From Complexity to Intelligence

The fundations of AI

Pierre-Alexandre Murena

A first question

Is there a difference between **Machine Learning** and **Artificial Intelligence** ?

A second question

Do birds fly ?

A third question

Is Michael Jackson dead ?

Table of contents

Table of contents

Analysis of deduction
Deduction examples (1)

1. All men are mortal.
2. Plato is a man.
3. Therefore, Plato is mortal.

Analysis of deduction

Deduction examples (2)

Cauchy-Schwarz inequality

Let $\alpha = (a_1, \ldots, a_n)$ and $\beta = (b_1, \ldots, b_n)$ be two sequences of real numbers. Then :

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geq \left(\sum_{i=1}^{n} a_i b_i\right)^2$$

Proof

Analysis of deduction
Deduction examples (2)

Cauchy-Schwarz inequality

Let $\alpha = (a_1, \ldots, a_n)$ and $\beta = (b_1, \ldots, b_n)$ be two sequences of real numbers. Then :

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geq \left(\sum_{i=1}^{n} a_i b_i\right)^2$$

Proof

For any $t \in \mathbb{R}$:

$$0 \leq \|\alpha + t\beta\|^2 = \|\alpha\|^2 + 2\langle\alpha, \beta\rangle t + \|\beta\|^2 t^2 = P(t)$$

The quadratic polynomial P is positive, so its discriminant is negative :

$$4|\langle\alpha, \beta\rangle|^2 - 4\|\alpha\|^2\|\beta\|^2 \leq 0$$

Analysis of deduction

What is deduction ?

A definition for deductive reasoning

Deductive reasoning is an approach where a set of logic rules are applied to general axioms in order to find (or more precisely *to infer*) conclusions of no greater generality than the premises.

Or, less formally :

- General $\longrightarrow$ Less general
- General $\longrightarrow$ Particular

Table of contents

Incompressibility method

Theorem

For $n > 0$, let $\pi(n)$ designate the number of primes lower than n. Then :

$$\pi(n) \leq \frac{\log n}{\log \log n} - o(1)$$

Table of contents

Table of contents

Limits of deduction

Will it rain today ?

Limits of deduction

We are hardly able to get through one waking hour without facing some situation (e.g. *will it rain or won't it ?*) where **we do not have enough information** to permit deductive reasoning ; but still we must decide immediately.
In spite of its familiarity, the formation of plausible conclusions is a very subtle process.

in [Edwin T. Jaynes, *Probability theory. The logic of science*, Cambridge U. Press, 2003]

Examples of conclusions of non-deductive reasoning

- It will rain today.
- All dogs bark.
- Everybody in this room knows that $1 + 1 = 2$
- The sun always rises in the East.
- Life is not a dream.
- $\ldots$

Inductive reasoning

Definition

Inductive reasoning is an approach in which the premises provide **a strong evidence** for the truth of the conclusion.

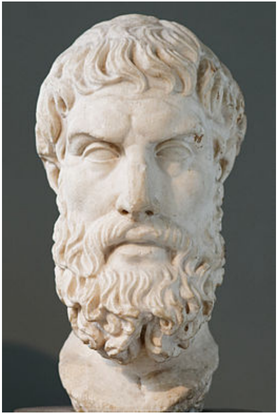
The conclusion of induction is not guaranteed to be true !

Let's play a game !

Table of contents

Philosophical treatment

Epicurus (342-270 B.C.)



Principle of Multiple Explanations : If more than one theory is consistent with the observations, keep all theories.

Philosophical treatment

Sextus Empiricus (160-210)

When they propose to establish the universal from the particulars by means of induction, they will effect this by a review of either all or some of the particulars. But if they review some, the induction will be insecure, since **some of the particulars omitted in the induction may contravene the universal** ; while if they are to review all, they will be **toiling at the impossible**, since the particulars are infinite and indefinite.

Philosophical treatment

Sextus Empiricus (160-210)

When they propose to establish the universal from the particulars by means of induction, they will effect this by a review of either all or some of the particulars. But if they review some, the induction will be insecure, since **some of the particulars omitted in the induction may contravene the universal** ; while if they are to review all, they will be **toiling at the impossible**, since the particulars are infinite and indefinite.

1. **It is impossible to explore all possible situations.**
2. **How is it possible to know that the chosen individuals are representative of the concept ?**

Philosophical treatment

Example of a wrong induction

Do birds fly ?

