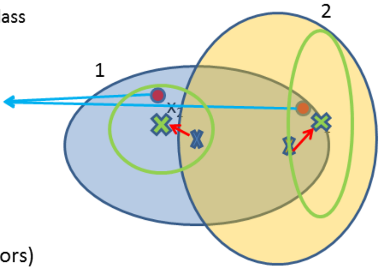
MFT-dynamic equations

1. Initialization of parameters (a priori knowledge)

2. E – step: compute similarities l(n|k) and class memberships f(k|n)

$$\left.\begin{aligned} &l(x_2|1)=0.1\\ &l(x_2|2)=0.2\\ &l(x_1|1)=0.3\\ &l(x_1|2)=0.3 \end{aligned}\right\} \quad \begin{aligned} &f(1|x_1), f(1|x_2),\\ &f(2|x_1), f(2|x_2) \end{aligned}$$

$$\begin{aligned} l_j(\vec{x}_i|\vec{m}_j, \vec{S}_j) \;=\; & (2\pi)^{-d/2} \vec{S}_j^{\;-1/2} \exp[-0.5(\vec{x}_i - \vec{m}_j)^T \\ & \vec{S}_j^{\;-1}(\vec{x}_i - \vec{m}_j)] \qquad (5) \end{aligned}$$

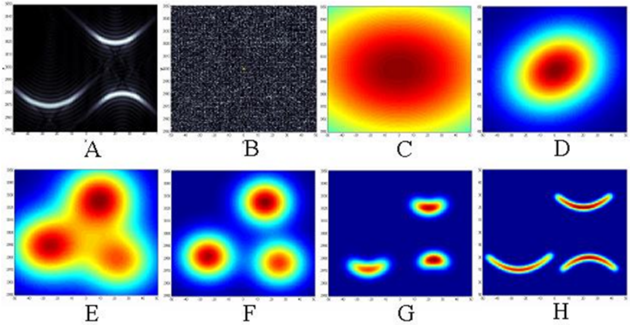


3. M-step:
- $d\mathbf{S}_k/dt$ (means, covariances, priors)
- $\mathbf{S}_k(t+dt)=\mathbf{S}_k(t)+ d\mathbf{S}_k/dt$

4. LL(t)-LL(t-dt) < threshold ?

$$LL(\vec{\theta}) = \sum_{i=1}^{n} \ln\Big(\sum_{j=1}^{K} r_j l_j(\vec{x}_i|\vec{m}_j, \vec{S}_j)\Big)$$

MFT-Evolution of concepts



Paralels - unsupervised clustering
- mixture models
- EM algorithm

Learning - EM algorithm

1, E-step: estimation of all probabilities $f_k(\mathbf{x}_i)$:

$$f_k(\mathbf{x}_i) = \frac{r_k l_k(\mathbf{x}_i|\mathbf{m}_k, \mathbf{S}_k)}{\sum_{k'=1}^{K} r_{k'} l(\mathbf{x}_i|\mathbf{\Theta}_{k'})}$$

2, M-step: choose the parameters which maximizes log-likelihood when the probabilities $f_k(\mathbf{xi})$ are known:

$$r_k = \frac{1}{N}\sum_{i=1}^{N} f_k(\mathbf{x}_i)$$

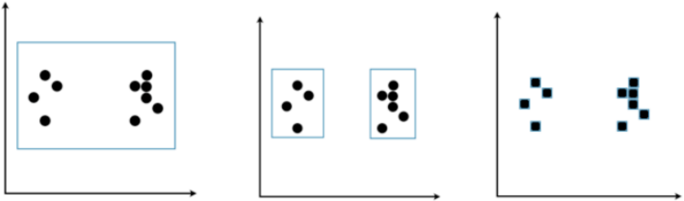
$$\mathbf{m}_k = \frac{\sum_{i=1}^{N} f_k(\mathbf{x}_i)\mathbf{x}_i}{\sum_{j=1}^{N} f_k(\mathbf{x}_j)}$$

$$\mathbf{S}_k = \frac{\sum_{i=1}^{N} f_k(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^T}{\sum_{j=1}^{N} f_k(\mathbf{x}_j)}$$

Some problems:

- Unknown number of clusters - stopping criteria
- Initialization

Hypothesis



How do we evaluate between these hypotheses?

