Occam's data-compressor

Justifying simplicity via algorithmic information theory

Tom Sterkenburg
(Algorithms & Complexity)

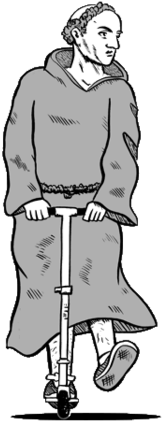
CWI
Centrum Wiskunde & Informatica

university of groningen

faculty of philosophy

CWI Scientific Meeting
May 29, 2015

The principle of simplicity in science



- **Occam's razor**: thou shalt not make things unnecessarily complicated!

The principle of simplicity in science

Occam's data-compressor

Tom Sterkenburg (Algorithms & Complexity)

Occam's razor

Compression

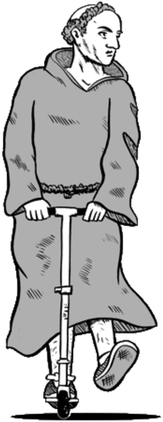
The argument

Bayesian prediction

The argument recast

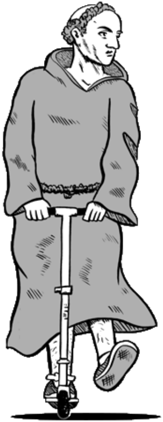
- **Occam's razor**: thou shalt seek simplicity!

The principle of simplicity in science



- **Occam's razor**: thou shalt seek simplicity!
  - Why? Can we *justify* that?

The principle of simplicity in science



- **Occam's razor**: thou shalt seek simplicity!

  - Why? Can we *justify* that?
  - "Simplicity"? Can we *measure* that?

Algorithmic information theory

simplicity $\sim$ **compressibility**

Prediction by datacompression

- Assign higher probability to more compressible future data.

- Let C be a computer.
- Let σ be a data sequence, and ρ be the shortest C-instruction for σ.
- Data sequence σ is more *compressible* as instruction ρ is *shorter*.
- Let $\lambda(\rho)$ be the *uniform probability* of ρ: it is higher as ρ is shorter.

Prediction by datacompression

- Assign higher probability to more compressible future data.

  - Let C be a computer.
  - Let σ be a data sequence, and ρ be the shortest C-instruction for σ.
  - Data sequence σ is more *compressible* as instruction ρ is *shorter*.
  - Let $\lambda(\rho)$ be the *uniform probability* of ρ: it is higher as ρ is shorter.

Definition (Solomonoff, 1964)

The algorithmic probability via C of data sequence σ is given by

$$Q_C(\sigma) = [\text{imposing definition}] \approx \lambda(\rho).$$

  - Let's refer to these probability distributions as the **predictors of type** $\mathcal{Q}$.

The argument to justify Occam's razor

1. predictors of type $\mathcal{Q}$ possess a simplicity preference
2. predictors of type $\mathcal{Q}$ are reliable (in essentially every case)

The argument to justify Occam's razor

1. predictors of type $\mathcal{Q}$ possess a simplicity preference

2. predictors of type $\mathcal{Q}$ are reliable (in essentially every case)

$\therefore$ predictors preferring simplicity are (essentially always) reliable

The argument to justify Occam's razor

1. predictors of type $\mathcal{Q}$ possess a simplicity preference

2. predictors of type $\mathcal{Q}$ are reliable (in essentially every case)

$\therefore$ justification of Occam's razor

Too good to be true?

Too good to be true?

1. predictors of type $\mathcal{Q}$ possess a simplicity preference

- Do they really?

Too good to be true?

1. predictors of type $\mathcal{Q}$ possess a simplicity preference

- Do they really?

Theorem

For every nonatomic computable measure μ,

$$\mathcal{Q}^{\mu} = \mathcal{Q}.$$

A different pair of glasses

- Let's recast the argument in terms of **Bayesian prediction**.

A different pair of glasses

- Let's recast the argument in terms of **Bayesian prediction**.

- Define *prior distribution* W over selected hypothesis class $\mathcal{H}$.
- The hypothesis class embodies our *inductive assumptions*.

A different pair of glasses

- Let's recast the argument in terms of **Bayesian prediction**.

- Define *prior distribution* W over selected hypothesis class $\mathcal{H}$.
- The hypothesis class embodies our *inductive assumptions*.

Definition

The Bayesian mixture distribution $P_W^{\mathcal{H}}$ *via prior* W *on hypothesis class* $\mathcal{H}$ *is given by*

$$P_W^{\mathcal{H}}(\sigma) := \sum_{P \in \mathcal{H}} W(P)P(\sigma).$$

A very special inductive assumption

- Consider the hypothesis class $\mathcal{H}^{\mathrm{eff}}$ of effectively approximable or simply *effective* hypotheses (Zvonkin & Levin, 1970).
- The Bayesian predictors with an effective prior over this hypothesis class are the predictors *operating under the inductive assumption of effectiveness*.
- ▷ We'll call these the **predictors of type** $\mathcal{R}$.

It's the same thing!

- The predictors of type $\mathcal{Q}$ and the predictors of type $\mathcal{R}$ **are the same**!

It's the same thing!

- The predictors of type $\mathcal{Q}$ and the predictors of type $\mathcal{R}$ **are the same**!

Theorem (Wood, Sunehag and Hutter, 2013)

$$\mathcal{Q} = \mathcal{R}.$$

  - The choice of computer is the choice of Bayesian prior over $\mathcal{H}^{\text{eff}}$.
- The defining simplicity preference of predictors of type $\mathcal{Q}$ is the adherence to the inductive assumption of effectiveness.

Recasting the argument

original

1. predictors of type $\mathcal{Q}$ possess a simplicity preference
2. predictors of type $\mathcal{Q}$ are reliable (in essentially every case)

recast

1. predictors of type $\mathcal{R}$ operate under the inductive assumption of effectiveness
2. predictors of type $\mathcal{R}$ are reliable under the assumption of effectiveness

Taken together

original predictors that possess a simplicity preference are (essentially always) reliable

recast predictors that implement the inductive assumption of effectiveness are reliable *under the same assumption* of effectiveness

Taken together

recast predictors that implement the inductive assumption of effectiveness are reliable *under the same assumption* of effectiveness

- This is just an instance of Bayesian *consistency*, that holds for every (countable) hypothesis class.

The (bittersweet) conclusion

The central element of algorithmic information theory is the constraint of *effectiveness*. Effectiveness is interesting as a highly general inductive assumption. Arguably, the predictors employing this assumption are universally optimal: they represent the best we can possibly do. As such, the theory presents a limit case to optimal prediction – an important topic both in philosophy and statistics.

But algorithmic information theory cannot do the job of justifying Occam's razor.

The (bittersweet) conclusion

The central element of algorithmic information theory is the constraint of *effectiveness*.
Effectiveness is interesting as a highly general inductive assumption. Arguably, the predictors employing this assumption are universally optimal: they represent the best we can possibly do. As such, the theory presents a limit case to optimal prediction – an important topic both in philosophy and statistics.

But algorithmic information theory cannot do the job of justifying Occam's razor.

www.cwi.nl/~tom