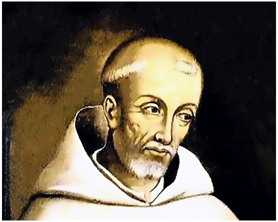
Philosophical treatment

Example of a wrong induction

Do birds fly ?

No !

Philosophical treatment
William of Ockham (1290-1349)



Occam's Razor Principle : Entities should not be multiplied beyond necessity

Philosophical treatment

Thomas Bayes (1702-1761)

Probabilistic point of view on inductive reasoning.

Bayes's Rule : The probability of hypothesis H being true is proportional to the learner's initial belief in H (the *prior probability*) multiplied by the conditional probability of D given H.

Philosophical treatment
David Hume (1711-1766)

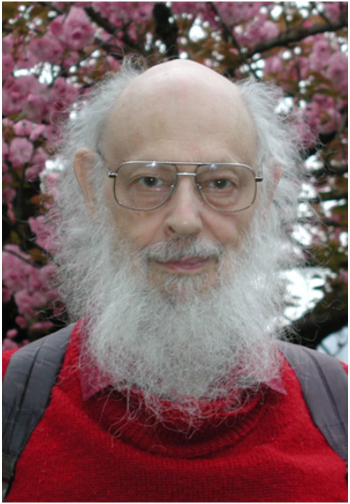
- Causal relations are not not found by deductive reasoning : just because a causal relation is stated in the past does not mean that it will be true in the future.
- Induction is based on a connection between the clauses "I have found that such an object has always been attended with such an effect" and I foresee that other objects which are in appearance similar will be attended with similar effects"
- Deduction cannot justify this connection ; but induction cannot justify it either.

A fundamental question

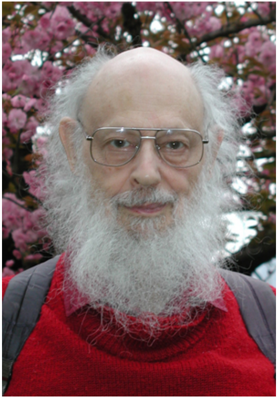
What is the justification for inductive reasoning ?

Table of contents

Ray J. Solomonoff (1926-2009)



General principle



Solomonoff's Lightsaber

Combining the **Principle of Multiple Explanations**, the **Principle of Occam's Razor**, **Bayes Rule**, using **Turing Machines** to represent hypotheses and **Algorithmic Information Theory** to calculate their probability.

Solomonoff's approach step by step

Step 1 : Principle of Multiple Explanations

Principle of Multiple Explanations

All hypotheses explaining the data have to be considered.

Only the hypotheses discarded by the data can be rejected.

Solomonoff's approach step by step

Step 2 : Simplicity Principle

Even if all hypotheses are considered, the most complex hypotheses must be dropped when we find simpler ones.
This idea is basically derived from Occam's Razor.

Solomonoff's approach step by step

Step 3 : Bayes Rule

To neglect complex hypotheses, Bayes rule can be used with high priors for simple hypotheses and low priors for complex hypothes :

$$Pr(H_i|D) = \frac{Pr(D|H_i) \times Pr(H_i)}{Pr(D)}$$

where the value of $Pr(H_i)$ is low if H_i is complex and high if H_i is simple.

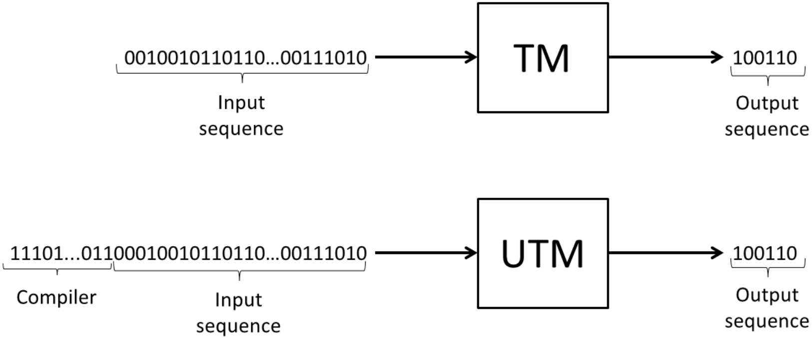
Solomonoff's approach step by step

Step 4 : Encoding hypotheses with Universal Turing Machines

- Data D are encoded as a sequence over a finite alphabet $\mathcal{A}$ (for example binary alphabet $\mathcal{A} = \{0, 1\}$).
- Hypotheses are processes : hence, they can be represented as Turing Machines (TM).
- Hypotheses are represented as input sequences of Universal Turing Machines (UTM).
- The set of possible inputs of a UTM corresponds to the set of hypotheses.