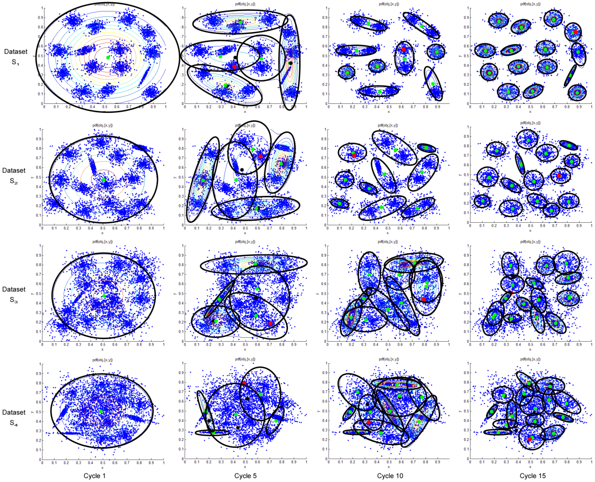


Figure: Evolution of the models during learning

Existing cognitive models based on MFT

basic models of language acquisition and category discrimination

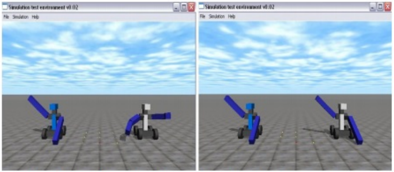


Figure 3 - Teacher and learner before (left) and after (right) the action is learnt.

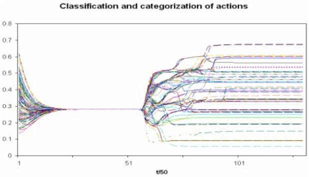
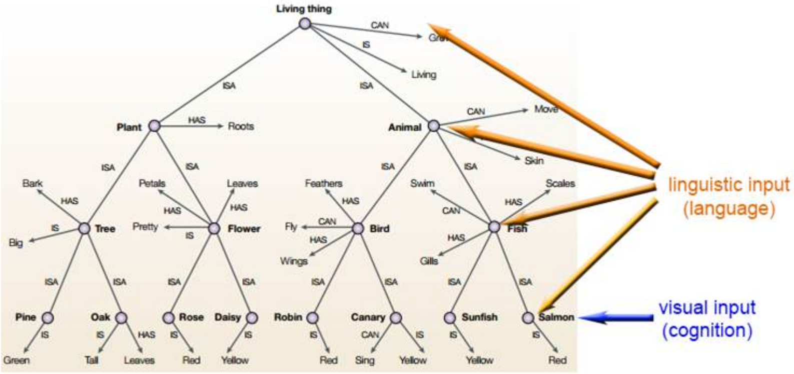


Figure 1 - Time evolution of the fields with 6 features being used as input: 112 different actions

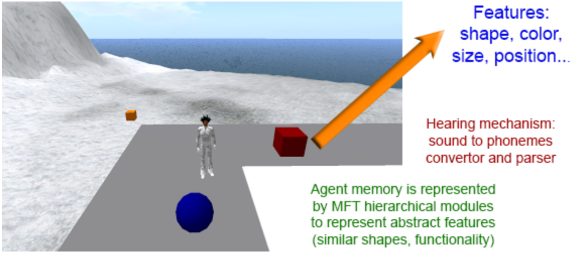
Figure: Tikhanoff 2007 - 6D, 112 actions, nonhierarchical

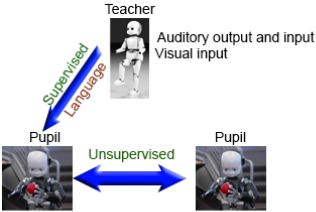
attention, emotional intelligence, integration of language and cognition, object representation and cognition - **mainly theoretical concepts**

Hierarchical Bayes



Future research: Agents in virtual environment





Further reading

- Coursera lecture by Idan Segev: Synapses, neurons and brain www.coursera.org
- Lectures: Computational cognitive science, http://www.compcogscilab.com/courses/ccs-2011/
- Reading list of Bayesian methods: http://cocosci.berkeley.edu/tom/bayes.html
- Ron Sun (2002). The Cambridge Handbook of Computational Psychology
- Lewandowsky, S. and Farrell, S.(2010):Computational Modeling in Cognition: Principles and Practice
- M.D.Lee and E-J.Wagenmakers :Bayesian Cognitive Modeling: A Practical Course (free chapter 1 and 2: https://webfiles.uci.edu/mdlee/BB_Free.pdf)