AN ERGODIC APPROACH TO OCCAM'S RAZOR

Aidan Rocke
aidanrocke@gmail.com

March 4, 2021

ABSTRACT

An algorithmic Occam's razor may be derived using the Expected Kolmogorov Complexity of a discrete random variable. This derivation is based upon a combination of Bayesian and ergodic perspectives that are both necessary and sufficient in the context of the scientific method as it is applied in the natural sciences.

1 A Bayesian perspective on computation

As computations are observer-dependent, computation is fundamentally Bayesian. In particular, the uncertainty associated with a discrete random variable is defined with respect to a predictive model. It follows that Occam's razor for a discrete random variable X is a measure of epistemic uncertainty or the memory requirements of the ideal model for predicting the behaviour of X.

In this setting, the minimum description length of X is given by the most parsimonious model Ω for predicting the behaviour of X:

$$\mathbb{E}[K(X)] = \mathbb{E}[-\ln P(X|\Omega)P(\Omega)] = H(X|\Omega) + H(\Omega) \tag{1}$$

where $H(\Omega)$ is the complexity of the model Ω and $H(X|\Omega)$ is the expected information gained by Ω from observing X.

The reason why the expression in (1) has a probabilistic representation is that the behaviour of a discrete random variable is described by its probability distribution. From an ergodic perspective, these probabilities also have a natural frequentist interpretation.

2 An ergodic analysis

Given that an event that occurs with frequency p generally requires modelling a sequence of length $\sim \frac{1}{p}$, in order to encode the structure of such an event, a machine would generally need a number of bits proportional to:

$$\ln(\frac{1}{p}) = -\ln(p) \tag{2}$$

But, how should we define the constant of proportionality? If we assume that the memory of the machine is finite and that the data-generating process is ergodic, an optimal encoding would use the expected number of bits:

$$-p \cdot \ln(p) \tag{3}$$

in order to encode an event that occurs with frequency p.

Regarding the assumptions, I may make a couple remarks. First, the ergodic assumption is equivalent to the premise that scientific experiments are repeatable in the natural sciences. Second, all Turing machines have finite memory.

3 Application: Ockham's razor and asymptotic incompressibility

Given a binary sequence $X_N = \{x_i\}_{i=1}^N$, we say that X_N is *asymptotically incompressible* if given the subsequence X_k and $N >> k$ on average we would not profit by gambling on the $N - k$ terms in X_N based on the partial knowledge provided by X_k. If X_N satisfies these assumptions then Ockham's razor applied to X_N scales as follows:

$$\mathbb{E}[K(X_N)] \sim N \tag{4}$$

which means that an effective betting strategy has infinite sample complexity and therefore the average size(in bits) of the smallest approximately correct predictive model, found using machine learning methods, scales with N.

4 Discussion

I would like to point out that mainstream Algorithmic Information Theory insists that the Expected Kolmogorov Complexity of a discrete random variable equals its Shannon entropy:

$$\mathbb{E}[K(X)] = H(X) \tag{5}$$

by appealing to contrived mathematical notions such 'Universal distributions' which may 'solve' the No Free Lunch problem [1,2]. What we may infer from this is that the mainstream theory is both incorrect and incomplete.

The implicit error in (4) is to assume that computations are observer-independent. Since we can't remove the observer in (1), we have:

$$X = \Omega \implies \mathbb{E}[K(\Omega)] = H(\Omega) \tag{6}$$

So information that is truly incompressible is self-referential.

References

[1] Peter Grünwald and Paul Vitányi. Shannon Information and Kolmogorov Complexity. 2010.

[2] Tom Everitt, Tor Lattimore, Marcus Hutter. Free Lunch for Optimisation under the Universal Distribution. 2016.

[3] Schnorr, C. P. (1971). "A unified approach to the definition of a random sequence". Mathematical Systems Theory.

[4] Shai Shalev-Shwartz and Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. 2014.