Solomonoff’s approach step by step

Step 4 : Encoding hypotheses with Universal Turing Machines



Solomonoff's approach step by step

Step 5 : Universal prior

The priors are chosen to be :

$$Pr(H_i) = 2^{-K(H_i)}$$

Solomonoff's Induction

1. Run any possible hypothesis H_i on the UTM :
   - If H_i produces the data D :
     1.1 Accept the hypothesis : $Pr(D|H_i) = 1$
     1.2 Calculate Kolmogorov complexity of H_i : $K(H_i)$
     1.3 $Pr(H_i) = 2^{-K(H_i)}$
   - Otherwise : Discard the hypothesis : $Pr(D|H_i) = 0$
2. $H^* = \arg\max_{H_i}\{Pr(H_i) \times Pr(D|H_i)\}$

This problem is intractable !

So what ?

The strongest result of this theory is that **a universal distribution can be used as an estimator *for all priors*.**

So what ?

The strongest result of this theory is that **a universal distribution can be used as an estimator *for all priors*.**

Theorem

If μ is the *concept* computable measure and the conditional semi-measure $\mu(y|x)$ is defined by $\mu(y|x) = \frac{\mu(xy)}{\mu(x)}$.
Let $\mathcal{B}$ be a finite alphabet and x a word over $\mathcal{B}$. The summed expected squared error at the n-th prediction is defined by :

$$S_n = \sum_{a \in \mathcal{B}} \sum_{l(x)=n-1} \mu(x) \left(\sqrt{\mathbf{M}(a|x)} - \sqrt{\mu(a|x)} \right)^2$$

Then $\sum_n S_n \leq K(\mu) \log(2)$

Table of contents

Induction and bias

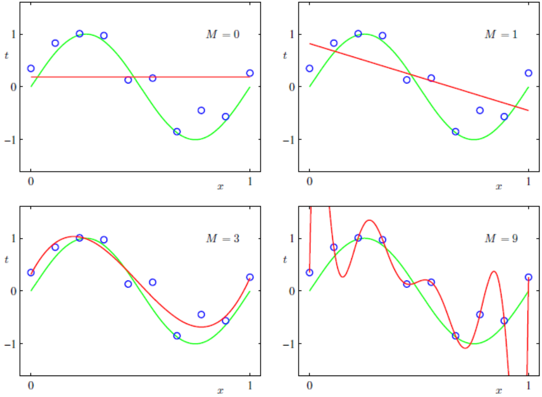
Remarks

1. An inductive algorithm is **biased** toward a given class of problems.
2. The performance of an algorithm is **necessarily** relative to a class of problems.
3. Induction does not create information : it only *transforms* a prior information contained in the algorithm.

Two classes of bias

1. **Representation bias** : a bias on the form of the concept
2. **Research bias** : a bias on how the concept is searched

Example : regression



Which model would you choose ?

Minimum Description Length Principle

The best theory to describe observed data is the one which minimizes the sum of the description length (in bits) of :

- the theory description
- the data encoded from the theory

Inductive principle
Minimum Description Length Principle

$$\hat{H} = \arg\min_{H_i} \quad C(H_i) + C(D|H_i)$$

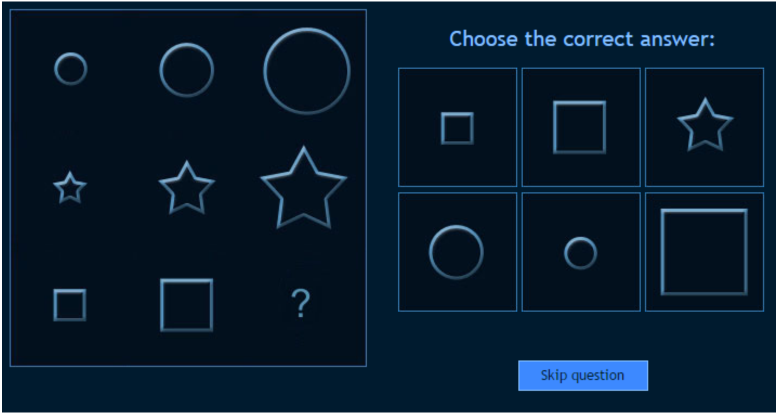
or

$$\hat{H} = \arg\min_{H_i} \quad K(H_i) + K(D|H_i)$$

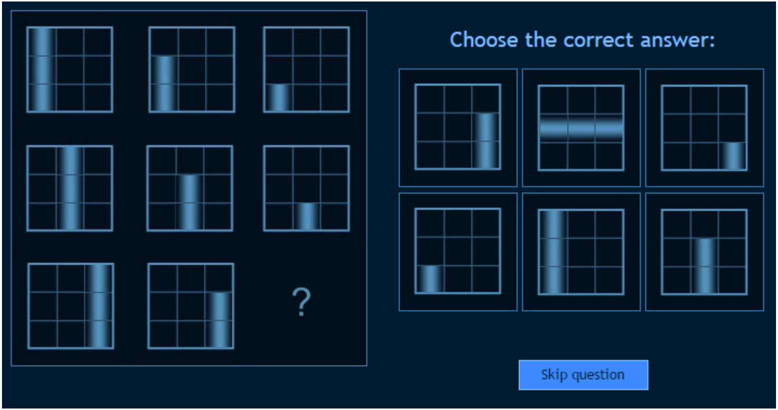
Table of contents

Table of contents

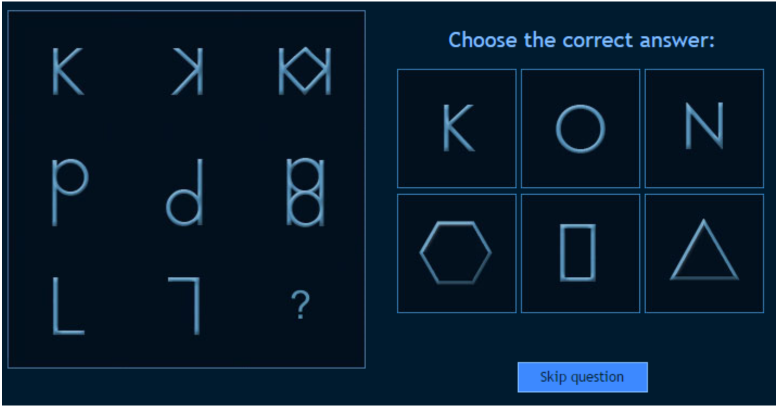
IQ tests



IQ tests



IQ tests



What to say about these problems ?

- Inductive problems
- Repetition of *similar* structures
- A question is asked about a missing state
- Search of regularity

An attempt of a definition

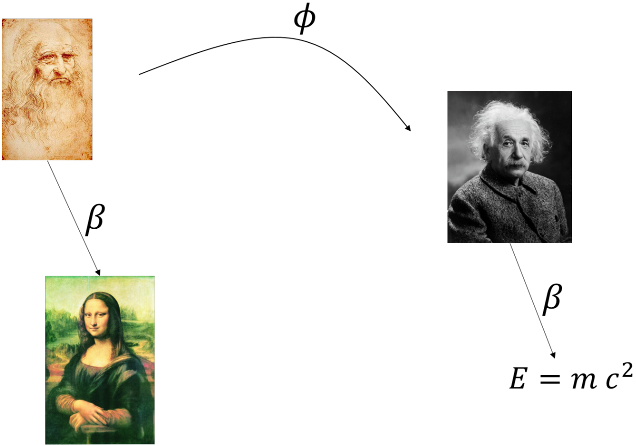
A definition (K. Holyoak, 2004)

Two situations are analogous if they share a common pattern of relationships among their constituent elements even though the elements themselves differ across the two situations.

Proportional analogy

Proportional Analogy concerns any situation of the form “A is to B as C is to D”.

Some examples



Some examples

- Gills are to fish as lungs are to man.
- Emmanuel Macron is to France as Vladimir Putin is to Russia
- Donald Trump is to Barack Obama as Barack Obama is to George Bush
- 37 is to 74 as 21 is to 42
- The sun is to Earth as the nucleus is to the electron

Question

Why is analogical reasoning so important ?

Why is analogical reasoning so important ?

Analogies in other domains

- **Mathematics and science** : used to discover new concepts, or to generalize notions to other domains.
- **Justice** : use of relevant past cases
- **Art** : metaphors, parody, pastiche...
- **Advertising** : use of ground knowledge to influence people
- **Humor** : jokes are often based on inappropriate analogies

Three axioms [Lepage 2003]

The following axioms are commonly accepted (but not always) :

1. **Symmetry :** $A : B :: C : D \Leftrightarrow C : D :: A : B$
2. **Exchange :** $A : B :: C : D \Leftrightarrow A : C :: B : D$
3. **Determinism :** $A : A :: B : x \Rightarrow x = B$ and $A : B :: A : x \Rightarrow x = B$

Table of contents

Comparing two analogies

Analogy 1

The *Mona Lisa* is to Da Vinci what $E = mc^2$ is to Albert Einstein.

Analogy 2

The *Mona Lisa* is to Da Vinci what *A Unified Field Theory Based on the Riemannian Metric and Distant Parallelism* is to Albert Einstein.

Question

How to characterize a good analogy ?

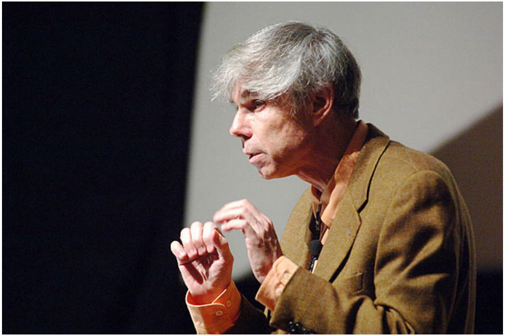
Relevance measure

A relevance measure has to be found to disqualify properties of little interest. Several criteria may be considered to measure relevance of a mapping :

- Number of common properties
- Abstraction level of the shared properties
- Structural alignment
- Pragmatic centrality (mappings are better when the goals expressed in the source and target are the same)
- Representational distortion
- Total description length (inspired by [Cornuéjols 1998])

Table of contents

Douglas Hofstadter (1945-now)



“*We are trying to put labels on things by mapping situations that we have encountered before. That to me is nothing but analogy.*”

A micro-world

- Alphabet $\Sigma = \{A, B, C, \ldots, Z\}$
- Elements of the analogy are words over Σ

A micro-world

- Alphabet $\Sigma = \{A, B, C, \ldots, Z\}$
- Elements of the analogy are words over Σ

Advantages of this micro-world

- Simplicity of the problems
- Human readibility
- Implies simple operations (predecessor, successor, add, remove, increment...)
- Covers a wide range of problems

Examples

- ABC : ABD : : IJK : x
- RST : RSU : : RRSSTT : x
- ABC : ABD : : BCA : x
- ABC : ABD : : AABABC : x
- IJK : IJL : : IJJKKK : x
- ...

Table of contents

Minimum Description Length Principle

... one more time ...

MDL Principle

The best theory to describe observed data is the one which minimizes the sum of the description length (in bits) of :

- the theory description
- the data encoded from the theory

Let's try to apply the MDL Principle to analogy reasoning !

Mathematical model

Consider the analogy equation $U : V :: W : x$

$$C(M) + C(D|M)$$

- D correspond to the data : $D = \langle U, V, W \rangle$
- M is a *global* model used to describe the data :
  - M can be the description of the data
  - M can be a description of a process generating data

We propose to find assumptions to simplify the complexity term

Hypothesis 1 : Separation of the models

The model M is split in two parts : a source model M_S and a target model M_T.

- $C(M) \leq C(M_S, M_T)$
- $C(D|M) = C(D|M_S, M_T)$

Hypothesis 2 : Model transfer

The target model is described with the help of the source model.

- $C(M) \leq C(M_S) + C(M_T|M_S)$
- $C(D|M) \leq C(D|M_S, M_T)$

Simplification of the MDL
Separation between source and target data

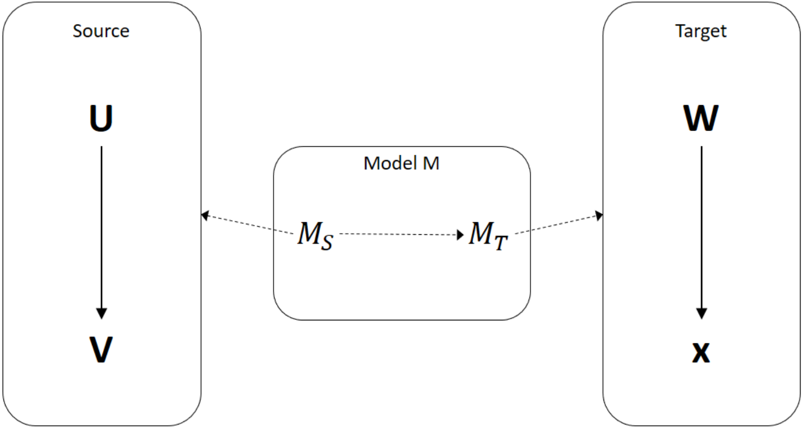
Hypothesis 3 : Separation between source and target data
The source and target data are described with the help of their corresponding model only.

- $C(M) \leq C(M_S) + C(M_T|M_S)$
- $C(D|M) \leq C(D_S, D_T|M_S, M_T) = C(D_S|M_S) + C(D_T|M_T)$

Important remark
The chosen simplification does not imply a transfer directly on the data, but on the models generating the data.

Simplification of the MDL

Summary



And now ?

Two approaches

- Find the X minimizing $C(M_S) + C(U, V) + C(M_T|M_S) + C(W, x)$
- Find the target model minimizing

$$C(M_S) + C(U, V) + C(M_T|M_S) + C(W)$$

and infer x from M_T and W

How to describe data with a model ?
New assumptions

Hypothesis 4 : Prevalence of inputs

Inputs are used to describe outputs.

- $C(M) \leq C(M_S) + C(M_T|M_S)$
- $C(D|M) \leq C(D_S|M_S) + C(D_T|M_T) \leq C(U|M_S) + C(V|M_S, U) + C(W|M_T) + C(x|M_T, W)$

How to describe data with a model ?

New assumptions

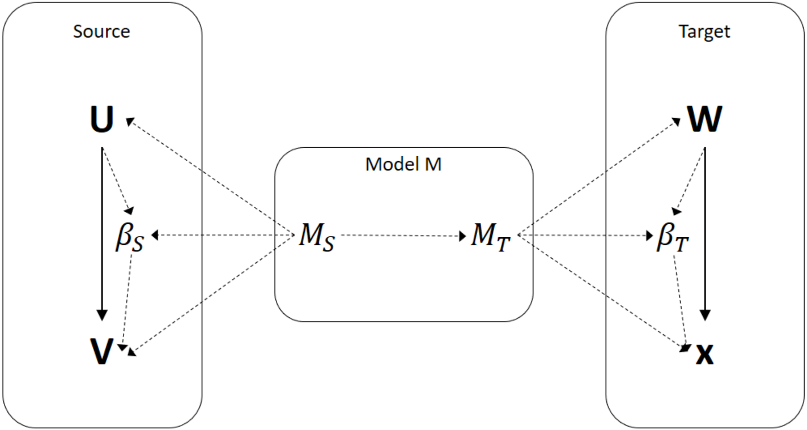
Hypothesis 5 : Decision function

For both source and target, there exists a decision function (resp. β_S and β_T).

- $C(M) \leq C(M_S) + C(M_T|M_S)$
- $C(V|M_S, U) \leq C(V, \beta_S|M_S, U) \leq C(\beta_S|M_S, U) + C(V|M_S, U, \beta_S)$
- $C(x|M_T, W) \leq C(\beta_T|M_T, W) + C(x|M_T, W, \beta_T)$

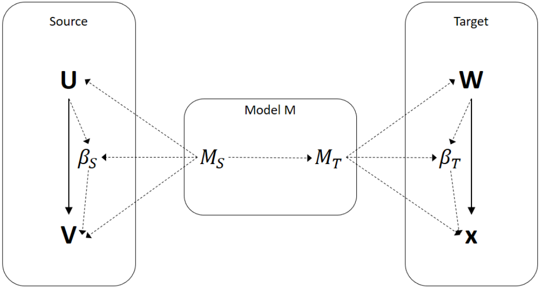
Simplification of the MDL

Summary



Final equation

$$C(M_S) + C(U|M_S) + C(\beta_S|M_S, U) + C(V|M_S, U, \beta_S)$$
$$+ C(M_T|M_S) + C(W|M_S) + C(\beta_T|M_T, W) + C(x|M_T, W, \beta_T)$$



Application : An example

Calculate **manually** the complexity of the proportional analogy :

ABC : ABD : : IJK : x

for the following values of **x** : IJL, ABD, IJK.

Application : An example

Calculate **manually** the complexity of the proportional analogy :

ABC : ABD : : IJK : x

for the following values of **x** : IJL, ABD, IJK.

Why not, but on which machine ?

Application : An example

Choice of the UTM

- Orientation ($\rightarrow$ or $\leftarrow$) : 1 bit
- Cardinality n : $\log(1 + n)$ bits
- Length l : $\log(1 + l)$ bits
- Type : 3 bits

- A letter : 5 bits
  Example : C('g') = 5

- A string : C(orientation) + Σ C(elements)
  Example : C('fci') = 1 + 3 $\times$ 5 = 16 bits

Application : An example

Choice of the UTM

- Ensemble : C(type of elements) + C(cardinality) + Σ C(elements)
  Example : C({ 'k', 'f', 'c' }) = 3 + 2 + 3 $\times$ 5 = 20 bits
- Group : C(type of elements) + C(number of elements) + Σ C(elements)
  Example : C({ 'u r l' }) = 3 + 2 + 3 $\times$ 5 = 20 bits
- Sequence : C(orientation) + C(type) + C(succession rule) + C(length) + C(first *or* last element)

Application : An example

Choice of the UTM

Example : length of the sequence 'abc'

- Orientation $\longrightarrow$: C(orientation) = 1
- Type : letters : C(type) = 3
- Succession rule : function taking a letter as input (C(type=letter) = 3 bits) and taking its first successor (C(successor) = 1)
  Hence C(succession rule) = 4 bits
- Length 3 : C(length) = 2
- First element 'a' : C(first element) = 5 bits

Hence C(sequence 'abc') = 1 + 3 + 4 + 2 + 5 = 15 bits

Application : An example

The models

ABC : ABD : : IJK : x

- Model 1 : Generate a sequence of 3 letters and replace the third element by its successor (solution : IJL)
- Model 2 : Generate a sequence of 3 letters and replace the last element by its successor (solution : IJL)
- Model 3 : Return ABD (solution : ABD)
- Model 4 : Generate a sequence of 3 letters and change the 'c' into a 'd' (solution IJK)

A related approach

$$C(M_S) + C(M_T|M_S)$$

```
// ABC : ABD :: IJK : IJL
let(alphabet, shift, ?, sequence, 3),
  let(mem,, ?, next_block, mem,, ?, last, increment),
  mem,,, next_block, mem,, 8;
```

$$C(X_S|M_S) + C(Y\_S|M_S)$$

$$C(X_T|M_T) + C(Y_T|M_T)$$

Description length as relevance index

Experimental validation

Problem	Solution	Proportion	Complexity
IJK 16.0 ± 0.085 s	IJL	93%	37
	IJD	2.9%	38
BCA 21.7 ± 0.12 s	BCB	49%	42
	BDA	43%	46
AABABC 23.8 ± 0.12 s	AABABD	74%	33
	AACABD	12%	46
IJKLM 24.7 ± 0.22 s	IJKLN	62%	40
	IJLLM	15%	41
123 6.39 ± 0.074 s	124	96%	27
	123	3%	31
KJI 18.6 ± 0.13 s	KJJ	37%	43
	LJI	32%	46
135 9.93 ± 0.10 s	136	63%	35
	137	8.9%	37
BCD 21.9 ± 0.30 s	BCE	81%	35
	BDE	5.9%	44

Problem	Solution	Proportion	Complexity
IJJKKK 13.7 ± 0.11 s	IJJLLL	40%	52
	IJJKKL	25%	53
XYZ 11.2 ± 0.093 s	XYA	85%	40
	XYZ	4.4%	34
122333 10.0 ± 0.098 s	122444	40%	56
	122334	31%	49
RSSTTT 10.4 ± 0.072 s	RSSUUU	41%	54
	RSSTTU	31%	55
IJJKKK 8.67 ± 0.071 s	IJJLLL	41%	52
	IJJKKL	28%	53
AABABC 12.2 ± 0.12 s	AABABD	72%	33
	AACABD	12%	46
MRRJJJ 22.1 ± 0.18 s	MRRJJK	28%	64
	MRRKKK	19%	65
147 13.6 ± 0.20 s	148	69%	36
	1410	10%	38

68 participants (36 female), ages 16-72

Table of contents

Conclusion

What to remember ?

- Difference between deduction and induction
- Non-universality of inductive reasoning
- Toward a universal solution ? Solomoff's theory of induction
- What is analogy reasoning ?
- Using complexity to solve analogy equations ?

What next ?

- Consider a large class of inductive problems : machine learning
- Apply MDL to machine learning problems

Licence de droits d'usage

Contexte public } sans modifications